Single nucleotide polymorphisms as tools in human genetics

Ian C. Gray[+], David A. Campbell and Nigel K. Spurr

Department of Biotechnology and Genetics, SmithKline Beecham Pharmaceuticals, New Frontiers Science Park, Third Avenue, Harlow, Essex CM19 5AW, UK

Received 26 June 2000; Accepted 4 July 2000

The development of detailed single nucleotide polymorphism (SNP) maps of the human genome coupled with high-throughput genotyping technologies may allow us to unravel complex genetic traits, such as multifactorial disease or drug response, over the next few years. Here we describe the current efforts to identify and characterize the large numbers of SNPs required and discuss the practicalities of association studies for the identification of genes involved in complex traits.

INTRODUCTION

In the 1980s, single nucleotide polymorphisms (SNPs) were detected using restriction enzymes to identify the presence or absence of cutting sites and scored by observing the resulting fragment length variation (1). In the 1990s, the SNP in the guise of the restriction site polymorphism was largely replaced by the simple tandem repeat (STR) as the marker of choice for linkage studies. STRs (di-, tri- or tetranucleotide repeats) show high levels of allelic variation in the number of repeat units, are widely and evenly distributed across the human genome and can be typed using PCR amplification. The combination of a highly polymorphic marker set and rapid typing technology led to the development of high-throughput semi-automated systems for STR genotyping during the 1990s (2,3).

The late 1990s saw a reversal from the use of STRs as SNPs regained favour amongst molecular geneticists. The main driving force behind the switch back to SNPs was a change in the type of genetic study that research groups are performing. STRs are ideal for the linkage studies that typified the genetics of the 1990s, where pedigree analysis is employed to identify a gene responsible for a monogenic disorder. However, there recently has been a shift away from monogenic disorders toward the analysis of complex multifactorial diseases such as osteoporosis, diabetes, cardiovascular and inflammatory diseases, psychiatric disorders and most cancers, which occur at a much higher frequency than single gene disorders and consequently are a greater social burden. There is also increasing interest in the genetics of drug response (pharmacogenetics), an understanding of which may allow the 'tailoring' of therapies on an individual basis. Pharmacogenetics is analogous to complex genetic disease in terms of the questions posed.

The broadly familial nature of complex diseases clearly indicates a significant genetic component. However, in contrast to monogenic conditions, this genetic element is comprised of multiple gene variants each contributing a small effect. This genetic complexity may also be compounded by heterogeneity, where different combinations of gene variants give rise to a similar phenotype. The extent of this problem is likely to be so great that the frequency of any polymorphism contributing to a disease phenotype will be only slightly elevated in a disease group compared with unaffected controls. Unfortunately, linkage analysis has limited power to detect such small effects (4), and attempts to identify genes involved in complex disease using linkage-based approaches have generally proved disappointing. Association studies with a large sample size, where cases of disease are compared with matched controls from the same population, are likely to give a greater chance of detecting small effects (see ref. 5 for an excellent review of the merits and pitfalls of the various methods of genetic dissection of complex traits).

Until now, population-based case–control studies have been limited, attempting to associate one or a few 'candidate genes' with disease. This restricted approach has been due largely to the lack of appropriate genetic markers and the inadequacy of the available genotyping tools for the high-throughput approaches required for large-scale genome-wide experiments. Ironically, the STR markers that have been so successful in the study of monogenic disease will probably have limited value for population-based studies due to the high levels of polymorphism which make them so useful for linkage studies. This high level of variation reflects high mutation rates (6) which are likely to confound population-based approaches. Furthermore, due to the large number of markers required, STR loci may be too sparse for association-based approaches.

In contrast, SNPs are highly abundant and are perceived as being more stable than STRs due to lower mutation rates. SNPs also have a further advantage: STR loci are 'surrogate' markers in the sense that polymorphism in the STR is used to locate an adjacent functional variant that contributes to the disease state; variation at the STR itself rarely contributes to the phenotype. Like STRs, SNPs can be used as surrogate markers, but many SNPs also have functional consequences if they occur in the coding or regulatory regions of a gene.

[+]To whom correspondence should be addressed. Tel: +44 1279 627 225; Fax: +44 1279 627 500; Email: ian_gray-1@sbphrd.com

Therefore, by using SNP markers, it is often possible to test for association between a phenotype and a functional variant directly. For these reasons, SNPs have been chosen as the basis for the high-density genetic marker maps required for the next phase of human genetics: the unravelling of complex genetic traits. This review will focus on: (i) SNP-based association studies; (ii) SNP identification; (iii) SNP frequency across the human genome; and (iv) SNP genotyping methods.

