**Review**

# Outline of disease gene hunting approaches in the Millennium Genome Project of Japan

By Teruhiko YOSHIDA*),†) and Kimio YOSHIMURA**)

(Communicated by Takashi SUGIMURA, M. J. A., Feb. 12, 2003)

**Abstract:** To capture the potentially enormous and vital opportunity in the post sequence era of human genome research, the Japanese government on December 19, 1999 drew up the basic outlines of the 5-year Millennium Genome Project (MGP). The "Disease Gene" Team of MGP aims to establish genome-based personalized medical care and to seed a revolutionary drug development inspired by the new information and technology, which the genome research provides. Five disease categories have been chosen as project targets: dementia, cancer, diabetes, hypertension and asthma, which are manifesting a major impact on the health and welfare of the rapidly graying Japanese society. These so-called "common diseases" are etiologically multifactorial, and the genetic components, if any, are considered polygenic with relatively high disease allele frequencies in the population. In the first year, 2000, a consortium or "Subteam" was established for each disease category, and several hypothesis-driven candidate gene approaches were launched. In 2001, a complementary strategy of disease gene hunting, a statistics-based genome-wide approach, was initiated. The Disease Gene Team decided to employ a genome-wide, gene-based SNP scan through close collaboration with another arm of MGP, the "Human Genome Variation" Team, which discovered 194,393 SNPs as of December 25, 2002 from genome of the Japanese people. Each Subteam reached a consensus regarding eligibility criteria and research protocols, which were reviewed by the institutional review boards in accordance with the new Ethics Guidelines For Human Genome/Genetic Analysis Research promulgated by the government on April 1, 2001. The Subteams organized a multi-institutional consortium, which in 2001 collaborated in the collection of quality germline DNA samples and clinical and life-style related information. In 2002, the third year of MGP, the genome scan started at two typing centers using the high-throughput SNP typing system established by RIKEN. The premise and prospect of the approach will be discussed.

**Key words:** Millennium Genome Project; cancer; microarray; SNP; genome scan; association study.

**Project overview.** *MGP and its goals*. On December 19, 1999, the Japanese government announced a so-called "Millennium Project" which aims to create the basic core for Japanese Science and economy in the 21st Century (official documents in Japanese are available at: http://www.kantei.go.jp/jp/mille/). The 120.6 billion Yen Project (US$ 1 billion, in FY 2000) was planned to target three major categories: information infrastructure, a graying society, and the environment. The human genome research and its clinical and industrial applications are expected to play crucial roles in addressing the issues of a graying society, and this part of the Millennium Project, 64 billion Yen in FY 2000, is called "Millennium Genome Project" (MGP). MGP is positioned in the so-called "post-sequencing era of human genome research". Soon after the rise of the molecular biology of human diseases in the early 1980s, it was widely recognized that an approach starting from an individual gene is not sufficient to understand most "common diseases" such as cancer, diabetes or cardiovascular diseases. Unlike monogenic diseases showing a classical mendelian inheritance, multiple genes are

*) Genetics Division, National Cancer Center Research Institute, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan.

**) Cancer Information and Epidemiology Division, National Cancer Center Research Institute, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan.

†) Correspondence to: T. Yoshida, e-mail: tyoshida@ncc.go.jp

obviously involved in the common diseases, and the genetic factors may interact significantly with environment/life-style factors and aging. This notion motivated the Human Genome Project, so that the researchers have the entire human nucleotide sequences as a guide to explore polygenic diseases. A draft-quality sequence was published in February 2001, and the "finished" sequence has now reached 95.8% of the human genome as of January 5, 2003.[1] The race to capture the enormous medical potentials of the human genome sequence information has already started as a post-sequencing era of genome research.

The goals of the Human Genome MGP have been described in a December 19, 1999 document by the then Prime Minister as follows: "Personalized medicine will be introduced, and revolutionary drug development will be launched, both based on the elucidation of the disease related genes of the major diseases of the elderly, such as dementia, cancer, diabetes, hypertension. Regenerative medicine, which utilizes the self-repair function of the body and is free from rejection reactions, will also be applied for tissues such as bone and blood vessels, based on the elucidation of the developmental and other functions of the organisms. The target date for Project completion is March 2005". The two keywords of MGP are therefore personalized medicine and new therapeutics development. Both, of course, have been the core agenda of medicine throughout its history, but what is newly expected for MGP is that genome research may bring personalized medicine and drug development to a new horizon, which benefits not only the elderly but also all segments of society.

*Organization.* To address these two tasks of personalized medicine and new therapeutics development, a research organization has been deployed as shown in Fig. 1. Headquarters of MGP is the "Evaluation and Advisory Board" formed of 12 senior leaders in bioscience and bioindustry. The business office of the Board is the Cabinet Office of the government. The Board is actively involved in the whole aspect of the Project, including setting the direction of the research, approval of research strategy, and the evaluation of research progress, and orders termination of the project when necessary. Thus, MGP is clearly a government-led "top-down" mega initiative. There may be at least two points to consider in light of the science policy of today and of the future. First, one of the major impetuses which spurred the Japanese government to launch MGP was the keen recognition of the enormous pharmaceutical and other medical industrial potentials of genome-based medical research in the post-draft sequencing era. The vital link between science and technology, intellectual properties, national interests and welfare level of the nation is obviously the expectation and justification of the MGP. Secondly, however, big top-down "project" type studies should be only a part of all research conducted in the country. Individual research initiated by each investigator's idea and interests is the bottom line of the science and culture of the country, and is, after all, the fundamental of the national well-being in the long-term, too. The Council for Science and Technology Policy, Cabinet Office (http://www8.cao.go.jp/cstp/), which is the major organization promoting big project type research, is also aware of the importance of the balance between a government-initiated research project and investigator-initiated individual research. Thus an important issue is what proportion of the public sector budget should be allocated as project type research and to which target. The active voices and decisions of senior leading scientists are increasingly crucial in conceptualizing and navigating national science policy.

MGP is further sub-divided into five Project Teams: Human Genome Variation, Disease Gene, Bioinformatics, Development/ Differentiation/ Regeneration, and Rice Genome. With respect to Project management, one major challenge being presented both to Japanese scientists and government officials as well is whether they can truly work together as a team, dissolving the boundaries among competing investigators and among the different Ministries. In project-type research such as MGP, strong leadership and a determined team spirit for a common aim are required to productively carry out the actual research work as a team. In the first half of the Project, Human Genome Variation and Disease Gene Teams are working closely in the genome-wide approach as described below both within each Team and between the two Teams, offering proof of the integrity of MGP.

In the latter half of the Project, more interdisciplinary fusion is required, especially between experimental molecular biology, epidemiology, biostatistics and informatics. As of this writing, MGP is about to enter the fourth year of the total 5-year project; each research plan was approved by the institutional ethics committees, appropriate clinical samples have been collected and the huge amount of data are being generated systematically. However, an essential part of genome research is information science and technology. In the last two years of the Project, much is to be expected from the tie with the Bioinformatics Team. Even as a branch of basic science,

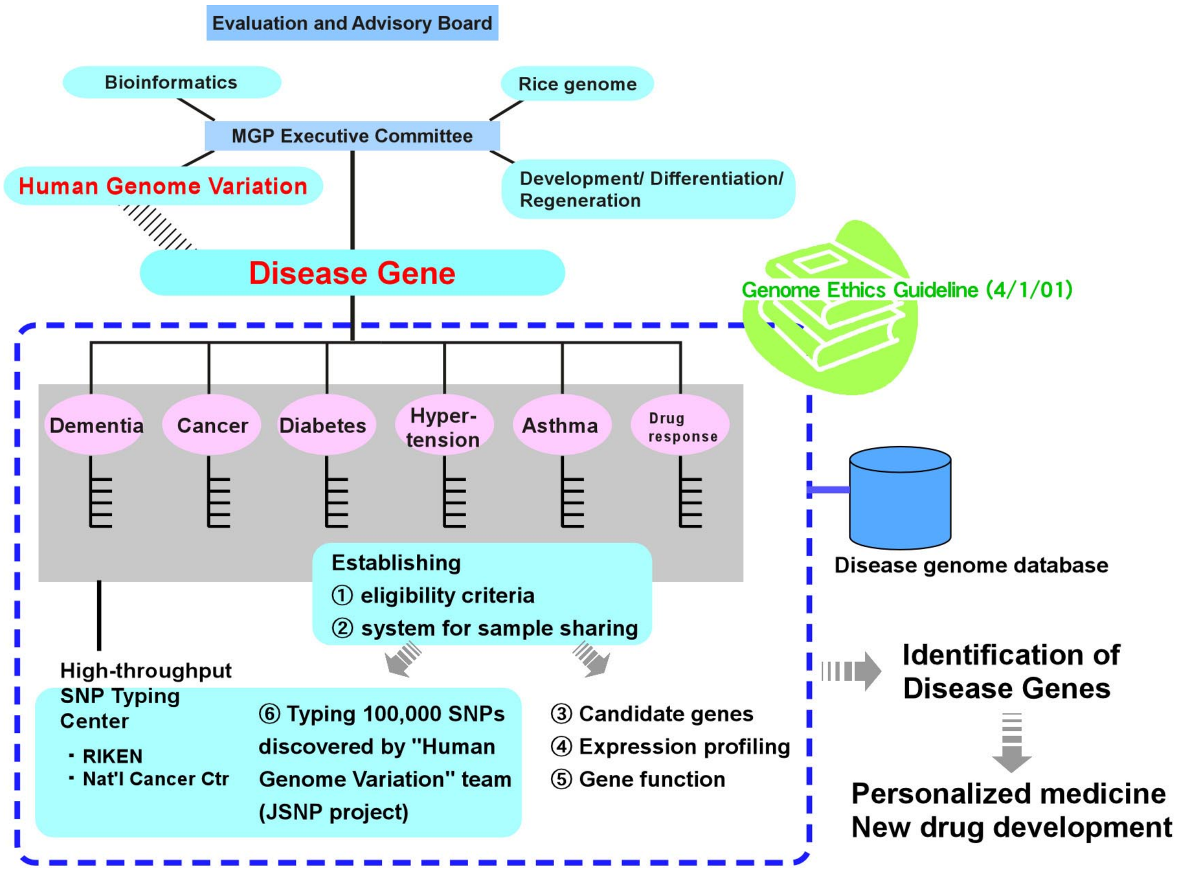

Fig. 1. Millennium Genome Project of Japan.

