ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

GENOMICS

Genomics 81 (2003) 510–518

www.elsevier.com/locate/ygeno

# High-resolution SNP scan of chromosome 6p21 in pooled samples from patients with complex diseases

Nicole Herbon,[a,1] Monika Werner,[a,1] Christine Braig,[a] Henning Gohlke,[a] Gaby Dütsch,[a] Thomas Illig,[a] Janine Altmüller,[a] Jochen Hampe,[b] Annette Lantermann,[b] Stefan Schreiber,[b] Ezio Bonifacio,[c] Annette Ziegler,[c] Sibylle Schwab,[d] Dieter Wildenauer,[d] Dirk van den Boom,[e] Andreas Braun,[e] Michael Knapp,[f] Peter Reitmeir,[g] and Matthias Wjst[a,h,*]

[a] *Institut für Epidemiologie, GSF Forschungszentrum für Umwelt und Gesundheit, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany*

[b] *Klinik für Allgemeine Innere Medizin der Christian Albrechts Universität, FG Mukosaimmunologie, Schittenhelmstr. 12, 24105 Kiel, Germany*

[c] *Institut für Diabetesforschung, Städtisches Krankenhaus Schwabing, Kölner Platz 1, 80804 München, Germany*

[d] *Klinik und Poliklinik für Psychiatrie und Psychotherapie, Universitätsklinikum Bonn, Sigmund-Freud-Straße 25, 53105 Bonn, Germany*

[e] *Sequenom Inc., 3595 John Hopkins Court, San Diego, CA 92121-1331, USA*

[f] *Institut für Medizinische Biometrie, Informatik und Epidemiologie, Universitätsklinikum Bonn, Sigmund-Freud-Straße 25, 53105 Bonn, Germany*

[g] *Institut für Gesundheitsökonomie und Management im Gesundheitswesen, GSF Forschungszentrum für Umwelt und Gesundheit, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany*

[h] *Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie (IBE), Ludwig-Maximilians-Universität München, Marchioninistr. 15, 81377 München, Germany*

Received 12 August 2002; accepted 18 November 2002

**Abstract**

We apply a high-throughput protocol of chip-based mass spectrometry (matrix-assisted laser desorption/ionization time-of-flight; MALDI-TOF) as a method of screening for differences in single-nucleotide polymorphism (SNP) allele frequencies. Using pooled DNA from individuals with asthma, Crohn's disease (CD), schizophrenia, type 1 diabetes (T1D), and controls, we selected 534 SNPs from an initial set of 1435 SNPs spanning a 25-Mb region on chromosome 6p21. The standard deviations of measurements of time of flight at different dots, from different PCRs, and from different pools indicate reliable results on each analysis step. In 90% of the disease-control comparisons we found allelic differences of <10%. Of the T1D samples, which served as a positive control, 10 SNPs with significant differences were observed after taking into account multiple testing. Of these 10 SNPs, 5 are located between *DQB1* and *DRB1*, confirming the known association with the DR3 and DR4 haplotypes whereas two additional SNPs also reproduced known associations of T1D with *DOB* and *LTA*. In the CD pool also, two earlier described associations were found with SNPs close to *DRB1* and *MICA*. Additional associations were found in the schizophrenia and asthma pools. They should be confirmed in individual samples or can be used to develop further quality criteria for accepting true differences between pools. The determination of SNP allele frequencies in pooled DNA appears to be of value in assigning further genotyping priorities also in large linkage regions.


*Keywords:* SNP; Genotyping; DNA pooling; 6p21; HLA; Asthma; Crohn's disease; Schizophrenia; Diabetes

## Introduction

Considerable effort has been directed toward detecting genetic loci contributing to complex human diseases [1]. Thus far, >100 genome scans have been completed [2], most of them using microsatellite markers to define linkage in kindreds with a particular disease. Although there are inconsistencies between studies, in most studies several consensus areas could be identified. Most of these consensus regions contain biologically interesting candidate genes but usually span a chromosomal region of 10–30 Mb, making it difficult to find the disease-associated variant.

Comparing individual genotypes of 1000 patients with

* Corresponding author. Fax: +49-89-3187-3533.
*E-mail address:* m@wjst.de (M. Wjst).
[1] These authors contributed equally to this work.

0888-7543/03/$ – see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S0888-7543(02)00035-6

1000 controls of 10,000 single-nucleotide polymorphisms (SNPs) in these regions remains a formidable task. This could be achieved by improved brute-force genotyping. The second option is to test only representative SNPs that are tagging linkage disequilibrium blocks (LDBs). Third, allele frequencies could be compared in pooled samples from cases and controls. Although tagging of LDBs has been successfully applied recently [3], complete LDB size information is only known for chromosome 22 [4]. Even worse, LDB size could be heterogeneous between different populations and between control and disease groups.

In seeking a rapid alternative approach, we have investigated the pooling of DNA samples with a fast and precise readout of genotypes [5]. DNA pooling has been successfully used in the identification of autosomal recessive genes in isolated populations, because affected individuals show an absence of heterozygosity at loci close to the disease-associated gene [6].

As the target we selected the human chromosome 6p21 region around the HLA complex. It was the first multimegabase region of the human genome to be completely sequenced [7] and is one of the most gene-rich regions, with >200 genes spread throughout the 4 Mb of the major histocompatibility complex (MHC) region. Between 10% and 40% of the genes have functions tied to defense and immunity [8]. Several characteristics, however, complicate research in this area: Paralogous regions of the MHC at chromosomes 1, 9, and 19 and many pseudogenes have suggested duplications in the past [9]. The extremely high number of polymorphisms underlines the biological variability. However, the LDB in this area is up to 170 kb long [10], with at least three hotspots, at *HLA DNA*, *DQB3*, and *TAP2* [11,12].

