Identification of Susceptibility Genes and Genetic Modifiers of Human Diseases

Kenneth Abel, Stefan Kammerer, Carolyn Hoyal, Rikard Reneland, George Marnellos, Matthew R. Nelson, Andreas Braun*
SEQUENOM, Inc., 3595 John Hopkins Court, San Diego, CA, USA 92121

ABSTRACT

The completion of the human genome sequence enables the discovery of genes involved in common human disorders. The successful identification of these genes is dependent on the availability of informative sample sets, validated marker panels, a high-throughput scoring technology, and a strategy for combining these resources. We have developed a universal platform technology based on mass spectrometry (MassARRAY) for analyzing nucleic acids with high precision and accuracy. To fuel this technology, we generated more than 100,000 validated assays for single nucleotide polymorphisms (SNPs) covering virtually all known and predicted human genes. We also established a large DNA sample bank comprised of more than 50,000 consented healthy and diseased individuals. This combination of reagents and technology allows the execution of large-scale genome-wide association studies. Taking advantage of MassARRAY's capability for quantitative analysis of nucleic acids, allele frequencies are estimated in sample pools containing large numbers of individual DNAs. To compare pools as a first-pass "filtering" step is a tremendous advantage in throughput and cost over individual genotyping. We employed this approach in numerous genome-wide, hypothesis-free searches to identify genes associated with common complex diseases, such as breast cancer, osteoporosis, and osteoarthritis, and genes involved in quantitative traits like high density lipoproteins cholesterol (HDL-c) levels and central fat. Access to additional well-characterized patient samples through collaborations allows us to conduct replication studies that validate true disease genes. These discoveries will expand our understanding of genetic disease predisposition, and our ability for early diagnosis and determination of specific disease subtype or progression stage.

Keywords: Genome-wide association, SNP, disease susceptibility genes, MALDI-TOF, mass spectrometry, genotyping, DNA pooling

1. INTRODUCTION

As a result of the Human Genome Project, genomic research is now entering an era where emerging data will permit investigators to decipher the genetic components of complex diseases. The focus of human genetics in recent years has shifted toward searching for genes that are involved in the development of common diseases such as cancer, diabetes, cardiovascular and Alzheimer's diseases. In the recent past, mainly two approaches have been applied: association studies using markers in candidate genes in samples of unrelated case and control individuals, and linkage analyses using markers of genome variation in large families with multiple affected individuals. Both of these approaches have critical limitations. Candidate gene association studies are generally limited to a small number of genes already expected to be involved in the disease pathway, and thereby provide little opportunity for novel gene discovery. While family-based linkage studies on the other hand have proven very successful for identifying genes with highly penetrant alleles segregating in a simple mendelian fashion within families[1], in very few cases have they resulted in identification of susceptibility genes for common, complex diseases. Successes using linkage methods have largely been restricted to rare diseases and to selected "familial" subsets of common disorders, accounting for only a small proportion of disease in the population. Additionally, the inherent low resolution of linkage studies, which frequently point to genomic regions containing hundreds of genes, usually precludes identification of the implicated susceptibility gene in a timely and cost-effective manner.

For common disease susceptibility genes, direct association approaches that compare the distribution of genetic marker frequencies between groups of unrelated individuals are expected to have greater power than traditional linkage studies[2]. Recently there has been increasing interest in the use of whole-genome association methods to

Imaging, Manipulation, and Analysis of Biomolecules and Cells: Fundamentals and Applications III, D. V. Nicolau, J. Enderlein, R. C. Leif, D. L. Farkas, R. Raghavachari, eds., Proc. of SPIE Vol. 5699 (SPIE, Bellingham, WA, 2005) · 1605-7422/05/$15 · doi: 10.1117/12.604620

identify genes that are involved in complex trait variation[3-5], although to date few large-scale studies have been reported. To discover genetic factors contributing to the etiology and pathophysiology of complex disorders, high-throughput analysis of a comprehensive set of genetic markers in suitable samples sets is a necessity. We have adopted an approach for disease gene identification by genome-wide association studies, using an automated chip-based method for mass spectrometric analysis of single nucleotide polymorphisms (SNPs). Here we discuss the development of the technology, reagent resources, and universal approach to identifying genes implicated in common diseases, and describe results from the successful implementation of this approach in several genome-wide case/control studies.

2. METHODOLOGY

2.1 MassARRAY: High-throughput DNA analysis by chip-based mass spectrometry

In numerous applications it has been demonstrated that the analysis of biochemically generated DNA products can be carried out using matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometry for separation and detection[5-10]. Briefly, the technology involves PCR amplification of the region containing the SNP of interest, an optimized primer extension reaction to generate allele-specific DNA products, and chip-based mass spectrometry for separation and analysis of the DNA analytes. Figure 1 illustrates an assay designed to distinguish mutations in codons 5 and 6 of the human β-globin gene. A single post-PCR primer extension reaction generates diagnostic products that, based on their unique mass values, allow discriminating between wildtype (HbA) and three mutant alleles potentially present at this site (HbS, HbC, and a two base pair deletion at codon 5).

The entire process has been designed for complete automation including assay design, PCR setup, post-PCR treatment, nanoliter transfer of diagnostic products onto silicon chips, serial reading of chip positions in the mass spectrometer, and final analytical interpretation. The MassARRAY™ system enables genetic analysis on an industrial scale, resulting in a streamlined process that has numerous advantages over many other genotyping technologies. First, chip-based mass spectrometry (MS) allows the separation and detection of a mixture of biomolecules within seconds without the need for labels or internal standards. The resultant mass signals provide absolute information on an inherent molecular property (the molecular weight), thus differentiating MS from other detection methods that are indirect. Second, the test strategy (MassEXTEND™) is designed to reveal only the few 'bits' of information relevant to the particular assay being performed. Third, the use of miniaturized chip-based sample preparation enables automated scanning of numerous samples on one array in the mass spectrometer. This combination of high-throughput with high definition signals is a strong improvement over other SNP genotyping technologies, drawbacks for which may include prohibitive speed and cost (gel-based methods), requirement for specific temperatures or reagents for specific hybridization, need for expensive

