SNP genotyping in drug development

SNP-based candidate gene analyses are likely to have a significant impact on predicting drug efficacy and will, undoubtedly, impact on the pharmaceutical industry's development strategies as a whole over the next few years.

The human single nucleotide polymorphism (SNP) map represents the third generation of human genome maps used for genetic linkage and association studies over the last 20 years. The first practical genetic maps were built from restriction fragment length polymorphisms (RFLPs) in the early 1980s. RFLPs result from nucleotide changes within restriction enzyme sites, and were detected by electrophoretic analysis of genomic DNA digested with specific restriction enzymes. The second generation genetic map was built with microsatellite markers in the early 1990s. Microsatellites are small stretches of dinucleotide or tetranucleotide repeat sequences that differ in size by 2-4 nucleotides and can be detected by electrophoretic analysis of genomic DNA amplified by PCR. Standardised sets of microsatellite markers were developed to provide complete genome scans for linkage analyses of both simple and complex human diseases. These microsatellite maps were widely used because the markers are highly informative, and a set of approximately 400 markers could be used to scan the genome using standardised, semi-automated methods. The recent completion of the human genome sequence has enabled the development of the third generation human genetic map using thousands of bi-allelic SNPs that are amenable to highly automated analyses using a diverse range of detection technologies[1]. Currently, public databases contain information on >2 million SNPs aligned to specific locations in the human DNA sequence.

These high-density SNP maps have enabled recent discoveries about the structure and organisation of the human genome and have revealed surprising patterns of genetic recombination. Comparisons between two different human genomes show that there is, on average, one SNP every 800-1,000 nucleotides[2]. Consequently, any two genomes will contain approximately three million differences. Comparison of SNPs along contiguous 'blocks' of DNA has shown that there are long stretches (30-100kb) of DNA with little recombination, separated by short intervals with much higher levels of recombination[3]. Each of these contiguous blocks of DNA show relatively little variability between the highly recombinant intervals. Typically a block containing six SNPs may have only 4-6 combinations, or haplotypes, of the 64 possible combinations. The observed combinations represent the ancestral haplotypes, many of which are shared among different populations and ethnic groups.

A current proposal exists to develop a comprehensive haplotype map of the human genome that can be used for genome wide association studies to identify genes underlying common complex human diseases. The haplotype map will cost <$80 million and will probably require a public-private consortium effort similar to the original

By Dr Nicholas C. Dracopoli and Dr Kim E. Zerba

SNP consortium involving the Wellcome Trust and 12 major pharmaceutical companies. While holding great promise, there is a great controversy as to whether haplotype-based association studies can overcome the problems observed with previous strategies, and whether there is sufficient signal to noise to detect associations at individual genes with common complex diseases that have not responded to previous approaches.

Linkage and association studies

Linkage and association analyses represent the two general approaches used to identify genes involved with human disease. Linkage refers to the physical and genetic relationships between sites on chromosomes. Linkage analysis assumes a near one-to-one correspondence between genotypes for single gene and the disease. For the simple, Mendelian, genetic diseases that segregate in families, co-inheritance patterns of a genetic marker and disease can provide evidence that a gene that is directly involved with the disease is somewhere in the vicinity of the marker on a particular chromosome.

Standardised screening panels of highly polymorphic microsatellite markers, uniformly spread across the genome, have been used to locate hundreds of Mendelian disease genes. Microsatellites have a large number of alleles (variants) which makes them highly informative in linkage studies for identifying alleles uniquely co-segregating with disease in families. While the costs of genotyping for such panels of markers have decreased dramatically, these panels are not amenable to full automation because they require electrophoretic separation to distinguish 2-4bp size differences, and complex data processing to prepare data for genetic analyses. In contrast to microsatellite genotyping, SNP genotyping does not require electrophoresis and can be automated for very high-throughput platforms. A single SNP is not as informative as a microsatellite marker because each SNP has only two forms. Despite having fewer variants, groups of SNPs in close proximity, can be genotyped to provide the same amount of information as that from a single microsatellite.