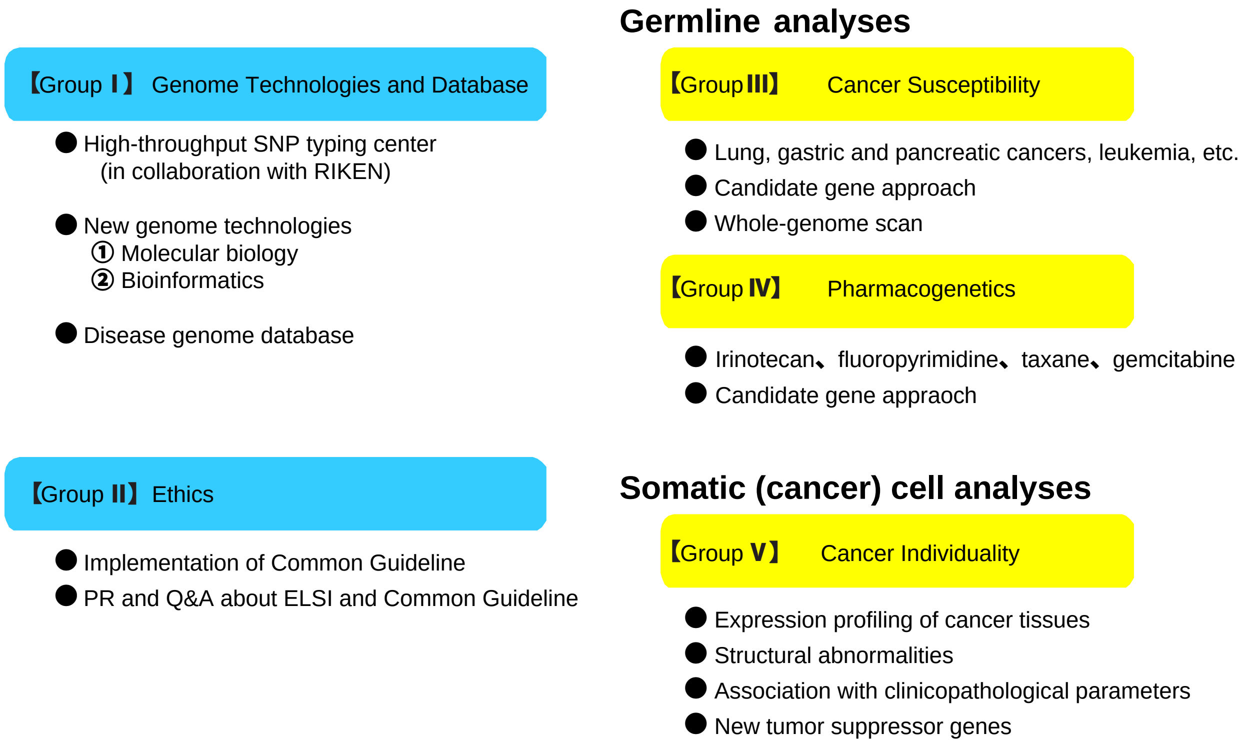
Germline analyses
【Group Ⅰ】 Genome Technologies and Database
● High-throughput SNP typing center (in collaboration with RIKEN)
● New genome technologies
① Molecular biology
② Bioinformatics
● Disease genome database
【Group Ⅲ】 Cancer Susceptibility
● Lung, gastric and pancreatic cancers, leukemia, etc.
● Candidate gene approach
● Whole-genome scan
【Group Ⅳ】 Pharmacogenetics
● Irinotecan、fluoropyrimidine、taxane、gemcitabine
● Candidate gene appraoch
【Group Ⅱ】 Ethics
● Implementation of Common Guideline
● PR and Q&A about ELSI and Common Guideline
Somatic (cancer) cell analyses
【Group Ⅴ】 Cancer Individuality
● Expression profiling of cancer tissues
● Structural abnormalities
● Association with clinicopathological parameters
● New tumor suppressor genes

Fig. 2. Organization of Millennium Genome Project at National Cancer Center.

genome research should deal with the 3 billion base pairs of the human genome harboring 20-30 thousands of often poorly annotated genes under complex patterns of transcriptional regulation, dynamically engaging in an intricate and mostly unidentified functional interaction. The genomes of other organisms may also need to be analyzed comparatively to find and understand the human genes. Moreover, in medical research, the genome-based approach is often necessary for the investigation of complex, multifactorial disease phenotypes, and the subjects, human beings, are far from the inbred laboratory organisms, with enormous genetic and epigenetic variations. All these factors inevitably place a crucial burden on the relatively new discipline in science, the so-called bioinformatics. Successful development of appropriate bioinformatics tools and databases may determine whether a genome-based approach can make sense out of the vast volume of seemingly senseless genomic raw data with statistical noise. However, the absolute paucity of bioinformaticists is a worldwide problem, and one of the major missions of the Bioinformatics Project Team is the education and training of experienced specialists in this field.

*Disease Gene Team.* The aim of this review is to outline the strategy of the Disease Gene Team of MGP (Fig. 1). The Team consists of six specialist “Subteams”: Dementia, Cancer, Diabetes, Hypertension, Asthma and Drug Response. The Subteam members have been selected from among scientists who already have an excellent record of research individually in each disease field. Therefore, the first year of MGP started with the natural extension and reinforcement of the ongoing individual research by closer collaboration within each Subteam, such as sharing unpublished information and cross validation of each member’s candidate genes using the samples owned by other members. In 2001, the second year of MGP, the Disease Gene Team and Evaluation and Advisory Board decided to launch a new project, a “genome-wide approach”, as a whole-team initiative, to be promoted in parallel with the “candidate gene approach”, which is planned and directed by each Subteam. The genome-wide approach is an adoption of the success of the Human Genome Variation Team, which has identified approximately 200,000 gene-associated SNPs (single nucleotide polymorphisms) in the Japanese population and has started to utilize the information for the genome-wide screening of disease-related genes by a proprietary high-throughput, low-cost, DNA-saving SNP typing system. The preceding effort has blossomed as the identification of a gene associated with increased risk for myocardial infarction, the first demonstration that a case-control association study by the SNP-based genome scan is a valid approach for a multifactorial “common” disease.[2)]

In the next section, an example of Subteam-specific research is shown for the Cancer Subteam. The other disease Subteams have established a similar research plan, except that cancer research, unlike that for other diseases, needs analyses of somatic (cancer tissue) mutation and abnormalities in gene expression as an essential part of the research in addition to the germline analyses.

**Cancer Subteam strategy.** The Cancer Subteam is composed of three member scientist groups, those from the National Cancer Center (NCC), the Institute of Medical Science, The University of Tokyo, and the Institute of Molecular and Cellular Biosciences, The University of Tokyo. Of these three institutions, the NCC is playing a central role within the Cancer Subteam of MGP and formed the following five research “Groups”: (I) Genome Technologies and Database, (II) Ethics, (III) Cancer Susceptibility, (IV) Pharmacogenetics and (V) Cancer Individuality (Fig. 2). To illustrate the scope of each Group, a few representative examples of its activity will be briefly introduced.

*(I) Genome Technologies and Database Group.* Genome analyses, by definition, intend to eventually cover the whole genome and/or genes. Although the reduction of the amount of DNA/RNA as well as the cost of each analysis have always been the top priority of the technology development, there is also an increasing demand for more DNA/RNA samples to apply to an increasing list of the new, powerful technologies. Moreover, there is extremely high research value in samples with a very limited amount and/or with a low yield of DNA/RNA, such as microdissected tissue fragments or old paraffin block archives. To make these invaluable samples available for a wide variety of genome analyses both at present and for the future, a new protocol of whole genome amplification has been developed by the researchers at NCC.[3)] Unlike several previous methods of similar whole genome DNA amplification, this method (code-named SATAN) combines a hydrodynamic shearing of DNA and high GC-content adapter primer for high annealing temperature. SATAN enables unbiased amplification of approximately > 90% of the genome as 0.3-1.5 kb fragments with excellent preservation of the complexity. For example, SATAN-amplified DNA was confirmed to show the same allelotyping pattern as the original DNA at 259 of 261

(99.2%) microsatellite markers on various chromosomes. DNA array CGH on the SATAN-amplified DNA correctly measured relative gene copy number at all 287 genomic loci examined. Typically, starting material as low as 1ng (corresponding to c.a. 100 cells) can be amplified several ten thousands fold, and SATAN was validated on 100-1000 laser-captured cells from cold methanol-fixed paraffin-embedded tissues. SATAN is now one of the basic protocols in cancer research in MGP, and a huge collection of invaluable DNA samples is now being archived as the SATAN libraries.

