Science and Engineering and Their Different Roles in Investigations of the Genetic Portion of the Etiology of Complex Human Traits

Joseph D. Terwilliger*

Department of Psychiatry and Columbia Genome Center, New York State Psychiatric Institute, Division of Molecular Genetics, Columbia University, 1150 St Nicholas Avenue, Room 504, New York, NY 10032, USA

Finnish Genome Center, Biomedicum Helsinki, University of Helsinki, Haartmaninkatu 8, FIN, 00251 Helsinki, Finland

Introduction

Owing largely to recent advances in technology, both in molecular biotechnology and computer science, the field of human genetics has been undergoing a series of rapid metamorphoses and "wide-eyed" optimism, some of which may eventually contribute to our understanding of the ephemeral relationships between genetic variation and phenotypic variation in man, while others may set us back decades by encouraging us to waste resources and energy pursuing blind technological pathways-all too often without focus on the underlying the scientific questions-to solving problems for which the technology is neither the problem nor the solution. The recent emphasis on population level biobanks and large cohort and case control studies for gene mapping are such an example of taking industrial and technology-driven approaches to fundamental biological questions, which are unlikely to yield much progress in scientific understanding (Le Fanu, 1999; Holtzman, 2001; Millikan, 2002; Weiss and Terwilliger, 2002; Terwilliger and Weiss, 2003). With the rapid influx of scientists with extremely heterogeneous background and expertise, or perhaps more worrisome, their extreme degree of narrow training specialization, there are also dangers of wasting substantial effort to document what has already been known in principle about complex traits for nearly a century, and in practice by hundreds of studies in the past 20 years. Are we reinventing the wheel, and more critically, charging down known "blind alleys" at a substantial cost in time and money, especially by being in too much of a hurry? It may be especially important for the likely readership of a journal entitled "Genomics and Informatics" to consider the background and nature of the underlying questions molecular biologists and computational scientists are looking to solve in humans. Methods are ways of making progress in science; both are important, but we should not make the mistake of equating the former with the latter (Terwilliger and Goring, 2000; Terwilliger *et al.*, 2002a, 2002b; Terwilliger and Weiss, 2003; Weiss, 1993).

Experimental Science vs Observational Science

When one desires to study the etiology of some trait in a mouse, experimental manipulations, through altering environmental exposures, mutagenesis, and breeding experiments are used to ask the question "if a mouse were exposed to A, what would happen to the mouse?" For example, so called "knockout mice"-in which a given gene is destroyed so that its product is not produced by the mouse at all-can be engineered to see what would happen to a given strain of mouse if a given protein were not produced at all. In the case of an environmental exposure, some mice from a given clonal population might be exposed to some industrial chemicals, while others are not, to examine the phenotypic consequences later in their lives. In the case of gene mapping in the mouse, one selects inbred strains on the basis of their phenotype, performs a backcross (or other controlled mating experiment) in which the amount of total genotypic variation is limited to a minimal amount (one chromosome per strain, since inbred strains are homozygous throughout their genomes), to try and correlate the observed phenotypic variation in the offspring to which strain certain chromosomal segments were derived from. This can be a powerful technique for identifying which genes harbor variation in given inbred strains that CAN have an influence on some phenotype when both background genetic variation and environmental exposures are rigidly controlled. Such breeding

* Corresponding author:
E-mail jdt3@columbia.edu,
Tel +1-917-287-7572, Fax +1-212-851-5176
E-mail jtej@ktl.fi,
Tel +358-40-7467-277, Fax +358-9-4744-8480

Accepted 18 March 2004

experiments often require hundreds of offspring from a given backcross for mapping to be successful, however, in a case where only two chromosomes segregate in the population, where linkage and linkage disequilibrium are the same thing, and where there is no environmental or cultural variation among organisms! Environmental studies typically subject mice to grossly increased levels of exposure to putative risk factors relative to the problem of avoiding harmful human exposures, in order to generate a strong enough signal to detect an effect clearly. In a similar way, experiments on inbred laboratory mice greatly simplify the causal environment. Even so, such breeding experiments by themselves are often insufficiently powerful for successful gene identification, to which end further breeding experiments are often required to generate so-called congenic strains in which mice are crossed back to one of the original strains until only a given small genetic region carries genetic variation from the other strain, thus essentially creating a new strain which is genetically identical to one of the parents except for a single gene region, selected on the basis of phenotypic effect of that gene, combined with genetic information that the remainder of the genome is from the other strain. Such techniques may be necessary to isolate which of the many genetic factors in a given region has a functional consequence, and furthermore, may be needed to examine the effects of a given variant form of a given gene in a variety of different genetic backgrounds. Such experiments demonstrate consistently that the same variant can have dramatically different phenotypic consequences in different strains, to the point where even many knockouts have no phenotypic consequences in some strains, while they can be devastating in others.

