Chapter 1

An Introduction to Gene Hunting

In human genetics, gene hunting is the process of finding genes that contribute to individual differences in phenotypes. The phenotypes may be diseases and disorders or part of normal variation. This module overviews the strategies and the associated technologies used for gene hunting. This introduction to gene hunting provides an overall perspective to the operation and introduces a number of concepts and terms. Many of these terms will be fully defined later in the relevant sections that describe a technique.

The process of gene hunting begins with a well-defined phenotype. The term "well-defined" is relative to what is already known about a phenotype. For example, many of those involved in the research and treatment of psychopathology suspect that our current view of schizophrenia really encompasses a number of different—but perhaps related—disorders. We simply have not been clever enough to distinguish one disorder from another based on phenotypic symptoms and signs. The fact that schizophrenia is still somewhat "ill-defined" in this way does not dissuade research from gene hunting for schizophrenia. *Au contraire*, many of these researchers hope that the discovery of genes for schizophrenia may be a Rosetta stone that can assist in solving the problem of heterogeneity.

Many human phenotypes relevant to social science are continuous or quantitative traits. A continuous trait is a phenotype that everyone possesses but we have different amounts of it. Height is a classic example. It is not that some people have height and others lack it. We all have height but we are not all equally tall. Gene hunting can be done—in fact, is currently being done—for continuous traits. A gene contributing to a continuous or quantitative trait has become known as a *quantitative trait locus* or QTL.

In the following, the phenotype is spoken of as if it were a disorder. This is purely a convention to avoid tortuous English such as "the phenotype of research interest" "a disorder or a continuous phenotype" and "relatives with high scores from those with low scores." Please recognize that the phenotype can be anything—a disorder, a dichotomous trait, or a quantitative trait.

1.1 An Historical Perspective

Before the 1980s, gene hunting was a laborious and expensive enterprise. The major technological problem was the relative paucity of polymorphic loci with known chromosomal locations. Most of the known polymorphisms were based on gene products and not the DNA itself. That is, the lab tests for genotyping distinguished one protein from another and not one DNA strand from another. Often, the results of testing proteins and enzymes could be equivalent with respect to the underlying genotype. For example, a person with blood group A could have either genotype AA or AO, and at that time there was no lab test that could distinguish between the two.

It would take an historian of genetics to count the number of known polymorphic loci with known chromosomal locations before 1980, but a rough guess would put them at no more than a few hundred. In contrast, the mathematical and statistical background and even some computer algorithms for gene hunting were remarkable advanced. The major technological breakthrough that permitted a "genetics revolution" occurred in the around 1980 and consisted of the ability to *detect polymorphisms directly in the DNA*.

To place the matter in perspective, George Huntington described the eponymous disorder in 1872 and emphasized its hereditary nature. Shortly after the rediscovery of Mendel, Huntington's disease was recognized as a dominant disorder. Still, attempts to find the chromosome that contained the Huntington's disease gene (let alone the gene itself) were completely fruitless until the new technology evolved. About two years later, in 1983, researchers localized the gene to somewhere on the short arm of chromosome 4. Ten years from then, they found the gene.

This pattern was repeated time after time in the 1980s and 1990 for phenylketonuria, cystic fibrosis, and a host of other Mendelian disorders recalcitrant to early research attempts. It was this revolution in DNA polymorphisms that made discovery of many Mendelian forms of DCGs. The nature and classification of these polymorphisms is discussed later in Section X.X and the techniques used to detect them are discussed in Section X.X.

1.2 Linkage Versus Association Versus Who Cares

The revolution in localizing genes for Mendelian disorders and traits began with a technique called *linkage* but used another technique called *association* to find the actual gene. Current technology blurs the difference between linkage and association, but it is useful to treat them as disparate methods for didactic purposes. Historically, linkage was u*sed to find the approximate location of a gene.* The linkage strategy began with a large number of *marker genes* or *markers.* These are polymorphic loci with known chromosomal locations. Although they are termed genes, the vast majority of them are not in protein-coding regions. Instead, they act as signposts or map co-ordinates in the linkage enterprise.