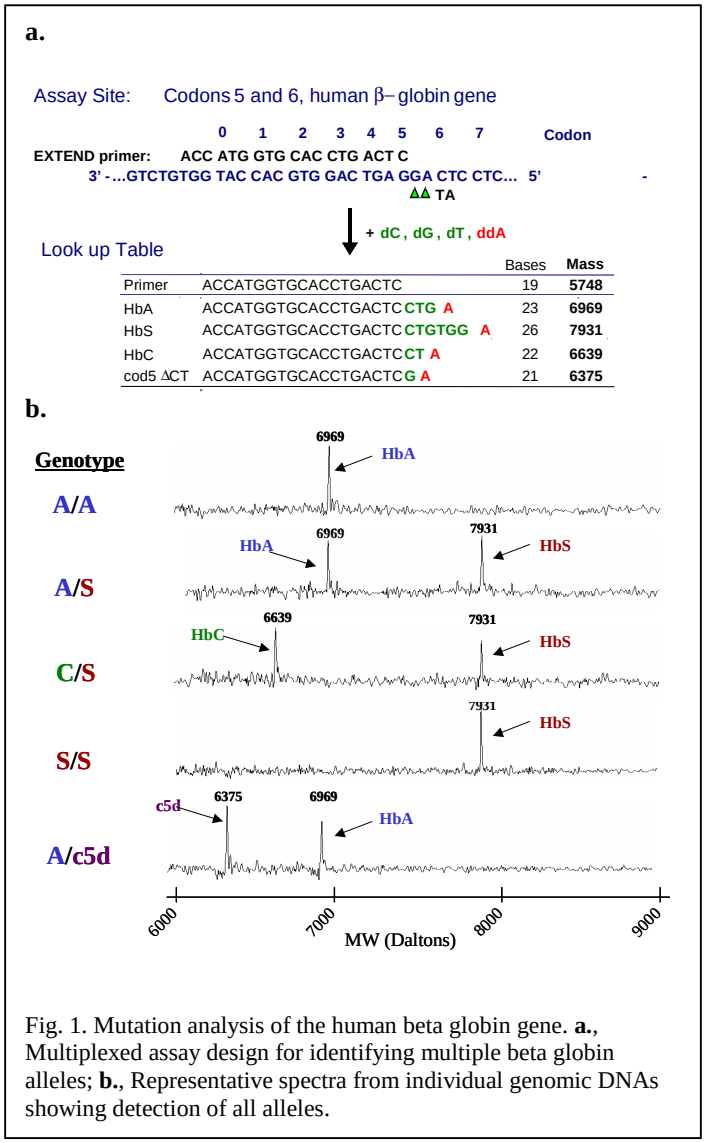

Fig. 1. Mutation analysis of the human beta globin gene. **a.**, Multiplexed assay design for identifying multiple beta globin alleles; **b.**, Representative spectra from individual genomic DNAs showing detection of all alleles.

labels or dyes which provide only indirect measurements of diagnostics DNAs, problems with accuracy, and low flexibility for multiplexing or for incorporating new SNPs.

2.2 A DNA Pooling strategy for large-scale association testing

The analysis of allele frequency distributions is a tool to study the abundance of particular alleles in sample sets and to compare these between different collections. Association testing may be regarded as a comparison of allele frequencies between case and control individuals, in which a statistically significant difference may implicate a locus in a disease etiology. The optimal approach to conducting an association study would require that we obtain the genotype of all common genetic variations in all individuals included in the study. Two key limitations prevent the implementation of such an optimal approach given the current state of the science. First, only half of the approximately ten million common SNPs (minor allele frequency > 10%) that would be most useful for association under the common disease/common variant hypothesis might currently be known[11-13]. Second, the cost per SNP of accurate individual genotyping remains prohibitively high to make the genotyping of more than a few thousand SNPs feasible on a sample set of sufficient size to offer reasonable statistical power. While we can expect that both of these limitations will be overcome at some point, current alternative approaches exist that make conducting large-scale association studies possible using available technologies.

For studies that involve comparisons of allele frequencies between selected groups of individuals, DNA pooling has been proposed to significantly reduce the number of reactions and the cost required to perform large-scale SNP association studies[14-17]. We have developed a typing technology for estimating allele frequencies from pooled DNAs from large numbers of individuals. When genotyping assays are applied to DNA samples combined in equimolar amounts, the quantitative nature of mass spectrometry allows us to generate mass peaks whose areas are directly proportional to the relative allele frequencies in the sample. This ability to use DNA pools means that, as opposed to genotyping individual DNAs, SNP assays can be tested far more efficiently with associated reductions in the amount of time, financial resources, and genomic DNA required. Recently several papers have been published that independently describe this approach to estimate allele frequencies in sample pools and its application in case-control style association studies[18-23].

With the promise of high-throughput testing of all genes in the human genome for disease or trait associations, we attempted to validate the MassARRAY for detecting statistically significant associations through comparison of allele frequency estimates from pooled DNAs. An early study involved testing 15 confirmed polymorphic SNPs in the cholesteryl ester transfer protein (CETP) gene for association with serum high density lipoprotein cholesterol (HDL-c) levels[24]. CETP is known to play an essential role in cholesterol metabolism, mediating exchange of cholesteryl ester in HDL-c for triglyceride in LDLs or VLDLs. Previously several SNPs in this gene had been reported to be associated with HDL-c levels[25-27], making this gene an ideal candidate for validating the pooling approach.

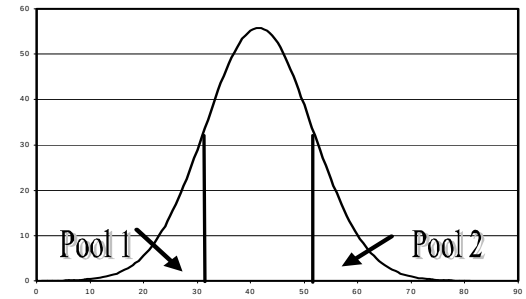

Fig. 2. Pooling strategy based on selecting individuals representing the low and high extremes of HDL-c levels.

