ARTICLE

Identification of the Genetic Basis for Complex Disorders by Use of Pooling-Based Genomewide Single-Nucleotide–Polymorphism Association Studies

John V. Pearson,* Matthew J. Huentelman,* Rebecca F. Halperin, Waibhav D. Tembe, Stacey Melquist, Nils Homer, Marcel Brun, Szabolcs Szelinger, Keith D. Coon, Victoria L. Zismann, Jennifer A. Webster, Thomas Beach, Sigrid B. Sando, Jan O. Aasly, Reinhard Heun, Frank Jessen, Heike Kölsch, Magdalini Tsolaki, Makrina Daniilidou, Eric M. Reiman, Andreas Papassotiropoulos, Michael L. Hutton, Dietrich A. Stephan, and David W. Craig

We report the development and validation of experimental methods, study designs, and analysis software for pooling-based genomewide association (GWA) studies that use high-throughput single-nucleotide–polymorphism (SNP) genotyping microarrays. We first describe a theoretical framework for establishing the effectiveness of pooling genomic DNA as a low-cost alternative to individually genotyping thousands of samples on high-density SNP microarrays. Next, we describe software called "GenePool," which directly analyzes SNP microarray probe intensity data and ranks SNPs by increased likelihood of being genetically associated with a trait or disorder. Finally, we apply these methods to experimental case-control data and demonstrate successful identification of published genetic susceptibility loci for a rare monogenic disease (sudden infant death with dysgenesis of the testes syndrome), a rare complex disease (progressive supranuclear palsy), and a common complex disease (Alzheimer disease) across multiple SNP genotyping platforms. On the basis of these theoretical calculations and their experimental validation, our results suggest that pooling-based GWA studies are a logical first step for determining whether major genetic associations exist in diseases with high heritability.

Genomewide association (GWA) studies that use hundreds of thousands of SNPs have the potential to revolutionize our ability to identify the genetic influences of complex traits and diseases. Although potentially allowing for the identification of common variants to complex disease, GWA studies often require millions of dollars to complete and, as such, are beyond the reach of many research groups. Despite their inherent high costs, these studies will remain one of the best ways to study the genetic basis of complex diseases in a hypothesis-free study design. GWA studies are typically designed with three phases: (I) individual genotyping of ⩾250,000 SNPs across hundreds to thousands of individuals, (II) validation of the most significant SNPs (typically tens to thousands of SNPs) by individual genotyping in new cohorts, and (III) fine-mapping SNPs adjacent to the validated SNPs (generally only a few regions) and/or validation in additional cohorts. One possible approach to reducing the overall cost of GWA studies is to replace individual genotyping in phase I with genotyping (or allelotyping) of pooled genomic DNA.

Several previous reports have investigated the feasibility of pooling on SNP genotyping microarrays (or related technologies). With a few exceptions, these reports have focused on predicting allelic frequencies across thousands of SNPs rather than on the effectiveness of pooling in identifying the genetic basis of complex disorders.[1–21] Indeed, it is not yet clear whether predicting allelic frequency to within 2% accuracy (as is frequently reported) is sufficient, when ⩾250,000 SNPs have incremental allelic frequency differences that vary only between 0% and perhaps a maximum of 10%–15%. Simply, the imprecision of ⩾250,000 pooled measurements may change a SNP ranked in the top 100 SNPs to a rank that misses a phase II cutoff—for example, to the top 1,000 SNPs. For instance, if the true allelic frequency difference between cases and controls is 11.0% for the best SNP (of 250,000 SNPs) and is 10.0% for the 1,000th best SNP, can we identify correctly the genetic loci when our measurement error is on the order of 2%? Simply by chance and with a 2% measurement error, we may predict the *true* best SNP at 9.5% or we may measure any one of several thousand other SNPs falsely at >11.0%. Clearly, multiple testing and imprecise measurements of allelic frequency make it difficult to ac-

From the Translational Genomics Research Institute (J.V.P.; M.J.H.; R.F.H.; W.D.T.; N.H.; M.B.; S.S.; K.D.C.; V.L.Z.; J.A.W.; E.M.R.; A.P.; D.A.S.; D.W.C.) and Arizona Alzheimer's Consortium (E.M.R.), Phoenix; Department of Neuroscience, Mayo College of Medicine, Jacksonville, FL (S.M.; M.L.H.); Sun Health Research Institute, Sun City, AZ (T.B.); Department of Neurology, St. Olav's Hospital (S.B.S.), and Department of Neuroscience, Norwegian University of Science and Technology (J.O.A.), Trondheim, Norway; Department of Psychiatry, University of Bonn, Bonn, Germany (R.H.; F.J.; H.K.); Department of Neurology, Aristotle University, Thessaloniki, Greece (M.T.; M.D.); Banner Alzheimer's Institute and Department of Psychiatry, University of Arizona, Tucson (E.M.R.); and Division of Psychiatry Research and Center for Integrative Human Physiology, University of Zurich, Zurich, Switzerland (A.P.)

Received August 21, 2006; accepted for publication November 7, 2006; electronically published December 6, 2006.

Address for correspondence and reprints: Dr. David Craig, Neurogenomics Division, TGen, The Translational Genomics Research Institute, 445 North Fifth Street, Phoenix, AZ 85004. E-mail: dcraig@tgen.org

* These two authors contributed equally to this work.

