

white paper

SNPs – Powerful Tools for Association Studies

August 2003

Sequence information from the completed human genome projects vastly increased our understanding of the human genome. Empowered by the expanded availability of genomic information, genetic studies have become an increasingly effective tool for identifying factors that influence disease susceptibility.

Physical differences between individuals have a strong genetic component. The goal of genetic studies is to identify specific genetic profiles that are associated with a disease (or trait) of interest. Although accurate phenotyping is essential for successful genetic studies, the availability of biological tools for genetic analysis is also necessary. In this document, we will discuss the evolution of the markers used to profile genetic differences, as well as explore the multiple approaches used in modern-day genetic studies. This analysis will clarify why single nucleotide polymorphisms (SNPs) are powerful tools for studying complex diseases or traits.

1. The evolution of genetic markers in scientific research

Scientific advancements have resulted in a series of genetic markers with ever-increasing information content and resolution power. In the past 30 years, restriction fragment length polymorphisms (RFLPs), short tandem repeats (STRs), and SNPs have played significant roles in genetic research.

Restriction fragment length polymorphisms

RFLPs were the first genetic markers to be widely used in genetic mapping studies. To identify RFLPs, genomic DNA is digested into small pieces, averaging several kilobases (kb), by restriction enzymes that recognize specific sequences (i.e., restriction sites) in the genome. When genomic variation between individuals eliminates or adds a restriction site it causes a change in the length of the

resulting restriction fragments that can be used as a genetic marker. Any change in the DNA has the potential to cause RFLPs, including SNPs, DNA insertions, and DNA deletions.

The usefulness of RFLPs for mapping genetic diseases was serendipitously discovered by researchers who were looking for the gene that causes sickle cell anemia. They discovered that when they used the restriction enzyme HpaI to cut the DNA of people with and without sickle cell anemia that people with sickle cell anemia had a 13-kb HpaI fragment, and people without sickle cell anemia had a 7.6-kb fragment. In other words, people in each group had restriction fragments of different and predictable sizes that could be used to distinguish the two groups from each other.

In this case, the nucleotide change that results in a RFLP is also the causative change (i.e., an A-to-T change in the number two nucleotide of the sixth exon in the β -globin chain of the hemoglobin gene, Kan and Dozy, 1978). However, RFLPs only very rarely cause phenotypes and the number of RFLP markers that are available is limited. Overall, RFLPs are primitive tools and provide a limited view of genetic diversity outside of the sequences of interest.

Short tandem repeats

The mid- to late-'80s brought the discovery of STRs (also called microsatellites). The most common forms are di-, tri-, and tetra-nucleotide repeats—such as (CA) n and (CAG) n , where “ n ” is the number of repeats—all of which are flanked by unique sequences. Although their biological function remains unclear, STRs are distributed throughout the genome and are maintained throughout generations.

Typically, each STR locus is scored by PCR amplification with locus-specific

primers. The amplification products, usually designed to be 100 bp or longer, are then separated on a polymer gel such as agarose or polyacrylamide. Each nucleotide repeat of different length at a locus is called an allele, and some loci can have more than ten alleles. The informativeness of the locus, also known as the degree of diversity, is measured by an index called heterozygosity that is based on both the number and proportion of individual alleles. Heterozygosity ranges from 0 to 1, where “0” is no variation and “1” is infinite variation. For markers to be highly informative, researchers generally agree that markers should have a heterozygosity of 0.7 or greater. For instance, STR marker D1S235 has a heterozygosity of 0.71. The score is based on the survey of 56 chromosomes (equivalent to 28 human samples) that revealed seven alleles at frequencies of 48%, 25%, 12%, 5%, 3%, 3%, and 1% respectively (Weissenbach 1992 & Gyapay 1994).

The current STR **genetic map** consists of 10,000 STRs distributed throughout the genome. Genetic distance between two markers is a function of the recombination frequency – the higher the recombination frequency, the farther apart two markers are on a genetic map. The unit of measure for genetic distance is Morgans (named in honor of Thomas Hunt Morgan, a pioneering fruit fly geneticist). A genetic distance of one centimorgan (cM) is equivalent to one percent recombination rate between two markers of interest. The actual physical distance between any two STRs that are a set genetic distance apart varies because genetic distance is a function of recombination frequency, which is variable throughout the genome. For example, in one region of the genome three cM of genetic distance could equal one Mb,

whereas in another region of the genome one cM of genetic distance could equal three Mb. Empirically, one cM is approximately equivalent to one Mb.

