

Introduction

The relentless pursuit of scientific knowledge in and of itself is driven by a unique human trait, the desire to analytically understand the world. In order to do so properly, to overcome the uncertainties and randomness of natural processes and to restrain the vagaries of the intuitive human mind, scientific studies must be carefully designed and statistical analyses rigorously applied.

In this Introduction I will first go over some background and context—conceptual and historical—to appreciate statistics in science, particularly biostatistics in biomedical research.

In the rest of the thesis, specific detailed backgrounds for each chapter will be provided in place. We first note that these projects represent different aspects of computational and statistical investigations of the genetic bases of human diseases. Currently, diseases with genetic components are often studied using methods of statistical genetics and expression analysis. In this work, under the rubric of statistical genetics, we have constructed and studied an extension of the classic family-based association test into exact tests of pedigrees (*Section Error! Reference source not found.. Error! Reference source not found.*) and applied it to study the genetic bases of Alzheimer's disease. In expression analysis, we describe a project of substantial scope that studied gene expression alterations in psoriasis (*Section Error! Reference source not found.. Error! Reference source not found.*). Based on an appreciation for the complexity of biological circuitry, we proposed a novel gene-to-gene relationship beyond co-expression that may be exploited by gene expression studies and used it to predict gene

function reliably (*Section Error! Reference source not found.. Error! Reference source not found.*). To explain expression alterations mechanistically, we have designed a method to computationally locate gene promoters (*Section Error! Reference source not found.. Error! Reference source not found.*) and a statistical method to predict the transcription factor binding sites they may contain (*Section Error! Reference source not found.. Error! Reference source not found.*).

The Unique Mandate of Statistics in Biomedical Research

Certainly few would disagree with the imperative to properly employ statistical techniques in modern scientific investigations. Generally, statistics is called upon when entities under study become too many to individually consider or when the phenomena of interest operate with apparent randomness.

The early developments of probabilistic techniques in science took place in astrophysics to deal with measurement error. But in other scientific disciplines, the adoption of statistical techniques was often necessitated by a change in perspective, a shift from the consideration of single, isolated events to one of populations of similar elements. Probabilistic techniques, with its ability to succinctly describe and operate on the properties of populations as wholes, are naturally adopted as a tool in this new outlook. This general transition was first evidenced in the latter half of the 19th century in the kinetic gas theory of Ludwig Boltzmann and James C. Maxwell and the social physics of Auguste Comte—who ironically foresaw the onslaught of statistics and detested it (Hacking 1990)—and of Lambert Quetelet in the 19th century (Quetelet 1835).

However, in this type of usage, randomness merely reflects an ignorance regarding the distinctions between members of a population, and not an appreciation of some essential aspect of the processes under question. While 20th-century physics found fundamental non-determinism in physical processes, without invoking the counter-intuitive concepts of quantum mechanics, numerous natural phenomena are already known to be practically random, in particular biological processes. An example in the field of genetics is Mendel's principle of random assortment. At larger scales, by virtue of nonlinearity compounded by quantity, even mechanistic interactions between simple constituents of a biological system are sufficient to endow it with monumentally unpredictable behavior. Our methods of observation and experimental intervention thusly meager in comparison, statistics will be needed to manage the unfathomable complexity.

Here, I will compare the development of statistical thinking alternately as a prescriptive and a descriptive model of science, and outline some of the challenges unique to biostatistics:

1. Modern biomedical investigations remain for the most part molecular and analytic. In fulfilling its role as an integral part of this research enterprise, biostatistics should strive for the interpretability of its results.
2. Statistics, as a fundamental component of the scientific method and as an effective *de facto* rhetorical device, is essential to the evolution of scientific ideas, in particular in the emotional debates of biomedical research.

3. The incorporation of scientific methods and statistical analyses into medical research met formidable psychological and sociological barriers, which, in one form or another, remain a challenge today.

