MINIREVIEW

Barkur S. Shastry

SNP alleles in human disease and evolution

Received: June 12, 2002 / Accepted: July 26, 2002

Abstract In two randomly selected human genomes, 99.9% of the DNA sequence is identical. The remaining 0.1% of DNA contains sequence variations. The most common type of such variation is called a single-nucleotide polymorphism, or SNP. SNPs are highly abundant, stable, and distributed throughout the genome. These variations are associated with diversity in the population, individuality, susceptibility to diseases, and individual response to medicine. Recently, it has been suggested that SNPs can be used for homogeneity testing and pharmacogenetic studies and to identify and map complex, common diseases such as hypertension, diabetes, obesity, and psychiatric disorders. These common diseases are likely to be caused by multiple genes and multiple nongenetic factors (environmental factors), each contributing a modest effect. Their cumulative effect results in the condition or trait. The traditional methods of identifying disease-related genes are not readily applicable to the detection of genes responsible for these multifactorial diseases. The availability of the entire genome sequence, therefore, may speed up the gene-hunting efforts in the near future, but what approaches are to be taken to accomplish this task and what are their limitations?

The human genome and the discovery of single-nucleotide polymorphisms (SNPs) as genetic markers

In two randomly selected human genomes, 99.9% of the DNA sequence is identical. The remaining 0.1% is thought to include some differences or variations in the genome between individuals. This variation, called polymorphism, arises because of mutations. The simplest form of these variations is the substitution of one single nucleotide for another (Fig. 1A), termed SNP. SNPs are more common than other types of polymorphisms and occur at a frequency of approximately 1 in 1000 base pairs (Brookes 1999) throughout the genome (promoter region, coding sequences, and intronic sequences). These simple changes in DNA sequence, most of which are probably located in intergenic spacers, are believed to be stable and not deleterious to organisms. SNPs that do not change encoded amino acids are called synonymous and are not subject to natural selection (Kimura 1983, snp.cshl.org). On the other hand, nonsynonymous SNPs alter amino acids and might be subject to natural selection. SNPs can be observed between individuals in a population, may influence promoter activity
or DNA and pre-mRNA conformation, and play a direct or indirect role in phenotypic expression (Krawezak et al. 1992; Lohrer and Tangen 2000; Pitarque et al. 2001; Spicker et al. 2001; LeVan et al. 2001). Because some SNPs are functional, comparative studies on identical twins, fraternal twins, and siblings suggest that genetic variation is one of the factors associated with susceptibility to many common diseases (Table 1) as well as every human trait such as tallness, curly hair, and individuality (Martin et al. 1997). Diversity in the population is also associated with these variations. Therefore, it may be possible to understand why some individuals are susceptible to common disorders by using the human genome sequence and the variations between individuals. However, there are limitations, and practical and ethical issues must be considered before undertaking such analyses (Chanock 2001).

### Table 1. A partial list of disorders associated with SNPs

<table>
<thead>
<tr>
<th>Disorder</th>
<th>Gene</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asthma</td>
<td>EDN 1 and NOS 1</td>
<td>Immervoll et al. 2001</td>
</tr>
<tr>
<td>POAG</td>
<td>Myocillin</td>
<td>Colomb et al. 2001</td>
</tr>
<tr>
<td>Systemic sclerosis</td>
<td>Fibrillin 1</td>
<td>Tan et al. 2001</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>MMP-1</td>
<td>Zhu et al. 2001</td>
</tr>
<tr>
<td>Arrhythmias</td>
<td>KCNQ1</td>
<td>Kubota et al. 2001</td>
</tr>
<tr>
<td>Idiopathic arthritis</td>
<td>MIF</td>
<td>Donn et al. 2001</td>
</tr>
<tr>
<td>Blood pressure</td>
<td>TAF1</td>
<td>Koschinsky et al. 2001</td>
</tr>
<tr>
<td>Biliary cirrhosis</td>
<td>MBL</td>
<td>Matsushita et al. 2001</td>
</tr>
<tr>
<td>Type II diabetes</td>
<td>Syntaxin 1A</td>
<td>Tsunoda et al. 2001</td>
</tr>
<tr>
<td>Systemic lupus erythematosus</td>
<td>Prolactin</td>
<td>Stevens et al. 2001</td>
</tr>
<tr>
<td>Eating disorder</td>
<td>Melanocortin</td>
<td>Adan and Vink 2001</td>
</tr>
<tr>
<td>Migraine</td>
<td>Insulin receptor</td>
<td>McCarthy et al. 2001</td>
</tr>
<tr>
<td>Ossification</td>
<td>Npps</td>
<td>Koshizuka et al. 2002</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>p53</td>
<td>Biros et al. 2001</td>
</tr>
<tr>
<td>Late-onset PD</td>
<td>tau</td>
<td>Martin et al. 2001</td>
</tr>
</tbody>
</table>

SNPs, Single-nucleotide polymorphisms; POAG, primary open-angle glaucoma; MMP-1, matrix metalloproteinase 1; EDN 1, endothelin 1; NOS 1, neuronal nitric oxide synthetase 1; MBL, mannose binding protein; Npps, nucleotide pyrophosphatase; TAF1, thrombin activatable fibrinolysis inhibitor; KCNQ1, potassium channel protein; MIF, macrophage inhibitory factor; PD, Parkinson disease.