SNP-BASED ASSOCIATION STUDIES

SNP-based association studies can be performed in two ways: direct testing of an SNP with functional consequence for association with a disease trait, or using an SNP as a marker for linkage disequilibrium (LD). LD is generally defined as a measure of the degree of association (co-segregation) of two genetic markers and can thus be used to identify those regions of the genome associated with disease in a population; surrogate markers are used to identify genetic regions showing association with the disease, allowing the subsequent identification of the adjacent causative gene. This is analogous to the use of linkage analysis to identify disease-related genes in families. However, due to the limited number of generations in a family study and consequently a limited number of recombination events, linkage can be detected over large genetic distances in pedigrees. Approximately 300 highly informative STR markers evenly spaced across the human genome (~1 every 10 cM) typically are required to localize the gene responsible for a monogenic disorder (see, for example, ref. 7). Conversely, LD in populations extends over far shorter distances due to erosion of inter-marker association as a result of recombination over successive generations. Furthermore, due to the extensive time periods relative to the three or four generation pedigrees used in a linkage study, LD in populations can reflect not only recombination, but also new mutation events and genetic drift (see below).

The question of the number of markers required for an LD genome scan to identify genes associated with complex disease has been (and still is) hotly debated, but is likely to be orders of magnitude higher than the 300 or so employed for linkage scans. The abundance of SNPs in the human genome coupled with their presumed stability in terms of mutation rate relative to STRs renders them the marker of choice for such studies. Purely theoretical approaches using computer simulations have predicted that LD is unlikely to extend beyond 3 kb in the human population, meaning that 500 000 SNPs would be required for a genome scan (8). However, the validity of such simulations has been questioned (for a succinct summary, see ref. 9). Predictions based on empirical data drawn from the literature are less pessimistic, suggesting that 30 000 SNPs would be sufficient (10).

With a detailed knowledge of the structure of the human genome likely to result from the current sequencing efforts, it should be possible to reduce the number of markers required for an LD genome scan by concentrating initially on those areas that are rich in genes. [This approach was proposed some years ago for linkage scans (see ref. 11).] A precise knowledge of the degree and pattern of fluctuation of recombination frequency across the genome would also allow us to distribute markers in an intelligent fashion and possibly reduce further the number of markers required for a genome scan (Fig. 1). A simple comparison of the genetic and physical maps for different genomic regions reveals large variations in the ratio of physical to genetic map distance, indicating wide differences in the levels of recombination in different parts of the genome. For example, a comparison of the recently published physical positions of chromosome 22 STRs (12) with their genetic map positions (13) reveals a considerable degree of variation in recombination rate across the chromosome, suggesting the presence of a number of 'recombination hotspots' (12). Jeffreys *et al.* recently have characterized two such regions, one adjacent to (and probably responsible for) the minisatellite *D1S8* (14) and a second in intron 2 of the *TAP2* gene within the major histocompatibility complex (MHC) (15). The extent of each of these hotspots is <2 kb. Surprisingly, a detailed comparison of crossover (assessed by PCR amplification of meiotic recombination products from sperm) versus LD at the *TAP2* hotspot revealed that for SNPs of low heterozygosity (i.e. with one rare allele), there is little correlation between recombination and LD, suggesting that new mutation followed by haplotype drift, rather than recombination, is the major factor dictating LD levels for low-frequency alleles. Although yet to be repeated for other loci, if this observation proves to be generally true, it should be considered when selecting markers for LD studies.

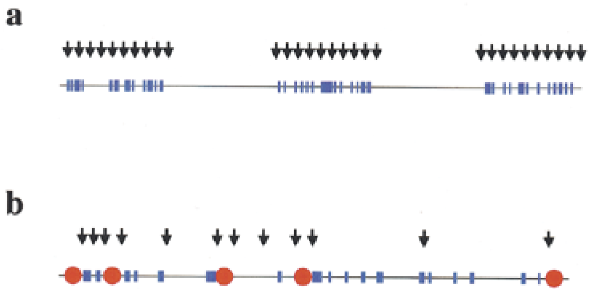

Figure 1. Intelligent genome-wide association scan. (**a**) Gene-rich regions are selected for marker saturation (genes are represented by blue vertical lines, and markers by arrows). (**b**) Within those gene-rich intervals, recombination hotspots (red cicles) define haplotypes and dictate the required marker density; regions with more hotspots require a larger number of markers.

