Single-Nucleotide Polymorphisms: an overview of the analytical power of SNP's in genomic research and the preliminary results of its application.

by

Lamont Tang

For

Biochemistry 118Q

INTRODUCTION

The Human Genome Project (HGP), a $3 billion federal program launched in 1990, is an effort designed to sequence the entire human genome by the year 2001. While there are diverse and profound implications of such an achievement, the HGP will actually produce very little information about DNA sequence variation within the human population. From a bioinformatics perspective, this issue clearly needs to be addressed since a key aspect of research in genetics is associating sequence variations with heritable phenotypes. Information about genetic variation is critical for understanding how genes function or malfunction and for understanding how genetic and functional variation are related. In addition, analysis of DNA sequence variation is becoming an increasingly important source of information for identifying the genes involved in both disease and in normal biological processes.

Since the most common sequence variations are single nucleotide polymorphisms (SNP's), it is not surprising that recent advents of mutation detection and highly efficient genotyping technologies (Chee et al. 1996) have coincided with the emergence of the new generation of markers based on single-nucleotide polymorphisms and has subsequently generated great interest in SNP discovery and detection. In addition to their frequency, SNP's have several other properties that make it an attractive candidate to be the primary analytical reagent in the study of human sequence variation. In this report, the value of SNPs in generating information about genetic variation will first be set in the historical context of positional cloning. In conjunction with a brief description of past generations of genetic markers, the potential powers and limitations of SNP will be discussed in context of genetic analysis. Lastly, a general perspective of the feasibility of low-cost and large-scale discovery and detection of SNPs will be discussed in light of the current state of technology.

POSITIONAL CLONING AND THE POTENTIAL OF SINGLE-NUCLEOTIDE POLYMORPHISMS

In our present state of scientific knowledge, genetic factors appear to contribute to virtually every human disease. Much of biomedical research in both the public and private sectors has been fuelled and is continuing to be motivated by the expectation that understanding the genetic component and contribution to disease will revolutionize diagnosis, treatment, and prevention. Within the past 10 years, the advent of molecular genetic technologies has already led to the identification of nearly 100 genes causing various genetic diseases by positional cloning.

The technique of positional cloning as a general strategy for the isolation of human disease genes is based upon the fact that any detectable differences in DNA sequences between individuals can be used as genetic markers in human DNA. It follows directly from this that linkage analysis can exploit these genetic markers as references to determine and map the position of other genes. In terms of discovering genes underlying Mendelian disorders, an important application of this approach is the identification and eventual cloning of a disease gene. This isolation of human disease genes has increased our understanding of the molecular and cellular basis of disease as well as providing insight into prevention and treatment of the disease.

Positional cloning has been of great use since it assumes no functional information. Thus it locates the disease gene by virtue of its location in the genome rather than by using knowledge of its biochemical features. This feature of positional cloning is perhaps what provided much of the general impetus for mapping and sequencing the human genome. It was first successfully applied in 1986 in which the first sequences used as genetic markers were ‘restriction fragment length polymorphism’ (RFLP). This first generation marker was based on the inherited differences in cleavage sites for restriction endonucleases. A single base change within a restriction endonuclease site is a readily detectable marker since the enzyme in question can no longer cleave the mutated site. A mutation that gives rise to an RFLP thus serves as a genetic marker that can be detected by a Southern blot.

Following the RFLP markers were the microsatellites. They were particularly useful genetic markers since they consist of tandem repeats of short nucleotide sequences. Microsatellites are distributed throughout the genome in a much higher frequency than RFLPs and therefore yield higher resolution maps and lend themselves to more precise linkage analysis. The current linkage map of the human genome consists of more than 10000 loci which are defined primarily by short tandem-repeat polymorphisms.

With the introduction of the third generation SNP marker, whole-genome scans in pedigree-based linkage analysis of families have shown that a map of about 2000 SNP’s has the same analytical power for this purpose as a map of 800 microsatellite markers. The higher resolution of the ‘third generation’ map will allow identification of disease genes with even greater precision.

In addition to their frequency of occurrence, SNPs have several other properties that make it an attractive candidate to be the primary analytical reagent in the study of human sequence variation. They are stable and have much lower mutation rates than to repeat sequences; detection methods for SNPs are potentially more amenable to being automated and used for large-scale genetic analysis, and most importantly, the nucleotide sequence variations that are responsible for the functional changes of interest will often be SNPs.

