# Application of SNP Technologies in Medicine: Lessons Learned and Future Challenges

Eric Lai

*Discovery Genetics, Genetics Research, GlaxoSmithKline, Research Triangle Park, North Carolina 27709, USA*

Over the past few years, single nucleotide polymorphisms (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). SNPs are frequently present in the genome with a density of at least one common (>20% allele frequency) SNP per kilobase pair (Lai et al. 1998; Sachidanandam et al. 2001). They are mostly biallelic (<0.1% of SNPs are triallelic; http://snp.cshl.org/) and thus easy to assay. More importantly, SNPs allow the unification of the candidate gene approach and association-based fine mapping to identify gene(s) of interest. They also aid in the association of linkage analysis to the phenotypic and genotypic data.

Although quantitative analysis and mathematical modeling have suggested that whole-genome association is more effective than linkage analysis for the identification of complex disease genes and in pharmacogenetics, the application of SNPs had been hindered by the lack of sufficient markers. In 1997, 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/1997news/pr072897.htm). The publicly funded efforts (NIH RFA: HG-98–001, 1998; http://www.nhgri.nih.gov/Grant_info/Funding/rfa-hg-98–001.html) got a jump-start in 1999 when 13 pharmaceutical companies and the Wellcome Trust formed The SNP Consortium (TSC) to accelerate SNP discovery and to ensure public accessibility to a minimum of 300,000 SNPs (http://snp.cshl.org/). The combined efforts of the public projects and TSC have been extremely productive and there are currently >1.6 million SNPs in the public databases (Sachidanandam et al. 2001). In this article, I will attempt to summarize what we know about SNPs and identify some of the challenges that await us in the application of SNPs in research and medicine.

The first questions most people would ask are, how many SNPs are there in the human genome and have we identified most of the SNPs? The frequently cited rate of 1 SNP/kb suggests that there are 3 million common SNPs in the human genome. However, recent data have indicated that the number of SNPs in the human genome is potentially much more than 3 million. The first indication came from the comparison of the Celera SNP database with the public data. Celera Genomics claimed to contain over 3.5 million putative SNPs in their database. However, only 400,000 of their SNPs were redundant when compared to the publicly available 1.6 million. The second line of evidence came from our own experiments. We have isolated >1000 SNPs in a 20-megabase region by re-sequencing eight individuals (not the same DNA source as the TSC SNPs). The overlap between our SNPs (~1,000) and the TSC SNPs in this region is ~5% (instead of the expected 50% if the total number of common SNP is around 3 million). These results suggest that there are potentially 10 million or more common SNPs in the human population. A theoretical modeling experiment has also predicted that there are more than 10 million SNPs in the genome (Kruglyak and Nickerson 2001).

There are two important implications in the usage of SNPs as a genetic tool if there are indeed over 10 million SNPs in the human genome. The first implication is that the SNP(s) you are looking for might not be discovered yet. The second implication is the need to select a representative set of SNPs out of the 1.6 million to cover the genome. The first problem is a difficult one since it is impossible to know whether the SNP(s) of interest is present in the current databases. There are two potential solutions. The first solution is to design experiments that combine SNP discovery and genotyping (Brenner et al. 2000). However, this approach has not been demonstrated for whole genome SNP scan and could be costly even if it is technically feasible. The second solution, which is suitable for both implications mentioned above, is the development of a comprehensive whole genome SNP marker set that has a high likelihood of detecting the SNP(s) of interest by linkage disequilibrium or association (see section below on marker set development) (Jorde 2000).

So how do we design a marker set that covers the genome as completely as possible? There are many suggestions and computer models using linkage disequilibrium (LD) as a guide and striking a balance between number of markers and information content (Kruglyak 1999; Jorde 2000). A number of recent studies have indicated that an average spacing of 30 kb provides a good balance (i.e., 100,000 SNPs for whole genome) (Collins 1999; Huttley et al. 1999; Goddard et al. 2000; Jorde 2000). In addi-

**E-MAIL ehl21107@GlaxoWellcome.com; FAX (919) 315-0113.**
Article and publication are at www.genome.org/cgi/doi/10.1101/gr.192301.

tion, we need to cover all known and putative genes. Thus, a comprehensive SNP map should contain at least 100,000 SNPs covering at least 2 SNPs per known or putative gene. We have to cover the gene-rich regions with more SNPs and be aware that there will be regions that are relatively devoid of polymorphisms (Miller et al. 2001). Computer modeling could also be useful in SNP selection (Hoh et al. 2000). The number of SNPs in the comprehensive map should not increase, but the content may improve as more information is available on the LD of the genome and the informative content of the individual marker. Regions with positive associations need to be followed up by exhaustive SNP discovery and genotyping and/or haplotyping (see below).

