Whose Hands Are On Your Genes?

David R. Karp, MD, PhD

Internal Medicine Grand Rounds
UT Southwestern Medical Center

Friday, June 27, 2008

This is to acknowledge that David R. Karp, MD, PhD, has disclosed financial relationships with commercial concerns related directly or indirectly to this program. Dr. Karp will be discussing off-label uses in the presentation.

David R. Karp, M.D., Ph.D.
Associate Professor and Chief,
Rheumatic Diseases Division
UT Southwestern Medical Center

Internal Medicine Grand Rounds
June 27, 2008

Dr. Karp is the Chair of one of the Institutional Review Boards at UT Southwestern. His research centers on the pathogenesis of autoimmune diseases such as systemic lupus erythematosus. He is also the Co-Principal Investigator on the NIH-sponsored Immunology Database and Analysis Portal (*ImmPort*), a data-sharing tool developed for large-scale immunogenetic and functional genomic studies.

Introduction

The completion of the Human Genome Project (HGP) was formally announced in April 2003, coinciding with the 50th anniversary of the description of the DNA double helix[1, 2]. This thirteen year, multinational project was coordinated by the National Institutes of Health and the US Department of Energy[3, 4]. While this date was an important milestone in the scientific progress of the Genome Project, it was dramatically upstaged by an event three years earlier. At a White House ceremony that took place on June 25, 2000, with British Prime Minister Tony Blair participating by videoconference, President Bill Clinton announced the completion of a "working draft" of the human genome sequence. He was accompanied by James Watson, co-discoverer of the helical structure of DNA, Francis Collins, director of the NIH-led effort, and J. Craig Venter, the founder and then president of Celera Genomics. Celera had begun a race with the public effort to sequence the genome, with a goal to produce data for sale to pharmaceutical companies and other interested parties. Bowing to pressure from the scientific community, Celera eventually combined their data with public project.

Statements made at that press conference suggested this was an event of nearly Biblical proportions. The President began by saying, "Today we learning the language in which God created life." "We have caught the first glimpses of our instruction book, previously known only to God," echoed Dr. Collins. The official press release for that event was no less sanguine[5]. It announced that, "decoding the human genome will lead to new ways to prevent, diagnose, treat, and cure disease." "Scientists will be able to: ...Alert patients that the are at risk for certain diseases... Reliably predict the course of disease... Precisely diagnose disease and ensure the most effective treatment... Develop new treatments at the molecular level." Even the most casual observer of the event got a sense of the monumental importance and untold promise that lay in our understanding of human genetics.

In the past five years, the human genome sequence has been refined and we know the extent to which we are all similar and in some ways, genetically different[6]. The analysis of 'Human Genetic Variation' was hailed as "Breakthrough of the Year" by Science Magazine in 2007[7]. Numerous studies have associated one or more variant genes with many complex diseases such as rheumatoid arthritis[8-11], systemic lupus erythematosus[12-14], Type 2 diabetes[15], and many others. As was reviewed in this forum last year, some of these genetic differences are potentially useful predictors of individual response to medications such as warfarin[16] or azathioprine[17]. The entire March 19, 2008 issue of *JAMA* was devoted to genomic medicine and contains both a set of original studies related to the association of paraoxonase variants and cardiovascular disease[18], osteoporosis[19], and DVT[20], as well as reviews of genomic medicine for the clinician[21, 22]. However, there is much more work to be done before the promises of the Genome Project are met. If we are to accurately and reliably use genetic information to predict, diagnose, and treat disease, thousands of people will need to be studied. In each case, genetic information, imbued with almost mystical importance, and clinical or phenotypic information will be collected, combined, shared, and analyzed over and over.

Whether this research is done by a single lab headed by the patient's own physician, or more likely, by international consortia of people the subject has never met, the risks of genetic research are different than those encountered when a new drug is tested. Here the risks are not the chance that there will be ineffective treatment, or that the subject will suffer physical harm

from an adverse drug reaction. Instead, the risks relate to improper use of information. Will the analysis of someone's DNA and clinical information be done in a way that respects their autonomy? Will the results of this analysis harm them socially or economically? Will this create a situation where whole groups of people are stigmatized by virtue of shared genetics?

Human Genetic Diversity

The human genome contains approximately 3 billion base pairs arranged across 24 pairs of chromosomes – 22 autosomes and the X- and Y-chromosomes. It is estimated that there are 20,000 – 25,000 genes in the human genome [2]. Given the average size of a functional gene, only about 2% of the genome directly codes for the production of proteins. Repetitive sequences – so called, "junk DNA" – make up almost 50% of the genome. While they do not code for proteins, they shape the structure and organization of genes on a chromosome throughout evolution, and may also code for "micro RNAs" that regulate gene expression.

Several different kinds of genetic variation exist between different individuals. The most basic is the single nucleotide polymorphism, or SNP. Each of these is a one base difference in the genome sequence that occurs naturally from individual to individual. They are present on average every 300 bases. Over 10 million commonly occurring SNPs have been described in the human genome. While some SNPs alter the coding sequence of the gene and thus the amino acid sequence of the protein, many occur in non-coding areas of the genome. Some of the non-coding SNPs may still have functions such as increasing or decreasing the transcription of a particular gene, although many are merely silent markers of genetic variation.

The combination of SNPs in a genome is not entirely random. Due to linkage disequilibrium, specific patterns of SNPs tend to be inherited in large blocks, termed haplotypes (Figure1). Using haplotypes, it is possible to predict large amounts of sequence variation from a smaller number of "tagging SNPs". Thus, it is possible to use a set of only 300,000 to 600,000 SNPs to predict the entire variation in the human genome instead of testing all 10 million. The description of the human haplotypes was the result of the International HapMap Project [23]. DNA from Utah families of European descent, Yoruban families from Ibadan, Nigeria, and individuals from Tokyo and Beijing was used. The second major analysis of human haplotypes was published in 2007, characterizing over 3.1 million SNPs[24]. This dataset is available on-line. It represents a wealth of information about the nature of human genetic diversity and evolution, and is an invaluable resource for planning and analyzing genetic association studies.

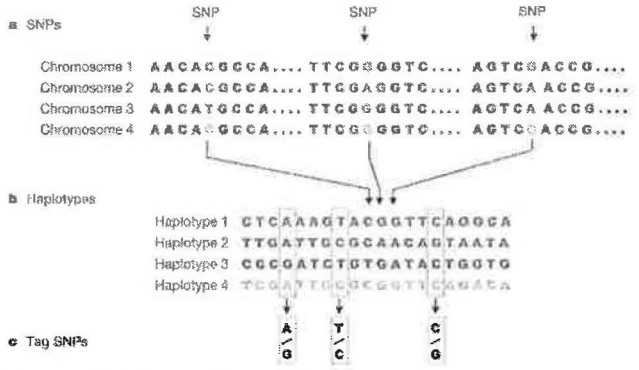

Figure 1. Formation of Haplotypes from tagSNPs
www.hapmap.org

One virtue of SNP typing is its amenability to large scale, high-throughput technologies. Using chip-based systems, it is possible to analyze from 100,000 to 1 million SNPs at one time from a single DNA sample. The cost to perform such an analysis is not cheap – on the order of $500 per sample. It is sufficiently inexpensive, though, to allow researchers to plan and execute studies involving thousands of patients

and controls. The amount of data produced has led to the development of new informatic tools that can analyze billions of pieces of information at one time. A thorough review of the HapMap project and its impact on medical genetics has recently been published[25].

Prior to the development of SNP typing, the most common method of assessing genomic variation was to measure the size of tandem repeats or microsatellite DNA. These are regions of repeating two to five nucleotide units. They are often located between the sequences of functional genes, but can occur within genes, as in the case of Huntington's Disease. The number of each repeat can vary from individual to individual and thus the size of a DNA fragment at that region is a genetic marker for the region as a whole. This type of genetic typing is used primarily for forensic analysis such as paternity testing and criminal identification.

Other forms of genetic variation include variation in the number of copies of a given gene in the genome, insertions and deletions of large amounts of genetic material, and translocations of whole segments of chromosomes. These non-SNP forms of genetic variation may account for up to 20% of the inter-individual differences in our genomes[7].