To build the STR genetic map, recombination frequencies between markers were determined by genotyping DNA samples from hundreds of people from large families (about 10 people per family) in the *Centre d'Etude du Polymorphisme Humain* (CEPH) sample collection. Then, the completion of human genome sequencing provided a **physical map** for the STRs by identifying the relative location of every STR on every chromosome.

When genome-wide STRs first became available they provided an infrastructure that had never before existed, including a genetic toolbox that could be used to pinpoint the genetic location of disease genes. Now, the availability of both the physical and genetic map makes STRs an even more powerful tool for genetic research.

The availability of STRs is credited for the rapid progress in research on *single gene defects*. One of the major STR-based tool sets is Applied Biosystem's ABI PRISM® Linkage Mapping Sets MD-10 and HD-5. These products provide 10 cM or 5 cM scanning density throughout the genome, respectively, and an average locus heterozygosity of greater than 0.7. These products are frequently

		Maternal	Paternal
Person A	T/T	(+) —CGTAACC—	(+) —CGTAACC—
		(-) —GCATTGG—	(-) —GCATTGG—
Person B	C/T	(+) —CG C AACC—	(+) —CGTAACC—
		(-) —GCGTTGG—	(-) —GCATTGG—
Person C	C/C	(+) —CG C AACC—	(+) —CG C AACC—
		(-) —GCGTTGG—	(-) —GCGTTGG—

Figure 2. Potential offspring from two heterozygous parents at a single locus.

used for **family-based linkage studies** and sometimes used for **sibpair studies**. Technically, genotyping STR loci requires sophisticated efforts because the polymorphisms result in subtle differences in length.

Single nucleotide polymorphisms

Millions of SNPs were discovered when DNA from multiple people was sequenced and compared for the human genome projects. Because of their high frequency throughout the genome, SNPs are believed to provide the highest resolution possible to measure genetic variation. (Figure 1.)

1.1. What are SNPs?

SNPs are single nucleotides that share identical flanking sequences and vary between individuals. These SNPs result from mutations that usually arose thousands of years ago and are stably passed down through generations. Interestingly, many RFLPs have turned out to be the result of SNPs.

Because mutation during replication is rare, SNPs are primarily bi-allelic (i.e., there are only two possible alleles at the locus). Because people have two copies of each chromosome, one inherited from their mother and one from their father, SNP analysis will result in one of the possible diplotypes. (Figure 2.)

At this SNP locus there are two alleles: T or C. Persons A and C are homozygous for T and C alleles, respectively; and person B is heterozygous, carrying both the T and C alleles.

1.2. Different kinds of SNPs

As previously discussed, most SNPs are bi-allelic. Tri-allelic SNPs do occur, but are uncommon because their genesis requires two independent mutation events at the same site. SNPs can also be categorized based on their location; for example SNPs in genes can be categorized as **coding** or **intronic**. Knowing a SNP's

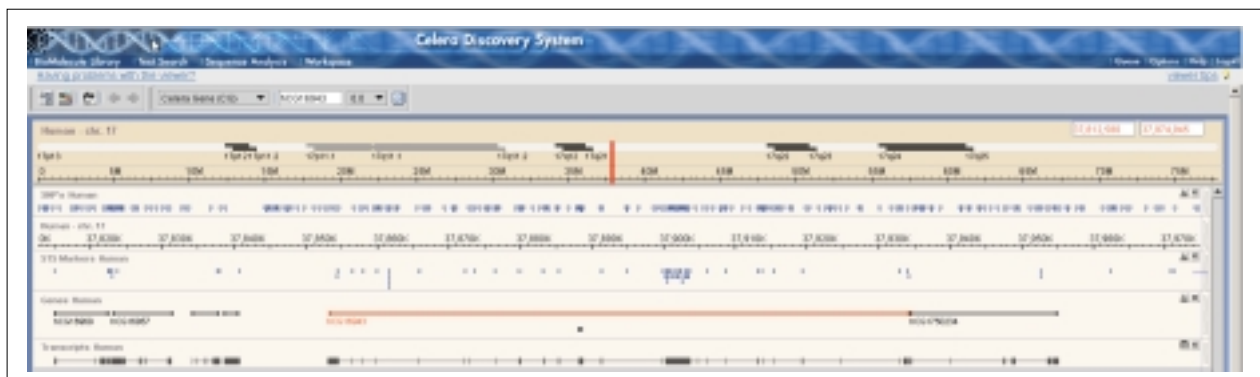


Figure 1. Celeris Discovery System™ graphical Map Viewer shows numerous SNP markers and fewer STS markers along a region of chromosome 17.

location can help focus the search for possible **surrogate** or **causative** SNPs in candidate gene or candidate region association studies. However, most SNPs are neutral functionally; only a small minority directly affect gene function by altering protein coding or regulatory element sequences.

