Do You Speak Genomish?

Holger Schwender, Sylvia Rabstein, and Katja Ickstadt

Writing this article in German and then translating it into English with a software tool such as Babel Fish would not be the best route to winning a Pulitzer. Even if the best translation software is used, some information would get lost. The same applies to statistical analyses of genetic data: Without having at least a little knowledge of the genetic terminology and concepts, the analysis might fail—even if done carefully. So, we need to understand the biologists, and they need to understand us. Unfortunately, this is not as easy as simply using Babel Fish to translate "Genomish" into your language of choice.

Genomics 101

The Human Genome

In the human body, there are billions of cells. The nucleus of almost every cell comprises the complete human genome. (That's why we belong to the eukaryotes, not the prokaryotes, whose cells do not have a nucleus.) The human genome is the blueprint for all cellular structures and activities in the human body. It consists of 23 pairs of chromosomes. In each pair, one chromosome comes from the mother and one from the father. While 22 of the chromosome pairs—the autosomes—are the same for both genders, there is one pair consisting of the two gender chromosomes (or genosomes)—X and Y—that differ between man and woman. Females carry two X chromosomes, whereas males have both an X and a Y chromosome.

Each chromosome is a huge chain of two intertwined strands of deoxyribonucleic acid (DNA), the double-helix. These DNA strands are sequences of nucleotides, where each nucleotide consists of a phosphate...
The DNA group, a deoxyribose sugar, and one of the four nitrogen bases—adenine (A), thymine (T), cytosine (C), and guanine (G). That's why DNA often is represented as a sequence of the four letters A, T, C, and G.

One sequence suffices to describe DNA, even though there are two strands. The reason for this is **complementary base-pairing**: A on strand one is always connected (via hydrogen bonds) to T on strand two, whereas C is always connected to G. So, if we know the sequence of one of the strands, we also know the sequence of the other strand. The leading end of each of these strands is called 5' end; the tail end is called 3' end. As they are complementary, one strand goes from 5' to 3', and the other from 3' to 5'.

Segments of the DNA, namely genes, contain construction information for proteins, which consist of amino acids. Because proteins are responsible for the structure and activity of a cell, hence for virtually everything that happens in an organism, it is important to understand how genes are translated into proteins. This is explained by the Central Dogma of Molecular Biology:

- **Transcription**
  - DNA
  - mRNA

The relationship between the codons and amino acids is known as the genetic code; the total process of converting a DNA sequence into a protein is known as **gene expression**.

Does this mean a gene always codes for the same protein? No, that would be too easy and might be an evolutionary disadvantage. For the translation step, only small parts of the genes—the exons—are needed. That's why in the transcription step, the introns (i.e., the noncoding parts of genes) are removed by a process called RNA-splicing. However, the same exons are not always retained. So different combinations of exons can be spliced to produce different mRNA isoforms of a gene that can lead to different proteins. It is believed that about 60% of genes are affected by this alternative splicing.

Noncoding regions, however, occur not only within the genes, but also between the genes. Even though these noncoding regions often are referred to as junk DNA, parts of them—namely the regulatory or **promoter** regions right before the genes—are assumed to play a role in gene regulation. Proteins called **transcription factors** can activate or repress the expression of a gene when they bind on small DNA sequences, called motifs, in the regulatory region of that gene.

Alternative splicing is not the only reason for a single gene to code for
different proteins. DNA, itself, differs between humans. Even though far less than 1% of the DNA varies between humans, these are still millions of base pair positions.

Genetic Variation

There are several forms and levels of genetic variation, including deletions or substitutions of bases and translocations of large segments of the chromosomes, or even changes in the number of chromosomes. Any of these changes is referred to as mutation, variant, or variation.

Some of the variations can affect your phenotype (e.g., how you look, how fast you metabolize the coffee you drank during breakfast, or your ability to roll your tongue). The phenotype may depend on not only the mutation, but also the number of affected chromosomes. If a mutation is dominant, a change in just one chromosome is sufficient to cause a different phenotype. But if it is recessive, both chromosomes must be affected to change the phenotype.