INTERPRETABILITY IS ESSENTIAL FOR BIOSTATISTICS

As statistics in general has become indispensable to all fields of science, so has biostatistics proven invaluable to the individual biomedical researcher. Thus biostatistics takes on the unique requirement of interpretability because of the inexorably analytic, top-down process that characterizes modern biomedical investigations. That is, because of our ability to manipulate biological systems at increasingly finer levels, observed regularities are rarely taken as they are and instead are often dissected further to their molecular bases. Hence, compared to the social sciences, where the construction of statistical models and the elucidation of numerical regularities may be an end in itself, in biomedical investigations, statistical analyses are expected to generate new leads for further research. The natural implication of this is the requirement that biostatistical analyses must render results that are *interpretable* in terms of intuitive biological concepts, such that they may be usefully incorporated into further investigations.

STATISTICS IN THE SCIENTIFIC METHOD AND IN THE EVOLUTION OF SCIENTIFIC IDEAS

While the indispensable roles that probabilistic and statistical arguments play in day-to-day learning is apparent, their proper places in models of idealized learning, i.e. models

of *scientific inference*, had been contentious. In this section, we first briefly discuss the incorporation of statistics into the scientific method. Next, given the acceptance of statistics into prescriptive models of human thought, we note recent findings in cognitive science research and adopt the sociological view of the scientific enterprise to make the argument that statistics can be seen to act as a rhetorical device in the evolution of scientific ideas.

Historically, epistemologists and philosophers of science have found it difficult to admit a role for chance in the formal enterprise of knowledge generation. The scientific method was first advocated by Sir Francis Bacon, who proposed as its principle method the process of *induction*. A seemingly intuitive process, induction and its validity were challenged in the 18th century by the British empiricist David Hume (1711-1776). Fundamental to Hume's argument was an unconcealed disdain for chance, which to him does not exist other than as a reflection of "our ignorance of the real cause of any event" (Hume and Beauchamp 1999). As did his contemporaries in philosophy, Hume regarded chance, unreason, and vulgarity as one and the same (Hacking 1990). What role is there for such an unappealing, worldly entity in the pursuit of true knowledge? Indeed, the gist of the *problem of induction* proposed by Hume is that even if our experiences have always taught us so, there does not exist any a priori justification for propositions such as "the sun will rise tomorrow" which can rule out the *possibility* that it may one day turn out otherwise. Granted, Hume did admit that chance and probability do play a role in actual learning from experiences. Yet to him the results of such learning constituted not true knowledge but mere "habit." That is, while frequentist probability is how we

learn about the world, it depends on the unappealing process of chance—it *ought* not be the basis of knowledge.

To the more practical-minded of Hume's day, the nascent probability theory found useful applications in such areas as gambling, annuities, and courtroom testimonies (Gigerenzer 1989), applications so fruitful that it was declared in 1736 that "chance is the very guide of life" (Butler 1736). Later developments in the 19th century of the so-called "granary of science" (Babbage 1856)—the vast body of accumulated statistics ranging from the movements of heavenly bodies to divorce rates—led inexorably to the philosophy of Charles S. Peirce at the turn of the 20th century. To Peirce, randomness characterizes all, not the least of which the human enterprise of science. How can we then guarantee the truth and validity of scientific inference in the face of chance? Inspired by the law of large numbers and the theory of evolution, he believed that as long as we properly manage and account for the influences of chance using statistical techniques, in the long run the knowledge that shall emerge from science will be *most likely* reflect the truth (Peirce, Houser et al. 1992). Peirce's thoughts have come to dominate current view of statistics in the scientific enterprise.

Having thus become an indispensable to the generation of scientific knowledge, statistics has also become essential to the evolution of ideas that is scientific discourse. In this context, it functions as a *de facto* rhetorical device, that is, it is appreciated largely as a powerful tool of persuasion. There are several possible reasons for making this assertion.

