[Frontiers in Bioscience S2, 1-4, January 1, 2010]

SNPs: At the origins of the databases of an innovative biotechnology tool

Anthony Corfield[1], Peter Meyer[2,3], Shelina Kassam[4], Gregor Mikuz[5], Consolato Sergi[6]

[1]Clinical Science at South Bristol, Bristol Royal Infirmary, University of Bristol, United Kingdom, [2]Charite University Hospitals, Institute of Human Genetics, Berlin, Germany, [3]Laboratories for Molecular Oncogenetics, Ulm, Germany, [4]Queen Medical Center, Nottingham University Hospital, United Kingdom; [5] Institute of Pathology, University of Innsbruck, Austria, [6]Pediatric Pathology Section, Institute of Pathology, University of Innsbruck, Austria

TABLE OF CONTENTS

1. ABSTRACT

The discovery that DNA sequence variations can influence the response of an individual to a drug or can predict the outcome of a disease has added a new dimension to evidence-based medicine. It is clear that the goals, risks, and benefits of drug therapy can be better assessed if the underlying genome of the patient is known. The relevance of identifying patients at increased risk of adverse drug reactions, the application of genomic technologies to drug development and the clarification of the mechanisms of drug action on cells will be important targets in the therapeutic approach to medicine in the 21st century. In this review, we summarize the development of single nucleotide polymorphisms (SNPs) and give computational biological data for SNPs databases.

2. PHARMACOGENETICS AND PHARMACOGENOMICS

Genetic factors are most probably the major determinants of the variability of drug action with many quantitative and qualitative differences in pharmacological activity (1). Although a high proportion of cases appear to show differences of a polygenic nature, monogenic factors may also play an important role. The interaction of pharmacology with the human genome is currently represented in pharmacogenetics and pharmacogenomics. The first discipline refers to the study of the inherited variations in drug metabolism and individual response, while the second details how drugs affect biological gene expression. Accordingly, if pharmacogenomics relates to how the human genome, as a whole, affects the organism's

response to drugs, then pharmacogenetics takes into account how the genes of an individual affect the organism's drug response. The genotype-based approach to pharmacology is a recent and probably essential step to improve the safety and efficacy of therapeutic substance uses and identifies y population groups able to respond to a specific therapy. This new approach has been promoted by recent knowledge acquired in the field of genomics and is being facilitated by the application of modern technological tools to biochemical assays and clinical diagnostics. One of the most important discoveries over the last 20 years has been the detection of single nucleotide changes in the human genome (2).

If mutations are DNA sequence variations in the human genome closely related to the onset of monogenic inheritable or acquired oncological and non-oncological diseases, there are some DNA sequence variations that will not determine a monogenic disease. In the past 5-10 years, it has become increasingly important to identify minor alterations of the human genome sequence occurring in every 100 to 300 bases along the 3-billion-base human sequence and in at least 1% of the population. These DNA sequence variations that will not determine a monogenic disease are called single nucleotide polymorphisms (SNPs). SNPs make up about 90% of DNA sequence variation as indicated by the Human Genome Project (http://www.ornl.gov, http://linkage.rockefeller.edu/soft/). The SNPs that are most likely to have a direct impact on the protein product of a gene are those that change the amino acid sequence of the proteins and changes in the gene regulatory regions, which control protein expression levels.

3. SINGLE NUCLEOTIDE POLYMORPHISMS (SNPS)

DNA genome variations include mutations and polymorphisms that may easily be distinguished by frequency in addition to the association with disease. Thus, if for example, a position in the genome where 93 % of people have a cytosine as nucleotide and the remaining 7 % have a guanine as nucleotide is a polymorphism. However, if one of the possible sequences is present in less than 1 % of the population (99.9 % of people have a C and 0.1 % have a G), then the variation is called a mutation. SNPs are DNA genome variations that involve just one nucleotide, or base. Any one of the four DNA bases may be substituted for any other – an A instead of a T, a T instead of a C, a G instead of an A, and so on. It has been suggested that, theoretically, a SNP could have four possible forms, or alleles, since there are four types of bases in DNA, in which transitions and transversions are possible. However, the majority of SNPs have two alleles only. To date, the number of the SNP is more than one and a half million. However, relatively few of these transform the gene product, because there is no amino acid change ("silent" changes).