The second step in linkage was to genotype members within families that have at least one person with a disorder. The final step is to test whether the inheritance pattern of a marker gene could predict affected and unaffected members within the families. If the marker genotypes were totally random with respect to disease status, then the gene for the disorder is not located close to the marker. On the other hand, if knowledge of the marker could predict who was and who was not affected within a pedigree, then the gene for the disorder must be somewhere close to the marker. Think of linkage as getting one into the right church but not necessarily the right pew.

Confronted with budgets in the billions and trillions of dollars, people can lose sight of the meaning of the size of the human genome–3.2 billion nucleotides. Imagine that the population of the United States increases something close to 10 fold, giving 3.2 billion people. The search for a gene akin to sicle cell anemia–where the problem is in one nucleotide–is analogous to searching through these 3.2 billion people to find a single person. Given the vast size of the human genome, the fact that a gene like the one for Huntington's could be isolated to the short arm of chromosome 4 makes the search for the actual gene much easier. The biological principles that make linkage analysis possible are discussed in Section X.X and the process of linkage analysis is treated in Section X.X.

Traditionally, the association design begins with a *candidate gene* and then tests whether those affected with a disorder have different allele frequencies than controls at that locus. After Huntington's was localized to chromosome 4, association strategies were used were used in that region until a protein-coding locus was identified that perfectly predicted who had the disorder. This is equivalent to trying the various pews in the Church of Eternal Linkage to find the right one. Association designs are discussed in Section X.X.

As the number of polymorphic loci expanded from the thousands to today's millions, linkage and association merged. The major reason for the merger is a phenomenon called *linkage disequilibrium* that will be fully discussed in Section X.X. Linkage disequilibrium is the nonrandom association of polymorphisms across short stretches of DNA, "short" meaning up to a hundred thousands base pairs. For example, suppose that we number the nucleotides in a fictitious section of DNA from 1 through 1,689. Suppose that there is a polymorphism in position 286 such that some DNA strands have an adenine in that position (the A allele) while others have a cytosine (C allele) in that position. Assume a second polymorphism at position 1,427, this time involving the nucleotides G and T.

There are four logical *haplotypes* across this whole section. (A haplotype is the concatenation of a series of alleles across a short section of DNA. The concept is elaborated in Section X.X). A DNA strand can have A at the position 286 and G at the position 1,427 giving the AG haplotype. The other haplotypes are AT, CG, and CT. If the AG haplotype (or, for that matter, any of the other three) occurs more or less often than predicted by chance, then the two polymorphisms are said to be in linkage disequilibrium. If the frequency of the haplotypes is consistent with chance expectations, then the two are in linkage equilibrium. Disequilibrium is the norm and across sections as short as the one in this example, the disequilibrium could even be complete. That is, there is a perfect (or very high) correlation between the allele at position 286 and the one at position 1,427. In common sense terms, if I know the genotype at position 286, I can predict the genotype at position 1,427 with a high degree of accuracy.

When there is strong disequilibrium, it is not necessary to genotype all polymorphisms. Just genotype a few and predict the others. This is the rationale behind the *genome-wide association study* with the acronyms GWA or GWAS. GWAS became possible because of technological breaktroughs that allowed tens of thousand and then hundreds of thousands of polymorphisms to be detected at relatively low cost. This technology is discussed in Section X.X.