To begin with, I believe that statistical arguments are often persuasive because of the some believe common statistical techniques to be canonical and indisputable, as natural and invariable as the laws of physics. To wit, novices in statistics sometimes confidently invoke the normal distribution without reservation or consideration of the underlying mechanisms. One possible source of this deep conviction in certain statistical concepts may be traced to the creation of the “hybrid theory.” As Gigerenzer *et al* argued, standard statistics “cookbooks” written for students and practitioners in other fields often overlook the subtle but fundamental disputes regarding the proper use of specific statistical techniques and their interpretations, in particular the differences between Egon Pearson and Jerzy Neyman on the one hand and Ronald A. Fisher on the other. This has the effect of presenting to the users of statistics a seemingly coherent view of statistical techniques that is to be considered the canonical method of scientific inference (Gigerenzer 1989), lending the methods an aura of mathematical inevitability and conceptual infallibility. This is particularly true for researchers in the biological and social sciences, whose knowledge and expertise have few necessary connections with statistical theories.

This de facto canonicalization of statistical methods leads not just to their sometimes mechanical usage but, more importantly, to its institutionalization in scientific discourse. Philosophers of science have long argued that the intellectual enterprise of science lends itself to sociological analyses. From this perspective, it has been found that in the evolution of scientific ideas, debates regarding competing explanations can be settled not on the merits alone but on sociological considerations

such as personal relationships. In this context, the ability of the scientists to present *compelling* arguments becomes particularly important. Hence in scientific discourse, scientists are additionally obliged to employ statistics to argue forcefully for their thoughts in the competitive evolution of scientific ideas.

There may be a more profound sense in which statistical arguments may be considered rhetorical. Compared to a decision analyst who makes choices based on precise calculations of risks and benefits, the imperfect human faculty of reason, while capable of predicting the consequences of actions, is not equipped to arrive at a balance by exact computation. Our day-to-day decision-making, in fact, is guided by our emotions. An emerging consensus in neuroscience, this view was clearly elaborated by the neurosurgeon Antonio Damasio (Damasio 1994). He found that patients with a specific brain damage retained the ability to reason clearly—as far as verbal exchanges and tests of logic were able to reveal—but not the association of emotions with thoughts or ideas. This turned out to be advantageous in certain situations—for example a patient was able to clearly think and respond appropriately while his car was spinning out of control on ice. Yet, on its own, the faculty of reason is not sufficient to arrive at complex decisions. To wit, when Damasio tried to schedule an appointment with one such patient, he found the perfectly logical patient to be able to give thorough evaluations of all the candidate dates and times in terms of potential conflicts and other prior engagements. Yet he was not able to make a decision because all the options and the arguments for and against each were altogether too disorienting. Without emotions, it

seemed that complex real life situations present too complicated a decision scenario to sort through with reason alone.

Taken together, then, the evidence shows that since the inborn faculty of reason is inherently incapable of truly appreciating statistical arguments in their full numerical details, the rhetorical power of statistical arguments stem then not from an appeal direct to reason but one made to the emotions. That is, statistical theories are a formalization and extension of human reason, and are viewed and revered as such—yet their role in scientific discourse is supplanted by that view, such that they are no longer persuasive by reason per se but its perceived association with reason. Indeed, arguments based on statistics can be so persuasive that it is sometimes regarded as a deceptive tool used to the detriment of the many (Huff 1982; Best 2001). However, this may be an unnecessarily grim view of statistics. The fundamental utility of probability and statistics cannot be denied. As systematizations of the imperfect intuitive reason, they can play a legitimate role in the resolution of complex decisions.

In particular, issues and debates in the life sciences are often heavily laden with ethical, religious, legal, and political concerns. Faced with the onslaught of the sometimes undisguised efforts to impede the pursuit of scientific knowledge, I believe it is incumbent on the conscientious scientist to learn to exploit the rhetorical power of statistics to speak, as a scientist, on behalf of science.

THE DIFFICULT INTRODUCTION OF EXPERIMENTATION AND STATISTICS TO MEDICINE

The transformation of medicine into the modern scientific enterprise we now know was not by any means unimpeded. We briefly discuss two examples of the difficulties encountered in the introduction of scientific experimentation and of experiment design and statistical analysis to medical research. We additionally note their implications for biomedical research today.

