A tutorial in genetic epidemiology and some considerations in statistical modeling

ERICK SUÁREZ, Act., Ph D*; CARLOS A. SARIOL, MD, MS†; ANA BURGUETE, MD, Ph D‡; GEOFFREY MCLACHLAN, Ph D, DSc.**

A new door has been opened to health professionals since the completion of the map of the human genome was announced in 2003, coinciding with the 50th anniversary of the discovery of the DNA helical structure by Watson and Crick in 1953. The continuous updating of the technology has enabled scientists to simultaneously analyze thousands of variables for genome analysis. These advances have created new opportunities to locate genes, to assess the gene-gene relationship, to measure the gene-environment interaction, to describe gene products, and to evaluate the gene-disease relationship. In epidemiology, new strategies have been developed to determine cause-effect relationship in case-control studies and cohort studies. With the information provided by the Human Genome Project, new epidemiological designs and new statistical methodology have been developed. The addition of molecular biology to traditional epidemiological approaches has given birth to a new discipline known as genetic epidemiology. The objective of this paper is to provide an introduction to concepts needed for assessing the association between genes and specific diseases in population based studies. Firstly, a description of the genetic concepts is presented as a framework for the epidemiological designs and the statistical procedures that have been utilized in genetic epidemiology. Then, a description of the different designs in genetic epidemiology is presented with the most recent publications. Finally, some considerations in the statistical analysis for genetic epidemiology are discussed.

Key words: Genetic Epidemiology, Genetic linkage, SNP, Mutations, Genetic Model, Association studies, Admixture, Penetrance, Logistic regression, Interactions.

I) The molecular biology of the gene

1.1 Intruduction

The human body is made up of millions of cells, where each cell contains a complete copy of a person's genetic plan or blueprint. This genetic plan is packaged in the cells in the form of chromosomes that are made up of strings of genes. The chromosomes, and therefore, the genes, are made up of DNA. The chromosomes and genes are contained in the nucleus of every cell, except for red cells, which have no nucleus and, thus, no chromosomes. A small number of genes are also contained in tiny packages in the cell called mitochondria which are the energy centers of the cell. The entire DNA in the human cell makes up what is called the human genome.

A gene is a specific segment of a DNA molecule that contains all the coding information necessary to instruct a cell to synthesize a specific product, such as an RNA molecule or a protein (enzymes, hormones, and antibodies) needed for the structural and metabolic functions of the cells, and thus of the entire organism. Each gene provides a blueprint for the synthesis (via RNA) of enzymes and other proteins and specifies when these substances are to be made [Watson et al., 2004].

The word gene was derived from Hugo De Vries's term pangen, itself a derivative of the word pangenesis which Darwin (1809-1882) had coined in 1868. The word pangenesis is made from the Greek words pan (a prefix meaning "whole", "encompassing") and genesis ("birth") or genos ("origin").

In 1953, James Watson (1928-) and Francis Crick (1916-2004), using X-ray diffraction analysis of crystallized DNA, discovered that native DNA consists of two long chains (strands) forming a double-stranded helix. The coiled polynucleotide chains of DNA are held together by

*Department of Biostatistics and Epidemiology, School of Public Health, University of Puerto Rico-Medical Science Campus, San Juan, Puerto Rico, †Caribbean Primate Research Center, Unit of Comparative Medicine, Internal Medicine Department, Department of Microbiology, University of Puerto Rico-Medical Science Campus, San Juan, Puerto Rico, ‡Instituto Nacional de Salud Pública de México, **Department of Mathematics, University of Queensland, Brisbane Australia.

Address correspondence to: Erick Suárez, Act., PhD, Department of Biostatistics and Epidemiology, School of Public Health, University of Puerto Rico-Medical Science Campus, P.O. Box 365067, San Juan, Puerto Rico, 00936-5067, Email: esuarez@rcm.upr.edu

hydrogen bonds between the bases of the opposite strands. The bases occur as a specific set of complementary pairs (Figure 1). Adenine (A) pairs only with thymine (T), and guanine (G) pairs only with cytosine (C). The number of complementary base pairs is often used to describe the length of a double stranded DNA molecule. For DNA molecules with thousands or millions of base pairs, the designation of kilobase pairs or megabase (Mb) are used, respectively. For example, the human chromosomes X and Y have approximately 155 Mb and 57 Mb, respectively (www.ensembl.org).

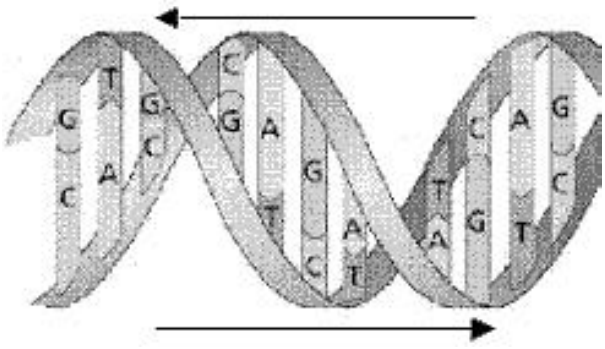


Figure 1. DNA Helix

The AT and GC base pairs lie within the interior of the molecule and the linked phosphorus and deoxyribose components form the backbone of each strand. The terms used to specify the directions of the strand are 3' and 5'. The two strands of a duplex DNA molecule run in opposite direction (antiparallel chains). One chain is oriented in a 3' to 5' direction and the other in a 5' to 3' direction. Because of base pairing requirements, when one strand of DNA has the sequence of bases 5'-TAGGCAT-3' the complementary strand must be 3'ATCCGTA-5'. In this case, the double-stranded form would be:

5'-TAGGCAT-3'
3'-ATCCGTA-5'

Chemically, each gene consists of a specific sequence of nucleotides. Each nucleotide is composed of three subunits: a nitrogen-containing compound, a sugar, and phosphoric acid. The genes have coding (exons) and not coding (introns) segments in the coding process. Genes may vary in their precise makeup from person to person. Also, different genes are "active" in different cell types, tissues, and organs; but not all the genes in the cell are "active" in every cell. However, despite the wide range of phenotypes observed in the human race, our DNA has little variability. Approximately, 99% of the nucleotides in DNA are the same in all humans. Those DNA locations, or loci, that vary from person to person are said to be polymorphic [Pasternak, 2005].

The process of DNA synthesis is called replication. As predicted by the Watson-Crick model of DNA, each strand of an existing DNA molecule acts as a template for the production of a new strand, and the sequence of nucleotides of a synthesized (growing) strand is determined by base complementary [Watson et al., 2004].

Proteins are required for the structure, function, and regulation of the body's cells, tissues, and organs. Proteins are essential components of muscles, skin, bones, and the body as a whole. A protein chain consists of a specific sequence of units called amino acids. All amino acids have the same basic chemical organization. There is a central carbon atom (a-carbon) with a hydrogen (H), carboxyl group ($C00^-$), amino group (NH_3^+), and an R group attached to it. An R group can be any one of 20 different side chain (groups) that make up the 20 different amino acids found in proteins. When R, for example, is a methyl group (CH_3), then the amino acid is Alanine.

The vast majority of genes encode information for the production of protein chains. Proteins are essential polymers (macromolecular) involved in almost all biological functions. They form structures within the cell such as the protein called keratin, from which hair is made; others are called enzymes which help to produce chemical reactions, such as digesting food. They catalyze chemical reactions; transport molecules within cells; escort molecules between cells; control membrane permeability; give support to cells, organs, and body structures; cause movements; provide protection against infectious agents and toxins; and regulate the differential production of other gene products. Proteins range in length from approximately 40 to more than 1000 amino acid residues. A protein folds into a particular shape (configuration) depending on the location of specific amino acid residues and the overall amino acid composition. In addition, many functional proteins consist of two or more polypeptide chains (a linear series of amino acids linked by peptide bonds). In some cases, a set of multiples of the same polypeptide chain is required for an active protein molecule (homomeric protein). In other instances, a set of different protein chains (subunits) assembles to form a functional protein (heteromeric protein).

The biological decoding of genetic information is carried out through intermediary RNA molecules synthesized from a segment of the DNA. Instead of thymine, the base uracil (U) is found in RNA. Uracil pairs with adenine in an RNA molecule. Most of the RNA molecules are single stranded, although often, within a single chain, segments of nucleotides are complementary to each other and form double-stranded regions. The production of RNA from

DNA is called transcription. As transcription proceeds, the newly synthesized RNA is released from DNA, and the DNA helix is reconstituted. The functional transcription product of a structural gene is an mRNA. Hundreds of different mRNAs can be in one cell. By contrast, there are four type of ribosomal RNA (rRNA); three of the rRNAs combine with a set of proteins to form a ribonucleic protein complex called the large ribosomal subunit, the other rRNA combines with another set of proteins to form a small ribosomal subunit. In the cytoplasm of the cell, one large and one small ribosomal subunit combine to form a ribosome. An active cell can have thousands of ribosomes [Watson et al., 2004].

There are approximately 50 different types of transfer RNA (tRNA) molecules in a cell that are actively synthesizing protein. The tRNAs range in length from about 75 to 93 nucleotides. There is at least one tRNA for each of the 20 amino acids found in proteins. An amino acid is linked enzymatically by its carbonxyl end to the 3'-end of a specific tRNA; after the binding of a particular amino acid to its tRNA, the tRNA is said to be charged. The tRNA molecule has three unpaired nucleotides, which together are called the anticodon sequence; this sequence plays an important role in the formation of the linear array of amino acids that constitute a protein. A codon is a nucleotide triplet of bases recognized by anticodons on transfer RNA and hence specifying an amino acid to be incorporated into a protein sequence. Each amino acid has more than one codon. The stop codon determines the end of a polypeptide. The process of decoding the information content of an mRNA into linear sequence of linked amino acids is called translation.

The codon in the mRNA that immediately follows AUG dictates the anticodon sequence and, therefore, which charged tRNA will bind to the ribosome complex. If the second triplet of nucleotides in the mRNA is CUG, then the charged tRNA will bind with the anticodon sequence GAC. This charged tRNA carries the amino acid leucine. Once in place, a peptide bond is formed between the carboxyl group of the methionine and amino group of the leucine. If the third codon is UUU, then the charged tRNA with an AAA anticodon will bind. In this case, the tRNA with an AAA anticodon carries the amino acid phenylalanine. Once in place, the linkage between the carboxyl group of leucine and its tRNA is broken; as a consequence a peptide bond is formed between the carboxyl group of the leucine and the amino acid of the phenylalanine.

The succession of operations including binding of a charged tRNA by means of anticodon-codon pairing, peptide bond formation, ejection of an uncharged tRNA, and translocation, continues until all the amino acids encoded by the mRNA are linked together. A coding region, exon only, of 1 kb gives rise to a protein with approximately 333 amino acids. The total coding region of a gene can be from .5 to about 15 kb in length. The complete genetic code consists of 64 codons. Three codons (UGA, UAG, UAA) are reserved for stop and one (AUG) for initiation. There is one codon (UGG) for the amino acid trytophan. For the rest of the amino acids found in proteins, there are least two to six codons. For example, leucine has six codons: UUA, UUG, CUU, CUC, CUA, and CUG. These characteristics define the degeneracy (changes of nucleotide in the third or second position of a codon) and redundancy (more than one codon per single amino acid) of genetic code; however, a codon specifies only one amino acid (no ambiguity) [Watson et al, 2004].

Once the sequence of amino acids that make up a particular protein is assembled, the protein dissociates from the ribosome and folds in to a specific three-dimensional form. The function of a protein ultimately depends on its amino acid sequence and its three-dimensional structure. Currently, functions have been assigned to only a small proportion of the genes in even the best understood of model organisms. In order to assign function to the remaining genes, it is helpful to examine the expression patterns of these genes in various tissues.