I have made two important assumptions in the previous section, namely that there are >10 million SNPs in the genome and that we need 100,000 for whole genome scans. This leads to the next logical question: How do we select the 100,000 out of the 10 million potential SNPs? The answer to this question is currently limited by assay technology and a lack of information regarding the SNPs. Let's assume that we are going to develop the first 100,000 whole genome SNP marker set from the current publicly available 1.6 million SNPs. The first step in the process is the enzymatic amplification of the loci. Since ~50% of the SNPs are located in repeat regions, only 800,000 of the SNPs would survive this in silico step. The false positive rate of the TSC SNP is ~5%–10%, so we expect ~720,000 of the SNPs to be polymorphic. Recent publications have shown that the possibility of any TSC SNP being polymorphic in a single major population (e.g., Caucasian, African-American, or Asian) is ~80% (Sachidanandam et al. 2001). Thus, if your sample population consists of mainly a single population, the number of available SNPs is ~500,000. Location and spacing of these SNPs would provide a good first-generation whole genome SNP marker set. This marker set can then be refined with genotyping data and LD information. Although this procedure is simple and logical, we do not have the necessary allele and population frequencies for all of the SNPs. TSC has funded the determination of allele and population frequencies for 100,000–150,000 SNPs. However, it is crucial to generate these data for 300,000–400,000 SNPs for the selection of a whole genome marker set.

Given that we could design and develop a comprehensive whole genome map of 100,000 SNPs, which genotyping platform should we use for the whole genome scans? In addition to the concurrent SNP discovery and genotyping approach by DNA sequencing mentioned above, there are three general approaches for genotyping (pooled sample reaction, individual reaction, and haplotyping). Pooled sample genotyping has the advantage of massive reduction in the number of total genotypes (Shaw et al. 1998). The size of the sample pool directly reduces the number of folds of genotyping required and thus the cost of the experiment. The disadvantages of pooled genotyping include the requirement of having all samples at the beginning of the experiment, the inflexibility in changing the design of the pools, and decreased statistical power for rarer alleles. Individual DNA sample genotyping is the standard accepted method with many commercially available platforms. The key parameters for consideration include cost (<$0.2 inclusive cost per genotype in 2001), no-call rate (should be <10%), error rate (should be <1%), throughput rate (at least 50,000 genotypes per day), and potential for multiplexing (at least 10×).

Haplotyping has promised to be the most sensitive method for association detection (Martin et al. 2000; Zollner and von Haeseler 2000; Fallin et al. 2001). In addition, phylogenetic analysis of the haplotypes might be even more sensitive or suitable for correlation of human genetic variations and complex phenotypes (Valdes and Thomson 1997; Service et al. 1999). However, most haplotyping approaches are still costly and tedious, and phylogenetic information is difficult to obtain. Although we are not ready to perform millions of genotypes per day with hundreds of thousands of markers, I am not concerned about the so-called "technology gap". For example, the Human Genome Project was initiated with Sanger's dideoxy sequencing and electrophoretic separation. At the time, most people were concerned about the lack of technology to complete the human genome project. As history has demonstrated, the human genome project was completely carried out by Sanger's dideoxy sequencing and electrophoretic separation (granted that it is now in capillaries instead of slab gels). We should be equally concerned about the lack of understanding of the origin, history, and proper use of genetic markers (especially SNPs). In summary, pooled genotyping would be a useful method for rapid first scanning of the genome followed by confirmation with individual genotyping and possibly haplotyping.

Another critical component of linkage or association studies is the availability of a sufficient number of patient DNAs with well-defined phenotypes. Collection of a large number of samples in many diseases for research based studies have been carried out routinely in many medical centers. A number of publications have addressed the issue of the number of patients required for various types of linkage and association studies (Cardon et al. 2000). Data collected from patients with diseases based on well-defined criteria or data based on biochemical tests are highly suited for genetic studies. However, well-defined and complete phenotypic information can be difficult to collect in certain diseases or pathological conditions. For example, there might be 12 symptoms related to hypersensitivity to certain types of antibiotics. A physician can make the diagnosis with only 3–4 symptoms even though different combinations of symptoms might appear. In this case, it is important to collect the presence or absence of all 12 symptoms after the diagnosis has been made. This will provide the maximum information content and allow the stratification of the patient population. The possibility of carrying out whole genome scans with 100,000 SNPs has important implications for patient informed consent. The human subjects in the studies have to fully understand that a whole genome scan is a

high resolution molecular fingerprinting technique that can be used to uniquely identify an individual. As more disease susceptibility genes and disease phenotypes are correlated with SNPs or SNP profiles in the future, it is possible to retrospectively determine one's susceptibilities to genetic diseases.