The same group who established SATAN also developed a method for high-fidelity amplification of whole cellular mRNAs.[3] The new protocol was designated TALPAT for T7 RNA polymerase promoter-attached Adapter Ligation-mediated PCR Amplification followed by *in vitro* T7-transcription. The optimized combination of the T7 transcription and PCR using high annealing temperature adapter primers was able to generate more than 10 mg of cRNA from 1 ng of total RNA (corresponding to c.a. 100 cells). TALPAT cDNAs are typically 0.3-3.0 kb fragments from the 3' end of the transcripts and serve as excellent samples for expression profiling by microarrays such as Affymetrix GeneChip. TALPAT has a highly important application in cancer research in particular; cancer tissue is a heterogeneous mixture of cancerous and non-cancerous components, and the cancer cells *per se* can also be highly heterogeneous. For instance, the cancer cells at the invasion front may have a distinct expression profile as compared to the less invasive, indolent portion of the tumor. Laser-capture microdissection (LCM) is a rapidly spreading technology, in which the cells of interest are selectively retrieved by laser-mediated dissection of the tissue section under a microscope. However, the conventional T7-based cRNA preparation for microarray analysis typically requires 5-10 µg of total RNA, or approximately $10^6$ cells, which is beyond the ordinary range of LCM. The power of TALPAT was in fact validated for the LCM samples obtained from cancer tissues, although it was found that the TALPAT-amplified samples should be compared with the TALPAT-amplified controls for reproducible results.

Another example of genome-characteristic technology developed by MGP researchers at NCC are the powerful sequence analysis algorithms, named Jessica and Polysy, for the sensitive and specific detection of polymorphisms and/or mutation. These algorithms are especially suited for data analysis for high-throughput multi-capillary automatic sequencers and have been developed by the researchers for their necessity in MGP at NCC, in which 10 ABI 3700 96-capillary sequencers generate about 2.5 Mb of sequence data per day for SNP/mutation screening. Jessica is unique among other existing sequence analysis algorithms in that it can handle small insertions and deletions (Indels) very effectively, and Polysy significantly reduces false positiveness in SNP detection. Jessica and Polysy have been packaged as a commercial software, *Nami-Hei*.

*(II) Ethics Group*. Although the Declaration of Helsinki by the World Medical Association[4] has been a widely recognized fundamental in medical research, and an ethics committee is a norm in leading medical institutions, it was evident that the rapidly developing genetic/genomic research on human materials needed a more detailed code specific to a particular field. In particular, the massive capture of germline polymorphism information on each individual can raise serious issues vis-à-vis privacy and information self-control rights. Just before the launch of MGP, the Japanese government convened a panel of specialists from diverse fields, including clinical medicine, basic science and social science, and the result was the “Guidelines for Ethical Issues in Genetic Research”, established on April 28, 2000. Obviously, unanimous agreement could not be expected for this kind of guideline, and the drafting members themselves were mostly divided between those who considered the guideline too prohibitive to necessary research, and those who criticized it as too pro-research. Although the guideline, often referred to as “Millennium Guideline”, was primarily intended for the MGP, which is conducted in the major research institutions with relatively ample resources, it still required considerable time for the MGP researchers to reorganize their system, such as the ethical committees, in accordance with the requirements of the new guideline. The ethics issues therefore did cause an initial delay in the Disease Gene Project of MGP, but the introduction of the Millennium Guideline, of a sort unprecedented in this country, was absolutely necessary and critical for any long-term healthy development of research-based national welfare. Immediately after the activation of the Millennium Guideline, the government started to extend it to cover all human gene/genome research in Japan, whether in academia, government or private sectors. The so-called “Common Guideline”[5] was put into effect on April 1, 2001, and the Millennium Guideline was integrated into the Common Guideline and deactivated.

In both Guidelines, two multi-disciplinary committees were organized, one primarily for drafting the Guideline and the other for primarily reviewing it, and each was respectively chaired by the then Deputy Director of NCC Research Institute and the then Director of NCC Central Hospital. Several other NCC staff members also played core roles in the preparation of the two guidelines, working closely with officials of the Ministry of Health, Labour and Welfare (MHLW). Thus, as a continuation of the work, the Ethics Group in the MGP is trying to gather information regarding the status of the Guideline implementation and to identify problems and difficulties in the current Common Guideline in preparation for the future revision scheduled for 2006. In addition, the members of the Ethics Group have been actively working to promote the Guideline in various media, by helping, for example, government officials establish a Q & A section on a website,[5] so that the Guideline will be widely known and correctly understood.

*(III) Cancer Susceptibility Group.*

*"Common cancer" as a target.* Cancer is essentially a disease of somatic mutation. The mutation arises, and the carcinogenesis process is spurred, in the context of the interaction of three major factors: germline genetic/epigenetic factor, life-style/environmental factor and aging factor. Depending on the types of the cancers, the relative contribution of the three factors varies significantly. For instance, external factors are obviously critical in cancers induced by occupational exposure to chemical carcinogens or cancers related to infectious agents. However, there can be substantial genetic differences among individuals in the metabolic profile of xenobiotics and susceptibility to infectious agents, such as receptor affinity for certain viruses. On the other hand, hereditary cancer syndromes showing a typical pattern of mendelian inheritance are considered monogenic diseases, but the increasing availability of genetic tests has identified more sporadic cases and asymptomatic carriers, revealing a highly variable phenotype and penetrance, which may be partly explained by the difference in non-genetic factors as well as the presence of "modifier" genes.[6] Nonetheless, in those relatively rare cancers where a single or only few dominant carcinogenic factor, either genetic or non-genetic, is operative, the etiology could be identified by a classical genetic or epidemiological approach, respectively. However, the majority of cancers lie between these two extremes of either environment factor-dominant or germline genetic factor-dominant categories and represent a complex interplay of multiple genetic and non-genetic factors in their carcinogenesis processes. The targets of the Cancer Subteam of MGP are those "common" cancers.

Cancer is not a single disease entity; rather, it is a collective term for a wide variety of etiologically, genetically and phenotypically distinct diseases. To focus the finite resources of MGP – most importantly the time limitation – several factors were considered in selecting the target cancers. In addition to such public health priorities as high prevalence/mortality in Japan, also considered were the availability of archive samples suitable for genomic analyses, collaboration with hospital personnel and our potential international competitive advantage based on the pre-MGP research activities. Finally, for the Cancer Susceptibility Group, we at NCC agreed to start with four major targets, lung, gastric and pancreatic cancers and leukemia.

*Candidate-gene approach.* The germline genetic factors which significantly influence an individual's cancer susceptibility may be identified by two distinct, but complementary, approaches. One is a "knowledge-based" or "hypothesis-driven" approach, in which one or more candidate genes are selected for the study, based on the existing information, rationale or hypothesis on the genes. This can be an intelligent and efficient approach, and the best candidates should obviously be scrutinized for the disease association first and at a high sensitivity. The other approach adopted in MGP is a "genetic statistics-based" approach and will be detailed in the "Genome-wide approach" section of this article.

Approximately 300 candidate genes have been selected from various pathways such as DNA repair, immune system, cell cycle regulation, transcriptional control and xenobiotics metabolisms and from oncogene and tumor suppressor gene groups. These candidates may not be necessarily specific to the initial four malignancies selected for the study; rather, many of them are expected to be relevant for the common pathways of cellular transformation in various types of cancers. The basic study design is a case-control study to detect a statistically significant association using SNPs as genetic markers. SNPs of the candidate genes are obtained in two ways: some are selected from the literature or public SNP databases, such as relatively common SNPs of the genes involved in the xenobiotics metabolisms. SNP typing on those genes is in progress on gastric cancers using the primer-extension/mass spectroscopy-based system.[7] Alternatively, the coding exons of the genes are resequenced to elucidate a more complete picture of the polymorphisms, including relatively rare

Case 1: Control 1

| | Genotype frequency | | | | | |
|---|---|---|---|---|---|---|
| Odds ratio | 0.01 | 0.05 | 0.10 | 0.20 | 0.30 | 0.40 |
| 1.2 | 58484 | 12281 | 6543 | 3750 | 2910 | 2594 |
| 1.4 | 15977 | 3378 | 1814 | 1057 | 834 | 755 |
| 1.6 | 7705 | 1639 | 887 | 524 | 420 | 385 |
| 1.8 | 4674 | 1000 | 545 | 326 | 265 | 246 |
| 2.0 | 3209 | 690 | 378 | 230 | 188 | 177 |
| 2.2 | 2380 | 514 | 284 | 174 | 144 | 137 |
| 2.4 | 1860 | 404 | 224 | 139 | 116 | 112 |
| 2.6 | 1510 | 329 | 184 | 115 | 97 | 94 |
| 2.8 | 1260 | 276 | 155 | 98 | 84 | 81 |
| 3.0 | 1076 | 236 | 133 | 85 | 73 | 72 |

Case 1: Control 2

| | Genotype frequency | | | | | |
|---|---|---|---|---|---|---|
| Odds ratio | 0.01 | 0.05 | 0.10 | 0.20 | 0.30 | 0.40 |
| 1.2 | 43582 | 9157 | 4882 | 2802 | 2177 | 1943 |
| 1.4 | 11838 | 2506 | 1348 | 787 | 623 | 565 |
| 1.6 | 5680 | 1210 | 656 | 390 | 313 | 288 |
| 1.8 | 3430 | 735 | 402 | 242 | 197 | 184 |
| 2.0 | 2346 | 506 | 278 | 170 | 140 | 133 |
| 2.2 | 1733 | 376 | 208 | 129 | 108 | 103 |
| 2.4 | 1350 | 294 | 164 | 103 | 87 | 84 |
| 2.6 | 1093 | 239 | 134 | 85 | 72 | 71 |
| 2.8 | 910 | 200 | 113 | 72 | 62 | 61 |
| 3.0 | 775 | 171 | 97 | 63 | 54 | 54 |

Fig. 3. Number of cases required for case-controls study. Power = 90%, level of significance = 5%, case:control = 1:1 or 1:2. Cells with fewer than 200 cases are shaded.

