INTRODUCTION

Gene Association Studies

In candidate gene association studies, one first identifies candidate genes that are hypothesized or known to be important in the pathogenesis of a condition. The next step is to identify polymorphisms within or close to the gene that could affect its regulation or function. Finally, one examines whether the polymorphisms occur more frequently in individuals who have a disease than in an appropriate control population. One of the major advantages of candidate gene association studies is that one uses knowledge of biologically plausible pathogenetic mechanisms to focus the search for genes on relatively few candidates. Another advantage is that the study subjects can be unrelated individuals so that genotypic and phenotypic data from multiple generations are not required. This is especially important in complex diseases such as chronic obstructive pulmonary disease (COPD) and atherosclerosis, in which the late age of onset makes it very difficult to ascertain DNA and phenotypic data from parents of affected individuals. However, there are several issues that negatively influence population-based genetic association studies that need to be addressed (1), including increased false positives due to hidden population stratification or admixture (2,3).

False-positive associations (type I errors) can occur if the frequencies of genetic markers and of the disease of interest vary across different population groups. Freedman et al. (4) and Marchini et al. (5) have recently reviewed and generated new data on the impact of population stratification on genetic association studies.

To avoid problems due to population stratification in association studies, both the cases and controls should be selected from the same population/ethnic group and geographic area. Because population genetic background is difficult to measure, it is impossible to guarantee genetic homogeneity. Therefore, alternative strategies based on the use of families have been employed. Such an alternative method of analysis is the transmission/disequilibrium test (TDT; Reference 6), which simultaneously tests linkage and association. The TDT evaluates the frequency of transmission of specific alleles at a single locus from heterozygous parents to their affected children. In the absence of association, each allele is expected to be transmitted with the Mendelian frequency of 50%. If a marker allele is transmitted significantly more often than 50% of the time, this implies that the allele must be linked to the disease-causing allele. The main advantage of the TDT is that it does not compare groups of cases and controls and therefore is not generally susceptible to population stratification. However, a major drawback of the TDT method is that it requires parental DNA, which is often unobtainable in studies of
late-onset diseases such as COPD. An example of this approach is the finding of an association between the IL4-590 polymorphism with asthma (7). The T allele of this polymorphism was transmitted from a heterozygous parent to an affected child in 64% of the informative meioses.

Confounding due to population stratification in population-based association studies can cause either biased and/or overdispersed test statistics, leading to false positives. While family-based association studies are not subject to the same problem, they do have their own limitations such as the recruitment of families. Therefore, methods estimating and correcting the test statistics for the effects of population substructure have been developed. Genomic control (GC) and structured association (SA) are two such methods that have been developed.

SA is a statistical method to test for association in the presence of hidden population substructure (8,9). It is a “latent-class” method, which assumes that the sample is composed of individuals from $K$ latent subpopulations, each having a characteristic set of allele frequencies at marker loci. Unlinked genetic marker loci are used to estimate subpopulation parameters. In the first step, a Markov Chain Monte Carlo method is used to estimate $K$, the allele frequencies in each subpopulation, and a set of vectors $q = (q_1, \ldots, q_K)$ representing the proportion of each individual’s genome from each of the $K$ subpopulations. In the second step, a test statistic that conditions on inferred subpopulation is calculated to account for the population substructure. Because SA actually attempts to infer ancestry, it can require up to several hundred unlinked “null” marker loci (4,10).

GC is another method to test for candidate gene-disease association in the presence of population substructure (11). GC corrects for the confounding due to population stratification, admixture, or cryptic relatedness. The GC approach uses a set of bi-allelic markers randomly chosen from the genome to model the overdispersion. In brief, the method involves computing the $\chi^2$ association test statistics at both the null and candidate loci. To correct for the overdispersion, the $\chi^2$ statistic at the candidate loci is divided by the variance inflation factor, $\lambda$, where $\lambda$ is estimated from the test statistics at the null loci. If there is no association between the candidate loci and the disease, the test statistic $\chi^2/\lambda$ has a $\chi^2_1$ distribution (11,12). In simulations, Bacana et al. (13) found that the median estimator of $\lambda$ (median of the $\chi^2$ statistics divided by 0.456) performed well with 50 or more null loci. The method was initially developed for case-control studies and bi-allelic candidate genes. It has been extended to multi-allelic markers/haplotypes and quantitative traits (14).