Information about SNPs can be used in two other ways in genetic analysis. When the genetics of a disease are studied in individuals in a population rather than in families, the haplotype distributions and linkage disequilibria can be used to map genes by association methods. For this purpose it has been estimated that 30000 to as many as 300,000 mapped SNPs will be needed.

Perhaps the most important information that SNP can offer to genetic analysis are the case-control studies that directly identify functional SNPs contributing to a particular phenotype. Currently, it is possible to genotype individuals and populations at >400 sites by simple sequence-length polymorphisms (SSLPs). While SNPs have only two alleles and are thus less informative than the typcial multi-allelic SSLP, this disadvantage can be overcome by using a greater density of SNPs. Wang et al. suggest that a genome scan of 1000 well-spaced SNPs will extract about the same linkage information as the current SSLP standard. In the near future, it will be possible to genotype at several thousand single-nucleotide polymorphic regions (Wang et al.)

According to the "Workshop on Human DNA Sequence Variation" on March 31 and April 1, 1997, "it was suggested that approximately 10-20 SNPs per block would be sufficient for characterizing the human genome. This could translate to a map of 30000-60000 markers to analyze blocks of a megabase in size, which might be the case in studying a unique population with recent ancestry."

It follows from this that the potential power of dense SNP maps are even greater for the more typical case of outbred populations. Since the blocks of outbred populations are smaller due to their inherent nature of containing a lot of mixed and old mutations, the blocks will be considerably smaller, thus requiring a comparably larger number of makers. Since block size is critically dependent on the population under investigation, such dense SNP maps allows the study of diseases from appropriate patient samples by association, without the necessity of family samples. This capability would reduce the cost of disease studies and, more importantly, genetic studies could be better designed with respect to phenotype (Workshop 97).

PRELIMINARY RESULTS

The development of DNA sequencing technology and the initiation of large-scale DNA sequencing projects has recently made it possible to directly measure variation in genomic DNA. Genomic sequence analysis show that when two random chromosomes are compared, they differ at ~ 1/1000 nucleotides (Kwok et al. 96). These results are consistent with the RFLP studies in the 1980's claiming that there is one variant nucleotide (nt) per 1000 nt screened.

However, most of the successes to date in identifying rare disease genes have been due to the fact that they are highly penetrant and monogenic in nature (e.g. Huntington's disease, cystic fibrosis, familial cardiomyopathy). These disease genes are located by performing linkage analysis on families, which requires 300-500 highly informative genetic markers spanning the entire human genome (Collins 98). Unfortunately, studies of common diseases (complex and polygenic) such as hypertension, diabetes, and artherosclerosis have proved to be much harder in locating the disease gene since the disease phenotype is affected by multiple genes in which each disease gene plays a small effect. These kinds of diseases are also much more amenable to environmental factors.

Risch and Merikangas proposed an alternative and more efficient study than linkage analysis of families in regards to polygenic diseases. Given that there exist hundreds and thousands of variants spread over the entire genome, they suggested performing association analysis on many affected and unaffected individuals. The result of their proposal led to the design of the DNA Polymorphism Discovery Resource.

According to Collins, Brooks, and Chakravarti (98 Genome Res) report on the DNA Polymorphism Discovery Resource, out of the 3 billion bases in human DNA, about 17 million SNPs are expected to be found when all chromosomes from 40 individuals are screened. Since coding regions are ~5% of the genome and are less likely to have SNPs (Nickerson et al. 1998), only a small proportion of these SNPs are expected to be in coding regions. The number of cSNPs is estimated to be approximately 500,000 or an average of about 6 per gene.

The preliminary results from Wang et. al. have demonstrated the feasibility of large-scale identification of human SNPs. The characterization of human diversity at the nucleotide level will be of great analytical power since the entire genome, and not just the recognized coding sequences (e.g. ESTs and STS databases) are accessible to analysis. Thus generating a high resolution third generation map of sequence variation at single nucleotide positions is of great interest in identification of genes underlying complex diseases.