The introduction of experimentation into medicine at the end of the 19th century was resisted by those who believed in *vitalism*, the thought that living substances contain a unique essence fundamentally different from inanimate objects, an essence that is not amenable to the usual methods of experimental manipulations and measurements. The most notable of the time was Xavier Bichat.

An appreciation of the unique properties of biological phenomena can be traced to the writings of philosophers centuries earlier. Perplexed by the inexplicable complexity of biological systems, philosophers had always assumed living beings, vis-à-vis inanimate objects, to contain an essence that is fundamentally different and that endows them with unique properties. This outlook underlies the *dualism* of René Descartes (1596-1650), who posited two types of substances in this world, physical matter and human souls (Descartes and Cottingham 1986). It is interesting to note that recent experimental findings in developmental psychology have revealed the Cartesian dualist view to be less the product of Descartes' meditations than it is an innate component of the human mind (Bloom 2004).

Whether philosophical or innate, this dualist view underlies the particular difficulty with introducing experimental techniques to biomedical research. To the vitalists, a science of life would have to invent a whole set of rules to account for that mysterious “life force.”

In response to vitalism, the physiologist Claude Bernard proposed that the “life forces” represent not a qualitatively different entity but a largely unobservable “internal milieu” in the body, a system that operates under the same ordinary rules of physics and chemistry and that is in principle equally amenable to experimental manipulations (Bernard 1961). His ideas are influential and are deemed critical to the eventual acceptance of experimentation in biomedical research.

Yet even today, it has been argued, our innate dualist outlook still presents barriers against biomedical research (Bloom 2004). A particularly prominent example is the restriction on stem cell research, an area of biomedical research with immense promises.

As much as Bernard is credited with bringing scientific experimentation to medicine, a different aspect of his outlook helped impede the adoption of statistical techniques. Unlike Quetelet who believed in the physical reality of statistical quantities, Bernard ridiculed such statistical concepts as averages, facetiously referring to the “average European urine” (Bernard 1961). This is in fact a recurrent theme not unprecedented in medicine. In 1836, in what may have been the first clinical trial in history, Jean Civiale compared two techniques of kidney stone removal quantitatively. While the probability theorist S. D. Poisson commended Civiale’s effort, he cautioned

against further use of probability in medicine, “in practical medicine the facts are far too few for them to enter into the calculus of probabilities... in applied medicine we are always concerned with the individual” (Poisson, Dulong et al. 1835).

Next we turn to the introduction to medicine of statistical experiment design and analysis, a transformation of medicine that was gradual and reluctant. This was initiated in the 1950s, during which practicing medical doctors saw a sudden rush of new medications such as antibiotics and steroids—each billed as the next breakthrough—that was altogether bewildering to say the least. In order for them to decide which medication to prescribe, it became clear that more efficient mechanisms were needed. At the same time, after the development of statistical theories of experiment design by R. A. Fisher, the 50s saw the adoption of statistics by various scientific disciplines. Accordingly, Fisher’s ideas on experiment design were starting to be incorporated into the scientific evaluation of therapeutics.

The difficulty with the adoption of statistical techniques in biomedical research—in particular clinical trials—was based less on conceptual differences and more on sociological factors. We illustrate this with a particularly bitter debate on a prematurely terminated clinical trial performed in the 1970s: the University Group Diabetes Program (UGDP) that studied the sulfonylureas, a class of drugs for type II diabetes mellitus. In contrast to the warm acceptance of the successful 1954 epidemiological study of the Salk polio vaccine, the UGDP study met acrimonious opposition (Marks 1997). Certainly the results of the UGDP were naturally controversial—the then widely-used first-generation sulfonylurea tolbutamide was shown to have statistically significant fatal cardiovascular

side effects. But the vociferous critics were remarkable in their persistence and their apparent objections against the *relevance* of the results of the UGDP. While concerns with relevance—in this case whether the effects observed on the UGDP subjects reflected real-world risks—may have merit, the inexplicable stubbornness of the medical critics may be best understood in the social context of the United States in the 1970s.