Each of the several forms a DNA segment can take is called an allele. If a population is analyzed with respect to genotype, then the combination of two alleles—one from each chromosome—is of interest. If this analysis is based on a specific locus in the DNA sequence, then the minor allele is the allele that occurs less often in the population. If the frequency of the minor allele is larger than 1%, the variant is called polymorphism. (This condition is necessary to distinguish inherited variation from spontaneous mutations.)

The most common—and, so far, the best investigated—type of genetic polymorphism is the single nucleotide polymorphism (SNP, pronounced “snip”). A SNP is characterized by the possibility of (usually two) different bases at one specific locus. However, deletions and insertions of several bases and repeats or inversions of small segments of the sequence also are called SNPs. Information about several million SNPs in different populations is available freely on databases such as dbSNP (www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=snp).

As an example of a SNP, imagine that most people have the base thymine (T) at a particular locus, but a few have adenine (A). In this case, the SNP can take either the homozygous reference genotype TT, the heterozygous variant TA or AT, as it is not possible to distinguish between the two chromosomes, or the homozygous variant AA.

Inherited variations in just a single gene can cause a disease. These diseases are different from relatively harmless (e.g., color blindness) to lethal (e.g., Tay-Sachs disease). For complex diseases and traits, such as cancer, it is, however, necessary to look at several genes and their interactions to understand their influence on the phenotype, as such diseases are caused by many factors, including genetic variants (and probably environmental factors as well).

Analyzing several genes at a time leads to consideration of haplotypes, which are combinations of alleles on the same chromosome. Assume, for example, that two loci on the same chromosome are analyzed and both have two possible alleles—(A1, A2) and (B1, B2)—respectively. Then, one question of interest is whether allele A1 of locus A is always inherited with allele B1 of locus B, or whether they are inherited independently. Thus, for haplotypes, the evaluation of the correlation structure of the alleles is of particular interest.

At this point, the process of recombination, i.e., DNA exchange between two chromosomes during the meiosis, which is the basis of sexual reproduction, plays an important role. If recombination occurs often, it is likely the alleles of several loci on a chromosome are independent from each other. If recombination between the loci, however, does not occur often, then the allele combinations in a population are not independent. This association of alleles is called linkage disequilibrium (LD). A goal of projects such as HapMap (www.hapmap.org) is to find recombination hotspots (i.e., places along the chromosomes that have a high recombination rate).

Genomic Data 101

As mentioned above, almost every cell in a human organism contains a complete set of the human genome. Thus, the same repertoire of proteins is included in each of these cells. The properties of the cells, however, are very different. This distinction is based on differences in the abundance, the distribution, and the state of the proteins. To understand the function of a cell, one must measure the abundance of the proteins in the cell. Even though there are methods for monitoring protein expressions, such as 2D gels and protein microarrays, measuring proteins is much more complex than measuring the levels of mRNA. Moreover, following the Central Dogma of Molecular Biology, the function of a cell also can be investigated by measuring the mRNA levels if the assumption behind DNA microarrays—that most of the mRNA is translated into proteins—holds. So, we will focus our interest on the analysis of DNA microarrays.
Table 1—Comparison of cDNA and Affymetrix Microarrays

<table>
<thead>
<tr>
<th>cDNA Microarray</th>
<th>Affymetrix Microarray</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two color chip</td>
<td>One color chip</td>
</tr>
<tr>
<td>Two target samples per array</td>
<td>One target sample per array</td>
</tr>
<tr>
<td>One probe per gene</td>
<td>11 probe pairs per gene</td>
</tr>
<tr>
<td>Probes of varying length (typically &gt; 500)</td>
<td>Probes are 25-mers</td>
</tr>
</tbody>
</table>

Measuring Gene Expression

DNA microarrays measure the expression levels of (tens of) thousands of genes simultaneously. The most widely used type of microarray are the Affymetrix GeneChip and the cDNA microarray. See Table 1 for a comparison.

The basic idea of most microarrays is complementary base-pairing. To measure the amount of a specific mRNA, and thus the expression level of a specific gene, one synthesizes a batch of the corresponding complementary (single-stranded) DNA probe and places it on the microarray. A stable mRNA-DNA hybrid will be formed when the mRNA comes into contact with the probe. If this mRNA has been labeled previously with fluorescent dye, the amount of hybridization can be measured by the intensity of fluorescence, which is assumed to be proportional to the abundance of the specific mRNA.