DNAs from a large cohort of generally healthy individuals evaluated for HDL-c were carefully quantitated and combined in equimolar amounts within two pools of approximately 400 individuals each, representing the highest and lowest extremes in a normal distribution of HDL-c levels (illustrated in Figure 2). MassARRAY was then used to estimate allele frequencies for the 15 CETP SNPs within the DNA pools. Mass spectra were processed by commercially available software (SpectroTYPER™) using baseline correction, peak identification, and peak area calculation algorithms. Normalized peak areas were computed as individual peak areas divided by the sum of total peak area. Based on the pool data, 14 of 15 SNPs exhibited significantly different frequencies between the low and high HDL-c groups at the 5% level (example in Figure 3, showing variable A and G allele frequencies between pools). Nine of these suggestive associations were subsequently confirmed ($p<0.05$) by genotyping the individual DNAs comprising the pools. Associated SNPs were found to be distributed across the length of the gene including a promoter SNP (rs1800775), an intronic SNP (rs708272), and a nonsynonymous I405V SNP (rs5882), all of which were

previously reported to be associated with HDL-c levels. Follow-up genotyping of selected SNPs offered two major benefits: confirmation of frequency differences observed between DNA pools, and reconstruction of haplotypes showing LD with the trait. Compared to individual SNPs, certain haplotypes derived from subsets of associated SNPs identified by pooling showed even stronger evidence for association with HDL-c levels.

These results clearly demonstrate that SNP typing using DNA pools can identify known associations, and confirm the idea that large-scale association studies could be accelerated by this approach. They provide a proof of concept of the use of pooling techniques and associated technologies for efficient initial screening of SNPs, and prioritizing them for subsequent analysis involving genotyping.

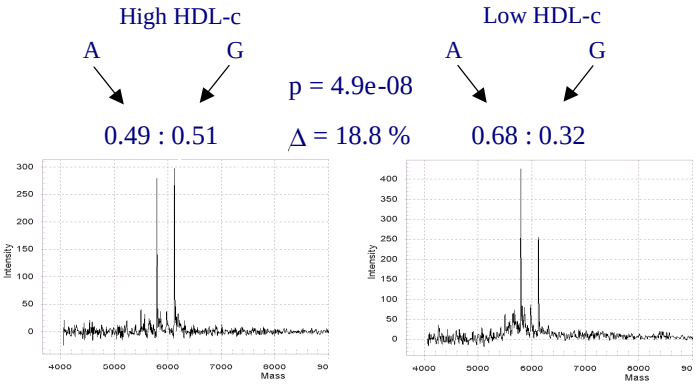

Fig. 3. Example spectra for the CETP intron 1 SNP rs708272 obtained from DNA pools (high vs. low HDL-c groups), showing significant frequency differences between the A and G alleles. Individual genotyping confirmed an allelic difference of 18% between the groups.

That SNP frequency estimates from a DNA pool can be obtained in several seconds provides support for the feasibility of investigating thousands of loci in a high-throughput manner. Further, the pooling approach offers the potential to reduce not only the number of SNPs that undergo follow-up individual genotyping, but also the quantity of each individual DNA used. In the following sections we describe the development of a strategy for expanding this approach to include virtually all human genes, beginning with the selection of informative SNPs throughout the genome and collection of suitable sample sets. We then describe in detail the results from the successful implementation of this strategy in several case studies.

3. APPROACH TO GENOME-WIDE ASSOCIATION TESTING

3.1 SNP selection

DNA sequence variations among individuals are an important tool for identifying genetic contributions to disease, and the selection of the most informative variations that offer the greatest opportunity for detecting association is critical to any strategy. Single nucleotide polymorphisms (SNPs), with their dense distribution across the genome, are potentially excellent markers for association studies. The power to successfully demonstrate association in a study of unrelated cases and controls is primarily dependent upon: 1) the size of the genetic effect at the biologically "causative" allele, 2) the sample size, and 3) the association between the causative allele and the marker allele (linkage disequilibrium, LD). An ideal marker panel for association mapping would contain one marker in complete LD (r^2 = 1) with a causative variant at each causative position in the human genome. As the LD between the marker and a causative allele decreases, the measured genetic effect at the marker decreases exponentially[16,28], as does the power to detect that effect using the marker for a given sample size. The International Haplotype Mapping (HapMap) Project was recently established to identify such a set of SNPs[29]. At present the HapMap project is likely some years from collecting sufficient data and understanding the results to produce an optimal mapping panel. In the absence of this information, we developed an assay panel targeting SNPs within regions of the genome containing known or predicted genes, that is, the portion of the genome where variations are most likely to influence cellular function and thus disease risk.

Publicly available candidate SNPs, mainly from the NCBI database dbSNP, represent a good resource for attempting to validate polymorphism in human populations, especially if judiciously selected based on available annotation[13]. We have created working assays for more than 200,000 putative SNPs annotated as residing within ten kilobases (Kb) of transcribed gene sequences. We assessed their polymorphism status by testing assays on pooled DNA samples from 92 CEPH Caucasian individuals[30] using the MassARRAY platform[24,31,32]. Of the public SNPs tested, the large majority of NCBI reference SNPs that were already annotated as "validated", or which were

submitted by multiple independent groups, were also found by us to be polymorphic. More than 100,000 SNPs mapping uniquely onto the genome were found to be polymorphic with minor allele frequencies greater than 3%. These confirmed polymorphic SNPs cover the genome with a density that reflects the distribution of genes.

To create sets of SNP markers for genome-wide association studies, we developed an algorithm to select subsets of these confirmed polymorphic SNPs with more regular spacing and satisfying other criteria, such as proximity to genes, location in coding versus non-coding regions, and higher minor allele frequency. As illustrated in Figure 4 , the algorithm sets up regularly spaced posts at a specified density across the genome, identifies the SNPs within each post neighborhood, ranks them by the above criteria and by their distance to the post, and picks the top-ranking SNP in each neighborhood. Using this algorithm we initially selected sets of SNPs located within 10 Kb of more than 65% of known and predicted genes. These sets were assembled without knowledge of LD in the regions containing the markers. Therefore some regions were expected to contain more markers than necessary to assure strong LD with possible common causative variations. Conversely some regions might contain too few markers, limiting our power to detect causative variations by association. Nevertheless, given that there are likely to be multiple genes possessing alleles of modest but measurable effects on common disease susceptibility, we anticipated the proposed strategy had sufficient power to identify many of these genes. These initial sets, ranging between 25,000 and 85,000 SNPs, have already been used successfully in more than a dozen genome-wide scans for association with common diseases and quantitative traits, including breast cancer, hypertension, and HDL-cholesterol levels.