Microarray technology developed over the past several years now allows the measurement of mRNA levels for tens of thousands of genes simultaneously. The applications of microarrays for the study of neurological diseases, like multiple sclerosis, Alzheimer's disease or neuromuscular diseases are promising, both for generating new pathophysiological hypotheses and for enabling new molecular classifications [Ducray et al., 2007]. Microarray data analysis on cancer research has opened new avenues for diagnosis and therapeutic interventions [Cowell et al., 2007]. Our capabilities for diagnosis and understanding of infectious diseases have also been enhanced by using microarrays [Palacios et al., 2007 and Sariol et al., 2007].

1.2) Nucleotide Sequence Alteration: Mutation

The replication of DNA is not a perfect process; errors that affect one or more base pairs can occur. In addition, external agents (radiation, radioactive compounds, ultraviolet, chemicals…) can permanently alter the sequence of nucleotides of a DNA molecule. A change in genetic material is called mutation. An agent that induces a mutation is a mutagen. In the absence of any evidence of a mutagenic effect, a naturally occurring mutation is considered spontaneous. Mutation can occur anywhere in the total DNA (single base pair or large region of a chromosome) of a organism. In humans, approximately 95% of DNA does not code for any gene products. As a

result, many mutations have no effect on the phenotype because they are located in regions of the genome that have no impact on cellular functions.

The bases guanine and adenine have the same fundamental chemical structures and fall into the class of compounds called purines. Similarly, the bases cytosine and thymine are chemically related to each other and are part of the compounds called pyrimidines. Any substitution of a purine with a different purine (G↔A) or a pyrimide with a different pyrimide (T↔C) in a DNA molecule is a transition mutation. A transversion mutation is any substitution of a purine by a pyrimidine or vice-versa (G↔T, A↔T, G↔C, A↔C). A base substitution within the coding region of a structural gene can change an mRNA codon and lead to the insertion of a different amino acid in the protein. The consequence of mutation of a DNA codon depends on which nucleotide pair is changed, the nature of the substitution, the specificity of the new codon, and the relative location of the mutated codon, among other factors. In general, DNA codon mutations are classified as silent, neutral, missense, or non-sense. A silent mutation occurs when there is a change in a DNA codon, but the amino acid that is inserted into a protein is not changed; for example, when the DNA codon UUU is altered by the mutation UUC, both produce the same protein, phenylalanine. A neutral mutation represents a nucleotide change at the DNA level that alters a codon, so that another amino acid is incorporated into the protein with no apparent loss of function; for example, when the DNA codon CUU (leucine) is altered by GUC (valine), both proteins have similar physicochemical properties. A base pair substitution producing a codon that specifies another amino acid is a missense mutation; the severity of this mutation depends on the nature of the substituted amino acid and whether the original amino acid plays an essential role in the function of the protein. A non-sense mutation occurs when a nucleotide substitution changes a codon that specifies an amino acid into one that is a stop codon; for example, when the DNA codon UGG is substituted by UAG (stop codon). The presence of a stop codon within a mRNA causes an incomplete or truncated protein to be produced [Watson et al., 2004].

When a base pair is either inserted into or deleted from the coding region of a structural gene, the sequence of codons can be changed, so that the new codons are translated into a completely novel sequence of amino acids bearing absolutely no resemblance to the original protein. These types of changes are called frameshift mutations because the reading frame of the normal array of a codon is shifted. A frameshift mutation usually has a devastating effect on the function of a protein, because of either protein truncation or the addition of an aberrant string of the amino acid. However frameshift is essential to guarantee the proper replication of some viruses like Hepatitis B virus and retroviruses including HIV. This mechanism allows the translation production of different essential proteins using the same coding genetic sequence [Knipe et al., 1996].

By strict definition, recessiveness and dominance are properties of the phenotype, although it is common to refer to genes as recessive or dominant. Most mutations have a recessive effect. In an homozygous recessive individual, insufficient functional gene product is produced, which, in turn, is responsible for an aberrant phenotype. Hereditary folate malabsorption, a rare autosomal recessive disorder that is caused by impaired intestinal folate absorption and impaired folate transport into the central nervous system, induces progressive neurological impairment. However, there are a number of human diseases that are the result of a single dominant allele. A mutation that alters the amount of a gene product (underproduction and overproduction), when a specific amount is needed for normal activity, can cause a dominant effect. For example, in metachondromatosis, the growth of bones is affected, leading to multiple enchondromas and osteochondromas mainly in tubular bones. On the other hand, underexpression of tumor suppression genes like p53 is heavily associated to head and neck cancer and lung cancer. In addition, dominant disorders occur when either a toxic gene product or a novel protein with an unusual mode of action is produced.

Other genetic alterations of DNA are the epigenetics modifiers. The term ‘epigenetics’ defines all meiotically and mitotically heritable changes in gene expression that are coded in the DNA sequence itself. Three systems, including DNA methylation, RNA-associated silencing, and histone modification, are used to initiate and sustain epigenetic silencing. Disruption of one or other of these systems can lead to inappropriate expression or silencing of genes, resulting in ‘epigenetics diseases’ [Egger et al., 2004]. DNA methylation is an enzymatic addition of a methyl group to cytosine residues at the C-5 position and occurs at the CpG sequences (meaning a C nucleotide followed by a G nucleotide). Although isolated CpG’s are usually methylated, the human genome contains regions rich in CpG’s, known as CpG islands. In humans, there are about 45,000 CpG islands, mostly found at the 5' ends of genes. They are unmethylated, except for those on the inactive X chromosome and some associated with imprinted genes. Detection of regions of genomic sequences that are rich in the “CpG” pattern is important because such regions are resistant to methylation and tend to be associated with promoter regions of genes which are frequently expressed [Watson et al., 2004]. Methylation of these

genes in cancer or perhaps with aging can lead to their irreversible silencing [Thomas, 2004]. Histone acetylation may regulate the expression of several genes in prostate cancer; it has been reported that treatment of prostate cancer cells with Histone Acetyl-transferases and deacetylases (HDAC) inhibitors increased expression of specific genes, and thus inferred a role for histone acetylation in gene regulation [Cheng et al., 2005].

1.3) Human Chromosomes

Biological connectivity from one generation of humans to the next is maintained by the fusion of a sperm from a male parent with an unfertilized egg from a female parent to produce a fertilized egg (zygote). It has been estimated recently that a human being inherits 19,000 to 27,000 genes from each parent (www.geneomics.energy.gov); however, these numbers still are under debate, when the Human Genome Project was completed, it was estimated around 35,000 genes. The strands of DNA are organized into chromosomes. There are, under normal conditions, 23 chromosomes in the nucleus of the fertilized human eggs, formed by 46 thread-like structures (23 pairs). After the first embryonic cell division, each nucleus of the daughter cells has 23 pairs of chromosomes and so on in the consecutive divisions. Every nucleus cell has 23 pairs of chromosomes. It is estimated that the human body has 3,237,658,234 base pairs (www.ensembl.org).

The cell division cycle (mitotic cycle) ensures that after each cell division, each daughter cell receives the same number of chromosomes. A process called meiosis ensures that each gamete (sperm cell or unfertilized eggs) receives only one member of each pair of 23 chromosomes.

Cells that contain one copy of the genome, such as sperm or unfertilized egg cells, are said to be haploid (1N). Fertilized eggs and most body cells derived from them contain two copies of the genome and are said to be diploid (2N). A diploid cell contains 46 chromosomes: 22 homologous pairs of autosomes and a pair of fully homologous (XX) or partially homologous (XY) sex chromosomes. Chromosomes can be seen under the microscope only during cell division. By stimulating cells to divide and then treating them with chemicals that block the last stages of cell division, one can observe and photograph individual chromosome. The cell division consists of the following phases: G1 phase, S phase (DNA synthesis), G2 phase, and M phase or mitosis. During mitosis, the mitotic cells can be divided in the following stages:

i) Prophase - Chromosomes become visible as extended double structure
ii) Prometaphase - Chromosome pairs become thicker and shorter
iii) Metaphase - Nucleus is replaced by spindle; becomes aligned on the spindle midpoint or equator
iv) Anaphase - Chromosome pairs split and move toward the spindle poles
v) Telophase - Chromosomes reach the spindle poles

After these phases, cytokinesis begins, which means that the original cell is divided into two cells; each daughter has a complete set of chromosomes, one member of each pair derived from each parent [Rappaport, 2005].

A comparison of the lengths and overall morphology of the mitotic metaphase chromosome of human males and females shows that 22 pairs are held in common. The chromosomes, not all the same length, are autosomes. The 23rd pair includes the sex chromosome. In females, both sex chromosomes have the same length and are designated as X chromosomes. In males, the sex chromosomes consist of one X chromosome and a smaller chromosome designated as Y. Autosomal refers to any chromosome other than a sex chromosome.

Chromosomes are often depicted as they would appear during cell division: as two duplicate chromosomes, joined to each other at the middle. The point where the two copies or chromatids are joined is called centromere. The position of the centromere aids in the identification of each one of the 23 chromosomes. The centromere rarely is in the middle of the chromosome, as a result, the short arm is identified by p and the long arm by q.

The presence of an extra chromosome 21 (Trisomy 21) in a zygote can lead to Down syndrome; Trisomy 18 to Edwards syndrome, and Trisomy 13 to Patau syndrome.

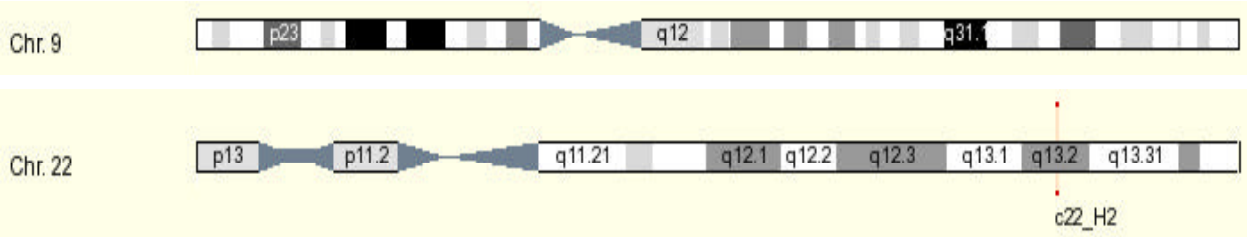
The presence of an extra sex chromosome has less biological impact than autosomal trisomy. The men with XXY constitution (47, XXY; Klinefelter syndrome) tend to be tall with long arms and large hands and feet; occasionally, they have decreased capabilities but are sexually competent. Females with Trisomy X (47, XXX) experience major learning problems. The only condition with just one sex chromosome, the X chromosome, is Turner syndrome (45, X). Women with Turner syndrome are short and infertile, have a thick neck, and in some cases, suffer from kidney and cardiovascular abnormalities. A single Y chromosome constitution (45, Y) has never been observed in a live birth [Rimoin et al., 2002].

Chromosomal structural changes occur when the DNA of a chromosome breaks and is rejoined to another broken piece of chromosome DNA forming an unusual rearrangement. Environmental agents, such as X-rays and chemicals, can induce chromosome breaks. These changes can occur between different (non-homologous) chromosomes. In some instances, parts of two non-homologous chromosomes are interchanged without any apparent loss of chromosomal material. The occurrence of

such mutual exchange is called reciprocal translocation (balanced translocation).

The ratio of the arms length (p/q), centromere index (p*100/(p+q)), and the length of each chromosome relative to the length of a haploid set were used initially to classify chromosomes. By convention, chromosome 1 is the longest autosomal chromosome, the next longest is chromosome 2, and so on. Genes can be mapped because they occupy a specific location (or locus) on a chromosome.

