

Logical Reasoning In Human Genetics

Human genetics is one of the most active fields in all of modern science, with very large research investment being made worldwide. A proliferation of journals and unprecedented flow of new papers reporting findings from across the spectrum of traits might suggest that this field is succeeding at a rate greater than almost any other in history.

Yet, this has turned out to be a complex area of science. There has been dramatic success in identifying genes associated with classical 'mendelian' diseases, in which it is clear in advance that a gene is responsible. There have probably been hundreds of such successes. In addition, genes that contribute to or even cause some cases of complex chronic diseases have been found. This course will present the strategies that have been used to find these genes, their rationale, and aspects of successful study designs.

At the same time, this success has been tempered by the etiological complexity even of the simplest traits, those dominated by effects of a single gene. There is a big difference between identifying a gene and using that knowledge successfully, which has proven to be much more difficult. Even for truly genetic disorders understood for decades, such as sickle cell anemia or cystic fibrosis, successful *genetically based* interventions are rare, though in some cases genetic counseling has been feasible and effective as a public health measure.

Success with simple diseases—at least in the sense of identifying their causal gene—has led to enthusiastic promises of major relief from disease that will come from the genetic pursuit of complex diseases. But the case has turned out to be rather different. For example, many genes have been claimed to affect the risk of these diseases, yet those effects that are confirmed are usually rare, and/or of much weaker effect than originally claimed. Effects of environmental exposures and genotypes of other loci probably play a major role in the difficulties being faced, and indeed modify the risk associated even with the most clear-cut etiological effects. But even such statements, though widely accepted, have not led to a full understanding of any of these disorders. Despite extensive effort and refinement of current approaches, there is still little empirical evidence or theoretical basis on which to base predictions of imminent success. Much remains to be learned and probably will require new thinking about the problem. Progress can be made by using an understanding of biology, evolution, and genetics, to understand why the attempt to understand them has been frustrating so far.

For example, multiple competing or complementary risk factors are poorly accounted for in genetic and epidemiological studies alike, for reasons that can be explored and understood. Similar claims about the environmental component of these traits, to those often made about their genetics, have motivated many epidemiologists to try to consider genetic factors as nuisance parameters in their studies of environmental risk, while geneticists do roughly the opposite in what is typically perfunctory inclusion of measures of environmental factors in their studies.

This course is designed to look at the overall issues in concept, data, and method, to consider the present state of affairs in terms of the origin and nature of human genetic variation, the nature of genotype-phenotype relationships, and the way that studies are being designed to understand them. We will use computer software to evaluate the power of different study designs to identify genetic or environmental risk factors under a variety of complex sets of assumptions about these genotype → phenotype relationships to try and make some conclusions about power and efficiency of studies of various sorts, and the robustness of study designs to the underlying sets of assumptions.

Beginning from evolutionary first principles about the origin and nature of human phenogenetic variation will give a conceptual basis for any investigator's choice about what kinds of biologically sound models might be reasonable for the trait s/he is interested in. This can help motivate decision-making about study design and power to make reliable inferences in studies, based on something more scientific than the usual calculations and models based on assumptions that are often used as the basis for a given research project, and lead, often predictably, to unsatisfying results.

It will be shown that methodological details are often not the critical issue that should be used to design genetic studies with optimal chances of success. Often, the biological problem is not adequately identified, around with to design family or population based ascertainment schema that can most powerfully lead to identifying and characterizing the important genetic (and environmental) factors. Sometimes, it may be better not to attempt such studies. The issue of importance is how (and when or whether) to design a study to find genes, not what statistical methods should be used to analyze the data.

Logic reasoning in advance, and in the interpretation of results, cannot provide miracle answers, but an improved basic perspective on these issues might be helpful in planning future studies, and to know better what the limitations of genetic and epidemiological approaches to complex traits and public health may be.