GC has the advantages of simplicity, robustness, and wide applicability, and promises (along with SA) to become one of the accepted methods of choice for population stratification control in gene association studies (4,5,10,15–17). However, both methods increase the genotyping burden because the research laboratory is forced to pay for genotypes that are not likely to be associated with the trait of interest.

Genotyping

A number of high-throughput genotyping methods [e.g., matrix-assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF), Sequenom (18), TaqMan® (19), and Pyrosequencing™ (20)] have been developed. In general, these techniques have been engineered to optimize the genotyping of large numbers of individuals for one single nucleotide polymorphism (SNP) at a time, as part of the discovery phase of genomics research to establish significant associations of genotype and a particular clinical outcome or response to treatment. An “ideal, gold standard” technology would be accurate, fast, robust, automated (including automated genotype calling), scaleable, inexpensive (both in operating reagent costs and in capital equipment costs), and applicable to single sample and patient use.

Microarray Genotyping

Of the many methods that have been developed for genotyping, those based on the use of microarrays offer the greatest potential for economic, patient-specific application, due to their ability to simultaneously interrogate multiple genetic markers (SNPs) using genetic material (template) that is PCR amplified from an individual. Genotyping microarrays are devices displaying specific oligonucleotide probes, precisely located on a small-format solid support. There are several microarray genotyping protocols, which include the following.

GeneChips®. Multiple, redundant microarray probes are hybridized to amplified and labeled patient template DNA. This technology relies on the discriminating power of hybridization alone to differentiate a single base change. It requires multiple probes for each SNP site due to low intrinsic signal to noise (poor sensitivity); the chip has to be a generic commercial Affymetrix design (expensive; Affymetrix, Santa Clara, CA, USA); and hybridization takes several hours or even overnight (21). An SA study using Affymetrix chips has recently been reported by Hinds et al. (10).

Tagged/ZipCode Arrays. Several methodologies rely on the hybridization of a dual-purpose oligonucleotide probe (carrying a tag sequence as well as an SNP-specific sequence) to a complementary probe on an array. Examples include single base extension-tag array on glass slides (SBE-TAGS; Reference 22) and the ultra-high-throughput beadarray system developed by Illumina (San Diego, CA, USA; Reference 23). These methods, although sensitive, require complex oligonucleotides and multiple reaction steps, including a final hybridization stage on the array that can take several hours.

Arrayed primer extension (APEX). In this method (24,25), the actual SNP probes themselves are attached to the microarray slide, and a single base extension (minisequencing) reaction is performed in silico using PCR-amplified patient DNA as a template. The combination of hybridization and single base extension yields greater discrimination and lower background. The actual genotyping assay itself is very fast, taking only 20 min.

APEX Microarrays

APEX is a resequencing method
based on a two-dimensional array of oligonucleotides, immobilized via their 5’ ends on a glass surface. The oligonucleotides (from 15- to 25-mers) are designed so that they are complementary to the gene up to, but not including, the base where the SNP exists. APEX combines the advantages of a highly parallel and miniaturized array-based method with the discriminatory power of the Sanger dideoxy terminator sequencing method (26). APEX has been used to detect β-thalassemia (27), TP53 (28), and BRCA1 mutations (29). APEX has recently been shown to be efficient at analyzing chromosome-wide SNP markers (30).

Here we describe the design, development, and testing of an APEX GC chip that can accurately determine the genotypes of up to 110 SNPs that are randomly scattered across all human autosomal chromosomes. This represents a useful resource for determining hidden population stratification in disease-association studies.

MATERIALS AND METHODS

GC SNP Selection

One hundred and 24 SNPs were selected for the experimental design of the genotyping chip (see Supplementary Table S1 at the BioTechniques web site at http://www.BioTechniques.com/December2004/TebbuttSupplementary.html). Information on 58 of these SNPs (including allele frequency data) was kindly given by David Reich and colleagues, who had previously tested these SNPs for use in assessments of population stratification (3,4), but who, for reasons of confidentiality, were unable to provide individual ethnic frequency data. We supplemented these 58 markers with 66 additional SNPs taken from Celera’s “allele_frequencies-CELERA. compact” file, downloaded from the SNP Consortium web site (http://snp.cshl.org/) and randomized. The mean frequencies were posted across multiple ethnic groups for the Celera data (as published on the SNP Consortium’s web site). We simply chose the first 66 SNPs in this randomized list, leaving out only SNPs that occur on the X or Y chromosomes plus any that are within 2–3 megabases of any previously selected SNPs.