Genome Wide Association Studies

SNP typing of individuals can focus on one or a few "candidate genes" obtained from previous genetic studies, from knowledge of the pathophysiology of a given condition, or just from educated guesses. These types of studies are looking to support a particular hypothesis. The ability to analyze most of the relevant tag SNPs in the genome of thousands of individuals has led to the development of different type of genetic analysis, the genome wide association study (GWAS), also called whole genome analysis (WGA). In this study, the DNA of many unrelated individuals with a condition or trait are compared to those without the condition. Usually, 100,000 or more SNPs are tested on each individual. These studies are hypothesis generating. That is, the goal of the study is to pick previously unknown genes for future analysis. An assumption of most GWAS is that complex diseases are the result of unfavorable combinations of common genetic variation throughout the population. In most conditions like rheumatoid arthritis, lupus, and diabetes, the number of genes found to be associated in this manner is typically ten or more. While many of these associations are statistically impressive, the contribution of each to the genetic risk is low, on the order of 1.2 to 2 fold. For example, variation in the gene for interferon response factor 5 (*IRF5*) has been associated with SLE in several different studies of multiple ethnic groups[14, 26-32]. Statistical significance is excellent with p values less than 10^{-20}. The prevalence of the risk allele in patients is 61% while in controls it is 51% giving an odds ratio of 1.47[26].

Genome wide association studies are also prone to identifying false positive results[33-37]. To begin with, one must mathematically correct for the 'false discovery rate' that will occur with hundreds of thousands of markers are tested simultaneously. The traditional $\alpha = 0.05$ must be lowered to 10^{-7}-10^{-8}. Second, the variants being tested are fairly common in the population. Thus, differences in geographic or ethnic origin in the controls versus the subjects can introduce findings as strong as true positive[38]. Therefore, replication in different populations is necessary before a given SNP association can be believed. These two characteristics of genome wide association studies – the need for large numbers of subjects to achieve statistical significance and

the need to replicate studies – are the prime reasons for the collection, combining, and sharing or large DNA banks.

Individual Genomes

What about just sequencing the whole genome from each person? Given that approximately 99.5% of the DNA sequence is the same for every human, this would be an enormous expense for very little return. Yet, this would yield an exact picture of genetic variation without relying on the statistical associations of tag SNPs or the unknown presence of insertions, deletions or copy number variation. The human genome sequences from both the Human Genome Project and Celera represented mixtures from several individuals. There are a few examples of published sequences from identified persons. Notably, Craig Venter had his genome sequenced and has published a book on what it tells us. The genome of DNA pioneer James Watson was sequenced by a company using him as the subject of a new method of DNA sequencing. These are both available on the Web for inspection. Each of these genomes cost several million dollars to produce, still less than the tens of millions spent by the Human Genome Project. Knome, Inc. (www.knome.com) is a private, for-profit company that will sequence your genome for $350,000. The X Prize Foundation, which is more well known for sponsoring challenges in spaceflight and innovative auto design has a challenge to create a technology capable of sequencing 100 individual genomes in 10 days for less than $10,000 per genome[39]. The prize is $10 million. To date, seven companies have registered to compete. The 1,000 Genomes Project (www.1000genomes.org) is an international consortium aimed at better defining the human genetic diversity by obtaining complete sequences from the samples in the HapMap project and others.

Why is genetic information different?

Why is genetic data about a person felt to be more important, more worthy of protection than, say a serum glucose or hemoglobin content? Even fairly unsophisticated people understand the idea that genes are the ultimate personal identifiers. News media report the conviction of criminals or the freedom of innocents based on DNA evidence. Television shows like 'CSI' show people being identified in minutes based on the analysis of a few epithelial cells left at a crime scene. The medical press reports on the ability of DNA analysis to predict or diagnose diseases. Thus DNA becomes the ultimate descriptor of the person. Because it is inherited, DNA also identifies us with our past – our families – our race and ethnicity – our religion and our tribe. It also connects us with our children. Whatever lies in our genes, known and unknown, is a reflection of our ancestry and a legacy for our heirs.

This has led to the concept that genetic data is a "diary of the future"[40]. Someone knowing your genomic sequence or SNP haplotypes could theoretically predict your chance of developing diseases ranging from Alzheimer's to alcoholism. Depending on the disease or condition that is being studied, there is a variable level of truth to that statement. Some conditions are monogenic with complete penetrance while others can and probably do arise from variation in many genes acting in concert. Epigenetic effects on gene expression and the influence of the environment change the degree of genetic determination so that even identical twins do not necessarily share the same destiny. Yet, for most purposes, the DNA sequence we are born with is the one we

have all our life. As more knowledge is gained an individual's genetic variation can be re-analyzed and re-interpreted. A DVD with your genomic sequence obtained at birth could actual be the "prequel" to your medical record.

Personal Genomics

The public fascination with genetics has led to the idea that the genomic sequence of an individual is a biomarker, a surrogate, for the combined self. In it we can find details of our ancestry, our medical past, predictions of our future, clues to our appearance, and even an estimation of our lifespan. That enthusiasm, coupled with an ever-expanding database of genetic associations – some stronger than others – and the genius of Internet marketing has resulted in the proliferation of companies ready to do personal genetic profiling. For ~$1000 and a vial of saliva, companies like deCodeMe, 23 and Me, and others (Table 1) will provide you with a 1,000,000 SNP analysis of your genetic variation. You can download and keep the raw data to peruse at your leisure, but they will also provide several online analyses for you. You can learn if you have variants associated with thousands of traits. Your genetic makeup can be compared to that in parts of Africa or Europe. You can enroll your whole family at a discount and see where you got the gene for nicotine addiction.

So the first answer to the question in the title of this talk is that one pair of hands on your genes is yours, if you like. So, what can you do with this information? Much of it is at the level of an expensive horoscope. Given that most of the associations are from non-replicated studies, and the relative risk is low, there really isn't much utility of the information a layperson can gain perusing their SNP genotype. Using DNA to trace your family's roots across Europe likely gives the same information your grandmother did. While some companies specialize in tracing the origins of African Americans, there is considerable variability from one company to another based on the quality and geographic origin of their reference samples. Tracing inheritance through your family tree is a great way to prove Gregor Mendel right, but there is the occasional, but real, risk that non-paternity will be revealed.

One of the most concerning aspects of personal genomics is the direct marketing of clinical

Table 1 Direct to Consumer DNA Testing (partial list)

Company	Web Site	Services
deCodeme	www.decodeme.com	Ancestry, disease risk, 1M SNP genome
23andMe	www.23andme.com	Ancestry, disease risk, 0.5M SNP genome
DNA Direct	www.dnadirect.com	Ancestry, disease risk, drug metabolism
National Geographic	www3.nationalgeographic.com/genographic	Ancestry
Family Tree DNA	www.familytreedna.com	Ancestry
Genelex	www.healthanddna.com	Ancestry, nutrition, drug metabolism, disease risk
Navigenics	www.navigenics.com	Disease risk, subscription updates, 1.8M genetic variants
Myriad	www.myriadtests.com	Hereditary cancers: BRCA1/2, HNPCC, APS, melanoma

informative tests[41, 42]. Companies will do testing for BRCA1 and BRCA2 mutations associated with breast and gynecological cancer, for HNPP variants associated with colon cancer, and ApoE alleles associated with a form of familial Alzheimer's Disease. In some of these cases, a consultation with one of their geneticists is needed before they will do the test. In others, such as the ApoE4 testing, no consultation is required[43]. Genelex (www.healthanddna.com), for example, advertises their services to both consumers and health care providers. Patients can be tested for pharmacogenomic variants such as cytochrome P 450 alleles and other variants, which is touted as necessary for the proper dosing of medications such as anti-depressants, beta blockers, anti-arrhythmics, and opioids. This is a free market and one that is only partially regulated by the FDA and the Federal Trade Commission, but one that has recently attracted the attention of both bioethicists[44] and state regulators[45]. New York and California officials are investigating these companies for the provision of genetic testing without the involvement of a physician. The companies argue that their direct to consumer services are 'educational' and not medical. While the purveyors of personal genomics tout the right of each of us to have this information, it is debatable whether much benefit will come from having it delivered in its current form.

