Current Research Status, Databases and Application of Single Nucleotide Polymorphism

1R. Javed and 2Mukesh
1DNA Sequencing Lab, National Bureau of Animal Genetic Resources, Karnal-132001, Haryana, India
2Department of Endangered Species Management, Wildlife Institute of India, Post Box 18, Chandrabani, Dehradun-248001, Uttarakhand, India

Abstract: Single Nucleotide Polymorphisms (SNPs) are the most frequent form of DNA variation in the genome. SNPs are genetic markers which are bi-allelic in nature and grow at a very fast rate. Current genomic databases contain information on several million SNPs. More than 6 million SNPs have been identified and the information is publicly available through the efforts of the SNP Consortium and other data bases. The NCBI plays a major role in facilitating the identification and cataloging of SNPs through creation and maintenance of the public SNP database (dbSNP) by the biomedical community worldwide and stimulate many areas of biological research including the identification of the genetic components of disease. In this review article, we are compiling the existing SNP databases, research status and their application.

Key words: Single nucleotide polymorphisms, SNPs databases, NCBI, SNP consortium, research status

INTRODUCTION

A single-nucleotide polymorphism (SNP, pronounced snip) is a DNA sequence variation occurring when a single nucleotide-A, T, C, or G- in the genome (or other shared sequence) differs between members of a species (or between paired chromosomes in an individual). For example, two sequenced DNA fragments from different individuals, AAGCTTA to AAGCTTA, contain a difference in a single nucleotide. In this case, there are two alleles-C and T. Almost all common SNPs have only two alleles. Single nucleotides may be changed (substitution), removed (deletions) or added (insertion) to a polynucleotide sequence. Ins/del SNP may shift translational frame. SNPs are mostly bi-allelic polymorphism and may occur in both coding and non-coding region of the genome. The SNPs found with in a coding sequence are of particular interest to scientist because they are more likely to alter the biological function of a protein (Shen et al., 2006). A SNP has a minor allele frequency of greater than 1% at least one population (Risch, 2000). The SNPs are essential for personalized medicine (Chang et al., 2007; Suh and Vigg, 2005; Erichsen and Chanoock, 2004). Many methodologies are reported or reviewed for genetic association studies (Lunn et al., 2006; Su et al., 2005). Database of SNPs provide a powerful resource for association studies that try to establish a relationship between phenotypes.

However, SNPs may play a useful role in niche applications such as mitochondrial DNA (mtDNA), Y-SNPs, Ancestry Informative Markers (AIMs), predicting phenotypic traits and other potential forensic case work applications (Bulbul et al., 2009). Coding region SNPs can fulfill a useful role for separating common HIV1/HIV2 mitochondrial DNA types (Coble et al., 2004, 2006) and assays have been developed to reliably examine mtDNA coding region SNP variation (Brandstatter et al., 2003; Vallone and Butler, 2004; Vallone et al., 2004). A total number of 10-30 million SNPs in the human genome are estimated (Bostein and Risch, 2003).

Over the past few years, SNPs have been proposed as the next generation of markers for the identification of loci associated with complex diseases and for pharmacogenetic applications (Lander and Schork, 1994; Lander, 1996; Risch and Merikangas, 1996; Kruglyak, 1997; Schafer and Hawkins, 1998). Association studies can genotype a limited set of SNPs that contribute to common haplotypes, with haplotype tagging SNPs (htSNPs) (Stram et al., 2003). Several biotech companies started the race and took the initiative to isolate 60,000 or more SNPs to develop a whole genome SNP-based map (http://www.abbott.com/news/1997/news/pr072897.htm).

The importance of SNPs in genetic studies comes from at least three different categories. First, SNPs can be used to reconstruct the history of genome. This is due to their abundance and most are inherited from one

Corresponding Author: Ruheena Javed, National Bureau of Animal Genetic Resources, Karnal-132001, Haryana, India
Tel: +919728589305, +91935463296

657
generation to the next, evolutionary stable, making them easier to follow in population studies. Studying the frequency and distribution of SNPs can lead to information on the evolution of the species. Second, SNPs can be directly responsible for genetic diseases since they may alter the genetic sequence of gene or of a regulatory region. Finally, SNPs may be utilized as markers to build the high-density genetic maps needed to perform association studies (Crapan and Suthat, 2007), which locate and identify genes of functional importance. It has been proposed that a set of 3,000 bi-allelic SNP markers would be sufficient for whole-genome-mapping studies in humans; a map of 100,000 or more SNPs has been proposed as an ultimate goal to enable effective genetic mapping studies in large populations (Gibbs, 2003). The SNP databases are important resources for performing genetic linkage, association and admixture studies, both academic and commercial groups are developing large number of genome-wide SNP dataset. These dataset now contain over 12.6 million SNPs (Baye et al., 2009).

