Genomewide linkage searches for Mendelian disease loci can be efficiently conducted using high-density SNP genotyping arrays

Gabrielle S. Sellick, Cheryl Longman[1], John Tolmie[1], Ruth Newbury-Ecob[2], Lynn Geenhalgh[2], Simon Hughes, Margo Whiteford[1], Christine Garrett[3] and Richard S. Houlston*

Section of Cancer Genetics, Institute of Cancer Research, Sutton, SM2 5NG, UK, [1]Department of Clinical Genetics, Royal Hospital for Sick Children, Glasgow, G3 8SJ, UK, [2]Department of Clinical Genetics, The United Bristol Healthcare NHS Trust, Bristol, BS2 8EG, UK and [3]North West Thames Regional Genetics Service, Kennedy-Galton Centre, North West London Hospitals NHS Trust, Harrow, HA1 3UJ, UK

ABSTRACT

Genomewide linkage searches aimed at identifying disease susceptibility loci are generally conducted using 300–400 microsatellite markers. Genotyping bi-allelic single nucleotide polymorphisms (SNPs) provides an alternative strategy. The availability of dense SNP maps coupled with recent technological developments in highly paralleled SNP genotyping makes it practical to now consider the use of these markers for whole-genome genetic linkage analyses. Here, we report the findings from three successful genomewide linkage analyses of families segregating autosomal recessively inherited neonatal diabetes, craniosynostosis and dominantly inherited renal dysplasia using the Affymetrix 10K SNP array. A single locus was identified for each disease state, two of which are novel. The performance of the SNP array, both in terms of efficiency and precision, indicates that such platforms will become the dominant technology for performing genomewide linkage searches.

INTRODUCTION

There have been considerable technological advances in conducting genomewide linkage searches since restriction fragment length polymorphisms (RFLPs) were used as genetic markers (1). Until recently, most genomewide linkage searches have been conducted using microsatellite markers spaced at ~10 Mb intervals across the genome (2,3). While these markers are robust and highly informative, genotyping is a time-consuming process. Furthermore, the process is not readily adaptable to changes to sample numbers, and the ability to scale up for very high-throughput analyses is severely limited.

Single nucleotide polymorphisms (SNPs) are abundant polymorphic markers uniformly distributed throughout the human genome (4). The high density of SNPs in mammalian genomes has led to the widespread proposal that they be may employed for genomewide association studies (5–7). However, less attention has been paid to their potential use in genomewide linkage searches. Although the heterozygosity of SNPs is lower than microsatellites, their global genomic distribution and adaptability to very high-throughput genotyping suggests that they will be widely amendable to genomewide linkage analyses (8). The availability of robust genetic map locations coupled with recent technological and cost developments in highly paralleled SNP genotyping makes it practical to now consider the use of these markers for genomewide linkage searches.

The case for using SNPs to undertake genomewide linkage searches has been forcibly made by the findings of Matise *et al*. (8) with an analysis of 56 CEPH (Centre d'Etude du Polymorphisme Humain) pedigrees genotyped over ~3000 SNPs. The Matise *et al*. SNP linkage set provided an effective map resolution equal to or better than the currently used Marshfield microsatellite mapping set (Set 13, 2003), which provides coverage of the human genome at ~8 Mb intervals. Few studies to date have, however, utilized SNPs for whole-genome genetic linkage analysis of Mendelian traits, although several studies have suggested that a high-density SNP marker set will be superior to the existing sets of microsatellite markers for genomewide linkage searches (9–11). Lindsey *et al*. (12) recently reported a follow-up SNP-based analysis of a family with dominantly inherited hereditary spastic paraplegia previously linked to *SPG4* on chromosome 2 by microsatellite markers (13). A total of 24 family members, 11 of which were affected were genotyped for 122 chromosome 2 SNPs. Multipoint linkage analysis using the SNP genotypes confirmed linkage to *SPG4* and provided logarithm of odds ratio (LOD) scores equal to or better than that obtained when using microsatellites (13).

The introduction of DNA genotyping technologies capable of accurately scoring thousands of SNPs in parallel potentially

---

offers a method for conducting cost-effective genomewide linkage searches with greater sample flexibility over conventional microsatellite searches. Here, we report the assay robustness, utility and performance of a recently introduced commercial SNP genotyping array to localize genes for three Mendelian diseases, autosomal recessively inherited neonatal diabetes, craniosynostosis and dominantly inherited renal dysplasia. Loci for each disease, two of which are novel, were identified and confirmed through the use of conventional microsatellite markers. Our experience indicates that the use of high-throughput SNP genotyping technologies represents a highly efficient method of conducting genomewide linkage searches.

MATERIALS AND METHODS

Patients and families

Three families were sampled with informed consent obtained from all patients in accordance with local ethical guidelines and the tenets of the Helsinki declaration. Two of the families were consanguineous with Family 1 segregating recessive neonatal diabetes (Figure 1A) and Family 2 segregating recessive craniosynostosis associated with calcification of the basal ganglia (Figure 1B). Detailed clinical descriptions of both the consanguineous families have been reported previously (14–16). The third family analysed segregated autosomal dominant renal dysplasia (Family 3; Figure 1C). Clinical presentation of phenotypes varied from stillbirth associated with bilateral renal agenesis to renal dysplasia associated with adult onset renal failure in 10 affected individuals. None of the affected individuals within Family 3 had clinical evidence of overt ophthalmologic abnormalities, indicative of renal-coloboma syndrome [Online Mendelian Inheritance in Man (OMIM^TM), Johns Hopkins University, Baltimore, MD. MIM Number: #120330: 24/03/2003 (http://www.ncbi.nlm.nih.gov/omim/)].

The family segregating recessive neonatal diabetes (Family 1) had been the subject of a previous unsuccessful commercially available microsatellite-based genomewide linkage search (unpublished data).