Perceived Risks of Genetic Research

For the promise of the Human Genome Project and the concept of personalized medicine to be realized, patients must choose to undergo genetic testing free of reprisal or discrimination. In a practical sense this means discrimination in employment and insurance. While this topic has received considerable attention, most of the data are both anecdotal and dated. A 2004 report from the Coalition for Genetic Fairness lists cases where health insurance was denied to children with asymptomatic Long QT Syndrome, carriers of alpha-1 anti-trypsin deficiency, and women with the BRCA-1 mutation[46]. A study published in 1992 looked at cases of denied or cancelled health insurance in response to surveys sent to over 1,000 genetic counselors as well as requests for information published in patient advocacy newsletters[47]. They found 29 cases for analysis. In many cases, the patients were asymptomatic or had mild phenotypes. The authors conclude that there is great misconception on the part of insurance companies and other social institutions regarding the seriousness of genetic conditions.

It is similarly difficult to find good data on the extent of genetic discrimination in employment. A 1998 review by Miller listed several non-systematic surveys revealing hundreds of un-verifiable cases of genetic discrimination[48]. That it happens rarely doesn't mean employers don't want to use genetic information. In the 1990's, Berkley Lawrence Laboratory was sued for conducting tests for pregnancy, syphilis, and sickle cell without employees' consent. While the case was really about testing without the employees' knowledge, it ignited state and federal legislative attempts to ban genetic discrimination. In 2001, the Fort Worth based Burlington Northern Santa Fe Railway (BNSF) settled a $2.2 million lawsuit with the Equal Employment Opportunity Commission (EEOC) based on BNSF's mandatory testing of employees who filed worker's compensation claims for carpal tunnel syndrome (CTS). Company physicians evaluating these claims obtained blood samples from the claimants without telling them what they were for [49]. One worker was threatened with dismissal if he refused to provide a sample. The vast majority of cases of CTS are idiopathic or work-related. The condition hereditary neuropathy with liability to pressure palsies (HNPP) is an autosomal dominant disorder due to a

deletion part of chromosome 17 containing the gene for peripheral myelin protein, *PMP22*. CTS is a common symptom of HNPP, which typically presents in childhood or adolescence[50]. The prevalence of work-related CTS is estimated to be 530/100,000 workers while the prevalence of HNPP and the *PMP22* deletion is unknown but estimated to be only 16/100,000. A study of 50 unrelated CTS patients failed to find any *PMP22* deletions setting the upper limit of only 6% of CTS cases caused by sporadic HNPP[51]. Thus, none of the ~20 workers tested would be expected to harbor this genetic marker.

Even if economic discrimination on the basis of genetic testing is a rare event, it is a major concern for the public. Those with a family history of breast or colon cancer have indicated that fears of employment or insurance discrimination would prevent them from being tested, thus losing their opportunity to undergo preventative treatment or monitoring[52-54]. For example, Apse, *et al.* reported on 470 unaffected relatives of patients in the Johns Hopkins Hereditary Colorectal Cancer Registry[55]. Only 18% had no concerns about genetic discrimination being used against healthy individuals; 45% rated their concern as high. 7% of respondents claimed they had already experienced genetic discrimination. Most of these related to inability to obtain health or life insurance, or higher premiums. 79% received information about genetic discrimination from media sources such as the Internet. The majority of people in the study were interested in genetic testing for themselves, yet 35% said they would pay for the tests out of pocket to avoid submitting claims to insurance companies and 15% would use an alias during testing. 47% wanted genetic test results excluded from their medical record. Even health care professionals have concerns about discrimination. Genetic counselors were surveyed about their attitudes toward BRCA1 and HNPCC testing if they or their families were at risk. 68% said they would not bill their insurance carrier and 26% would use an alias to avoid discrimination[56].

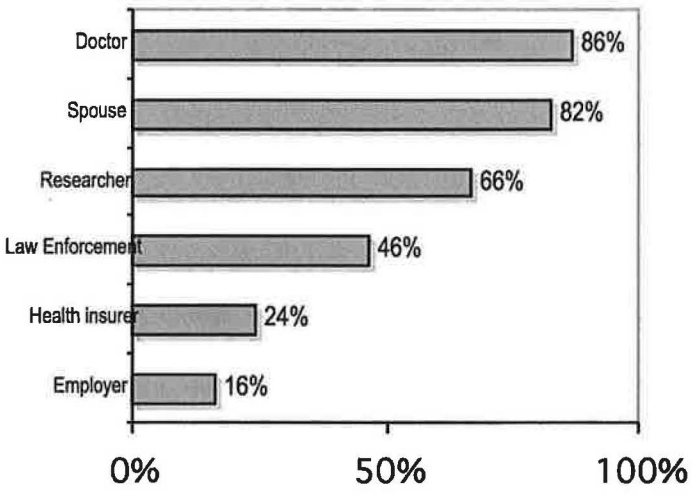

Figure 2. Percent of groups trusted with genetic test results Adapted from ref 50.

A 2007 online survey of 1,199 of randomly selected US adults was conducted by the Genetics and Public Policy Center[57]. In general, there was strong support of genetics to support health care, while nearly all (92%) of subjects feared the test results could be used in a manner to harm subjects. 93% of participants felt that neither employers nor insurers should be able to use genetic testing results to make decisions about hiring or insurability. Approximately three-fourths of them wanted laws that prevent these actions.

Fortunately, there is a remedy for the possibility of genetic discrimination in health insurance and the workplace. As of January 2008, 35 states have laws that prevent an employer from using genetic information in the hiring, firing, or setting workplace conditions for employees. 45 states

have laws prohibiting the use of genetic information for health insurance purposes. Both the Texas Labor Code and Insurance Code contain provisions that prohibit the use of genetic information by employers (more than 15 employees), health plans, and licensing agencies in making decisions. As Dr. Igarashi announced in this forum, Congress finally passed the Genetic Information Non-Discrimination Act of 2007 which was signed by President Bush on May 21[58]. This legislation prohibits the use of genetic information by group health plans for underwriting purposes, and prohibits employers, employment agencies and labor organizations from using genetic testing to fail to hire, discharge, or unfairly classify employees. Genetic information is now included as protected healthcare information (PHI) under the Health Insurance Portability and Accountability Act of 1996 ("HIPAA"). Genetic information obtained by participation in clinical research protocols is specifically included in the Act. Since HIPAA does not allow future, unspecified, uses of PHI without authorization, the ongoing use of stored DNA samples for genetic research beyond what was put in the original informed consent forms will require that the DNA be de-identified. As discussed below, this is likely to be a controversial issue going forward.

The Scope of Genomic Research

To understand the ethical concerns regarding population genomic research, it is useful to look at the size and scope of the effort worldwide. Unlike genetic studies of the past where a single investigator collected DNA from families or individuals who were their own patients or with whom they had established a close relationship, biobanks by their design collect specimens from many investigators or health care providers and aggregate them for future study. There are some distinctions between specimens such as DNA (limited quantity, physical ownership, etc) and data derived from the specimen, but many of the ethical concerns are the same.

A 1999 Rand study estimated that there were over 300 million samples in biobanks in the United States with over 20 million samples added each year[59]. Biobanks have been created by academic institution such as the NUGene Project at Northwestern University that has a goal to obtain DNA and selected medical information from 100,000 individuals. At Vanderbilt University, DNA is extracted from left over blood specimens and is linked through an encrypted code to a derivative of the subject's medical record[60]. Similar projects are in place at the Marshfield Clinic (40,000 subjects), the Mayo Clinic, Duke University, and others[61].

There are private (commercial) biobanks designed to provide data to for-profit corporations. Nearly three-fourths of all investigational new drug submissions to the FDA include some provision for storing samples by the sponsoring pharmaceutical or biotechnology company. Perhaps the most famous commercial biobank is deCode Genetics, a company that originally had an exclusive license to link the health records of all 270,000 Icelanders with genetic data and family relationships[62, 63]. The ethical issues associated with commercial biobanks are multiple, including the rights of subjects to benefit from commercialization efforts using their data. Private companies typically assure subjects that their privacy will be protected, but they are not under any legal obligation to do so. Moreover, specimens and data are corporate assets. When DNA Sciences went bankrupt in 2003, Genaissance Pharmaceuticals bought the DNA samples and medical information on 18,000 individuals for $1.3 million[61]. Typically, purchasers of such assets regard any contract between subjects/customers and the seller as non-binding on their part.