2001 and http://www.ncbi.nlm.nih.gov/SNP/). As a result of this worldwide intensive effort, more than 2.8 million SNPs have been identified and a high-density map has been constructed in some cases (Iida et al. 2001a–d; Iwasaki et al. 2001; Saito et al. 2001; Osier et al. 2001). Although several mapping methods (Table 2), such as single-strand conformational polymorphism (Orita et al. 1989), denaturing gradient gel electrophoresis (DGGE), enzymatic mutation detection (Youil et al. 1995), microarray or variant detector arrays (Wang et al. 1998; Marshall and Hodgson 1998; Ramsay 1998; Hacia et al. 1999; Hacia and Collins 1999; Dong et al. 2001; Qi et al. 2001; Yoshino et al. 2001), and heteroduplex analysis (Lichten and Fox 1983) are available, so far none of them has supplanted DNA sequencing as the method of choice. SNP mapping requires a tremendous amount of time and resources (hundreds and thousands of individuals must be studied to eliminate false-positive and false-negative results). Additionally, many of the methods detect only about 80% of mutations. In some methods, mutations in GC-rich regions may not be detected (e.g., DGGE), and other methods involve expensive instruments and kits or toxic chemicals (Table 2). Therefore, new, faster methods must be developed and old methods refined so that, at least, cost is not a key consideration. In this respect, recently developed high-throughput SNP genotyping (Jenkins and Gibson 2002; McClay et al. 2002) and molecular beacon methods (Mhlanga and Malmberg 2001) are highly valuable. According to some estimates, 50% of mutations are likely to be in a noncoding sequence, 25% lead to amino acid substitution, and 25% are silent (Cargill et al. 1999; Halushka et al. 1999). However, it should be noted that the above estimates are only for cSNPs (SNPs found in protein-coding regions) and SNPs in untranslated regions depend on sample size because they have different allele frequency distributions.

**Mapping and characterization of SNPs**

Because most sequence variants are SNPs, a massive effort has been undertaken by several private and public organizations, such as Celera Genomics, Incyte Genomics, the Wellcome Trust Sanger Institute in the United Kingdom, and Washington University in the United States, to generate a high-density SNP map of the genome (Marth et al. 2001; Irizarry et al. 2000; Altshuler et al. 2000; Mullikin et al. 2000; also see the International SNP Map Working Group 2001 and http://www.ncbi.nlm.nih.gov/SNP/). As a result of this worldwide intensive effort, more than 2.8 million SNPs have been identified and a high-density map has been constructed in some cases (Iida et al. 2001a–d; Iwasaki et al. 2001; Saito et al. 2001; Osier et al. 2001). Although several mapping methods (Table 2), such as single-strand conformational polymorphism (Orita et al. 1989), denaturing gradient gel electrophoresis (DGGE), enzymatic mutation detection (Youil et al. 1995), microarray or variant detector arrays (Wang et al. 1998; Marshall and Hodgson 1998; Ramsay 1998; Hacia et al. 1999; Hacia and Collins 1999; Dong et al. 2001; Qi et al. 2001; Yoshino et al. 2001), and heteroduplex analysis (Lichten and Fox 1983) are available, so far none of them has supplanted DNA sequencing as the method of choice. SNP mapping requires a tremendous amount of time and resources (hundreds and thousands of individuals must be studied to eliminate false-positive and false-negative results). Additionally, many of the methods detect only about 80% of mutations. In some methods, mutations in GC-rich regions may not be detected (e.g., DGGE), and other methods involve expensive instruments and kits or toxic chemicals (Table 2). Therefore, new, faster methods must be developed and old methods refined so that, at least, cost is not a key consideration. In this respect, recently developed high-throughput SNP genotyping (Jenkins and Gibson 2002; McClay et al. 2002) and molecular beacon methods (Mhlanga and Malmberg 2001) are highly valuable. According to some estimates, 50% of mutations are likely to be in a noncoding sequence, 25% lead to amino acid substitution, and 25% are silent (Cargill et al. 1999; Halushka et al. 1999). However, it should be noted that the above estimates are only for cSNPs (SNPs found in protein-coding regions) and SNPs in untranslated regions depend on sample size because they have different allele frequency distributions.
SNPs in gene discovery