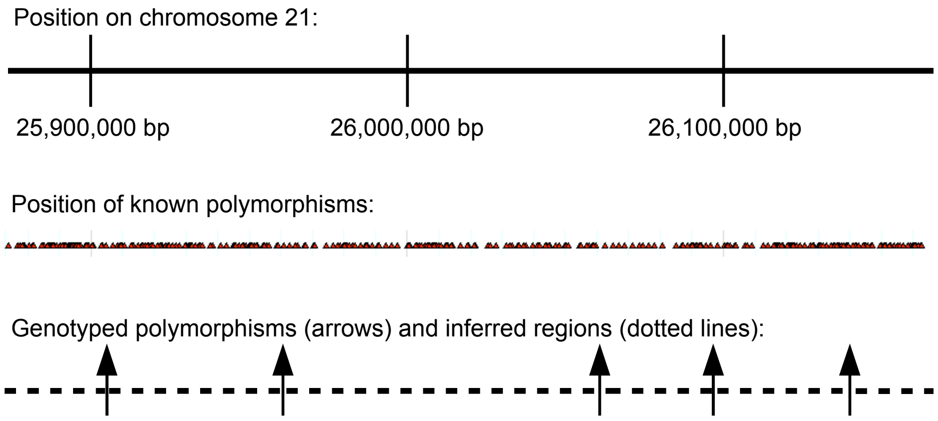
Like the association study, the GWAS starts with candidate genes, but the vast majority of them have no theoretical connection with the disorder. Instead, they are "candidates" because they give good information about the polymorphisms in linkage disequilibrium around them. Hence, if you know the genotype at a candidate, you have a pretty good guess about the genotypes surrounding that polymorphism. Like the linkage design, finding that a candidate polymorphism is associated with schizophrenia does not mean that the specific polymorphism is directly causal for the disorder. The actual polymorphism that contributes to schizophrenia could be at another place in the region. Again, just like linkage, the GWAS narrows down the list of suspects but does not arrest the actual culprits.

The GWAS, however, has decided advantages over traditional linkage. In classic linkage, the search for a gene might be narrowed from 3.2 billion to several tens of millions of base pairs. GWAS dramatically lowers the search to tens or hundreds of thousands of base pairs. Moreover, it does so in an efficient manner.

Consider the APP (amyloid precursor protein) gene that, if mutated in certain sections, can lead to Alzheimer's disease. The length of this DNA segment is roughly 300,000 base pairs. At the time of this writing, well over 100 polymorphisms have been catalogued within this region. Should researchers genotype each and every one of these polymorphisms? If we were to do this for every protein-coding region, the cost of genotyping would be prohibitive. Instead, the approach in current genetics is to genotype certain sections of the APP gene and infers the intermediate areas. A schematic of this process is provided in Figure X.X.

Here, the positions on chromosome 21 and the position of known polymorphisms give real data based on the Haplotype Mapping Project results as of October, 2009. The arrows indicating which polymorphisms to genotype and which polymorphisms can be inferred are hypothetical. Still, the figure illuminates the logic of the process–genotype few loci to save money and then infer genotypes in between. Technically, the inferred genotypes (or haplotypes) are called *imputed* genotypes (or haplotypes). With current technology, this approach–judiciously applied–can lead to over 90% accuracy in imputation.

Figure 1.1: Schematic of genotyping polymorphisms in the APP gene for a GWAS.


Position on chromosome 21:
25,900,000 bp
26,000,000 bp
26,100,000 bp
Position of known polymorphisms:
Genotyped polymorphisms (arrows) and inferred regions (dotted lines):

1.3 Gene Expression

Gene hunting usually involves the search for DNA spelling variations that are associated with the disorder. A gene could still be involved in the pathogenesis of a disorder, but not because of its polymorphisms. For example, suppose the contributing factor of a gene to schizophrenia is its over or under expression. Everyone has the same protein. It is just that some people have too much (or too little) of it. Furthermore, the difference in expression could come about for environmental reasons. Part of the technology used in GWAS can be adapted for the purpose of studying differential gene expression. This is discussed in Section X.X.

1.4 The Human Genome Project and its Contribution to Gene Hunting

In 1985, Robert Sinsheimer, a renowned biologist who was then Chancellor of the University of California at Santa Cruz, organized a meeting to discuss a shocking—some would say "irrational"—idea. He and the participants mulled over the possibility of biology embarking on a "big science" venture akin to building a giant particle accelerator in physics or manned space exploration. That project was sequencing the human genome or finding the nuclotide sequence of all human chromosomes..