Clearly a detailed knowledge of the location of each recombination hotspot in the genome would be of enormous help in designing marker panels for LD studies. Unfortunately, the two hotspots identified by Jeffreys *et al.* (14,15) share no sequence similarity, nor do they resemble any previously identified motifs, rendering identification of further hotspots by simple sequence searches unlikely. However, the use of families to construct haplotypes should allow hotspots to be uncovered and subsequently defined using approaches similar to those described by Jeffreys *et al.* Indeed, the *TAP2* hotspot was identified originally by linkage analysis (16), and 'recombination breakpoint' maps, highlighting potential hotspots, already exist (17). With the accelerated pace of genome characterization that is likely to follow the completion of the human genome sequence, together with a rapidly increasing SNP marker pool to draw on, it is conceivable that population-specific (but overlapping) marker sets of, say, 10 000–20 000 SNPs arranged in haplotypes will be optimized and available for LD genome scans for the identification of genes involved in complex disease within the next 2–3 years. A consequence of using so many markers simultaneously (essentially conducting many independent tests) is the requirement for large sample sizes to detect small genetic effects at a statistically significant level. Therefore, such an experiment would require the collection of large patient and control cohorts and the generation of many millions of genotypes. At present, such large-scale experiments are not economically feasible (see below).

In the meantime, are there any practical methods that can be used to identify genes with a role in complex disease? Candidate gene analysis, where a gene is selected on the basis of biological function and tested for association with the disease phenotype, currently is in vogue. However, this approach is generally unsatisfactory as we presently know little of the complete functional spectrum of all but a few gene products and less about the detailed molecular biology of most complex diseases, rendering selection of appropriate candidates something of a lottery. A more pragmatic approach is the use of linkage analysis to identify tentatively linked regions in families and then extend the analysis to the relevant population, simultaneously increasing confidence in the linkage and narrowing the critical interval harbouring the culprit gene. The most well known example of this strategy is the association of the apolipoprotein E gene with late-onset Alzheimer's disease, following equivocal linkage to the genetic interval harbouring *ApoE* on chromosome 19q13.2 (18,19). However, this approach also has limitations as the initial linkage analysis is unlikely to detect small effects, even at low confidence levels (4).

SNP IDENTIFICATION

Clearly the identification and characterization of large numbers of SNPs are necessary before we can begin to use them extensively as genetic tools. It is likely that a pool of several hundred thousand SNPs will be required as a resource for the construction of optimized marker sets for association studies. How are these to be identified? Four methods commonly are used for SNP (or mutation) detection: (i) identification of single strand conformation polymorphisms (SSCPs) (20); (ii) heteroduplex analysis (21); (iii) direct DNA sequencing; and (iv) the recently developed variant detector arrays (VDAs) (22). For SSCP detection, the DNA fragment spanning the putative SNP is PCR amplified, denatured and run on a non-denaturing polyacrylamide gel. During the gel run, the single-stranded fragments adopt secondary structure according to sequence. Fragments bearing SNPs are identified as a result of their aberrant migration pattern and confirmed by sequencing. Although a widely used and relatively simple technique, SSCP gives a variable success rate for SNP detection, typically ranging from 70 to 95% (23–25), is labour intensive and has relatively low throughput, although higher capacity methods are under development using capillary- rather than gel-based detection (26).

Heteroduplex analysis relies on the detection of a heteroduplex formed during reannealing of the denatured strands of a PCR product derived from an individual heterozygous for the SNP. The heteroduplex can be detected as a band shift on a gel, or by differential retention on a high-performance liquid chromatography (HPLC) column (27). HPLC rapidly has become a popular method for heteroduplex-based SNP detection due to simplicity, low cost and a high rate of detection [95–100% (23–25,28)]. Throughput is reasonable at ~10 min per sample using commercially available systems such as the Transgenomic Wave. At SmithKline Beecham, we routinely use HPLC for detection of SNPs in genes encoding potential drug targets.