The promise of high-throughput automated genome screens using SNPs for linkage analysis has motivated efforts to construct a genetic map with SNPs and then to develop standardised third generation SNP screening panels at higher density and resolution than existing microsatellite maps. The first SNP linkage map[4] was presented at the 2002 Cold Spring Harbor Laboratory meeting earlier this year. Strategies using uniformly spaced or clusters of SNPs are now being considered to develop cost-effective standardised SNP screening panels[5]. Despite these advancements much additional effort is required to develop these panels and thus, depending on progress, microsatellites will still be more useful for linkage analysis for the immediate future.

Simple Mendelian monogenetic diseases represent only a small fraction (<2%) of the non-infectious human disease burden. Most of this burden (>98%) is due to non-Mendelian, multifactorial diseases like heart disease, cancer, hypertension, non-insulin dependent diabetes and asthma. Unfortunately, family studies have proven almost useless for the discovery of complex disease genes. For monogenetic diseases, the impacts of genetic variations on the disease are expected to be large enough to be detectable in relatively small samples from family studies. In contrast, the impacts of variations in individual genes on complex, multifactorial diseases are expected to be much smaller and too difficult to detect in small family samples. Moreover, family studies are simply not applicable to standard drug intervention trials. Linkage analysis will therefore not be useful for identifying and predicting subjects at risk for adverse reactions to drug exposure or predicting variation in efficacy of drug response.

Most studies seeking to identify genes involved in complex diseases and drug response are focused on population-based study designs and association analysis. Population-based clinical trials represent an enormous resource and potential for success using genetic association studies. Association analysis attempts infer relationships between genetic variation and phenotypic variation. Phenotypic variation is represented by interindividual variation in disease and drug response endpoints or intermediate physiological and biochemistry biomarkers of disease and drug response.

The high-density SNP map offers tremendous opportunities for candidate gene and eventually genome-wide SNP association studies but many difficult challenges remain. Most of the 2.4 million known SNPs need assay development and determination of allele frequencies among populations. The known SNPs represent less than a quarter of the expected number of human SNPs. Many SNPs will be restricted to particular populations and many will vary in allele frequency among populations or ethnic groups[6].

Variation in allele frequency presents a key challenge for clinical trial-based genetic association studies. Multiple ethnic groups are a common feature of drug intervention clinical trials. Ethnic groups commonly differ in SNP allele frequencies,