SNPs, small insertions and deletions. Moreover, some normal-tumor tissue pairs are included among the resequencing samples, so that both germline polymorphisms and somatic mutations are searched simultaneously. Approximately 3,000 SNPs and other minor variations such as small insertions and deletions have been identified so far from more than 200 cancer patients and non-cancer volunteers. From this category, several non-synonymous missense SNPs on the DNA repair genes are selected and examined in the initial case control study on lung cancers.

An important part of the effort of the Cancer Susceptibility Group is the ascertainment of the DNA samples. The number of cases necessary to achieve 90% power and 5% significance level in a case-control study at case:control = 1 : 1 or 1 : 2 ratio is shown in Fig. 3. From the standpoint of practical cancer prevention by capturing a high-risk group for cancer development, the genetic risk factors with 10% or higher population genotype frequency and 2.0 or higher relative risk are the main target of our research (examples of relative risk = 2: smoking for pancreatic or gastric cancers, positive family history of breast cancer in the first degree relatives). Approximately 200-300 cases are then needed for the initial case-control association studies for the candidate genes. It is estimated that the familial relative risk for many common cancers is around 2, and the data from a population-based series of individuals with breast cancer were compatible with a log-normal distribution of genetic risk in the population that is sufficiently wide enough (standard deviation = 1.2) to discriminate high- and low-risk groups.[8)]

*(IV) Pharmacogenetics Group*. This Group seeks the most direct and immediate application of genetic information to cancer clinics, and in that sense it is one of the most important areas in MGP. It is expected to embody personalized medicine by selecting the most effective and safe anticancer drug regimen for each patient with cancer. Although the malignancy of the cancer cells, and thus the somatic mutation and expression profile of the cancer cells, and the local drug delivery environment of the cancer tissue may be the major determinants of the anticancer effect, germline polymorphisms affect both the anticancer effect and adverse drug reaction causing each individual's differences in ADME (absorption, distribution, metabolism and excretion), activity of the receptor and following signal transduction pathway of the drug.[9)]

The Pharmacogenetics Group focuses on the identification of germline polymorphisms that are associated with the development of adverse reactions of four anticancer reagents: irinotecan, fluoropyrimidines, taxanes and gemcitabine. These drugs were selected because they 1) are approved drugs actively used in oncology clinics in Japan, 2) show a promising therapeutic effect and will be in active duty for the next several years at least, but 3) have some adverse effect in a certain fraction of the patients. At this moment, the candidate gene approach is ongoing by resequencing the genes considered to be involved in the metabolism, transport, receptor and signal transduction of the drugs. The phenotypes to be associated with the genotype are not only the symptoms, signs and clinical laboratory data showing the adverse reactions (pharmacodynamics, PD), but also the concentration of the drug and its major metabolites in the serum and urine (pharmacokinetics, PK). Although the clinical control of the PD parameters is the final goal of the research, PD will be subject to many variables other than the chemotherapy *per se*. PK analysis demands a substantial resource and effort in the study, not to mention the patient's burden of the scheduled multiple blood drawing, but it should greatly enhance the sensitivity and power of the genotype-phenotype association study involving several hundreds of patients.

The crux of this research is, however, that it is a prospective clinical genetic research involving a follow-up of dynamic phenotypic change before and after chemotherapy. In addition to detailed informed consent, all the specimens and voluminous clinical data need to be collected in a setting of very busy in-patient or out-patient clinics. A central case registration, data management and protocol monitoring system is mandatory in this kind of clinical research, and SMO (site management organization) was introduced to assist the doctors, CRCs (clinical research coordinators), and researchers of the Pharmacogenetics Group.

The research of the Group is a collaboration with the pharmacology/pharmacogenetics specialists at the National Institute of Health Sciences (NIHS) in Japan. In addition to the limited candidate gene approach, the necessity of the more comprehensive, even genome-wide, approach has been repeatedly discussed to explore any unknown genes related to the drug effect.

*(V) Cancer Individuality Group.*

*Structural abnormalities*. A plethora of genetic abnormalities have been found in cancer cells. In the early phase of molecular oncology, the structural alterations of the cancer cell genome were analyzed mostly on few candidate loci to detect their amplification, rearrangement or deletion including loss of heterozygosity (LOH). The methods of choice have been Southern blot hybridization, restriction fragment length polymorphism (RFLP), microsatellite allelotyping, PCR-single strand conformation polymorphism (SSCP), FISH and its several evolution forms. Comparative genomic hybridization (CGH) was introduced in 1992[10] as a powerful method to map the regional gain or loss of DNA sequences along the entire chromosomes. Unlike other genome scan methods for DNA copy number changes such as AP-PCR, CGH gives an excellent bird's-eye landscape over the entire genome by a single experiment. However, the weakness of CGH includes; 1) the translocations may not be detected, 2) the signals tend to be unstable at the centromeric and telomeric regions, and 3) resolution is generally poor. If an amplicon is longer than 5 Mb, as low as 2-fold amplification and even deletion may be identified, but if an amplicon is approximately 300 kb, at least 5 to 10-fold amplification may be necessary for detection by a conventional CGH. Array CGH or matrix-based CGH[11] can increase the resolution and simplicity significantly. In array CGH, the metaphase chromosomes as targets of hybridization are substituted with arrayed DNA probes such as BAC, cDNA or even oligonucleotides. In MGP, the SATAN method was used to amplify BAC DNA efficiently, and BAC-arrayed CGH chips covering the whole genome have been prepared. To further increase the sensitivity and specificity of the analysis, paraffin-embedded cancer tissues are microdissected by LCM, and their DNAs are analyzed by the array CGH to search mainly for gene amplification at a higher resolution.

Cloning and characterization of tumor suppressor genes have been one of the major molecular carcinogenesis projects at NCC. In addition to the positional cloning based on LOH or homozygous deletion, a functional complementation assay by YAC transfer has been employed to identify candidate tumor suppressor genes. However, proof of a tumor suppressor gene in human cancer *in vivo* is not an easy task in many cases, and an extensive search for somatic mutation is one of the agenda. MGP has accelerated the mutation search in some of the tumor suppressor gene studies at NCC by offering its high-speed sequencing capacity.[12]

*Expression profiling*. Array CGH with whole genome coverage is still mostly a technology being developed in a few research laboratories because it requires a specific hybridization to a target sequence embedded in the entire human genome, which has a complexity about two orders of magnitude higher than the mRNA pool and is composed of approximately 50% interspersed repeats.[13],[14] In contrast, microarrays for mRNA expression profiling are rapidly becoming a standard option in many laboratories, and many high-quality and reliable chips are available commercially. There are two basic types of high-density microarrays: those generated by on-chip synthesis of oligonucleotide probes and those made by spotting prefabricated oligonucleotides or DNA fragments.[15] The latter type of microarrays with cDNA fragments spotted on them are often called the "Stanford type", after the group at Stanford University led by Patrick Brown, and a typical example of the former type is GeneChip by Affymetrix (http://www.affymetrix.com/). In the Stanford-type cDNA arrays, gene expression is typically analyzed by comparing two RNA samples, giving rise to the relative expression level of each gene. GeneChip, on the other hand, is a highly reproducible and quantitative platform, and after a proper normalization, the signal intensity of each probe set can be directly compared among different

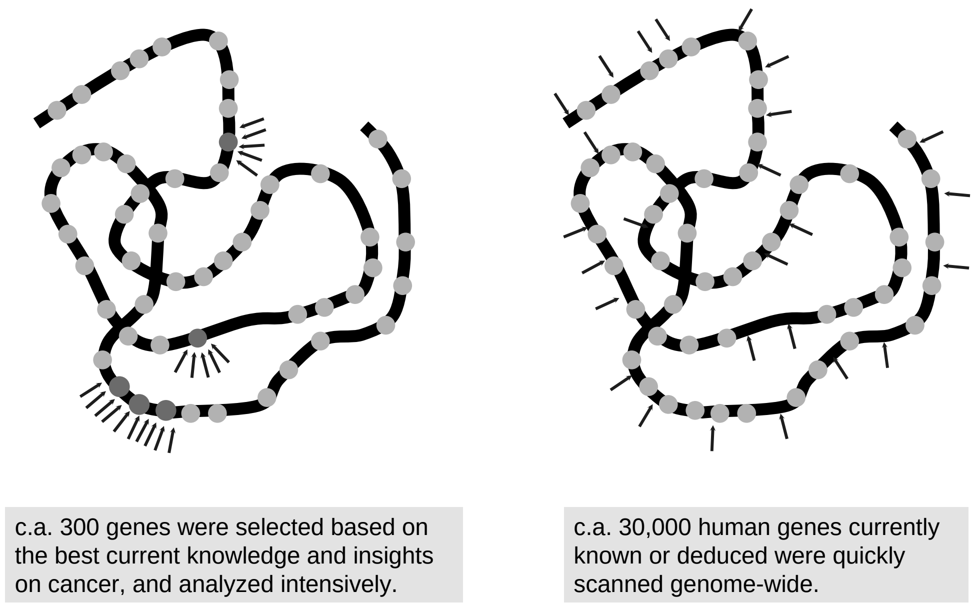
c.a. 300 genes were selected based on the best current knowledge and insights on cancer, and analyzed intensively.
c.a. 30,000 human genes currently known or deduced were quickly scanned genome-wide.

Fig. 4. Candidate gene and whole-genome approaches.

probe sets and also among different chips. At NCC, GeneChip was selected and is being used for comprehensive expression analysis on c.a. 12,000 annotated genes.