Currently, the favoured high-throughput method for SNP detection is direct DNA sequencing. Once the sequencing reactions have been completed, a single Applied Biosystems 3700 capillary system can generate sequence from >1500 DNA fragments of 500 bp in 48 h with minimal human intervention. Dye–terminator sequencing chemistry will detect ~95% of heterozygotes; the more expensive and labour-intensive dye–primer chemistry will identify 100%. The recently formed SNP consortium (TSC), a non-profit foundation sponsored by the 10 major pharmaceutical companies and the UK's Wellcome Trust, has used dye–terminator sequencing to identify and map rapidly >100 000 SNPs (see http://snp.cshl.org/). SNPs may also be detected *in silico* at the DNA sequence level. The wealth of redundant sequence data deposited in public databases in recent years, in particular expressed sequence tag (EST) sequences, allows SNPs to be detected by comparing multiple versions of the same sequence from different sources.

VDA technology is a relatively recent addition to the high-throughput tools available for SNP detection. This technique allows the identification of SNPs by hybridization of a PCR product to oligonucleotides arrayed on a glass chip and measuring the difference in hybridization strength between matched and mismatched oligonucleotides. The VDA detection rate is comparable to that of dye–terminator sequencing and allows rapid scanning of large amounts of DNA sequence; Wang *et al.* (22) used this technique to identify ~2500 SNPs in 2 Mb of human DNA and, more recently, Halushka *et al.* (29) have used the same method to identify 874 SNPs in 75 candidate genes for hypertension.

In addition to choosing a method for SNP detection, the population in which the SNPs are to be detected must also be decided; SNP allele frequencies vary considerably across human ethnic groups and populations. The SNP consortium has opted to use an ethnically diverse panel to maximize the chances of SNP discovery. Halushka *et al.* (29) chose to analyse African and Northern European populations due to the differing prevalence and phenotype of hypertension, the disease under study, in these two ethnic groups. Other studies describe the use of disease populations for SNP discovery, following the logic that variants contributing to the disease state should be represented at a higher frequency in a disease cohort (30–32). Each approach has its merits. However, any polymorphism is likely to make a small individual contribution to the disease phenotype and will be found at only a slightly elevated frequency in the disease cohort compared with a control group. Therefore, identification of the same polymorphism by searching in a matched non-diseased population is highly probable. Different SNP panels will be required for different studies, based on the origin of the population and the study design, but the diverse approach should lead to the generation of a large pool of SNPs from which to draw the most appropriate panel for the study under consideration.

SNP FREQUENCY ACROSS THE HUMAN GENOME

There is now a good deal of information regarding SNP frequency within specific genes and the dispersal pattern of SNPs across the human genome. Nearly 300 genes recently have undergone a detailed analysis of SNP content (29–35). Although the methods and populations used in each study are different, rendering detailed comparisons impractical, several common observations can be made; all are predictable but none the less provide valuable confirmation of many of our assumptions. Changes in non-coding sequence and synonymous changes in coding sequence are generally more common than non-synonymous changes, reflecting greater selective pressure reducing diversity at positions dictating amino acid identity. Transitional changes are more common than transversions, with CpG dinucleotides showing the highest mutation rate, presumably due to deamination (36). There is enormous diversity in SNP frequency between genes, reflecting different selective pressures on each gene as well as different mutation and recombination rates across the genome. The degree of linkage disequilibrium varies widely across different genes, again reflecting different recombination and mutation rates.

The identification and study of SNPs in specific genes has provided us with a wealth of data and ratified many of our beliefs regarding gene and genome dynamics. However, as such studies are biased deliberately towards coding regions, the data generated from them are unlikely to reflect the overall distribution of SNPs throughout the genome. In contrast, the SNP consortium protocol was designed to identify SNPs with no bias towards coding regions, and the 100 000 TSC SNPs mapped at the time of writing should generally reflect sequence diversity across the human chromosomes (although the data set will not be completely free of bias; for example, selection will occur against sequences that are unclonable using the TSC protocol). Figure 2 depicts TSC SNP distribution across the long arm of chromosome 22; fluctuations in SNP density are apparent, possibly reflecting variation in mutation and recombination levels and differing selection pressure. The TSC aims to expand the number of SNPs identified across the genome to 300 000 by the end of the first quarter of 2001. Data are released quarterly via both the TSC's own web page (see above) and the SNP database dbSNP, hosted by the National Center for Biological Information (NCBI; http://www.ncbi.nlm.nih.gov/SNP/index.html).

