2004 CURT STERN AWARD ADDRESS
The SNP Endgame: A Multidisciplinary Approach[*]

Neil Risch

Department of Genetics, Stanford University School of Medicine, Stanford; and Division of Research, Kaiser Permanente, Oakland

Neil Risch

I would like to thank Dr. Rosenberg and the ASHG Awards Committee for this exceptional honor. It is a particular privilege to receive this award from Dr. Rosenberg, who played such a significant role in my career development by recruiting me to Yale 20 years ago. When Lee first contacted me regarding the Curt Stern Award, he told me that the best part of being chair of the Awards Committee is being able to inform the award recipients. Well, as he found out, informing me turned out to be no modest task. His first mistake was trying to contact me by e-mail. Many of you know the challenge of trying to reach me by e-mail. But he was persistent and followed up with phone calls. I did call back, but inevitably after he had left the office (he has normal office hours; I do not). Eventually, we did make contact, and the call was predictably celebratory.

Receiving an honor such as this provides an oppor-tunity to reminisce and examine not only the past decade (for which the Stern Award is designated) but also previous decades of research in human genetics and in my particular area, genetic epidemiology, as well. I am particularly aware of the fact that I am not a bench scientist but a population and statistical researcher, which makes the decision of the Awards Committee that much more meaningful to me, in recognizing the contributions that statistical types make and have made to the field of human genetics.

In tracing my own personal history and development in this field and to gain some perspective, it was useful to examine the parallel developments going on in human genetics, genetic epidemiology, and statistical genetics. In doing so, I also came to appreciate that both deliberate as well as random events shape one's career and that both are important. Perhaps an analogy is provided in the population-genetics concepts of natural selection and genetic drift, both of which contribute to the molding of gene frequencies in a population over time.

In table 1, I have outlined my own personal career course in parallel with concurrent events in human genetics and genetic epidemiology. I have started with my graduate-school training, since, up to that point, I was a pure mathematician. Having discovered the field of biomathematics and a newly kindled interest in biology, I was fortunate to find the recently established graduate program in biomathematics at UCLA. In my second quarter in that program, I enrolled in a course in human genetics taught by Anne Spence and John Merriam. After a few weeks, I knew I had found my discipline. I was fascinated not only by the natural logic of genetics and its relevance for human development but also by the mathematical elegance underlying its principles. I also soon learned that, not surprisingly, many of the early developments in the field of statistics derived from problems arising in genetics. I was also lucky to have selected a graduate program whose faculty included two leading statistical geneticists, Ken Lange and Anne Spence, who became my graduate advisers. While at UCLA, I also was fortunate to have exposure to the field of population genetics through course work taught by Ayesha Gill in the Biology Department. It became another passion of mine, although it would be years before I was able to act upon it.

[*] Previously presented at the annual meeting of The American Society of Human Genetics, in Toronto, on October 30, 2004.

Table 1

A Personal History

Dates	Location	Field
1973–1979	Graduate school, UCLA, Department of Biomathematics. Advisers: Ken Lange, Anne Spence, and Ayesha Gill	Family, twin, and adoption studies; path analysis, segregation analysis, and pedigree analysis; Morton-Elston debate over major-gene evidence
1979–1984	Columbia University, Department of Psychiatry, School of Medicine, and Department of Biostatistics, School of Public Health. Advisers: Zena Stein and Mervyn Susser (Epidemiology)	Linkage analysis in pedigrees (Elston and Stewart [1971] algorithm, LIPED computer program [Ott 1974]); RFLPs
1984–1994	Yale University, Department of Epidemiology and Public Health (Biostatistics) and Department of Human Genetics	Linkage analysis in pedigrees continues; microsatellites, positional cloning, Human Genome Project inception; linkage analysis of complex traits (small families and sib pairs)
1995–2004	Stanford University, Department of Genetics, Department of Statistics, and Department of Health Research and Policy	Resequencing of the human genome; SNPs; association studies; linkage disequilibrium; HapMap; population genetics and epidemiology; admixture mapping