Molecular analyses

SNP genotyping and candidate gene analysis. Genomewide linkage searches of the three families were undertaken using either a pre-release version of the GeneChip® Mapping 10k Xba Array containing 10 044 SNP markers or the final release version of the GeneChip®Mapping 10k Xba Array containing 11 555 SNP markers (Affymetrix Inc., Santa Clara, CA). Analysis of the family segregating neonatal diabetes was undertaken using the pre-release version of the array. SNP genotypes were obtained by following the Affymetrix protocol for the GeneChip® Mapping 10k Xba Array (17). Briefly, 250 ng of genomic DNA isolated from peripheral blood was digested per sample with the restriction endonuclease XbaI for 2.5 h. Digested DNA was mixed with Xba adapters and ligated using T4 DNA ligase for 2.5 h. Ligated DNA was added to four separate PCR reactions, amplified, pooled and purified to remove unincorporated ddNTPs. The purified PCR product was then fragmented

with DNase I, end-labelled with biotin and hybridized to an array for 18 h in a standard Affymetrix 640 hybridization oven. After hybridization, the arrays were washed, stained and scanned using an Affymetrix Fluidics Station F400 with images obtained using an Affymetrix GeneArray® scanner 2500. Affymetrix MicroArray Suite software was used to obtain raw microarray feature intensities (RAS scores). RAS scores were processed using the Affymetrix Genotyping Tools software package GCOS/GDAS (Affymetrix Inc.) to derive SNP genotypes, marker order and linear chromosomal location.

Conformation of linkage and generation of genotypes for regions of the genome underrepresented by SNP markers was undertaken using fluorescence-labelled microsatellite markers (Invitrogen, Carlsbad, CA) referenced to the UCSC Human genome database (http://genome.ucsc.edu, July 2003 release). Analysis was undertaken on an ABI Prism 3100 genetic analyzer with allele sizes determined using the ABI PRISM® Genotyper software package (Applied Biosystems, Foster City, CA).
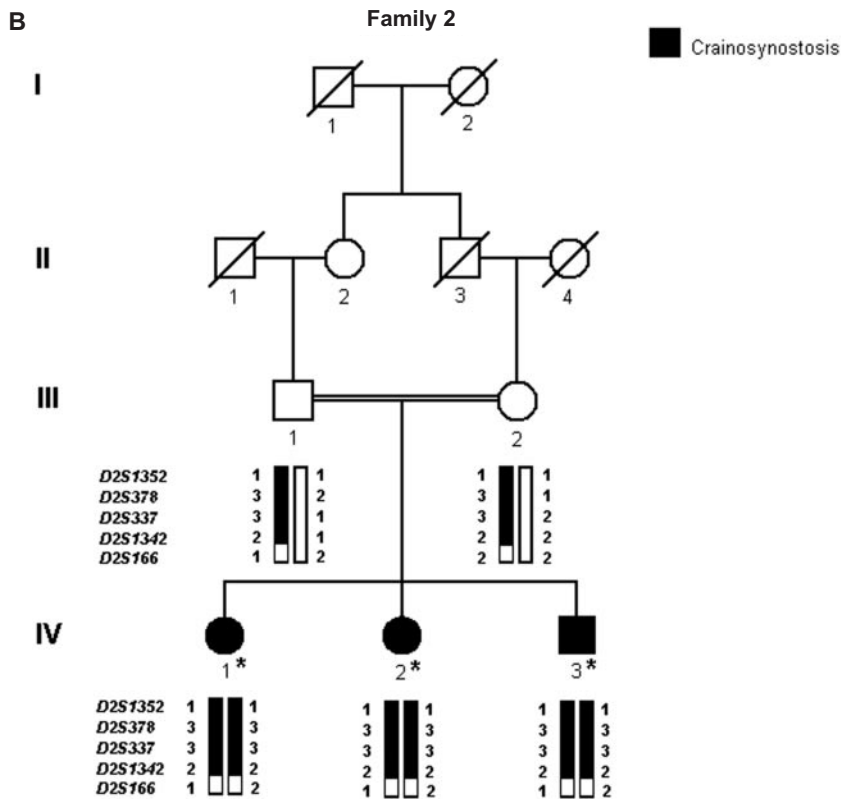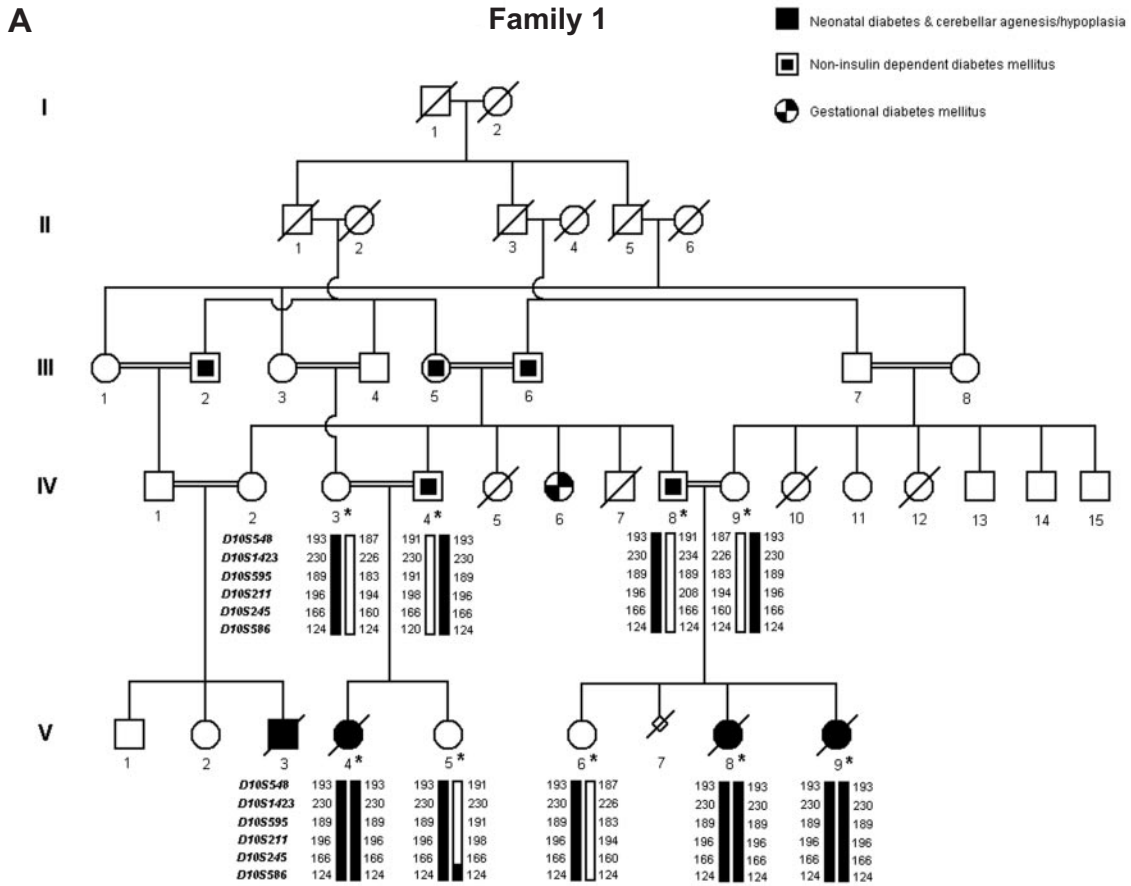
Sequencing of the candidate gene for renal dysplasia (*PAX2*), including all exons and intronic–exonic boundaries was undertaken using PCR primers designed by Primer 3 software (http://www.broad.mit.edu/cgi-bin/primer/primer3_www.cgi). Amplified PCR products were purified (Qiagen Inc., Valencia, CA) and bi-directionally sequenced using BigDye Terminator chemistry implemented on an ABI Prism 3100 sequencer. Sequences were aligned and compared with consensus sequences obtained from the human genome databases (http://genome.ucsc.edu; http://www.ncbi.nlm.nih.gov) using the software package Sequencher (Version 3.1.1; Gene Codes Corp., Ann Arbor, MI). All PCR primers and conditions are available upon request. Control DNA used in the *PAX2* sequencing was isolated from 120 individuals of Northern European decent with no medical history of chronic illness.

