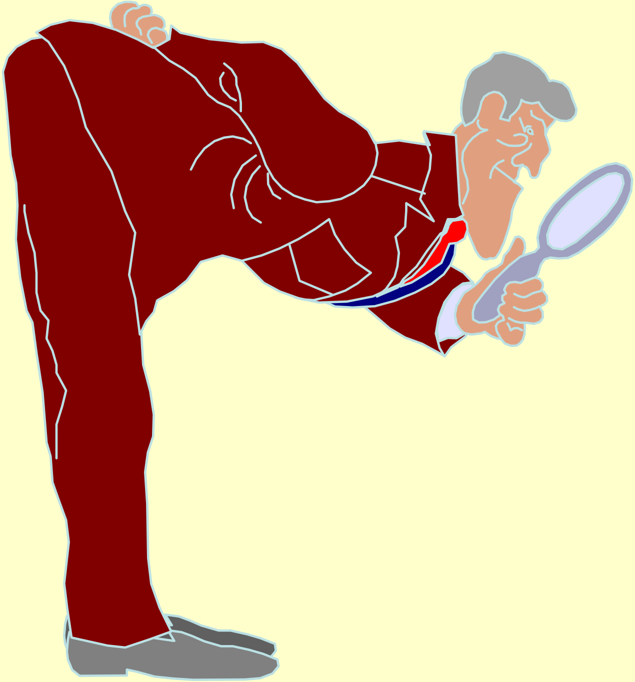
BIOSTATISTICS FOR TRANSLATIONAL & CLINICAL RESEARCH



Genetic Variation

& CANCERS

One man who drinks alcohol and smokes cigarettes lives to age 90 without getting liver or lung cancer; another man who smokes and drinks the same amount **gets cancer** at age 60.

One woman's breast cancer responds to chemotherapy, and her **tumor shrinks**; another woman's breast cancer shows no change after the same treatment.

How can these differences be explained?

Scientists think that tiny **variations** in the human **genome** called Single Nucleotide Polymorphisms, or **SNP**s (snips) for short, can help them to answer these questions.

They believe **SNPs can help them catalogue the unique sets of changes involved in different cancers**. They see **SNPs as a potential tool** to improve cancer diagnosis and treatment planning.

They suspect that SNPs may play a **role in the different responses to treatments** seen among cancer patients.

And they think that SNPs may also be involved in the **different levels of individual cancer risk** observed.

The human genome is the complete **set of instructions for life** as we know it. Except for red blood cells, which have no nucleus, the **human genome is located in the nucleus of every cell in the body**. There it is organized into 22 pairs of very **large molecules** called **chromosomes** and one pair of sex chromosomes (total: 23 pairs).

Chromosomes are made of deoxyribonucleic acid (DNA). **DNA contains only four chemical bases** or building blocks: Adenine, Thymine, Cytosine, and Guanine - called **A, T, G, and C**, for short. There are roughly 3.2 billion chemical bases (A, T, C, G) in the human genome.

Each DNA molecule is made up of two long complementary (related) **strands**, which are **intertwined** like a rope. This is the famous "**double helix**." Since **A always pairs with T**, and **C with G**, the order on one strand dictates the order on the other.

Only about **3 percent** of the human genome is **actually used** as the set of instructions. These regions are called **coding regions**, and they are scattered throughout the chromosomes.

Scientists can find no function for most of the remaining 97 percent of the genome - yet! These regions are called **non-coding regions**.

A **coding region contains genes**. A gene is a unique **DNA sequence** within a chromosome that ultimately directs the building of a specific protein with a specific function. Close to each gene is **a "regulatory" sequence of DNA**, which is able to **turn the gene "on" or "off."** There are **at least 35,000 genes** in the human genome, and there may be more.

An amazing aspect of the human genome is that there is so little variation in the DNA sequence when the genome of one person is compared to that of another. Of the 3.2 billion bases, **roughly 99.9 percent are the same between any two people.**

It is the variation in the remaining tiny fraction of the genome, **0.1 percent** -- roughly several million bases--that makes a person unique. This small **amount of variation determines attributes** such as how a person **looks**, or the **diseases** he or she develops.

Variation occurs whenever the **order of the bases** in a DNA sequence changes. Variations can involve only one base or many bases. If the **two strands of a chromosome are thought of as nucleotides threaded on a string**, then, for example, a string can break and the order of the beads can vary. One or more nucleotides may be **changed, added, or removed.** In chromosomes, **these changes are called polymorphisms, insertions, and deletions.**

In addition to these changes, some DNA sequences called **repeats** like to **insert extra copies of themselves several times**. Chromosomes can also undergo more dramatic changes called **translocations**. These occur when an **entire section of DNA on one chromosome switches (or Swaps) places with a section on another**.

Some of the variations that occur in the coding and regulatory regions of genes **have "harmless" effects.** They can, for example, change the way a person "looks." Some people have blue eyes, others brown; some are tall, others short; and some faces are oval, others round.