Pharmacogenetics

References

1 Tsuchihashi, Z and Dracopoli, NC 2002. Pharmacogenetics Journal 2: 103-110.

2 Cargill, MD, Altshuler, J, Ireland, P, Sklar, K, Ardlie, N et al 1999. Nature Genet. 22: 231-238.

3 Gabriel, SB, Schaffner, SF, Nguyen, H, Moore, JM, Roy, J et al 2002. Science 296: (in press).

4 Matise et al 2002. Genome Mapping and Sequencing Meeting. Cold Spring Harbor Laboratory.

5 Goddard, KAB and Wijsman, WM 2002. Genet. Epidemiol. 22: 205-220.

6 Fullerton, SM, Clark, AG, Weiss, KM, Nickerson, DA, Taylor, SL et al 2000. Am. J. Hum. Genet. 67: 881-900.

7 Wilson, JF, Weale, ME, Smith, AC, Gratrix, F, Fletche, B et al 2001. Nature Genet. 29: 265-269.

8 Templeton, AR, Weiss, KM, Nickerson, DA, Boerwinkle, E et al 2000. Genet. 156: 1259-1275.

9 Tiret, L, Poirier, O, Nicaud, V, Barbaux, S, Herrmann, S-M et al 2002. Hum. Mol. Genet. 11: 419-429.

10 Zerba, KE, Ferrell, RE and Sing, CF 2000. Hum. Genet. 107: 465-475.

11 Emahazion, T, Feuk, L, Jobs, M, Sawyer, SL, Fredman, D et al 2001. TIG 17: 407-413.

12 Tabor, HK, Risch, NJ, and Myers, RM 2002. Nature Rev. Genet. 3: 1-7.

13 Nelson, MR, Kardia, SLR, Ferrell RE, and Sing, CF 2001. Genome Res. 11:458-470.

14 Reich, DE and Lander, ES 2001. TIG 17: 502-510.

15 Weiss, KM and Clark, AG 2002. TIG 18: 19-24.

16 Couzin, J 2002. Science 296: 1391-1392.

17 Mallal S, Nolan, D, Witt, C, Masel, G, Martin, AM et al 2002. Lancet 359: 727-732.

disease frequency or susceptibility to adverse drug reactions. This combination of differences from population stratification within a study is a well-known problem plaguing genetic association studies that can inflate the rate of false positive associations. Moreover, commonly-used labels for ethnicity may not reflect the underlying genetic heterogeneity of stratified clinical trial studies[7].

Association studies rely on the concept that most SNPs are not functionally related to phenotypic variation and often associations that are detected will be due to allele frequency correlations (linkage disequilibrium) with one or more unmeasured functional polymorphisms on the same chromosome. Studies often choose a small subset of SNPs to first consider in a candidate gene study. To exhaust the possibility that significant associations observed might be due to an unmeasured functional polymorphism on the same chromosome, follow-up studies are then considered using a more comprehensive set of SNPs in the candidate region. Unfortunately, there is usually significant variation in the linkage disequilibrium within and among gene regions making critical and difficult the choice of which subset of SNPs to initially genotype to avoid inflation of the rate of false negative associations[8,9].

The combination of low signal to noise ratio between variation at individual genes contributing to complex diseases, small sample sizes and linkage disequilibrium between adjacent SNPs makes commonly used methods of multiplicity adjustment of p-values to control for false positive associations difficult. Replication of genetic associations among studies is particularly difficult given the population heterogeneity described above and the possibility that it may be common for genetic effects to be dependent on the context of other biological or genetic factors that define subgroups of a population[10-12]. New analytical methods are needed to deal with these complex genetic relationships[13]. These methods need to be scalable so they can manage analyses of 102-103 SNPs in each of 103s of DNA samples. Analytical strategies involving simultaneous consideration such numbers of SNPs, even one at a time, are non-trivial and present significant problems for the current generation of software packages.

Strategies are now being considered to reduce the number of SNPs needed to analyse based on blocks of completely correlated SNPs occurring together as haplotypes. Despite the controversy of whether the common haplotype map is sufficiently deep or appropriate to detect associations with complex diseases[14-16] or drug response, new ana-

lytical methods will be needed for haplotype-based high-throughput analyses in the context of clinical trial designs and genetic heterogeneity at many levels.

Pharmaceutical applications

SNP genotyping has a broad range of applications in pharmaceutical development. Traditionally, the major impact of genetic studies has been in the successful discovery of monogenic disease genes such as Huntington's disease and cystic fibrosis. Many more recent efforts to identify novel genes underlying complex non-Mendelian diseases, such as non-insulin dependant diabetes (NIDDM) or asthma have been less successful because of the underlying heterogeneity of the disease and the weak impact of individual genes on the disease phenotype. The development of a comprehensive SNP and haplotype database from many human populations may significantly improve the chances of discovering the genetic basis of the common human diseases that have, up to now, resisted this approach. The combination of very high throughput and inexpensive genotyping technology along with the comprehensive haplotype map of the human genome will enable, for the first time, genome wide association studies for many complex diseases. The hope is that the highly refined maps will permit very sensitive comparisons of haplotype frequencies in case-control studies that will identify candidate genes with positive associations to various complex disease phenotypes. A key to the success of this approach will be the collection of samples from large, clinically-defined populations. Despite the optimism arising from a new understanding of the organisation of the human genome, it is still unclear if the signal from individual genes will be large enough to be reliably detected above the noise after scanning thousands of SNPs and haplotypes in large genomic association studies of complex diseases.