Most studies so far aim for class prediction through a supervised learning approach,[16] and genes associated with distinct histopathological features are being screened for various cancers such as liver, kidney, pancreas, stomach, lung and leukemia. Another important application of this approach is identification of an expression profile that predicts a different outcome to cancer therapeutics. Such a biomarker should be particularly useful for the treatment choice for esophageal cancer. For esophageal cancer that extends deeper than the muscularis mucosae but remains within the esophagus, either chemoradiotherapy or surgery will be performed, showing a similar overall 5-year survival rate of approximately 50%. Although the conventional clinicopathological parameters failed to segregate the subgroup that will be benefited significantly by chemoradiotherapy, supervised learning analysis on GeneChip data identified an expression pattern on a set of genes associated with an extended survival. Researchers are now designing a custom "minichip" harboring a spotted oligonucleotide array corresponding to selected genes to launch a prospective clinical study to test the validity of the chip. Unlike conventional diagnostics, which typically measures a single molecule, a class-predicting expression profile depends on the hundreds of genes selected by computational statistics. Gene set selection may be progressively refined by accumulating data, which are used for additional training sets. Thus, even after release for bedside use, a scheme of feeding back the chip data to some sort of central database, together with properly anonymized clinical information, should be considered for microarray- and other genome technology-based diagnostics.

**Genome-wide approach.** As outlined above for common cancers, "common diseases" generally develop and progress through a complex interplay of multiple factors, including those which are genetic, environment/life-style and aging-related. Identification of the genetic component is important in two aspects: 1) recognition of a genetically high-risk population for effective and efficient prevention and/or early diagnosis and treatment, and 2) elucidation of the molecular pathogenesis of the disease to identify target molecules for diagnosis, treatment and prevention. For the quest of the disease genes of common diseases, each Subteam has gathered its expertise and has selected candidate genes as illustrated in the Cancer Subteam strategy section.

Although the candidate gene approach can be an efficient and necessary approach, its success largely depends on the extent to which the researchers understand the entire spectrum of the pathogenesis and the genes involved in that process. A draft human genome sequence was published in February 2001,[13),14] and one for mouse on December 5, 2002.[17] The arrival of the mouse genome sequence and the comprehensive collection of the 60,770 full-length mouse cDNA clones, which identified total 37,086 representative transcript and protein sets,[18] is a huge step forward in unveiling all human genes and their functions; the mouse is an experimentally tractable surrogate organism with huge biological experimental data being accumulated in a number of human disease models. In fact, 28,000-30,500 protein-coding genes have been identified in the mouse genome, and 99% of them have direct counterparts in humans, and 96% lie within a similar conserved syntenic interval in the human genome. Moreover, a comparison of the genome sequences showed a surprising similarity between the two mammals, even in the non-genic regions of the chromosome.[19] However, even with those remarkable advances in our knowledge about genomes, the latest catalogue of human and mouse genes leaves approximately 40-50% of them with little or no functional annotation.[13),14),18] Together with the fact that any information we have on the remaining 50-60% of the "known" genes is often fragmentary and far from complete, a more comprehensive genome-wide, statistics-based search was therefore adopted in MGP as an essential complement to the candidate gene approach (Fig. 4).

Points of the genome-wide approach for disease gene hunting in MGP are discussed below.

*1. CD-CV hypothesis and association study*. The basic premise of the genome-wide approach in MGP is "common disease-common variant" (CD-CV) hypothesis.[20] In mendelian diseases with significant phenotypic impact (e.g. high mortality rate at young age), there may be many rare functionally deleterious alleles often generated by *de novo* germline mutation on a single causative gene. In contrast, there may be multiple risk-modifier genes for common diseases, each with only a moderate effect (e.g. relative risk = 1.5-2.0) on the phenotype. For instance, segregation analyses suggested that the polygenic model of several common low penetrance genes with multiplicative effects on risk fits well with the data on a population-based series of breast cancer as well as the data on high risk families.[21),22] Such weak risk-conferring disease alleles might well-survive any natural selection, or even have had a selective advantage in the past (e.g. "thrifty genotype" hypothesis for type II diabetes[23]) and may expand by genetic drift during a population bottleneck and may exist in a substantial fraction in the current source population[24]; the result is a few common disease alleles (variants) causing a common disease (Fig. 5). This CD-CV model seems to be applicable to some common disorders at least.

There are two types of genome-wide search based on statistical genetics: linkage analysis and association study (Fig. 6). For a typical monogenic mendelian disease, a genetic model detailing penetrance of the disease genotype or mode of inheritance, can be inferred from a segregation analysis on few large-size pedigrees. A traditional linkage analysis may then successfully track down a disease locus. However, such a parametric linkage analysis is not feasible for common diseases, for which penetrance of the disease allele cannot be reliably estimated. One solution is a nonparametric or model-free linkage analysis, such as the affected sib-pair method, in which loci with a high frequency of allele sharing among multiple pairs of affected siblings are searched over the entire genome. Major disadvantages of the affected sib-pair analysis are 1) candidate regions defined by the method are usually still very large for the following positional cloning, because many chromosomal segments are shared by siblings anyway, and not all affected sib pairs may share an allele at a candidate locus in polygenic common diseases, 2) for a relatively rare "common" disease, collection of a necessary number of sib pairs is often unrealistically difficult. It was estimated that the number of sib pairs necessary for the identification of a disease gene with a relative risk of 2 or less is more than 2,500.[25] In the case of cancer, such a number is practically impossible to achieve even for the commonest one, gastric cancer, within the 5-year MGP. A whole-genome association study was thus selected in MGP as a common strategy for the five disease categories.

*2. JSNP as a genetic marker*. There are two popular genetic markers for the genome scan: microsatellite and SNP, which are roughly compared in Table I. Many studies have relied upon commercially available 400 or 800 microsatellite markers to narrow down the disease locus to a range of approximately 10 cM. In Japan, a unique effort of systematic search for polymorphic microsatellite markers has successfully identified approximately 30,000 markers, which theoretically enable mapping at a 100 kb resolution by association study. The tracking down to a disease gene can then be done efficiently through a regional fine mapping and haplotype analysis by SNP. On the other hand, SNP collection has exploded in the post sequencing era, and more than 3 million SNPs have already been deposited in dbSNP at NCBI.[26] Unlike those SNPs scattered anywhere in the human genome, the Human Genome Variation Team of MGP focused on gene-associated

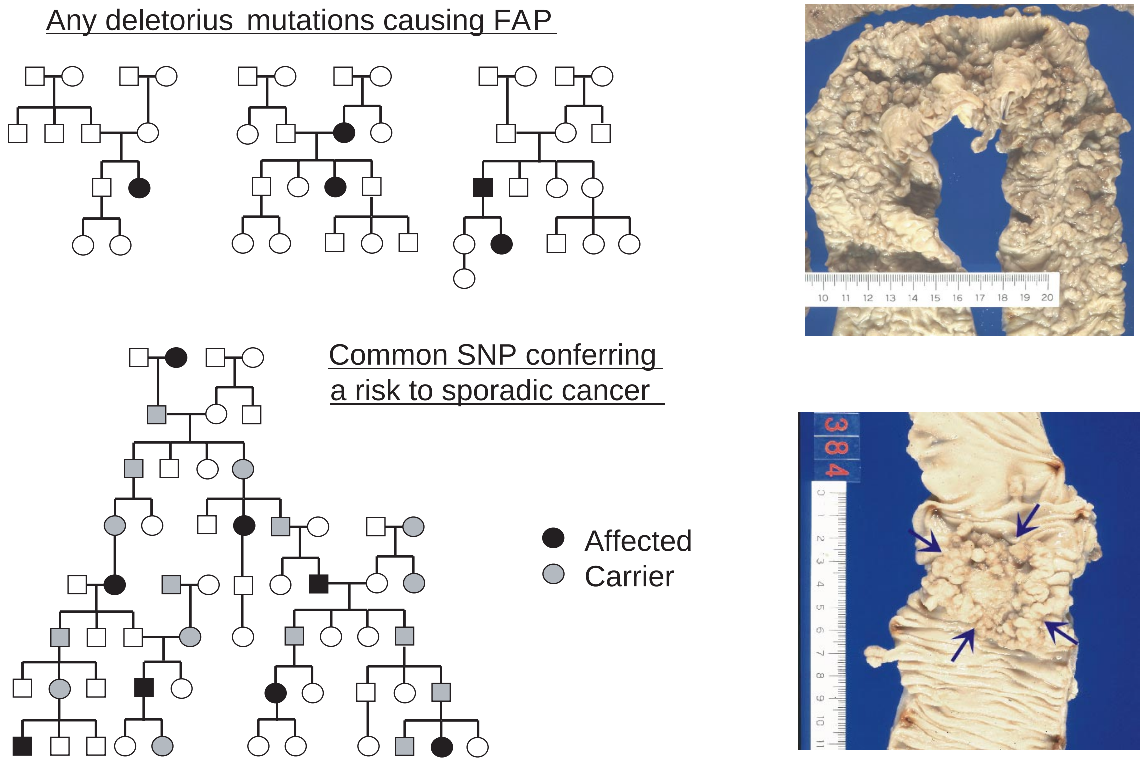
Any deletorius mutations causing FAP
Common SNP conferring a risk to sporadic cancer
● Affected
● Carrier
384

Fig. 5. Common disease-common variant (CD-CV) hypothesis. Squares denote males, circles females.