Orthologs of human PAX2 were identified for mouse, zebrafish, Japanese madaka and sea lamprey using BLASTP and were compared with GenBank (NCBI; http://www.ncbi.nlm.nih.gov). Protein sequences were aligned using ClustalW (EMBL; http://ebi.ac.uk/clustalw/).

Bioinformatics

A search for non-Mendelian inconsistencies was initially undertaken using the statistical package STATA version 7.0 (Stata Corp., College Station, TX). In addition, MERLIN and PedCheck were used to search for unusual patterns of Mendelian inheritance consistent with potential genotyping errors (18,19).

The search for regions of haplotype sharing was initially undertaken by using STATA. Regions identified as being consistent with linkage were further analysed by multipoint linkage analyses using the Homoz or Genehunter programs (20,21). Allele frequencies of SNP markers for the analysis of the renal dysplasia family were based on Caucasian data generated by the typing of unrelated individuals by Affymetrix (NetAffx^TM; http://www.affymetrix.com/analysis/index.affx). As population frequency data for markers was not available for the specific ethnic groups of the two consanguineous pedigrees, LOD scores were evaluated over a range of marker allele frequencies.
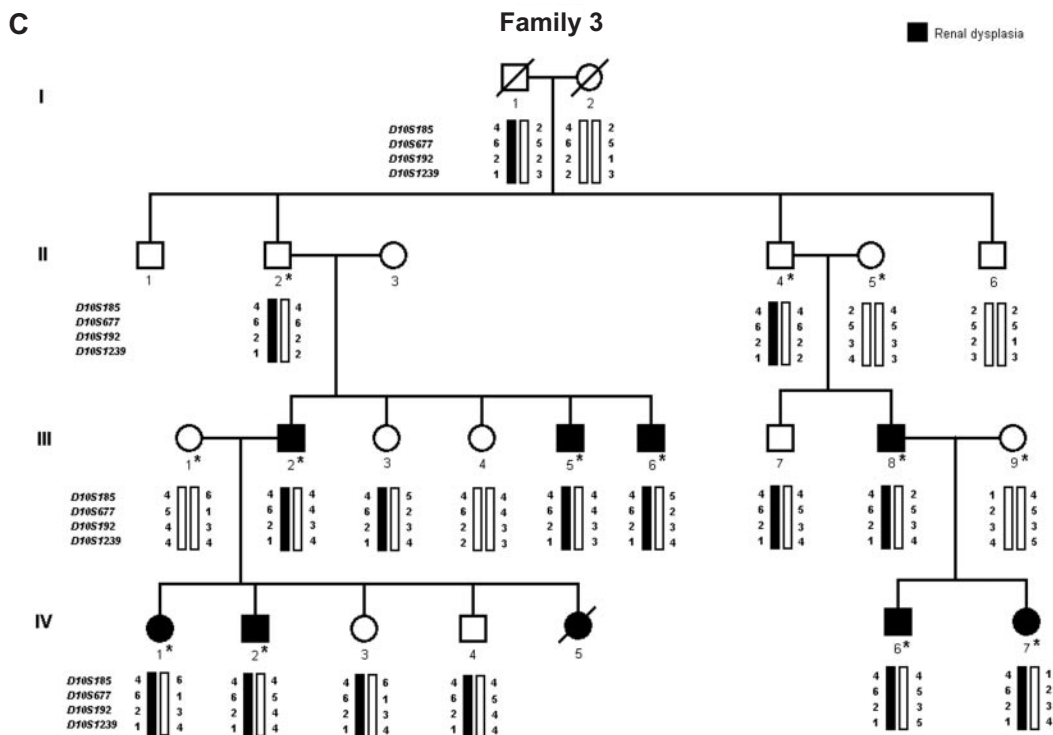
A   **Family 1**

■ Neonatal diabetes & cerebellar agenesis/hypoplasia

▣ Non-insulin dependent diabetes mellitus

◕ Gestational diabetes mellitus

B   **Family 2**

■ Crainosynostosis

Figure 1. Pedigrees of the three families showing microsatellite marker genotypes and the segregating disease-causing haplotype. Blackened symbols indicate affected individuals. Individuals with asterisk were subject to whole-genome SNP genotyping. (**A**) Neonatal diabetes pedigree (Family 1). Symbols with internal black squares indicate individuals affected with non-insulin dependent diabetes mellitus. Stripped symbols denote individuals affected with gestational diabetes. Black rectangles denote the shared disease haplotype on chromosome 10p13-p12.1. The marker order is telomere, *D10S548*, *D10S586*, centromere. (**B**) Craniosynostosis pedigree (Family 2). Black rectangles denote the shared disease haplotype on chromosome 2p16.3-p14. The marker order is telomere, *D2S1352*, *D2S166*, centromere. (**C**) Renal dysplasia pedigree (Family 3). Black rectangles denote the shared disease haplotype on chromosome 10q23.31-q25.1. The marker order is telomere, *D10S185*, *D10S1239*, telomere.

Neonatal diabetes and craniosynostosis were modelled as fully penetrant recessive diseases with the population frequencies of the disease alleles set to 0.001. Renal dysplasia was modelled as an autosomal dominant trait with the population frequency of the disease allele set to 0.001. As the penetrance of the disease allele was not known, an affected-only analysis was undertaken. The map order and distances between SNP and microsatellite markers were based on the UCSC Human Genome browser (http://genome.ucsc.edu/, July 2003 release).