Human DNA Samples and Validated Genotypes

D. Reich, M. Freedman, and colleagues generously provided independent genotypes for 48 of the 58 SNPs that they had recommended to us, across a number of African American Coriell DNA samples [NA17101–NA17112, inclusive (http://coriell.umdnj.edu/)]. We obtained DNA aliquots of these 12 Coriell samples for APEX genotyping. Of the 66 SNPs selected from Celera, we were able to download genotypic information for all 66 markers across the same Coriell individuals (except for NA17110, which had only limited genotypes) from the SNP Consortium and NCBI web sites (http://www.ncbi.nih.gov/SNP/).

GC SNP Chip-Probe and PCR Design

We employed the services of Biodata Ltd. (Tartu, Estonia) to design the single base extension genotyping probes and PCR primers for each of the 124 SNPs. Six oligonucleotide probes for each SNP were designed: two classical APEX probes (one probe per DNA strand), plus four additional allele-specific arrayed primer extension (AS-APEX) probes per SNP marker (two probes per strand) that include the actual SNP site at the 3’ end of the probe. Allele-specific single base extension of these AS-APEX probes during the reaction is contingent on the presence of the actual complementary base at the SNP site in the patient template DNA (27,31). This provides a useful redundancy in our probe set that gives us considerable robustness in our genotyping and thus confidence in the data.

PCR primers were designed based on a melting temperature ($T_m$) of 62° ± 3°C (at 20 mM monovalent salt concentration in PCR buffer). All primers were computationally tested against the human genome and found to amplify single product. During the computational design process, one of the initial 58 SNPs from Reich and Freedman (SNP: rs1362189) was found to be located in a repeated region that has three highly similar regions in the genome. This SNP was removed from any further analysis and thus from the chip because Biodata were unable to find primers that amplify a unique PCR product for this marker.

All APEX probe and PCR primer oligonucleotide sequences and associated data for the final 123 SNPs are listed in supplementary material in Tables S2 and S3. Oligonucleotides were synthesized at a 50-nmol scale and aliquoted into 96-well plates at 200 pmol/µL stock concentration (MWG-Biotech, High Point, NC, USA). Each of the APEX and AS-APEX probes were “AminoLink” (MWG-Biotech)-modified at the 5’ end. Biodata algorithms suggested the grouping together of multiple PCR primer pairs that would multiplex the entire 123 PCR product set in just 16 PCRs. Details of these multiplexed pairs are listed in the supplementary material in Table S4.

Microarray Printing

Using BioRobotics MicroGrid™ spotters (Matrix Technologies, Cambridge, UK), the APEX and AS-APEX probe oligonucleotides (50 pmol/µL in 150 mM sodium phosphate print buffer, pH 8.5) were printed to specific grid positions on CodeLink™-Activated microarray slides (Amersham Biosciences, Piscataway, NJ, USA), following the manufacturer’s recommended protocols. The 5’ end of each probe oligonucleotide is amino-modified, allowing its covalent attachment to the slide’s preapplied surface chemistry. Each grid consisted of duplicate spots of each of the six probes per SNP as well as multiple buffer-only spots and positive-control spots. The latter comprised two types of positive controls: multiple combinations of self-extending control oligonucleotides, designed to extend to one or more of the four DNA bases, A, C, G, and T (SeqN probes; Reference 25), and an oligonucleotide probe based on a plant-specific gene sequence (Npg1; Reference 32) that will extend by a single N base due to the presence of an
exogenous complementary template oligonucleotide in the APEX reaction mixture. The Npg1-positive control probe sequence and its corresponding template oligonucleotide sequence are listed in the supplementary material in Table S2. Two replicate grids were printed on each slide, enabling two samples to be genotyped per slide.

Each Npg1- and SeqN-positive control probes were spotted at least 80 times onto the grids, at regular physical intervals (Figure 1). Each of the six probes for each SNP was printed at a reasonably wide distance apart from any other probe for the same SNP within the grid (as were their duplicate spots). This enables a useful degree of robustness in the system, which is especially helpful in cases of high local background and hybridization problems. Each spot was approximately 110 µm in diameter.