As a result of experiences during World War II, the field of medicine saw wide and drastic structural changes in the 1950s and 1960s: the rising importance of scientific research in general and medical research in particular, the ballooning financial support for medical research from public institutions and private advocacy groups, the expansion of the Veterans Affairs hospital system, and the federal government-sponsored constructions of community hospitals. Together they fueled a dramatic expansion of medical schools and teaching hospitals, both of which rapidly gained influence and power. At the same time, existing medical doctors at these hospitals found themselves displaced by young research-oriented faculty members, and felt their prestige and influence to be drastically reduced. In addition, in the early 1970s, there was a general sense of a “crisis” developing in the escalating cost of medical care, a situation that brought unprecedented federal interventions and deprived medical practitioners of the independence they had enjoyed for so long (Starr 1982). For the doctors who felt caught in the middle of these developments, the UGDP certainly represented a menace: the harmful side effects of tolbutamide revealed by the UGDP not only suggested the fallibility of medical doctors but also compelled the Food and Drug Administration to take the then-unprecedented step of putting a warning on

tolbutamide packaging—another significant invasion of the autonomy of medical doctors (Marks 1997).

One common theme that runs across the oppositions against statistics in medicine, from Poisson to Bernard to the opponents of the UGDP, is the legitimate concern with the *relevance* of scientific studies to medical practice. As the UGDP critic Alvan Feinstein noted, clinical relevance of a scientific study is assessed by asking “How are the patients in the study like my patients? How are they different? How are patients’ lives affected by the choice of treatment ‘a’ over treatment ‘b’?” (Feinstein 1971). (Feinstein 1971) Today we find in *evidence-based medicine* an attempt to systematically answer the questions of clinical relevance.

Science and Statistics in Modern Medicine

The acrimony that greeted the findings of the UGDP (see Section 0) can be compared with recent findings of the unexpected risks of hormone replacement therapy (HRT).

While for decades HRT has been recommended for post-menopausal women, it has only recently been found in large studies to endow little benefit on the women while at the same time slightly increase their risks for some diseases. HRT was first adopted in spite of its theoretical potential risks and the small observational studies that served as its scientific basis. It is reasonable to say that the HRT had always had an intuitive appeal: if post-menopausal women have lowered levels of sex hormone, just replace them to ameliorate the effects of menopause. Now this has been shown to be untrue for the most commonly used HRT formulation in a randomized clinical trial (Rossouw,

Anderson et al. 2002), a trial follow-up (Grady, Herrington et al. 2002), and a meta-analysis of observational studies (Nelson, Humphrey et al. 2002). Recent evidence even suggests added risks for cognitive function decline (Rapp, Espeland et al. 2003). Overall, the risk conferred by HRT over 5-year periods is about a 1% increase in major medical events (Vogel 2003).

Yet it is informative that 30 years after UGDP, both biomedical researchers and the public has come to accept the gross incompleteness of medical knowledge and the risk-benefit trade-off that seems to pervade scientific medicine. This is all the more remarkable since in comparison, the HRT has been in use longer and by more people than tolbutamide.

In hindsight, what the HRT trials has taught us is that the basic job of biostatisticians is to act as gatekeepers against unreliable information, to avoid turning sincere good-will into unintended harm at the cost of institutional credibility, to appeal not to the emotion but to the intellect.

Statistics in Modern Studies of Genetics

The causes of human diseases are many, but as biomedical research met the many challenges of common infectious diseases, physical traumas, congenital abnormalities, and metabolic disorders, the illnesses that have come to the fore are the chronic conditions that are slow in onset. These conditions represent years of accumulated interactions between the environment and the body, processes that amplify individual

genetic differences. This thesis will focus on the study of the genetic bases of such diseases.