The Affymetrix GeneChip consists of a small glass plate (typically 0.5" by 0.5"), comprising hundreds of thousands of probe cells. In almost all of these probe cells, millions of copies of a specific oligonucleotide probe (abbreviated "oligo") are placed on the surface. Each oligo on an Affymetrix chip is a 25-mer (i.e., a chain of 25 bases). Each gene is represented typically by 11 pairs of such probes. Each of these pairs consists of a perfect match (PM) oligo, for measuring the abundance of the mRNA of interest, and a mismatch (MM), for adjusting the PM intensity for background noise and nonspecific binding.

After labeling the mRNA with fluorescent dye, it is passed over the array and stimulated with a laser. After a time, the material that has not bound to the surface is removed. What is left is the hybridization of the mRNA sample to the oligos on the microarray. At each spot, the (relative) abundance of the genetic material can be measured now by the intensity of the fluorescence. (More about the idea behind PMs and MMs and how the 22 probe intensities are summarized to one gene expression value can be found in the article starting on Page 15).

The cDNA microarray consists of a glass plate with thousands of spots, each corresponding to a specific mRNA sequence. The expression value of a gene is measured using competitive hybridization of two mRNA samples to the sequences on the chip. One of these samples is a control, typically identical for all chips. The other is the sample of interest, such as a tumor sample. However, instead of measuring mRNA directly, this technology measures complementary DNA (cDNA) because of the higher stability of DNA over RNA. cDNA is a long chain of single-stranded DNA usually consisting of more than 500 bases obtained from a full-length string of mRNA via reverse transcription.

After reverse-transcribing the two mRNA samples into cDNA, one of the samples is labeled with green fluorescent dye, and the other with red dye. They are then mixed and applied to the microarray. After the cDNA of a particular gene has hybridized to its particular spot, the microarray is imaged twice—one image for each dye—using a scanner. For each spot, the log2-ratio of the two images (i.e., of the red and green intensities) typically represents the relative expression of the corresponding gene.

For measuring gene expression, knowledge of the DNA sequence is needed. This information is available for some species, such as humans. However, for many other species—such as plants—little or nothing is known. If a genetic study focuses on such a species, it is necessary to first gain genetic information about it. Altmann et al. discuss the aspects that have to be considered in such a study in the article beginning on Page 27.

Measuring Single Nucleotide Polymorphisms

Like genes, SNPs can be measured using microarrays. Chips such as the Mapping 500K Array Set of Affymetrix can genotype tens or even hundreds of thousands of SNPs. However, here we would like to focus on another method for measuring SNPs which can also be applied to measure gene expression. This method called polymerase chain reaction (PCR) has enabled the boom in molecular genetics since the early 1980s. PCR is a technology to amplify (i.e., produce multiple copies of) a specific DNA sequence to get enough genetic material to analyze with standard biochemical procedures. In each of the typically 20 to 30 cycles of a PCR process, the amount of DNA is doubled. The original amount of the DNA sequence is thus amplified by a factor of 2^n to 2^30. Each cycle of the PCR takes about five minutes and consists of three steps:

1. Denaturation. The DNA double strand is separated by heating it to a temperature of 94°-96° Celsius.
2. Annealing. The temperature is decreased slowly to 50°-60° Celsius so that two specific primers—short oligos that match the beginning and end of the DNA sequence of interest—are able to anneal to the single-stranded DNA sequences.
3. Extension. DNA components (i.e., the four types of DNA nucleotides) and a DNA polymerase—a special enzyme that catalyzes the synthesis of DNA—are added to the probe. The reagent is then heated to 72° Celsius to enable the synthesis of two copies of the specific DNA sequence.

If the genotyping of a SNP is of interest, the two primers are annealed near the known location of the SNP, one on each side. After amplifying the DNA sequence of interest, another method is needed for evaluating the genotype of the sample. Many methods have been derived to detect which DNA sequences have been amplified by the PCR. As an example, a method called TaqMan is presented that—contrary to other procedures—simultaneously amplifies a specific sequence and measures the amount of this specific DNA sequence. For this, a third oligo—besides the two primers—complementary to the target sequence is added to the PCR mixture. This oligo is labeled with two fluorescent dyes, the reporter and the quencher, one at each end of the oligo. As long as the third oligo is annealed (between the two primers) to the target sequence, the reporter and quencher are close to each other and the quencher absorbs the fluorescence emission of the reporter. But if the region targeted by the two primers is reached by an enzyme—the Taq DNA polymerase—during the copying process, both the reporter and quencher are detached from the DNA sequence. Because they are not in proximity anymore, the reporter’s fluorescence is not absorbed. The intensity of the fluorescence is proportional to the amount of detached reporter and therefore measures the amount of the DNA sequence for which it is specific.