Following the printing of the arrays, the slides were incubated overnight at room temperature at 75% relative humidity to drive the covalent coupling reaction between the probes’ 5’ amino group and the CodeLink slide chemistry to completion. Blocking of the arrays was in 50 mM ethanolamine, 0.1 M Tris, pH 9.0, 0.1% sodium dodecyl sulfate (SDS) at 50°C for 15 min, according to the manufacturer’s protocol.

**Coriell DNA Sample PCR Amplification**

One hundred and 23 SNP primer pairs were designed, and a multiplex PCR amplification was performed on Coriell genomic DNA samples (plus a negative PCR control sample that contained no genomic DNA). The multiplex PCR group had a unique combination of the primer pairs among 16 reactions (see Supplementary Table S4). Each PCR was performed in a total volume of 25 µL containing 2.5 µL 10× PCR buffer [Tris-Cl, (NH₄)₂SO₄, 15 mM MgCl₂, pH 8.7], 3 mM MgCl₂, 200 µM dNTPs without dTTP, 160 µM dTTP, 40 µM dUTP, 1.25 U HotStarTaq DNA polymerase (5 U/µL; Qiagen, Valencia, CA, USA), 1 µL 7.7 mM primer mixtures, and 10 ng genomic DNA. The incorporation of the dUTP allows the amplified DNA to be enzymatically sheared by uracil N-glycosylase (UNG; EPICENTRE, Madison, WI, USA) to produce a DNA size of about 100 bases, which is optimal for hybridization to the oligonucleotides on the microarray. PCRs were initiated by a 15-min polymerase activation step at 95°C and completed by a final 10-min extension step at 72°C. The cycles for PCR were as follows: 25 cycles of a 30 s denaturation at 95°C, 30 s annealing at 60°C, and a 50 s extension at 72°C; followed by another 10 cycles of 30 s at 95°C, 30 s at 58°C, and 50 s at 72°C. Aliquots of all PCR products were visualized with ethidium bromide staining under ultraviolet light (UV) on a 2.5% agarose gel following electrophoresis in 0.5× Tris-borate EDTA (TBE) buffer.

**DNA Purification and Fragmentation**

The 16 multiplex PCR products were pooled for each individual Coriell sample (and negative control), and the total volume was measured. After adding 2.5 volumes of ice-cold 100% ethanol and 0.25 volumes of 10 M ammonium acetate solution, the PCR products were precipitated at -20°C overnight. The mixture was centrifuged at 20,800 × g at 4°C for 20 min. The supernatant was carefully removed, and the DNA pellet was washed with 400 µL of ice-cold 70% ethanol. The DNA pellet was then dissolved in 15 µL double-distilled water. Three micro-\riators of the precipitated DNA were run on a 2.5% agarose gel.

The remaining precipitated DNA was then fragmented, and unincorporated dNTPs were inactivated by digestion with 1 U UNG and 1 U shrimp alkaline phosphatase (SAP; Amersham Biosciences) for 1 h at 37°C, followed by enzyme inactivation for 10 min at 95°C in a 20-µL reaction mixture containing 2 µL 10× digestion buffer [0.5 M Tris-HCl, pH 9.0, 0.2 M (HN₄)₂SO₄]. Three microliters of digests were visualized by electrophoresis on a 2.5% agarose gel followed by ethidium bromide staining.

**Microarray-Based Genotyping: APEX**

The APEX reaction was performed in a total volume of 50 µL by the addition of 17 µL fragmented DNA template, 1 µL 2 pmol/µL Npg1-positive control template oligonucleotide, 1 µM each fluorescently labeled dideoxy nucleotide triphosphate (Texas Red®, ddATP, Cy₃-ddCTP, Cy5-ddGTP, R110-ddUTP; Perkin Elmer Life Sciences, Boston, MA, USA), 5 U Thermo Sequenase™ DNA polymerase (Amersham Biosciences) diluted in its dilution buffer, to 2× Thermo Sequenase reaction buffer (10×, 260 mM Tris-HCl, pH 9.5, 65 mM MgCl₂). The reaction mixture was applied to the array of APEX and AS-APEX probes previously printed on the CodeLink slide, which had been washed twice in 95°C double-distilled water and placed on a HyPro100 incubation plate.
(Thermo Electron, Milford, MA, USA) set at 58°C. The reaction mixture was covered with a small piece of Parafilm®, and the APEX reaction allowed to proceed at 58°C with agitation (setting 1) for 20 min. Following the incubation period, the slides were washed with 95°C double-distilled water to remove the template DNA, enzyme, and excess ddNTPs. Further washing in 0.3% alcanox and 95°C double-distilled water ensured low background on the array images.