When trying to understand the phenotypic consequences of variation in genetic and environmental exposures in human, it is naturally impossible to perform experiments as would be scientifically desirable. With the exception of controlled clinical trials, it is very difficult to alter the kinds of natural environmental exposures we encounter in life, in a satisfactorily directed way; of course, we cannot do mutagenesis experiments or designed breeding(especially not inbreeding) to assist in our mapping efforts. The science of epidemiology has focused for generations on how to identify environmental risk factors from observational data, where-for example-cases of a given disease and controls might be ascertained from a population and subjected to questioning about their earlier environmental exposures, a practice which has led to some important findings, but mostly of the "sledgehammer" variety, for instance demonstrating that most lung cancer victims smoked at some point in their lives, while only a small portion of the healthy population did. Indeed, even factors that have been thought to be understood regularly turn out not to be so clear after all, as one clearly sees in the daily epidemiological study reports in major media outlets.

In the absence of controlled experiments, human geneticists are forced to become detectives, and look for natural experiments, in which the sorts of matings we would like to observe have been done by people of their own free will (and without regard to our biomedical interests), or where environmental exposures have been intrinsically altered in such a way as to make for a useful resource for epidemiological investigation. Thus rather than being a formally experimental science, as would be the ideal for hypothesis testing and unraveling of the complexities of life and illness, human genetics must rely on statistical inference from observed data.

These are the primary contrasts between the more pure scientific culture of research in lower organisms and the more anthropological detective work we are relegated to in human studies. This necessity to play detective has motivated whole industries of sophisticated technological advances, spawning the young sciences of genomics and bioinformatics to help us perform our detective work, or to make the most out of the undersized, underpowered, and suboptimally designed natural experiments with which we are stuck. Thus even in possession of these fine new tools, geneticists must still search the world for optimal data to apply them to. The real challenge in the genomic era is to remember that fundamentally our questions are scientific, largely requiring selection of appropriate study designs, and not matters of technology per se.

Genomics vs Genetics

The *American Heritage Dictionary* (2000) defines genetics as "The branch of biology that deals with heredity, especially the mechanisms of hereditary transmission and the variation of inherited characteristics among similar or related organisms," and genomics as "The study of all of the nucleotide sequences, including structural genes, regulatory sequences, and noncoding DNA segments, in the chromosomes of an organism." While the above definitions are not universal, they at least provide a starting point for considering why people who self-identify with each of the above disciplines approach genetic questions from a different perspective.

Originally, genetics was primarily a theoretical and statistical science, based on Mendel's observations that heredity characters are correlated in families because discrete units of hereditary information (later termed "genes") are inherited from one's parents according to very simple principles which, in today's terminology are that each individual has two copies of each gene which may vary, one received from each of that individual's parents, and a random choice of which the individual will transmit to each of his offspring. Because of parent-offspring transmission, units of information are shared among relatives, so that genetic variation that impacts on variation in phenotype, the phenotype will be correlated among relatives, in ways that can be specified by the probabilistic laws of inheritance. R.A. Fisher, Sewall Wright and other scientists developed the field of genetics as a quantitative statistical science, and described the consequences of Mendel's laws, and deviations therefrom (such as linkage and linkage disequilibrium), and their potential effects on correlations in phenotypic variation within and between species (Fisher, 1939, 1949, and 1960; Wright, 1968). Much of this early literature goes untouched by modern geneticists, many of whom are either wholly unaware of this scientific heritage, or who do not realize its relevance. Not the least of the reason for this is the great difficulty in understanding this highly technical material. An additional problem is that often without thinking deeply about it, molecular biologists assign "genetic" terminology to the "molecular processes" they discovered to explain the genetic phenomena. For example, traits are equated with genes underlying them, even when the latter are strictly matter of hypotheses based on a poor understanding of theoretical biology and genetics. But Mendel carefully chose his traits, whereas variation in most traits of interest in biomedicine is not highly correlated with a specific variant in a single gene. The confusion probably is the result of the rapid growth of molecular biology in the second half of the 20th century out of classical genetics that was largely restricted to 'Mendelian' traits (Elston, 1968; Terwilliger and Goring, 2000).