The term polymorphism was first introduced to imply a variation (3). It was not implicit to have either harmful or beneficial effects for the individual. To date, it is known that many polymorphisms actually do affect a person's phenotype, though in more complex and sometimes unexpected ways. SNPs play an important role in epidemiology and the knowledge of DNA variation is helpful in determining how individuals respond to disease or contact with environmental factors, such as bacteria, toxins, chemicals, drugs and therapeutic procedures. In medicine, these DNA sequence variations can have a major impact in addressing the multiple gene-associated diseases, such as cancer, diabetes, vascular disease, and neurologic and psychiatric disturbances. In these diseases, SNP maps will be a powerful research tool to enhance the understanding of the pathogenesis of many diseases and their reaction to drugs.

4. SNPs DATABASES

At the center of pharmacogenetics and pharmacogenomics are the databases of SNPs which are a vital tool to deal with relevant questions. An interesting research project was initially started to identify how many and which SNPs are present in the human genome and their significance in different populations. Several groups worked to find SNPs and create SNP maps of the human genome. Two major groups were involved in identifying SNP maps: the U.S. Human Genome Project (HGP) and a large group of companies called the SNP Consortium (AstraZeneca Group PLC, Aventis, Bayer Group AG, Bristol-Myers Squibb Co., F. Hoffmann La-Roche, Glaxo Wellcome PLC, IBM, Motorola, Novartis AG, Pfizer Inc., Searle, Smith Klein Beecham PLC, and UK Wellcome Trust philanthropy) headed by Dr Arthur L. Holden (http://www.ornl.gov and http://snp.cshl.org). The program aimed to release SNP data at quarterly to monthly intervals. This program ceased in the autumn of 2002 when the acquisition of genotype and allele frequency data was completed. Currently, there are several SNPs databases of which two major human gene-based polymorphism databases HGVbase and the dbSNP are the probably the major players.

HGVbase (Human Genome Variation Database – http://hgvbase.cgb.ki.se/) is a web-based database providing an accurate and comprehensive catalog of normal human genes and genome variation as a research tool for determining the genetic component of the human phenotypic variation. HGVbase was first created as a joint venture between two research teams, one located at the Karolinska Institute in Sweden and the other one based at Interactiva GmbH in Germany. The first database was released in August 1998 and a European consortium was formed shortly after, involving teams at the Center for Genomics and Bioinformatics at the Karolinska Institute (Sweden), the European Bioinformatics Institute (UK) and at the European Molecular Biology Laboratory (Germany). The primary objective of this Consortium was to facilitate access to scientists for the analysis of genotype-phenotype associations, exploring how SNPs may influence phenotypes related to e.g. common disease risk and drug response differences.

Table 1. Presentation of a selection of commercial companies with world-wide web addresses producing databases covering pharmacogenomics and pharmacogenetics areas

Company	Sequence DB	Pharmacogenetics	Pharmacogenomics
www.applera.com	Yes	Yes	No
www.curagen.com	Yes	Yes	No
www.decode.com	No	Yes	No
www.biospace.com	No	No	Yes
www.genelogic.com	No	No	Yes
www.dnavision.be	No	Yes	No
www.genomatrix.de	No	Yes	No
Genset	No	Yes	No
Human Genome Sciences	Yes	Yes	No
Hyseq	Yes	Yes	No
Incyte	Yes	Yes	Yes
Inpharmatica	Yes	No	No
Large Scale Biology	No	No	No
Lynx Therapeutics	No	No	Yes
Oxford GlycoSciences	No	No	No
Oxagen	No	Yes	No
Pharmagene	No	No	Yes
Phase I	No	No	Yes
Proteome	Yes	No	No
Rosetta Inpharmatics	No	No	Yes
Variagenics	No	Yes	No

Note: Pharmacogenetics deals with drug metabolizing enzymes and the effects that genetics variation within these genes has on the biological metabolism of xenobiotic agents in the body, whereas the term pharmacogenomics is here used to describe the effects of xenobiotic agents on patterns of biological gene expression. Gene expression data can be generated by several techniques, such as high-throughput cDNA sequencing, differential display reverse transcription PCR, microarrays of oligonucleotides or cDNAs.

The second most important web-based database of SNPs is probably dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) based at the National Center for Biotechnology Information (NCBI) that is an international resource for molecular biology information and computational biology. dbSNP is a database built from sequences submitted by individual laboratories and by data exchange with the international nucleotide sequence databases, the European Molecular Biology Laboratory (EMBL) and the DNA Database of Japan (DDBJ). Arrangements with the U.S. Patent and Trademark Office enable the incorporation of patent sequence data.