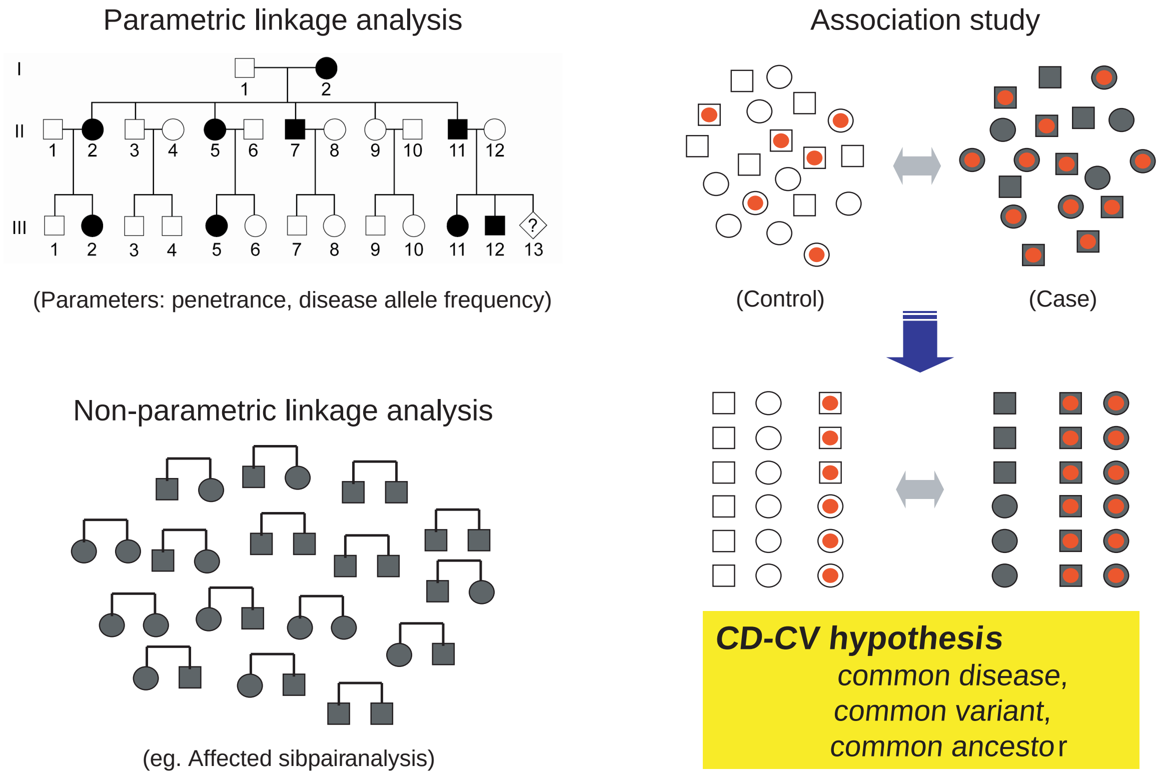
Parametric linkage analysis
I
II
III
?
(Parameters: penetrance, disease allele frequency)
Non-parametric linkage analysis
(eg. Affected sibpairanalysis)
Association study
(Control)
(Case)
CD-CV hypothesis
common disease,
common variant,
common ancestor

Fig. 6. Two strategies in statistical genetics. Filled squares (males) and circles (females) are diseased people. In the scheme for an association study, red dots mean those who have a disease allele of a SNP under study.

Table I. Characteristics of SNPs and microsatellites (MSs).

| | SNP | Microsatellite |
|---|---|---|
| Extent of LD | Short, few kb-100 kb | Long, 100 kb-1 Mb |
| Haplotype | Can increase effective number of alleles for gene mapping, but most haplotype analysis depends on statistical inference | One MS may correspond to several SNP haplotypes |
| Statistical computation | Biallelic, simple statistical model and computation | Highly polymorphic nature may reduce the sample size for each genotype in association study |
| Association with gene | Haplotype can be constructed within a gene | Usually located in intergenic regions or introns |
| Direct cause of phenotype | SNP itself can be a direct cause of phenotype | Usually a marker outside genes, and SNP or other polymorphism/mutation responsible for the phenotype needs to be eventually searched |
| Discovery | Needs a large amount of experimental work | Sequence-based prediction is possible |
| Mutation rate | Low, applicable for association study to identify alleles derived from very ancient founder | High, not applicable for association study to identify very ancient alleles |
| Marker and info availability | Already many databases available world wide, including allele frequency data<br>*JSNP DB is a great advantage in Japan | 5-10 cM commercial markers widely used, but a high density marker set is not yet a standard |
| Automation | Already many high through-put typing technologies available world wide | Not currently amenable to highly automated analyses |
| Pooled sample analysis | Possible only in special platforms | Already a standard protocol |
| Study of junk DNA sequence | Mostly for gene-finding research | Mapped in> 10 Mb gene-poor desert of the genome, good for study on junk DNA |

LD, linkage disequilibrium; DB, database.

SNPs as the most valuable landmarks for gene discovery and resequenced about 5% of the human genome comprising the promoter, exons and some introns using DNA from 24 unrelated Japanese volunteers.[27] On December 25, 2002, a total of 194,393 SNPs were made publicly available (so-called JSNP), together with the allele frequency data for 78,570 SNPs.[28),29] The latter information is important for the study of common diseases among the Japanese population, as a substantial ethnic variation has been noted in the SNP allele frequency.[30]

The rapid decay of linkage disequilibrium (LD) among SNPs, probably due to their biallelic nature and thus, a weak allele marking capability, has been considered a major drawback of SNP as a genome scan marker, because a huge number of SNPs are then necessary to cover the genome.[31] However, more recently, it was found that the human genome is structured into discrete islands of high LD separated by hot spots of recombination.[32)-35] In European and Asian populations, 85% of the genome lies in LD blocks of 10 kb or larger, and 50% in blocks of 44 kb or larger.[36] The average distance between JSNPs is approximately 800 bp,[27] which should be a sufficient length for LD mapping in the most genic region of the genome.

The serious difficulty or feasibility of genome-wide association study on unrelated subjects to map common disease loci has been a matter of hot discussion,[20),37] but the researchers who developed JSNP database successfully identified the lymphotoxin-$\alpha$ gene as a susceptibility gene for myocardial infarction (odds ratio = 1.78, p = 0.0000033, 1,133 cases vs. 1.006 controls).[2] This was the first report clearly showing that properly designed genome-wide case-control association studies are powerful tools for identifying genes related to common diseases.

The human chromosomes may be classified into a limited number of haplotypes at the LD blocks. In October 2002, International Consortium launched a haplotype-mapping (HapMap) project[38),39] to identify SNP-based haplotypes common to the major ethnic groups. SNPs which can distinguish each haplotype (haplotype-tagging SNPs, or htSNPs) will be selected, enabling a more efficient (i.e., a fewer SNP typing without loss of statistical power) haplotype-based association study for disease gene mapping.

*3. High-throughput SNP typing platform by RIKEN*. RIKEN, one of the core institutions in the Human Genome Variation Team, established a highly efficient, high-throughput SNP typing system[40] combining 96-plex PCR, Invader assay (Third Wave Technologies Inc., WI, U. S. A.), 384-microwell card and bar code-based LIMS (laboratory information management system). In particular, the typing requires as little as 0.1 ng of genomic DNA per SNP, an important feature as a genome-scan platform. The RIKEN technology has been transferred to NCC as a part of MGP collaboration, and the two institutions jointly formed the SNP Typing Center for the Disease Gene Team (Fig. 1).

*4. Confounders and population stratification*. While association study is a powerful approach for hunting genes of common diseases, probably the most critical or vulnerable part of the strategy is the case-control selection. There can be two major potential pitfalls in the genetic association studies: 1) poor phenotype-genotype correlation, and 2) cases and controls from different source populations (population stratification). The first difficulty, which is specific to genetic association study, arises because eligibility criteria for case selection depends on a certain disease phenotype. However, the phenotype may not be correlated one to one with a genotype.[20),37] There may be locus heterogeneity (a similar phenotype produced by different genes), allele heterogeneity (multiple founders or genocopy) or phenocopy (a similar phenotype produced by a nongenetic cause), all reducing the power of the association study. This may especially be an issue for cancer, because many cancer phenotypes may be largely determined by somatic mutations and/or alteration in an expression profile. Histopathological phenotypes may not be able to narrow down a group of patients sharing a common ancestral mutation (common variant).

The second issue, population stratification, is often argued as a potential confounder in a genetic association study. The concept of confounding is briefly explained here. Assume an association study that tests an association between an SNP as a predictor variable and a disease as an outcome variable. A confounding factor is an extraneous factor responsible for the difference in disease frequency among different genotypes or alleles, and defined as follows: 1) it is a risk factor for the disease, 2) it is associated with the SNP, 3) it is not a consequence of the SNP.[41),42] For instance, if the SNP under study has nothing to do with a smoking habit, and thus there is no significant difference in the proportion of smokers among the subjects with different genotypes at the SNP, then smoking is not a confounder in the association study for lung cancer. Alternatively, if the risk-conferring effect of the SNP is solely due to its positive effect on smoking behavior, and thus the smoking is an intermediate factor in a causal chain, then adjustment by matching or stratification for smoking will miss the SNP as a risk factor for lung cancer. Genetic association studies are often regarded as being prone to false positive- or poorly-reproducible results, even after careful consideration of known extraneous risk factors such as smoking for cancer.