The chromosome band is part of a chromosome which is clearly distinguishable from its adjacent segments by appearing darker or lighter by one or more banding techniques. The bands may represent different DNA sequences (nucleotide composition) along a chromosome, localized structural features such as DNA loop formation within sectors of a chromosome, or the presence of specific proteins that bind to designated sections of a chromosome. Bands are denoted with Arabic numerals starting from the centromere and proceeding to the end of each chromosome arm. For example, the bands of chromosome 9 and 22 are (www.ensembl.org):



There are examples of disease-specific chromosome rearrangements. Many of them are found in somatic cells that have become cancerous. For example, the chromosomal rearrangement t(9,22)(q34,q11, know as Philadelphia chromosome) is present in about 90% and 95% of cases of chronic myeloid leukemia, a cancer that causes an overabundance of certain types of white cells called granular leukemia. t(9,22)(q34,q11) means a translocation between chromosome 9 at band q34 and chromosome 22 at band q11. This was the first chromosome abnormality found in any kind of malignancy.

A phenomenon in which the disease phenotype depends on which parent passed on the disease gene is called imprinting. For instance, both Prader-Willi and Angelman syndromes are inherited when the same part of chromosome 15 is missing. The Prader-Willis syndrome is a disorder characterized by short stature, obesity, and mild to moderate learning difficulties. The Angelman syndrome is characterized by gait disturbance, epilepsy, and sever learning difficulties. When the father's complement of 15 is missing, the child has Prader-Willis syndrome, but when the mother's complement of 15 is missing, the child has Angelman syndrome [Rimoin et al., 2002].

The location of a gene on a chromosome is called a locus (loci, plural). Variants of a single gene that occupy the same locus on the two homologous chromosomes are known as an allele. Differences in alleles may give rise to differences in traits or physical structure of an individual, for example, eye color. A gene's most common allele is called the wild type allele, and rare alleles are called mutants. The appearance of a trait is called phenotype and the genetic constitution is the genotype. An organism with different paternal and maternal alleles of a gene is a heterozygote. The term homozygote refers to a gene in which both maternal and paternal alleles are identical. If a gene behaves as a dominant in the heterozygous, then an organism with two dominant genes is said to be homozygous dominant. An organism with two of the same recessive genes is homozygous recessive. In humans, some diseases result from dominant genes and some from recessive genes [Watson et al., 2004]. The autosomal dominant polycystic kydney disease (ADPKD) is the most common inherited form of an autosomal dominant disease.

It is one of the most common hereditary diseases and the fourth leading cause of kidney failure. It seems to affect all races and both genders equally. Symptoms usually develop between the ages of 30 and 40, but they can begin earlier, even in childhood. About 90% of all ADPKD cases are autosomal dominant but they can be autosomal recessive [Beer et al., 2006].

1.4) Mendel's Laws of Inheritance

The first coherent description of the inheritance of genes was presented by Gregor Mendel in 1865, based on breeding experiments with pea plants, which were summarized in the following principles [Watson et al., 2004]:

- Segregation of alleles -Each person carries two copies of each gene, one inherited from each parent. Alleles are transmitted randomly and with equal probability (transmission probabilities). If we have two alleles for a single major gene, then the transmission probabilities will be ½.
- Independent assortment - The alleles of different

genes are transmitted independently. Today it is known that this does not apply when loci are located near each other on the same chromosome.

A third concept considered part of the Mendelian framework assumes that the expression of the genes is independent of which parent they come from. In recent decades, however, exception to these principles, such as imprinting and other forms of parent-of-origin effects have been recognized. Another assumption commonly made to compute probabilities on pedigree is random mating, that is, the probability that any two individuals will mate is independent of the genotype. This assumption is sometimes difficult to preserve due to the fact that individuals are more likely to mate within their ethnic groups [Thomas, 2004].

During the gamete formation, the allele of a gene pair is segregated from each other. The fusion of the gametes during fertilization is completely random. And, if there are enough gametes, all possible combinations (fertilization) will occur. The expected frequency of gametes combination (fertilized eggs) is the product of the individual frequency of the gametes that unite. For example, suppose that there are only two alleles (*r, R*) for a single major gene. If a parent is homozygous for either *r* or the *R* allele, then that is the only allele he or she can transmit; so, the transmission probability will be one. On the other hand, if the parent is heterozygous, then either the r or the R allele can be transmitted, both with ½ probability. So, the joint effect of two heterozygous parent's genotype on the offspring's full genotype (the combination of the two transmitted gametes) will be as follows:

Gametes production		Mother *Rr*	
		1/2*R*	1/2*r*
Father *Rr*	*1/2R*	*¼ (RR)*	*¼ (Rr)*
	1/2r	*¼ (Rr)*	*¼ (rr)*

The offspring's genotypes are RR, Rr, and rr, while the offspring's phenotypes are R and r. The probability of the offsprings' genotype follows the Hardy-Weinberg equilibrium:

$P\{RR\}=p^2$, $P\{Rr\}=2p(1-p)$, and $P\{rr\}=(1-p)^2$, where $p=1/2$.

The Hardy-Weinberg model describes and predicts genotype and allele frequencies in a non-evolving population, under the following basic assumptions: i) the population is large (i.e., there is no genetic drift); ii) there is no gene flow between populations, from migration or transfer of gametes; iii) mutations are negligible; iv) individuals are mating randomly; and v) natural selection is not operating on the population. Given these assumptions, a population's genotype and allele frequencies will remain unchanged over successive generations, and the population is said to be in Hardy-Weinberg equilibrium (HWE). The Hardy-Weinberg model can also be applied to the genotype frequency of a single gene. Then, the HWE is a simply prediction of population genotypic frequencies based on allele frequencies:

Alleles: frequencies	Genotypes: frequencies
A: p	*AA*: p^2
a :1-p	*Aa* : 2p(1-p)
	aa : $(1-p)^2$

So, the Hardy-Weinberg model consists of two equations: one that calculates allele frequencies and one that calculates genotype frequencies. Because this model is dealing with frequencies, in terms of probabilities, both equations must add up to 1. The equation for allele frequencies for a gene with two alleles is: p + (1-p) = 1. If we know the frequency of one allele (p) we can easily calculate the frequency of the other allele (1-p). This is the simplest case, but the equation can also be modified and used in cases with three or more alleles. In a diploid organism with alleles *A* and *a* at a given locus, there are three possible genotypes: *AA*, *Aa*, and *aa*. If we use p to represent the frequency of *A* and (1-p) or q to represent the frequency of *a*, then we can write the genotype frequencies as p^2 for *AA*, q^2 for *aa*, and 2(p)(q) for *Aa*. The equation for genotype frequencies is $p^2+ 2pq + q^2 = 1$. Testing for HWE is useful since there are several biological and methodological explanations for deviation from the expected HWE, such as:

(i) typing errors (e.g. missing heterozygotes), (ii) assortative mating (e.g. negative assortative may result in excess of heterozygotes), (iii) selection (e.g. heterozygote advantage may result in excess of heterozygotes), and (iv) population structure (e.g. two merged populations).

Under the HWE, the alleles frequencies can be used to compute the frequencies of phenotype. For example, the frequencies of the *ABO* blood group, which has three alleles *A,B*, and *O*, where the allele *O* is dominant over *A* and *B*, while *A* and *B* are co-dominant, four possible phenotypes can be defined: *O* (corresponding to genotype *OO*), *A* (genotypes *AA* and *AO*), *B* (genotypes *BB* and *BO*), and *AB* (genotype *AB*). Assuming HWE, the population frequency of the four phenotypes is:

$$\Pr(A) = \Pr\{\text{genotype}=AA \text{ ó } AO \text{ ó } OA\} = q_A^2 + 2q_Aq_O$$

$$\Pr(B) = \Pr\{\text{genotype}=BB \text{ ó } BO \text{ ó } OB\} = q_B^2 + 2q_Bq_O$$

$$\Pr(O) = \Pr\{\text{genotype}=OO\} = q_O^2$$

$$\Pr(AB) = \Pr\{\text{genotype}= AB \text{ ó } BA\} = 2q_Aq_B$$

where q_A, q_B, and q_O are the allele frequencies of *A*, *B*, and *O* respectively.

Based on this example, the offspring of *AxB* mating would produce the following phenotype probabilities:

Genotype		Offspring Phenotype Probability			
Father	Mother	*A*	*B*	*O*	*AB*
AA	***BB***	0	0	0	1 (*AB*)
AA	***BO***	½ (*AO*)	0	0	½ (*AB*)
AO	***BB***	0	½ (*BO*)	0	½ (*AB*)
AO	***BO***	¼ (*AO*)	¼ (*BO*)	¼ (*OO*)	¼ (*AB*)

As a consequence, the joint population probabilities for each of the genotype combinations of the *AxB* mating would be computed using Bayes's theorem, as follows:

Genotype		
Father	Mother	Joint probabilities given *AxB*
AA	***BB***	$\Pr\{AA,BB\|AxB\}=P\{AA,BB\}/P\{AxB\}= = q^2_A * q^2_B / K$
AA	***BO***	$\Pr\{AA,BO\|AxB\}=P\{AA,BO\}/P\{AxB\}= = 2*q^2_a * q_B * q_O / K$
AO	***BB***	$\Pr\{AO,BB\|AxB\}=P\{AO,BB\}/P\{AxB\}= = 2* q^2_B * q_A * q_O / K$
AO	***BO***	$\Pr\{AO,BO\|AxB\}=P\{AO,BO\}/P\{AxB\}= = 4* q_A * q_B * q^2_O / K$

where $K=\Pr\{AxB\}=\Pr\{A\}*\Pr\{B\}=(q^2_A + 2q_A q_O)(q^2_B + 2q_B q_O)$

1.5) Genetic Linkage

One of the main objectives of genetic studies is to determine the linear order of genes along a chromosome. It is expected that if the genes of a chromosome always remained together (completely linked) then they would be passed on to the sex cells as an intact block, with no new genetic combinations formed during the meiotic process. When two consecutive markers derive from different parental chromosomes, the event is called a recombination. Genetic linkage is symbolized with either a single straight line or a single forward slash separating the genes that reside on different members of a pair of homologous chromosomes, for example,

$$\frac{AB}{ab}, \frac{Ab}{aB}, \ldots\ldots \text{ AB/ab, Ab/bB,}\ldots.$$

The notation AB/ab signifies that the dominant alleles (*A*,*B*) of two different gene loci are on one chromosome and the recessive alleles (*a*,*b*) are on the homologous chromosome. With complete linkage, a double heterozygote (*AB/ab*) would produce only two types of gametes:

Parents	Gametes production	Probability
AB/ab	*AB*	**½**
	ab	**½**

Crossing *AB/Ab* with ab/ab would yield only two kinds of genotypes (AB/ab and ab/ab) among the offsprings, with no new genetic combinations. As mentioned above for ADPKD, some alleles have a recessive effect, when two copies of the same allele are required to produce certain characteristic (trait). Other alleles have a dominant effect, where a single copy of such an allele in the presence of a recessive allele is sufficient to create the same biological effects as two copies of the dominant allele. A cross with complete dominance between a double heterozygote (SsTt) and a doubly homozygous recessive (sstt) produce four phenotypic classes representing four genotypes among the offsprings:

		Mother *ss/tt*
	Gametes production	*St*
Father *Ss/Tt*	*ST*	*STst*
	St	Stst
	sT	sTst
	St	*Stst*

In this case, there are two new genetic combinations (*Sstt*,*ssTt*). If the frequency of these new genetic combinations is less than 50% (not all possibilities have the same probability), then the two gene loci are not assorting independently, which means they are genetically linked on the same chromosome; linkage can occur only if two loci are on the same chromosome. The proportion of new genetic combinations among the progeny of a cross is the consequence of the frequency of reciprocal physical

exchanges between the two gene sites on no sister chromatics during meiosis. Chromatic is one of the two chromosome strands of a duplicated chromosome [Pasternak, 2005].