Because SNPs are usually only present in two forms, the allele that is more rare (i.e., present in less than 50% of the population) is referred to as the **minor allele**. Allele frequencies of SNPs vary greatly, in part dependent upon how long ago the mutation occurred in the population. Allele frequencies also vary among populations that have different ancestors and histories. For example, there are cases where the minor allele in the African-American population is the major allele in the Caucasian population, although this is extremely rare for minor alleles that occur at low frequencies. In the Applera genomic initiative, SNP assays in the initial release of the off-the-shelf TaqMan® Assays-on-Demand™ SNP Genotyping Products are restricted to SNPs with minor allele frequencies higher than 5% in either Caucasian or African-Americans in order to deliver the most useful set of assays.

The availability of over four million putative SNPs and a well-defined SNP map provides researchers with the most informative and dense set of genetic markers to date. In addition, because most SNPs are biallelic and technologies such as 5' nuclease assays enable one-step genotype calling, SNPs are also the

easiest genetic markers to score, analyze, and coordinate with high-throughput systems.

2. Haplotypes

Haplotypes are an ordered set of alleles located on one chromosome. They reveal whether a chromosomal segment was maternally or paternally inherited and can be used to delineate the boundary of a possible disease-linked locus.

Genomic DNA is diploid, containing two pairs of each chromosome. Genotyping methods (including TaqMan® 5' nuclease assays and direct sequencing) detect both alleles simultaneously, yielding unphased genotype results. Because it is generally impractical to analyze SNP alleles only on a single chromosome in the laboratory (except for the X and Y chromosomes, which are not part of a chromosomal pair in males) the most common way of determining haplotypes is by computational inference. The algorithms used for these inference methods are still being refined, but are quite effective when large numbers of samples are available for analysis.

In the example below (Figure 3), Smith and Jones both have the same genotype (Aa, Bb, Cc, Dd). However, Smith has ABCD/abcd haplotypes and Jones has AbCd/aBcD haplotypes. Haplotypes can be made up of different alleles of genetic markers including SNPs and STRs.

STRs at a 5 cM to 10 cM density can be sufficient when genotyping familial samples, where limited recombination events—one every 50 cM on average—occur per

meiosis. In other words, haplotypes in families can be represented by genetic markers that are 5 cM to 10 cM apart. On the other hand, when genotyping a population of unrelated individuals, a higher density, approximately one SNP in 5 kb to 200 kb, is required to properly characterize haplotypes. This is because more meioses, and therefore more recombination, could have occurred between a SNP and the putative disease allele. Under such circumstances, only a higher density map, such as a SNP map, will be able to follow the recombination and segregation of alleles (also see “Choosing SNPs for an Association Study” below). This density is defined by the observed average size of co-inheritance blocks, generally described as linkage disequilibrium (LD) blocks, haplotype blocks, or haploblocks.

3. Linkage Disequilibrium

Linkage disequilibrium (LD) is the non-random association between alleles at two loci (Goldstein, 2001), and is primarily the result of a physical association. The extent of LD is typically measured by **D'** and/or **p value**. **D'** ranges from 0 to 1, where 1 indicates complete LD; in other words, the two alleles are always inherited together. A **D'** value of 0.8 or greater indicates significant LD. **P value** is a general probability measure used to calculate whether the observed outcome was likely to occur by chance. Typically, in genetic studies, **p value** is used to measure the likelihood that two loci are not in LD. Therefore, a **p value** of 1 indicates a significant likelihood that the two loci are not in LD.

Smith haplotype		Jones haplotype	
Maternal	___ A ___ B ___ C ___ D ___	Maternal	___ A ___ b ___ C ___ d ___
Paternal	___ a ___ b ___ c ___ d ___	Paternal	___ a ___ B ___ c ___ D ___

Figure 3. Example of two individuals with the same diplotypes who can be discriminated via haplotype analysis.