**Status of current research:** Current estimates are that SNPs occur as frequently as every 100-300 bases. This implies in an entire human genome there are approximately 10 to 30 million potential SNPs. More than 6 million SNPs have been identified and the information has been made publicly available through the efforts of the SNP consortium and others databases. Many of these SNPs have unknown associations. Compilation of public SNPs by NCBI (http://www.ncbi.nlm.nih.gov/projects/SNP/) has produced a subset of SNPs defined as a non-redundant set of markers that are used for annotation of reference genome sequence and are thus referred to as reference SNPs (rsSNPs). At present there are several SNP public databases (Dvornik et al., 2004). The largest are dbSNP and HGVRbase containing together several million SNPs (Aerts et al., 2002). There also exist other relatively small or specific SNP databases, e.g., TSC (Thorisson and Stein, 2003) JNSNP (Hirakawa et al., 2002), HOWDY (Hirakawa 2002), GenoSNPs (Marsh et al., 2002). While the databases have been continuously expanded, the quality and completeness of the deposited SNP data remain to be of particular importance and to be assessed. New technologies make possible for the first time to genotype hundred of thousands of SNPs simultaneously. In present time data mining is a reasonable approach to investigating the number of SNPs that are informative for ancestry information (Baye et al., 2009). One can investigate distribution and density of SNPs across the genome by using HapMap, Affymetrix illumina SNP databases. The SNP databases in their current status might have some limitations for studies of complex disorders, especially in different ethnic groups, due to incomplete and/or uneven representation of SNPs and/or candidate genes in these groups (Dvornik et al., 2004). Several new SNP genotyping techniques have been developed viz. Illumina Infinium Assay, Sequenom iPLEX, Illumina Golden Gate, APEX-2, Taqman assay. All these above genotyping techniques are primer extension based SNP genotyping techniques (Goert et al., 1999; Syvanen, 2001). SNP-RFLPing assay is also a new web based SNP genotyping technique (Hisueh et al., 2006). Recently, SNP-Blast http://www.ncbi.nlm.gov/SNP/snp_blastByOrg.cgi was developed by coupling the NCBI dbSNP (Sherry et al., 2001) with a BLAST program of NCBI. SNP-BLAST is designed to perform the BLAST function among various SNP databanks for many species.

**Available SNPs databases:** In present days several SNP databases are available to search and choose the SNPs (Table 1).

**The SNP consortium (TSC):** The publicly funded efforts (NIH RFA: HG-98-001, 1998; http://www.nhgri.nih.gov/Grant_info/Funding/raf-hg-98-001.html) got a jump-start in 1999 when 13 pharmaceutical companies and the welcome trust formed the SNP Consortium (TSC) to accelerate SNP discovery and to ensure public accessibility to a minimum of 300,000 SNPs (http://www.snp.cshl.org/) (Lai, 2001). The goal was to generate a widely accepted, high-quality, extensive, publicly available map using SNPs as markers evenly distributed throughout the human genome. In the end, many more SNPs (1.8 million totals) were discovered (Thorisson and Stein, 2003). Now that the SNP discovery phase of the TSC project is essentially complete, emphasis has shifted to studying SNPs in populations. Various TSC member laboratories are genotyping a subset of SNPs as part of the Allele frequency project. The goal of the TSC allele frequency/genotype project is to determine the frequency of certain SNPs in three major world populations (http://www.snp.cshl.org). SNP data were made available through the consortium website at quarterly intervals during the project's first year and at monthly intervals during the second year. This cycle of releases ceased in fall 2001 once the discovery phase was finished, but with recent additions of genotype and allele frequency information, new data were released in fall 2002.