1.6) Genetic map

A genetic map is essential for determining the chromosome locations of disease-carrying genes and a physical map is required for isolating them. A physical map can be constructed for human chromosomes. A genetic map (linkage map, meiotic map) shows the order of sites derived from meiotic recombination frequencies. High frequencies of two or more alleles at one locus increase the likelihood that parents will have different genotype, making a linkage analysis feasible. It is estimated that there are one million DNA base pairs in 1 centiMorgan (unit of recombination, or 1% probability of crossover). Overall, the human female genetic map is 1.5 times longer than the male genetic maps.

The term "frequency of an allele" denotes the proportion of a particular allele to the total number of alleles of a locus on a particular population. For example, for a locus with two alleles (*A1,A2*) in a population of 13,000 persons, with 3800 individuals who are *A1A1*, 6400 who are *A1A2*, and 2800 who are *A2A2*, the frequency of allele *A1* would be:

$$\frac{\#\text{ of persons with A1}}{\#\text{ of alleles}} = \frac{3800*2+6400}{2*13000} = 0.54$$

The frequency of allele *A2* would be:

$$\frac{\#\text{ of persons with A2}}{\#\text{ of alleles}} = \frac{2800*2+6400}{2*13000} = 1-0.54 = 0.46$$

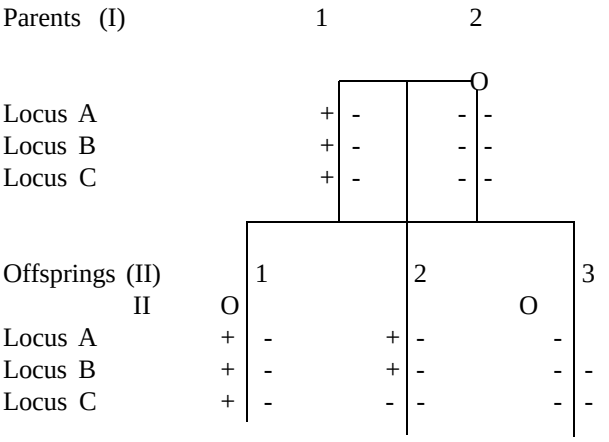
In a large population, if the frequency of one allele at a locus is .999 and the other is 0.001, the vast majority (99.8%) of the individuals would be homozygous for the more frequent allele. If two alleles were equally frequent (.5 each one), then half the population would be heterozygotous at the locus. As a consequence, human genetic analysis depends on loci with frequently occurring alleles.

When two or more alleles of a locus occur with a frequency of 0.01 or greater in a population, a genetic polymorphism exists, and the locus is said to be polymorphic. Polymorphism is a naturally occurring variation in the sequence of genetic information on a segment of DNA among individuals.

At the DNA level, a single nucleotide base (A or C or G or T) difference between two homologous genes is sufficient to create an allele. In many instances, a single base pair change can cause a gene product to differ drastically from the normal product. There should also be a large number of single base pair differences, within a gene, that have no effect on the gene product, and others that occur without any biological consequences in segments of the DNA that do not code for proteins. The actual analysis of DNA samples from a group of individuals is slightly more complicated because chromosomes occur as pairs. However, each genotype (++, +-,—) produces a distinctive pattern of fragments after hybridization (process of base pairing of two complementary strands). The pattern of DNA fragments that is the result of the presence and/or absence of a mutated restriction endonuclease site occurring frequently in a population is called a restriction fragment length polymorphism (RFLP).

A set of alleles on a particular chromosome transmitted from parent to child is called a haplotype. A diplotype is defined for two haplotypes carried by an individual. When a single site is examined, there are two possible haplotypes (+,-); when two different sites on the same chromosome are examined, there are four haplotypes (++, +-,-+,—), and with n loci, there are 2^n haplotype. Medelian framework and the inheritance of an RFLP locus can be traced within a pedigree. The determination of the alleles of an RFLP locus in an individual is called haplotyping (geno-typing, DNA typing). When two or more linked RFLP loci are followed in a pedigree, it is possible to determine the occurrence of a recombinant event, a new arrangement of linked gene loci occurred in the offsprings as a result of the crossing over (see figure 2).

Figure 2. Occurrence of a recombinant event in three Locus



The father (I-1) is heterozygous for three different RFLP loci on the same chromosome, and the mother (I-2) is triply homozygous; the offspring II-2 received from his parents a chromosome tha had undergone a crossover event at locus C; the others offsprings inherited non-crossover chromosomes from their parents.

Although RFLPs have provided a useful set of loci for genetic studies, additional studies revealed that these loci are not distributed uniformly throughout every

chromosome. Fortunately, other polymorphic loci consisting of simple repeating units of two, three, or four nucleotide pairs (short tadem repeats) occur in large numbers (>100, throughout the human genome and can be scored readily with the polymerase chain reaction. The dinucleotide repeat CA/GT occurs about 100,000 times throughout the human genome:

CACACACACACACACACACACA..
GTGTGTGTGTGTGTGTGTGTGTG…

Overall, these blocks consist of repeating CA/GT units ranging in length from 2 to 40 units, with any block at a particular chromosome location retaining more or less the same number of units. In addition to CA dinucleotide repeats, there are tri- and tetranucleotide repeats scattered throughout the human genome. The entire strand of a short tadem repeat is fewer than 400 bases in length (www.rsc.org).

1.7) Genotyping Single-Nucleotide Polymorphisms

The most common type of polymorphism in the human genome and easiest to measure is the single nucleotide polymorphism (SNP) or point mutation. It thus, has been the primary focus of recent genetic epidemiological studies [Bhatti et al., 2006]. Transitions are the most common type of SNP; transversions are less common [Thomas, 2004]. Many hemoglobinopathies are due to point mutations that cause the replacement of an amino acid (misense), and, are, consequently, abnormal protein products. The most common, causing Tay-Sachs disease, is a 4-base pairs insertion (frameshift); this disease is a fatal genetic lipid storage disorder in which harmful quantities of a fatty substance called *ganglioside* G_{M2} build up in tissues and nerve cells in the brain (www.ninds.nih.gov/disorders/taysachs.htm).

Each individual has many single nucleotide polymorphisms that, together, create an unique DNA pattern for that person. SNP represents a site at which two different nucleotide pairs occur at a frequency of 1% or greater. In the human genome, the number of SNPs in chromosomes 1, 10, and 20 are: 923864, 571051, and 315702, respectively. Various strategies have been used to discover SNPs [Bhatti et al., 2006]. Initially, differences within a sequence tagged sites from different individuals revealed candidate SNP sites. More recently, detailed analysis of overlapping DNA sequence derived from the Human Genome Project provide huge numbers of potential SNPs. Additional studies are required to establish the extent of the polymorphism; whether the flanking sequence acts as effective primer regions; and whether a site is a reliable DNA marker **(**a polymorphic gene whose physical location is known) [Greg et al., 2004].

Single nucleotide polymorphisms (SNPs) are becoming widely used as genotypic markers in genetic association studies of common, complex human diseases. For such association screens, a crucial part of study design is determining what SNPs to prioritize for genotyping. The data are accessible through search and browse in different interfaces, allowing users to select proteins or SNPs of interest using a number of identifiers and phenotypic effects, selecting SNPs or proteins of interest based in an overview of SNP properties and phenotypic effects and links to SNP and protein entries in related databases like dbSNP, UCSC browser, Hap map database, JSNP, HGV base, etc.

Two general strategies for selecting SNPs in association studies include haplotype tagging methods or targeted selection of candidate genes and candidate variants. Whole-genome scans will become increasingly technologically efficient and economically feasible in the near future. Meanwhile, scientists using candidate gene, SNP, or haplotype approaches face the challenge of choosing among 10 million possible SNPs or smaller numbers of haplotype-tagging SNPs. In the context of prioritization of candidate SNPs that are most likely to be biologically relevant, numerous criteria are useful (Bhatí et al., 2006). The candidate SNPs and genes should be selected from publicly available sources based on some of the following criteria: 1) functional relevance and importance for biologically events; 2) degree of heterozygosity, i.e., allele frequencies, as reported in literature or databases; 3) position in or around the genes; and 4) their use in previous genetic epidemiology studies.

II) Genetic Epidemiology

2.1) Introduction

The term genetic epidemiology was originally coined by Neel and Schull in 1954 to describe the confluence of two fields required for the study of common diseases, their population distribution and etiology. Because of this hybrid nature, genetic epidemiology draws from several distinct related fields: population genetics, quantitative genetics, epidemiology, and biostatistics. The heart of genetic epidemiology understands the genetic and environmental contribution to a disease and how they relate to one another [Risch, 2002].

The understanding of the relationship between a phenotype, whether it is normal or abnormal, and its corresponding genotype depends, ultimately, on the ability to isolate (clone) and characterize an individual's gene(s). The DNA sequence of a gene reveals the domain of the encoded protein, how mutations in different exons are responsible for a disease. With a cloned gene, experiments

can be developed to determine how various mutant gene products destabilize the normal process. In addition, diagnostic tests for specific gene mutations can be developed from the DNA sequences of normal and mutated genes [Pasternak J., 2005]. However, there are considerable amounts of important ancillary details concerning genetic, physical, and cytogenetic maps; location of known genes and transcribed regions; base pair composition; single-nucleotide polymorphism; repeated sequences; as well as other features. With unambiguous information, a human geneticist selects from only four different familial patterns to determine the mode of inheritance of a trait that resides at a single gene locus. Included in this category of a single-gene (monogenic) traits are autosomal dominant (eg., Huntington's disease, polycystic kidney disease, retinitis pigmentosa,..), autosomal recessive (eg., cystic fibrosis, b-Thalassemia, …), X-linked recessive (eg., hemophilia A, color vision defect, ...), and X-linked dominant conditions [Knipe et al., 1996].

2.2) Methodology

In genetic epidemiology, classical epidemiological designs (i.e., case-control study and cohort study) have been used to identify, directly or indirectly, the most probable gene(s) to be causing a particular disease. Due to the complexities in finding a particular gene, in genetic epidemiology, different methods are used in a sequence to better understand the location of particular gene(s), usually the "starting point" depends on the resources and information available. Some of these methods will be carried out using DNA information, while others may only use the phenotype characteristics and no DNA samples (Thomas, 2004). A description of these methods is given with the most recent publications, as follows:

Descriptive epidemiology - Use of routinary data, such as standardized rates to compare groups, can provide clues to whether genetic or environmental factors are involved. For example, the difference of the age-adjusted incidence rate of female breast cancer between Black and White women population in the USA, that showed that Black women have 10% less incidence than White [Althuis et al., 2005], provided the clues to investigate for genetic explanation in a particular ethnic group.

Familial Aggregation - The first step in pursuing possible genetic epidemiology is usually to demonstrate that the disease tends to "run in families" more than would be expected by chance. Geneticists refer to such clustering as familial aggregation. Family studies have a central role in genetic epidemiology. Although epidemiology generally involves studies of unrelated individuals, often using population-based sampling, genetic epidemiology focused on related individuals in the form of family histories or opportunistically identified and sampled pedigrees [Hopper et al., 2005]. This is often based on case-control studies using family history or on twin or adoption studies. For example, a family history has shown that the odds of having breast cancer in women with a first degree relative with breast cancer is 2.4 (95%CI: 1.84, 3.06) times this odds in women with no first degree relative with beast cancer [Slattery et al., 1993]. Another example is in prostate cancer, where it has been estimated that 10%-15% of patients with this cancer have at least one relative who is also affected and first degree-relatives of patients with prostate cancer have a two-fold to three-fold increased risk for developing this disease [Gronberg, 2003].