RESULTS

Genotyping

Available individuals from each of the three families were subject to whole-genome SNP linkage analysis and subsequent conformational microsatellite genotyping (Figure 1). The average SNP genotype call rate obtained from the 25 samples analysed was 93.6% (range 84.1–96.2%), providing ∼10 800 genotypes per individual from the 10K XbaI final release array. The observed concordance rate between the SNP genotypes obtained from the repeated sample (Family 1, V:5) was 99.7%. None of the eleven males typed was assigned a heterozygous state for any of the SNPs mapping to the X chromosome. Overall 0.07% of the markers displayed obvious non-Mendelian transmission and a further 0.18% of markers exhibited presumptive transmission errors. This implies that

the overall genotyping error rate associated with the 10K SNP genotyping is of the order of ∼0.25%.

Linkage analysis and loci identification

Autozygosity mapping is a highly efficient method for the discovery of autosomal recessive disease loci. The methodology seeks homozygous regions in consanguineous families. The greater the number of affected individuals who have a shared homozygous region defined by the largest number of informative markers mapping to the region, the more likely the region is to harbour the disease-causing mutation within the family. We applied this strategy to identify disease loci for recessive neonatal diabetes and recessive craniosynostosis.

For the family segregating recessive neonatal diabetes (Family 1), two regions of homozygosity >6 Mb were identified. Haplotype analysis excluded other possible homozygosity-by-descent regions. One region of potential linkage was defined by six polymorphic SNPs (7q32.3-7q33) and the other by 26 polymorphic SNPs (10p13-p12.1). Both regions of potential linkage were further evaluated for homozygosity with microsatellite markers. Only the 8.5 Mb interval on 10p13-p12.1 defined by SNPs rs1398431 and rs2149636 and containing the microsatellites *D10S548*, *D10S1423*, *D10S595*, *D10S211*, *D10S245* and *D10S586* was compatible with linkage (Figure 1A). None of the unaffected family members were homozygous for the disease haplotype. Multi-point LOD scores were conservatively generated assuming a

first-cousin inbreeding coefficient. Under this inheritance model, the multipoint LOD score based solely on SNP genotypes was 3.2 assuming equal allele frequencies. The SNP-based multipoint LOD score was stable over a range of disease and marker allele frequencies. The multipoint LOD score generated from the six microsatellites was only marginally higher. Combining the SNP and microsatellite marker data allowed positional refinement of the linked region to a 7.7 Mb interval between rs1398431 and *D10S596* with a resultant LOD score of 3.3.

For the family segregating craniosynostosis (Family 2), involvement of any of the known disease genes *FGFR1, FGFR2, FGFR3, MSX2, TWIST or TCOF1* (22–24) had previously been excluded by mutational analysis. Haplotypes generated from SNP data within our study confirmed these earlier results. Whole-genome linkage analysis identified three regions of homozygosity, each >10 Mb. The three regions of potential linkage were defined by 76 SNPs (1q25.3-1q32.1), 73 SNPs (2p14-p16.3) and 20 SNPs (19q13.31-q13.42), respectively. Each of the three regions was further evaluated with microsatellite markers. Only the 16.2 Mb interval of shared homozygosity on chromosome 2p14-16.3 defined by SNPs rs1073981 to rs719293 and containing the microsatellites *D2S1352*, *D2S378*, *D2S337* and *D2S1342* was compatible with linkage (Figure 1B). The multipoint LOD score at 2p16.3-p14 generated from the SNP data was 2.4 assuming equal allele frequencies. Although the LOD score does not

attain the classical threshold of significance, it corresponds to a *P*-value of 0.0004. The SNP-based multipoint LOD score was stable over a range of disease and marker allele frequencies. The multipoint LOD score based on the five microsatellites was identical.