Alternatively, a p value of 0.001 indicates that there is only a one in 1,000 chance that the two loci are not in LD; in other words, the two loci are highly likely to be inherited together. A p value of 0.005 is considered a threshold of significant LD between two loci.

Because LD is primarily the result of a physical association between markers, an LD block contains the genetic markers that are largely inherited together through multiple generations; thus, one marker can represent an entire LD block no matter how large a physical distance the LD block covers. For an association study, each LD block in the regions (or genes) of interest should be represented by at least one genetic marker.

So far, studies on three chromosomes (De La Vega, 2002 and 2003) and on a collection of other chromosome segments (Gabriel, 2002) indicate that LD patterns vary across the genome and differ among ethnic groups. For example, sample data from chromosome 22 demonstrate that LD blocks in Caucasians extend an average 36 kb. The same study indicates that LD blocks in African-

Americans are generally smaller, averaging approximately 29 kb. Based on these studies it is evident that the SNP density required for LD studies will vary based on the specifics of both the sample population and the genetic region being studied. Applied Biosystems is currently completing the construction of LD maps and block data for the entire genome. It is conceivable that investigators will use the LD and haplotype information generated from Applied Biosystems SNP validation studies to determine the specific density required for their own studies. Access to this information will be provided to guide the cost-effective selection of markers.

However, it is important to remember that although LD blocks are useful for mapping, they do not represent an absolute unit of inheritance. They are blocks that are statistically likely to be inherited together; recombination does occur within LD blocks, just not as frequently as in regions outside of LD blocks. Furthermore, only certain regions of the genome contain LD blocks; other regions undergo high levels of recombination and require higher marker densities for genetic mapping.

An alternative to LD blocks is a metric which measures linkage disequilibrium across the genome. An example of this is the LD unit described in Maniatis, 2001. This has the advantage of covering the whole genome irrespective of the existence of blocks and allows more efficient selection of SNP markers (Figure 4).

4. Types of Genetic Studies

Genetic markers are merely tools that can be used in genetic studies to identify genes responsible for disease susceptibility. There are typically two approaches that researchers take to mapping: linkage studies and association studies. Both of these studies can be implemented for various genetic diseases in different population structures.

Linkage Studies

Linkage studies typically use families with multiple affected individuals ideally including three or more generations to identify genetic regions that are more likely to be inherited with a disease or biological response than would be expected by random chance. Because samples are family based, the blocks of co-inherited parental DNA tend to vary, but