Table 2. Characteristics of Several Population-Genomic Studies

Project	Location	Description
CARTaGENE	Quebec, Canada	50,000+ adult participants
Estonian Genome Project	Tartu, Estonia	DNA and health data from 1 million adults and children
Latvian Genome Project	Riga, Latvia	60,000 samples
UK Biobank	Manchester, United Kingdom	DNA, health data, environmental data from 500,000 adults. 30 year follow-up
Medical Biobank	Umeå Sweden	DNA and health records on 85,000 adults
Singapore Tissue Network	Biopolis, Singapore	250,000 participants
Biobank Japan	Kanagawa, Japan	300,000 adults with 30 common illnesses

Adapted from Maschke KJ. Nat biotechnol 2005;23:539-45

Finally, regional or national efforts constitute the largest group of holders of biological specimens and medical data (Table 2). The governance structure, consent processes, and intellectual property aspects of each of these biobanks is slightly different. A large-scale population genomic project in the US has been proposed[64]. The collection of samples and data from 500,000 and 1 million subjects who would be followed prospectively for 10 years received support from an advisory committee to the Secretary of Health and Human Services[65]. Given the current funding climate at the NIH, however, it is unlikely that a project of this magnitude will happen soon, if at all. In the meantime, the NIH has developed other mechanisms to leverage as much population genetic data from existing studies as possible (*vide infra*).

Evidence of the widespread nature of genomic bio/databanks is the existence of groups that "collect the collections". For example, the Public Population Project in Genomics (P^3G) is an international consortium designed to catalog not only the characteristics of the database, but also the specific features such as questionnaires used and types of genetic study, along with software tools and models of ethical guidance[66]. They hope to promote interoperability of these databanks so that statistical power and replication can be facilitated. Currently, 109 studies with target accrual of 11 million participants are part of this project.

Privacy, Confidentiality, and Autonomy

The risks of participating in genetic research are centered on the issues of privacy, confidentiality and autonomy. These risks are mitigated though attention to the design of the research, the research infrastructure, and most importantly, the information given to subjects when they enroll. Privacy is the ability of the individual to withhold information about themselves from others. The ultimate in privacy is anonymity where no information can be linked to an identifiable person. In the context of clinical research, including genetic studies, confidentiality refers to the efforts taken by the investigators to protect the privacy of subjects. Autonomy is the control that subjects have over their personal information. This includes the control that subjects have over the use of their specimens and data for future research.

The concern that most people have when they participate in genetic research is the possibility that their identities will become known and associated with the clinical information they provided. This is not difficult provided there is a reference genotype to compare to. It takes less

than 100 SNPs, to uniquely identify an individual[67]. If a subject is present in two databases, a match can easily be made between an identified list and de-identified clinical characteristics [68]. This possibility is only limited by the lack of publicly accessible genomic data. With the explosion of large research databases it is increasingly possible that match will be made between de-identified and identified datasets, including those held by the military and law enforcement.

Table 3. Terminology of Sample Identifiablity

Term	Synonyms
Identified	Nominative
	Personally identifiable
Coded	Linked
	Reversibly anonymized
	Traceable
Non-Identifiable	Unlinked anonymized
	Unidentifiable
	Anonymous

Adapted from ref 69.

It is worth discussing the terminology used to describe various levels of data identifiablity (Table 3). First, there is identified data where a name, medical record number, etc., is part of the data record. Second, coded (or identifiable) data has the obvious identifiers removed and replaced with a code that can often be linked back to the individual. Third, anonymized data is similar to coded data but the link has been destroyed. Finally, there is truly anonymous data that never had personal identifiers attached to it in the first place. There is little agreement on the actual terminology used to describe these types of data, and less so on the policies that govern their use[69]. In the United States, one of the relevant issues is that research using data (or specimens giving rise to data) that are anonymized is not human subjects research according to a 2004 Office of Human Research Protections (OHRP) guidance document[70]. In effect, this makes such research outside the statutory regulation of an IRB. By policy, most IRBs, including ours, reserve the right to make the determination on a case-by-case basis.

Phenotype

The risk of subject re-identification is not just related to the matching of genotypes in a de-identified database with ones in database with identifiers. There is significant risk associated with the phenotype linked to the genotype. In some sense, this is a greater risk, depending on the richness of the data. In the simplest association study, the only phenotype present may be gender and affection, i.e., either the condition under study is present or not present. With complex genetic studies there may be associations with subsets of disease: Lupus patients with thrombocytopenia or nephritis, for example, or HIV patients co-infected with HCV. There may be significant genetic associations that are only apparent when certain environmental exposures are taken into account. In rheumatoid arthritis, the at-risk HLA alleles have the strongest effect in subjects with a history of significant cigarette smoking, which in turn may be a risk factor for autoantibodies to citrullinated proteins[71, 72]. Thus to understand the full impact of gene-environment interactions, it is necessary to accumulate a detailed clinical phenotype.

In some cases, the ability of phenotype to identify individuals is obvious. The listing of specific job titles like, "Chairman of Internal Medicine", or combinations of sufficiently rare descriptors such as, "five-year old Ethiopian boy with Mucolipidosis II," will easily identify individuals within the organization collecting the data. However, combinations of data elements that seem to be so common as to be un-linkable to a single individual can be used with great certainty. Approximately 87% of the US population is uniquely identified by the combination of date of

birth, gender, and 5-digit ZIP Code[73]. In one case, privacy expert Latanya Sweeney was able to combine public information on the health care expenses for employees of the Commonwealth of Massachusetts and City of Cambridge voter registration data to uniquely identify the medical records of the then Governor, William Weld. Over 67% of the population of Cheyenne, Wyoming could be uniquely identified though their family relationships published in the obituary section of the *Wyoming Tribune-Eagle*[74].

While the Health Insurance Portability and Accountability Act (HIPAA) may control some of this information, it is not perfect. Over 75% of individuals in a sample of 23 'de-identified' records in the Illinois state cancer registry could be uniquely identified using month/year of diagnosis, type of cancer, and 5-digit ZIP code. Using an analysis of 'trails', the supposedly de-identified hospital discharge information that is available in public or quasi-public sources, Malin and Sweeney were able to uniquely identify between 30% and 100% of individuals with genetic disorders such as cystic fibrosis, phenylketonuria, and Huntington's Disease[75].

How Can Privacy Be Protected?

It is possible to preserve individual privacy while still promoting or even enhancing research. Likely, several different strategies will need to be employed: regulatory, technological, and policy driven. HIPAA is the primary regulation in the US affecting the de-identification of samples placed in databases. Explicit identifiers (name, address, SSN, etc.) are removed and replaced with a code that may or may not be linked to the individual. Privacy is then the responsibility of the code holder. As noted above, other phenotypic features may be sufficient to establish identity.

Computational strategies can be used to protect genomic or phenotypic data. On a person-by-person basis, a decision can be made to suppress or generalize (e.g., substituting 'North Texas' for 'Dallas') items of data. There are currently methods to accurately suppress the ability of location visit trails to link de-identified DNA to personal data[75, 76] or to produce an accurate account of an electronic medical record specifically altered to eliminate identifiers while preserving both the clinical information and context. Unique genetic records from individuals can be combined so that they are indistinguishable from the others. For example, a polymorphic sequence, AACT or AATT becomes AAYT where Y is either pyrimadine. While this protects individual privacy, it still allows the recovery of aggregate data for statistical analysis. Another computational method to protect privacy is to keep all data within a very secure computing environment. Users are allowed to query the database for aggregate information (e.g., gene frequencies) and to do analyses, but not to download complete records on individuals.

Table 4. Methods to Protect Privacy in Genomic Databases

Methodology	Examples
Regulatory	• HIPAA de-identification • Genetic Information Non-Discrimination Act • Certificates of Confidentiality • Freedom of Information Act limits
Computational	• Masking or binning of data • Secure server prevents data release
Policy	• Data use agreements • IRB oversight

Finally, database policies will protect individual privacy, albeit only to the extent that they are

enforced. Access to potentially identifiable data can be tightly controlled by use agreements supervised by data use committees. These agreements limit the persons who can see or use the data, the allowed uses, and prohibitions on re-distribution. Typically, such agreements require that an IRB review the proposed use and agree that it will be consistent with the terms of the original data or specimen collection. It isn't clear how database holders can enforce such data use agreements, but sanctions by them or by funding agencies (e.g., the NIH) would seem likely.