In addition to these remarkable genomic features, many diseases have been clinically associated with the HLA complex. We have therefore set up a multidisease panel of which asthma, schizophrenia, Crohn's disease (CD), and type 1 diabetes (T1D) were initially selected. Although the principal function of class I and class II HLA molecules, that of binding antigenic peptides and presenting them for recognition to antigen-specific T-cell receptors, is well known, the contribution of additional genes is likely.

The linkage peak of these diseases at chromosome 6p21 is quite broad, giving evidence of more disease loci [13–15]. This was illustrated with hemochromatosis, in which the initial association with the HLA A3 antigen was just a marker for the *HFE* gene >4 Mb telomeric [16], or with congenital adrenal hyperplasia, which is due to defective alleles of the *CYP21A2* gene juxtaposed in the class III genes of the HLA complex [17]. Similar effects are expected for T1D, in which the strongest linkage peak with a lod score up to 60 was observed for a locus called *IDDM1* [8,18,19].

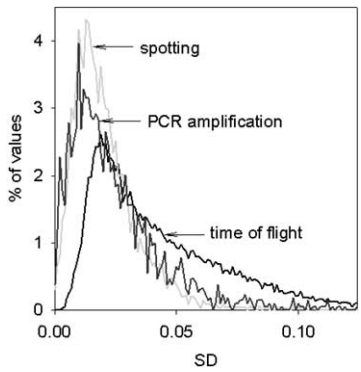


Fig. 1. Distribution of standard deviation of the mean (SD) of time-flight measurements of the same spots, measurement of different spots within the same PCR, and different PCRs of the same DNA pool.

## Results

Of the initial 1435 SNPs selected from the databases, 31 were lost because of insufficient sequence information or impossible primer design. Of those tested, >25% ($n = 383$) were nonpolymorphic in the nondiseased population. Further assays were lost because of either one of the two quality criteria of the spectra: major allele intensity < 70 ($n = 170$) or minor allele signal-to-noise-ratio < 4.0 ($n = 19$). Other assays had a peak of the biotinylated primer only ($n = 117$), a peak of the unextended primer ($n = 52$), no signal ($n = 72$), or the wrong peaks ($n = 21$). From the validated 570 SNPs, we used 534 unique SNPs in the final analysis. According to Ensembl annotation, 306 (57.3%) of these 534 SNPs are located in intergenic regions, 11 (2.1%) in the 5′-upstream and 5 (0.9%) in the 3′-downstream region of a gene, 162 (30.3%) in introns and 47 (8.8%) in exons. The noncoding region is therefore not fully represented.

All 534 SNPs were then tested in the five DNA pools. A total of 22,646 single mass spectra obtained from 6896 PCRs could be recorded. The distribution of standard deviation of the mean (SD) of time-of-flight recording (TOF) of the same spots, measurement of different spots within the same PCR, and different PCRs of the same DNA pool are displayed in Figure 1. The average SD of TOF was slightly higher (0.043) than the average deviation for single spots (0.021) or different PCRs (0.026).

Only a few control pool estimates have been replicated so far in individual samples. rs1028411 had a "C" pool frequency of 0.57. After taking into account unequal allele amplification as being observed in "CA" carriers [20], this was corrected to a value of 0.46, whereas single genotyping gave an overall frequency of 0.47. TSC0879839 had an "A" pool frequency of 0.12, corrected to 0.11 and a true frequency of 0.08. rs11244 had a "C" frequency of 0.74, corrected to 0.69 and a true value of 0.77. TSC0899479 had a "T" frequency of 0.49 corrected to 0.42 (true 0.31). TSC0504774 had an "A" frequency of 0.67 corrected to 0.54 (true 0.54).

Table 1
Allele prevalences in disease pools. Given are only differences that were either significant compared to the control pool or compared to all other pools

| | SNP | Disease | NCBI 26/ Ensembl position(bp) | Annotation by Sanger position* | Disease allele** prevalence | Control group allele prevalence | *P* value | All other combined allele prevalence | *P* value |
|---|---|---|---|---|---|---|---|---|---|
| 1 | rs1536054 | asthma | 21.535.002 | — | 0,89 | 0,59 | $7{,}57 \times 10^{-7}$ | 0,67 | $2{,}76 \times 10^{-5}$ |
| 2 | TSC0113430(rs127486) | asthma | 22.604.498 | — | 0,71 | | | 0,54 | $2{,}76 \times 10^{-4}$ |
| 3 | rs1281896 | asthma | 47.277.859 | — | 0,75 | | | 0,94 | $1{,}27 \times 10^{-10}$ |
| 4 | TSC0210562(rs2005423) | Crohn disease | 36.906.706 | *MICA* | 0,74 | 0,54 | $1{,}19 \times 10^{-4}$ | | |
| 5 | rs1729 | Crohn disease | 47.324.630 | *HLA-DRB1* | 0,70 | | | 0,90 | $3{,}02 \times 10^{-9}$ |
| 6 | rs11244 | T1D | 27.702.101 | *HLA-DOB* | 0,57 | 0,74 | $6{,}84 \times 10^{-4}$ | 0,72 | $4{,}53 \times 10^{-4}$ |
| 7 | TSC0099063(rs1016472) | T1D | 34.305.477 | *RFP* | 0,70 | 0,53 | $9{,}22 \times 10^{-4}$ | | |
| 8 | rs928815 | T1D | 36.981.644 | *NFKBIL1* | 0,74 | 0,57 | $4{,}71 \times 10^{-4}$ | 0,59 | $5{,}02 \times 10^{-4}$ |
| 9 | TSC0276264(rs968154) | T1D | 37.798.074 | *BTNL2* | 0,42 | 0,62 | $5{,}45 \times 10^{-4}$ | | |
| 10 | rs1064663(=rs707955) | T1D | 37.886189 | *HLA-DRB1* | 0,80 | | | 0,60 | $2{,}86 \times 10^{-5}$ |
| 11 | TSC0382794(rs196601) | T1D | 37.979.392 | *HLA-DQA1* | 0,95 | 0,81 | $4{,}04 \times 10^{-4}$ | | |
| 12 | TSC0116798(rs2003800) | T1D | 38.011.655 | *HLA-DQA1* | 0,34 | 0,59 | $1{,}85 \times 10^{-5}$ | 0,63 | $6{,}84 \times 10^{-9}$ |
| (3) | rs1281896 | T1D | 47.277.859 | — | 0,97 | 0,76 | $1{,}15 \times 10^{-7}$ | 0,81 | $7{,}41 \times 10^{-7}$ |
| (5) | rs1729 | T1D | 47.324.630 | *HLA-DRB1* | 0,96 | 0,76 | $1{,}99 \times 10^{-6}$ | 0,81 | $1{,}25 \times 10^{-5}$ |
| 13 | TSC0879839(no rs#) | T1D | 47.412.644 | *HLA-DQB1* | 0,55 | 0,88 | $3{,}59 \times 10^{-9}$ | 0,86 | $3{,}15 \times 10^{-10}$ |
| 14 | rs206984 | schizophrenia | 21.508.014 | *HLA-E* | 0,91 | 0,58 | $4{,}03 \times 10^{-8}$ | 0,61 | $7{,}90 \times 10^{-8}$ |
| 15 | TSC0901066(rs1884948) | schizophrenia | 27.814.176 | *ABCB3* | 0,87 | | | 0,66 | $1{,}57 \times 10^{-4}$ |
| 16 | rs1632447 | schizophrenia | 35.142.959 | *HLA-F* | 0,67 | | | 0,83 | $1{,}58 \times 10^{-4}$ |
| (10) | rs1064663 (=rs707955) | schizophrenia | 37.886189 | *HLA-DRB1* | 0,53 | | | 0,74 | $3{,}98 \times 10^{-5}$ |
| 17 | rs663310 | schizophrenia | 38.399.893 | *HLA-DOA* | 0,82 | | | 0,64 | $4{,}43 \times 10^{-4}$ |