Coincident with the increased importance of such complex genetic diseases is the completion of the sequencing of the genomes of multiple species¹ and the emergence of large-scale measurement technologies. Biomedical researchers are now in a unique position to correlate the behaviors of genes, their products, and their interactions in the context of the genome for the purpose of understanding diseases mechanisms. In this paper, work will be presented on the two major approaches to the study of the genetic bases of human diseases, *statistical genetics* and *gene expression array analysis*. Broadly, the goal of statistical genetics is to correlate genotype with phenotype, that is, the conditions and the outcomes of Nature's genetic experiments. Gene expression microarray analysis, on the other hand, studies in detail the activities of individual genes averaged over collections of cells. Changes in the activities of genes related to diseases may be fundamental to the diseases or may reflect their roles in the manifestations thereof. Given the interconnectedness of biological processes and the amplification inherent to many biological pathways, it is not unreasonable to assume that secondary effects dwarf primary effects in number and in magnitude. Thus, whereas statistical genetics focuses on the *etiological*, gene expression analysis is best suited for studying the *symptomatic*. In addition, by studying cellular processes at such a detailed level, gene expression

¹ <http://www.ensembl.org>

analysis requires proper and automatic incorporation of biological information, not to mention a generous amount of human interpretation.

These two approaches have their respective intrinsic and extrinsic strengths and weaknesses, which we will illustrate in the context of two diseases, Alzheimer's disease and psoriasis. While it is sometimes claimed that an "undeclared dispute" exists between the "classical geneticists" and the "proponents of gene expression analysis" (Darvasi 2003), a more reasonable view is that this antagonism is unwarranted and the dichotomy false. Studies that combine these two approaches have proven fruitful in dissecting complex diseases. We will briefly review one such recent work that has gracefully combined these two complementary approaches statistically using large-scale genotyping and expression profiling data (Schadt, Monks et al. 2003). We will also present our results on a similar attempt in the chapter on psoriasis (see *Section Error! Reference source not found.. Error! Reference source not found.*)

ALZHEIMER'S DISEASE AND STATISTICAL GENETICS

The common form of Alzheimer's disease (AD) is a late-onset progressive neurodegenerative disorder that surfaces after the age of 65. No specific clinical feature exists to definitively differentially diagnose AD, ante-mortem, from senile dementia, although imaging techniques are being explored for their potential in this application. Pathologically, the brains of AD patients show selective neuronal loss associated with characteristic neurofibrillary tangles in the neurons and deposits of amyloid substances

in senile plaques and cerebral blood vessels. As the demography of developed nations around the world shifts and life expectancy lengthens around the world, AD will be a major medical and societal challenge in the coming decades.

It has been observed for some time that AD can be clustered in families (“familial AD”), suggesting a strong genetic component. In fact within some families AD can be inherited in an autosomal dominant manner. As a set of methods especially appropriate for such simple genetic diseases, *linkage analysis* has revealed that in a small number of families the associated mutation lies in the amyloid precursor protein (APP) gene (Goate, Chartier-Harlin et al. 1991). However subsequent analyses did not uncover APP mutations in most familial AD patients. Given that autosomal dominant AD is genetically heterogeneous, further linkage analysis and positional cloning studies identified missense mutations in the gene presenilin 1 (Schellenberg, Bird et al. 1992; Sherrington, Rogaev et al. 1995). Its homolog, presenilin 2, was later identified as a candidate AD gene in certain European families (Levy-Lahad, Wasco et al. 1995). Consistent with the strong disease penetrances of these mutations, experimental studies have implicated these genes in the pathogenesis of AD. Linkage analysis have been indispensable to these developments in the understanding of AD, a mental disorder made that much more difficult to study biologically due to its late onset and its lack of animal models.