Population stratification occurs when the source population comprises unrecognized subpopulations that have different allele frequencies of the SNPs under study, and these subpopulations also have different risks of disease, thus satisfying the above criteria for a confounder. However, even for the studies in western countries, there is not much evidence that bias from stratification can be a major problem[43] except under extreme conditions such as the famous study of type II diabetes in Pima Indians with genetic admixture.[44] In fact, publication biases toward positive results and underpowered non-significant studies of real associations seem to account for the majority of the inconsistency.[45]

On the other hand, several statistical methods have been developed to detect and correct for q cryptic population structure, so that valid case-control tests of association are possible even in the presence of population structure (ref. 46) for review). The methods are classified into two categories: one is a "genomic control" method, in which marker loci unlinked to the candidate locus are included to adjust the distribution of a standard $\chi^2$ statistics.[47] The other is a "structured association" method, which first uses data from random, unlinked markers to infer the details of cryptic subpopulations, and tests of association are then essentially performed in each of the identified subpopulations.[48]

These methods may be employed to analyze the SNP scan association data in MGP, because the subjects of the genome-wide association study in MGP have been ascertained from many hospitals and medical centers in accordance with eligibility criteria established by the disease Subteam. All the patients are ethnic Japanese, but their demographic features may vary significantly. At this moment, there is no comprehensive genome-wide information regarding SNP allele frequency difference among different areas in Japan. The genome-wide JSNP typing in MGP will offer such basic information in genetic statistics on the Japanese popu-

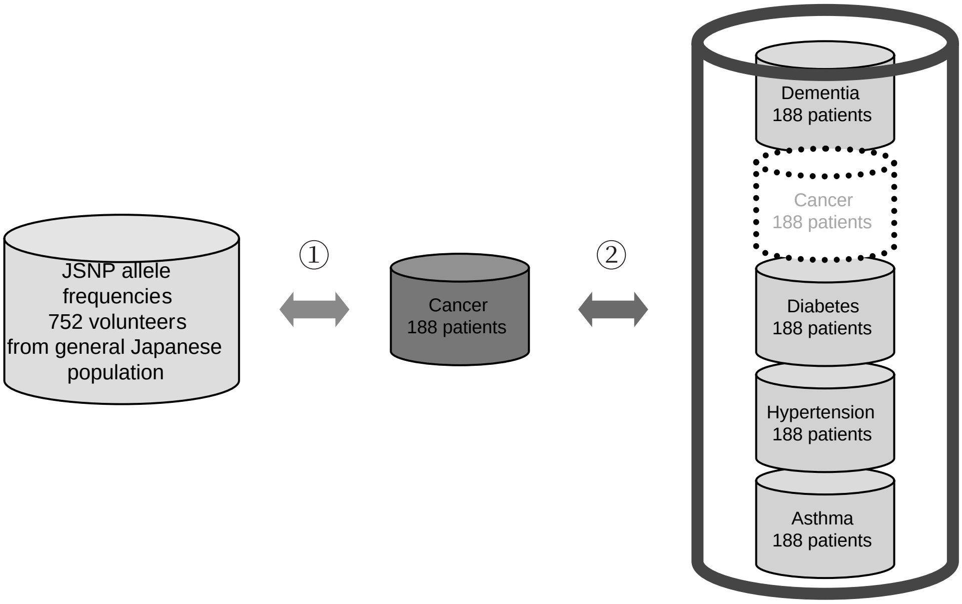

Fig. 7. Two sets of case-reference comparisons in MGP.

Table II. Power of two-stage genome-wide association study in MGP

■Assuming:

Number (#) of SNPs to be typed = 100,000
Proportion of SNPs with true association = 0.001
Dominant model
Disease allele frequency in control = 0.2
Significance level for the 1st stage = 0.01
Significance level for the 2nd stage = 0.05 with Bonferroni correction

# of Samples
1st stage Case = 188, Control = 188 × 4 = 752
2nd stage Case = Control = 752

■Simulation:

| Odds ratio of the SNP | 2 | 1.7 | 1.5 |
|---|---|---|---|
| Power of the 1st stage | 0.95 | 0.73 | 0.44 |
| Expected # of positive SNPs at the 1st stage | 1,094 | 1,073 | 1,043 |
| Expected # of false-negative SNPs at the 1st stgae | 5 | 27 | 56 |
| Expected # of "true" SNPs surviving the 1st stage | 95 | 73 | 44 |
| Power of the 2nd stage | 0.99 | 0.84 | 0.41 |
| Expected # of positive SNPs at the 2nd stage | 94 | 61 | 18 |

lation.

*5. Two-stage screening scheme*. In order to perform a SNP-based genome scan efficiently on the five MGP diseases within the limitations of budget and time, the Disease Gene Team of MGP designed a two-stage screening scheme. In the 1st "exploratory" stage, a total of 940 patients comprising 188 cases from each of the five Subteams are now being typed at the two typing centers, RIKEN and NCC. The target number of SNPs to be typed by the end of FY 2003 is 100,000 from the gene-associated JSNP collection. In the 1st stage, the same set of 100,000 SNPs is analyzed for the five diseases for the purpose of a quick scan across the whole genome to flag approximately 1,000-2,000 SNPs for each disease. In the 2nd "confirmatory" stage, the set of 1,000-2,000 SNPs selected for each disease will be analyzed separately on 752 cases (different patients from those analyzed in the 1st stage) and 752 controls. In the 1st stage, two kinds of statistical comparisons are performed to select SNPs worthy of 2nd stage analysis; 1) comparison with the JSNP allele frequency on 752 people from the general Japanese population publicized on the JSNP web site,[28] and 2) cross-comparison with four other diseases (188 cases vs. 4 × 188 = 752 "references") (Fig. 7). The allele frequency odds ratio and its p value are calculated, and those SNPs with a p value less than 1% will be picked up

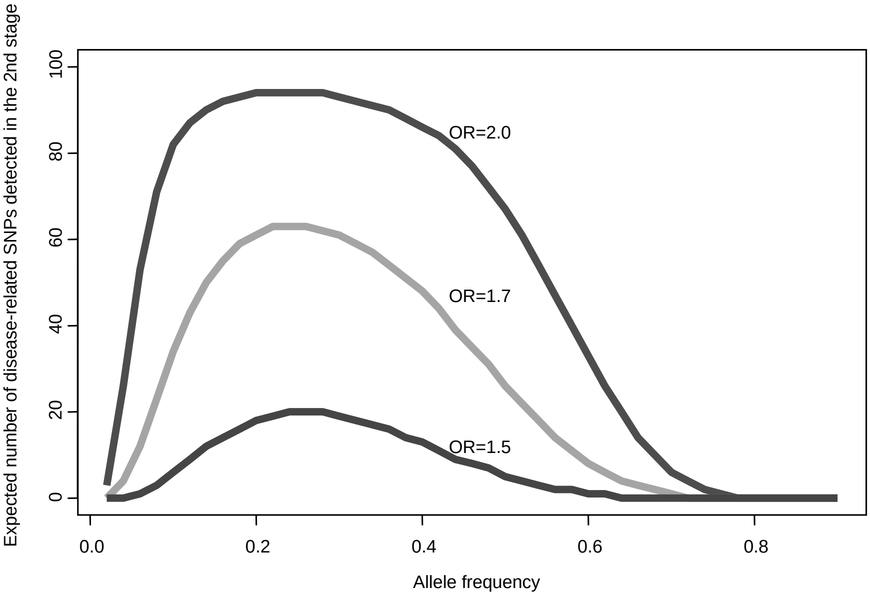

Fig. 8. Power calculation of two-stage genome-wide association study in MGP.

for the 2nd stage. In the 2nd confirmatory stage, an appropriate correction for multiple testing will be applied, such as the Bonferroni correction, which *per se* may be too conservative[49] in the presence of LD among 100,000 SNPs.

Table II shows a simulation of the two-stage screening assuming a dominant model, risk allele frequency = 0.2, and the fraction of the "true" SNPs = 0.1% or 100 SNPs in the 100,000 SNPs to be analyzed. Estimating the number of SNPs showing true association with the disease that fits CD-CV hypothesis is difficult; a rough assumption would be 10-100, probably corresponding to less than 30 genes or so. For instance, it was suggested that at least 100 loci affect susceptibility to coronary artery disease,[50] and a comparable number of the disease genes is also suggested for some other common diseases.[24),51] Because we assumed 100 true SNPs present in the initial 100,000 target SNPs, the last row of the Table II, "Expected number of positive SNPs at the 2nd stage", roughly corresponds to "Expected number of positive SNPs at the 2nd stage, expressed as % of the true SNPs initially present in the 100,000 SNPs analyzed in the 1st stage". Fig. 8 shows the power of the two-stage screening scheme of MGP.

**Disease database and perspectives.** The traditional products of scientific research are publications and, more recently, intellectual properties. In addition to those outcomes, an important mission of MGP is the establishment of a database for genome-based medicine and medical research. The 12/19/99 document by the then Prime Minister stated that the construction of a "Disease Database" at the NCC should be one of the human genome-related databases to be established in MGP. The details of the Disease Database, or Disease Genome Database, have not yet been decided, but three points are envisioned at this moment: 1) JSNP genome scan data on five MGP diseases (dementia, cancer, diabetes, hypertension and asthma) will be included, although to what extent "raw" data can be deposited in the database awaits further discussion, because 100,000 SNP typing data *per se* will be a comprehensive genomic finger print of an individual. 2) Strict access restriction, security and privacy protection precautions will be installed. 3) There will be a seamless link with the Integrated Database being constructed by the Bioinformatics Team of MGP, especially its H-Invitational initiative to coordinate human full-length cDNA sequencing and annotation projects worldwide.[52]

Huge human and budgetary resources have been invested in MGP with high expectations from various sectors of this country. Powered by the latest high-throughput genome technology, and, most importantly, by establishing a nation-wide multidisciplinary research team, MGP will certainly trigger a new breakthrough in our understanding and control of common diseases. However, as emphasized in the first part of this article, big project-type research is adept at solving only a portion of the questions and problems; a lifeline of science and medicine is, after all, investigator-initiated, proposal-based research. Only from the rich soil of such idea-driven research will a huge national project like MGP crystallize, and its mission in turn should include an offering of seeds to inspire and boost future individual research.