SNP GENOTYPING METHODS

Even by the most modest estimates, whole-genome approaches to LD mapping will require several million genotypes for each case–control study, creating the need for high-throughput SNP genotyping systems. The most common SNP typing chemistries currently available are hybridization, primer extension and cleavase methods, or variations on these themes. Hybridization involves annealing one strand of a PCR product spanning the SNP to oligonucleotides complementary to each of the alternative SNP states and measuring relative hybridization efficiency (37) in an analogous fashion to the VDA technology described above. Primer extension measures the ability of a DNA polymerase to extend an oligonucleotide across the polymorphic site in the presence of nucleotides that will only allow extension across one of the two variants (38). Cleavase technology measures SNP status according to the ability of a cleavase enzyme to cut a matched or mismatched three-strand hybridization structure (39).

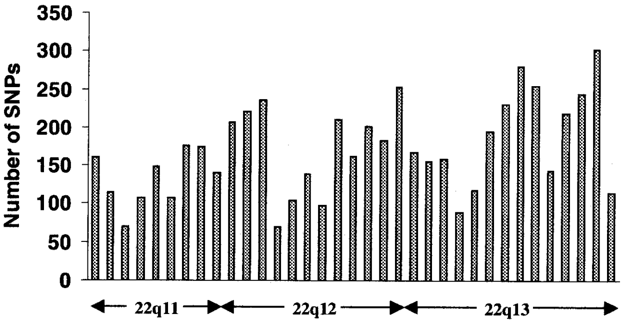

Figure 2. TSC SNP distribution along the long arm of chromosome 22 (taken from the TSC website at http://snp.cshl.org/). Each column represents a 1 Mb interval; the approximate cytogenetic position is given on the *x*-axis. Clear peaks and troughs of SNP density can be seen, possibly reflecting different rates of mutation, recombination and selection.

The chemistry of SNP genotyping is only half of the assay: each chemistry must be married to an appropriate detection system. In recent years, numerous biotechnology companies have begun to develop high-throughput systems for SNP detection. It is beyond the scope of this article to review all of the current technologies, but they include fluorescent microarray-based systems (Affymetrix), fluorescent bead-based technologies (Luminex, Illumina, Q-dot), automated enzyme-linked immunosorbent (ELISA) assays (Orchid Biocomputer), fluorescent detection of pyrophosphate release (Pyrosequencing), fluorescence resonance energy transfer (FRET)-based cleavase assays (Third Wave Technologies) and mass spectroscopy detection techniques (Rapigene, Sequenom).

Although there is room for further development for each of the chemistries and detection systems available for high-throughput SNP typing, most, if not all of them will be capable of delivering the genotyping capacity required in coming years. The overriding consideration at present is not the ability to achieve the required numbers of genotypes but the associated cost; even affluent pharmaceutical companies would soon find their pockets empty should they choose to undertake large-scale genetic studies at the current cost of ~US$0.50 per genotype! A 10-fold or greater cost reduction is required before such studies truly become feasible. One potential way of reducing the number of assays required (and hence cost), together with preserving valuable DNA resources, is to genotype equimolar pools of DNA from affected individuals and corresponding pools from controls and quantify the results (40). However, DNA pooling is not appropriate for every study (e.g. those in which haplotype construction is required).

CONCLUDING REMARKS

We are on the cusp of a new era of human genetic analysis. The human genome sequence together with high-throughput methods for genetic analysis will put us in the unique position of being able to begin to unravel complex genetic traits, a situation long envisaged but unattainable until now. Like most scientific advances, our current position is built on the work of the last 20 years; despite the hype, SNP genotyping is simply an elderly technique supercharged by high-throughput detection and assay systems. The population-based association study is not a new idea. The current excitement comes not from the development of new methods of genetic analysis, but rather from the technological advances that will allow us to use adaptations of existing methods to decipher the subtle conundrums of human genetics.

ACKNOWLEDGEMENT

We would like to thank Zhong Li for assistance with Figure 2.