Once the map of these SNPs is confirmed, they can be used for evolutionary biology studies, gene discovery and mapping, prediction of drug and environmental response, diagnostic tests, heterogeneity testing, and association studies (Gray et al. 2000; Schork et al. 2000). For the purpose of gene discovery, SNPs are considered to be the most predominant segregating form of variation at the molecular level because of their frequent occurrence throughout the genome, and they can be useful in association studies. However, they are less informative, in the sense that humans have relatively low nucleotide diversity compared with Drosophila and maize, than another type of marker called a microsatellite. Microsatellites are simple sequence repeats, the most common classes being dinucleotide, trinucleotide, and tetranucleotide, and they occur at a rate of 1 in every 10 kb in a wide range of eukaryotic genomes. Human microsatellites are used for linkage studies and they average at least ten alleles with heterozygosity per locus over 80% (very high). Although they have a much higher mutation rate than the standard sequence, they are not densely distributed. Linkage mapping focuses on the small number of meiotic events within a family and association between marker alleles and traits. It does not require a very dense map of markers at the initial stage. On the other hand, linkage disequilibrium mapping explores family associations and requires a dense map of markers. In association studies, the marker is prevalent in patients versus those without the disease and this is considered evidence of association between the disease and the marker (SNP). Although there are many limitations (described following), association studies are perhaps the best for mapping of polygenic complex disease loci. However, this type of study requires a large number of patients and an adequate control group to achieve over 80% power to detect a locus. Once a significant site is identified, one can use either a pedigree-based transmission disequilibrium test (TDT), which measures the transmission of alleles from a heterozygous parent to the affected offspring (an unequal transmission of SNP alleles to affected and unaffected siblings), or a case-control population sampling, which measures the association between SNPs and the disease in a large population (Keavney et al. 2000; Spielman et al. 1992). TDT detects both linkage and association, and SNPs are usually used. These methods are based on the assumption that SNP variants account for population susceptibility to certain disorders; however, it is unknown how many SNPs are needed. These methods have several limitations, such as difficulties associated with population structure, different levels of linkage disequilibrium in loci, allelic and nonallelic heterogeneity of phenotypes, and epistatic interaction of alleles, all of which have been previously discussed by others (Schork et al. 2000; Chakravarti 1999; Weiss and Terwilliger 2000). An allele frequency database for human polymorphic sites for multiple populations (Osier et al. 2001; Hirakawa et al. 2002) can be found on the Kidd lab home page (http://info.med.yale.edu/genetics/kkiddd/). Additionally, a recently proposed haplotype (a distinct combination of single nucleotide types on a single chromosome at a locus) map (Fig. 1B) of the genome may speed up or simplify the hunt for the association between DNA variations and complex diseases (Helmuth 2001). This is because haplotype diversity, which is greater than SNP diversity, may be generated by new SNP alleles that can arise because of mutation at different loci, and hence can be studied by both linkage and association methods. In haplotype association studies, multitype genotypes are reduced to haplotypes and this has proved to be a more efficient mapping technique than that of SNP analysis.

SNPs in pharmacology

Another potential application of SNPs is the development of individualized medicine. Inherited genetic differences between individuals appear to determine each patient’s response to medicine. For instance, some patients show response to the prescribed drug without any serious side effects, whereas others do not respond and experience adverse reactions. In fact, it is estimated that properly prescribed medications cause 2 million Americans to get sick and result in 100 000 deaths each year because of adverse