In the mid-1970s, the field of genetic epidemiology focused primarily on the estimation of heritability of traits via family, twin, and adoption studies; path analysis; and inferences regarding the role of major genes in the familial aggregation of traits through segregation analysis. Complex models for segregation analysis—including major-gene and polygenic components and parametric transmission parameters—evolved, along with intense debates about the sensitivity and specificity of methods focusing on nuclear families that allowed for genetic alternatives to major-gene inheritance (i.e., polygenic inheritance, as promoted by Newton Morton and colleagues) versus extended family pedigrees that did not allow alternative genetic models but did allow for transmission patterns to be non-Mendelian (promoted by Robert Elston and his colleagues). Analysis of extended pedigrees was made possible by Elston and Stewart's recursive algorithm (Elston and Stewart 1971). Their seminal paper also showed how to efficiently calculate LOD scores for linkage analysis in extended pedigrees, perhaps its most far-reaching impact. Soon afterward, Jurg Ott developed the program LIPED (Ott 1974), which served as a mainstay for performing parametric linkage analysis for nearly a decade.

Upon my move to my first position at Columbia in 1979, I was well trained in a variety of disciplines, including mathematics, statistics and biostatistics, biomathematics, biology, and genetics. The rigors of the Ph.D. program at UCLA also required a year of medical-school course work. Thus, I felt I was well prepared to take on the role of statistical geneticist, applying mathematical models to the analysis of genetic and family data to better understand the inheritance of human diseases and traits. Although segregation analysis was still flourishing, linkage analysis was becoming more prominent because of the availability of computer programs for performing the complex calculations that previously were cumbersome. However, running programs such as LIPED in those days was still not without challenge, since computers had not yet evolved to the point that data input was electronic. The program and the pedigree information needed to be read from hundreds of computer cards, which were often dropped, torn, or mangled in the card readers. How things have changed!

The greatest limitation to linkage analysis in those days, however, was the paucity of genetic markers that could be employed, since they were phenotypically based on blood groups and serum proteins. Thus, the most significant developments influencing the course of human genetics were the recognitions that the DNA itself could serve as the basis for genetic markers (Kan and Dozy 1978) and that restriction-fragment–length polymorphisms that spanned the entire human genome could be derived and used for linkage analysis (Botstein et al. 1980). The early identification of the location of the gene for Huntington disease on chromosome 4 (Gusella et al. 1983) demonstrated that linkage analysis in humans was no longer an arcane, low-yield dalliance but rather the first step in a realistic effort to identify mutations underlying human genetic disease. At this point, there was also the recognition that simultaneous analysis of adjacent markers could provide greater precision in locating disease genes, and statistical methods and software for generating multilocus LOD scores for linkage analysis were derived (Lathrop et al. 1985).

During my years at Columbia, other random events occurred that would greatly influence my career direction. After my first 2 years, the funding with which I was brought there unfortunately and unexpectedly disappeared; I was advised to apply for grant money to support myself and specifically was directed to apply for a career-development award. At Columbia in those days, there were no true genetic epidemiologists or statistical geneticists, so it was entirely unclear who could serve as a mentor on my grant. One of my colleagues in psychiatry recommended that I contact Zena Stein, an epi-

demiologist with an interest in birth defects who resided in the School of Public Health. It was a fortuitous recommendation. After meeting with Zena, I recognized that, despite all the extensive training I had had at UCLA, the one area that was lacking in my background was epidemiology. This gap, previously unrecognized, now seemed particularly prominent, given my self-description as a genetic epidemiologist (i.e., an epidemiologist with no formal training in epidemiology). Zena generously offered to be my mentor, and both she and her husband, Mervyn Susser, spent the next 2 years providing training in epidemiology for me. Again, I was fortunate to have landed in an outstanding environment to obtain such tutelage. It didn't take me long to realize that epidemiology had much in common with population genetics, since both disciplines involve the study of patterns of disease in populations, with epidemiology focusing heavily on measurement and environmental components and population genetics focusing on explanations for observed patterns of gene frequencies. More than anything else, studying epidemiology broadened my perspective about the multifactorial nature of disease causation and the need for a framework to consider host and exposure factors together.