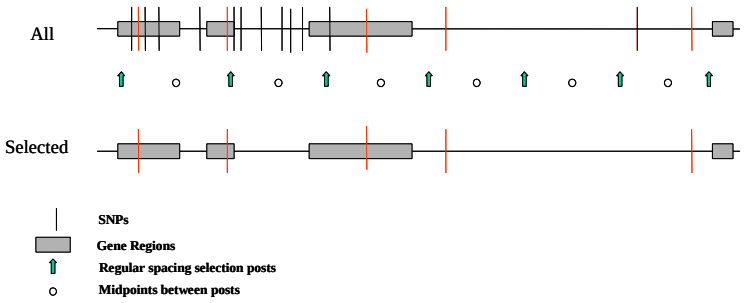

Fig. 4. Selection strategy for gene-based SNPs based on regularly spaced marker intervals.

As more public SNPs and annotation continue to become available, we are using this algorithm to update and expand our gene-based SNP sets to increase gene coverage, the goal being one SNP approximately once per 10 Kb within all known and predicted human genes. For SNPs not yet validated, our algorithm selects SNPs with features we have found to be predictive of true polymorphisms. As public haplotype information becomes available along with the tools for analyzing haplotypes[33], that information is also being used to select the more informative SNPs for genome-wide association tests. With a recent expansion of our SNP panel targeting gene regions not represented in earlier sets, our current panel contains approximately 110,000 assays for gene-based SNPs with minor allele frequencies of at least 3%. Figure 5 shows the distribution of all SNPs with regard to marker spacing. These SNPs have median and mean inter-marker distances of 10.4 Kb and 26.3 Kb, respectively, within gene-containing regions. Ninety-eight percent (98%) of the currently estimated 22,287 human genes reported in the Ensembl database[34] are represented by these SNPs; the coverage increases to 99.5% for genes mapped to assembled chromosomes (genome build 34). SNPs in the remaining genes were either unavailable from public SNP sources, or failed to meet minimum selection criteria as described above.

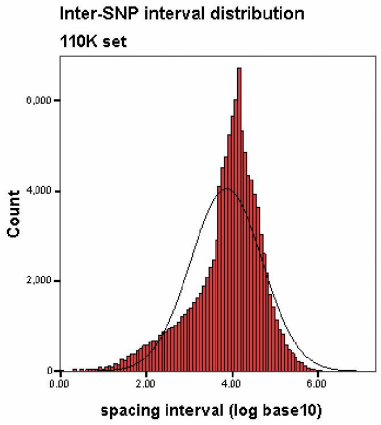

Fig. 5. Distribution of numbers of selected SNPs based on inter-marker spacing.

3.2 Collection of suitable sample sets

Our sample resource for genome-wide association studies consists of more than 50,000 samples with extensive clinical databases. This repository contains several collections for the analysis of a variety of diseases and traits. Among these, there are 6,600 twin samples from England and Australia, for which we have clinical data regarding several diseases as well as a variety of biochemical and biophysical measurements. These samples were used to conduct genome-wide association studies for hypertension, central obesity, HDL-cholesterol levels, osteoarthritis, and osteoporosis. In addition, we collected approximately 11,000 samples from healthy donors from blood donations centers in California and used those age-stratified for identifying morbidity-related genes.

For disease-specific association studies we established an international network with clinicians and geneticists. For instance, the type 2 diabetes collections consist of about 2,900 samples from Germany, Denmark, and Newfoundland. Additional cohorts from Newfoundland comprise about 3,300 individuals with obesity, osteoarthritis, and inflammatory bowel disease. The cancer sample repository contains about 7,200 cases and controls for breast, prostate, lung, and skin cancer from Germany, Italy, England, Australia, and the United States. Clinical information including family cancer history, lymph node and organ metastases, tumor grading, and method of treatment is available for many of the samples. Other large sample collections are useful for the identification of susceptibility genes causing neuropsychiatric disorders like schizophrenia or Alzheimer disease.

3.3 High-throughput strategy using DNA pools

These SNPs are now being routinely employed in a multi-stage strategy that relies on estimating allele frequencies from pools of individual DNA samples in a high-throughput setting to rapidly and cost effectively screen large numbers of SNPs to identify those associated with disease. The multi-staged strategy we employ for a simple case-control is illustrated in Figure 6. Typical percentages for SNPs proceeding to the subsequent stage are shown. The first step is to combine equimolar quantities of DNA from each subject within each group, forming one pool for cases and one for controls. Allele frequencies for each marker SNP are estimated within each pool using the primer extension genotyping and chip-based mass spectrometry methods as described above. The allele frequencies are compared between cases and controls using a statistical procedure that accounts for sampling and measurement uncertainty, yielding a p-value for each tested SNP that supports the significance of observed differences.

Based on the first stage association results, a proportion of the most significant SNPs that satisfy pre-specified criteria are taken into the second stage to undergo a repeated pooled analysis. Possible criteria for selecting SNPs from the first stage include marginal significance levels (e.g. $p = 0.05$ or 0.01), or proportions of all SNPs (e.g. the 5% with the smallest p-values). At the second stage the selected SNPs are subjected to the same procedure as in the first stage, but in triplicate. The repeated PCR provides a more precise estimate of the allele frequencies within each pool, as well as an assay-specific estimate of the PCR variability, thereby yielding a more powerful comparison between the two groups. SNPs are selected from the second stage using pre-specified statistical criteria (e.g. $p = 0.01$). Last, the associations of these SNPs are confirmed by genotyping each selected SNP in each individual comprising the pools, thus removing the influence of measurement variability in subsequent hypothesis tests.