SNP genotyping platforms are being developed at an exponential rate. The importance of genomics-based pharmaceutical approaches to disease is clearly enormous (4). Applied Biosystems are developing third generation 5' nuclease assays with TaqMan® minor groove binder probes and high quality instrumentation software systems based on the TaqMan 7900HT system (www.appliedbiosystems.com). Primer extension products detected using MALDI-TOF (MassARRAY™) technology has been targeted by Sequenom to validate over 200,000 human SNPs (www.sequenom.com) (5). The MALDI-TOF technology has been introduced at the Centre Nationale de Genotypage (France) (www.cng.fr). This is a variant of primer extension based genotyping with allele discrimination by matrix-assisted laser desorption/ionization time-of-flight. The use of SpectroDESIGNER™ assay software has notably improved Sequenom's success rate for assay development (www.sequenom.com). A method to increase sensitivity with mass spectrometry is the "GOOD Assay" (6). A further step designed to increase the density of SNPs is being elaborated by Orchid BioSciences who are investing in technology to increase the density of SNP assays per well. Thus, they have presented array-based platforms, including SNPcode™, SNPstream®, and ultra-high throughput (UHT) SNPstream™ (www.orchid.bio).

The website facility at the data-coordinating center of the SNP Consortium (http://snp.cshl.org) has shown incredible development within a few months and computational biology tools are extraordinary in achieving their target of reducing work time. Many websites, such as the SNP Consortium website are served by professional web servers. The service provided by a professional website is clearly not trivial, because of the numerous queries that can be solved. The Apache web server (www.apache.org) running on a Linux operating system platform supports the SNP Consortium. Originally there was an FTP (File Transfer Protocol) access only to whole database as well HTTP-based graphical interface to browse data with ASeDB/AcePerl serving as the database back end (7). Subsequently, MySQL (SQL: Standard Query Language) and Oracle text-table format were used for the web browsing interface (8). A Java applet allowed users to view individual traces. The web-programs that perform searches and display chromosome features in the website of the SNP Consortium are written in the Perl programming language, making use of the GMOD (Generic Model Organism Database) and Bioperl software libraries. Perl is a high-level programming language deriving mostly from the ubiquitous C programming language and to a lesser extent from mostly sed, awk, and Unix shell. The Bioperl Project is an international association of developers of open source Perl tools for bioinformatics, genomics and life science research.

An SNP report shows considerable data useful for clinicians and scientists, including the alleles observed, the flanking sequences, genomic location reported allele frequencies, the laboratory submitting these data, and additional molecular biology data. Many SNPs can be dumped to text file or in a variety of formats. Recently, the effective organization of the GBrowse package for genome browsing on the website of the Consortium has enormously simplified the SNPs database (9). An additional facilitation has been achieved by introducing the XML technology (XML: Extensible Markup Language). XML is an electronic document exchange standard. Practically, it is a markup-design language from which markup languages are derived. XML documents are made up of elements which are located between opening and closing tags, attributes, which can be applied to elements, and empty elements, that have no text between the opening and closing tags. The HapMap project, in which the SNP Consortium is a participant, contains additional important information. Data downloads are available at http://hapmap.cshl.org. Further human genome platforms and their browsers may be found on the World Wide Web (10-11). Table 1 shows a short presentation of a selection of commercial companies with world-wide web addresses producing databases covering pharmacogenomics and/or pharmacogenetics areas.

5. FINAL CONSIDERATIONS

Evaluating a disease carefully remains a difficult diagnostic problem. SNP genotyping may probably one of the first steps of the diagnostic process. When coupled to population-based genetic knowledge, it is an efficient way to detect DNA copy number abnormalities, scan candidate

regions for homozygosity, and determine shared haplotypes among as few as two patients with a similar phenotype. It is important to realize that human genetic variability can affect one or more stages of pharmacokinetics: absorption, metabolism, transport to the target molecule, structure of the intended and/or unintended target molecules, degradation of the drug, and the excretion of the degradation products. All these potential points of variability can be affected and determine a different drug response from individual to individual. It is not possible to predict medical advances over the next decade, but diagnostic tests will probably allow extremely valuable screening in many populations and the introduction of individualized therapy is just starting.