REFERENCES

1. Botstein, D., White, R.L., Skolnick, M. and Davis, R.W. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.*, **32**, 314–331.
2. Livak, K.J., Marmaro, J. and Todd, J.A. (1995) Towards fully automated genome-wide polymorphism screening. *Nature Genet.*, **9**, 341–342.
3. Hall, J.M., LeDuc, C.A., Watson, A.R. and Roter, A.H. (1996) An approach to high-throughput genotyping. *Genome Res.*, **6**, 781–790.
4. Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.
5. Lander, E.S. and Schork, N.J. (1994) Genetic dissection of complex traits. *Science*, **265**, 2037–2048.
6. Chakraborty, R., Kimmel, M., Stivers, D.N., Davison, L.J. and Deka, R. (1997) Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proc. Natl Acad. Sci. USA*, **94**, 1041–1046.
7. Nelen, M.R., Padberg, G.W., Peeters, E.A., Lin, A.Y., van den Helm, B., Frants, R.R., Coulon, V., Goldstein, A.M., van Reen, M.M., Easton, D.F. *et al.* (1996) Localization of the gene for Cowden disease to chromosome 10q22–23. *Nature Genet.*, **13**, 114–116.
8. Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.*, **22**, 139–144.
9. Ott, J. (2000) Predicting the range of linkage disequilibrium. *Proc. Natl Acad. Sci. USA*, **97**, 2–3.
10. Collins, A., Lonjou, C. and Morton, N.E. (1999) Genetic epidemiology of single-nucleotide polymorphisms. *Proc. Natl Acad. Sci. USA*, **96**, 15173–15177.
11. Antonarakis, S.E. (1994) Genome linkage scanning: systematic or intelligent? *Nature Genet.*, **8**, 211–212.
12. Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J. *et al.* (1999) The DNA sequence of human chromosome 22. *Nature*, **402**, 489–495.
13. Dib, C., Faure, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Hazan, J., Seboun, E. *et al.* (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature*, **380**, 152–154.
14. Jeffreys, A.J., Murray, J. and Neumann, R. (1998) High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. *Mol. Cell*, **2**, 267–273.
15. Jeffreys, A.J., Ritchie, A. and Neumann, R. (2000) High resolution analysis of haplotype diversity and meiotic crossover in the human *TAP2* recombination hotspot. *Hum. Mol. Genet.*, **9**, 725–733.
16. Cullen, M., Noble, J., Erlich, H., Thorpe, K., Beck, S., Klitz, W., Trowsdale, J. and Carrington, M. (1997) Characterization of recombination in the HLA class II region. *Am. J. Hum. Genet.*, **60**, 397–407.
17. Cox, S.A., Attwood, J., Bryant, S.P., Bains, R., Povey, S., Rebello, M., Kapsetaki, M., Moschonas, N.K., Grzeschik, K.H., Otto, M. *et al.* (1996) European Gene Mapping Project (EUROGEM): breakpoint panels for human chromosomes based on the CEPH reference families. Centre d'Etude du Polymorphisme Humain. *Ann. Hum. Genet.*, **60**, 447–486.
18. Pericak-Vance, M.A., Bebout, J.L., Gaskell Jr, P.C., Yamaoka, L.H., Hung, W.Y., Alberts, M.J., Walker, A.P., Bartlett, R.J., Haynes, C.A., Welsh, K.A. *et al.* (1991) Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage. *Am. J. Hum. Genet.*, **48**, 1034–1050.
19. Strittmatter, W.J., Saunders, A.M., Schmechel, D., Pericak-Vance, M., Enghild, J., Salvesen, G.S. and Roses, A.D. (1993) Apolipoprotein E: high-avidity binding to β-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proc. Natl Acad. Sci. USA*, **90**, 1977–1981.
20. Orita, M., Iwahana, H., Kanazawa, H., Hayashi, K. and Sekiya, T. (1989) Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. *Proc. Natl Acad. Sci. USA*, **86**, 2766–2770.
21. Lichten, M.J. and Fox, M.S. (1983) Detection of non-homology-containing heteroduplex molecules. *Nucleic Acids Res.*, **11**, 3959–3971.
22. Wang, D.G., Fan, J.-B., Siao, C.-J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J. *et al.* (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, **280**, 1077–1082.
23. Choy, Y.S., Dabora, S.L., Hall, F., Ramesh, V., Niida, Y., Franz, D., Kasprzyk-Obara, J., Reeve, M.P. and Kwiatkowski, D.J. (1999) Superiority of denaturing high performance liquid chromatography over single-stranded conformation and conformation-sensitive gel electrophoresis for mutation detection in TSC2. *Ann. Hum. Genet.*, **63**, 383–391.