Linkage analysis and positional cloning played crucial roles in locating the genetic bases of simple Mendelian diseases, in particular where, as in AD, the cellular processes underlying the pathology is unknown. Today the body of knowledge that has

been accumulated on simple genetic diseases is impressive. As of July 2004, the online database for such mutations hosted by the National Center for Biotechnology Information (NCBI²), the Online Mendelian Inheritance in Man (OMIM³) contains 15,490 genes/loci and their allelic variants and corresponding Mendelian phenotypes. While the number is not directly comparable, the human genome is only about twice the size at around 25,000 genes according to the annotation by Ensembl⁴. Important lessons have been learned regarding the nature of such mutations in evolution and in diseases (Botstein and Risch 2003).

Although linkage analysis has helped the discovery of the genetic causes for many familial ADs, autosomal dominant forms attributable to mutations in the APP, PS1, or PS2 genes constitute only 0.4%, 2-3%, and 0.1% of all AD cases, respectively. Together with familial cases of unknown genetic etiology, only 5% of all AD cases are attributable to autosomal dominant forms (Richard and Amouyel 2001). The non-familial form of AD, the “sporadic cases,” in fact represents the majority of AD cases.

It is widely agreed that if there is any genetic component to the sporadic ADs, they are not disease-causing like the missense mutations described previously but are disease-predisposing. In an influential paper based on theoretical analysis, Risch (Risch and Merikangas 1996) found that linkage analysis is not likely to provide enough power

² <http://www.ncbi.nlm.nih.gov/>

³ <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM>

⁴ http://www.ensembl.org/Homo_sapiens/

to detect the rather subtle risks conferred by susceptibility alleles. In these cases linkage disequilibrium (LD) studies will be needed.

The most potent AD-susceptibility gene now known is the apolipoprotein E (APOE) gene $\epsilon 4$ allele, which confers in Caucasian populations relative risks for AD of 12.5 and 2.7 to homozygous and heterozygous carriers, respectively, compared to non-carriers (Farrer, Cupples et al. 1997). The identification of APOE was motivated by biological studies in which ApoE from the cerebrospinal fluid was found to bind to APP with high avidity (Strittmatter, Saunders et al. 1993) and studies that showed ApoE co-localizes with senile plaques in AD patient brains (Namba, Tomonaga et al. 1991). ApoE seems to be intimately involved in the pathogenesis and disease progression of AD. In fact for certain candidate AD treatments only $\epsilon 4$ carriers show positive response compared to patients on placebo treatments (Richard, Helbecque et al. 1997).

Other genes that have been proposed as candidate susceptibility genes include the ApoE receptors LRP (Beffert, Arguin et al. 1999) and VLDL-receptor (McIlroy, Vahidassr et al. 1999; Helbecque, Berr et al. 2001), the protease gene α_2 -macroglobulin (Blacker, Wilcox et al. 1998), and the angiotensin I converting enzyme (Crawford, Abdullah et al. 2000; Farrer, Sherbatich et al. 2000; Narain, Yip et al. 2000). It should be noted that contradictory results have been reports for each of these candidate genes. This may be a result of the significant role played by different genetic backgrounds, sampling biases, and the complex and multi-layered sets of interactions that translate genotype to phenotype.

PSORIASIS AND GENE EXPRESSION PROFILING

Psoriasis (PS), literally an “itchy condition,” is a skin disorder that causes localized reddening of the skin accompanied by itchiness and flaking of the skin. There are four sub-types, plaque-type, guttate, erythrodermic, and pustular. It affects 0-5% of the population, depending on ethnicity. While not life-threatening, psoriasis has no known efficacious treatment. Recent evidence suggests that PS is not a pure skin condition and may have immune-related etiology. Detailed studies on PS may shed light on other autoimmune diseases such as type I diabetes mellitus.

Like AD, the genetic component of PS has also been long recognized. In a twin study, monozygotic twins were found to have a concordance rate of 70% and dizygotic twins 23%; where concordant twins showed similar age of onset, body distribution, and severity and course of PS (Farber, Nall et al. 1974). However unlike AD, to date there are only a few reported cases of familial forms of PS with dominant inheritance patterns (Tomfohrde, Silverman et al. 1994; Matthews, Fry et al. 1996). Linkage analyses have found 16 PS susceptibility loci (more details in *Section Error! Reference source not found.*), the strongest of which, PSORS1, lies on the short arm of chromosome 6 in the MHC, a tract of the human genome seemingly dedicated to coding the immune system (Capon, Munro et al. 2002). Genetic heterogeneity has also been demonstrated in a PS study where the susceptibility MHC locus was shown to be associated with all PS subtypes except pustular psoriasis of the palms and soles (Asumalahti, Ameen et al. 2003).