The feasibility of generating high density SNP maps (with <100kb spacing) in an efficient manner was also demonstrated by Lai et al. In their study, they assessed the efficiency of current SNP discovery approaches and technologies by comparing YAC-based versus BAC/PAC-based maps, sequencing individual DNAs versus a pooled DNA sample, and evaluated three different software applications for polymorphism detection. Their results showed that high-density SNP maps can be efficiently generated using existing technologies and that a genome-wide map with 60000-100000 is achievable in a reasonable time frame (4 months). The cost ranged from $3000 per gene-based SNP to $1500 for random SNPs.

CONCLUSION

Ultimately however, the efforts to reduce the costs and increase feasibility of discovering and detecting SNPs rests on the promise that the availability of a large collection of SNPs will aid in identifying candidates for polymorphisms with functional significance. This promise of genomic medicine has already been fulfilled in part by Geistererfer-Lowrance et al. who identified specific mutations of the beta-myosin heavy chain, tropomyosin, myosin binding protein, and troponin T genes in familial hypertrophic cardiomyopathy. They were able to predict early mortaility due to sudden cardiac death based on the information of a single amino-acid mutation. The results illustrate the potential predictive power of the genetic variants—especially if they lead to monogenic mendelian disorders.

Unfortunately, as stated before, the more complex polygenic diseases do not lend themselves to such easy and simple genetic analysis. While SNPs are powerful tools for mapping and discovering the genes associated with common diseases, the techniques and methods for discovering and scoring SNPs efficiently are just beginning to emerge. It is clear that SNPs are definitely amenable to automation and large-scale genetic analysis but the question of interpreting variation across the genome still remains. Once the sequencing of the human genome is finished, the genetic diversity of a population and how those differences correspond to therapeutic outcomes remains to be understood. Such investigations in the future will hopefully fulfill the promises of improving drug development and therapeutics tailored to individual genotypes.

BIBLIOGRAPHY

Chee, M.,;Yang, R., Hubbel, E., Berno, A.,Huang, Xiaohua C., Stern, D., Winkler, J., Lockhart, D.j., Morris, M.S., Fodor, S.P.A. *Science* 1996 October 25; **274**: 610-614.

Collins, F. S., Mark S. Guyer, Chakravarti, A. *Science* 1997 November 28; 278: 1580-1581.

Collins, F. S., Brooks, L. D., Chakravarti, A. (1998). A DNA Polymorphism Discovery Resource for Research on Human Genetic Variation. *Genome Res.* 8: 1229-1231

Geisterfer-Lowrance AA, Kass S, Tanigawa G, Vosberg HP, McKenna W, Seidman CE, Seidman JG. 1990. *Cell.* **62:** 999-1006.

Kerem, B.-S., J.M. Rommens, J.A. Buchanan, D. Markiewicz, T.K. Cox, A. Chakravarti, M. Buchwald, and L.-C. Tsui. 1989. *Science* **245:** 1073-1080

Kwok, P.-Y., Q. Deng, H. Zakeri, S.L. Taylor, and D.A. Nickerson. 1996. *Genomics* **31:** 123-126

Lai, E., Riley, J., Purvis, I., Roses, A.. 1998. *Genomics* **54**: 31-38.

Nickerson, D.A., S.L. Taylor, K.M. Weiss, A.G. Clark, R.G. Hutchinson, J. Stengård, V. Salomaa, E. Vartiainen, E. Boerwinkle, and C.F. Sing. 1998. *Nat. Genet.* **19:** 233-240

Risch, N., Merikangas, K. 1996. *Science* **273:** 1516

Rommens, J.M., M.C. Iannuzzi, B.-S. Kerem, M.L. Drumm, G. Melmer, M. Dean, R. Rozmahel, J.L. Cole, D. Kennedy, N. Hidaka 1989. *Science* **245:** 1059-1065

Schafer, A.J., and Hawkins, J.R. 1998. DNA variation and the future of human genetics. *Nat. Biotech*. **16**: 33-39.

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglut T, Hubbell E, Robinsin E., Mittman M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lipshutz R, Chee M, Lander ES. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*. 1998; **280**: 1077-1082.

INTERNET SOURCES:

“Workshop on Human DNA Sequence Variation”. 1997.
http://www.nhgri.nih.gov/98plan/variation_report.html