My recruitment to Yale in 1984 provided me an opportunity to flourish, with a joint appointment in the two departments that most closely reflected my academic interests—the Department of Epidemiology and Public Health (Division of Biostatistics) and the Department of Human Genetics. For the next 10 years, my work reflected the developments occurring in molecular biology and human genetics. The development of short-tandem-repeat markers, or microsatellites, spanning the genome (Litt and Luty 1989; Weber and May 1989) greatly facilitated human linkage analysis, and the era of positional cloning had come to the fore. Parametric multipoint linkage analysis was established as an essential tool for disease-gene mapping. Because my appointment at Columbia had been in psychiatry and I had worked on psychiatric disorders—and this remained true also at Yale—my attention was focused on the role linkage analysis could play in identifying the genetic contribution to complex, or non-Mendelian, disease. To that point, the dogma was that linkage analysis could be employed only for traits that had been previously subjected to segregation analysis and had shown evidence for a major-gene contribution and that such genetic contribution could then be modeled in linkage analysis based on the results of segregation analysis. However, my experience with segregation analysis at the time was that results were often equivocal—sometimes producing no or poor evidence of a major gene when one exists, other times producing evidence for a major gene when none exists. I also had learned from the example of HLA-associated diseases such as type 1 diabetes that clear linkage evidence could be obtained, even from relatively small samples of sib pairs, when prior segregation analysis was equivocal in accurately demonstrating the mode of inheritance. In fact, for complex traits, segregation analysis rarely, if ever, has the power, when based simply on family trait patterns, to illuminate complex genetic architectures. I therefore examined the range of models of inheritance and the magnitude of genetic effects that could be detected through linkage analysis, using small constellations of affected family members, even in the absence of prior segregation evidence. I also realized that understanding inheritance of complex diseases would require the use of genetic markers. It was clear that the power of linkage analysis depended primarily on the magnitude of effect of the locus under consideration, most simply measured by its impact on familial aggregation (for a discrete trait) or by proportion of variance explained (for a continuous trait). Because the molecular and statistical approaches for identifying Mendelian genes were well in hand, the field generally started to refocus attention on the more common complex diseases (often characterized as multifactorial), since the public health impact of these disorders is markedly greater. Sibship-based linkage studies for common diseases started to abound, but the success rate did not match the spectacular history of linkage analysis in the Mendelian setting.

While at Yale, my collaboration with epidemiologists, both faculty and students continued. One colleague I worked with extensively, both in teaching and research, was Kathleen Merikangas, a psychiatric epidemiologist with interests in genetics. We spoke frequently about the state of the field of genetic epidemiology and where it was and should be going. We continued these discussions even after my move to Stanford in 1995. We began to develop an awareness that the linkage approach, although having some modest success in complex diseases, was unlikely to identify the large majority of genes. We were influenced by a news item that appeared in *Science* on July 14, 1995, entitled "Epidemiology Faces Its Limits." The article described the difficult state of affairs experienced by epidemiologists in trying to reliably identify disease risk factors when the relative risks are not large. Although the article did not discuss human genetics or genetic epidemiology, we realized that many of the comments could apply to the developing situation in human genetics as well.

We wrote a perspective for *Science*, at first reflecting the parallel trends occurring in epidemiology and human genetics. The original title of the article was "Human Genetics Facing Its Limits." However, upon reflection, we realized that the tone of the article was too pessimistic. It was our belief that if we were to publish such an article, it would not be sufficient simply to point out the limitations of current approaches; we needed to pro-

vide an alternative. The answer, once again, came from developments in molecular biology—namely, the Human Genome Project. If we could have any tool to use for mapping disease genes, we wondered what would it be? Again, on the basis of my experience with HLA-associated diseases and my knowledge about disease associations with other blood-group systems, I knew that many of these associations, although highly significant statistically, would not produce substantial or robust linkage signals. Therefore, why not reverse the process of positional cloning? Instead of searching randomly through the genome by location, why not start with genetic variants and test them directly as candidates? The problem with candidate-gene association studies had been the limited number of candidates and, therefore, the low prior probability of a "hit." But what if we could compile a list of all polymorphisms in the human genome? The number would certainly be large, but what impact would that have on power if we needed to use a very strict level of significance to protect against false positives? By use of simple power calculations, we showed that, even with use of an extremely strict level of significance, the power of association studies would be dramatically greater to detect genes of modest effect than would linkage analysis. Indeed, in many cases, the sample sizes required for linkage would be unrealistically large. Figure 1 reproduces the table we presented in the *Science* perspective that shows the dramatic difference in required sample size. Therefore, we concluded that the main impediment to the identification of disease genes of modest effect was not statistical but the lack of the appropriate reagent-—namely, a compendium of the polymorphisms present in the human population. To provide a more optimistic vision, we changed the title of the perspective to "The Future of Genetic Studies of Complex Human Diseases" (Risch and Merikangas 1996).