* As Ensembl positions are not correct in the MHC region, SNPs have also been mapped to the extended MHC consensus sequence RFP_to_KNSL2.fasta (4 April 2001) obtained from http://www.sanger.ac.uk/HGP/Chr6/ and annotated with RFP_to_KNSL2_gene_list.txt (The MHC Sequencing Consortium 1999).

** Allele frequencies are given for the major allele in the control pool.

The control pool overall contributed 4041 (17.8%) measurements, asthma 4760 (21.0%), CD 4863 (21.5%), T1D 4704 (20.8%), and schizophrenia 4278 (18.9%). Mean allele frequency estimates were 0.54 (SD 0.048) in the control pool, 0.54 (0.041) in asthma, 0.54 (0.056) in CD, 0.53 (0.037) in T1D, and 0.53 (0.032) in schizophrenia, indicating similar amplification of DNA and detection efficiency in all pools.

The lower detection limit for minor alleles usually was in the 0.05-0.10 range (Fig. 2) although one actually measured with a 0.03 frequency (Table 1). Figure 3 summarizes the difference of adjusted allele frequencies in the four pools. A total of 138 (8.8% of all comparisons) had a difference >10% and; 278 (17.7%) showed >7.5% difference.

Raw allelic estimates for all pools are given in Figure 4. At least five major patterns emerge from this picture: The first is a more or less identical allelic estimate in all pools. This could be expected in samples with no population structure present and any disease variant absent in all pools. Pattern 2 is a single-point, single-pool outlier, for example at 29.7 Mb, TSC0244678 in the schizophrenia pool. In a dense SNP panel (where LDBs are covered several times), this seems to be most likely an artifact of sample selection, pooling, or spotting. In a wide distance SNP panel, however, it could point toward a single causal variant (or at least one in LD with a causal variant). Pattern 3 consists of single-point, multiple-pool differences, for example at 38.9 Mb, TSC0397534. Again, the most likely causes here are amplification or spotting artifacts. Pattern 4, multipoint, single-pool outliers (for example, 47.2–47.4 Mb, T1D), are likely to be real differences that should be followed up by individual genotyping with high priority. An alternative explanation is the unequal representation of individuals in a DNA pool in which some with an alternative haplotype dominate the overall result. Finally, pattern 5 is seen as multipoint, multiple-pool differences (for example, 37.8–37.9 Mb, HLA region). This is probably the most difficult

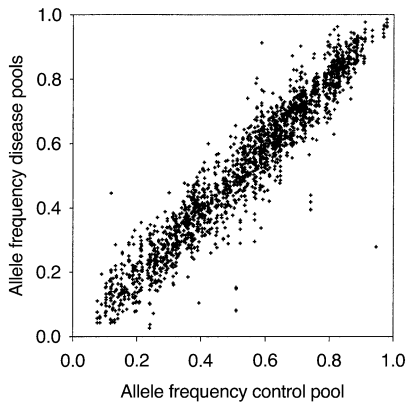


Fig. 2. Allele frequencies of SNPs. Control pool versus four disease pools (asthma, schizophrenia, type 1 diabetes, Crohn's disease).

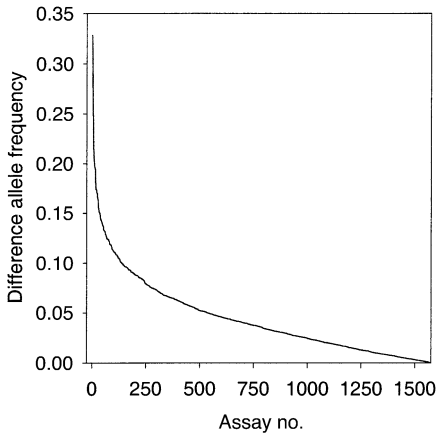


Fig. 3. Distribution of adjusted mean allele differences. Given are 1575 two-point case-control comparisons of 534 SNPs.

pattern to interpret and requires further experimental data.