Informed Consent is a Problem for Population Genomics

Respect for subject autonomy is one of the guiding principles of research involving human subjects. Typically that is embodied in the informed consent process where subjects are told exactly what is going to be done in the course of a particular research project and they agree – usually in writing – to participate. Participants are informed of all the risks associated with the study, both common and rare. They are told of the process to discontinue participation in the research at any time, the chance that commercial products may come from the research, and how their confidentiality will be protected. It is usually obvious how to apply these principles to the typical clinical trial or physiological study. It is not clear whether and how they apply to modern population-based genetic research. As noted above, genome-wide association studies consist of many thousands of participants. Often, researchers will aggregate their DNA specimens or the data derived therefrom (there are some differing issues relating to biological specimens versus data, but they will be treated the same in this discussion), or pass the specimens/data on to other investigators. Not every research question will be known at the time of original informed consent, and the original investigators cannot predict how a specimen/data will be used. Thus, it is probable that a person's DNA and whatever clinical information is attached to it will be studied by researchers unknown to them in order to answer questions unrelated to the original donation.

It has been argued that this type of research carries with it no more than minimal risk to subjects[77]. The physical risk is merely peripheral venipuncture or collecting saliva. There is virtually no chance that medically useful data will be discovered for an individual participant. The sheer number of subjects in a study minimizes the importance of any one. Still, there is chance that future research will take place on topics that current subjects would not approve of, such as mental illness, racial or ethic variation, or sexually transmitted diseases. Subjects may have a legitimate claim to both useful information and commercial applications that come from their participation[78].

This tension between protecting subject autonomy and facilitating research can be dealt with in several ways. First, biobank or database research could be presumed to be no different from any other research study, and the consent thought of in the same way. All the principles from the Nuremberg Report, the Declaration of Helsinki and the Common Rule apply, giving individual rights supremacy over public good. This strict protectionist approach advocated by some ethicists mandates that if a future use isn't spelled out explicitly in the consent form, then subjects must be re-consented before the new research can go forward. This is an unwieldy approach when thousands of participants would have to be contacted. Moreover, this can actually be a burden to subjects when if they are contacted regularly for consent on studies A, B, and C using the DNA they donated years before. While this option seems unworkable to most

researchers, a 2000 study by the UK Human Genetics Commission found that this was the option favored by 82% of the public they surveyed[77].

Another method of respecting subject autonomy is to garner community consent. This may be an important option when specified groups, such as children with a given genetic disorder or individuals of a single nationality or ethnic origin are studied. This method is used by the US Department of Veterans Affairs to allow veterans' organizations to have input into the research done using samples and information from their DNA bank[79], as well as the UK Biobank and CARTaGENE.

The most controversial method of addressing this issue is the concept of blanket or general consent[77, 80, 81]. In this way, subjects are asked to agree to any and all uses of their DNA or genetic information in the future. Several groups and authors have argued that this major revision of generally accepted features of informed consent is the only way that population genetic research can realistically be done. Implicitly or explicitly, the traditional notion of informed consent is replaced by a less strict standard – one that some have said should not be called informed consent at all. As Arnason notes, "The more general the consent is, the less informed it becomes."[62]

Hank Greely, a bioethicist at Stanford Law School who recently gave Ethics Grand Rounds here at UT Southwestern, has summarized the arguments against blanket or broad consent[78]. He points out that asking subjects to consent to unspecified future research with the promise of confidentiality is ethically suspect given the ability of subject identities to be learned though a combination of genotype and phenotype, and legally improper as it contradicts the Common Rule, the Federal Regulation governing human subjects research. He further argues that Common Rule takes precedence over the OHRP guidance saying that research involving appropriately coded or de-identified specimens or data is not human subjects research (*vide supra*). Lastly, the moral constructs of justice and subject protection by researchers may require that genomic research be done in a way that subjects could, in fact, be identified. If truly clinically important information were discovered though a genome wide survey, isn't it the right of the subject to receive it? If samples or data are not linked to the subjects in some manner, how can they withdraw their participation or limit their use?

If blanket consent is not ethically or legally permissible, how can subject rights and desires be respected while promoting science? One option is to allow subjects to exercise their autonomy by first consenting to the collection of DNA and medical information and waive their right to consent to future research. Instead, they *authorize* the use of their DNA and information for certain purposes and not others. This often takes the form of options the subject must choose in the consent form:

> "In addition to allowing Dr. X to study disease Y, I give my permission for Dr. X to share my de-identified DNA with other approved researchers to study other conditions. ___ Yes; ___ No."

To be accurate, these statements are necessarily vague, and do not meet the standards of true informed consent. These authorizations could also include conditions when the subjects wish to

be re-contacted with new information, or if significant commercial developments are contemplated.

This method is neither as restrictive as traditional informed consent nor as liberal as blanket consent, but it implies an ongoing relationship between investigators and subjects. To make this relationship both meaningful and workable, there has to be a process to monitor and approve use of specimens or data. This could take the form of individual IRBs, but given the regional or national nature of genomic databases, it often requires an oversight body such as a data access committee. Most large-scale biobanks have such a committee, e.g., the UK Biobank Ethics and Governance Council[82]. The solution proposed by the NIH will be discussed below.

Table 4. Models of Informed Consent for Genomic Research

Method of Consent	Issues
Specific (one time) Consent	Limits research
Re-Consent	Difficult to do; may limit research; may actually annoy subjects
Presumed Consent	Lacks respect for persons or groups; subjects have to 'opt-out'
Community Consent	Hard to define; requires high degree of organization
Blanket Consent	Requires subjects to give up future rights and concerns
Authorization model	Spells out exactly what can be done without re-contacting subjects or providing them with more information

An Example of How Presumed Blanket Consent Goes Wrong

Failure to respect the autonomy of subjects can have long-ranging consequences. One of the most striking examples involves research done with the Havasupai, a small Native American tribe whose members live in the western Grand Canyon. Like many native peoples, the Havasupai have a high prevalence of Type 2 diabetes mellitus. In the early 1990's, researchers from Arizona State University approached the tribe to obtain blood samples to examine the genetic contribution to diabetes in their community. The informed consent the tribe members signed mentioned research into "behavioral/medical" problems, while the oral presentations given to the tribe stressed diabetes. In 2003, tribe members learned of research on schizophrenia and genetic anthropology done using their samples. The Havasupai oral history places the creation of their people in the canyon. The research with their samples was used to support the hypothesis that the indigenous peoples of North America migrated from eastern Asia. Outraged tribe members sued the university for $50 million[83, 84]. Although the suit has been dismissed by state courts[85], the case has been appealed and the ability to do genetic research with any of the native people of North America has been severely compromised[86], as tribal IRBs have essentially halted all prospects for this type of study.

Even the National Geographic Society has suffered from the lack of trust that native people have with biomedical research. They have launched the Genographic Project to collect 100,000 DNA samples from indigenous people around the world in an attempt to map human diversity and migration before the combination of modern mobility and admixture makes this impossible. US researchers have collected a pitifully small number of DNA samples from people in Alaska and other native peoples before participant mistrust basically halted the project[87]. These cases

illustrate that while genetic research may not carry physical harm, or even risk of individual identification, the risk of stigmatization of families or groups is real and can have far reaching effects on both subjects and researchers alike.

The National Institutes of Health 'Solution'

As the major source of bioscience research funding in the United States, the National Institutes of Health (NIH) also has a major influence on how research is conducted. In 2003, the NIH instituted a Data Sharing Policy[88]. Any investigator-initiated projects with direct costs of $500,000 or more per year are expected to make their final research data available to the public. The rationale behind this decision seems laudable. As promoted by the Office of Extramural Research, data sharing reinforces open scientific inquiry, encourages diversity of analysis and opinion, enables exploration of topics not originally investigated, and allows the creation of new datasets from multiple existing ones[89]. It also demonstrates that the NIH is a good steward of public resources by promoting practices that avoid costly duplication of research efforts, or fail to completely exploit the utility of unique research findings.

While noble in purpose, the implementation of this policy has been met with many opinions from the scientific community. Some researchers are making every effort to comply, some are in fierce opposition to data sharing, and many are just confused. Some investigators feel that data they have generated, no matter what the funding source, is theirs to control in perpetuity. Allowing other researchers to look at their data before they are done with it potentially opens them to being "scooped" with research findings and publications, or put at a disadvantage when it comes to using their proprietary information as preliminary findings in a new grant application. Their feeling is that they will share their data by publishing selected pieces of information over time.