Segregation analysis - The objective in this step is to determine whether the pattern of disease among relatives is compatible with one or more major genes or shared environmental factors. For this purpose, no molecular data are used, the aim being to test hypotheses about whether one or more major genes and/or polygenes can account for the observed pattern of familial aggregation, the mode of inheritance, and to estimate the parameters of the best-fitting genetic model [Thomas, 2004]. For example, a case-control study of breast cancer risk [Claus et al., 1998] developed a model to estimate the probability of carrying a mutation in BRCA1 or BRCA2 genes using 4730 cases (20-54 years old) with histological confirmed breast cancer and 4688 control subjects. In each instance, a woman's probability of being a gene carrier is calculated conditional to the breast and ovarian cancer status of first-degree and second degree female relatives as well as the current age or age at death of any unaffected female relatives. The model used Bayes Theorem within the Mendelian framework, assuming an autosomal dominant transmission for both BRCA1 and BRCA2 genes. Women with a probability of less than or equal to 1% of carrying mutations in either BRCA1 and BRCA2 genes were defined as noncarriers. The results showed that among BRCA1 noncarriers, case subjects were 2.1 (95%CI: 1.2,1.3) times more likely to report a first-degree or second-degree family history of breast cancer, than were control subjects. Noncarriers were predicted to have a lifetime risk of 9% of developing breast cancer compared with a 63% risk for carriers. Thus, a woman who is identified as a carrier of the risk marker and who also has a strong family history of breast cancer is likely to possess a much higher risk for breast cancer than a carrier with no known family history of the disease.

Linkage Analysis - The objective is to find the location of a major gene by looking for evidence of cosegregation with other genes whose locations are already known (i.e., markers genes). Cosegregation is a tendency for two or more genes to be inherited together, and hence for

individuals with similar phenotypes to share alleles at the marker locus. Recombination is very unlikely to occur between two loci that are immediately adjacent to each other. The probability of recombination increases with the physical distance between two loci, from zero for adjacent loci to a limiting value ½. It is this gradient in recombination probabilities that allows linkage analysis to determine the probable location of a gene. Thus, either a nonrecombination haplotype *ab* or *AB* might be transmitted, each with probability (1-q)/2, or a recombination haplotype *aB* or *Ab*, each with probability q/2. The *lod score* (lods), which is equivalent to the likelihood ratio test, has been used to assess the hypotheses H_o: q=1/2, as follows:

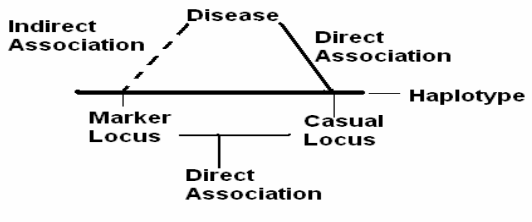
$$lod(q) = \log_{10}[\ell(\hat{q}) - \ell(1/2)]$$

where is the likelihood function using the Binomial distribution to determine the probability of the number of recombinations in one family [Thomas, 2004]. In genetic epidemiology, blood samples are taken from potentially informative members of multiple case families and typed for genetic markers at a known location. Extended families with many cases are particularly informative for this purpose and do not need to be population-based, although large series of pairs of affected siblings can also be used. Linkage analysis was used to localize BRCA1 and subsequently BRCA2, using extended pedigrees with many cases of breast and ovarian cancer [Thomas, 2004]. Also, it was used to assess the association between the body mass index (BMI) and chromosomes 6 and 11, across six examinations of the Framingham Heart Study; where it was shown that linked regions on these chromosomes remained significantly associated [Atwood et al., 2006].

Fine mapping - The objective is to locate a gene using haplotypes and linkage disequilibrium (LD). LD is defined as the non-random association between two alleles at the two different loci on the same chromosome. LD is often termed "allelic association." LD captures a deviation from probabilistic independence among alleles or genetic markers. For instance, LD between two alleles, say A and B, can be quantified by measuring the following difference:

$$Pr\{AB\}-P\{A\}P\{B\}$$

where P{AB} is the probability of observing haplotype AB, P{A} is the probability of observing haplotype A, and P{B} is the probability of observing haplotype B. Haplotypes, however, are not directly available in most cases and their frequencies must be inferred probabilistically from genotype data [Montana, 2006]. The essential idea is that a marker locus in strong LD with a disease causal locus is expected to be located nearby, as follows:



LD mapping is carried out after genetic linkage between a polymorphic locus and the disease gene is determined. Then, members of families with a genetic disease within a founder population are haplotyped with a number of additional polymorphic markers on the same chromosome, and statistical tests are run to determine which loci are in linkage disequilibrium with the gene disease [Cheng I., et al., 2006]. The statistical tests to measure LD are based on the following measurement:

$$D' = \frac{D}{D_{max}} = \frac{Pr\{AB\} - Pr\{A\}Pr\{B\}}{D_{max}}$$

where D_{max} is the maximum difference between Pr{AB} and Pr{a}Pr{b}. D'=1 denotes complete LD, and historical recombination results in the decay of D' towards zero [Wang et al., 2005].

Association with gene candidate - The objective is to test different candidate genes with a particular disease. The linked region may include a number of genes with known functions that could be relevant to the etiology of the disease. By comparing the genotype at these candidate loci using the case-control epidemiological design (population-or family based), one can test hypotheses about whether they are actually associated with the disease. The fundamental study designs and statistical analysis methods for testing associations between genetic polymorphisms and a disease are similar to classical epidemiological designs; however, in order to homogenize the genotype background, some control in the familial background is recommended. In the web site Online Mendelian Inheritance in Man at www.ncbi.nlm.nih.gov, there are more than 15,000 known disease-causing Mendelian disorders listed, and provides links to genes that have been implicated in the etiology of complex multifactorial diseases [Greg et al., 2004].

Cloning the gene and identifying mutations - The objective is to determine the molecular sequence of the disease. When a candidate region is sufficiently narrow and no candidate gene has been found in that region, DNA from that region can be exhaustively searched for

polymorphisms. Polymorphisms in diseased persons that are rare in non-diseased persons are considered to be possibly causal mutation. For example, in the L-myc EcoRI polymorphisms, that produces the S and L alleles, it has been shown that individuals carrying the S allele tend to have poor cancer prognosis; i.e., in lung cancer, the S/S genotype was significantly associated with lymph node metastasis (R=2.8, 95%CI: 1.8, 4.3), distant metastasis (R=4.7, 95%CI: 2.4, 9.2), and clinical stage (R=2.3, 95%CI: 1.2,4.4) using the L allele (L/L genotype) as the reference group [Spinola et al., 2004].

Characterization of the gene- The objective is to describe the effect of the gene in an experimental setting. Genetic epidemiology is applied to estimate the frequency of the various mutations and the disease risk of these mutations, including confounding and interaction assessment; particularly with age, host, and environmental factors. The DNA sequence of the cloned gene reveals the domain of the encoded protein, how mutation disrupts its function, and the extent to which mutations in different exons are responsible for a disease. After a putative disease-causing gene has been cloned and sequenced, it is screened for a nucleotide change involving one or a few base pairs. The underlying principle of a mutation detection assay is that the nucleotide sequence of the gene in affected individuals will differ from a sequence content of the same gene in individuals with a normal phenotype. In order to understand the role and function of the genes, one needs the complete information about their mRNA transcripts and proteins. Unfortunately, exploring the protein functions is very difficult due to their unique 3-dimentional complicated structure and a shortage of efficient technologies. To overcome this difficulty one may concentrate on the mRNA molecules produced by the genes of interest (gene expression) and use this information to investigate specific questions of the functional roles of the genes. Several statistical methods are currently used for the analysis of gene expression in microarray data samples [Speed et al., 2003; McLachlan et al., 2004]. These methods can be classified in two major groups: i) methods that identify differentially expressed genes, and 2) methods that classify the functional dependency of genes. The objective of the first method is to identify those genes that are consistently expressed at different levels under different conditions using the classical statistical test (t-test, ANOVA, Mann-Whitney test,…) controlling the probability of false declaration [McLachlan et al., 2006]. The second method pretends to identify the shared patterns of expression across genes to classify new diseases of subtype of diseases for subsequent validation and prediction, and ultimately to develop individualized prognosis and therapy, using cluster analysis methods [McLachlan et al., 2004].

2.3) Analytical studies

Analytical study in classical epidemiology refers to an observational study where at least two individual groups (exposed and un-exposed or diseased or no-diseased) are observed, prospectively or retrospectively, in order to provide evidence of cause-effect relationship between an exposure and a disease under study. For this purpose, a comparison group is used as a reference to determine the effect of the exposure factor or the occurrence of the disease. The comparison groups (exposure vs. non-exposure or disease vs. non-diseases) could be selected from different sources (random survey of the population, registries, death certificates,…) with the condition of having similar genetic backgrounds. In genetic epidemiology, the hereditary diseases are the diseases under study and exposure is a genetic factor (directly or indirectly the presence of a particular chromosome, DNA regions, SNP,…).

One of the concerns in genetic epidemiology is to find the comparison group, because the genetic information is usually not available at the time of subject selection; so, controls are often matched instead on race or ethnicity, so they are more representative of the source population of cases [Witte et al., 1999]. For example, a population-based study of prostate cancer required that controls have three of four grandparents from the same ethnic group as the case to which they were matched [Whittemore et al., 1995]. However, race and ethnicity are of great concern for potential confounding because of the variability within the ethnic groups (i.e., Caucasian, African American, Latino, Asian). Geneticists call such confounding "population stratification (admixture)". Classically, confounding is the distortion of the relationship between the exposure of interest and disease due to the effect of a true risk factor that is related to the exposure. Similarly, population stratification is the distortion of the relationship between a genotype of interest and disease due to the effect of a true risk factor that is related to the genotype. In population stratification, ethnicity acts as a surrogate for the true risk factor, which may be environmental or genetic; as such, controlling for ethnicity can reduce the confounding bias. However, a recent study has shown that self-reporting ancestry may not be a reliable method to reduce the possible impact of population stratification in genetic association studies [Burnett et al., 2006]. If the population structure is recognized, it can be accounted for either at the design or the analysis stage of a study. Thus, the most important potential threat from population structure arises when the structure is unknown, so-called

cryptic substructure (Devlin and Roeder, 1999).

As a comparison group, spouses and adopted children are of interest in family studies, since they are genetically unrelated but are likely to share a common adult environment and similar demographic characteristics (other than gender). For example, in a study of multiple sclerosis (MS), the spouse did not experience an increased risk of MS, suggesting no major role for environmental factors acting in adulthood [Nielse et al., 2005; Thun et al., 1999]. In another study related with colorectal cancer, it was shown that the risk of colorectal cancer was 1.8 (95%CI: 1.2, 2.7) higher for the parents and siblings of the patients with adenomas, as compared with the spouse control, after being adjusted by confounding variables; suggesting that genetic factors are related with this cancer [Winawer et al., 1996].

One of the complications in genetic epidemiology arises from the need to sample families rather than individuals, which are harder to frame than individuals. In the study of dichotomous diseases, a further complication is that multiple-case families are the most informative, and these are not efficient by simple random sampling of families. For these reasons, most studies of familial aggregation of disease are based on ascertainment of probands, followed by the identification of their family member. The proband is the individual who caused a family to be identified and included in a genetic analysis, usually a person with the disease under study. This selection procedure has been called kin-cohort design; this design has several practical advantages, including comparatively rapid execution, modest reduction in required sample size compared with cohort or case-control designs; however, this design is subject to several biases, including the following: selection bias that arises if a proband's tendency to participate depends on the disease status of relatives, information bias from inability of the proband to recall the disease histories of relatives accurately, and biases that arise in the analysis if conditional independence assumption is invalid or if samples are too small to justify standard asymptotic approaches [Gail et al., 1999].

Ascertainment concerns the manner by which families are selected for genetic analysis and how to correct for it in likelihood models. Because such families are often neither drawn at random nor selected according to well-defined rules, the problem of ascertainment correction in the genetic analysis of family data has proved hard-wearing.

Two extreme ascertainment methods are complete ascertainment and single ascertainment. In complete ascertainment, all families in a defined population with at least one case are included; meanwhile, in single ascertainment, the families are included with probability proportional to the number of affected members.