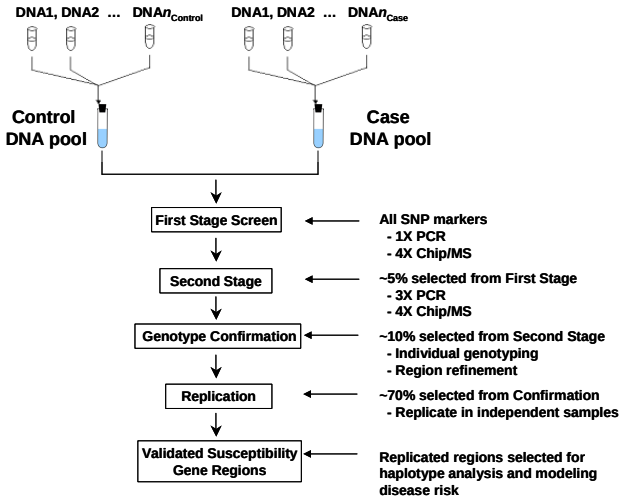

Fig. 6. Strategy for genome-wide SNP association testing. Comparisons between DNA pools are used to filter large numbers of gene-based SNPs for suggestive associations. Reproducibility from repeated pool testing is then used to select SNPs for confirmation by individual genotyping, followed by replication testing in independent sample sets.

This selection procedure results in the identification of many SNPs that are associated with the disease following genotype confirmation. Due to the relatively liberal selection criteria employed at each stage to reduce the number of false negatives, most of the resulting associations are expected to be false positives. False positives can be excluded in either of two ways. One approach applies stringent significance criteria that accounts for the large number of independent tests carried out throughout the process. An alternative approach, commonly applied in settings where a large number of hypotheses must be considered, is to validate the results by testing the selected SNPs in an independent sample from the same population of inference[35]. Genetic effects that are significantly associated with disease in one or more independent collections, after accounting for multiple testing at this stage, are likely in strong linkage disequilibrium with true susceptibility alleles[36-38]. Further work can be done to fine map the replicated regions by genotyping other SNPs in the region and identify the minimal set of individual SNPs and/or haplotypes that account for disease susceptibility in that region, and to guide the search for causative genetic variations.

4. RESULTS FROM GENOME-WIDE ASSOCIATION STUDIES

4.1 Search for additional genes influencing HDL-cholesterol levels

The results from the analysis of the CETP gene encouraged the application of this approach on a much larger scale to identify additional genes that may be involved in regulating HDL-cholesterol levels. Here we describe results of one large-scale association study employing the approach described above, in greater detail and with additional considerations that have come into focus through experience in many such studies.

As for the CETP study, we again created two groups of individuals defined by the extremes of a normal trait distribution[28]. The "discovery collection" included unrelated females selected from a collection of twin pairs[39] with age-, BMI-, and alcohol consumption-adjusted HDL-c values within the upper or lower 19th percentile of the adjusted HDL-c distribution. The final collection included 304 subjects with adjusted HDL-c levels below 1.18 mmol/L and 295 subjects above 1.89 mmol/L. DNA samples from each subject were quantified and similarly combined in equimolar amounts to form one pool for low and one pool for high HDL-c individuals.

A collection of 25,494 public SNP markers was used in this association study, representing a subset of the complete SNP panel described previously. These SNPs provided a roughly even distribution over approximately 46% of known and predicted genes (in 2002), had a median inter-marker spacing of 36 Kb, and were mostly common with minor allele frequencies greater than 10%. Following a single PCR and mass extension reaction for each SNP, allele frequencies within a pool were determined from four mass spectrometric analyses of the extension products. The distributions of the resulting allele frequencies (for the high mass allele) of the combined pools are shown in Figure 7a. The left skew of these distributions are typical of all such distributions of common SNPs estimated from DNA pools using this chip-based MS method, a phenomenom partially due to decreased sensitivity of the mass spectrometer to detect higher mass products[40]. Such skewing has also been observed with most other applicable pooling methods[41]. The distribution of odds ratios and p-values from the test of association between alleles and HDL-c grouping is presented in Figure 7b. The p-value distribution is mildly right skewed, differing only slightly from the expectation of a uniform distribution assuming that the null hypothesis of no association is true for nearly all of the SNPs tested. Tests of association between phenotype and each SNP using pooled DNA were carried out in a similar fashion as described elsewhere (Barratt et al. 2002).

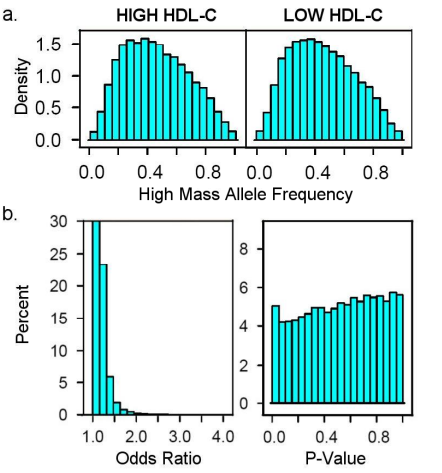

Fig. 7. Results from testing >25,000 SNPs in high- and low-HDL-c pools. **a.**, Distribution of numbers of SNPs based on frequency estimates for the higher mass allele; **b.**, Distributions of odds ratios and p-values from association tests.

1,032 SNPs with the most significant associations with HDL-c high/low status (p-value < 0.03) were selected from the first stage for re-measuring the DNA pools, but in triplicate. Reproducibility of observed pool comparisons in this stage is regarded as critical for selecting SNPs for subsequent analysis. This stage is routinely performed using newly synthesized oligonucleotides to account for any potential variance introduced as a result of oligo quality. With the availability of triplicate PCR spotted onto four chips each, the PCR-to-PCR and chip-to-chip variability were estimated directly, rather than applying a laboratory average as was done for the PCR variance in the first stage. The distributions of allele frequencies for each SNP within each pool are given in Figure 8a. The shapes of these distributions show a reduction in the skew observed the first stage of analysis. The decrease of frequencies around 0.5 is due to the higher sampling variance (and subsequent decreased power) in this range. The odds ratios and p-values of the SNPs typed in the second stage display the enrichment of significant associations compared to the results on the complete set of SNPs in the first stage.