PROPOSED GENERAL AGENDA

Day 1: Introduction to the problem

AM Philosophical overview

Some important conceptual questions

Logic, reasoning, and the development of theory

Induction, deduction

Popper and positivism, falsifiability

Concepts of causation

Statistical inference and statistical causation

The Scientific Method (do we actually use it?)

Contingent reasoning, probability, and the logic of science (Bayes etc., Jaynes etc.)

Ad hoc, post hoc and emotional factors in science

Use and extrapolation rather, than testing, of theory: applied vs basic science?

Does human genetics solve problems?

The genetic theory of evolution and the rise and patterning of variation

The history of evolutionary thinking

Darwin, Mendel, and the Modern Synthesis

The discovery of DNA and the Central Dogma

Population genetics introduction

First principles: random mating, mutation, migration, natural selection

Building a model: finite populations with drift

Mutation, recombination and the generation of variation

Consequences of genetic evolution: age-area correlations, admixture phenomena and allelic associations within and between chromosomes, LD and allelic association along chromosomes

Phenotypic and phenogenetic drift

The resulting nature and structure of the genome

Basic gene and chromosome structure

Modules, repeat elements, and the like

Evolution by duplication: Gene origin by exon shuffling

Gene families and gene duplication: gene ontologies

Gene regulation by *cis*-factors and enhancer evolution by mutation

Divergence and redundancy in function

The consequences of different types of mutation

Basic concepts of linkage and linkage disequilibrium mapping.

What is linkage and how does it work.

Linkage disequilibrium is linkage in the population.

How do we map a gene with known genotype against a map of markers?

Lod scores, recombination fractions and parametric linkage analysis.

Linkage and linkage disequilibrium are measures of correlation among genotypes of different loci, and have nothing to do with phenotypes.

Discussion and/or student work

Day 2: Simple traits

AM - Introduction to the notion of simple “Mendelian” traits

What did Mendel show (and what did he not show)

Archibald Garrod and the origins of human genetics

Power is determined by detectance, or $P(\text{Genotype} | \text{Phenotype})$, rather than by penetrance, or $P(\text{Phenotype} | \text{Genotype})$.

Conditional probability and conversion of penetrance (risk) models to detectance models

Linkage mapping of simple traits – detectance convoluted with linkage between loci

Linkage disequilibrium mapping of simple traits – detectance convoluted with LD.

Introduction to software for analytical and simulation based power analysis.

What evidence is needed before you do a 'genetic' study?

PM - Methodological approaches to simple traits and how they work.

How studies have been designed

How things have been found

What has been found: simple traits aren't so simple after all

The uncertainties and generalizations

Illustrated by examples from 'simple' (single-gene) traits

(detectance is estimable from data, but what about penetrance?)

Simulation of simple datasets with simple models

Lod score analysis methods and study designs

Model-free analysis methods and study designs

Linkage disequilibrium analysis methods and study designs

Ascertainment and ascertainment bias

Discussion and/or student work

Day 3: Complex traits

AM - Phenogenetics and the consequences of evolution by phenotype

Information feed-forward from the genome: The protein code

Organisms as differentiated entities

Feedback to the genome. cis-regulation and differential gene expression

Functional arbitrariness, pleiotropy, and multiple layers of 'coding'

Single-gene function

Genome structure mapped onto phenotype: families of function

Process as trait

Development and homeostasis

Relevance of model systems

All is complex, but not all is chaos

"Goings on in Mendel's Garden"

PM Detectance, study design, and linkage and LD analysis of complex traits

How studies are designed

How things have been found

The uncertainties and generalizations

Illustrated by examples from 'complex' (multifactorial) traits

Effects of reductions in penetrance

Effects of allelic heterogeneity

Effects of locus heterogeneity

Effects of multiple gene models

Why are complex traits mappable in animal models? How to interpret the results?

Discussion of methods of experimental genetics and their use in human genetics

Natural experiments as approximations to experimental methods

Effects of study design on detectance and power for linkage and/or association mapping

Though we do not know what truth is, do we know anything about what it is NOT?