There have been a number of reported expression profiling studies of PS in the literature (Bowcock, Shannon et al. 2001; Oestreicher, Walters et al. 2001). Gene expression microarray profiling measures the abundances of different mRNA species and provides snapshots of the activities of the transcriptome. The underlying technology allows for unprecedented parallelism in such measurements and represents a breakthrough in biotechnology. In this paper, we will present an expression profiling study of PS that is unique in several respects. First, by using all five chips from the Affymetrix U95A-E series as opposed to using just the A chip as is usually done, this study is unprecedented in its comprehensiveness of measurement. Secondly, by sampling both the PS-involved skin along with PS-uninvolved skin, we may be able to distill the subtle signals of susceptibility genes at the mRNA level. Thirdly, since the skin is a relatively simple organ, this study may serve as an example of the study of complex interplays between two biological systems, in particular the epidermis and the immune response. On the other hand, by sampling the whole depth of the epidermis, the mRNA abundances we measure represent averages over a large heterogeneous collection of cells. In addition, due to the localized inflammation in PS-involved skin, the concentration of cells, cell types, and mRNA may not be directly comparable across samples. We will also present an example where differences in cellular/mRNA concentrations may resolve a contradiction with results previously reported in the literature.

In order to study the effects of genetic susceptibility factors, the common polymorphism HLA-Cw6 within the MHC susceptibility locus was genotyped. We will

present some results from an analysis of expression data in combination with HLA-C genotype status. While the study was case-control in design, it is interesting to consider the possibility of studying the genetics of PS in a family-based design to fine-map the known psoriasis susceptibility loci (*Section Error! Reference source not found.. Error! Reference source not found.*).

By virtue of the comprehensiveness of this transcriptome survey, we have an opportunity to characterize the activities of biological processes in the various skin samples. We make use of the annotations in GeneOntology⁵ to provide a global view of the transcriptome. We interpret the results based on our previous experience in the analysis of the yeast transcriptome (*Section Error! Reference source not found.. Error! Reference source not found.*). Our global measurement of human mRNA abundances in addition allows us to study the regulations of subgroups of genes. We will present some results that bring together the gene regulation results that build on the collection of promoter sequences we have derived (*Section Error! Reference source not found.. Error! Reference source not found.*) and uses the transcription factor binding site prediction methods described in *Section Error! Reference source not found.. Error! Reference source not found.* to study regulatory mechanisms that may play roles in the psoriatic skin.

⁵ <http://www.geneontology.org>

COMBINING STATISTICAL GENETICS AND GENE EXPRESSION ANALYSIS

There are several reports in the literature that successfully combined genotype status with expression analysis. One notable example was in the identification of the complement factor C5 as a candidate susceptibility gene for a murine model of allergic asthma (Karp, Grupe et al. 2000). Here QTLs identified in previous linkage analysis studies were used to be *logically* combined with differential expression. In particular, the genomic locations of differentially expressed genes were overlaid with QTLs derived from previous studies (Ewart, Kuperman et al. 2000). The C5 gene surfaced in this intersection analysis and its plausibility as a susceptibility gene was further supported experimentally.

In a recent study, linkage analysis and expression analysis were combined *statistically* in a study on genes related to obesity (Schadt, Monks et al. 2003). By using the expression levels directly as quantitative phenotypes, *expression QTLs* (eQTLs, in contrast to *clinical QTLs*, or cQTLs) were derived that are in linkage with changes in expression level. In fact, many of the eQTLs uncovered are located in the chromosomal neighborhood of the corresponding genes.