The 100 most significant associations from the second stage (p-value < 0.02) were selected for confirmation by genotyping each subject DNA comprising the high and low HDL-c groups. Individual genotypes permitted sample allele frequencies to be determined with extremely low error, and thus high and low HDL-c groups could be compared taking only sampling variability into account. Tests of association using individual genotypes were carried out using a chi-square test of heterogeneity based on allele and genotype frequencies. P-values were derived using the log odds of each contrast and their standard errors. Over 90% of the SNPs genotyped were significantly different between groups (p = 0.05), ten with p-values less than 0.001 and three less than 0.0001. The allelic odds ratios for the significant SNPs ranged from 1.2 to 2.5.

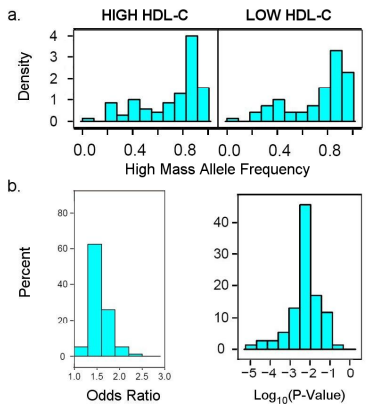

Fig. 8. Results from repeated testing of >1,000 SNPs selected from the initial phase. **a.**, Distribution of numbers of SNPs based on frequency estimates for the higher mass allele; **b.**, Distributions of odds ratios and p-values from association tests.

In order to separate the relatively few expected genes with true and reproducible genetic effects from the expected large number of false positives, approximately 75 of the confirmed SNPs were chosen for genotyping in three independent sample collections of similar ethnic backgrounds for which serum HDL-c values were available. Various analyses of the association between each SNP and the continuous HDL-c variables were carried out as appropriate for each collection. For comparisons among the four collections (including the discovery collection), individuals from the replication collections were selected from within the top and bottom quartiles of HDL-c values and SNPs were compared between the high and low groups using a random effects meta analysis method[42].

Three SNPs were found to be highly statistically significant after adjustments for multiple testing. One SNP was observed to have consistently modest allelic effects but did not reach the level of global significance. The summary of these tests is shown in Table 1 along with a description of the gene located near the marker SNP, the high HDL-c group allele frequency (Freq. – standardized to the allele that increases in the high to low HDL-c group), and the odds ratio (OR) and p-value estimated from the initial discovery sample, the replication samples, and all samples analyzed together. The top three highly significant SNPs are located within or near three of the most recognizable genes known to be involved in HDL-c metabolism: cholesterol ester transfer protein, lipoprotein lipase, and hepatic lipase[43]. Again, variations in CETP are well known to influence serum HDL-c levels, however studies have also found consistent associations with variants in LPL or HL[44]. The fourth SNP (SQHC072) is located within the exon of a gene that has not previously been associated with HDL-c metabolism either genetically or biologically. However, the protein product is vesicle-associated and known to be involved in exocytosis, a process that would be central to cholesterol efflux and reverse cholesterol transport.

Table 1.

			Discovery		Replication		Combined	
Gene	Description	Freq.	OR	P-value	OR	P-value	OR	P-value
CETP	Cholesterol ester transferase protein	0.50	1.7	9.0E-06	1.7	1.7E-09	1.7	6.8E-14
LPL	Lipoprotein lipase	0.86	1.7	2.8E-03	1.7	7.7E-04	1.7	4.7E-06
HL	Hepatic lipase	0.72	1.7	1.4E-04	1.3	8.4E-03	1.4	9.6E-05
SQHC072	Vesicle-associated protein involved in exocytosis; expressed in macrophages	0.34	1.5	4.5E-04	1.2	3.5E-02	1.3	1.8E-04

In summary, beginning with a panel of only 25,000 SNP markers, we identified four SNPs associated with HDL-c levels. The SNPs were located within three genes with well-known involvement in HDL-c metabolism and one gene not previously connected to lipid metabolism that has likely involvement on the basis of the genetics and known biology. Without knowing how many other genetic variations exist having an effect on HDL-c levels as strong as

those identified here, it was not possible to estimate the overall sensitivity of this approach using only 25,000 SNPs. A comprehensive survey of the genome will require testing the current 110,000 gene-based SNP panel.

4.2 Experience with other large-scale association studies

As another example of the successful application of this approach, we recently reported results from a large-scale association test for alleles conferring susceptibility to breast cancer[45]. By comparing the allele frequencies for more than 25,000 SNPs in nearly 16,000 genes, in DNA pools derived from breast cancer cases versus appropriate controls, strong evidence was obtained implicating a small region on chromosome 19p13.2 containing the genes ICAM1, ICAM4, and ICAM5. An association was seen not only with disease susceptibility, but also with certain indicators of disease prognosis. These findings were in agreement with previous reports on the involvement of ICAM1 in tumor progression, although our data could not formally exclude any of the three closely linked ICAM genes. Further, the association was also seen between this region and prostate cancer susceptibility. The identification of a region with variants increasing risk to both breast and prostate cancer is particularly interesting since they have certain common features, such as hormone-sensitivity, parallel incidence rates in various countries, and other common genetic alterations[46].

Over the last two years our research group has carried out more than a dozen genome-wide association studies. The goal of these studies has been to identify genes that influence common disease susceptibility or disease-related traits, including other cancers, type 2 diabetes, schizophrenia, osteoporosis, osteoarthritis, and several cardiovascular disease-related traits. By applying the same strategy, many compelling disease associations and genes have been identified. For example, among the SNPs showing evidence for association with melanoma were at least three in the gene BRAF, recently reported to be somatically mutated in this disease[47,48]. Among the associations confirmed from other studies were genes previously known to be involved in those traits, including: PPARG and HNF3B in type 2 diabetes; AGC1 in osteoarthritis; and DDC in schizophrenia. These initial studies have given us valuable experience and insights into improving the design, execution, and analysis of such scans.