For the analysis of the family segregating dominant renal dysplasia (Family 3), identification of the disease locus was based on demonstrating co-segregation of a marker haplotype with affection. Two regions of potential linkage were identified (nine polymorphic SNPs mapping to 9q33.2-q34.13 and 55 polymorphic SNPs mapping to 10q23.31-q25.1). Only the 16.1 Mb haplotype on chromosome 10q23.31-q25.1 was supported by microsatellite data (Figure 1C) and was defined by the SNPs rs2077946 and rs1361356. The multipoint LOD score generated independently from both the SNP and microsatellite data was 2.4. Of the RefSeq genes mapping to the region of linkage, *PAX2* represented an obvious candidate for further analysis. A novel G→A single base pair substitution at nucleotide position 874 within exon 3 was identified in all affected members and obligate gene carriers within the family. The G874A mutation resulted in an alanine to threonine substitution at position 111 of the expressed protein (Figure 2). The sequence change was not identified in 240 alleles from healthy control individuals. The A111T amino acid change resides within the Paired DNA-binding domain of the expressed PAX2 protein and is evolutionarily highly conserved between human, mouse, sea lamprey and zebrafish
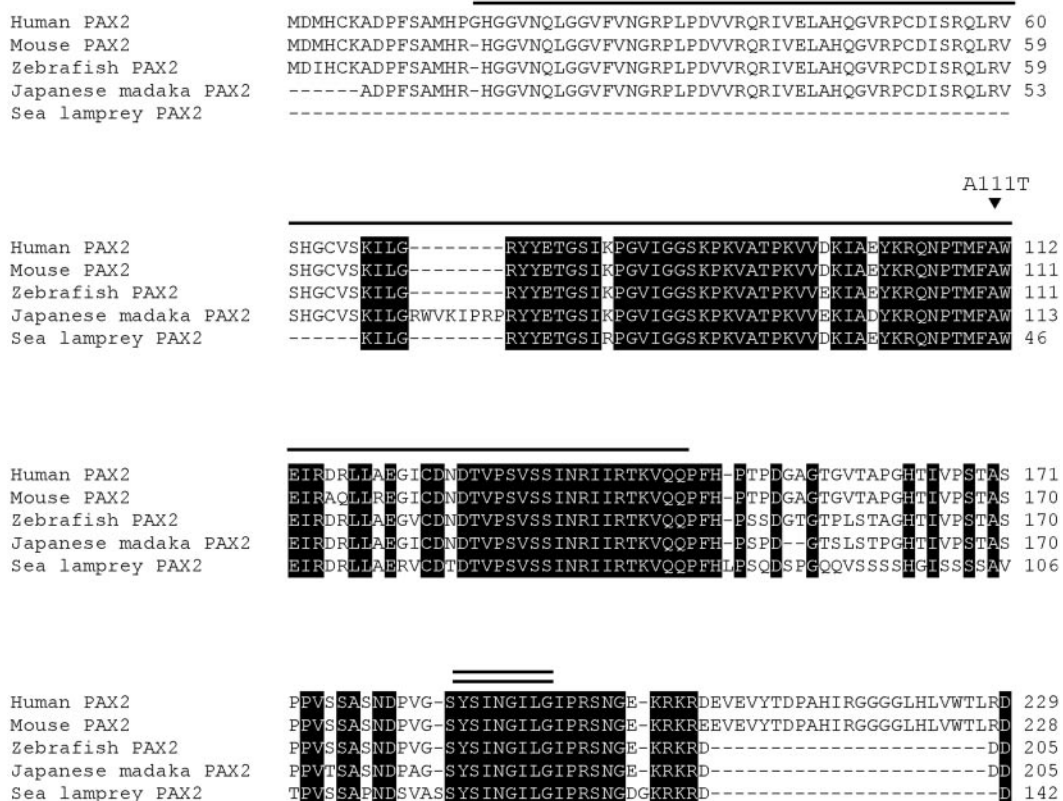
```
                                                                            A111T
                                                                              ▼
Human PAX2         MDMHCKADPFSAMHPGHGGVNQLGGVFVNGRPLPDVVRQRIVELAHQGVRPCDISRQLRV 60
Mouse PAX2         MDMHCKADPFSAMHR-HGGVNQLGGVFVNGRPLPDVVRQRIVELAHQGVRPCDISRQLRV 59
Zebrafish PAX2     MDIHCKADPFSAMHR-HGGVNQLGGVFVNGRPLPDVVRQRIVELAHQGVRPCDISRQLRV 59
Japanese madaka PAX2 ------ADPFSAMHR-HGGVNQLGGVFVNGRPLPDVVRQRIVELAHQGVRPCDISRQLRV 53
Sea lamprey PAX2   -----------------------------------------------------------

Human PAX2         SHGCVSKILG--------RYYETGSIKPGVIGGSKPKVATPKVVDKIAEYKRQNPTMFAW 112
Mouse PAX2         SHGCVSKILG--------RYYETGSIKPGVIGGSKPKVATPKVVDKIAEYKRQNPTMFAW 111
Zebrafish PAX2     SHGCVSKILG--------RYYETGSIKPGVIGGSKPKVATPKVVEKIAEYKRQNPTMFAW 111
Japanese madaka PAX2 SHGCVSKILGRWVKIPRPRYYETGSIKPGVIGGSKPKVATPKVVEKIADYKRQNPTMFAW 113
Sea lamprey PAX2   ------KILG--------RYYETGSIRPGVIGGSKPKVATPKVVDKIAEYKRQNPTMFAW 46

Human PAX2         EIRDRLLAEGICDNDTVPSVSSINRIIRTKVQQPFHPTPDGAGTGVTAPGHTIVPSTAS 171
Mouse PAX2         EIRAQLLREGICDNDTVPSVSSINRIIRTKVQQPFHPTPDGAGTGVTAPGHTIVPSTAS 170
Zebrafish PAX2     EIRDRLLAEGVCDNDTVPSVSSINRIIRTKVQQPFHPSSDGTGTPLSTAGHTIVPSTAS 170
Japanese madaka PAX2 EIRDRLLAEGICDNDTVPSVSSINRIIRTKVQQPFHPSPD--GTSLSTPGHTIVPSTAS 170
Sea lamprey PAX2   EIRDRLLAERVCDTDTVPSVSSINRIIRTKVQQPFHLPSQDSPGQQVSSSSHGISSSSAV 106

Human PAX2         PPVSSASNDPVG-SYSINGILGIPRSNGE-KRKRDEVEVYTDPAHIRGGGGLHLVWTLRD 229
Mouse PAX2         PPVSSASNDPVG-SYSINGILGIPRSNGE-KRKREEVEVYTDPAHIRGGGGLHLVWTLRD 228
Zebrafish PAX2     PPVSSASNDPVG-SYSINGILGIPRSNGE-KRKRD----------------------DD 205
Japanese madaka PAX2 PPVTSASNDPAG-SYSINGILGIPRSNGE-KRKRD----------------------DD 205
Sea lamprey PAX2   TPVSSAPNDSVASSYSINGILGIPRSNGDGKRKRD-----------------------D 142
```

Figure 2. Partial multiple protein alignment of human PAX2 with orthologs of mouse, zebrafish, Japanese madaka and sea lamprey showing the disease-associated A111T mutation segregating in Family 3. The human protein sequence was used as the basis for amino acid numbering beginning from the initiating methionine. Black shading of residues shows fully conserved amino acids. Above sequences, the Paired-box functional domain is shown as a horizontal bar, while the octapeptide functional domain is indicated by a double horizontal bar. The A111T mutation found in individuals with renal dysplasia is indicated by a blacked triangle.

(Figure 2). An analysis of all *PAX2* variants discovered till date shows that 75% of all mutations occur within the Paired functional domain (The Human PAX2 Allelic Variant Database Web Site, http://pax2.hgu.mrc.ac.uk). These findings are consistent with the sequence change being pathogenic and causative of disease in the family.

DISCUSSION

Considerable technical developments in array design and assay development have led to methods allowing efficient high-throughput SNP genotyping (25,26). Some of these methods have achieved a level of affordability and ease of use such that genomewide linkage scans can now be performed using dense sets of SNPs. The Affymetrix 10K SNP platform is one of the several recently introduced platforms for high-throughput SNP genotyping, others include those developed by Applied Biosystems (http://www.appliedbiosystems.-com/), Motorola Life Sciences (http://www.motorola.com/lifesciences/) and Illumina (http://www.illumina.com/). These platforms allow a whole-genome scan for a disease locus to be completed with greater efficiency than most laboratories can currently achieve using conventional marker sets. This improvement in genotyping efficiency is especially relevant in the analysis of complex traits, which require large numbers of families to be continually collected and processed.

Studies utilizing SNPs as markers for genetic linkage searches have until now been predominately based on the evaluation of specific genomic regions or a comparison of the informativeness of SNP markers compared with microsatellite markers (5,8,17). Here, we have evaluated the use of a high-density SNP genotyping platform to identify loci for three Mendelian diseases without any a priori knowledge of the genomic location of the disease predisposing mutation. Moreover, it is noteworthy that in one of the families (Family 1) a microsatellite-based genomewide linkage search had previously failed to identify a disease locus.