While publication of a summary of research results in a peer-reviewed journal is one way to share data, it only begins to address the spirit of the policy. The NIH really expects that the "final research data", meaning any factual information needed to replicate the study analysis, are made available no later than the first publication of the major research findings. Investigators are not expected to have prolonged and exclusive use of the data, although provisions are made for the appropriate protection of intellectual property.

All data is potentially subject to sharing, including genetic information about individuals and extensive health information collected in the process of a clinical trial or genetic mapping study. The NIH expects the privacy of individual subjects to be protected in any data-sharing scheme. It is up to the investigator, their IRB, and the grantee institution, however, to develop mechanisms that do this. HIPAA regulations must be followed, and previously collected data can only be shared to the extent described in the consent form signed by the subjects at the time. It is expected that prospective research will be done with new consents that specifically mention publication and sharing of data, as well as the privacy protections that will be used to protect subjects.

Adding to the confusion of investigators is the fact that individual NIH institutes and programs may implement the Data Sharing Policy differently, with regard to types, amounts, and timing of data publication. Some programs have chosen to facilitate the policy directly. The Division of

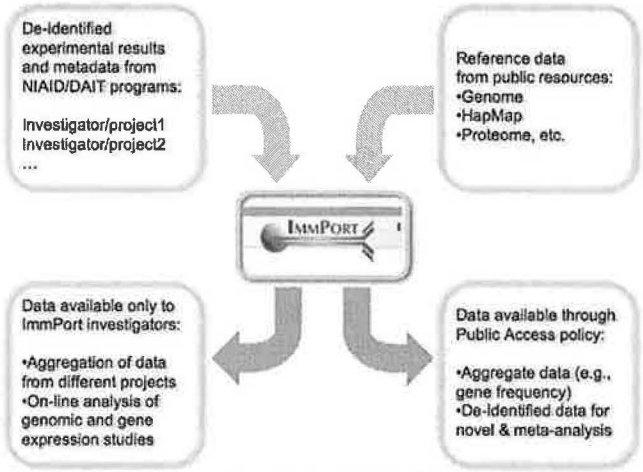

Figure 3. Schematic of the NIAID/DAIT ImmPort Database

Allergy, Immunology and Transplantation (DAIT) has contracted with Northrop Grumman Information Technologies and UT Southwestern Medical Center to create *ImmPort* the Immunology Database and Analysis Portal[90]. The goal of this web-based system to provide DAIT supported researchers a way to archive and share their research data. It includes the ability for investigators within consortia to collaborate by combining their datasets. A large reference database provides a compilation of public information on the human genome, genetic variation, and protein pathways. Data is initially deposited in ImmPort into private workspaces accessible only by the investigator and their staff. They can use web-based analytical tools to perform genetic association tests and gene expression analyses. When the major research findings from the project have been published, the investigator and DAIT program staff agree to move the data into a quasi-public workspace where other DAIT scientists can view and analyze suitably de-identified information. Lastly, aggregate data (e.g., gene frequencies) will be made available in a form that can be viewed by the interested public.

The NIH Data Sharing Policy has been specifically applied to genome-wide association studies (GWAS)[91]. All NIH supported research, both on the NIH campus and extramural, that includes GWAS now must include a plan to deposit information in a centralized data repository within the National Center for Biotechnology Information (NCBI). This repository is termed the database of genotypes and phenotypes (dbGaP)[92]. It is expected that data submissions to dbGaP will include not only the individual records of genotypes and whatever health information (phenotype) is collected in the course of the study, but also the study protocol, the questionnaires, study manuals, a list of variables measured, and any other supporting documents. All of this information is converted into a format that can be searched and analyzed online. The statistical analysis of the genetic data is available to anyone through a sophisticated web browser interface. Currently, data from 22 studies are available online in are in process. Many more are planned. These include shared data from the Framingham Study, studies of age-related eye disease, neurological, autoimmune, and psychological disorders.

There are several ethical concerns that have to be dealt with in the operation of dbGaP. First is the original deposition of data. This is to occur in a timely manner as the data are collected and curated. At that time, the Institutional Official (e.g., Dean for Research) will provide certification that all Federal and State laws and regulations have been followed, particularly the Common Rule and HIPAA, as well as any institutional policies, and in addition, they will certify the appropriate research uses of the dataset as well as any uses excluded by the consent form(s), and that the identities of the research participants will not be disclosed to dbGaP. In addition, the institution must certify that an IRB has reviewed and verified the fact that this particular data submission to dbGaP and the subsequent data sharing are consistent with the informed consent

Table 6. Projects Participating in dbGaP

Condition	Institute/Sponsor	Participants
Age-related eye disease	NEI	600
Parkinsonism	NINDS/NIA	2,573
ADHD	GAIN[1]	2,874
Diabetic Neuropathy	GAIN	1,835
Stroke	NINDS	1,555
ALS	NINDS	1,876
Depression	GAIN	3,720
Framingham Study	NHLBI	9,500
Psoriasis	GAIN	2,898
Type I Diabetes (DCCT)	NIDDK	1,441
Schizophrenia	GAIN	2,909
Bipolar Disorder	GAIN	2,400
Alzheimer's Disease	NIA	10,000
Women's Health Study	NHLBI	28,000

[1]GAIN: Genetic Association Information Network
Adapted from reference 92.

the subjects signed, that the data are de-identified according to standards set forth in Federal regulations, that the IRB has considered the risks to individuals, families, and groups associated with this GWAS repository submission, and the genotype and phenotype were collected in accordance with 45 CFR 46 (the Common Rule)[93].

This obviously puts the onus of verifying research subject protection squarely on the grantee institution, not on dbGaP. It meets the requirements of the OHRP 2004 guidance in that dbGaP is technically **not** doing human subjects research and cannot be directly regulated by the Common Rule. As stated above, the actual OHRP guidance may, itself, contradict the Federal regulations and thus be immaterial. Nevertheless, the policies of dbGaP with regard to the access to data stored in the repository suggest that they view use of the data to provide potential risks to the original subjects, their families, or ethnic groups.

Institutional review boards seldom look for more work to do, and it is unlikely that any IRB relishes the chance to review data that have been collected in a GWAS and compare them to consent forms. In many cases, the original DNA samples could have been collected years ago by investigators who are not part of the current study. Obtaining the original consent forms may be impossible. Except for consents executed with dbGaP in mind, it is unlikely that they even speak to the kind of data sharing contemplated here. IRBs will be put in a position of guessing the intent of subjects and making decisions as to whether data can be submitted to dbGaP knowing that an incorrect decision may engender bad feelings between the investigator, the institution, and the NIH program staff. Still, the substantial number of studies officially part of dbGaP to date shows that this process is far from impossible.

The second concern relates to the access to the data stored in dbGaP. As mentioned above, study documentation and pre-computed genotype-phenotype associations are available to the public via the Web. Record level data, i.e., de-identified genotypes and phenotypes on individual study subjects as well as pedigrees are available to qualified researchers who are approved by a Data Access Committee (DAC). The DAC could be specific to the NIH Institute sponsoring the study, or may be an NIH-wide DAC. Through a process similar to applying for an NIH research

grant, an investigator can ask for access to the data from the dbGaP studies. If the proposed database research is approved, a Data Use Certification (DUC) will be issued and the data can be downloaded from a secure site. The DUC requires the investigator and their institution to promise not to use the data for unapproved purposes, not to share or sell the data, and not to try to identify study participants. While there could be penalties to institutions whose researchers do not abide by the terms of the DUC, it is still largely up to the recipient investigator to honestly safeguard the data.

The NIH GWAS site does not minimize the risks associated with contribution of data to dbGaP[93]. It points out that technology available now or in the future could be used to identify subjects from a combination of genotype and phenotype. They explicitly point out the risk of stress, embarrassment, and emotional harm that could result from inadvertent release of information about families, stigmatizing conditions, and ethnic groups. They also verify that the data stored in dbGaP is subject to Freedom of Information Act requests, although those are likely to be denied. Law enforcement agencies could search the dbGaP databases looking for complete or partial matches to forensic DNA samples. Each of these potential risks is to be considered by IRBs as they certify data for submission to the repository.