Other variations in coding regions are **harmless** because they occur in regions of a gene that do **not affect the function of the protein made.**

There are a group of variations in coding and regulatory regions that result in **harmful effects**. **These are called mutations. They cause disease because changes in the genome's instructions alter the functions of important proteins that are needed for health**. For example, diabetes, cancer, heart disease, Huntington's disease, and hemophilia all result from variations that cause harmful effects.

Finally, there are genetic variations that have **"latent" effects**. These variations, found in coding and regulatory regions, are not harmful on their own, and the change in each gene only becomes apparent under **certain conditions.** Such changes may eventually cause some people to be at higher risk for cancer, but **only after exposure** to certain environmental agents. They may also explain why one person responds to a drug treatment while another does not.

There is part of the genome from two people who are both smokers and drinkers, but only one of them gets cancer. The zoom into the chromosomes of these two men shows just a sampling of the differences in variation that are responsible for their individual cancer risk. **The variations themselves do not cause cancer. They only affect each person's susceptibility to tobacco smoke and alcohol after exposure.**

A SNP is defined as a single base change in a DNA sequence that occurs in a significant proportion (more than 1 percent) of a large population. The single base is replaced by any of the other three bases. Here is an example: in the DNA sequence **TAGC**, a SNP occurs when the G base changes to a C, and the sequence becomes **TACC**.

SNPs are scattered throughout the genome and are found in both coding AND non-coding regions. **SNPs can cause silent, harmless, harmful, or latent effects**. They occur with a very high frequency, with estimates ranging from about 1 in 1000 bases to 1 in 100 to 300 bases. This means that there could be millions of SNPs in each human genome. The abundance of SNPs and the ease with which they can be measured make these genetic variations significant.

Most SNPs occur in non-coding regions and do not alter genes. Scientists are finding that some of these SNPs have a useful function. If a SNP is frequently found close to a particular gene, **it acts as a marker for that gene**.

The remaining SNPs occur in coding regions. They could alter the **protein** made by that coding region, which in turn could influence a person's health. To understand how a SNP could do this requires a brief **look at proteins and their building blocks, amino acids.**

Amino Acids

Humans use **20 different amino acids** as building blocks to make the thousands of proteins found within the body. Each amino acid has the same basic structure with a **central carbon atom that has four groups attached**. Three of the groups are the same in every amino acid: one is a single hydrogen, one is an amino group that **acts like a hook**, and one is a carboxyl group that acts like an eye. **The fourth group is a side chain, each side chain has a unique size and shape which is unique for every amino acid.**

When two amino acids are lined up end-to-end, and the hooks and eyes are allowed to react together, a peptide bond forms. As more amino acids attach, a chain begins to form that has a strong but flexible backbone with side chains sticking out at regular intervals. **This is protein-building in progress.**

Base pairing is used to create mRNA from one of the DNA strands. **Every T in the DNA strand pairs with an A in the mRNA strand, and Cs and Gs in the DNA pair with Gs and Cs in the mRNA**. However, if there is an A in the DNA sequence, it pairs with a new base called Uracil (U) in the mRNA strand. Thus, if **the DNA sequence is TACGCAATATGCATT, the mRNA sequence becomes AUGCGUUAUACGUAA.**

The sequence of the bases of an mRNA (A, U, G, C) directly spells out the sequence of amino acids in the protein. **Every three bases** in the mRNA sequence codes for a single amino acid and is called a “codon”; some amino acids have more than one codons.

The 3-D structure determines the function of the protein. **When there is a change in one or more amino acids, then the ability of the protein to function may be affected.** The protein's function may be unchanged or it may become sluggish, hyperactive, or inactive.

And SNPS lead to changes in codons which, in turn, lead to changes in amino acids.

When a SNP occurs, even within a coding region, **it is possible that there is no effect**. For example, when the DNA sequence **GAC** becomes **GAG**, the codon in the mRNA changes from CUG to CUC. Since **both codons stand for the same amino acid** leucine (different codons in the same acid), **there is no change in the protein.**

The change in the DNA is called a silent change.

There are also SNPs within coding regions that lead to **very subtle changes in proteins**. For example, if a SNP causes the codon GAU to change to GAG, the amino acid changes from aspartic acid to glutamic acid. **These amino acids have very similar chemical properties**, but glutamic acid is just a little bigger. If the SNP causes a change in a part of the protein that is not important to its function, the result may be very subtle or totally harmless.

The cell continues to function normally.

In another change for codons, the sequence GUA becomes GUU, and the amino acid called aspartic acid changes to valine in the protein. **These two amino acids have very different chemical properties.** The substitution of one for the other may, or may not, severely alter how the protein folds and functions, **depending on where the change occurs in the protein**. **When the change in protein lead to disease symptoms in the patient, SNP is harmful and is called a mutation.**

An example of a SNP causing disease is found in sickle cell anemia. The change in one nucleotide base in the coding region for the hemoglobin beta gene causes the amino acid **glutamic acid** to be replaced by **valine**. As a result, the hemoglobin molecule can no longer carry oxygen as efficiently because the structure of the protein is changed dramatically.