**Members of the SNP consortium:** The international member companies, which together committed at least $30 million to the consortium's efforts, are AP Biotech, AstraZeneca Group PLC, Aventis, Bayer Group AG, Bristol-Myers Squibb Co., F. Hoffmann-La Roche, Glaxo
Table 1: Presently available SNP databases for Human genome and SNPs involved in diseases

<table>
<thead>
<tr>
<th>Database</th>
<th>Emphasis</th>
<th>Details</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbSNP</td>
<td>SNPs from the complete genome</td>
<td>More than 6 million validated SNPs</td>
<td>Savanen (2001)</td>
</tr>
<tr>
<td>HapMap</td>
<td>Whole genome SNPs in four populations</td>
<td>More than 1 million SNPs</td>
<td>Thorisson et al. (2005)</td>
</tr>
<tr>
<td>HGVbase</td>
<td>SNPs from the complete Genome</td>
<td>More than 2.8 million SNPs</td>
<td>Fredman et al. (2002)</td>
</tr>
<tr>
<td>GVS</td>
<td>Access to dbSNP and HapMap SNPs</td>
<td>4.5 million SNPs</td>
<td>Weiner (2007)</td>
</tr>
<tr>
<td>Perlegen genotype data</td>
<td>Whole genome SNP in three populations</td>
<td>More than 1.5 million SNPs</td>
<td>Hinds et al. (2005)</td>
</tr>
<tr>
<td>JSNP</td>
<td>Common SNPs within Japanese population</td>
<td>More than 197,000 SNPs</td>
<td>Hirakawa et al. (2002)</td>
</tr>
<tr>
<td>PharmGKB</td>
<td>Genes involved in drug metabolism</td>
<td>SNPs from 167 genes</td>
<td>Hewett et al. (2002)</td>
</tr>
<tr>
<td>SNP500Cancer</td>
<td>Genes involved in cancer</td>
<td>More than 13,400 SNPs</td>
<td>Packer et al. (2004)</td>
</tr>
<tr>
<td>NIEHS SNP's program</td>
<td>Environmental response genes</td>
<td>More than 83,000 SNPs</td>
<td>Livingston et al. (2004)</td>
</tr>
<tr>
<td>Human Cytochrome P450 (CYP) Allele Nomenclature Committee</td>
<td>Human cytochrome P450 genes</td>
<td>SNPs from 25 CYP 450 genes</td>
<td>Ingelman et al. (2001)</td>
</tr>
<tr>
<td>Cytokine gene polymorphism</td>
<td>Cytokine gene polymorphisms in human disease</td>
<td>SNPs from more than 40 cytokine genes</td>
<td>Hollegaard and Bidwell (2006)</td>
</tr>
</tbody>
</table>

Wellcome PLC, IBM, Motorola, Novartis AG, Pfizer Inc., Searle and SmithKline Beecham PLC. The Wellcome Trust contributed at least $14 million. Laboratories funded by these companies identify SNPs are located at the Whitehead Institute, Sanger Centre, Washington University (St. Louis) and Stanford University. Data management and analysis take place at Cold Spring Harbor Laboratory.

The Single Nucleotide Polymorphism database (dbSNP): The National Center for Biotechnology Information (NCBI) has established the public SNP database dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) and plays a major role in facilitating the identification and cataloging of SNPs through its creation and maintenance of the dbSNP database (Sherry et al., 2001). Database submissions can include a broad range of molecular polymorphisms: single base nucleotide substitutions, short deletion and insertion polymorphisms, microsatellite markers and polymorphic insertion elements such as retrotransposons (Benson et al., 1999). Once such variations are identified and catalogued in the database, additional laboratories can use the sequence information around the polymorphism and the specific experimental conditions for further research applications (Smigielski et al., 2000).

To facilitate research efforts, NCBI’s dbSNP is included in the Entrez retrieval system which provides integrated access to a number of software tools and databases that can aid in SNP analysis. For example, each SNP record in the database links to additional resources within NCBI’s Discovery Space (We et al., 2003) as shown in Fig. 1. Resources include: GenBank, NIH’s sequence database; Entrez Gene, a focal point for genes and associated information; dbSTS, NCBI’s resource containing sequence and mapping data on short genomic landmarks; human genome sequencing data and PubMed, NCBI’s literature search and retrieval system. SNP records also link to various external allied resources. SNPs are also thought to be key enablers in realizing the concept of personalized medicine (The United States Congress

![Fig. 1: NCBI's discovery space facilitating SNP research](image-url)
defines personalized medicine as the application of genomic and molecular data to better target the delivery of health care, facilitate the discovery and clinical testing of new products. However, their greatest importance in biomedical research is for comparing regions of the genome between cohorts (in statistics and demography, a cohort is a group of subjects who have shared a particular experience during a particular time span).