Novel loci were identified for the two recessive diseases studied. First, a locus for neonatal diabetes was identified on chromosome 10p13-p12.1 in Family 1. Forty known genes (based on Swiss-Prot, TrEMBL, mRNA and RefSeq) map to the 7.7 Mb region of linkage (UCSC Genome browser, http://genome.ucsc.edu/, July 2003 release). Of these, 17 are predicted or hypothetical genes with little or no associated information regarding their biological function. At least three genes, phophatidylinositol 4-phosphate 5-kinase, type II, alpha (*PIP5K2A*), pancreatic transcription factor 1 alpha (*PTF1A*) and calcium channel, voltage-dependent, beta 2 subunit (*CACNB2*), represent plausible candidates on the basis of either pancreatic and cerebellar expression in mammals and implied biological function of the expressed protein. Second, a locus for recessive craniosynostosis associated with calcification of the basal ganglia was mapped to chromosome 2p16.3-p14. The phenotype is distinctive from other documented forms of craniosynostosis that have been reported previously (16). Potential candidates for the mutated protein causing disease in Family 2 include targets for the transcription factor TWIST or molecules whose biological pathway counteracts the normal fibroblast growth factor receptor signalling pathway (22). Seventy-nine transcripts map to the 16.2 Mb region of linkage. Excluding the 44 predicted or hypothetical

genes mapping to the region, there are no obvious candidates at present. The family segregating autosomal dominant renal dysplasia was linked to chromosome 10q23.31-q25.1 and a novel mutation in *PAX2* demonstrated to be causative of disease. Although there is considerable variability in the severity of the disease phenotype within Family 3, previously identified missense mutations within *PAX2* have been shown to be causative of equivalently sized phenotypic differences segregating within single families (OMIM^TM, MIM Number #120330: 24/03/2003 (http://www.ncbi.nlm.nih.gov/omim/); The Human PAX2 Allelic Variant Database Web Site, http://pax2.hgu.mrc.ac.uk).

The relationship between the information provided by SNP and microsatellite markers is not straightforward, however, it has been estimated that 1.7–2.5 SNP markers provide the same information as that of one microsatellite marker (9,11). The information content (IC) averaged across the entire 10K SNP marker set is greater than that of the ABI medium density (*MD10*) screening set of microsatellite markers (Applied Biosystems) and has been previously shown to contribute significantly to differences between microsatellite and SNP-based genome scans (26). Since expected LOD scores correlate with IC, the 10K SNP array provides at least equal power to detect linkage compared with a search based upon a 5 Mb microsatellite screen.