The successful identification of gene regions associated with susceptibility to various diseases or with quantitative traits, using SNP sets representing only a fraction of all human genes, supports the promise of more comprehensive genome-wide studies. Currently we are engaged in our largest associations studies to date, employing our 110,000 SNPs representing at least 98% of all known and predicted human genes. One study aims to identify alleles conferring susceptibility to lung cancer. In a somewhat different application of the case-control approach, another study aims to find alleles associated with severity of beta^0-thalassemia/HbE disease, which is especially highly prevalent in southeast Asia. Patients with this disease carry one null allele of the beta-globin gene (HBB), and one HbE allele encoding a Glu>Lys substitution in codon 26 of the beta globin peptide. A secondary consequence of the HbE mutation is the potential for alternative splicing. The puzzling feature of this disease is the extreme variation in disease severity observed among these patients, suggesting the existence of additional genetic modifiers. To carry out this association study, from a much larger collection have been selected patients presenting with either the mildest or most severe forms of this disease. DNAs from approximately 200 patients in each group were used to construct pools which are being used in high-throughput SNP typing. While underscoring the flexibility for pooling subjects based on any phenotype that can be compared between groups, this study offers a unique opportunity to evaluate the approach for its sensitivity to detect contributions from modifier genes.

5. CONCLUSIONS

The high-throughput approach presented here using pooled DNAs and genome-wide SNPs to filter the entire gene content of the human genome is useful for discovering candidate susceptibility genes for various common diseases, or genes underlying other complex traits. Given the complexity of the genetic architecture underlying trait variation, such genetic analyses alone are not likely to unambiguously identify the genes or genetic variations responsible. However they can quickly and cost-effectively identify the most likely regions for further validation in independent samples to support genetic evidence, and subsequently for functional experiments to identify which variations in which genetic and environmental contexts are truly influencing disease susceptibility and prognosis. The observed association of genetic variants to disease susceptibility or outcome has promising implications for patient management. It is expected that these association studies will provide novel insights into the biology of many common diseases, as well as create new opportunities for diagnostics and therapeutic intervention.

ACKNOWLEDGMENTS

We would like to thank Christian Jurinke and Kai Tang for helpful contributions. For the beta-thalassemia study we acknowledge support from NIH grant DK61883.

REFERENCES

1. Botstein D and Risch N, "Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease." Nature Genet, **33** Suppl, 228-237, 2003.
2. Risch NJ, "Searching for genetic determinants in the new millennium." Nature, **405**, 847-856, 2000.
3. Risch N and Merikangas K, "The future of genetic studies of complex human diseases." Science, **273**, 1516-1517, 1996.
4. Collins FS, Guyer MS, and Charkravarti A, "Variations on a theme: cataloging human DNA sequence variation." Science, **278** 1580-1581, 1997.
5. Goldstein DB, Ahmadi KR, Weale ME, and Wood NW, "Genome scans and candidate gene approaches in the study of common diseases and variable drug responses." Trends Genet, **19**, 615-622, 2003.
6. Braun A, Little DP, and Koster H, "Detecting CFTR gene mutations by using primer oligo base extension and mass spectrometry." Clinical Chemistry, **43:** 1151-1158, 1997.
7. Freking BA, Murphy SK, Wylie AA, Rhodes SJ, Keele JW, Leymaster KA, Jirtle RL, and Smith TP, "Identification of the single base change causing the callipyge muscle hypertrophy phenotype, the only known example of polar overdominance in mammals." Genome Res **12:** 1496-1506, 2002.
8. Little DP, Braun A, Darnhofer-Demar B, and Koster H, "Identification of apolipoprotein E polymorphisms using temperature cycled primer oligo base extension and mass spectrometry." Eur J Clin Chem Clin Biochem **35:** 545-548, 1997.
9. Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, and Lander ES, "Detecting recent positive selection in the human genome from haplotype structure." Nature **419:** 832-837, 2002.
10. Sklar P, Gabriel SB, McInnis MG, Bennett P, Lim YM, Tsan G, Schaffner S, Kirov G, Jones I, Owen M, Craddock N, DePaulo JR, and Lander ES, "Family-based association study of 76 candidate genes in bipolar disorder. BDNF is a potential risk locus. Brain-derived neutrophic factor." Mol Psychiatry **7,** 579-593, 2002.
11. Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L, and Nickerson DA, "Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans." Nature Genet, **33,** 518-521, 2003.
12. Kruglyak L and Nickerson DA, "Variation is the spice of life." Nature Genet, **27,** 234-236, 2001.
13. Reich DE, Gabriel SB, and Altshuler D, "Quality and completeness of SNP databases." Nature Genet, **33,** 457-458, 2003.
14. Barcellos LF, Klitz W, Field LL, Tobias R, Bowcock AM, Wilson R, Nelson MP, Nagatomi J, and Thomson G, "Association mapping of disease loci, by use of a pooled DNA genomic screen." Am J Hum Genet **61,** 734-747, 1997.
15. Carmi R, Rokhlina T, Kwitek-Black AE, Elbedour K, Nishimura D, Stone EM, and Sheffield VC, "Use of a DNA pooling strategy to identify a human obesity syndrome locus on chromosome 15." Hum Mol Genet**, 4,** 9-13, 1995.
16. Risch N and Teng J, "The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. I. DNA pooling." Genome Res**, 8,** 1273-1288, 1998.
17. Shaw SH, Carrasquillo MM, Kashuk C, Puffenberger EG, and Chakravarti A, 1998 "Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes." Genome Res**, 8,** 111-123, 1998.
18. Barratt BJ, Payne F, Rance HE, Nutland S, Todd JA, and Clayton DG, "Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design." Ann Hum Genet**, 66,** 393-405, 2002.
19. Herbon N, Werner M, Braig C, Gohlke H, Dutsch G, Illig T, Altmuller J, Hampe J, Lantermann A, Schreiber S, Bonifacio E, Ziegler A, Schwab A, Wildenauer D, van den Boom D, Braun A, Knapp M, Reitmeir P, and Wjst M,