SNPedia: SNPedia (pronounced snipedia) is a wiki-based bioinformatics web site that serves as a database of SNPs (Atanassova, 2009). Each article on a SNP provides a short description, links to scientific articles and personal genomics web sites, as well as microarray information about that SNP (Carioso, 2007) http://www.bioworld.com/issues/2007/dec-jan/inside-the-box-snpedia/. Thus SNPedia may support the interpretation of results of personal genotyping from, e.g., 23 and Me, Navigenics, deCODEme or Knome.

SNPs applications

SNPs and disease diagnosis: The SNPs have been used as molecular markers in many studies of monogenic or complex diseases owing to their high frequency and simple (binary) variation pattern (Kim and Mishra, 2007). Each person's genetic material contains a unique SNP pattern that is made up of many different genetic variations (Mircea et al., 2005). Kang et al. (2007) screened and scanned mitochondrial gene polymorphisms to determine the significance of mitochondrial DNA mutations in Korean migraineurs. According to Orpana and Sutath (2007) occasionally, a SNP may actually cause a disease and therefore, can be used to search for and isolate the disease-causing gene.

SNPs in association studies: Comparison of disease affected and unaffected group of individuals DNA for SNP patterns is called associate on study (Howard, 2005). Through association studies we can detect differences between the SNP patterns of the two groups; thereby indicating which pattern is most likely associated with the disease-causing gene (Mircea et al., 2005).

An example is the gene associated with the late onset Alzheimer’s apolipoprotein E or ApoE (Strittmatter and Roses, 1996). This gene contains two SNP that results in three possible alleles for the gene E2, E3 and E4. Most SNPs are not responsible for a disease state instead; they serve as biological markers for pinpointing diseases. Because they are located near a gene found to be associated with certain disease (Barnes and Gray, 2003; Chakravarti, 2001). An association study of a major autism linkage region on human chromosome 17 was done by Stone et al. (2007). According to Matthias (2004), only 10% of all gene-based SNPs have sequence-predicted functional relevance making them a primary target for genotyping in association studies.

SNPs and drug development: Single Nucleotide Polymorphism (SNP) technologies can be used to identify disease-causing genes in humans and to understand the inter-individual variation in drug response. By establishing an association between the genetic make-up of an individual and drug response it may be possible to develop a genome-based diet and medicines that are more effective and safer for each individual (Shastry, 2007). In a recent study, Tang et al. (2005) analyzed genetic data from 3,636 people who identified themselves as belonging to one of four major racial and ethnic groups in the United States (Caucasian, African American, East Asian and Hispanic). The appeal of pharmacogenetics lies in the possibility of personalized medicine (Hunt, 2008).

SNPs in paternity investigation: Based on the 52 SNP-plex developed by the SNPforID Consortium, Costa et al. (2008) designed two 10-plex to study Single Nucleotide Polymorphisms (SNPs) for human identification and to establish its usefulness in paternity casework. Twenty autosomal SNP set was studied in 56 paternity investigation cases from south Portuguese resident population, also analyzed with 17 Short Tandem Repeats (STRs). The use of SNPs as a complement to the analysis of autosomal STRs in paternity casework can result in paternity index and paternity probability values equivalent or higher than those obtained with more STR loci, but with lower costs. Instead of using additional STR loci, the analysis of 20 autosomal SNPs, as a complement technique to standard methodologies, is an appealing alternative in paternity investigation cases (Paulo et al., 2009). Forty autosomal most informative and population-independent SNPs were used by Whittle et al. (2009) in paternity investigation and as a prelude for forensic human identification.

SNPs in functional proteomics and gene therapy: SNP related functional proteomics involve the identification of functional SNPs that modify proteins and protein active sites structure and function. Several studies have attempted to predict functional impact by mapping nsSNPs onto protein structures and to discover properties of structural context that distinguish disease-associated and neutral nsSNPs (Stützle et al., 2003, Sunyaev et al., 2001; Wang and Moul, 2001). Functional proteomics is closely tied to modern (post-genomic) drug design and function SNP information helps to discover new
ACKNOWLEDGMENTS

The authors are hearty thankful to Dr. M.S. Tantia for critical reading and valuable comments and Dr. R.K. Vih for helpful suggestions and kind support to write this study. Authors also acknowledge the moral boost up provided by our colleagues.

REFERENCES