The writing of this perspective brought me back full circle to my graduate training. After having spent the prior 15 years focusing on genetic epidemiology and the study of disease, I realized fully the implications of what we were proposing: that it would bring population genetics into greater prominence in human genetics. This is because genetic association studies require knowledge of both epidemiological and population-genetic principles. It is important to know how allele frequencies vary in the population and how linked variants are associated with each other (i.e., patterns of linkage disequilibrium). Equally, it is important to avoid confounding (an epidemiological principle) due to nonrandom mating.

It has been 8 years since our perspective was published, and >7 million confirmed SNPs (single-nucleotide polymorphisms) are currently listed in the SNP database dbSNP. Molecular technology for sequencing multiple genomes and for identifying and characterizing genetic variation in the human population has proceeded apace. Patterns of linkage disequilibrium are being characterized and attention being paid to how such patterns—and genetic structure itself—vary across populations of differing geographic origin. Large-scale association studies are being planned. Much has also been written about different strategies for conducting genomewide association studies—namely, a map-based approach based on patterns of linkage disequilibrium, sometimes referred

Genotypic risk ratio (γ)	Frequency of disease allele A (p)	Linkage		Probability of transmitting disease allele A $P(tr-A)$	Association			
		Probability of allele sharing (Y)	No. of families required (N)		Singletons		Sib pairs	
					Proportion of heterozygous parents (Het)	(N)	(Het)	(N)
4.0	0.01	0.520	4260	0.800	0.048	1098	0.112	235
	0.10	0.597	185	0.800	0.346	150	0.537	48
	0.50	0.576	297	0.800	0.500	103	0.424	61
	0.80	0.529	2013	0.800	0.235	222	0.163	161
2.0	0.01	0.502	296,710	0.667	0.029	5823	0.043	1970
	0.10	0.518	5382	0.667	0.245	695	0.323	264
	0.50	0.526	2498	0.667	0.500	340	0.474	180
	0.80	0.512	11,917	0.667	0.267	640	0.217	394
1.5	0.01	0.501	4,620,807	0.600	0.025	19,320	0.031	7776
	0.10	0.505	67,816	0.600	0.197	2218	0.253	941
	0.50	0.510	17,997	0.600	0.500	949	0.490	484
	0.80	0.505	67,816	0.600	0.286	1663	0.253	941

Figure 1   Reproduction of the table by Risch and Merikangas (1996). Used with permission from the publisher.

Table 2

Some Fields That Will Contribute to the Future of Complex-Disease Genetics

Field	Contribution
Human genetics	Correlate genotypes with in vivo phenotypes
Molecular biology	Make genotyping and sequencing more efficient and cost-effective
Human evolution and population genetics	Understand the distribution of genetic variation in the human population and how it is structured
Comparative evolution	Predict functional elements of the human genome, especially outside coding regions
Functional genomics and cell biology	Establish relationships between sequence variation and in vitro phenotypes
Statistics	Develop and apply novel methods for analyzing large volumes of SNP data
Epidemiology	Develop and maintain large, population-based clinical databases for application of genomic technology and epidemiological methods; environmental risk factors
Medicine	Further refine clinical entities to reduce etiologic heterogeneity
Bioinformatics	Develop, maintain, and manipulate large databases storing clinical, genetic, and epidemiological data
Pharmacogenetics	Discover and characterize the role of genetic variation in treatment response and side effects

to as "the HapMap" (Collins et al. 1997; International HapMap Consortium 2003)—versus a sequence-based approach, with focus on the variation occurring in coding and adjacent regions previously implicated in the vast majority of positionally cloned Mendelian genes (Risch 2000; Botstein and Risch 2003). Also, optimal epidemiologic study designs for this endeavor have been extensively discussed, with focus on nested case-control studies versus cohort studies (Caporaso et al. 1999; Langholz et al. 1999; Clayton and McKeigue 2001). Plans are under way to develop large-scale cohorts or their equivalent for executing these studies in an epidemiologically rigorous and reproducible way.