While the median inter-marker distance between SNPs in the Affymetrix 10K array is only 104 kb with a mean genetic map distance of 0.31 cm (17), markers are not uniformly spaced through the genome and a number of regions are significantly underrepresented (chromosomes 16p, 17q, 19p and 22q) (Figure 3). This is especially problematic if only a single marker maps between two distant regions of higher IC but is itself non-informative within the individuals being genotyped. In each of the three genomewide linkage searches we conducted using the 10K SNP array, it was necessary to 'infill' such regions with additional microsatellite markers mapping to the corresponding genomic regions of the UCSC genetic map. As predicted, many of the areas of the genome that are underrepresented by SNPs are in telomeric and centromeric chromosomal regions (Figure 3). The underrepresentation of informative SNP markers in these repetitive-rich regions of the genome closely mirrors a similar scarcity of markers in conventional commercial microsatellite screening sets, such as the medium density ABI (*MD10*) set of markers (Figure 3), which is a direct consequence of a global paucity of mapped polymorphisms within such regions. As SNP arrays offer the potential to conduct single-experiment genomewide linkage searches, it is highly desirable that marker coverage is improved specifically in the regions of low informativity, thus obviating the need for subsequent infilling with extra DNA markers. In addition to ensuring that coverage of the genome is more uniform, there is the issue of whether having a dense map of uniform SNPs compared to a less dense map of SNP clusters will enhance the overall assay performance for detecting true genetic linkage. This will of course depend on the type of linkage analysis performed and the size of the pedigrees analysed. Clusters of SNPs provide the maximal information for linkage mapping if they exhibit minimal linkage disequilibrium. Selecting a group of clustered SNPs therefore provides the likelihood for detecting the greatest number of haplotypes within a sample set (9).
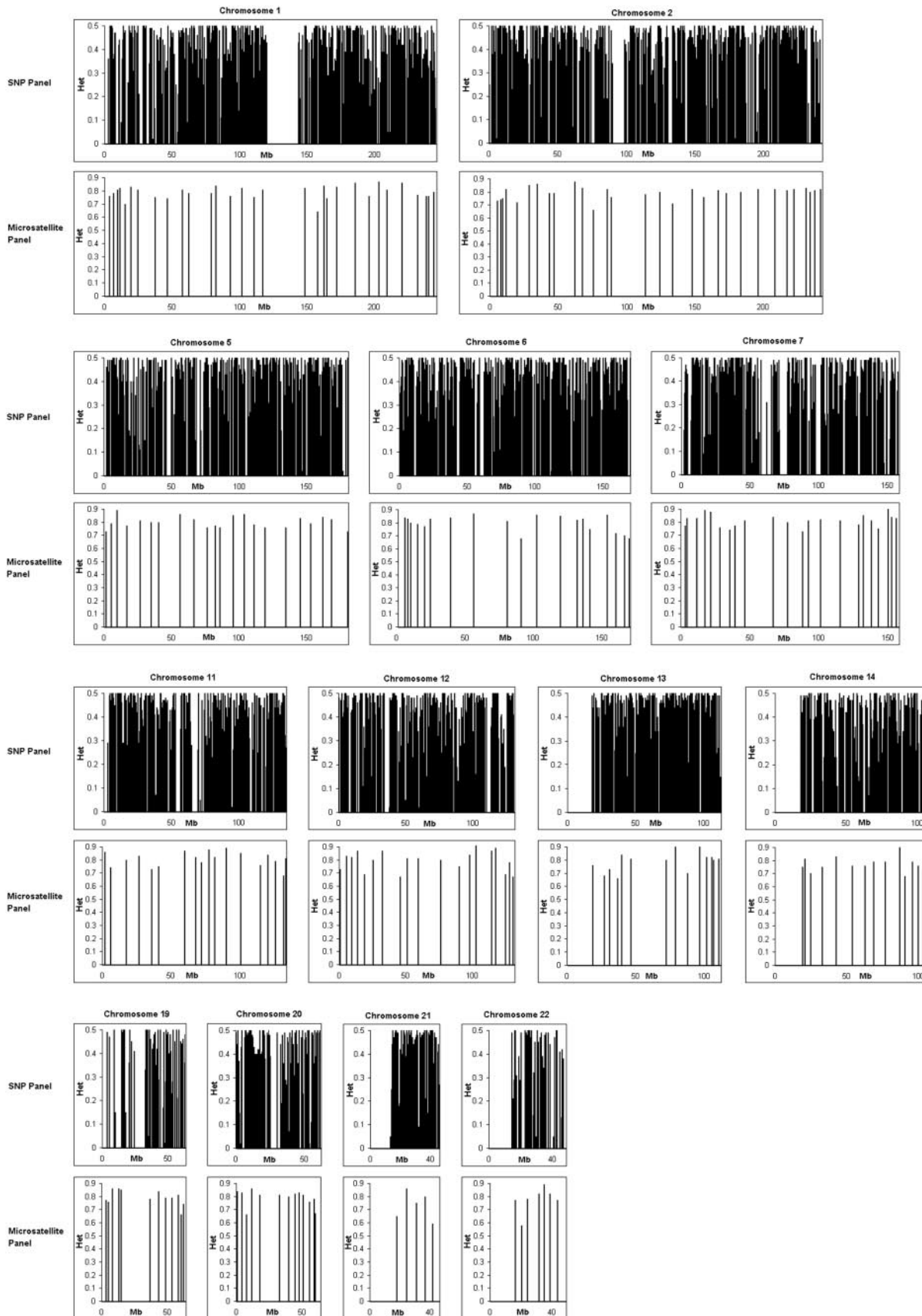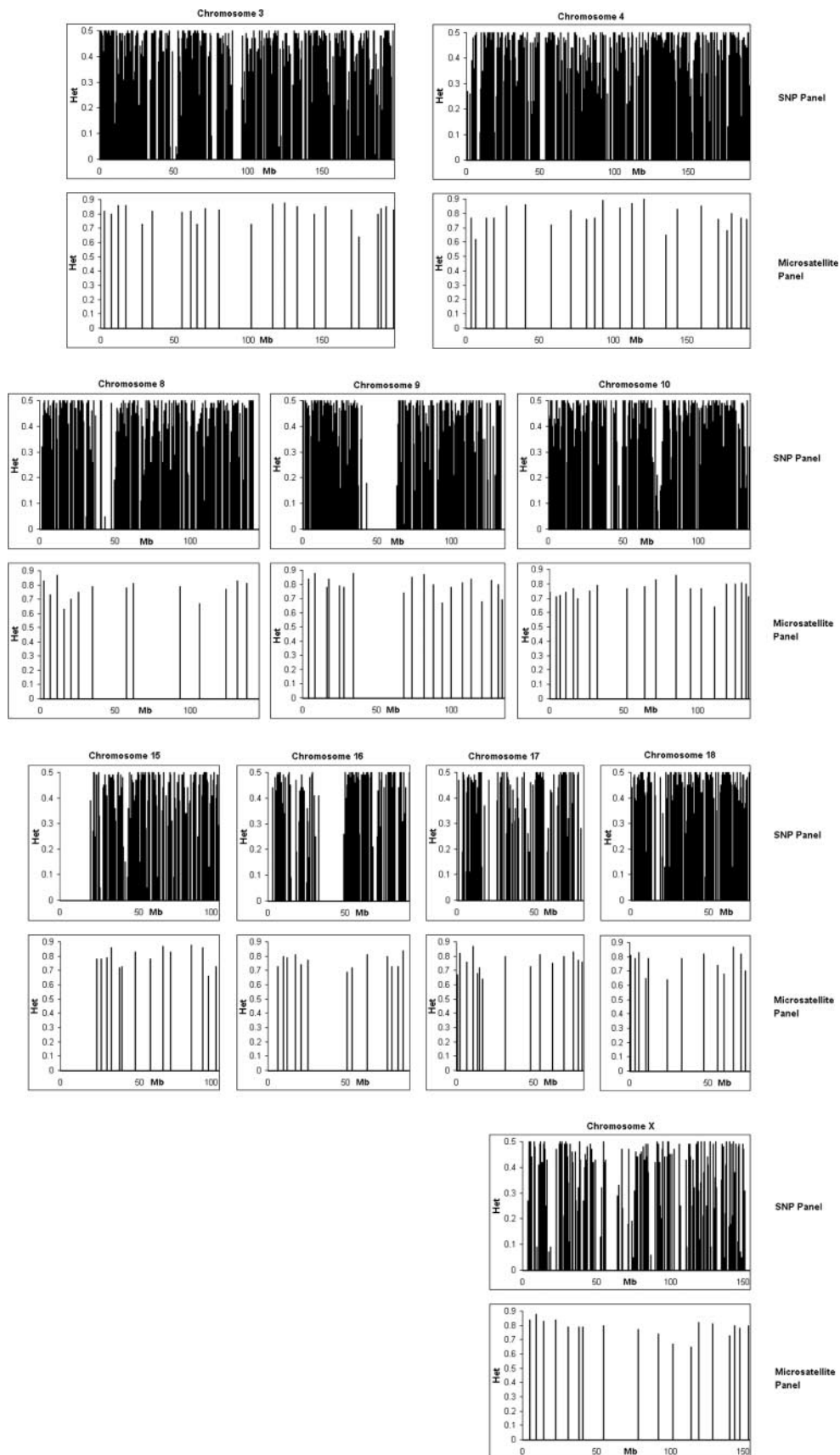
Figure 3. Comparison of the chromosome location and heterozygosity of SNPs within the Affymetrix 10K XbaI SNP array and microsatellites within the ABI medium density (*MD10*) marker set. Marker locations were taken from the UCSC database (http://genome.ucsc.edu, July 2003 release).

The observable genotyping error rate in our genomewide linkage searches using the SNP array was vastly better than that achievable in microsatellite searches. However, it is more difficult to detect bi-allelic SNP-based genotyping errors by checking for Mendelian inconsistencies as compared with multi-allelic microsatellite markers, especially if only single generation, nuclear families are interrogated.

The volume of data generated from SNP-based genomewide scans is vast compared with conventional microsatellite-based searches. Robust software for archiving, manipulating and integrating marker data is essential if searches based on SNPs are to become commonplace. The recent upgrade of GDAS^{TM} 3.0 software (Affymetrix Inc.) and the latest release of the pedigree database program ProgenyLab 6.0 (Progeny Software, South Bend, IN) permit automated export of linkage format files with the option of selecting a set number of markers per file for the entire genome. ProgenyLab also allows integration of SNP and microsatellite marker data into a single overlapping set and automated removal of non-Mendelian errors before linkage file export. These programs combined with the latest release of GENEHUNTER v2.1_r5 (http://www.fhcrc.org/labs/kruglyak/Downloads/) permit processing of more than 300 SNPs in a single run. Such improvements in software go some way to address the inherent problems when using SNPs, however, further advances are required to produce seamless transition of SNP platforms to linkage statistics.