6. REFERENCES

1. Vesell, E. S. and Penno, M. B. Assessment of methods to identify sources of interindividual pharmacokinetic variations. *Clin. Pharmacokinet.* 8:378-409 (1983)

2. Sachidanandam, R.; Weissman, D.; Schmidt, S. C.; Kakol, J. M.; Stein, L.D.; Marth, G.; Sherry, S.; Mullikin, J.C.; Mortimore, B.J.; Willey, D. L.; Hunt, S. E.; Cole, C.G.; Coggill, P. C.; Rice, C.M.; Ning, Z.; Rogers, J.; Bentley, D. R.; Kwok, P.Y.; Mardis, E. R.; Yeh, R.T.; Schultz, B.; Cook, L.; Davenport, R.; Dante, M.; Fulton, L.; Hillier, L.; Waterston, R.H.; McPherson, J. D.; Gilman, B.; Schaffner, S.; Van Etten, W. J.; Reich, D.; Higgins, J; Daly, M. J.; Blumenstiel, B.; Baldwin, J.; Stange-Thomann, N.; Zody, M. C.; Linton, L.; Lander, E. S. and Altshuler, D. (International SNP Map Working Group) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409:928-933 (2001)

3. Kassam S, Meyer P, Corfield A, Mikuz G, Sergi C. Single Nucleotide Polymorphisms (SNPs): History, Biotechnological Outlook and Practical Applications. *Current Pharmacogenomics, 3,* 237-245 237 (2005)

4. Ambrose, H. J. SNPs and Pharmacogenomics. *Pharmacogenomics* 3, 583-586 (2002)

5. Buetow, K. H.; Edmonson, M.; MacDonald, R.; Clifford, R.; Yip, P. and Kelley, J. Little, D.P.; Strausberg, R.; Koester, H.; Cantor, C. R. and Braun, A. High-throughput development and characterization of a genome wide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* 98:581-584 (2001)

6. Drysdale, C. M.; McGraw D. W.; Stack, C. B.; Stephens, J. C.; Judson, R. S.; Nandabalan K.; Arnold K.; Ruano G. and Liggett, S. B. Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. *Proc. Natl. Acad. Sci. U. S. A.* 97:10483-10488 (2000)

7. Thorisson, G. D. and Stein, L. D. The SNP Consortium website: past, present and future. *Nucleic Acids Res.* 31:124-127 (2003)

8. Stein, L.D. and Thierry-Mieg, J. Scriptable access to the Caenorrhabditis elegans genome sequence and other ACeDB databases. *Genome Res* 8:1308-1315 (1998)

9. Stein, L.D. The Generic Genome Browser: A building block for a model organism system database. *Genome Res* 12:1599-1610 (2002)

10. Hubbard, T.; Barker, D.; Birney, E.; Cameron, G.; Chen, Y.; Clark, L.; Cox, T.; Cuff, J.; Curwen, V.; Down, T.; Durbin, R.; Eyras, E; Gilbert, J; Hammond, M; Huminiecki, L.; Kasprzyk, A.; Lehvaslaiho, H.; Lijnzaad, P.; Melsopp, C.; Mongin, E.; Pettett, R.; Pocock, M.; Potter, S.; Rust, A.; Schmidt, E.; Searle, S.; Slater, G.; Smith, J.; Spooner, W.; Stabenau, A.; Stalker, J.; Stupka, E.; Ureta-Vidal, A.; Vastrik, I. and Clamp, M. The Ensembl genome database project. *Nucleic Acids Res.* 30:38-41 (2002)

11. Kent, W. J.; Sugnet, C. W.; Furey, T. S.; Roskin, K. M.; Pringle, T. H.; Zahler, A. M. and Haussler, D. The human genome browser at UCSC. *Genome Res.* 12:996-1006 (2002)

Abbreviations: DDBJ: DNA Database of Japan, EMBL: European Molecular Biology Laboratory, GMOD: Generic Model Organism Database, HGP: Human Genome Project, MALDI-TOF: Matrix-Assisted Laser Desorbtion Ionisation - Time Of Flight, SNP: single nucleotide polymorphism

Key words: Single Nucleotide Polymorphisms, Polymerase Chain Reaction, Genotyping, Pharmacogenomics, Pharmacogenetics, Children, Cardiology, Review

Send correspondence to: Consolato Sergi, Institute of Pathology, University of Innsbruck, Muellerstrasse 44 A-6020 Innsbruck, Austria, Tel: 43 512 9003 71316, Fax: 43 512 582 088; E-mail: consolato.sergi@i-med.ac.at

http://www.bioscience.org/current/volS2.htm