**Microarray Imaging and Spot Intensity Calculation**

Slide microarrays were imaged using an arrayWoRx® Auto Biochip Reader (Applied Precision, LLC, Issaquah, WA, USA), fitted with the following filter sets: (i) A488-Ex. 480/30–Em. 530/40 (R110 dye); (ii) Cy3 (narrowband)-Ex. 546/11–Em. 567/15 (Cy3); (iii) Texas Red-Ex. 602/13–Em. 631/23 (Texas Red); and (iv) Cy5-Ex. 635/20–Em. 685/40 (Cy5). Exposure times for each dye were set up to give 60%–70% pixel saturation for selected Npg1-positive control probes. The resolution of the imager was set to 10 µm.

Four 16-bit TIFF files for each sample were obtained (one from each channel), and these were analyzed using Genorama™ software (version 4.2; Asper Biotech, Tartu, Estonia). Spot intensity values (using default settings except for “local background subtraction” not selected) and probe name/grid coordinates were exported to Microsoft® Excel®.

**RESULTS AND DISCUSSION**

**GC SNP Selection**

For the assessment of hidden population stratification by GC, it is important to select random, effectively unlinked SNPs from the genome (3,4). SNPs are not chosen with regard to their allelic frequencies comparing different populations around the world. At the same time, it is important that the minor allele frequencies of the SNPs be substantial, at least 15%. Going to frequencies as low as 5% is acceptable if the SNP allele that is being assessed for stratification has a similarly low frequency.

Supplementary Table S1 shows the mean minor allele frequencies for the 123 SNPs that were selected for the GC APEX chip, showing a reasonable distribution from 5% to 50%. Table S1 also lists the chromosomal location of each SNP, with potential functional implication with respect to genes. Although our selection does have intrinsic bias with respect both to the original 57 SNPs provided by Reich and Freedman and colleagues (4) and also to the fact that we only used the Celera data file for the remaining 66 SNPs, it is clear...
that there is still useful “randomness” in the final selection of 123 SNPs. For example, the chromosomal distribution of these SNPs relates to the sizes of the autosomes, with 11 SNPs occurring on the largest chromosome (chromosome 1), while only 2 SNPs occur on the smallest chromosome (chromosome 22). In addition, the majority of the 123 SNPs selected are non-gene coding-sequence associated, consistent with the low percentage of gene-related sequence in the genome (despite likely bias in the preselection of the SNPs by both Celera and other SNP discovery researchers who might have originally favored gene-rich regions for study). In fact, 83 SNPs were classified as intergenic, 32 SNPs were intronic, 5 SNPs occurred in the untranslated region of mRNAs, and only one SNP was in a gene-coding region (synonymous polymorphism). The two remaining SNPs were unclassified.

**GC APEX Chip Genotyping Performance**

Genotyping for all of the 123 GC SNPs was carried out on 12 Coriell DNA samples (NA17101–NA17112, inclusive). One of the samples (NA17111) was genotyped twice on duplicate arrays at different times using independent PCR template. The genotypes from these duplicated NA17111 sample experiments were treated independently in the analysis. SNP Chart software (http://www.snpchart.ca) was used to simultaneously display the four-channel spot intensity data for all of the APEX and AS-APEX probes assayed for a given SNP and a particular Coriell individual. The caller then used this signal information to manually score

Figure 2. SNP Charts for rs1474537. Examples of SNP Chart outputs easily interpreted as specific genotypes. The SNP is rs1474537, which is a CT (based on the “LEFT” DNA strand). Coriell sample NA17108 is CC; NA17106 is TT; NA17104 is CT. Spectral overlap of the R110 dye (conjugated to ddUTP) into the Cy3 channel, due to suboptimal microarray scanner filter sets, results in a non-sequence-specific C signal whenever a T is present. However, the robustness of the experimental design can cope with this. SNP, single nucleotide polymorphism.
either a genotype or a null call. Two callers independently scored genotypes for each of the Coriell DNA samples across all 123 SNPs. Figure 2 shows output SNP Charts for genotypes at the rs1474537 SNP for homozygous major allele, homozygous minor allele, and heterozygous individuals.