Population-based ascertainment of probands (both diseased and non-diseased) is highly desirable for generalization and validity. If population-based disease and population registries are not available, then it is essential to establish the representativeness of the individuals with and without the disease, or, with and without the exposure [Thomas, 2004]. When it is difficult to ascertain all individuals in an area, a secondary base is chosen. For example, if one were to select all children diagnosed with cancer from a particular hospital, the proper control group would be those children who would have gone to that hospital had they developed cancer. The difficulty of defining this group will vary with the complexity of hospital referral patterns, which is especially true for childhood cancer because individual hospitals may have expertise related to specific diagnoses [Ross et al., 2004].

Family- based studies remain widely used, and are being further developed. However, family-based designs typically imply higher genotyping costs and can face difficulty in recruiting enough families. The simplest established method for adjusting for the effects of cryptic substructure is Genomic Control (Devlin and Roeder, 1999), which considers the distribution over the null markers of Y^2, the Armitage test statistic that compares average allele counts in cases and controls. Since few, if any, of the null markers are expected to have a causal association with the disease phenotype, any inflation of the empirical Y^2 values above their nominal distribution may be attributed to demographic effects, such as cryptic substructure. Also the statistical procedures for genomic control require the genotypes of cases and controls at several "null" markers that are not in linkage disequilibrium with the gene being tested for association and may imply additional genotyping costs, this is modest compared with the cost of implementing a family-based design. In 2007, Setakis demonstrated that using a logistic regression model including the null markers like covariates in the model protect against false positives and mitigate population stratification effects under extreme ascertainment bias, without significantly compromising power. These methods do not require an estimate underlying subpopulations like genomic control and it is easier to statistical modeling and interpretation; despite the fact that population structure is not explicitly modeled (Setakis et al., 2007).

2.3.1) Cohort Study

The cohort study is one of the important analytical studies in epidemiology. In this study, there are at least two groups of individuals that are exposure and un-exposure to a specific factor, without the disease of interest

at baseline. In genetic epidemiology, the exposure group is composed of individuals who have a specific phenotype or genotype characteristic (trait). The participants of the cohort design are followed in time to determine the risk of developing a specific disease. This study is desirable because exposure precedes the health outcome — a condition necessary for causation — and is less subject to bias because exposure is evaluated before the health status is known. As a consequence, it limits the possibility of investigator preferences, or "bias," affecting the selection of study subjects. It is most useful for estimating the incidences (risk of developing a disease) in the exposed and un-exposed groups. The cohort approach enables us to look for different outcomes because the study subjects are selected on the basis of their exposure only. The main disadvantages are the costs, the time involved when the incidence is low, and the rate of lost to follow-up.

An example of a recently published cohort study aimed to determine the role of the hepatic lipase gene (LIPC-480C>T) in predicting coronary heart disease and the modified effect of physical activity, using a population-based prospective study in the San Luis Valley of Colorado. Hispanic and non-Hispanic White (n=966) were followed for 14 years (1984-1998). The results showed that LIPC-480 TT genotype predicted an increase in coronary heart disease in both ethnic groups, and physical activity altered this relation; in normal levels of physical activity, the hazard ratio was 2.6 (95%CI: 1.4, 4.8); while in persons with vigorous physical activity, the hazard ratio was not significant (est. HR=0.5, 95%CI: .1,2.2) [Hokanson et al., 2003].

Another cohort study was developed to investigate the role of Androgen receptor gene polymorphisms in predicting the pathogenesis of benign prostatic hyperplasia among 510 men randomly selected from Olmsted County (Minnesota) from 1990 through 2000. Androgen receptor CAG and GGN genotyping was performed. A CAG repeat length of <21 was associated with enlarged prostate (est. HR=1.4, 95%CI: 1.0,1.9) and serum prostate-specific antigen level >1.4 ng/ml (est. HR=1.5; 95%CI: 1.1,2.0) [Rosebud et al., 2004].

2.3.2) Case-control studies

The case-control design is another important analytical study in epidemiology. At the beginning of this design, there are at least two groups of individuals that are already diagnosed with and without the disease of interest, in order to compare their level of exposure to a specific factor in the past. Usually, the diseased individuals are identified as cases; and the non-diseased individuals as control. In genetic epidemiology, the exposure group is the individuals which have a especific phenotype or genotype (trait). The case-control design is recommended when the incidence of the disease is low. The majority of the diseases in genetic epidemiology are relatively rare, so case-control studies are the most recommended design.

The main advantage of the case-control study is that it enables us to study rare health outcomes without having to follow thousands of people and is, therefore, generally quicker, cheaper, and easier to conduct than the cohort study. One of the major drawbacks of case-control in conventional risk factor epidemiology is recall bias due to the retrospective collection of exposure; however, in genetic epidemiology this is not a concern, since a subject's constitutional genotype does not vary over time and is not subject to the individual's memory [Thomas, 2004].

In the conventional case-control, one randomly selects controls from the source population of cases. The objective in selecting controls is to sample individuals representative of those who, had they developed disease, would have been selected as cases, and to sample these controls independently of exposure. In particular, controls should be sampled from the set of subjects meeting any matching criteria who have attained the age at which the case was diagnosed, and who were still disease-free at that age instead of population controls one could match each case to his or her younger non-diseased sibling(s) by: 1) constructing a likelihood which allows for the possibility that the sibling will develop the disease before the case's age; and 2) choosing a reference date that is sufficiently prior to the age at diagnosis of the case to include the entire exposure period of the control [Witte et al., 1999].

Another matching control could be an affected cousin of each case; this control might allow for closer matching on age than siblings control. In addition, there will be more cousins than siblings available as potential controls. However, cousin's control does not provide absolute protection from population stratification like sibling's control does, since cousins each have one parent that typically did not descend from a common ancestor [Witte et al., 1999].

Also, spouse has been chosen as a control in case-control studies, particularly in family case-control studies. Among the advantages of this type of control are feasibility, higher response rate, and comparability of recall among relatives in the case and control groups. However, the choice of spouse control will often lead to a different sex distribution among cases and controls if the risk of the condition is sex-specific. Another concern is that men tend to marry younger women, which implies that the age of siblings and parents of case and control may differ as well. The main difference between the conventional case-

control study and a family case-control study designed to study familial aggregation is that the comparison takes place between the relatives of cases and controls and not between cases and controls [Verhage et al., 2003].

An alternative to case-control designs is matching each case to a hypothetical control (pseudosibs) having the possible combinations of parental alleles not inherited by the case. For example, assume that a case's parents have genotype A/B and C/D, respectively, at a locus of interest, and that the genotype A/C was transmitted to the case; so, there would be three types of pseudosibs with genotype A/D, B/C, and B/D. This study is also called case-parent-trios. This design was recently used to determine the relationship of the polymorphism, Val34Leu in factor XIII, and intrauterine growth restriction (birth weight below the 10th percentiles, according to gestational age and sex); the results showed that this polymorphism increased the risk of intrauterine growth restriction approximately 70% when the parent of origin was the father as opposed to the mother [Infante-Rivard et al., 2005].

One of the variations of a case-control study, not in terms of the comparison group definition but in terms of the controls selection, is the design called nested case-control study (or the case-control in a cohort study). The cases in this study that occur in a defined cohort are identified and, for each case, a specified number of matched controls is selected from among those in the cohort who have not developed the disease by the time of disease occurrence in the case. An example of this design was applied in a cohort of Taiwanese men who were carriers of hepatitis B virus; this condition has been associated with hepatocellular carcinoma with higher incidence in males than in females. The results of this study showed that BMI modified the association of hepatocellular carcinoma with testosterone and SRD5A2 genotype in men with low BMI (<23.2), the adjusted OR for the SRD5A2 polymorphism VV versus the LL was 8.64 (95%CI: 2.8, 27.1) [Ming-Whei et al., 2001].

III) Statistical considerations

3.1) Consideration 1: Correlated data

In genetic epidemiology, the *Odds Ratio* is the measurement used to assess strength of association between exposure and disease, usually estimated by the logistic regression model [Xu & Shete, 2006; Zou, 2005]. This measurement will provide the clues for finding the major gene(s) related with a specific disease, particularly in case-control designs. Other measurements have also been used for this purpose; for example, the Relative Risk and Hazard Ratio, particularly in longitudinal studies, using a Poisson regression model and Cox proportional hazard model, respectively [Burton et al., 2005]. Usually, these measurements are adjusted by the following factors: age, sex, and ethnic group. In order to estimate these measurements, the assumptions of correlated and non-correlated data are considered; usually, when the design involved the participation of all members of the family, correlated data are expected. When non-correlated data are assumed, the classical assumption of independent observations is used to estimate the parameters of the mentioned models, usually identified as Generalized Lineal Model [McCullagh & Nelder, 1995]. In genetic epidemiology, due to the population stratification bias, it is sometimes difficult to support the assumption of independent observations. So, to account for these similarities in the participants, which may induce correlation in the observations, generalized linear and latent mixed model (GLLAMM), has been used [Snijders et al., 2003; Kreft et al., 2004; Rabe-Hesketh et al., 2005]. In order to exemplify this approach, we describe the logistic regression (LR) model with two levels in a family-based design, as follows:

$$\text{logit}(p_{ij}) = a_j + b_j * G_i$$

where

p_{ij}.- indicates the penetrance of the disease given ith genotype (or ith risk allele) in family jth family, that is the conditional probability that a randomly selected individual in the study population possesses the disease, given the data.

α_j.- indicates the intercept in the linear combination of the model for jth family. When a random sample of families is considered and in order to take into account the correlation among members of the family, the intercept could be defined as follows:

$$a_j = g_0 + u_{0j}; \quad u_{0j} \sim N(0, s^2_{uo})$$

where g_0 is the average of $a's$ for all families, known as the fixed part; u_{oj} is the effect of the *j*th family (latent variable), known as the random part; s^2_{u0} is the variance of u_{oj}

b_j.-indicates the slope in the linear combination of the model for jth family. When a random sample of families is considered and the strength of the association between the disease and the genetic factor is different between families, this parameter could be defined, as follows:

$$b_j = g_1 + u_{1j}; u_{1j} \sim N(0, s^2_{u1})$$

where g_1 is the average of $b's$ for all families, known as the fixed part; u_{1j} is the effect of the jth family (latent variable) for the b_j estimation, known as the random part; s^2_{u1} is the variance of u_{1j}.

G_i.- indicates the genetic risk factor under evaluation

The methods to estimate the parameters of the GLLAMM is a complex one, usually it is an interactive process (one solution generates another solution). The most frequently used methods are based on the a first- or second- order Taylor expansion of the link function (in logistic regression, the link function is the logit(p)=log(p/(1-p)). When the approximation is around the estimated fixed part, this is called marginal quasi-likelihood (MQL); when is around an estimate of the fixed and random part, it is called penalized or predictive quasi-likelihood (PQL). Therefore, specialized software has been developed for this purpose (such as MLwiN, Winbugs, HLM, and VARCL), or new routines have been created in the classical statistical programs (such as SAS, BMDP, STATA, and R) (Snijders, 2003 Rabe-Hesketh, 2005).

One of the concerns, when the assumption of correlated data is not considered is the possibility of producing biased estimates of variance components and standard errors. For binomial data, one potential cause of extra-binomial variance is through a failure to identify correctly the different level with the model. Omitting an important level from the hierarchical structure implies that the within-group (e.g., within family) clustering of responses will not be adequately modeled and this can cause over-dispersion, since the observed number of successes (e.g., number of diseased persons) can only be assumed to belong to a binomial distribution when the observations are assumed to be independent (Leyland et al, 2004).