Genetics proper is the study of genes and their inheritance, and perhaps we need a separate term-phenogenetics?-to refer to the study that currently is called 'genetics', that is, the study of how genetic *variation* influences phenotypic *variation*. Ultimately this is the real question that many scientists using genomics and informatics are hoping to address. In fact, the older sense of the word 'gene' to refer to protein-coding sequence, is clearly being blurred by other kinds of 'gene'(the terminology is not yet standard) that are used to transcribe RNA, to regulate coding-gene expression, as signals for chromosome packaging, and the like-all of which are important and of great interest in informatics and genomics.

Thus, "genomics" is focused on identifying the tasks performed by different DNA sequences, variation that affect how well a given task is performed (which is really what genetics is all about). For most people in the genomics field, the focus is on technologies and algorithms for identifying sequence variation and technologies to measure it. These differences in perspective between "genomic approaches" and "genetic approaches" show the substantial areas where multidisciplinary efforts of geneticists and genomics experts will be essential. However, this will somehow require getting past the training and specialization narrowness of most people in the field, and some willingness not to race to hastily to the end problem of identifying the causes of complex diseases, when the important middle conceptual and study-design steps are given short shrift (with predictably problematic results).

Statistical Genetics vs Bioinformatics

The same *American Heritage Dictionary* (2002) defines informatics as "The science that is concerned with the gathering, manipulation, classification, storage, and retrieval of recorded knowledge," implying that bioinformatics by extension would refer to the science of gathering, manipulation, classification, storage, and retrieval of recorded biological information. Statistical genetics refers to the study of probabilistic quantitative relationships among inherited sequence variations, and the quantitative statistical relationships between such variation and phenotypic variation. The major distinction between these fields, much as that between genomics and genetics is between description of structure and its evolution, versus the study of the functional effects of variation.

Bioinformatics is largely focused on sifting through masses of haphazardly collected data in an attempt to build models of what the information in DNA sequences describes. In such approaches, genes, proteins, and other biological data are treated as discrete units of information, for example, in describing networks of genes that might work together to perform some task, much like the Krebs cycle and other biochemical pathways we all had to laboriously memorize in graduate school (Krebs and Kornberg, 1957). Just because we can write down the molecules participating in the Krebs cycle, however, does not mean that variation in the efficiency of the Krebs cycle must necessarily be related to variation in the DNA sequence of the genes that encode the various proteins used in the Krebs cycle. The latter is a question of statistical genetics-designing experiments to test whether variation in those genes might relate to variation in phenotype, whereas the former would fall under the rubric of bioinformatics, assuming such information were assembled from existing literature and data of various types. Essentially, the emphasis in genetics and statistical genetics is on elucidating and quantifying the stochastic relationships between genotype and phenotype, while genomics and bioinformatics are more focused on elucidating more deterministic relationships or descriptions of phenomena involving DNA sequences, or generating hypotheses for geneticists or other biologists to consider testing, if such hypotheses make biological sense.

Data Mining vs Statistical Experimentation

Contemporary genomics is focused on generating as much sequence data as possible from as many people as possible, without much regard for what specific scientific questions are to be asked. If biotechnology makes it possible to sequence or genotype thousands of individuals at thousands of sites in the genome, then genomics researchers suggest that it should be done, and immediately, as the best way to search for sites in the DNA sequence with functional consequences. Informaticians try to find ways to mine these enormous sets of existing data to look for patterns of relationships that distinguish sets of individuals, perhaps on the basis of phenotype. But typically the emphasis of these technology and engineering-related fields is on obtaining as much data as possible and identifying patterns using as much data as possible, without much regard to how the *data are ascertained*, or *what sort of patterns might* be predicted on the basis of what is known about quantitative genetics. This is sometimes referred to as "hypothesis-free" or "hypothesis generating" research, in contrast to most (of the best) genetics research that is designed to test specific hypotheses about genetic causation.

Hypotheses can be developed because we have formal theory of inheritance, and the evolutionary processes responsible for genetic variation and its distribution, originally developed by Fisher, Wright, and others for the likely relationships between phenotypic and genotypic variation, especially conditional on different models of etiology and ascertainment. We have less specific theory for the structure of genomes, so that much of genomics and informatics is currently much more empirical and ad hoc.