24. Dobson-Stone, C., Cox, R.D., Lonie, L., Southam, L., Fraser, M., Wise, C., Bernier, F., Hodgson, S., Porter, D.E., Simpson, A.H. and Monaco, A.P. (2000) Comparison of fluorescent single-strand conformation polymorphism analysis and denaturing high-performance liquid chromatography for detection of EXT1 and EXT2 mutations in hereditary multiple exostoses. *Eur. J. Hum. Genet.*, **8**, 24–32.
25. Gross, E., Arnold, N., Goette, J., Schwarz-Boeger, U. and Kiechle, M. (1999) A comparison of BRCA1 mutation analysis by direct sequencing, SSCP and DHPLC. *Hum. Genet.*, **105**, 72–78.
26. Wenz, H.M., Baumhueter, S., Ramachandra, S. and Worwood, M. (1999) A rapid automated SSCP multiplex capillary electrophoresis protocol that detects the two common mutations implicated in hereditary hemochromatosis (HH). *Hum. Genet.*, **104**, 29–35.
27. Underhill, P.A., Jin, L., Lin, A.A., Mehdi, S.Q., Jenkins, T., Vollrath, D., Davis, R.W., Cavalli-Sforza, L.L. and Oefner, P.J. (1997) Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res.*, **7**, 996–1005.
28. O'Donovan, M.C., Oefner, P.J., Roberts, S.C., Austin, J., Hoogendoorn, B., Guy, C., Speight, G., Upadhyaya, M., Sommer, S.S. and McGuffin, P. (1998) Blind analysis of denaturing high-performance liquid chromatography as a tool for mutation detection. *Genomics*, **52**, 44–49.
29. Halushka, M.K., Fan, J.B., Bentley, K., Hsie, L., Shen, N., Weder, A., Cooper, R., Lipshutz, R. and Chakravarti, A. (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet.*, **22**, 239–247.
30. Cambien, F., Poirier, O., Nicaud, V., Herrmann, S.M., Mallet, C., Ricard, S., Behague, I., Hallet, V., Blanc, H., Loukaci, V. *et al.* (1999) Sequence diversity in 36 candidate genes for cardiovascular disorders. *Am. J. Hum. Genet.*, **65**, 183–191.
31. Yamada, R., Tanaka, T., Ohnishi, Y., Suematsu, K., Minami, M., Seki, T., Yukioka, M., Maeda, A., Murata, N., Saiki, O. *et al.* (2000) Identification of 142 single nucleotide polymorphisms in 41 candidate genes for rheumatoid arthritis in the Japanese population. *Hum. Genet.*, **106**, 293–297.
32. Ohnishi, Y., Tanaka, T., Yamada, R., Suematsu, K., Minami, M., Fujii, K., Hoki, N., Kodama, K., Nagata, S., Hayashi, T. *et al.* (2000) Identification of 187 single nucleotide polymorphisms (SNPs) among 41 candidate genes for ischemic heart disease in the Japanese population. *Hum. Genet.*, **106**, 288–292.
33. Nickerson, D.A., Taylor, S.L., Weiss, K.M., Clark, A.G., Hutchinson, R.G., Stengard, J., Salomaa, V., Vartiainen, E., Boerwinkle, E. and Sing, C.F. (1998) DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nature Genet.*, **19**, 233–240.
34. Rieder, M.J., Taylor, S.L., Clark, A.G. and Nickerson, D.A. (1999) Sequence variation in the human angiotensin converting enzyme. *Nature Genet.*, **22**, 59–62.
35. Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C.R., Lim, E.P., Kalyanaraman, N. *et al.* (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.*, **22**, 231–238.
36. Duncan, B.K. and Miller, J.H. (1980) Mutagenic deamination of cytosine residues in DNA. *Nature*, **287**, 560–561.
37. Hacia, J.G. (1999) Resequencing and mutational analysis using oligonucleotide microarrays. *Nature Genet.*, **21** (Suppl. 1), 42–47.

38. Syvanen, A.C., Aalto-Setala, K., Harju, L., Kontula, K. and Soderlund, H. (1990) A primer-guided nucleotide incorporation assay in the genotyping of apolipoprotein E. *Genomics*, **8**, 684–692.
39. Lyamichev, V., Mast, A.L., Hall, J.G., Prudent, J.R., Kaiser, M.W., Takova, T., Kwiatkowski, R.W., Sander, T.J., de Arruda, M., Arco, D.A. *et al*. (1999) Polymorphism identification and quantitative detection of genomic DNA by invasive cleavage of oligonucleotide probes. *Nature Biotechnol.*, **17**, 292–296.
40. Breen, G., Harold, D., Ralston, S., Shaw, D. and St Clair, D. (2000) Determining SNP allele frequencies in DNA pools. *Biotechniques*, **28**, 464–466.