## Discussion

We have shown that a pooled DNA approach permits an association screening of SNPs across a large linkage region. Although samples were drawn from different studies at different areas within Germany, allelic frequencies are quite similar across the DNA pools. This indicates that >90% of potential SNP assays can be postponed until those with more prominent differences have been tested.

The reproducibility and accuracy of the MALDI-TOF technique is vital for the detection of even small differences in allele frequencies between cases and controls. It allows the efficient and reproducible estimation of allele frequencies in DNA pools [5,20]. For studies in which differences between pools and combined single genotypes are relevant, one should correct estimations of absolute allele frequencies for unequal allele representation, which occurs mainly as a result of the decreasing sensitivity of the mass spectrometer for higher molecular masses [21]. Compared to other pooling techniques, MALDI-TOF mass spectrometry offers the advantage of low setup costs for new assays and automated sample preparation [22,23].

### *False-negative and false-positive results*

The "common disease-common variant" hypothesis proposed by Chakravarti [24] assumes relatively high frequencies of disease alleles in complex disorders. Only a few examples are known at the moment, like the *ApoE e4* allele in Alzheimer's disease, Factor V Leiden in deep venous thrombosis, *PPAR* $\gamma$ Pro12A1a in type 2 diabetes [25] or *HCR*WWCC/*PSORS1 in psoriasis [26]. A pooling approach would be suitable only for detecting abundant variants. Low-frequency alleles, however, in combination with allele variants of other genes ("multiple-pathway hit" hypothesis) would probably be missed. Exemplifying such a low-frequency allele with a strong impact on disease manifestation are the recently described variants in *CARD15,* which are associated with CD in an autosomal recessive pattern [27].

Important positive differences may be explained by population stratification. This effect is usually introduced by a founder whose alleles are found at a high frequency in the descendent population. Because in this study case and control pools have been selected from the same outbred background population, stratification did not appear to be a major problem. It is unclear, however, if there are genomic regions that are more prone to stratification effects as may be expected from this study. The *CARD15* in CD illustrates the possibility of several disease genes contributing to linkage in one chromosomal region [27]. Because the number of pooled individuals was limited, differences may also reflect some chance fluctuation.

Which of the SNPs with substantial allele frequencies are really tagging a disease-causing gene could to some extent be determined by observation of clustering at certain positions. A single pool single difference could indicate a disease variant (or chance event) on a short LDB, whereas a single pool multiple difference probably reveals a longer LDB (with fewer chance events). A definite decision can only be made on the basis of individual genotypes and the haplotype structure of the tested SNPs.

Any spatial refinement is complicated by uncertain physical placement of many SNPs. A direct record linkage of the SNP identifiers was possible only for a small number of SNPs. Even a BLAST search leaves a group of SNPs with multiple hits or moderate scores; others have changed their position or they have been removed from the databases within the course of this study. This problem is evident also in recent literature data [28], in which only 296 (78%) of 380 genetic markers of the Marshfield panel could be placed on the UCSC April 2001 assembly, and 17 of those identified on the assembly were not supported by previous genetic maps. Corroborating this tendency is another evaluation of the draft human genome sequence [29] that indicated a large discrepancy between the number of observed and expected expressed sequence tags matching the genomic sequence. Differences even increased with the next release, suggesting that known errors had not been removed.

A comparison of the original map used to set up the panel and the current map position (data not shown) revealed a considerable offset of many SNP positions between 25 Mb and 27.5 Mb. In addition, several inversions and misplacements occur between the 37-Mb and 47-Mb. This could also be found by examining the Celera 2001 genome assembly (data not shown). The correct position of the SNP on an assembly, however, is critical for such a large-scale association scan. Further efforts are needed to fill the remaining gaps, validate the current assemblies, and correct existing errors.

*Disease-specific results*

T1D showed the most significant allelic differences, with five SNPs between *DQB1* and *DRB1* (range $P = 4.4 \times 10^{-4}$ down to $P = 3.15 \times 10^{-10}$). This is in agreement with the observation that polymorphisms in exon 2 of *DQB1* and *DRB1* form the DR3 and DR4 haplotypes, which are strongly associated with T1D patients of European descent. Even minor associations can be reproduced with the pooling approach. Examples include the *HLA-DOB* gene [18], which is supported by the marker rs11244 ($P = 4.53 \times 10^{-4}$), and an association with a microsatellite in the first intron at *LTA* [18], which is supported by rs928815 ($P = 4.53 \times 10^{-4}$), close to the tumor necrosis factor (*TNF*) superfamily. In total, at least 7 of 10 associations discovered in the T1D sample pool are replications of earlier results.

The two SNPs associated with CD are at the well-known *HLA-DRB1* gene and in the MHC class I within a cluster between *MICA*, and *TNF* gene. *MICA* is a member of the natural killer cell lectinlike receptor subfamily expressed predominantly in the gastrointestinal epithelium and associated earlier with CD [30] as well as *TNF* [31].

As in the CD sample [13], the linkage peak of schizophrenia [14] was also bimodally distributed on chromosome 6p with a peak telomeric of the *MHC* and another peak inside of the *MHC* region. The known *DRB1* association is also confirmed here [31], but others (*HLA-E, HLA-F* in class I and *ABCB3* and *HLA-DOA* in class II) need further verification.

The two SNPs that show a significant difference in the asthma pool are within 27 kb of each other (Table 1). According to recent assembly data, their physical position is assumed between two predicted genes telomeric to the MHC, a region that is also supported by the maximum lod score in our previous linkage study [15]. A significant association with *DRB1* could not be found thus far [32]; however, nearly all HLA associations have been described with specific sensitization, which is seen as an invariant trait in asthma patients.

In conclusion, many known associations were confirmed and several others were newly identified. It will be necessary to confirm them in individual samples, and even if they are not verified they can be used to develop further quality criteria for accepting true differences between pools.