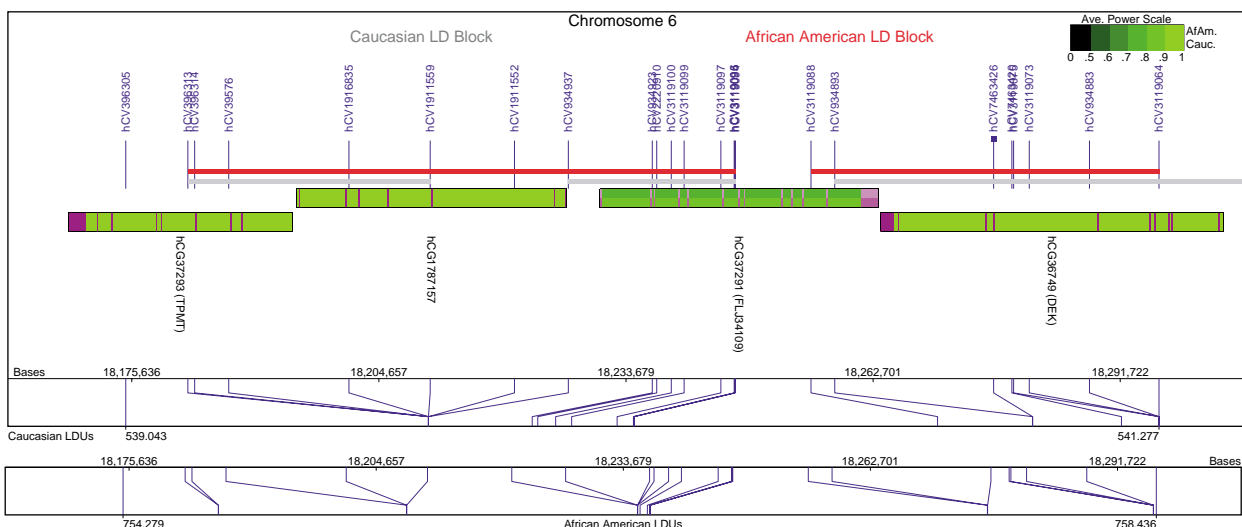


Figure 4. Example of the distribution of SNP genotyping assays across a region of chromosome 6. Validated SNPs are indicated by vertical lines with Celera identifiers, and gene regions as green horizontal rectangles, with exons in purple. Horizontal bars represent haplotype blocks calculated for the African-American (red) and Caucasian (gray) populations. The two axes below indicate the physical scale in base-pairs, and the metric linkage disequilibrium units scale calculated with the LDMAP software (Maniatis et al., 2002) for the Caucasian and African-American populations.

are large, averaging 50 cM. This estimate is based on an average recombination rate of one event per 50 cM between generations. Therefore, a relatively small number of markers are required to achieve 5 cM average density for linkage studies, for example 2500 SNPs (Zubritsky, 1999). Because individual markers such as STRs are more informative, approximately 800 STRs have been used to provide a similar level of resolution in the past. Linkage studies are typically the most effective means for mapping single-gene, highly penetrant phenotypes and are well suited to uncover new correlations between regions and phenotypes. However, because of the large size of the genetic region identified, it is often impossible to use linkage studies to identify individual genes within the linked region that are correlated with the phenotype being studied.

LOD scores are used “to measure the degree of linkage between a marker locus and a putative disease locus. The higher the LOD score, the more likely the two are linked. In humans, a LOD score > 3 is often considered significant and a LOD score < 2 is considered evidence of lack of linkage.

4.1 Successful Linkage Mapping Studies

Linkage studies were proposed by David Botstein in 1980 and have been used successfully ever since. The discoveries of BRCA1 and Crohn’s disease gene locus represent the triumph and difficulties of linkage studies in general.

4.1.1. Discovery of the BRCA1 cancer gene

Researchers attempting to identify genes included in breast cancer knew that in some families susceptibility to develop breast cancer at a young age was passed from gen-

eration to generation. However, a distinctive Mendelian pattern of inheritance was not evident and researchers did not know whether they were dealing with one gene of low penetrance, or many different genes that each contributed to susceptibility in a small way. In 1990, Mary-Claire King’s group published a seminal paper, identifying a 600-kb region around D17S34 on chromosome 17q21 as the location of BRCA1, a putative gene that is associated with the early onset form of breast cancer (Hall et al., 1990).

However, it was not possible through linkage analysis to identify the specific gene in the 600-kb region responsible for the increased breast cancer risk. Instead, researchers then switched to a candidate gene association approach, searching the sequences within the region to look for predicted genes that might be the culprit and conducting association studies with new families. It was four years later when Mark Skolnick’s group identified the specific BRCA1 gene (Miki et al., 1994).

4.1.2. Crohn’s Disease

Starting with epidemiological data showing that siblings of people with Crohn’s disease had a 35-fold increased risk of developing the disease, Rioux and his group undertook a genome-wide linkage study in 158 sib-pair families. They identified a number of regions with compelling linkage, and an 18-cM region on chromosome 5 that contributes to Crohn’s disease susceptibility in families with early-onset disease. An analysis of the genetic regions revealed numerous genes that, based on their known biology, were potential candidates (Rioux et al., 2000). The researchers then created a dense genetic map of STRs and SNPs in the region on chromosome 5, and identified a haplotype contained within a 250-kb linkage block

that confers susceptibility to Crohn’s disease (Rioux et al., 2001). Unfortunately, the strong LD in the region that made mapping more efficient, also resulted in 11 SNPs having equivalent association with disease susceptibility. The researchers were unable to narrow the gene-hunt any further with a genetic mapping approach (Rioux et al., 2001), and intend to turn to molecular biology techniques to identify the gene (Halim, 2001).

The road to the BRCA1 gene discovery and the difficulty in the Crohn’s disease research underscore the need for the tools for fine mapping and candidate-gene-based association studies. Specifically, there is a missing link between broad linkage regions and specific genes. As demonstrated above, disease-linked regions identified by a linkage study tend to be broad—5 cM to 10 cM—and too large to find a specific phenotype-associated gene. The conventional approach for following up on a linked region is a hybrid of physical mapping and candidate gene picking, a laborious and time-consuming process that does not guarantee success. TaqMan® Assays-on-Demand™ SNP genotyping products provide a SNP map at 10 kb density in gene regions, and thus are ready-to-use tools for linkage follow-up studies.