3.2) Consideration 2: Risk factor definition

The genetic risk factor under evaluation (G_i) can be defined in different ways. For example, the genotype (set of SNPs) at certain locus (i.e., *AA,Aa,aa*), combination of geneotypes at different loci (i.e., *AA/BB, AA/Bb, AA/bb, Aa/BB, Aa/Bb, Aa/bb, aa/BB, aa/Bb, aa/bb*), alleles at risk (yes/no), or combination of alleles at risk. As a consequence, different association studies can be defined for the same genetic risk factor [Balding, 2006].

3.3) Consideration 3: Selection of the Candidate polymorphism

If only one polymorphism is being implicated in disease causation, usually, the classical approach of Logistic Regression can be used [Hosmer & Lemeshow, 2000]. So, the *familial Relative Risk* can be approximated by the *OR* estimation, particularly when the penetrance is low. For example, assuming that the correlation among members of the same family was close to zero, the OR is estimated with 95% confidence as follows:

$$\hat{FRR} \sim \hat{OR} = e^{\hat{g}_1 \pm 1.96 * SE(\hat{g}_1)}$$

where $SE(\hat{g}_1)$ is the standard error of $\hat{g}_1$.

The main concerns are the criteria to select the polymorphisms as predictors of the logistic regression model and the adjustments to estimate the FRR that should be performed when several polymorphisms that could be highly correlated between them are present in the model. Another concern is that the number of candidate polymorphisms in the regression models should be less that the number of observations to have a unique solution.

3.4) Consideration 4: Definition of the Genetic Model

To estimate adequately the strength of the association using the OR's, it is recommended to define the genetic model (i.e., dominant, recessive, multiplicative) previous to determining the type of statistical description that will be performed and the type of OR that will be estimated. For example, a single SNP with alleles *A* and *B*, tested in an unmatched case-control design, the following OR's can be computed according to the type of genetic model [Lewis, 2002]:

i) Full Genotype

	AA	AB	BB	OR_{BBvsAA}	OR_{ABvsAA}
Cases	a	b	c	$c*d/f*a$	$b*d/e*a$
Control	d	e	f		

ii) Dominant Model: allele B increase risk

	AA	AB+BB	OR
Cases	a	b+c	$\frac{(b+c)*d}{(e+f)*a}$
Control	d	e+f	

iii) Recessive Model: two copies of allele B are required to increased risk

	AA+AB	***BB***	**OR**
Cases	a+b	c	$\frac{c*(d+e)}{f*(a+b)}$
Control	d+e	f	

iv) Multiplicative Model: analyzed by alleles, not by genotype

	A	B	OR
Cases	2a+b	b+2c	$\frac{(b+2c)*(2d+e)}{(e+2f)*(2a+b)}$
Control	2d+e	e+2f	

The aims of the study and the information available will determine the genetic model and the statistical analysis. The complexities of the analysis are increased when the number of alleles for each SNP is increased.

3.5) Consideration 5: Unphased data

The other concern, apart from the genetic model, is the lack of information about the evolutionary history (unphased data); for example, whether the marker allele has paternal or maternal origin. Unphased data contain less information about the evolutionary history of the sample and increase the risk of inferring nonexisting hotspots or, oppositely, failing to infer existing hotspots and actual recombination [Wiuf, 2004].

3.6) Consideration 6: Number of Genes

Candidate gene. These studies might involve typing 5-50 SNPs within a gene. The gene can be either a positional candidate that results from a prior linkage study or a functional candidate that is based, for example, on homology with a gene of known function in a model species. In this case, again, classical approach of LR can be used; however, the problem of interaction among SNPs or Epistasis has to be considered. The presence of epistatasis is a particular cause of concern, since, if the effect of one locus is altered or masked by effects at another locus, power to detect the first locus is likely to be reduced and elucidation of the joint effects at the two loci will be hindered by their interaction [Cordell, 2002]. For example, the following penetrance distribution for two loci interacting epistatically in a heterogeneity disease model (individual becomes affected, Y=1, through possessing a predisposing genotype at either locus A or locus B):

Y	Genotype at locus B		
Genotype al locus A	b/b	b/B	B/B
a/a	0	0	1
a/A	0	0	1
A/A	1	1	1

However, another biological phenomenon would be if the "effect" of locus B is in a recessive disease model (so that two copies of allele B are required to cause disease) then having two copies of alleles A at locus A is sufficient to "mask" this effect, i.e., given genotype A/A at locus A, the effect of locus B is not observed. So, direct biological inference from the results of the statistical test is very difficult. The degree to which statistical modeling can elucidate the underlying biological mechanisms is likely limited, and may require prior knowledge of the underlying aetiology [Cordell, 2002].

Fine mapping. Often refers to studies that are conducted in a candidate region of perhaps 1-10Mb and might involve several hundred SNPs. The candidate region might have been identified by a linkage study and contain perhaps 5-50 genes. The concern for the statistical modeling is that the number of predictors could be higher than the number of subjects. One alternative that has been proposed is to use principal components (PCs) analysis to compute combinations of SNPs that capture the underlying correlation structure within the locus, and then, estimate the OR's based on the PCs. The PC approach captures linkage-disequilibrium information within a candidate region, but does not require the difficult computing implicit in the haplotype analysis [Gauderman et al., 2007].

Genome-wide. These studies seek to identify common causal variant throughout the genome. A typical genomewide association study is now expected to contain data on ^{3} 500k assayed SNPs for several thousand of individuals; however, increasing the marker density is not a guarantee to detect association if the penetrance is low [Thomas et al., 2005]. Cluster Analysis have been recommended for GWA to initially group together genes with similar pattern of expression and then test for a dense set of markets [Eisen et al., 1998].

3.7) Consideration 7: Interaction terms

In complex diseases, such as diabetes, asthma, hypertension, and multiple sclerosis, environmental and socio-demographic effects have to be considered in the model due to the multifactorial assumption of disease causation. Therefore, adjusted OR's can be estimated if no significant interaction terms are part of the logistic model. The interaction terms can be between SNPs or between SNPs with environmental or socio-demographic predictors, in particular, ethnic groups; for example, assuming E_i is a risk allele (0,1) and two ethnic groups (0,1), the most simple logistic regression model will be:

$$\text{logit}(p_{ij}) = a_j + b_j * G_i + b_{ethnic} * ethnic + d_{G*ethnic}$$

where $d_{G*ethnic}$ is the interaction term for the genetic risk factor and ethnic group. However, to validate the adjusted *OR* by ethnic group, it is necessary to previously assess the interaction term ($d = 0$?). If the interaction term does not show significant effect ($d = 0$), then *OR* adjusted by ethnic group can be estimated. On the contrary, if this term was significant ($d \neq 0$), then an *OR* has to be estimated for each ethnic group. So, the number of complications

will arise when several confounding variables are considered with a large number of SNPs, particularly in the assessment and interpretation of interaction terms.

IV) Conclusions

In most countries, the study of genetic epidemiology is becoming an area of priority in public health. The Institute of Medicine in the USA [IOM, 2002] has reported that genetics is a critical area for public health education in the 21st century. Nine of the ten top causes of death in the USA have a genetic component, among them are: heart diseases, cancer, chronic lower respiratory disease, diabetes, and Alzheimer's disease. As a consequence, most of the schools of public health are incorporating bioinformatics and genetic epidemiology courses in their curricular programs.

Genetic epidemiology is a "hunter process" to locate the most probable gene causing a heritability disease. It is a hierarchical process that combines observational design (case-control and cohort) with experimental procedures (DNA data), controlling the population stratification effect. One of the major drawbacks of case-control in conventional risk factor epidemiology is recall bias due to the retrospective collection of exposure; however, in genetic epidemiology this is not a concern, since a subject's constitutional genotype does not vary over time and is not subject to the individual's memory [Thomas, 2004]. The problem is that there are more than 19,000 possible genes and millions of combination of base pairs for different mutants.

Population stratification (admixture) is of great concern in genetic epidemiology that can lead to three distinct problems: confounding; cryptic relatedness, resulting in overdispersion of the test statistics (significance of any test could be either increased or decreased); and selection bias. Family-based case-control designs reduce the occurrence of these problems, but at the expense of some loss of power from "overmatching" on genotype [Thomas et al., 2005].

Statistical modeling offers an instrument to quantify the association between the disease and the genetic risk factor(s), using the *Odds ratio*, relative risks, and hazard ratio. Classical statistical modeling has been used to estimate these measurements, specially when the number of SNPs is small, despite the fact that some interactions can be difficult to interpret due to the Epistasis effect. Multilevel modeling has been an alternative to control correlation within subjects of the same class for family-based studies. However, the problem of analyzing these models could become very complex when fine-mapping and genome wide association studies are considered, particularly to assess interaction terms.

This introduction to the main aspects of genetic epidemiology has been written to encourage new investigators to get involved in this fascinating area, particularly biostatisticians, epidemiologists, and other scientists of related fields. We hope this objective has been accomplished. Welcome to a new challenge!!

Resumen

Nuevas opciones han surgido para los profesionales de la salud desde que se anunció en el 2003 que el genoma humano se había completado, coincidiendo con el 50vo aniversario del descubrimiento de la estructura elíptica del ADN por Watson y Crick en 1953. La actualización continua de la tecnología ha permitido analizar simultáneamente miles de variables para el análisis del genoma humano. Estos avances han creado nueva oportunidades para localizar genes, evaluar la relación entre genes, medir la interacción entre genes y el medio ambiente, describir el producto de los genes, y evaluar la relación entre gen y enfermedad. En epidemiología se han desarrollado nuevas estrategias para evaluar la relación de causa y efecto en estudios de casos y controles y estudios de cohorte. Con la información provista por el proyecto del Genoma Humano se han desarrollado nuevos diseños epidemiológicos y técnicas estadísticas. La incorporación de la biología molecular a los métodos tradicionales de epidemiología ha dado nacimiento a la disciplina conocida con epidemiología genética. El objetivo de este manuscrito es proveer una introducción a los conceptos necesarios para evaluar la asociación entre gen y enfermedad en estudios poblacionales. Primero, se presenta una descripción de los conceptos genéticos, como parte del marco teórico correspondiente para los diseños epidemiológicos y procedimientos estadísticos que se han utilizado en epidemiología genética. Posteriormente, se presenta una descripción de los diferentes diseños de estudio en epidemiología genética y las publicaciones más recientes. Finalmente, se discuten algunas consideraciones estadísticas para el análisis de los estudios de epidemiología genética.

Acknowledgements

We thank Laura Bretaña for her help in editing this document. We appreciate the initial comments and suggestions of Dra. Isabel do Santos. In addition, the first author wants to thank the Department of Mathematics,

University of Queensland (Brisbane, Australia) for their hospitality during his sabbatical year.

This study was supported by the School of Public Health of the University of Puerto Rico and by the following awards: (i) RCMI clinical research infrastructure initiative (RCRII) award, 1P20 RR11126, from the National Center for Research Resources (NCRR), NIH (ii) U54CA096297, MDACC/PRCC Partnership for Excellence in Cancer Research award, NIH/NCI**.**

References

1 Althuis M. , Dozier J, Anderson W., Devesa S., and Brinton L. 2005. Global trends in Breast cancer incidence and mortality 1973-1997. Int J Epidemiol;34:405-412.

2 Atwood L, Heard-Costa N., Fox C., Jaquish C., and Cupples A. 2006. Sex and age specific effects of chromosomal regions linked to body mass index in the Framingham Study. BMC Genetics, 7:7.

3 Balding D. 2006. A tutorial on statistical methods for population association studies. Nature Reviews|Genetics, vol. 7, oct..

4 Beer M., Poter R., Jones Th. (editors). 2006-2007. The Merck Manual, 18th edition. Merck Research Laboratories, Whitehouse station, N.J. Merck & Co., Inc.

5 Bhatti P., Church D., Rutter J., Struewing J., and Sigurson A. 2006. Candidate Single Nucleotide Polymorphism Selection using Publicly Available Tools: A Guide for Epidemiologists. Am J Epidemiol 164: 794-804.