## Materials and methods

*DNA samples*

*Asthma*

We selected samples from a recent family study [15] using one diseased individual from each family. Entry criteria for the study consisted of at least two children with confirmed clinical asthma, whereas premature or low-birthweight children were excluded, as were those with other pulmonary disease. The number of participants is higher than that in the original publication because additional families have been examined; included are 63 members from the first and 59 from the second part of the study. The final asthma pool consisted of 122 individuals (61 male, 61 female) with an average age of 12 years. All study methods were approved by the Ethics Commission of Nordrhein-Westfalen.

*Crohn's disease*

Recently, the first susceptibility gene for CD, *CARD15*, was identified in independent populations [33–35]. For the DNA pool used in this study we selected only diseased individuals from the German study. Because the *NOD2* genotype influences linkages on other chromosomes, we have stratified the pool analyses on the presence of the three coding mutations in *NOD2* associated with disease susceptibility: SNP8, 12, and 13. The *NOD2* "positive" pool used here comprised 76 unrelated patients carrying one of the three mutations. The study was approved by the Ethics Committee of the University Hospital Kiel, and written informed consent was obtained from all patients.

*Schizophrenia*

Genomic DNA of 125 patients with schizophrenia or schizo-affective disorder (diagnosis based on DSM IIIR criteria) with a family history of psychiatric disorders were used for the pool. Family history of psychiatric disorders was defined as having at least one second-degree relative with either schizophrenia, schizo-affective disorder, bipolar disorder, or recurrent depression. Family ascertainment is described in detail [14]. Recruitment was approved by local ethical committees.

*Diabetes*

DNA samples of 384 unrelated Caucasian patients with type 1 diabetes (T1D, $n = 275$ female, 109 male) from Germany were analyzed in two DNA pools. Pools were constructed according to age of T1D onset, with the first pool consisting of patients with onset of T1D $< 15$ years of age ($n = 192$; 137 female, 55 male) and the second pool 15 years of age ($n = 192$; 138 female, 54 male). Data are given here only for the late-onset patients. Patients were recruited between 1989 and 2000 in the context of the German BABYDIAB study [36]. BABYDIAB is a prospective study, covering from birth, offspring of parents having T1D. Within this study DNA samples from all family members, including the proband with T1D, were collected. All patients had insulin-dependent and ketosis-prone diabetes. Written informed consent was obtained from all patients who participated in the study. The study was approved by

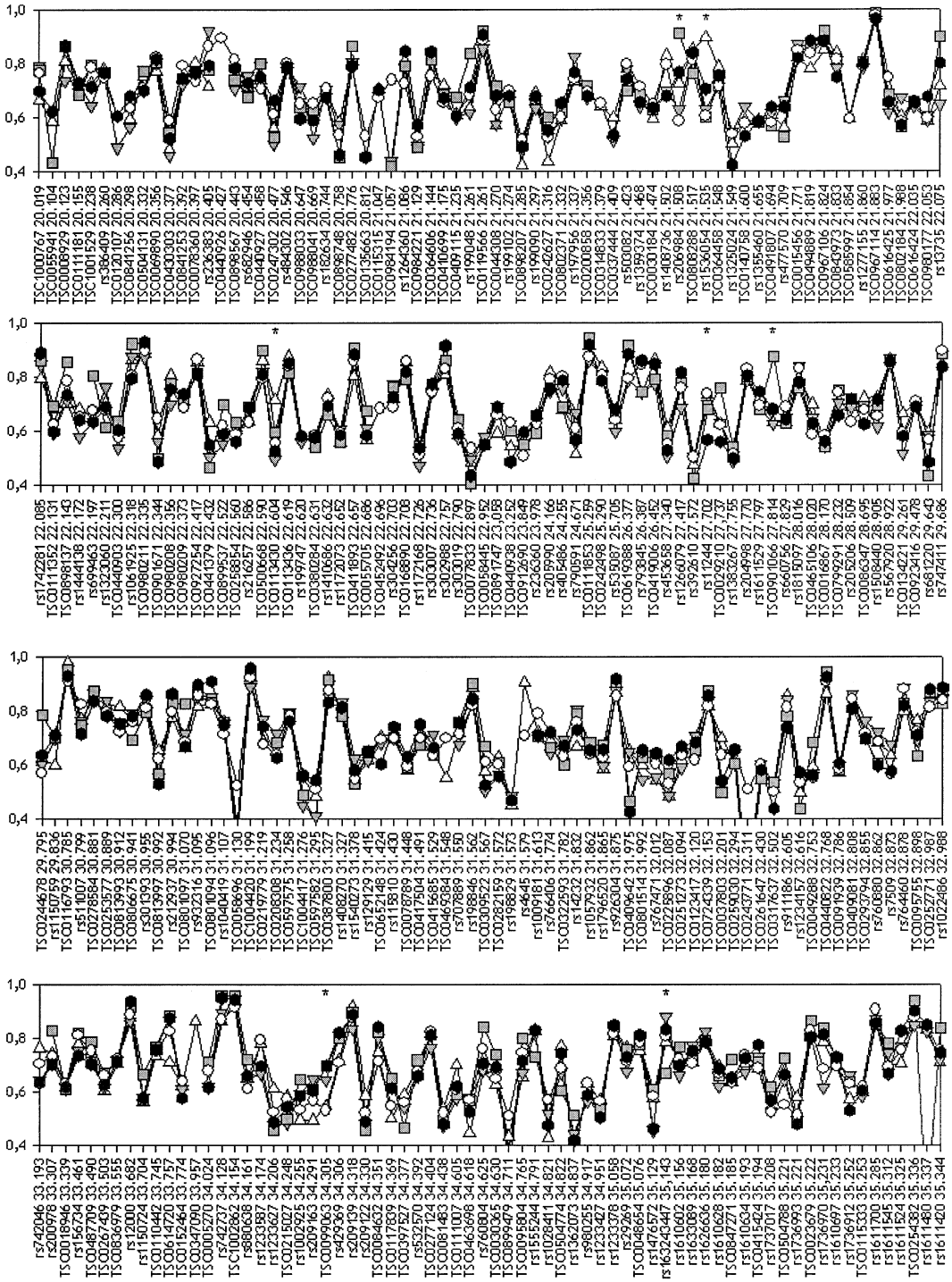

Fig. 4. Mean of allele frequencies in five DNA pools. Arbitrary placement by forcing physical order by best BLAST score on the current NCBI 26 (Ensembl) assembly between bp 20,000,000 and 50,000,000. Ensemble may give different results for a few markers, because their alignment is not restricted to the 6p21 area.

the ethical committee of the Bayerische Landesärztekammer.