Although the confluence of human population genetics and epidemiology for the future of human-genetics studies appears on the horizon, it is clear that many disciplines must contribute to this effort if we are to be successful on a large scale; some of these disciplines have been listed in table 2. Identification of functional elements in the human genome and their relevance for human traits and disease will require a concerted effort among scientists from a broad array of disciplines.

In summary, I must also recognize the many people who have contributed to and influenced my career and development—my many teachers, students, and colleagues. I am deeply indebted to my graduate-school advisers and mentors, Ken Lange and Anne Spence, for introducing me to the fields of human genetics and statistical genetics; Ayesha Gill, for teaching me population genetics; and all three, for their support and guidance through graduate school. I am also grateful to Zena Stein and Mervyn Susser for the excellent postgraduate supervision and training they provided me in epidemiology. I must also acknowledge my lifelong mentors, my parents, Frank and Sonya Risch, who were constant and uncompromising in the support and love they showed me in all of my academic endeavors.

Finally, I would I also like to comment about the man

for whom this award is named, Curt Stern. Much has been said by previous recipients of this award about Stern's seminal research contributions to genetics. I would like to focus on another aspect of his career; that is, the positive impact he had in mentoring his trainees. It has been noted by many who had the good fortune to sit in on his classes at UC Berkeley that Stern was a gifted teacher, and many have also written about how generous and supportive Stern was as a mentor and adviser. I would like to reflect on one story that has particular meaning for me, an anecdote reported by one of his former advisees, the great human geneticist James Neel. After the publication of the *Science* perspective I wrote with Kathleen Merikangas (Risch and Merikangas 1996), a number of colleagues contacted us to say that the numbers in the table (see fig. 1) did not conform to the formulas given in the paper. Re-examination revealed to me that I had created an error in translating the formulas into the computer program I had used to generate the number of families required for linkage analysis, with the effect that the numbers given in the table were too small by a factor of ~2. Apparently, during his graduate career, Neel had also made some embarrassing error and had to confront his adviser, Curt Stern, about it. The following is Neel's description of the incident (Neel 1983):

> During my first year of graduate work with Stern I did some incredibly stupid thing. I do not remember what it was, but it would be impossible to forget how Stern handled it. Having straightened me out, he paused a minute. I was waiting for the next blow to fall when he smiled and said, "Jim,...great men make great mistakes."

References

Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for Mendelian disease,

future approaches for complex disease. Nat Genet Suppl 33: 228–237

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331

Caporaso N, Rothman N, Wacholder S (1999) Case-control studies of common alleles and environmental factors. J Natl Cancer Inst Monogr 26:25–30

Clayton D, McKeigue PM (2001) Epidemiological methods for studying genes and environmental factors in complex diseases. Lancet 358:1356–1360

Collins FS, Guyer MS, Chakravarti A (1997) Variations on a theme: cataloging human DNA sequence variation. Science 278:1580–1581

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21:523–542

Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, et al (1983) A polymorphic DNA marker genetically linked to Huntington's disease. Nature 306:234–238

International HapMap Consortium (2003) The international HapMap project. Nature 426:789–796

Kan YW, Dozy AM (1978) Polymorphism of DNA sequence adjacent to human β-globin structural gene: relationship to sickle mutation. Proc Natl Acad Sci USA 75:5631–5635

Langholz B, Rothman N, Wacholder S, Thomas DC (1999) Cohort studies for characterizing measured genes. J Natl Cancer Inst Monogr 26:39–42

Lathrop GM, Lalouel JM, Julier C, Ott J (1985) Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. Am J Hum Genet 37:482–498

Litt M, Luty JA (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac-muscle actin gene. Am J Hum Genet 44:397–401

Neel JV (1983) Curt Stern, 1902–1981. Annu Rev Genet 17: 1–10

Ott J (1974) Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. Am J Hum Genet 26:588–597

Risch N (2000) Searching for genetic determinants in the new millennium. Nature 405:847–856

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Weber JL, May PE (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet 44:388–396