Despite skepticism in the scientific community and the usual fits and starts surrounding a novel idea, there was sufficient enthusiasm among scientists that they began planning how to accomplish the goal. In 1990, both the Department of Energy (DOE) and the National Institutes of Health (NIH) presented their plan to the U.S. Congress and received funding for a 15-year project. The Human Genome Project (HGP) had begun.

Wisely, the initial effort eschewed the headline-grabbing science that might have occurred had the project jumped straight into sequencing. Instead they favored the less glamorous approach of developing the technology that could do the sequencing first and then leave the actual sequencing for the later years of the project.

Despite that approach, there was considerable progress in genetic sequencing. In 1995, the very first genome of an organism other than a virus was sequenced, two years later the genome of the molecular geneticist's "work horse" organism *E. coli*'s was sequenced, followed within a year by that of the round worm *C. elegans* (another work horse). In 1999, the first human chromosome (chromosome 22) was sequenced. International teams, particularly through the support of the UK's Wellcome Foundation, had also joined in the sequencing effort.

Toward the end of 2000, U.S. President Clinton, Francis Collins (head of the NIH part of the HGP), Ari Patrinos (chief of the DOE effort), and Craig Venter (CEO of a private company involved in gene sequencing) announced successful completion of a draft version of the sequence. In February of 2001, that sequence was jointly published in *Science* and in *Nature*. In 2003, the goals of the project were completed—two years early and under budget.

Contrary to popular belief, scientists did not discover ***the*** human genome sequence. Indeed, there is no such thing as ***the*** human genome sequence. There are billions of unique human genome sequences. What happened was that scientists in many different—but collaborating—laboratories working on individual human chromosomes, developed a ***consensus*** or ***reference*** sequence. The consensus sequence is not necessarily a "good" sequence—or, for that matter, a "bad" one—and it certainly is not a representative sequence. Instead, it is the first sequence developed, and it serves as a primitive map that everyone can agree upon in order to chart the unknown areas of the human genome.

Work on sequencing human DNA did not stop—and in many ways accelerated—with development of the consensus sequence. The technology developed in the HGP was adapted for other uses, most notably detecting and cataloguing human polymorphisms. A series of other "genome projects" developed, notably the Human Epigenome Project (intended to identify positions of methylation in the human genome), the Human Genome Structural Variation Project (find CNVs), the SNP consortium (map human single nucleotide polymorphisms–see Section X.X), as well as specialized projects aimed as specific diseases and as well as cell types (the *transcriptome* or a catalogue of the amounts of different types of mRNA produced in various cell types). Much of the technology used today for gene hunting can trace its origin to the HGP, especially its decision to invest heavily in technological advancements.

1.5 The New Holy Grail

Sequencing the human genome was sold to the U.S. Congress as well as the public as the "holy grail" of biology. There certainly is an element of truth to that, but the stark fact is that the establishment of a consensus sequence was not an end in itself. Instead, it was a primitive map of virgin territory. Meaningful results come from exploring that territory, not looking at the map. In other words, the HGP developed a very important *tool* for further research. It did not establish an end in itself so that geneticists could say, "Wow! We found it! Now we can retire and relax."

The cost of sequencing individual genes on a large number of individuals is not prohibitive, but the cost of sequencing a whole human genome–your or mine–is still too great to make it practical. The consensus sequence cost about $300 million (less the $3 billion to develop the technology to do so). Today, the cost has droped to the many thousands of dollars. Abetted by private prize money (and later, grants from the NIH), the current ambition is to sequence the entire genome of a person for under $1,000. This challenge has spurred efforts in both the public and private sectors into "next-generation DNA sequencing" and progress has been so brisk that it is impossible to capture in a text like this. Currently five different technologies are vying for supremacy (see Shendure & Ji, 2008, p. 1140). Which ones will emerge as superior (for which purposes)? That is akin to guessing the winner of the Kentucky Derby five years from now.

Today's "holy grail" is the $1k genome. We should all applaud this effort, But we should never overlook the fact that this is really the development of another *tool*, albeit an exceptionally important tool.