If the genotype of a specific SNP is being investigated, two allele-specific oligos, one for each of the two possible bases, with different fluorescent dyes will be added to the PCR mixture and the fluorescence intensities will shed light on the genotype of the specific sample.

The name TaqMan refers to the heat-resistant Taq DNA polymerase generally used in the PCR processes on the one hand and to the popular 1980s computer game Pac-Man on the other, as the bases of the TaqMan oligo are eaten by the Taq DNA polymerase just as the dots in the game are eaten by Pac-Man.

Genomic Data Analysis 101

Analyzing Gene Expression Data

The most common and most interesting tasks in the analysis of microarrays are:

- The identification of differentially expressed genes (i.e., genes whose expression levels differ strongly between several groups [e.g., cancer vs. non-cancer])

- Finding genes whose expression levels vary together (i.e., clustering genes)

- Constructing a classification rule based on gene expression data for assigning new observations to a class (e.g., cancer or noncancer)

Before one can start with any of these high-level analyses, it is necessary to preprocess the data generated in the microarray experiment. On Affymetrix microarrays (e.g., each gene typically is represented by a set of 11 pairs of probes), the goal of the low-level analysis (i.e., the preprocessing) is to reduce the 22 probe intensities to one value for each gene on each chip.

Affymetrix provides an algorithm called Plier (Probe Logarithmic Intensity Error estimation) for performing this data reduction. There also exist academic preprocessing methods, such as RMA (Robust Multiarray Analysis) and PLM (Probe Level Models). These preprocessing procedures are compared in the article starting on Page 15.) Many other competing algorithms are available (e.g., MBEI [Model-Based Expression Indices], which was the first academic alternative to the Affymetrix software, and GCRMA, a further development of RMA that takes the GC-content [i.e., the amount of the bases guanine and cytosine in the probes], into account).

The correct measuring of the probe intensities can be obscured by both SNPs and alternative splicing. Here, the really interesting (and still open) question is not how these genetic variations influence the probe intensities, but how they influence the gene expression values and the detection of differentially expressed genes.

The identification of differentially expressed genes—one of the high-level analyses of interest—is a massive multiple testing problem in which thousands, or even tens of thousands, of tests—one for each gene—are performed simultaneously. Historically, a typical multiple testing problem consisted of at most five to 20 tests, and controlling the probability of one or more false positives (i.e., the family-wise error rate [FWER]) at a prespecified level of significance by a procedure such as the Bonferroni correction was reasonable. But when several thousands of null hypotheses are tested, controlling the FWER is usually much too conservative. The goal of multiple testing in microarray analyses is to identify a fairly large number of interesting genes. It does not matter if a few of these genes are false positives as long as their number is relatively small compared to the total number of identified genes. In this case, the false discovery rate (FDR) seems to be a more reasonable error measure, as the FDR is, loosely speaking, the expected proportion of false positives among all genes identified as differentially expressed. Another idea for an error measure in massive multiple testing is to control the probability that the proportion of false positives among all detected genes is less than a prespecified threshold at a prespecified level of significance. This error measure is known as the tail probability of the proportion of false positives (TPPFP).

Following the paradigm of multiple testing, the level of significance at which an error rate should be controlled has to be specified prior to any testing. However, this can lead to complications. Setting the level of significance (e.g., to $\alpha = 0.05$) might result in the identification of too few or too many genes. Alternatively, fixing the rejection region (e.g., so that 250 genes are identified) can lead to an undesirably large error rate. Methods that attempt to find a balance between the number of identified genes and the controlled error rate are, for example, the significance analysis of microarrays (SAM) and the empirical Bayes analysis of microarrays.