SNP analyses will also have a significant impact on pharmacogenetics. It will soon be feasible to carry out large-scale analyses of hundreds of candidate genes (if 100 selected genes can really be referred to as candidate genes) and eventually genome wide analyses to search for markers that are predictive of drug efficacy or of the risk of adverse events. Candidate gene studies are being routinely applied to the analysis of samples from clinical trials for both purposes. For example, a recent report has identified a single HLA haplotype with high sensitivity and specificity for the hypersensitivity reaction after treatment with abacavir (an HIV-1 nucleoside-analogue reverse-transcriptase inhibitor)[17]. In this

case HLA-analyses could be used to identify those subjects at high risk of the potentially fatal hypersensitivity reaction and provide more than three-fold reduction in the risk of the adverse event. In this example, the hypersensitivity reaction results primarily from a single genetic risk factor detected in a simple test. This may not be the case for other adverse events which may have a complex genetic basis. In these cases, the sensitivity and specificity of a test for an individual marker, or set or related markers, may not be sufficient to identify many, and certainly not all, of the at risk subjects. Even if it is possible to develop a pharmacogenetic test for a complex adverse event, then how much of the relative risk do we need to be able to identify? Is it necessary to identify all cases, or to reduce the risk to an equivalent level as that seen with other, perhaps less efficacious, therapies in the same class. Furthermore, the consequences of making a wrong call using a pharmacogenetic marker are much more serious when predicting an adverse event than when predicting drug efficacy. In this case, there is risk of serious harm to the patient who would be incorrectly

treated with a drug that induces an adverse event, as opposed to being treated with a safe therapeutic agent that is incorrectly predicted to work optimally in that particular patient. For these reasons, we believe that the impact of pharmacogenetics on drug development will be much greater in predicting drug efficacy, but that it will have an impact on both predicting efficacy and adverse events when applied to existing, marketed drugs.

Conclusions

Comprehensive SNP and haplotype maps will clearly have a significant impact on biology and medical sciences. While the short term impact of pharmacogenetics is still unclear, it is already having a significant influence on drug development strategies. The co-development of therapeutics and pharmacogenetic tests will become increasingly common. The promise of pharmacogenetics cannot, however, be met without significant improvements in technology. In contrast to transcription profiling where whole genome analyses can be completed on only two DNA

Pharmacogenetics

chips, it is still necessary to select a limited number of candidate genes and SNPs for genetic analysis. Another 100-fold reduction in genotyping cost is still necessary before widespread genome scanning with >104 SNPs per sample becomes economically feasible. However, in the current phase, we believe that SNP-based candidate gene analyses are likely to have a significant impact on predicting drug efficacy and that these analyses will have an increasing impact on pharmaceutical companies' development strategies over the next few years.   **DDW**

Dr Nicholas Dracopoli is Vice-President of Clinical Discovery Technologies at Bristol-Myers Squibb. In this role he is responsible for several research areas including pharmacogenomics, proteomics, bioimaging, biomarker assay development and the BMS clinical laboratories. Previously he was Vice-President at Genos Biosciences, a joint venture between Sequana Therapeutics and the Memorial Sloan-Kettering Cancer Center, and Vice-President of Molecular Genetics at Sequana Therapeutics. Dr Dracopoli obtained his BSc and PhD degrees from the University of London and

completed post doctoral fellowships at the Memorial Sloan-Kettering Cancer Center and the Massachusetts Institute of Technology (MIT). Subsequently he served as an Assistant Director at the Whitehead/MIT Genome Center and as a Section Chief at the National Center for Human Genome Research at the NIH before moving to the biotechnology industry. Dr Dracopoli has authored more than 70 scientific publications and has extensive experience in the fields of genomics, molecular biology and cancer research.

Dr Kim Zerba is currently Associate Director of Statistical Genetics and Biomarkers in Biostatistics and Data Management for Clinical Discovery at Bristol-Myers Squibb. He received his PhD in Zoology, specialising in evolutionary biology, ecology and biostatistics, from Arizona State University in 1989. His postdoctoral research focused on statistics, genetics and the role of genetic variation in complex human diseases, particularly cardiovascular disease, in the Department of Human Genetics at the University of Michigan until 1993. He then continued with the University of Michigan as a research scientist conducting research on the genetics of complex diseases until 1999 when he joined the Bristol-Myers Squibb Pharmaceutical Research Institute.