The performance of using a dense SNP marker set for conducting genomewide linkage screens is apparent from our analysis. In addition to the improved genetic resolution offered by these platforms over conventional microsatellite markers, they are inherently convenient as very small amounts of DNA are required ($\sim$250 ng) and additional samples can be readily typed at any given stage of a project. It is therefore clear on the basis of data presented herein that the platform, with certain caveats, provides a suitable alternative to using microsatellite markers for genomewide linkage searches. Our experience supports the view that genotyping platforms based on high-density SNP markers will shortly become the dominant technology for high-throughput linkage analyses.

ACKNOWLEDGEMENTS

REFERENCES

1. Botstein,D., White,R.L., Skolnick,M.H. 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. Weissenbach,J., Gyapay.G., Dib,C., Vignal,A., Morissette,J., Millasseau,P., Vaysseix,G. and Lathrop,M. (1992) A second-generation linkage map of the human genome. *Nature*, **359**, 794–801.
3. Sheffield,V.C., Weber,J.L., Buetow,K.H., Murray,J.C., Even,D.A., Wiles,K., Gastier,J.M., Pulido,J.C., Yandava,C. and Sunden,S.L. (1995) A collection of tri- and tetranucleotide repeat markers used to generate high quality, high resolution human genomewide linkage maps. *Hum. Mol. Genet.*, **4**, 1837–1844.
4. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
5. Risch,N. and Merikangas,K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.
6. Collins,F.S., Guyer,M.S. and Charkravarti,A. (1997) Variations on a theme: cataloging human DNA sequence variation. *Science*, **278**, 1580–1581.
7. 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.
8. Matise,T.C., Sachidanandam,R, Clark,A.G., Kruglyak,L., Wijsman,E., Kakol,J., Buyske,S., Chui,B., Cohen,P., de Toma,C. *et al.* (2003) A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkage map and screening set. *Am. J. Hum. Genet.*, **73**, 271–284.
9. Kruglyak,L. (1997) The use of a genetic map of biallelic markers in linkage studies. *Nature Genet.*, **17**, 21–24.
10. Wilson,A.F. and Sorant,A.J. (2000) Equivalence of single- and multilocus markers: power to detect linkage with composite markers derived from biallelic loci. *Am. J. Hum. Genet.*, **66**, 1610–1615.
11. Goddard,K.A. and Wijsman,E.M. (2002) Characteristics of genetic markers and maps for cost-effective genome screens using diallelic markers. *Genet. Epidemiol.*, **22**, 205–220.
12. Lindsey,J.C., Lusher,M.E., McDermott,C.J., White,K.D., Reid,E., Rubinsztein,D.C., Bashir,R., Hazan,J., Shaw,P.J. and Bushby,K.M. (2000) Mutation analysis of the spastin gene (SPG4) in patients with hereditary spastic paraparesis. *J. Med. Genet.*, **37**, 759–765.
13. Reid,E., Grayson,C., Rogers,M.T. and Rubinsztein,D.C. (1999) Locus-phenotype correlations in autosomal dominant pure hereditary spastic paraplegia: a clinical and molecular genetic study of 28 United Kingdom families. *Brain*, **122**, 1741–1755.
14. Hoveyda,N., Shield,J.P., Garrett,C., Chong,W.K., Beardsall,K., Bentsi-Enchill,E., Mallya,H. and Thompson,M.H. (1999) Neonatal diabetes mellitus and cerebellar hypoplasia/agenesis: report of a new recessive syndrome. *J. Med. Genet.*, **36**, 700–704.
15. Sellick,G.S., Garrett,C. and Houlston,R.S. (2003) A novel gene for neonatal diabetes maps to chromosome 10p12.1-p13. *Diabetes*, **52**, 2636–2638.
16. Longman,C., Whiteford,M., Koppel,D., Donaldson,M., Paterson,W. and Tolmie,J. (2003) Craniosynostosis associated with intracranial calcification: a novel recessive syndrome. *Clin. Dysmorphol.*, **12**, 215–220.
17. Matsuzaki,H., Loi,H., Dong,S., Tsai,Y.Y., Fang,J., Law,J., Di,X., Liu,W.M., Yang,G., Liu,G. *et al.* (2004) Parallel genotyping of over 10 000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Res.*, **14**, 414–425.
18. Abecasis,G.R., Cherny,S.S., Cookson,W.O. and Cardon,L.R. (2002) Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genet.*, **30**, 97–101.
19. O'Connell,J.R. and Weeks,D.E. (1998) PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am. J. Hum. Genet.*, **63**, 259–266.
20. Kruglyak,L., Daly,M.J. and Lander,E.S. (1995) Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. *Am. J. Hum. Genet.*, **56**, 519–27.
21. Kruglyak,L., Daly,M.J., Reeve-Daly,M.P. and Lander,E.S. (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.*, **58**, 1347–1363.
22. Wilkie,A.O. (1997) Craniosynostosis: genes and mechanisms. *Hum. Mol. Genet.*, **6**, 1647–1656.
23. Winter,R.M. and Baraister,M. (2003) *London Dysmorphology Database v10. Busney, UK: London Medical Databases.* Oxford University Press, Oxford, UK.
24. Horiuchi,K., Ariga,T., Fujioka,H., Kawashima,K., Yamamoto,Y., Igawa,H., Sakiyama,Y. and Sugihara,T. (2004) Treacher Collins syndrome with craniosynostosis, choanal atresia, and esophageal

regurgitation caused by a novel nonsense mutation in TCOF1. *Am. J. Med. Genet.*, **128A**, 173–175.

25. Tsuchihashi,Z. and Dracopoli,N.C. (2002) Progress in high throughput SNP genotyping methods. *Pharmacogenomics J.*, **2**, 103–110.

26. John,S., Shephard,N., Liu,G., Zeggini,E., Cao,M., Chen,W., Vasavda,N., Mills,T., Barton,A., Hinks,A. *et al.* (2004) Whole-genome scan, in a complex disease, using 11 245 single-nucleotide polymorphisms: comparison with microsatellites. *Am. J. Hum. Genet.*, **75**, 54–64.