In addition to these univariate variable selection methods, there are also multivariate procedures for analyzing
gene expression. One of the most popular approaches is principle component analysis (PCA), which can be used to find genes whose expression levels vary together. A procedure, called gene shaving that makes use of this feature of PCA can be applied in either a supervised or unsupervised way (i.e., information about the class membership of the observations [e.g., cancer vs. non-cancer] can be included or not). Two-way hierarchical clustering is another popular method for clustering genes and observations simultaneously. In this issue, Haiyan Huang and Kyungil Kim apply PCA, hierarchical clustering, and a modified version of the nonhierarchical clustering method K-Means to gene expression data and discuss advantages and disadvantages of these procedures in the article beginning on Page 49.

Genes, however, cannot only be clustered by their expression but also based on the motifs in the regulatory region to which the transcription factors bind. Erin M. Conlon et al. show how clusters of genes co-regulated by the same transcription factors can be found and presented as genetic regulatory networks in the article starting on Page 45. They also compare these networks with pathways (i.e., molecular interaction networks in biological processes).

Having reduced the number of genes from several thousand to several tens or a few hundreds, usual suspects such as Random Forests and Support Vector Machines (SVMs) are applied often to gene expression data to construct a classification rule. Usually only a subset of the genes is needed to build a classifier, so such classification methods also can be used to recursively eliminate genes by computing a measure of importance and by deselecting either the most unimportant gene or a specified percentage of the genes. For other procedures, such as partial least squares (PLS) or the prediction analysis of microarrays (PAM), no previous data reduction is necessary, as they can handle a vast number of variables directly. PAM is explained more thoroughly by A. Wise et al. in their article beginning on Page 39.

Analyzing Genotype Data

The goals of the analysis of SNP data are similar to the goals of the analysis of gene expression data:

- SNPs and SNP interactions whose distributions differ substantially between several groups should be identified.
- The class of new observations should be predicted by a rule based on the SNPs.
- SNPs showing a coherent pattern should be found.

Not only are the goals the same, but the methods are also similar. For example, all the procedures for multiple testing described in the previous section can be applied analogously to SNP data. Only the test statistic and its (estimated) null distribution need to be changed, as SNP data are categorical, whereas gene expression data are continuous.

The same applies to hierarchical clustering. Cluster analysis of DNA microarray data typically is based on Euclidean distance. But for SNP data, the distance measure has to be able to handle categorical data. Thus, typically matching coefficients or measures based on Pearson's χ²-statistic are used.

As mentioned above, susceptibility to disease is determined often by combinations or interactions of SNPs, rather than by individual SNPs. So, suitable classification methods attempt to find such SNP interactions. It has turned out that logic regression, which bases its predictions on Boolean combinations of binary variables, is an interesting discrimination method for the analysis of SNP data. An application of hierarchical clustering and logic regression to a case-control study of breast cancer is presented in the article by Katja Ickstadt et al., beginning on Page 21.

SNPs and other genetic variants cannot only be used to predict the outcome of binary variables, such as the presence of a disease, but also the outcome of continuous traits. Locations on the DNA sequence that influence such traits are called quantitative trait loci (QTL). The article by Susan J. Simmons and Ann E. Stapleton, beginning on Page 11, addresses the identification of QTLs using Bayesian hierarchical regression.

Another interesting task is to analyze the dependencies between several SNPs on the same chromosome by inferring the haplotypes in the study population. The major problem in the analysis of haplotypes is that the allelic information for each SNP does not tell which chromosome carries which base (i.e., the genotype of the SNP is known, but for none of the bases is it known to which of the two chromosomes it belongs). With two or more heterozygous SNPs, one needs to deduce which of the bases are carried by the same chromosome. Approaches such as the EM algorithm or an algorithm called Phase can be used to estimate the haplotype frequencies and the most likely haplotype pairs for each individual.

More to Come

Congratulations! You have just finished reading an article with lots of genetic stuff in it. We have given you a wild ride through the basic terminologies and concepts of genetics and a taste of the statistical procedures used to analyze different types of genomic data. We hope our appetizer has been delicious enough for you to want to continue reading the remaining articles in this special issue. The authors of the following articles give more insight into some of the methods mentioned only briefly here.