Null and Discrepant Calls

A null call (NN) would be made if the individual researcher felt that there was not enough information to accurately call the genotype or if the signals across the six probes gave inconsistent results. Whether an individual chose to call a genotype or not also depends on their experience with SNP Chart and their “conservativeness.” For 10 SNPs, null calls were made by both callers for greater than 90% of the Coriell DNA samples, indicating either PCR failure to amplify those templates or total probe failure. These SNPs were rs1033917, rs1259606, rs1318933, rs1344617, rs1372727, rs1452635, rs1460540, rs1547958, rs760998, and rs975302. When these SNPs are removed from the analysis, the number of null calls drops dramatically.

The proportion of discrepant calls is defined as the number of discrepant calls between the two callers divided by the total calls made by both callers, excluding null calls. Of 1244 non-null calls made by both, 57 were discrepant, giving a proportion of 0.046. The proportion of discrepant results can be used as one measure of the error due to caller subjectivity because the two callers cannot both be correct. However, it will underestimate the true error rate because it is possible that both callers made the same genotype call but that that call is incorrect (see below). If we assume that one caller is more experienced and always correct, the error rate for the other caller is at least 0.046. Assuming equal error rates between the two, each has an error rate of at least 0.023. So the estimated error rates for each caller would be somewhere in the middle; however, as mentioned, this underestimates the true error rate because it assumes that both callers do not make the same error. For 34 of the SNPs, there was 100% concordance between both scorers for all samples.

Table 1 summarizes the null call analysis for the 113 SNPs that gave callable intensity data from the microarrays. These remaining null calls and the discrepancies are measures of a combination of array data quality and caller subjectivity (including eye fatigue). The total number of identical scores made by the two callers was 1187.

The number of null and discrepant calls can be heavily influenced by intrinsic problems associated with spotted microarrays, including pin failure on the printing robot and localized hybridization failure (both leading to missing data), spot morphology problems, and high local background issues (affecting channel intensity calculations in Genorama). Figure 3 shows an example of one of the four color images from a selected Coriell sample array experiment, clearly illustrating some of these problems. Both of the callers were blinded to these array inconsistencies and were reliant on the basic SNP Chart data that comprised the redundant and duplicated probe spots for each SNP. In addition, there was an initial error in the SNP Chart display of one of the probes for each SNP. This often led to inconsistency in the signal patterns that were observed, potentially increasing the errors or the percentage of null calls.

Final Genotyping

A third researcher used SNP Chart (with the software display error corrected) to make a final call for genotypes showing either discrepancies or null calls between the original two callers. This researcher was also given access privileges to the four color micro-

<table>
<thead>
<tr>
<th>Table 1. Summary of Null Calls</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No. of Null Calls</strong></td>
</tr>
<tr>
<td>--------------------------</td>
</tr>
<tr>
<td>Caller 1</td>
</tr>
<tr>
<td>Caller 2</td>
</tr>
<tr>
<td>Common</td>
</tr>
</tbody>
</table>

Figure 3. Genotyping microarray image. An example of one of the four color images from a selected Coriell sample array experiment, clearly illustrating some of the intrinsic issues with microarray data, including pin failure on the spotting robot and localized high background problems. The robustness of our experimental design is somewhat able to cope with these issues.
array images for each of the Coriell samples, enabling a quality-control step to be made with respect to inconsistencies in the arrays that could have led to the original null and discrepant calls. This researcher was unable to change any of the 1187 genotypes that had already been independently scored the same by the original two callers. A total of 282 genotypes (1469 minus 1187; 19%) had to be reassessed. Following this analysis, three additional SNPs (rs1039680, rs1156335, and rs966514) were deemed to have failed APEX genotyping because of a prevailing high number of null calls across multiple Coriell samples.

Therefore, of the original 123 SNPs represented on the GC APEX chip, we failed 13 (just over 10%). These failures were due to either initial PCR failure or to APEX/AS-APEX probe sequence issues, leading to low true signal intensities and/or high false signals. We have not included these 13 SNP failures in our overall accuracy rate calculation because they likely represent “systematic” failures that would be reproduced for any DNA sample genotyped, rather than “random” failures that would indeed need to be taken into account for accuracy rate calculation. A genotype was scored across all the remaining 110 SNPs for 13 DNA samples (1430 possible genotypes), except for a single null call (NN) for Coriell NA17105 for the rs1003399 SNP and for three null calls (NA17104, NA17111_A, and NA17111_B) for rs1362195 (see Supplementary Table S5).