Simulation studies of more complex models – effects of complexity on detectance and

power
Family size, multiplex pedigrees, random sampling and so on.
Discussion of assigned project, and/or examples (students can present their own?)

Day 4: Non-genetic complications and some inferential issues

AM - Environmental and cultural cofactors and the like

Standard notions and approaches in environmental epidemiology.
Point source risk factors (bugs) to quantitative risk factors (environment) to point source assuming (genes).
General strategies: matching, dietary recall, interview, random digit dialing, surveying grocery bills, physiological monitoring (e.g., glucose levels, anthropometrics), medical records (on probands, and on their families), record linkage. How many cigarettes? Where did you work? How much education, income, telephones, x-rays.
Hard to measure things that may be overwhelmingly important: exercise, stress, SES, culture.
Cohort vs case-control vs case-only studies.
The 'new' idea of Mendelian whatever (Davy Smith??).
Agricultural models: insecticide, fertilizer, etc.
Experimental designs and approaches in agriculture and what they tell us.
Ames testing and what it tells us.

Evolutionary aspects – we evolved in dramatically different environments from what we live in today... Consequences and ramifications

PM Effect size estimation, attributable risk, and their equivalents in genetics.

What we do in genetics (major challenges and misperceptions)
Study designs in human genetics to look at genetics and environment jointly
Sampling on exposure variables and detectance distributions
Population cohorts and detectance distributions
Korean diaspora project

Why one might be able to estimate risk models in epi studies but cannot in genetics.
Retrospective data to make prospective risk assessments.
Secular trends (unpredictable).
Feedback (e.g., sudden switch to Atkins diet),
Admixture and its complexities
Migrant studies.

Study design differences between epidemiology and genetics.
Ascertainment bias and confounding as useful in genetics, perilous in epidemiology
Effects of multiple testing.
Locus specific heritability and attributable risk
Relative risks and measures of familiality.
Epistasis bleeds into the marginals

Simulation studies of more complex models including environmental cofactors

Detectance and gene-environment interaction models

Detectance under different ascertainment schemes, including sampling on environmental exposure status

Prospective risk estimated from retrospective data despite secular trends and adaptive behavior.

Day 5

AM Conceptual issues and strategies: Causation, association, inference, and truth

Using biological theory to conceive the problem

Where does evolution come into the picture?

Assumptions about the trait

Is a study designed to fit assumptions or test them?

What are the hypotheses?

1. Some gene(s) affect the trait
2. Some gene(s) affect the trait *in this sample* (POPULATION?)
3. The same gene(s)/alleles affect the trait in *all* samples (POPULATIONS?)
4. The effect of a particular gene/allele is *e in this sample* (POPULATION?)
5. The effect of a particular gene/allele is *e in all samples* (POPULATION?)

What kind of information is legitimate to include as *prior* information?

What does it MEAN to include this information? Power calculations vs interpretation.

What is our actual prior information?

How is that information used in (a) designing and (b) interpreting our study?

What are the possible action criteria, after the study is done?

Replication

Statistical or deterministic causation?

What we know truth is vs what we know truth is NOT.

Simulation of truly complex models and datasets

Joint analysis of linkage and LD on datasets of arbitrary structure.

Effects of ascertainment scheme on detectance and power

Relative power as generalizable

Absolute power a function of assumptions

Power decreases rapidly with increasing complexity, even compared with power under exactly correct marginal models

Analysis with wrong models is more powerful than analysis under the true marginal model in the presence of ascertainment bias, which is always present in good genetic studies

PM Wrap-up: Conceptual issues and strategies: Interpretation and inference

Overview of things

More complicated mapping strategies

Evolution of thought from families to association to HapMapping to admixture to beyond

Meta-analysis, etc.

Evolution of scale and method over biology.
Microarrays and expression profiling and what that promises and has found.
Philosophical recapitulation and discussions.