Although we have far exceeded our initial goal of isolating 300,000 SNPs, the practical issues in the application of SNPs in medicine have actually been expanded by orders of magnitude. Not only do we have many more SNPs in the genome than we expected. but the amount of work that is required to understand these polymorphisms and to correlate with functions is staggering. We have to be realistic about the timing of the impact of genetic information on medical care, and manage our expectations accordingly. Although a genome scan is now possible, high genotyping costs will prohibit routine full-scale utilization of the currently available SNPs for at least a few years. The major obstacle is an information gap, not a technology gap. Our challenges are to design and develop whole genome SNP marker sets that are informative and cost-effective and to develop a method of rapid collection of phenotypically well characterized case/control populations. Future availability of these reagents will allow major changes in the speed of discovery of the underlying mechanisms of common complex diseases and sufficiently improve patient care.

## REFERENCES


Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S.J., McCurdy, S., Foy, M., Ewan, M., et al. 2000. *Nat. Biotech.* **18:** 630–634.

Cardon, L.R., Idury, R.M., Harris, T.J.R., Witte, J.S., and Elston, R.C. 2000. *Pharmacogenetics* **10:** 503–510.

Collins, A.L., Lonjou, C., and Morton, N.E. 1999. *Proc. Nat. Acad. Sci.* **96:** 15173–15177.

Fallin, D., Cohen, A., Essioux, L., Chumakov, I., Blumenfeld, M., Cohen, D., and Schork, N. J. 2001. *Genome Res.* **11:** 143–151.

Goddard, K.A., Hopkins, P.J., Hall, J.M., and Witte, J.S. 2000. *Amer. J. Hum. Genet.* **66:** 216–234.

Hoh, J., Wille, A., Zee, R., Cheng, S., Reynolds, R., Lindpaintner, K., and Ott, J. 2000. *Annals Hum. Genet.* **64:** 413–417.

Huttley, G.A., Smith, M.W., Carrington, M., and O'Brien, S.J. 1999. *Genetics* **152:** 1711–1722.

Jorde, L.B. 2000. *Genome Res.* **10:** 1435–1444.

Kruglyak, L. 1997. *Nat. Genet.* **17:** 21–24.

———.1999. *Nat. Genet.* **22:** 139–144.

Kruglyak, L. and Nickerson, D.A. 2001. *Nat. Genet.* **27:** 234–236.

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

Lander, E.S. 1996. *Science* **274:** 536–539.

Lander, E.S. and Schork, N.J. 1994. *Science* **265:** 2037–2048.

Martin, E.R., Lai, E.H., Gilbert, J.R., Rogala, A.R., Afshari, A.J., Riley, J., Finch, K.L., Stevens, J.F., Livak, K.J., Slotterbeck, B.D., et al. 2000. *Amer. J. Hum. Genet.* **67:** 383–394.

Miller, R. D., Taillon-Miller, P., and Kwok, P.Y. 2001. *Genomics* **71:** 78–88.

Risch, N. and Merikangas, K. 1996. *Science* **273:** 1516–1517.

Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Mullikin, J.C., Mortimore, B.J., Willey, D.L., Hunt, S.E., Cole, C.G., et al. 2001. *Nature* **409:** 928–933.

Schafer, A.J. and Hawkins, J.R. 1998. *Nature. Biotech.* **16:** 33–39.

Service, S.K., Lang, D.W.T., Freimer, N.B. and Sandkuijl, L.A. 1999. *Amer. J. Hum. Genet.* **64:** 1728–1738.

Shaw, S.H., Carrasquillo, M.M., Kashuk, C., Puffenberger, E.G., and Chakravarti, A. 1998. *PCR Meth. Applic.* **8:** 111–123.

Valdes, A.M. and Thomson, G. 1997. *Amer. J. Hum. Genet.* **60:** 703–716.

Zollner, S. and von Haeseler, A. 2000. *Amer. J. Hum. Genet.* **66:** 615–628.

# Application of SNP Technologies in Medicine: Lessons Learned and Future Challenges

Eric Lai

*Genome Res.* 2001 11: 927-929
Access the most recent version at doi:10.1101/gr.192301