Take Home Messages

1. Every physician needs to understand the power and limitation of modern genetic testing. Patients will ask them for advice; whether to get genetic tests for *bona fide* clinical purposes, or whether to participate in genetic research. At some point, price, speed, reliability and predictive value will make individual pharmacogenomic tests important considerations. The primary care physician may be the first person a patient turns to for an explanation of their personal genotype ordered on-line[94], or to explain the results of genome studies reported in the media[21, 95].

2. The collection of data associated with genetic research represents a risk to privacy and confidentiality. A combination of technology, policy, and regulation will be necessary to maintain the public trust and encourage participation in research.

3. Getting meaningful and legal informed consent for genetic research remains the subject of ongoing debate. The best way to obtain permission for studies done now and in the future is still not known. The consequences of abusing consent are clear, however.

4. The impact of funding agency decisions has created new problems. Researchers and IRBs struggle to keep up with new regulations while respecting the trust given to them by their subjects. Making data available on public web sites, even with privacy safeguards, raises questions of data ownership, scientific priority, and intellectual property.

References

1. Collins FS, Morgan M, Patrinos A. The Human Genome Project: lessons from large-scale biology. Science 2003;300:286-90.

2. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 2004;431:931-45.

3. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 2001;409:860-921.

4. Venter JC, Adams MD, Myers EW, et al. The Sequence of the Human Genome. Science 2001;291:1304-51.

5. President Clinton announces the completion of the first survey of the entire human genome. 2000. (Accessed April 16, 2008, at http://www.ornl.gov/sci/techresources/Human_Genome/project/clinton1.shtml.)

6. Kidd JM, Cooper GM, Donahue WF, et al. Mapping and sequencing of structural variation from eight human genomes. Nature 2008;453:56-64.

7. Pennisi E. Breakthrough of the year. Human genetic variation. Science 2007;318:1842-3.

8. Costenbader KH, Chang SC, De Vivo I, Plenge R, Karlson EW. PTPN22, PADI-4 and CTLA-4 genetic polymorphisms and risk of rheumatoid arthritis in two longitudinal cohort studies: evidence of gene-environment interactions with heavy cigarette smoking. Arthritis Res Ther 2008;10:R52.

9. Plenge RM, Cotsapas C, Davies L, et al. Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. Nat Genet 2007;39:1477-82.

10. Remmers EF, Plenge RM, Lee AT, et al. STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. N Engl J Med 2007;357:977-86.

11. Plenge RM, Seielstad M, Padyukov L, et al. TRAF1-C5 as a risk locus for rheumatoid arthritis--a genomewide study. N Engl J Med 2007;357:1199-209.

12. Harley JB, Alarcon-Riquelme ME, Criswell LA, et al. Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci. Nat Genet 2008;40:204-10.

13. Hom G, Graham RR, Modrek B, et al. Association of systemic lupus erythematosus with C8orf13-BLK and ITGAM-ITGAX. N Engl J Med 2008;358:900-9.

14. Kelly JA, Kelley JM, Kaufman KM, et al. Interferon regulatory factor-5 is genetically associated with systemic lupus erythematosus in African Americans. Genes Immun 2008;9:187-94.

15. Zeggini E, Scott LJ, Saxena R, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 2008;40:638-45.

16. Schwarz UI, Ritchie MD, Bradford Y, et al. Genetic determinants of response to warfarin during initial anticoagulation. N Engl J Med 2008;358:999-1008.

17. Corominas H, Domenech M, Laiz A, et al. Is thiopurine methyltransferase genetic polymorphism a major factor for withdrawal of azathioprine in rheumatoid arthritis patients? Rheumatology (Oxford) 2003;42:40-5.

18. Bhattacharyya T, Nicholls SJ, Topol EJ, et al. Relationship of paraoxonase 1 (PON1) gene polymorphisms and functional activity with systemic oxidative stress and cardiovascular risk. Jama 2008;299:1265-76.

19. van Meurs JB, Trikalinos TA, Ralston SH, et al. Large-scale analysis of association between LRP5 and LRP6 variants and osteoporosis. Jama 2008;299:1277-90.

20. Bezemer ID, Bare LA, Doggen CJ, et al. Gene variants associated with deep vein thrombosis. Jama 2008;299:1306-14.

21. Pearson TA, Manolio TA. How to interpret a genome-wide association study. Jama 2008;299:1335-44.

22. Scheuner MT, Sieverding P, Shekelle PG. Delivery of genomic medicine for common chronic adult diseases: a systematic review. Jama 2008;299:1320-34.

23. The International HapMap Consortium. A haplotype map of the human genome. Nature 2005;437:1299-320.

24. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007;449:851-61.

25. Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. J Clin Invest 2008;118:1590-605.

26. Graham RR, Kozyrev SV, Baechler EC, et al. A common haplotype of interferon regulatory factor 5 (IRF5) regulates splicing and expression and is associated with increased risk of systemic lupus erythematosus. Nat Genet 2006;38:550-5.

27. Demirci FY, Manzi S, Ramsey-Goldman R, et al. Association of a common interferon regulatory factor 5 (IRF5) variant with increased risk of systemic lupus erythematosus (SLE). Ann Hum Genet 2007;71:308-11.

28. Graham RR, Kyogoku C, Sigurdsson S, et al. Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus. Proc Natl Acad Sci U S A 2007;104:6758-63.

29. Kozyrev SV, Lewen S, Reddy PM, et al. Structural insertion/deletion variation in IRF5 is associated with a risk haplotype and defines the precise IRF5 isoforms expressed in systemic lupus erythematosus. Arthritis Rheum 2007;56:1234-41.

30. Shin HD, Sung YK, Choi CB, Lee SO, Lee HW, Bae SC. Replication of the genetic effects of IFN regulatory factor 5 (IRF5) on systemic lupus erythematosus in a Korean population. Arthritis Res Ther 2007;9:R32.

31. Kawasaki A, Kyogoku C, Ohashi J, et al. Association of IRF5 polymorphisms with systemic lupus erythematosus in a Japanese population: support for a crucial role of intron 1 polymorphisms. Arthritis Rheum 2008;58:826-34.

32. Siu HO, Yang W, Lau CS, et al. Association of a haplotype of IRF5 gene with systemic lupus erythematosus in Chinese. J Rheumatol 2008;35:360-2.

33. Ioannidis JP. Non-replication and inconsistency in the genome-wide association setting. Hum Hered 2007;64:203-13.

34. Ott J. Association of genetic loci: Replication or not, that is the question. Neurology 2004;63:955-8.

35. Salanti G, Sanderson S, Higgins JP. Obstacles and opportunities in meta-analysis of genetic association studies. Genet Med 2005;7:13-20.

36. Shen H, Liu Y, Liu P, Recker RR, Deng HW. Nonreplication in genetic studies of complex diseases--lessons learned from studies of osteoporosis and tentative remedies. J Bone Miner Res 2005;20:365-76.

37. Sullivan PF. Spurious genetic associations. Biol Psychiatry 2007;61:1121-6.

38. Tian C, Plenge RM, Ransom M, et al. Analysis and application of European genetic substructure using 300 K SNP information. PLoS Genet 2008;4:e4.

39. Archon X Prize for genomics. 2008. (Accessed June 2, 2008, at http://genomics.xprize.org/.)

40. Annas GJ. Privacy rules for DNA databanks. Protecting coded 'future diaries'. Jama 1993;270:2346-50.

41. Mouchawar J, Laurion S, Ritzwoller DP, Ellis J, Kulchak-Rahm A, Hensley-Alford S. Assessing controversial direct-to-consumer advertising for hereditary breast cancer testing: reactions from women and their physicians in a managed care organization. Am J Manag Care 2005;11:601-8.

42. Gollust SE, Hull SC, Wilfond BS. Limitations of direct-to-consumer advertising for clinical genetic testing. Jama 2002;288:1762-7.

43. Couzin J. Once shunned, test for Alzheimer's risk headed to market. Science 2008;319:1022-3.

44. Katsanis SH, Javitt G, Hudson K. A case study of personalized medicine. Science 2008;320:53-4.

45. States crack down on online gene tests. Forbes, 2008. (Accessed May 9, 2008, at http://www.forbes.com/2008/04/17/genes-regulation-testing-biz-cx_mh_bl_0418genes_print.html.)