6 Burnett M., Strain K., Lesnick T., de Andrade M., Rocca W., Maraganore D. 2006. Reliability of Self-reported Ancestry among Siblings: Implications for Genetics Association Studies. Am J Epidemiol 163: 486-492.

7 Burton P., Scurrah K., Tobin M., and Palmer L. 2005. Variance Components Models for Longituddinal Family Data. International Journal of Epidemiology 34:1077-1079.

8 Claus E., Schildkraut J, Iverse E, Berry D, Parmigiani G. 1998. Effect of BRCA1 and BRAC2 on the Association Between Breast Cancer Risk and Family History. Journal of the National Cancer Institute 90: 1824-29.

9 Cheng I, Penney KL,Stram DO, Le Marchand L, Giorgi E, Haiman CA, Kolonel LN, Pike M, Hirschhorn J, Henderson BE, Freedman ML. 2006. Haplotype-based association studies of IGFBP1 and IGFBP3 with prostate and breast cancer risk: the multiethnic cohort. Cancer Epidemiol Biomarkers Prev. 15(10):1993-7.

10 Cheng L, Carrolll P, Dahiya R. 2005. Epigenetic Changes in Prostate Cancer: Implication for Diagnosis and Treatment. Journal of the National Cancer Institute 97, 2: 103-15.

11 Cordel H. 2002. Epistasis: what it means, what it doesn't means, and the statistical method to detect it in humans. Human Molecular Genetics. Vol. 11, No. 20, 2463-2468.

12 Cowell JK, Hawthorn L. The application of microarray technology to the analysis of the cancer genome. 2007. Curr Mol Med 7(1):103-20.

13 Devlin B. and Roeder K. 1999. Genomic Control for Association Studies. Biometric 55, 997-1004.

14 Ducray F, Honnorat J, Lachuer J. 2007. DNA microarray technology: principles and applications to the study of neurological disorders. Rev Neurol (Paris) 163(4):409-20).

15 Egger G, Liang G, Aparicio A., Jone P. 2004. Epigenetics in Human Disease and Prospects for Epigenetic Theraphy. Nature. 429: 457-63.

16 Eisen M., Spellman P., Brown P., Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. Proc. NAtl. Sci., vol. 95, pp. 14863-14868.

17 Kraft P, Palmer Ch., Woodward A., Turunen J, Minassian S., Paunio T., Lonnqvist J, Peltonen L. and Sinsheimer. 2004. RHD maternal-fetal genotype incompatibility and schizophrenia: extending the MFG test to include multiple siblings and birth order. Eur J Hum Genet 12: 192-198.

18 Gail M., Pee D., Carroll R. 1999. Kin-Cohort Designs for Gene Characterization. J Natl Cancer Inst Monogr No. 26, 55-60.

19 Gauderman J., Murcray C., Gilliland F., Conti D. 2007. Testing Association Between Disease and Multiple SNPs in a Candidate Gene. Genet Epidemiol. DOI 10.1002/gepi.

20 Greg G. and Spencer M. 2004. A Primer of Genome Science. Sinauer Associates, Inc. Publishers.

21 Gronberg H. Prostate cancer epidemiology. Lancet; Mar 8, 2003; 361: 859-864.

22 Hopper J., Bishop T., Easton D. 2005. Genetic Epidemiology: Population –based family studies in genetic epidemiology. Lancet 366: 1398-1406.

23 Hokanson J., Kamboh M., Scarboro Sh., Eckel R., and Richard Hamman. 2003. Effects of the Hepatic Lipase Gene and Physical Activity on Coronary Heart Disease Risk. Am J Epidemiol 158,9, 836-43.

24 Hosmer D. and Lemeshow S. 2000 Applied Logistic Regression, John Wiley and Sons.

25 Infante-Rivard C., Weinberg C. 2005 Parent-of-Origin Transmission of Thrombophilic Alleles to Intrauterine Growth-Restricted Newborns and Transmission-Ratio Distortion in Unaffected Newborns. Am J Epidemiol Vol. 162, No. 9.

26 Knipe D., Howley P., Griffin D., Lamb R., Martin M. (Editors). 1996. Fields Virology, fifth edition. Lippincott-Raven publishers.

27 Kreft I., Leeuw J. 2004. Introducing Multilevel Modeling. Sage Publication, reprinted.

28 Lawson A. Browne W., Vidal C. 2003. Disease Mapping with WinBUGS and MLwiN. John Wiley and Sons, Inc.,

29 Lewis C. 2002. Genetics association studies: Design, analysis and interpretation. Brief Bioinform Vol. 3, No. 2, 146-153.

30 Leyland A., Goldstein H. 2004. Multilevel Modelling of Health Statistics. John Wiley and Sons, Inc., Reprinted.

31 Montana G. 2006. Statistical Methods in genetics. Brief Bioinform. Vol. 7,No. 3, 297-308.

32 McCullagh P. and Nelder J. 1995. Generalized Linear Models. Second Edition, Chapman and Hall.

33 McLachlan G., Kim-Anh Do, Ambroise Ch. 2004. Analyzing Microarray Gene Expresión Data. John Wiley & Sons, Inc.

34 McLachlan G. Bean R., Joves B. 2006. A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. Bioinformatics Vol. 22, no. 13, pages 1608-1615.

35 Ming-Whei Y., Cheng Sh., Ming-Wei L., Yang Sh., Liaw Y., Chang H., et al. 2000. Androgen-Receptor Gene CAG Repeats, Plasma Testosterone Levels, and Risk of Hepatitis B-Related Hepatocellular Carcinoma. J Natl Cancer Inst, Vol, 92, No. 24.

36 Nielse N., Westergaard T., Rostgaard K., Frisch M., Hjalgrim H., Wohlfahrt J., Koch-Henriksen N., and Melbye M. 2005. Familial Risk of Multiple Sclerosis: A nationwide cohort study. Am J Epidemiol 162:774-778.

37 Palacios G, Quan PL, Jabado OJ, Conlan S, Hirschberg DL, Liu Y, Zhai J, Renwick N, Hui J, Hegyi H, Grolla A, Strong JE, Towner JS, Geisbert TW, Jahrling PB, Buchen-Osmond C, Ellerbrok H, Sanchez-Seco MP, Lussier Y, Formenty P, Nichol MS, Feldmann H, Briese T, Lipkin WI. 2007. Panmicrobial oligonucleotide array for diagnosis of infectious diseases. Emerg Infect Dis. Jan;13(1):73-81.

38 Pasternak J. 2005. An Introduction to Human Molecular Genetics: Mechanisms of Inherited Diseases, second edition. John Wiley & Sons, Inc.
39 Rabe-Hesketh S. Skrondal A. 2004. Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models. Chapman & Halls/ CRC.
40 Rabe-Hesketh S. Skrondal A. 2005. Multilevel and Longitudinal Modeling Using STATA. STATA Press.
41 Rappaport R. 1996. Cytokenesis in Animal Cells. Developmental and Cell Biology Series. Editors: Barlow P., Bard J., Green P. Kirk D. Cambridge University Press.
42 Risch N. 2002. Genetics epidemiology. Chapter 15, Vol I of Emery and Rimoin's Principles and Practice of Medical Genetics, fourth edition. Rimoin D. et al (editors). Published by Churchill Livinston,
43 Rimoin D., Connor J., Pyeritz R., Krrf B. (Editors). 2002. Emery and Rimoin's Principles and Practice of Medical Genetics, fourth edition. Published by Churchill Livinston.
44 Rosebund R., Bergstralh E., Cunningham J, Hebbring S, Thibodeau S., Lieber M., and Jacobsen S. 2004. Androgen Gene Polymorphisms and Increased Risk of Urologic Measuires of Benign Prostatis Hyperplasia. American Journal of Epidemiology; 159,3; 269-76.
45 Ross J., Logan G. Spector, Andrew F. Olshan and Greta R. Bunin. 2004. Invited Commentary: Birth Certificates—A Best Control Scenario? Am J Epidemiol 159:922-924.
46 Sariol CA, Munoz-Jordan JL, Abel K, Rosado LC, Pantoja P, Giavedoni L, Rodriguez IV, White LJ, Martinez M, Arana T, Kraiselburd EN. 2007. Transcriptional activation of interferon stimulated genes but not of cytokine genes after primary infection of rhesus macaques with dengue virus type 1. Clin vaccine immunol apr 11; [epub ahead of print].
47 Setakis E., Stirnadel H., Balding D. 2007. Logistic Regression protects against population structure in genetic association studies. Genome Research, Downloaded from www.genome.org on October 21, 2007.
48 Slattery M., Kerber R. 1993. A comprehensive evaluation of family history and breast cancer risk: The Utah population database. JAM Vol 270(13): 1563-1568.
49 Snijders T., Bosker R. 2003. Multilevel Analysis: An Introduction to Basic and Advanced multilevel Modeling. Sage Publication, reprinted.
50 Speed T (editor). 2003. Statistical Analysis of Gene Expression Microarray Data. Chapman & Halls/ CRC.
51 Spinola M., Pedotti P., Dragan T., Taioli E. , 2004. Meta-Analysis Suggest Association of L-myc EcoRI Polymorphism with Cancer Prognosis. Clin Cancer Res Vol. 10, 4769-4775.
52 Thomas D. 2004. Statistical Methods in Genetic Epidemiology. Oxford University Press.
53 Thomas D., Haile R., Duggan D. 2005. Recent Developments in Genomewide Association Scans: A workshop Summary and Review. Am. J. Hum. Genet. 77:337-345.
54 Thun M, Henley J, Apicella L. 1999. Epidemiologic studies of fatal and nonfatal cardiovascular disease and ETS exposure from spousal smoking. Environ Health Perspect. Suppl 6:841-6.
55 Verhage B., Aben K., Straatman H., Verbeek A., Beaty T. and Kiemeney L. 2003. Spouse controls in family case-control studies: a methodological consideration. Familial Cancer 2: 101-108.
56 Verzilli C., Stallard N., Whittaker J. 2006. Bayesian Graphical Models for Genomewide Association Studies. Am. J. Hum. Genet., Vol. 79.
57 July Wang W., Barrat B., Clayton D., Todd J. 2005. Geneome-Wide Association Studies: Theoretical and Practical Concerns. Nature Reviews|Genetics. Vol. 6.
58 Watson J., Baker T., Bell S., Gann a., Levine M., Losick R. 2004. Molecular Biology of the Gene, fifth edition. Benjamin Cummings Publisher.
59 Winawer S., Zauber A., Gerdes H., O'Brien M., Gottlieb L., et al. 1996. Risk of Colorectal Cancer in the Families of Patients with Adenomatous Polyps. N Engl J Med. 334(2):82-7.
60 Witte J, Gauderman J, and Duncan Th. 1999. Asymptotic Bias and Efficiency in Case-Control Studies of candidate Genes and Gene-Environment Interactions: Basic Family Designs. Am J Epidemiol Vol. 149, No. 8.
61 Wittermore SS, Kolonel LN, Wu AH, John EM, Gallagher RP, Howe GR, Burch J, Hankin J, Dreon DM, West DW. 1995. Prostate cancer in relation to diet, physical activity , and body size in blacks, white, and Asians in United States and Canada. J Natl Cancer Inst 87: 652-61.
62 Wiuf C. 2004. Inference on Recombination and Block Structure Using Unphased Data. Genetics 166: 557-545.
63 Woodworth G. 2004. Biostatistics: A Bayesian Introduction. John Wiley and Sons, Inc.,
64 Xu H and Shete S. Mixed-effects Logistic Approach for Association following Linkage Scan for Complex Disorders. 2006. Annals of Human Genetics 230-237.
65 Zou G. Statistical Methods for the Analysis of Genetics Association Studies. 2006. Annals of Human Genetics 70, 262-276.