*Control sample*

Samples were pooled from a recent population-based survey (KORA 2000) in the region of Augsburg, located in southern Germany. The pool consisted of 288 unselected individuals randomly drawn from all 4261 survey participants. Within this pool, 58 individuals were in the age range 25–35 years (29 males), 58 were 36–45 years (29 males), 57 were 46–55 years (29 males), 58 were 56–65 years (29 males), and 57 were 66–75 years of age (29 males). The study was approved by the ethical committee of the Bayerische Landesärztekammer.

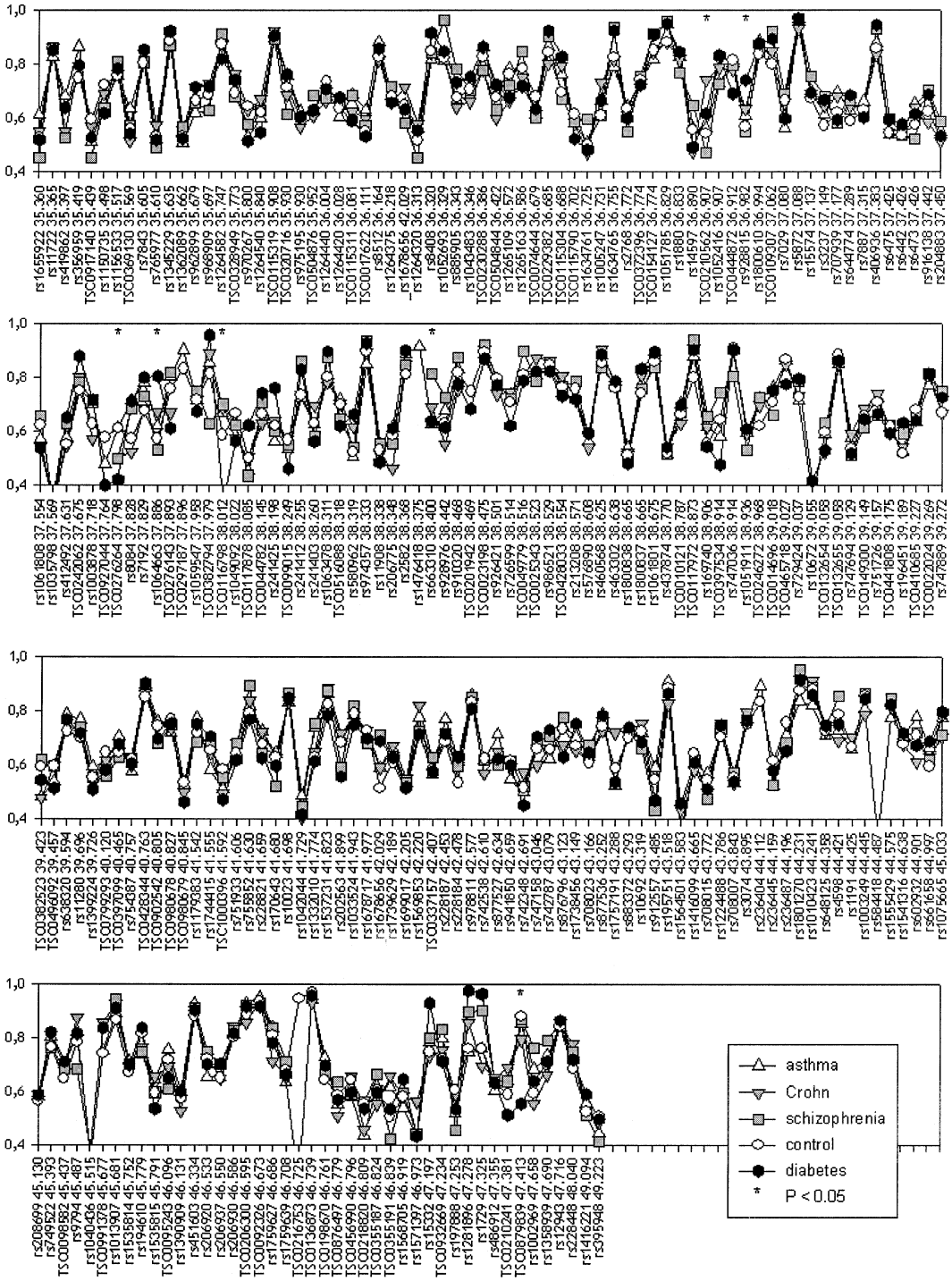


Fig. 4 (*continued*)

*Preparation of DNA pools*

For quantification of all individual DNA samples, we used a dsDNA-specific PicoGreen fluorescent dye (Molecular Probes, Eugene, Oregon) with a Genios fluorescence plate reader (Tecan). Before pooling, all samples were carefully adjusted to the same concentration. The particular DNA pools of sizes ranging from 76 to 288 individuals were constructed by mixing equimolar amounts of genomic DNA. The final concentration of the pooled DNA was adjusted to 1.7 ng/$\mu$l.