"High-resolution SNP scan of chromosome 6p21 in pooled samples from patients with complex diseases." Genomics**, 81,** 510-518, 2003.
20. Le Hellard S, Ballereau SJ, Visscher PM, Torrance HS, Pinson J, Morris SW, Thomson ML, Semple CA, Muir WJ, Blackwood DH, Porteous DJ, and Evans KL, "SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage and analysis." Nucleic Acids Res**, 30,** e74, 2002.
21. Mohlke KL, Erdos MR, Scott LJ, Fingerlin TE, Jackson AU, Silander K, Hollstein P, Boehnke M, and Collins FS, "High-throughput screening for evidence of association by using mass spectrometry genotyping on DNA pools." Proc Natl Acad Sci USA, **99,** 16928-16933, 2002.
22. Visscher PM and Hellard SL, "Simple method to analyze SNP-based association studies using DNA pools." Genet Epidemiol**, 24,** 291-296, 2003.
23. Werner M, Sych M, Herbon N, Illig T, Konig IR, and Wjst M, "Large-scale determination of SNP allele frequencies in DNA pools using MALDI-TOF mass spectrometry." Hum Mutat**, 20,** 57-64, 2002.
24. Bansal A, van den Boom D, Kammerer S, Honisch C, Adam G, Cantor CR, Kleyn P, and Braun A, "Association testing by DNA pooling: an effective initial screen." Proc Natl Acad Sci USA, **99,** 16871-16874, 2002.
25. Freeman DJ, Packard CJ, Shepherd J, Gaffney D, "Polymorphisms in the gene coding for cholesteryl ester transfer protein are related to plasma high-density lipoprotein cholesterol and transfer protein activity." Clin Sci (Lond), **79**, 575-581, 1990.
26. Corbex M, Poirier O, Fumeron F, Betoulle D, Evans A, Ruidavets JB, Arveiler D, Luc G, Tiret L, Cambien F, "Extensive association analysis between the CETP gene and coronary heart disease phenotypes reveals several putative functional polymorphisms and gene-environment interaction." Genet Epidemiol, **19**, 64-80, 2000.
27. Dachet C, Poirier O, Cambien F, Chapman J, Rouis M, "New functional promoter polymorphism, CETP/-629, in cholesteryl ester transfer protein (CETP) gene related to CETP mass and high density lipoprotein cholesterol levels: role of Sp1/Sp3 in transcriptional regulation." Arterioscler Thromb Vasc Biol, **20**, 507-515, 2000.
28. Schork NJ, Nath SK, Fallin D, and Chakravarti A, "Linkage disequilibrium analysis of biallelic DNA markers, human quantitative trait loci, and threshold-defined case and control subjects." Am J Hum Genet, **67,** 1208-1218, 2000.
29. The International HapMap Consortium, "The International HapMap Project." Nature **426**, 789-796, 2003.
30. Dausset J, Cann H, Cohen D, Lathrop M, Lalouel JM, and White R, "Centre d'etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome." Genomics, **6**, 575-7, 1990.
31. Buetow KH, Edmonson M, MacDonald R, Clifford R, Yip P, Kelley J, Little DP, Strausberg R, Koester H, Cantor CR, and Braun A, "High-throughput development and characterization of agenomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrixassistedlaser desorption/ionization time-of-flight mass spectrometry." Proc Natl Acad Sci USA, **98**, 581-4, 2001.
32. Nelson MR, Marnellos, G, Kammerer S, Hoyal CR, Shi MM, Cantor CR, and Braun A, "Large-scale validation of single nucleotide polymorphisms in gene regions." Genome Res, **14**, 1664-8, 2004.
33. Barrett JC, Fry B, Maller J, Daly MJ, "Haploview: analysis and visualization of LD and haplotype maps." Bioinformatics, [Epub ahead of print, PubMed ID: 15297300], 2004.
34. International Human Genome Sequencing Consortium, "Finishing the euchromatic sequence of the human genome." Nature **431**, 931-45, 2004.
35. Ripley BD, *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, 1996.
36. Hirschhorn JN, Lohmueller K, Byrne E, and Hirschhorn K, "A comprehensive review of genetic association studies." Genet Med**, 4,** 45-61, 2002.
37. Ioannidis JP, Ntzani EE, Trikalinos TA, and Contopoulos-Ioannidis DG, "Replication validity of genetic association studies." Nat Genet**, 29,** 306-309, 2001.
38. Lohmueller KE, Pearce CL, Pike M, Lander ES, and Hirschhorn JN, "Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease." NatureGenet, **33,** 177-182, 2003.
39. Andrew T, Hart DJ, Snieder H, de Lange M, Spector TD, and MacGregor AJ, "Are twins and singletons comparable? A study of disease-related and lifestyle characteristics in adult women." Twin Res, **4,** 464-477, 2001.
40. Jurinke C, Oeth P, and van den Boom D, "MALDI-TOF mass spectrometry: a versatile tool for high-performance DNA analysis." *Mol Biotechnol* **26**, 147-164, 2004.
41. Sham P, Bader JS, Craig I, O'Donovan M, and Owen M, "DNA Pooling: a tool for large-scale association studies." Nat Rev Genet, **3,** 862-871, 2002.
42. DerSimonian R and Laird N, "Meta-analysis in clinical trials." Control Clin Trials**, 7,** 177-188, 1986.

43. von Eckardstein A, Nofer JR, and Assmann G, “High density lipoproteins and arteriosclerosis: role of cholesterol efflux and reverse cholesterol transport.” Arterioscler Thromb Vasc Biol**, 21,** 13-27, 2001.
44. Knoblauch H, Bauerfeind A, Krahenbuhl C, Daury A, Rohde K, Bejanin S, Essioux L, Schuster H, Luft FC, and Reich JG, “Common haplotypes in five genes influence genetic variance of LDL and HDL cholesterol in the general population.” Hum Mol Genet**, 11,** 1477-1485, 2002.
45. Kammerer S, Roth RB, Reneland R, Marnellos G, Hoyal CR, Markward NJ, Ebner F, Kiechle M, Schwarz-Boeger U, Griffiths LR, Ulbrich C, Chrobok K, Forster G, Praetorius GM, Meyer P, Rehbock J, Cantor CR, Nelson MR, and Braun A, “Large-scale association study identifies ICAM gene region as breast and prostate cancer susceptibility locus.” Cancer Res, **64**, 8906-8910, 2004.
46. Lopez-Otin C and Diamandis EP, “Breast and prostate cancer: an analysis of common epidemiological, genetic, and biochemical features.” Endocr Rev, **19**, 365-396, 1998.
47. Davies H, Bignell GR, Cox C et al., “Mutations of the BRAF gene in human cancer.” Nature, **417**, 949-954, 2002.
48. Pollock PM and Meltzer PS, “Lucky draw in the gene raffle.” Nature**, 417,** 906-907, 2002.

*abraun@sequenom.com; phone 1 858 202-9018; fax 1 858 202-9020; www.sequenom.com