Association Studies

Linkage studies employing family-based samples are largely successful in delineating highly penetrant single-gene disorders and broad genetic regions, however, as illustrated above, linkage studies have only met with limited success in identifying genes involved in polygenic disorders. In fact, almost all complex diseases are polygenic and identifying the underlying genes has been difficult because the multiple disease genes tend to diminish the

statistical significance of linkage to any one gene. It is believed that association studies that compare genetic differences in case and control samples will provide more statistical power to identify disease susceptibility genes.

Traditional association studies begin with a candidate gene or genetic region that researchers already suspect is associated with the phenotype of interest. These studies assess the genetic area of interest in case (affected) and control (unaffected) populations that are as closely matched as possible. For example, researchers might examine the same genetic region in individuals with high cholesterol levels and individuals with desired cholesterol levels to determine if there are genetic differences between these two groups. One challenge associated with traditional association studies is their dependence on prior knowledge about the gene or region. Thus, traditional association studies are not suitable for identifying new genes or genetic regions that might be associated with a phenotype.

Prior to the completion of the human genome project, another barrier to successful implementation of association studies in the past was that there were only a very limited set of genetic markers, which made it difficult to establish meaningful associations. However, today, because of the explosion of knowledge about biological pathways and the increase in availability of informative genetic markers such as SNPs, association studies have a better chance of success. Modern association studies can be categorized into three groups: candidate-gene-based association studies, candidate-region-based association studies,

and whole genome association studies.

4.2. Candidate-Gene-Based Association Studies

Candidate-gene-based association studies are the most common approach used in gene identification research. Such studies are useful when a researcher has collected samples from patients with a specific disease phenotype and has a belief or evidence that the underlying genetic defect resides in a specific biological pathway. The biological pathway is often identified through literature searches and/or other previous knowledge. Before the genomic era, candidate-gene-based association studies were done based on limited information about genes and genetic markers and met with only limited success. Today, with the completion of human genome sequencing and annotation, where 30,000 genes in the human genome have been identified and over four million SNPs discovered, candidate-gene association studies should be an effective approach for identifying disease-causing genes. The TaqMan® Assays-on-Demand™ SNP genotyping products, which were designed with a gene-centric focus, provide one SNP approximately every 10 kb, are fully validated, and provide powerful tools to maximize the chance of success for candidate-gene-based association studies.

A candidate-gene-based association study might involve more than 1,000 samples including cases and controls. The number of genes under investigation could be 1 to 2 (single-gene based), 10 to 20 (gene-family based), or 50 and up (biological pathway-based). Through TaqMan® Assays-on-Demand™ products, researchers can access approximately five SNPs per gene;

therefore, the number of SNPs per study could range from 5 to 250 or more.

4.3. Candidate-Region-Based Association Studies

Candidate-region-based association studies are often useful when previous literature and/or a linkage study establishes a link between a disease phenotype and a specific region on a chromosome. The so-called candidate-region is often defined by STRs, a cytogenetic band, or combination of both. Choosing genetic markers in a target region with a density of about one SNP per 10 kb (in gene regions) maximizes the chance of identifying a disease-associated locus. Candidate-region-based association studies provide an opportunity for discovering novel genes based on their location information.

Candidate-region-based association studies often require a larger number of samples than a linkage study. This is because more samples means more chromosomes, which provide more recombination events in the target region (as well as genome-wide). Recombination that creates haplotypes coinciding with phenotypic changes is called informative **recombination**; such events are the key to narrowing down the target region to a single gene.

A candidate-region-based association study, like the candidate-gene based-association study, might involve more than 1,000 samples. Candidate regions, often derived from a STR mapping study, are typically at 5 to 30 cM in size. The number of SNPs necessary for fine mapping will vary based on the target region, however most candidate-region-based association studies can be carried out with between 500 and 3000 SNPs from the TaqMan® Assays-on-Demand™ SNP map.