Genotype Validation

We were able to obtain independent genotypic data for 101 of the 110 successful SNPs, from a combination of data provided by D. Reich, M. Freedman, and colleagues, and by mining the NCBI database (http://www.ncbi.nlm.nih.gov/SNP/). Our laboratory Coriell sample NA17101 was found to have a high (>25%) number of genotype errors, and it was subsequently determined that this sample was not in fact NA17101, rather it had been wrongly labeled at the source. The data from this sample were removed from further analysis. Although there were some missing genotypes for particular Coriell samples, we could still validate our microarray-based data against 1141 genotypes. Of these 1141, we found 1124 to be identical to our data, with a single null call (0.1%) and 16 miss calls (1.4%) for a combined error rate of 1.5%. These 17 errors occurred across 17 different SNPs, although 8 of them were in a single Coriell sample (NA17112). A complete repeat of the PCR and array APEX assay for this Coriell NA17112 sample showed accurate genotypes. Nine of the 16 miss call errors (5 of which were in NA17112) had been identically scored by the two original scorers and thus were not subsequently looked at by the third researcher in this study. It does represent a 0.8% intrinsic error rate. Supplementary Table S5 shows the complete list of the genotypes scored.

Interestingly, within the 1141 genotypes to which we were able to compare our data, we identified just two inconsistencies between the two database genotype resources. The first one was for Coriell sample NA17109 at SNP rs940870: the NCBI database denotes this as GG, while Reich and Freedman call it AG. Our own data were consistent with the AG call. The second inconsistency was again for Coriell sample NA17109, but this time at SNP rs987007: Reich and Freedman call this TT, while the NCBI database actually has two independent scores of TT and TG. Our own data were consistent with the TT call.

SNP Chart generates a visualization of spot intensity values from multiple channels from a multiple probe set specific for a given SNP and DNA sample. This visualization can be easily interpreted by the trained human eye as a specific SNP genotype. However, this is a time-consuming and not very practical proposition for high-throughput genotyping. Automation of the calling of genotypes, based on the information displayed in SNP Chart, is currently being developed. This will allow faster analysis and reduce user-subjectivity issues. It is unlikely that any automatic scoring algorithm will be perfect for all SNPs, however, and the data visualization of SNP Chart will provide a useful “manual override” in cases of null calls.

SUMMARY

In comprehensive statistical reanalyses of multiple disease-gene association data sets, Freedman et al. (4) argue the critical importance of reporting not only positive associations but also the significance of detected associations when corrected for the maximum stratification consistent with the data. Such additional information will be essential to give more appropriate confidence limits to any positive gene-disease associations and will surely help to reduce the plethora of unrepeatable (and therefore unvalidated) gene associations in the literature. Importantly, Freedman et al. (4) and Marchini et al. (5) suggest that even lower levels of stratification will need to be accounted for as future disease-gene association studies involve ever-increasing numbers of individuals and genetic markers.

The 98.5% accuracy rate (with both null calls and miss calls included in the 1.5% error rate) for our GC APEX chip is comparable to other published array-based single base extension genotyping assays (34–37), demonstrating its effectiveness for up to 110 SNPs that are widely distributed across the human genome. We believe that this represents a useful resource for determining hidden population stratification in gene-disease association studies, by way of GC.

ACKNOWLEDGMENTS

J.Q.H. and K.M.B. contributed equally to the work summarized in this paper. We would like to thank David Reich and Matt Freedman for initial advice and access to extensive genotype data prior to publication, Dean English for assistance with the manuscript figures, and Peter Paré for continued support. This research was supported by the Canadian Institutes of Health Research, CANARIE, and the Michael Smith Foundation for Health Research.

COMPETING INTERESTS

The authors declare no competing interests.
REFERENCES


Received 1 June 2004; accepted 29 July 2004.

Address correspondence to:
Scott J. Tebbutt
James Hogg iCAPTURE Centre for Cardiovascular and Pulmonary Research
St. Paul's Hospital
University of British Columbia
Room 388 Burrard West-1081 Burrard Street
Vancouver, BC, V6Z 1Y6, Canada
email: stebutt@mlr.ubc.ca