---

**Table 2. A comparison of selected mutation screening methods**

<table>
<thead>
<tr>
<th>Method</th>
<th>Fragment length (bp)</th>
<th>Advantage</th>
<th>Disadvantage</th>
<th>Efficiency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-strand conformational polymorphism</td>
<td>~300</td>
<td>No expensive equipment</td>
<td>Small fragments, temperature variation</td>
<td>80</td>
</tr>
<tr>
<td>Heteroduplex</td>
<td>300–600</td>
<td>No expensive equipment</td>
<td>Conditions to be determined</td>
<td>80</td>
</tr>
<tr>
<td>Denaturing gradient gel electrophoresis</td>
<td>100–1000</td>
<td>Simple, long and short fragments</td>
<td>Gradient gel required, mutation in GC region may not be detected</td>
<td>100 with GC clamps</td>
</tr>
<tr>
<td>Enzymatic mismatch detection</td>
<td>300–1000</td>
<td>Long and short fragments</td>
<td>Identifies all kinds of mutations</td>
<td>100</td>
</tr>
<tr>
<td>Base excision sequence scanning</td>
<td>50–1000</td>
<td>Accurate</td>
<td>Expensive instruments</td>
<td>100</td>
</tr>
<tr>
<td>RNAase cleavage</td>
<td>1.6 kb</td>
<td>Longer fragment and fast</td>
<td>Requires special kit</td>
<td>100</td>
</tr>
<tr>
<td>Chemical cleavage</td>
<td>1–2 kb</td>
<td>Large fragment</td>
<td>Multistep, labor intensive, hazardous chemicals</td>
<td>100</td>
</tr>
<tr>
<td>DNA sequencing</td>
<td>500</td>
<td>Rapid and easy, no additional sequencing</td>
<td>Labor intensive</td>
<td>100</td>
</tr>
</tbody>
</table>
drug reactions (Pirmohamed and Park 2001; Bader 2001; Wolf et al. 2000; Lazarou et al. 1998). Antitumor agent 6-mercaptopurine, for example, has been used for children with leukemia. Those children who do not have its metabolizing enzyme — thiopurine methyltransferase — show a severe hematopoietic toxicity, whereas children with the more active enzyme require higher doses of the drug. Similar anti–drug reactions were also observed in asthma and other conditions (Drazen et al. 1999; Roses 2000). The most important of the several hundred thousand SNPs in the human genome for this purpose are SNPs in drug-metabolizing enzymes, particularly in the coding and the promoter regions, because changes there may have functional significance (Risch 2000). The SNP method can also be applied to crop (corn and soybean) genetics (Rafalski 2002). Haplotype analysis is possible for crop genetics because of the availability of many genes and expressed sequence tags, as well as their intraspecific nucleotide diversity.

SNPs and evolution

SNPs can be used to study DNA sequence variation among species. Because such variations are present at all levels of evolution, including branching points of speciation, they may provide an understanding of how the modern genome evolved. Most SNPs are not in protein-coding regions but are elsewhere in the genome; therefore, their distribution is not under selective pressure. Variations in the protein-coding region that affect the phenotype might be subject to natural selection, but if these variations were retained in the gene over time, then they must have some benefit for the individual for successful reproduction. The variants that are selected for retention by natural selection may represent an important step in evolution. Thus, by calculating a ratio between variants in noncoding and coding regions of a series of protein families found in different species, it may be possible to trace the branching point of an evolutionary tree. At this branch point, the variant must have become advantageous for the species and hence fixed in the gene pool (Osier et al. 2001; Stephens et al. 2001; Liberles 2001). While this gene pool has been continuously expanding during evolution, it might have resulted in the modern human genome. It is very well known from several studies that humans are similar to chimpanzees at the genomic level. However, there are differences between humans and chimpanzees. For instance, malaria; rheumatoid arthritis; and breast, colon, and lung cancers are extremely rare in chimpanzees but common in humans. In this regard, SNPs may provide important health clues.

However, it is not clear at present how genetic variation alone determines the susceptibility of an individual to some diseases or to adverse drug interactions. This is because most common traits and phenotypes are the result of long-term interaction between genetic and nongenetic environmental factors. Factors such as lifestyle and diet may contribute to disease susceptibility by altering gene expression. Frequency of polymorphism may also vary among different populations (Wakeley et al. 2000; Nielsen 2000; Nielsen and Slatkin 2001). Therefore, even after complete sequencing and identification of SNPs in an individual’s DNA, it is not simple to associate these variations to disease without knowing the functional significance of the identified SNPs. Additionally, researchers are struggling to understand disease heterogeneity even in monogenic disorders. In multigenic disorders, the contribution of susceptible individual genes to the disorder is very weak. Therefore, genotype alone is not sufficient to predict or of susceptibility to disease (Martin et al. 1997) nor is phenotype variation alone necessarily linked to DNA sequence variation (Lander and Schork 1994; Weeks and Lathrop 1995). On the other hand, finding out how SNPs affect an individual’s health and transforming this knowledge into the development of new medicine, which requires the correlation of SNPs with specific diseases, will revolutionize the treatment of most common killer diseases. This is because such an understanding of the relationship between SNPs and the disease will allow clinicians to determine whether an individual will respond to a medicine or experience serious side effects. Drug companies will be able to design different drugs for each patient with similar clinical symptoms or disease phenotypes. In future, this knowledge may give clinicians more insight into the disease and change the definition of some disorders.

Conclusion

There is no doubt that the identification of genes underlying polygenic and complex diseases such as psychiatric disorders, diabetes, hypertension, and asthma is of paramount interest for clinicians, geneticists, patients, and the public.

References

Bader JS (2001) The relative power of SNPs and haplotype as genetic markers for association tests. Pharmacogenomics 2:11–24
TIGR/MYOCOLIN gene promoter with the severity of primary open-angle glaucoma. Clin Genet 60:220–225


Tsunoda K, Sanke T, Nakagawa T, Furuta H, Nono K (2001) Single nucleotide polymorphism (D68D, T to C) in the syntaxin 1A gene correlates to age at onset and insulin requirement in type II diabetes patients. Diabetologia 44:2092–2097