*SNP selection and verification*

We selected a first set of 418 SNPs from the TSC consortium (http://snp.cshl.org/) on ctg15907 in ~ 15-kb inter-

vals on the UCSC Golden Path Assembly (December 2000 freeze, http://genome.ucsc.edu/) from bp 25,050,386 to bp 44,467,453. In July 2001 we selected a second set both from the TSC consortium and the NCBI database. Of these 1435 SNPs (800 from NCBI, 617 from TSC, and 18 from other sources), 570 SNPs could be validated. In January 2002, the positions of 534 SNPs were confirmed by a BLAST search (ftp://ftp.ncbi.nih.gov/blast/) on the Ensembl 3.2.1 (http://www.ensembl.org/Homo_sapiens/) sequence from bp 20,000,000 to bp 50,000,000, which relies on the NCBI 26 assembly of the human genome and is annotated with known transcripts. Because the genomic map expanded considerably during the study interval, the median distance between SNPs increased from 15 kb to 24 kb.

*SNP analysis*

The determination of allele frequencies in pooled DNA was based on MALDI-TOF mass spectrometry of allele-specific primer extension products [37,38]. All assays for the PCR and associated extension reaction were designed by the SpectroDESIGNER software (Sequenom Inc., San Diego, California). Primer data are available at http://cooke.gsf.de/wjst/papers/GenomicsSupp103), and all reactions were run under the same conditions. Primers were obtained from MWG Biotech AG (Ebersberg, Germany) and Metabion GmbH (Planegg-Martinsried, Germany). The reaction volume of 50 $\mu$l contained 17 ng of pooled genomic DNA, 2 pmol of the first sequence-specific primer with a universal sequence at the 5′ end, 25 pmol of the second primer, 10 pmol of a biotinylated universal primer, 0.2 mM of each dNTP, 1.5 mM $MgCl_2$, reaction buffer, and one unit of Thermo-Start DNA Polymerase according to the manufacturer's protocol (ABgene, Epsom, UK). PCR conditions were an initial denaturation step for 10 minutes at 95°C, and then 45 cycles of 20 s at 95°C, 30 s at 56°C, 30 s at 72°C, and a final extension step for 10 minutes at 72°C. Each PCR was replicated three times. The biotinylated universal primer produced DNA strands complementary to the PROBE (primer oligo base extension) primer. Allele-specific primer extensions were conducted using the Mass EXTEND Reagents Kit based on biotin-streptavidin binding of the generated PCR products to paramagnetic beads on the MULTIMEK 96 automated 96-channel robot (Beckman Coulter, Fullerton, California). Primer extension products were loaded onto four positions of a 384-element chip nanoliter pipetting system (SpectroCHIP, SpectroJet; Sequenom) and analyzed using a MassARRAY mass spectrometer (Bruker Daltonik GmbH, Bremen, Germany). The resulting mass spectra were processed and analyzed for peak identification, peak area calculation, and allele frequency estimation using the SpectroTYPER RT 2.0 software (Sequenom).

*Data handling and analysis*

For each SNP position, differences in allele frequencies between a specific disease pool and the reference pool and differences between a specific disease pool and all other nondisease pools (up to four pools) were tested. By applying a linear mixed regression model with a random effect due to the single PCR, and measurement errors due to spotting primer extension products onto the chips, one can derive an estimation of the allele frequency of the disease pool or reference pool. In addition to the mean estimators of the allele frequencies, the corresponding variance estimations are obtained. The allele frequency for a nondisease pool is calculated as weighted sum over these pools. As weights, the inverse of corresponding variances are used. In a succeeding step, the usual $\chi^2$ statistic for testing independence for a $2 \times 2$ table is modified to control the uncertainty in estimating the true allele frequency. This is done by adding the two variance terms of the estimated allele frequencies to the denominator of the $\chi^2$ statistic (variance of the allele frequencies difference). To adjust for extra variation due to the construction of the pools the modified $\chi^2$ statistic is then multiplied by 0.456 and divided by the median of modified $\chi^2$ statistic over all SNP positions.

The resulting test statistic is an approximately distributed $\chi^2$ with 1 df under the null hypothesis [39]. To account for the multiple testing problem, the method of false discovery rates is applied [40]. In the analyses all SNP positions are tested if at least four of the five pools are available, and for each pool four and more measurements of the allele frequencies are derived from different PCRs. All analyses are done using SAS version 8.2 (procedure MIXED). Raw allele counts as well as the derived parameters from both methods are available from http://cooke.gsf.de/wjst/papers/GenomicsSupp103

## Acknowledgments

We thank all patients for their participation, Liane Thaller for secretarial assistance, Margret Bahnweg and Bettina Wunderlich for help with all laboratory work, Andreas Jendretzke for technical support, and Michelle Emfinger for proofreading of the manuscript. The project was funded by the Deutsche Forschungsgemeinschaft DFG WI621/5-1, GSF FE 73922, National Genome Network UW S15T01, UW S12T01, NV S02T08, Juvenile Diabetes Research Foundation JDRF 1-2000-619. BMBF competence network "Inflammatory Bowel Disease", and an EU-RTD grant.

## References

[1] D.C. Rao, D.A. Province (Eds.), Genetic Dissection of Complex Traits: an overview Academic Press, London, 2002, pp 13–34.
[2] J. Altmüller, L.J. Palmer, G. Fischer, H. Scherb, M. Wjst, Genome-wide scans of complex human diseases: true linkage is hard to find, Am. J. Hum. Genet. 69 (2001) 936–950.