4.4. Whole Genome Association Studies

Whole genome association studies, which are similar to candidate-region-based association studies, provide the opportunity to discover novel genes because they do not depend upon prior knowledge of a candidate gene. In addition, such studies are the most promising for identifying susceptibility alleles with small relative risks that currently cannot be identified by linkage studies. The density of genetic markers required, although still undetermined, will depend on the makeup of the samples, and is likely to vary from region to region and from ethnic group to ethnic group, with some regions requiring a higher density than others because of varying sizes of LD blocks. A survey of chromosome 22 segments revealed an average LD block size of 25.8 kb. If this size holds true for the rest of genome, and assuming researchers would use two SNPs per LD block in whole genome association studies, just over 200,000 SNPs would be required to cover the entire genome. Although the cost of whole genome association studies is still prohibitive, and none have been completed to date, many geneticists remain hopeful that they will soon be possible, perhaps as a result of advances in automation and highly parallel genotyping and data analysis.

Applied Biosystems is developing a new, ultra-high-throughput genotyping technology called SNPlex™ System, which is expected to enable these types of studies.

4.5. Choosing SNPs for an Association Study

With the availability of the Applied Biosystems LD map in the form of off-the-shelf TaqMan® Assays-on-Demand™ products corresponding

to each SNP, and the custom TaqMan® Assays-by-Design™ service, researchers have the option of choosing which SNPs to use in their studies based on the needs of the experiment, rather than designing experiments around a small number of available assays.

One consideration when selecting SNPs for genetic studies is whether to select markers based on density only, the so-called “picket fence” approach, or to bias the selection to SNPs in genes. In developing the TaqMan® Assays-on-Demand™ SNP genotyping products, we have chosen to use a gene-centered picket fence approach. This approach includes SNPs in regions based on genes that are informative (minor allele frequency > 5%) and densely located (roughly 10 kb apart). This is based on the likelihood that most disease, and susceptibility to disease, is a result of gene function, including the amount and activity of gene products.

We are living at an exciting time, where the desire to improve the quality of life fuels the research of disease susceptibility and treatment of disease. SNPs have provided us with the most powerful tools to date for studying the genetics of complex diseases and traits in order to reach these lofty goals.

5. References

De La Vega, F.M., et al. 2003. “A Whole-Genome Gene-Centric Linkage Disequilibrium SNP Map.” XIX Congress of Genetics, July 6-11, Melbourne, Australia.

De La Vega, F.M., et al. 2002. “Selection of single nucleotide polymorphisms for a whole-genome linkage disequilibrium mapping set.” CSH Genome Sequencing & Biology Meeting, May 7-11, Cold Spring Harbor, NY.

Gabriel, S.B., et al. 2002. “The structure of haplotype blocks in the human genome.” *Science*. June 21; 296(5576):2225-9.

Goldstein, D. 2001. “Islands of linkage disequilibrium.” *Nature*. 29(2):109-111.

Gyapay, G., et al. 1994 “The 1993-94 Genethon human genetic linkage map” *Nature Genetics*. 7(2):246-339.

Hall, J.M., Lee, M.K., Newman, B., et al. 1990. “Linkage of early-onset familial breast cancer to chromosome 17q21.” *Science*. 250, 1684-1689.

Halim, N. 2001. “Scientists build case for ‘haplotype’ map of human genome, find new gene for Crohn’s disease.” *Eureka Alert*. October 3.

Kan, Y.W. and Dozy, A.M. 1978. “Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation.” *Proc Natl Acad Sci USA*. 75(11):5631-5.

Maniatis, N., et al. 2002. “The first linkage disequilibrium (LD) maps: Delineation of hot and cold blocks by diplotype analysis.” *PNAS* 99: 2228-33.

Miki, Y., et al. 1994. “A strong candidate for the breast and ovarian cancer gene BRCA1.” *Science*. 266, 66-71.

Rioux, J.D., et al. 2000. “Genome wide search in Canadian families with inflammatory bowel disease reveals two novel susceptibility loci.” *Am. J. Hum. Genet.* June; 66 (6):1863-70.

Rioux, J.D., et al. 2001. “Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn’s disease.” *Nat. Genet.* Oct; 29 (2):223-8.

Weissenbach, J. et al. 1992. "A second-generation linkage map of the human genome." *Nature*. 359(6398):794-801.

Zubritsky, E. 1999. "SNP mining." *Analytical Chemistry News & Features*. October 1, 683A-686A.

Glossary

Allele – One of the variant forms of a gene or a genetic locus.

Causative SNPs – Changes in a single nucleotide that cause a disease or trait.

Coding SNPs – SNPs that occur in regions of a gene that are transcribed into RNA (i.e., an exon) and eventually translated into protein. Coding SNPs include synonymous SNPs (i.e., confer identical amino acid) and non-synonymous SNPs (i.e., confer different amino acid).