Genetic theory cannot tell us a priori which (if any) genetic variation is associated with any specific phenotypic variation, but specifically testable hypotheses can be framed to search for that variation because the theory of inheritance is generic; it specifies what sorts of correlation are likely to be identified for a genetically mediated trait, based on analysis of the trait correlations in families, and the sorts of populations from which the data are ascertained. Randomly searching for patterns in data without regard for predicted correlations due to genetic phenomena, and the constraints imposed thereby is easy to do because of the powerful genomics and informatics methods now available. However, in the end it proves to be inefficient at best, and there is ample history to show that this approach is plagued by high rates of false discovery, and failure to identify real structure in the datasets. For this reason it is important for the genomics and bioinformatics researchers to gain a thorough knowledge of quantitative genetics and its methods, in order to know where and how their kinds of data and approaches will be most relevant to the practical genetic questions.

Engineering vs Basic Science

The contrast between genetics and genomics/ bioinformatics are essentially the same as that separating engineering and basic science. Engineers are extremely talented and intelligent people interested in finding ideal ways to accomplish certain tasks, either by development of new technological solutions, or by applying existing technology to accomplish some task or solve some problem. Basic science, on the other hand is aimed at answering a question, whereas too often in human genetics what is done is to assume the answer on the basis of poorly understood theory and use highly technical methods to find the specifics of the desired answer. This is very common in human genetics, but is an illegitimate mix of basic science and engineering worldviews.

There is a way to think of the problem more constructively, and that is for geneticists to ignore present technical limits, assuming the biotechnology and informatics communities will solve them, and to be more critical in considering the questions they expect those tools to assist in answering. I have written many times (eg, Terwilliger and Goring, 2000) that geneticists should assume they have the full sequence data from any individual they would like, and the technical capability to perform any quantitative analysis they can formally describe. I firmly believe that technological solutions will be forthcoming in both areas, and that leaves only the scientific question of where to apply them and how, which needs much greater attention, and is fundamentally much more conceptually difficult. Despite tremendous advances in biotechnology and informatics, we are not close to solving the problem of how to correlate genetic variation with phenotypic variation in other than simple Mendelian situations, and these clearly do not present the most pressing health problems of our day.

Interdisciplinary Collaboration is Essential

Ultimately, to unravel the etiology of human complex disease is an extremely difficult yet important task requiring the input of people with enormously heterogeneous background and expertise. Clearly epidemiologists are expert in identifying how variation in environmental exposures relate to phenotypic variation, genetic epidemiologists and statistical geneticists are expert in identifying the relationships between heritable exposures, and phenotypic variation. Population geneticists and evolutionary biologists are experts in identifying and quantifying the relationships among genes on a population level, and the relationships between phenotypes and genotypes on an evolutionary level. Anthropologists, historians, genealogists and demographers are essential to analyzing family structures, identifying connections among affected individuals in historical time, and locating populations with interesting population histories that may be more useful in genetic studies. Social scientists, nutritionists, and environmental health researchers are essential to quantifying and measuring the environmental risk factors hypothesized to be related to the traits of interest. Bioinformaticians, medical experts, physiologists, biochemists and the like are essential to helping develop hypotheses worth testing, both from clinical and biological perspective, as well as from sifting through extant data for latent structure which may potentially be of biological relevance. Numerous clinical and laboratory experts are essential to making sure the phenotypes are measured in the most accurate way possible, to make weak correlations with genotypes as identifiable as possible. Genomics experts and biotechnologists are obviously critical to generating the sequence and genotype data one would hope to correlate with phenotypic variation. And in the end, all of these roles need to be treated with equivalent importance in designing and carrying out a study, since no expert in any one of these subareas can possibly be sufficiently expert in the others to design a study without collaboration and cooperation. This is what makes genetic research particularly appealing, but also makes it exceptionally complex to start and to organize. Mere collection of extant data and genotyping them will not lead to a powerful study, but rather one needs to combine expertise from different areas to design an appropriate experiment to apply the genomics and informatics technology to, and this can take many years of planning and fieldwork before the actual genotype-phenotype correlations can even be explored.