[3] G.C. Johnson, et al., Haplotype tagging for the identification of common disease genes, Nat. Genet. 29 (2001) 233–237.
[4] N. Patil, et al., Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21, Science 294 (2001) 1719–1723.
[5] K.H. Buetow, et al., High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry, Proc. Natl. Acad. Sci. USA 98 (2001) 581–584.
[6] N.C. Arbour, et al., Homozygosity mapping of achromatopsia to chromosome 2 using DNA pooling, Hum. Mol. Genet. 6 (1997) 689–694.
[7] MHC sequencing consortium, Complete sequence and gene map of a human major histocompatibility complex, Nature 401 (1999) 921–923.
[8] C.M. Milner, R.D. Cambell, J. Trowsdale, Molecular genetics of the human major histocompatibility complex, in: R. Lechler, A. Warrens (Eds.), HLA in Health and Disease, Academic Press, London, 2000, pp. 35–50.
[9] R.E. Bontrop, The evolution of the major histocompatibility complex: insights from phylogeny, in: R. Lechler, A. Warrens (Eds.), HLA in Health and Disease, Academic Press, London, 2000, pp. 163–169.
[10] D.E. Reich, et al., Linkage disequilibrium in the human genome, Nature 411 (2001) 199–204.
[11] M. Carrington, Recombination within the human MHC, Immunol. Rev. 167 (1999) 245–256.
[12] P. Zavattari, et al., Major factors influencing linkage disequilibrium by analysis of different chromosome regions in distinct populations: demography, chromosome recombination frequency and selection, Hum. Mol. Genet. 9 (2000) 2947–2957.
[13] J. Hampe, et al., Association between insertion mutation in NOD2 gene and Crohn's disease in German and British populations, Lancet 357 (2001) 1925–1928.
[14] S.G. Schwab, et al., A genome-wide autosomal screen for schizophrenia susceptibility loci in 71 families with affected siblings: support for loci on chromosome 10p and 6, Mol. Psychiatry 5 (2000) 638–649.
[15] M. Wjst, et al., A genome-wide search for linkage to asthma. German Asthma Genetics Group, Genomics 58 (1999) 1–8, doi:10.1006/geno.1999.5806.
[16] J. Boretto, et al., Anonymous markers located on chromosome 6 in the HLA-A class I region: allelic distribution in genetic haemochromatosis, Hum. Genet. 89 (1992) 33–36.
[17] M. Amor, K.L. Parker, H. Globerman, M.I. New, P.C. White, Mutation in the CYP21B gene (lle172Asn) causes steroid 21-hydroxylase deficiency, Proc. Natl. Acad. Sci. USA 85 (1988) 1600–1604.
[18] D.E. Undlien, E. Thorsby, HLA associations in type 1 diabetes: merging genetics and immunology, Trends Immunol. 22 (2001) 467–469.
[19] N.J. Cox, et al., Seven regions of the genome show evidence of linkage to type 1 diabetes in a consensus analysis of 767 multiplex families, Am. J. Hum. Genet. 69 (2001) 820–830.
[20] M. Werner, et al., Large-scale determination of SNP allele frequencies in DNA pools using MALDI-TOF mass spectrometry, Hum. Mutat. 20 (2002) 57–64.
[21] E.W. Schlag, J. Grotemeyer, R.D. Levine, Do large molecules ionize?, Chem. Phys. Lett. 190 (1992) 521–527.
[22] M.S. Bray, E. Boerwinkle, P.A. Doris, High-throughput multiplex SNP genotyping with MALDI-TOF mass spectrometry: practice, problems and promise, Hum. Mutat. 17 (2001) 296–304.
[23] I.G. Gut, Automation in genotyping of single nucleotide polymorphisms, Hum. Mutat. 17 (2001) 475–492.
[24] A. Chakravarti, Population genetics-making sense out of sequence, Nat. Genet. 21 (1999) 56–60.
[25] D.E. Reich, E.S. Lander, On the allelic spectrum of human disease, Trends Genet. 17 (2001) 502–510.
[26] K. Asumalahti, et al., Coding haplotype analysis supports HCR as the putative susceptibility gene for psoriasis at the MHC PSORS1 locus, Hum. Mol. Gen. 1 (2002) 589–597.
[27] D.P. McGovern, D.A. van Heel, T. Ahmad, D.P. Jewell, NOD2 (CARD15), the first susceptibility gene for Crohn's disease, Gut 49 (2001) 752–754.
[28] A.T. DeWan, A.R. Parrado, T.C. Matise, S.M. Leal, The map problem: a comparison of genetic and sequence-based physical maps, Am. J. Hum. Genet. 70 (2002) 101–107.
[29] N. Katsanis, K.C. Worley, J.R. Lupski, An evaluation of the draft human genome sequence, Nat. Genet. 29 (2001) 88–91.
[30] S.S. Seki, et al., Stratification analysis of MICA triplet repeat polymorphisms and HLA antigens associated with ulcerative colitis in Japanese, Tissue Antigens 58 (2001) 71–76.
[31] A. Kawasaki, N. Tsuchiya, K. Hagiwara, M. Takazoe, K. Tokunaga, Independent contribution of HLA-DRB1 and TNF- promoter polymorphisms to the susceptibility to Crohn's disease, Genes Immun. 1 (2000) 351–357.
[32] M.F. Moffatt, et al., Association between quantitative traits underlying asthma and the HLA-DRB1 locus in a family-based population sample, Eur. J. Hum. Genet. 9 (2001) 341–346.
[33] J. Hampe, et al., Evidence for a NOD2-independent susceptibility locus for inflammatory bowel disease on chromosome 16p, Proc. Natl. Acad. Sci. USA 99 (2002) 321–326.
[34] J.P. Hugot, et al., Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease, Nature 411 (2001) 599–603.
[35] Y. Ogura, et al., A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease, Nature 411 (2001) 603–606.
[36] A.G. Ziegler, M. Hummel, M. Schenker, E. Bonifacio, Autoantibody appearance and risk for development of childhood diabetes in offspring of parents with type 1 diabetes: the 2-year analysis of the German BABYDIAB Study, Diabetes 48 (1999) 460–468.
[37] D.P. Little, et al., Detection of RET proto-oncogene codon 634 mutations using mass spectrometry, J. Mol. Med. 75 (1997) 745–750.
[38] D.P. Little, A. Braun, B. Darnhofer-Demar, H. Köster, Identification of apolipoprotein E polymorphisms using temperature cycled primer oligo base extension and mass spectrometry, Eur. J. Clin. Chem. Clin. Biochem. 35 (1997) 545–548.
[39] B. Devlin, K. Roeder, Genomic control for association studies, Biometrics 55 (1999) 997–1004.
[40] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. B57 (1995) 289–300.