46. Faces of genetic discrimination: How genetic discrimination affects real people. 2004. (Accessed January 4, 2008, at http://www.geneticalliance.org/ksc_assets/documents/facesofgeneticdiscrimination.pdf.)

47. Billings PR, Kohn MA, de Cuevas M, Beckwith J, Alper JS, Natowicz MR. Discrimination as a consequence of genetic testing. Am J Hum Genet 1992;50:476-82.

48. Miller PS. Genetic discrimination in the workplace. J Law Med Ethics 1998;26:189-97, 78.

49. Schulte PA, Lomax G. Assessment of the scientific basis for genetic testing of railroad workers with carpal tunnel syndrome. J Occup Environ Med 2003;45:592-600.

50. Chance PF. Overview of hereditary neuropathy with liability to pressure palsies. Ann N Y Acad Sci 1999;883:14-21.

51. Stockton DW, Meade RA, Netscher DT, et al. Hereditary neuropathy with liability to pressure palsies is not a major cause of idiopathic carpal tunnel syndrome. Arch Neurol 2001;58:1635-7.

52. Bluman LG, Rimer BK, Berry DA, et al. Attitudes, knowledge, and risk perceptions of women with breast and/or ovarian cancer considering testing for BRCA1 and BRCA2. J Clin Oncol 1999;17:1040-6.

53. Lerman C, Audrain J, Orleans CT, et al. Investigation of mechanisms linking depressed mood to nicotine dependence. Addict Behav 1996;21:9-19.

54. Lerman C, Narod S, Schulman K, et al. BRCA1 testing in families with hereditary breast-ovarian cancer. A prospective study of patient decision making and outcomes. Jama 1996;275:1885-92.

55. Apse KA, Biesecker BB, Giardiello FM, Fuller BP, Bernhardt BA. Perceptions of genetic discrimination among at-risk relatives of colorectal cancer patients. Genet Med 2004;6:510-6.

56. Matloff ET, Shappell H, Brierley K, Bernhardt BA, McKinnon W, Peshkin BN. What would you do? Specialists' perspectives on cancer genetic testing, prophylactic surgery, and insurance discrimination. J Clin Oncol 2000;18:2484-92.

57. U.S. public opinion on uses of genetic information and genetic discrimination. 2007. (Accessed May 31, 2008, at http://www.dnapolicy.org/resources/GINAPublic_Opinion_Genetic_Information_Discrimination.pdf.)

58. The genetic information non-discrimination act of 2007. In: PL 110-233; 2008.

59. Eiseman E, Haga SB. Handbook of human tissue sources: A national resource of human tissue samples. Santa Monica, CA: RAND; 1999.

60. Roden D, Pulley J, Basford M, et al. Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. Clin Pharmacol Ther 2008.

61. Maschke KJ. Navigating an ethical patchwork--human gene banks. Nat biotechnol 2005;23:539-45.

62. Arnason V. Coding and consent: moral challenges of the database project in Iceland. Bioethics 2004;18:27-49.

63. Greely HT. Iceland's plan for genomics research: facts and implications. Jurimetrics 2000;40:153-91.

64. Collins FS. The case for a US prospective cohort study of genes and environment. Nature 2004;429:475-7.

65. Policy issues associated with undertaking a new large U.S. population cohort study of genes, environment, and disease. 2007. (Accessed May 21, 2008, at http://www4.od.nih.gov/oba/sacghs/reports/SACGHS_LPS_report.pdf.)

66. Knoppers BM, Fortier I, Legault D, Burton P. Population Genomics: The Public Population Project in Genomics (P(3)G): a proof of concept? Eur J Hum Genet 2008;16:664-5.

67. Lin Z, Owen AB, Altman RB. Genomic research and human subject privacy. Science 2004;305:183.

68. Lowrance WW, Collins FS. Identifiability in genomic research. Science 2007;317:600-2.

69. Knoppers BM, Saginur M. The Babel of genetic data terminology. Nat biotechnol 2005;23:925-7.

70. Guidance on research involving coded private information or biological speciments. 2004. (Accessed June 2, 2008, at http://www.hhs.gov/ohrp/humansubjects/guidance/cdebiol.htm.)

71. Kallberg H, Padyukov L, Plenge RM, et al. Gene-gene and gene-environment interactions involving HLA-DRB1, PTPN22, and smoking in two subsets of rheumatoid arthritis. Am J Hum Genet 2007;80:867-75.

72. Padyukov L, Silva C, Stolt P, Alfredsson L, Klareskog L. A gene-environment interaction between smoking and shared epitope genes in HLA-DR provides a high risk of seropositive

rheumatoid arthritis. Arthritis Rheum 2004;50:3085-92.

73. Sweeney L. Weaving technology and policy together to maintain confidentiality. J Law Med Ethics 1997;25:98-110, 82.

74. Malin B. Re-identification of familial database records. AMIA Annual Symposium proceedings / AMIA Symposium 2006:524-8.

75. Malin B, Sweeney L. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. Journal of biomedical informatics 2004;37:179-92.

76. Malin B. A computational model to protect patient data from location-based re-identification. Artificial intelligence in medicine 2007;40:223-39.

77. Caulfield T. Biobanks and blanket consent: The proper place of the public good and public perception rationales. King's Law Journal 2007;18:209-26.

78. Greely HT. The uneasy ethical and legal underpinnings of large-scale genomic biobanks. Annu Rev Genomics Hum Genet 2007;8:343-64.

79. Lavori PW, Krause-Steinrauf H, Brophy M, et al. Principles, organization, and operation of a DNA bank for clinical trials: a Department of Veterans Affairs cooperative study. Control Clin Trials 2002;23:222-39.

80. Caulfield T, Upshur RE, Daar A. DNA databanks and consent: a suggested policy option involving an authorization model. BMC Med Ethics 2003;4:E1.

81. Hansson MG, Dillner J, Bartram CR, Carlson JA, Helgesson G. Should donors be allowed to give broad consent to future biobank research? Lancet Oncol 2006;7:266-9.

82. Tutton R, Kaye J, Hoeyer K. Governing UK Biobank: the importance of ensuring public trust. Trends Biotechnol 2004;22:284-5.

83. Dalton R. When two tribes go to war. Nature 2004;430:500-2.

84. Indian Givers: The Havasupai trusted the white man to help with a diabetes epidemic. Instead, ASU tricked them into bleeding for academia. Phoenix New Times, 2004. (Accessed March 3, 2008, at http://www.phoenixnewtimes.com/2004-05-27/news/indian-givers.)

85. Tribal suit over blood samples dismissed. USA Today, 2007. (Accessed May 21, 2008, at http://www.usatoday.com/test/test4/news/2007-05-04-test_4_bobxx_N.htm.)

86. Havasupai v. Arizona Board of Regents. 2008. (Accessed May 21, 2008, at http://www.cofad1.state.az.us/casefiles/cv/cv070454.pdf.)

87. DNA gatherers hit a snag: Tribes don't trust them. The New York Times, 2006. (Accessed February 28, 2008, at http://www.nytimes.com/2006/12/10/us/10_dna.html.)

88. NIH data sharing policy. 2007. (Accessed May 26, 2007, at http://grants.nih.gov/grants/policy/data_sharing/.)

89. Frequently asked questions on data sharing. 2004. (Accessed May 26, 2007, at http://grants.nih.gov/grants/policy/data_sharing/data_sharing_faqs.htm.)

90. Immunology database and analysis portal. 2008. (Accessed May 26, 2007, at https://www.immport.org/immportWeb/home/home.do.)

91. Genome wide associations studies (GWAS) web site. 2007. (Accessed May 26, 2008, at http://grants.nih.gov/grants/gwas/.)

92. Mailman MD, Feolo M, Jin Y, et al. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 2007;39:1181-6.

93. NIH points to consider for IRBs and institutions in their review of data submission plans for institutional certifications under NIH's policy for sharing of data obtained in HIH supported or conducted genome-wide association studies (GWAS). 2007. (Accessed May 26, 2008, at http://grants.nih.gov/grants/gwas/gwas_ptc.pdf.)

94. Hunter DJ, Khoury MJ, Drazen JM. Letting the genome out of the bottle--will we get our wish? N Engl J Med 2008;358:105-7.

95. Feero WG. Genetics of common disease: a primary care priority aligned with a teachable moment? Genet Med 2008;10:81-2.