The arguments presented here may seem strange given the common rhetoric from the leaders of the human genome project. But a critical look at what has actually been achieved shows that it is unlikely that major progress in public health or even in understanding phenogenetic relationships will suddenly come upon us out of scaled-up, hypothesis-free technology without appropriately focused genetic questions and applying great technology to suboptimal scientific experiments. More thorough integrative collaboration goes against many of the trends toward academic micro-specialization, but is needed and will provide both the challenge as well as the excitement of this area of research in the future, if we take our recent experiences with current approaches to heart.

It would be wrong to minimize the communication problem in any multidisciplinary field of endeavor. Such problems are made even more difficult when the literature in genetics is mostly available only in English, which is a second language at best to most researchers outside the US and UK, making reliance on review papers and media accounts a more important source of information for such scientists than it might otherwise be. As someone who tries to read articles and books from the anthropological literature written in Korean, I can attest to how complicated in can be to read material in a different language, especially when it is not in your own major field of expertise. And problems also come about from reading material that has been translated and filtered through someone else's perceptions of meaning into your own language as well. As an example of the latter, I have been misquoted in the Korean media (Kyunghyung Sinmun, 2003; Joongang Ilbo, 2004) as having claimed that the Korean population was a "genetically pure" or "homogeneous" population (whatever that actually means...) In fact, I never said or thought anything remotely like this. Of course there are about 80 million Koreans in the world, which is an absolutely enormous population size, and it is clear that there is necessarily an enormous amount of genetic variation in such a large, ancient, and historically substructured population.

There are many ways one can be misled by relying on secondary sources of information and numerous reasons they can misrepresent the true situation. Communication and collaboration barriers show down work, and greatly reduce the number and variety of scientists whose creative imagination, experience, and insight could be brought to bear worldwide on these problems. Of course, there are economic and political reasons why things are the way they are, but there is no legitimate scientific rationale for rushing as we have been doing. There will be a lot of work to do in this area for decades to come, and a sincere desire actually to solve the problems should lead to a comparable willingness to be inclusive in the ways I have suggested.

References

American Heritage Dictionary of the English Language (2000). Houghton-Mifflin Publishers.

Elston, R. C. (1998). Linkage and association. *Genet. Epidemiol.* 15, 565-576.

Fisher, R. A. (1939). *The genetical theory of natural selection.* Oxford: The Clarendon Press.

Fisher, R. A. (1949). *The theory of inbreeding.* Edinburgh: Oliver and Boyd.

Fisher, R. A. (1960). *The design of experiments.* New York: Hafner Pub Co.

Holtzman, N. A. (2001). Putting the search for genes in perspective. *Int. J. Health Serv.* 31, 445-461.

Joongang Ilbo (2004). Dissecting Korean Genes: "There were ancestors from the South". Jan 28.

Krebs, H. A. and Kornberg, H. L. (1957). *Energy Transformations in Living Matter.* Berlin: Springer-Verlag.

Kyunghyang Sinmun (2003). "The Korean people have struggled to maintain their genetic purity". Jan 24.

Le Fanu, J. (1999). *The Rise and Fall of Modern Medicine.* New York: Carroll & Grf Publishers.

Millikan, R. (2002). The Changing Face of Epidemiology in the Genomics Era. *Epidemiology* 13, 472-480.

Terwilliger, J. D. and Goring, H. H. H. (2000). Gene mapping in the 20th and 21st centuries: Statistical methods, data analysis, and experimental design. *Hum. Biol.* 72, 63-132.

Terwilliger, J. D., Haghighi, F., Hiekkalinna, T. S., and Goring, H. H. (2002a). A biased assessment of the use of SNPs in human complex traits. *Curr. Opin. Genet. Dev.* 12, 726-734.

Terwilliger, J. D., G ring, H. H. H., Magnusson, P. K. E., and Lee, J. H. (2002b). Study design for genetic epidemiology and gene mapping: The Korean diaspora project. *Shengming Kexue Yanjiu (Life Science Research)* 6, 95-115.

Terwilliger, J. D. and Weiss, K. M. (2003). Confounding, ascertainment bias, and the blind quest for a genetic 'fountain of youth'. *Ann. Med.* 35, 532-544.

Weiss, K. M. (1993). *Genetic Variation and Human Disease: Principles and evolutionary approaches.* Cambridge: Cambridge University Press.

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

Wright, S. (1968). *Evolution and the Genetics of Populations,* Vol.1-4. Chicago: University of Chicago Press.