As there are almost as many papers on the statistical analysis of genomic data as there are bases in the human genome (which might be a little bit underestimated), neither this article nor the whole issue can explain all the genetic concepts and introduce all the statistical procedures applicable to genomic data. We have therefore selected a small set of basic concepts and methods and have explained them briefly.

At the beginning of this article, we said biologists and statisticians speak different languages. In our collaboration with biologists, we have seen that it takes a while for a statistician to understand biologists, and vice versa. But at the end of the day, speaking the other language leads to very fruitful and exciting collaborations. We get much insight into genetics and biotechnology, and biologists understand that statistics plays an important role in genomics and will have an important role to play in the future.

References

References available online at www.amstat.org/publications/chance.
**Glossary**

**Adenine (A).** One of the four bases in DNA. The other three bases are cytosine (C), guanine (G), and thymine (T). Adenine is always paired with thymine (see complementary base-pairing).

**Affymetrix GeneChip.** One of the most widely used DNA microarrays produced by Affymetrix. Consists of hundreds of thousands of probe cells, each containing several million copies of a specific oligonucleotide probe.

**Allele.** One of the several forms a DNA segment (e.g., a gene) can take.

**Carcinogen.** Any substance or agent that promotes cancer.

**cDNA.** Acronym for complementary DNA. Single-stranded DNA that is complementary to messenger RNA or DNA that has been synthesized from messenger RNA by reverse transcriptase.

**cDNA Microarray.** A DNA microarray that comprises thousands of spots, each corresponding to a specific mRNA sequence. A mixture of single-strand DNA, complementary to the mRNA of two samples (usually one from a patient and one from a control), is hybridized to the cDNA microarray. The log ratio of the two fluorescent dyes with which the two samples have been labeled usually is used as the expression value of the gene.

**Chromosome.** A large, threadlike macromolecule in the cell nucleus that carries the genes in a linear order. The human genome consists of 23 pairs of chromosomes; in each of these pairs, one chromosome comes from the mother and the other from the father.

**Coding Region.** The portion of DNA that is transcribed into mRNA and translated into proteins.

**Comparative Genomics.** The study of relationships between the genomes of different species or strains.

**Complementary Base-Pairing.** The two complementary strands of DNA are connected via hydrogen bonds between base pairs: A is always paired with T, and G is always paired with C.

**Complex Disease.** A disease that is caused by many genetic (and possibly environmental) factors.

**Complex Trait.** A genetically inherited feature of an individual, caused by many genetic (and possibly environmental) factors.

**Cytosine (C).** One of the four bases in DNA. The other three bases are adenine (A), guanine (G), and thymine (T). Cytosine always is paired with guanine (see complementary base-pairing).

**Diploid.** A genome is diploid when it contains two copies of each chromosome. The human genome is diploid.

**DNA.** Acronym for deoxyribonucleic acid. DNA carries the genetic instructions, in very long sequences of nucleotides, for making living organisms. Two long strands of DNA in the form of a double helix make up each chromosome.

**DNA Microarray.** A small glass slide that comprises thousands—or even hundreds of thousands—of spots, or probe cells. Each of these spots contains specific genetic material for measuring the expression of a single gene. The most prominent examples are the Affymetrix GeneChip and cDNA microarrays.

**Dominant.** A mutation is called dominant if a change in just one of the two chromosomes that build a pair is sufficient to cause a different phenotype.

**Epistasis.** The suppression of a gene by the effect of an unrelated gene, or the interaction of two or more genes to produce an effect different from the effect of the individual genes.

**5' End.** (Read as “five-prime end.”) The leading end of each of the DNA strands. The conversion of a DNA sequence into a protein is processed from the 5' end to the 3' end.

**Gene.** A segment of DNA involved in producing a polypeptide chain. It can include regions preceding and following the coding DNA as well as introns between the exons. Considered a unit of heredity.

**Gene Expression.** The total process of converting a DNA sequence into a protein.

**Genome.** The full DNA sequence of an organism, constituting a blueprint for all cellular structures and activities in that organism. Virtually all cells of an organism contain a copy of the complete genome.

**Genotype.** The genetic constitution of an organism specifying the particular alleles at defined loci in the genome. Can refer to one gene, a specific group of genes, or the entire set of genes within an organism.