SNPs may cause subtle changes in a group of genes that under normal conditions are latent, i.e., they are switched "off." But when a person is exposed to precarcinogens or carcinogens, they can be switched "on."

Since the proteins from these genes regulate how fast or how slowly the harmful agents are absorbed, bound, metabolized, and excreted from the body, even a small or subtle change in any one of them may alter a person's risk for cancer.

It is important to remember that the SNP itself does no harm under normal circumstances. When a person is exposed to an environmental agent that is carcinogenic, does the SNP exert an influence (**exposure activates the “switch”**).

When a person smokes cigarettes, **precarcinogens** in the form of tobacco smoke enter a person's lungs and lodge in the fat-soluble area of the cells. They become bound to **proteins that convert the precarcinogens into carcinogens**. These reactive molecules are quickly handed over to detoxifying proteins that make the carcinogens water-soluble and **allow the body to eliminate them into the urine**, before they can damage the cell.

Because of SNPs, a person's genome may express a very active carcinogen-making protein, or a sluggish one, or something in between.

A protein with a very active binding site can "grab" more precarcinogen than usual. Or the protein may convert the precarcinogen to the carcinogen at a faster rate. In both cases, more carcinogen molecules pile up in the lungs, causing damage to the cells' DNA, which can lead to cancer.

On the other hand, if the SNPs result in a slow carcinogen-making protein, the lung is exposed to fewer DNA-damaging carcinogens, and the chance of cancer is reduced.

SNPs may also influence detoxifying enzymes that prepare carcinogens for elimination from the human body. SNPs that result in very active forms of detoxifying enzymes will remove the carcinogens quickly from the body, allowing less time for damage.

SNPs that produce sluggish detoxifying enzymes will permit carcinogens to remain in the body for a longer time.

Some workers in the dye industry, after occupational exposure to arylamines, develop an increased risk of bladder cancer. Scientists suspect that SNPs may be involved; more complicated case involving more than one gene.

SNPs may also explain why some patients respond well to a specific drug treatment, while others have minimal or no response.

SNPS may also be involved when patients have different side effects in response to the same drug.

Many proteins interact with the drug - involved in its transportation throughout the body, absorption into tissues, metabolism into more active forms or toxic by-products, and excretion. If a patient has SNPs in any one or more of these proteins, they may alter the time the body is exposed to active forms of the drug or any of its toxic byproducts. **Some of these are not yet discovered!**

Scientists are trying to identify all the different SNPs in the human genome. They are sequencing the genomes of a large number of people and then comparing the base sequences to discover SNPs. The sequence data is being stored in computers that can generate a single map of the human genome containing all possible SNPs.

STATISTICAL ISSUES

- Two major approaches/ procedures are:

(1) Analysis of DNA sequence data

(2) Statistical analysis of marker associations; mostly in the form of Case-control SNP-disease association

- Major issues? Multiple-decision problem

SNP-disease association analyses

Case-control analyses are based upon the use of contingency tables and unconditional logistic regression adjusted for age and sex . It is important to restrict the type 1 error in genetic studies and to account for multiple testing issues; however, Multiple testing issues are ignored in power calculations.

Each SNP will be classified in two ways:

(1) As a three-level variable (homozygous variant versus heterozygous vs. homozygous wild type). **Example**: CC versus CG versus GG

(2) As a two-level variable (homozygous variant vs. heterozygous/homozygous wild type). **Example**: CC versus {CG, GG}

“Normality/ wild type” (GG) determined by frequency (largest).

Then use (a) Chi-square test (DiseasexSNP: 2x2 or 2x3 contingency tables) or (b) Logistic regression (with 1 or 2 dummy variables for each SNP)

DNA sequence data

Hardy-Weinberg equilibrium was tested at each SNP locus on a contingency table of observed versus predicted genotype frequencies using a modified Markov-chain random walk algorithm [Guo and Thompson, 1992].

Pairwise linkage dis-equilibrium between each pair of SNP loci was analyzed using a likelihood-ratio test, whose empirical distribution was obtained by a permutation procedure [Slatkin and Excoffier, 1996].

Maximum likelihood haplotype frequencies were imputed within each group of subjects using an Expectation-Maximization (EM) approach [Excoffier and Slatkin, 1995], as implemented in the program Arlequin v2.0 which could be used to manage and analyze the data. Software is available from:

http://anthro.unige.ch/software/arlequin/

Hardy-Weinberg Equilibrium: The stable frequency distribution of genotypes, AA, Aa, and aa, in the proportions p^2, 2pq, and q^2 respectively (where p and q are the frequencies of the alleles, A and a) that is a consequence of random mating in the absence of mutation, migration, natural selection, or random drift.

References:

Guo SW, Thompson EA. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* **1992**;*48*:361-372.

Slatkin M, Excoffier L. Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm. *Heredity* **1996**;76:377-383.

Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genet* **1995**;11:241-247.