Diplotype – Genotyping data for which no phase or haplotype is known, e.g. A/a: there is no knowledge which of the pair of chromosomes the A allele resides on.

Genetic map – Also known as a linkage map. A genetic map shows the position of genes and/or markers on chromosomes relative to each other, based on genetic distance (rather than physical distance). The distance between any two markers is represented as a function of recombination.

Genetic marker – A DNA sequence whose presence or absence can be reliably measured. Because DNA segments that are in close proximity tend to be inherited together, markers can be used to indirectly track the inheritance pattern of a gene or region known to be nearby.

Genotype – The combination of alleles carried by an individual at a particular genetic locus.

Homologous genes or regions – Genes or regions found in different species that are very similar. The homology indicates that these genes or regions are evolutionarily conserved, and implicates their functional importance.

Intronic SNPs – Single nucleotide polymorphisms that occur in non-coding regions of a gene that separate the exons (i.e., introns). Introns are transcribed into RNA, but are not translated into protein.

Linkage map – See genetic map.

Meiosis – Meiosis is the process by which one diploid cell gives rise to four haploid cells, each containing half of the chromosome complement of the diploid cell. In mammals, meiosis produces sperm cells and egg cells. It is during meiosis that recombination occurs.

Mendelian pattern of inheritance – Refers to the predictable way in which single genes or traits can be passed from parents to children, such as in autosomal dominant, autosomal recessive, or sex-linked patterns.

Mutation – A change in the DNA sequence. A mutation can be a change from one base to another, a deletion of bases, or an addition of bases. Typically, the term mutation is used to refer to a disease-causing change, but technically any change, whether it causes a different phenotype or not, is a mutation.

Penetrance – Penetrance describes the likelihood that a mutation will cause a phenotype. Some mutations have a high penetrance, almost always causing a phenotype, whereas others have low penetrance, perhaps only causing a phenotype when other genetic or environmental conditions are present. The best way

to measure penetrance is phenotypic concordance in monozygotic twins.

Phenotype – Visible or detectable traits caused by underlying genetic or environmental factors. Examples include height, weight, blood pressure, and the presence or absence of disease.

Physical map – A map that shows the specific physical points where genes and/or markers on chromosomes reside. A physical map typically marks distance by measures of physical distance (e.g., kilobases).

Polygenic disorders – Disorders that are caused by the combined effect of multiple genes, rather than by just one single gene. Most common disorders are polygenic. Because the genes involved are often not located near each other, inheritance does not usually follow Mendelian patterns of inheritance in families.

Sib-pair analysis – Sib-pair analysis involves studying siblings (with at least one of them showing a phenotype) and often their parents (for investigating maternal or paternal inheritance).

Single gene defects – Defects caused by the action of only one gene. With single gene defects, it is typically easy to identify a Mendelian pattern of inheritance, unless the mutation is of low penetrance.

Surrogate SNPs – Single nucleotide polymorphisms that do not cause a phenotype, but can be used to track one because of their strong physical association (linkage) to a SNP that does cause a phenotype.

Susceptibility – The likelihood of developing a disease or condition.

Worldwide Sales Offices

Applied Biosystems vast distribution and service network, composed of highly trained support and applications personnel, reaches 150 countries on six continents. For international office locations, please call the division headquarters or refer to our Web site at www.appliedbiosystems.com.

Applera is committed to providing the world's leading technology and information for life scientists. Applera Corporation consists of the Applied Biosystems and Celera Genomics businesses.

Headquarters

850 Lincoln Centre Drive
Foster City, CA 94404 USA
Phone: 650.638.5800
Toll Free: 800.345.5224
Fax: 650.638.5884

For Research Use Only.
Not for use in diagnostic procedures.

ABI PRISM and Applied Biosystems are registered trademarks and AB (Design), Applera, Assays-by-Design, Assays-on-Demand and SNPlex are trademarks of Applera Corporation or its subsidiaries in the U.S. and/or certain other countries.

TaqMan is a registered trademark of Roche Molecular Systems, Inc. The PCR process and 5' nuclease process are covered by patents owned by Roche Molecular Systems, Inc. and F. Hoffmann-La Roche Ltd.

All rights reserved.

©2003 Applied Biosystems. All rights reserved.

Printed in the USA, 8/2003, LD
Publication 127AP01-01