On the other hand, what is clear from our experience in MGP is that the collection of a large number of samples with clinicopathological, demographic and life-style information and informed consent is truly critical for genome-based medical research in the post-sequencing era. Genome epidemiological study on a large-scale cohort is ideal for many common life-style related diseases and should be launched as soon as possible. For a relatively rare "common" disease like cancer with a defined histopathological type, hospital based case-control studies are also mandatory, but at the same time, ascertainment of family members such as affected relatives should be emphatically promoted.[53] It is also expected that we will witness an increasing number of multi-institutional clinical trials for new therapeutics or diagnostics in the next decades. These clinical studies will offer an invaluable opportunity to collect specimens with high-quality clinicopathological information. A scheme should be developed to centralize and share such clinical specimens and information as a national bioresource archive.

All these sorts of tasks are obviously beyond the scope of individual research and may need a nation-wide long-term effort guided by a proper central project management with clear and solid vision. It is our hope that MGP will pave the way to valid and productive human genome research at the crossroads of epidemiology, population genetics, bioinformatics and clinical and basic medicine.

**Acknowledgements.** We thank President Emeritus Takashi Sugimura, M. J. A., for his critical reading of the manuscript and for his advice. The work presented in this article was mainly supported by the Program for Promotion of Fundamental Studies in Health Sciences of the Organization for Pharmaceutical Safety and Research of Japan.

## References

1) http://www.ncbi.nlm.nih.gov/genome/seq/

2) Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., Sato, H., Sato, H., Hori, M., Nakamura, Y., and Tanaka, T. (2002) Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat. Genet. **32**(4), 650-654.

3) Sasaki, H., and Tanabe, C. (2001) Application of LCM to the unbiased whole genome amplification and expression profiling with cDNA microarray. Igaku-no-Ayumi **197**(13), 979-985 (in Japanese).

4) http://www.wma.net/e/policy.html

5) http://www2.ncc.go.jp/elsi/

6) Eng, C., and Ponder, B. A.(1993) The role of gene mutations in the genesis of familial cancers. FASEB J. **7**(10), 910-919.

7) Leushner, J., and Chiu, N. H. (2000) Automated mass spectrometry: a revolutionary technology for clinical diagnostics. Mol. Diagn. **5**(4), 341-348.

8) Pharoah, P. D., Antoniou, A., Bobrow, M., Zimmern, R. L., Easton, D. F., and Ponder, B. A. (2002) Polygenic susceptibility to breast cancer and implications for prevention. Nat. Genet. **31**(1), 33-36.

9) Relling, M. V., and Dervieux, T. (2001) Pharmacogenetics and cancer therapy. Nat. Rev. Cancer **1**(2), 99-108.

10) Kallioniemi, A., Kallioniemi, O. P., Sudar, D., Rutovitz, D., Gray, J. W., Waldman, F., and Pinkel, D. (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science **258**(5083), 818-821.

11) Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., Dohner, H., Cremer, T., and Lichter, P. (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. Genes Chromosomes Cancer **20**(4), 399-407.

12) Nishioka, M., Kohno, T., Tani, M., Yanaihara, N., Tomizawa, Y. *et al.* (2002) MYO18B, a candidate tumor suppressor gene at chromosome 22q12.1, deleted, mutated, and methylated in human lung cancer. Proc. Natl. Acad. Sci. U. S. A. **99**(19), 12269-12274.

13) International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature **15**(409), 860-921.

14) Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J. *et al.* (2001) The sequence of the human genome. Science **291**(5507), 1304-1351.

15) Marshall, A., and Hodgson, J. (1998) DNA chips: an array of possibilities. Nat. Biotechnol. **16**(1), 27-31.

16) Ramaswamy, S., Golub, T. R. (2002) DNA microarrays in clinical oncology. J. Clin. Oncol. **20**(7), 1932-1941.

17) Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. Nature **420**(6915), 520-562.

18) Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H. *et al.* (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature **420**(6915), 563-573.

19) Dermitzakis, E. T., Reymond, A., Lyle, R., Scamuffa, N., Ucla, C. *et al.* (2002) Numerous potentially functional but non-genic conserved sequences on human chromosome 21. Nature **420**(6915), 578-582.

20) Wright, A. F., and Hastie, N. D. (2001) Complex genetic diseases: Controversy over the Croesus code. Genome Biol.

**2**(8), COMMENT 2007.1-2007.8.

21) Antoniou, A. C., Pharoah, P. D., McMullan, G., Day, N. E., Ponder, B. A., and Easton, D. (2001) Evidence for further breast cancer susceptibility genes in addition to BRCA1 and BRCA2 in a population-based study. Genet. Epidemiol. **21**(1), 1-18.

22) Antoniou, A. C., Pharoah, P. D., McMullan, G., Day, N. E., Stratton, M. R., Peto, J., Ponder, B. J., and Easton, D. F. (2002) A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. Br. J. Cancer **86**(1), 76-83.

23) Neel, J. V. (1962) Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? Am. J. Hum. Genet. **14**, 353-362.

24) Wright, A. F., Carothers, A. D., and Pirastu, M. (1999) Population choice in mapping genes for complex diseases. Nat. Genet. **23**(4), 397-404.

25) Risch, N., and Merikangas, K. (1996) The future of genetic studies of complex human diseases. Science **273**(5281), 1516-1517.

26) http://www.ncbi.nlm.nih.gov/SNP/index.html

27) Haga, H., Yamada, R., Ohnishi, Y., Nakamura, Y., and Tanaka, T. (2002) Gene-based SNP discovery as part of the Japanese Millennium Genome Project: identification of 190,562 genetic variations in the human genome. Single-nucleotide polymorphism. J. Hum. Genet. **47**(11), 605-610.

28) http://snp.ims.u-tokyo.ac.jp/index.html

29) Hirakawa, M., Tanaka, T., Hashimoto, Y., Kuroda, M., Takagi, T., and Nakamura, Y. (2002) JSNP: a database of common gene variations in the Japanese population. Nucleic Acids Res. **30**(1), 158-162.

30) Marth, G., Yeh, R., Minton, M., Donaldson, R., Li, Q., Duan, S., Davenport, R., Miller, R. D., and Kwok, P. Y. (2001) Single-nucleotide polymorphisms in the public domain: how useful are they? Nat. Genet. **27**(4), 371-372.

31) Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat. Genet. **22**(2), 139-144.

32) Goldstein, D. B. (2001) Islands of linkage disequilibrium. Nat. Genet. **29**(2), 109-111.

33) Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., and Lander, E. S. (2001) High-resolution haplotype structure in the human genome. Nat. Genet. **29**(2), 229-232.

34) Jeffreys, A. J., Kauppi, L., Neumann, R. (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat. Genet. **29**(2), 217-222.

35) Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M. *et al.* (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science **294**(5547), 1719-1723.

36) Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J. *et al.* (2002) The structure of haplotype blocks in the human genome. Science **296**(5576), 2225-2229.

37) Weiss, K. M., and Terwilliger, J. D. (2000) How many diseases does it take to map a gene with SNPs? Nat. Genet. **26**, 151-157.

38) Couzin, J. (2002) Human genome. HapMap launched with pledges of $100 million. Science **298**(5595), 941-942.

39) http://genome.gov/page.cfm?pageID=10005336

40) Ohnishi, Y., Tanaka, T., Ozaki, K., Yamada, R., Suzuki, H., and Nakamura, Y. A. (2001) High-throughput SNP typing system for genome-wide association studies. J. Hum. Genet. **46**(8), 471-477.

41) Rothman, K. J., and Greenland, S. (1998) Modern Epidemiology. 2nd ed., Lippincott Williams & Wilkins, Philadelphia, p. 120.

42) Schlesselman, J. J. (1982) Case-Control Studies. Monograph in Epidemiology and Biostatistics, Oxford University Press, New York, p. 58.

43) Wacholder, S., Rothman, N., and Caporaso, N. (2000) Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. J. Natl. Cancer Inst. **92**(14), 1151-1158.

44) Knowler, W. C., Williams, R. C., Pettitt, D. J., and Steinberg, A. G. (1988) Gm3; 5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. Am. J. Hum. Genet. **43**(4), 520-526.

45) Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S., and Hirschhorn, J. N. (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat. Genet. **33**(2), 177-182.

46) Pritchard, J. K., and Donnelly, P. (2001) Case-control studies of association in structured or admixed populations. Theor. Popul. Biol. **60**(3), 227-237.

47) Pritchard, J. K., and Rosenberg, N. A. (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am. J. Hum. Genet. **65**(1), 220-228.

48) Satten, G. A., Flanders, W. D., and Yang, Q. (2001) Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. Am. J. Hum. Genet. **68**(2), 466-477.

49) Simes, R. J. (1986) An improved Bonferroni procedure for multiple tests of significance. Biometrika **73**(3), 751-754.

50) Sing, C. F., Haviland, M. B., and Reilly, S. L. (1996) Genetic architecture of common multifactorial diseases. In Variation in the Human Genome (Ciba Foundation Symposium 197). John Wiley and Sons, Chichester, pp. 211-229.

51) Ponder, B. A. (2001) Cancer genetics. Nature **411**(6835), 336-341.

52) Cyranoski, D. (2002) Geneticists lay foundations for human transcriptome database. Nature **419**, 3-4 (News).

53) Terwilliger, J. D. (2001) On the resolution and feasibility of genome scanning approaches. Adv. Genet. **42**, 351-391.