**Guanine (G).** One of the four bases in DNA. The other three bases are adenine (A), cytosine (C), and thymine (T). Guanine always is paired with cytosine (see complementary base-pairing).

**Haploid.** Characteristic of a cell containing only one set of chromosomes.

**Heterozygous.** Characteristic of a genotype having two alleles that differ from each other at corresponding chromosomal loci.

**Homozygous.** Characteristic of a genotype having both alleles identical to each other at corresponding chromosomal loci.

**Hybridization.** The process of joining two complementary strands of DNA.
Linkage Disequilibrium. The nonrandom association of alleles at two or more loci.

Locus. Plural: loci. A location (i.e., a single base-pair position) in the DNA sequence.

Minor Allele Frequency. The frequency of the less frequent allele.

Mismatch (MM). An oligonucleotide probe used on an Affymetrix microarray to adjust the corresponding perfect match (PM) for background noise and nonspecific binding. Each MM oligo is a sequence consisting of 25 bases, which is almost identical to the sequence of the corresponding perfect match. Only the thirteenth base of the MM is complementary to the thirteenth base of the corresponding PM.

mRNA. Acronym for messenger RNA. One specific type of RNA into which the information in the region of a gene that codes for a protein is transcribed.

Mutation. Any structural change in the DNA (e.g., deletions or substitutions of bases as well as translocations of large segments of the chromosomes). Synonymous to variation or variant.

Nonspecific Binding. Hybridization of a mRNA or cDNA sequence to a spot that actually corresponds to a different mRNA or cDNA sequence.

Nucleotide. The DNA is built up by sequences of thousands of nucleotides. Each nucleotide consists of a phosphate group, a deoxyribose sugar, and one of the four bases—adenine (A), thymine (T), cytosine (C), and guanine (G).


Ortholog. Genes in different species that evolved from the same ancestral gene.

Pathway. A molecular interaction network in a biological process.

Perfect Match (PM). An oligonucleotide consisting of 25 bases used on Affymetrix GeneChips to measure gene expression. Each PM is used with its corresponding mismatch (MM).

Phenotype. The observable characteristics of an organism, resulting from genetic characteristics influenced by a particular environment.

Polymorphism. A variation occurring in more than 1% of a population.

Probe. A piece of labeled RNA or DNA used to measure the expression of a gene.

Probe Cell. A location on a microarray that typically contains millions of copies of a specific probe.

Probe Set. A set of probes that represents a particular gene. On Affymetrix GeneChips, a probe set typically consists of 11 pairs of perfect matches and mismatches.

Promoter. The region of a DNA molecule to which RNA polymerase binds to initiate the process of transcription (i.e., the synthesis of RNA whose sequence is determined by the sequence of the DNA adjacent to the promoter site). Also, the sequence of bases in the DNA at such a promoter site.

Protein. A very large, complex molecule. Proteins are responsible for virtually everything that happens in an organism.

Recessive. A mutation is called recessive if both chromosomes must be affected to cause a change in the phenotype.

Regulatory Sequence. A promoter, enhancer, or other segment of DNA where regulatory proteins such as transcription factors bind preferentially. They control gene expression and, thus, protein expression. Regulatory sequences also can be found on messenger RNA, but they generally are not as well-studied as those in DNA.

RNA. Acronym for ribonucleic acid. A chemical similar to a single strand of DNA but carrying a different sugar molecule (ribose instead of deoxyribose) and a different base (uracil [U] instead of thymine).

SNP. Acronym for single nucleotide polymorphism (referred to as “snip”). A polymorphism occurring at a single base-pair position.

Spot. A location on a DNA microarray.

Susceptibility. The responsiveness to become infected by a disease.

3’ end. (Read as “three-prime end.”) The tail end of each of the DNA strands. The conversion of a DNA sequence into a protein is processed from the 5’ end to the 3’ end.

Thymine (T). One of the four bases in DNA. The other three bases are adenine (A), cytosine (C), and guanine (G). Thymine always is paired with adenine (see complementary base-pairing).

Trait. A genetically inherited feature of an individual.

Transcription Factor. A protein that binds to DNA at a specific promoter or enhancer region or site where it regulates transcription. Transcription factors can be activated selectively or deactivated by other proteins, often as the final step in signal transduction.

Variation. Synonym for mutation.

Variant. Synonym for mutation.