Am. J. Hum. Genet. 2007;80:126–139. 0002-9297/2007/8001-0013$15.00

Table 1. Single-Marker Analysis of Pooled GWA Data from Three Disorders with Known Associated Genetic Loci

Analysis	Disorder	Variant	OR	No. of Cases, No. of Controls (Ethnicity)	Pooling Rank (by GenePool)	Approximate No. of SNPs in LD ($r^2 > .5$)	Platform	Arrays per Cohort
1	AD[a]	*ApoE-ε4*	8.3	280, 169 (US)	6/500,568	1	Affymetrix 500K	9
2	AD[a]	*ApoE-ε4*	8.3	280, 169 (US)	18/317,208; 125/317,208	2	Illumina 300K	2
3	AD[b]	*ApoE-ε4*	~3–8	199, 191 (Norwegian); 214, 129 (German); 168, 69 (Greek)	1/500,568; 21/500,568; 34/500,568	1	Affymetrix 500K	9
4	PSP	*MAPT*	3.3	288, 344 (US)	2/500,568 (best); 32 *MAPT* SNPs in top 1,000 SNPs	168	Affymetrix 500K	10
5	PSP	*MAPT*	3.3	288, 344 (US)	1/116,110 (best); 15 *MAPT* SNPs in top 1,000 SNPs	38	Affymetrix 100K	10
6	SIDDT	*TSPYL*[c]	...	3, 100 (Amish)	6/10,555	13	Affymetrix 10K	3

NOTE.—In each case, SNPs in LD with the previously published associated locus were in the top 50 SNPs overall and would have been flagged for validation.

[a] Diagnosed postmortem.

[b] Variously clinically diagnosed.

[c] Data are provided in the work of Puffenberger et al.[30]

curately rank and identify associated SNPs by use of a pooling-based GWA design.

We first investigate the factors influencing pooling-based and individual genotype–based GWA studies. For individual genotyping, there are a number of factors that influence the ability of GWA studies to detect genetic associations. These include but are not limited to: (1) the allele frequency of the causal variant; (2) its odds ratio (OR) or genetic relative risk; (3) the linkage disequilibrium (LD) between the causal variant and probed SNPs; (4) the number of individuals in each cohort; (5) the number of probed SNPs in LD with the causal variant; and (6) the analysis approach taken. Specific to pooling, there are additional factors that influence the ability to detect a true association. These additional factors include: (7) the precision of allele frequency measurements made by the SNP genotyping microarray; (8) the accuracy of pool construction by pippetting; (9) the integrity of the pooled genomic DNA; (10) the number of individuals pooled or overall pooling strategy; and (11) the number of microarray technical replicates. Furthermore, population stratification and admixture can mask true associations in all studies. Beyond these additional factors, by pooling one loses the abilities to compare subphenotypes of pools, to directly measure genotype, and to detect gene-gene interactions.

However, perhaps the most important factor in favor of pooling-based GWA studies is that this study design can be completed for thousands of dollars, whereas individual genotyping may require millions of dollars simply to complete the first phase. There are numerous orphan diseases and many small populations which cannot realistically be studied using individual genotyping at this time, and a pooling-based GWA study is an attractive, cost-effective alternative. Unfortunately, the following questions have not been fully addressed in the context of >300,000 markers: (1) whether a pooling-based GWA can be effective; (2) how one should design a pooling-based GWA study; (3) what is the resolution of the study; and (4) how can one analyze the data.

In this article, we investigate the factors that influence effectiveness of a pooling-based GWA study, develop analysis tools for completing pooling-based GWA studies (GenePool), and establish the practical capability of pooling-based studies to identify the correct genetic locus using actual case-control pooling data with published associated loci. We show that pooling-based GWA studies are a logical first step for studying many diseases with high heritability and that they provide an opportunity to screen for major genetic associations at a substantially lower cost.

Methods

Sample Pooling for Alzheimer Disease and Progressive Supranuclear Palsy

Before quantitation, all DNA samples were checked for quality using 1% agarose gel electrophoresis, and obviously degraded samples were excluded from the pooling analysis. Individual genomic DNA concentrations of each subject were determined in quintriplicate with the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen) according to the manufacturer's instructions. The median concentration was calculated for each individual DNA. Alzheimer disease (AD [MIM 104300]) pools were constructed as four subpools divided by region or population, as shown in table 1. Individual DNA samples were then added to their respective pools in equivalent molar amounts. Each AD subpool was created de novo a total of three times, to control for pippetting errors, whereas each progressive supranuclear palsy (PSP [MIM 601104]) subpool was created five times. Each subpool contained identical samples per cohort to better assess variance. In the "Discussion" section, we describe potential advantages of each subpool containing independent samples. Once created, each pool was diluted to 50 ng/μl with sterile water, in preparation

for the high-density SNP genotype assay. Sample DNA amplified through the use of available whole-genome amplification technologies was avoided because uneven amplification in some samples may substantially reduce power at regions of high amplification.

High-Density SNP Genotyping

Pools were assayed on the Affymetrix 500K platform following the Affymetrix protocol for individual genotyping. Each AD case/control subpool was assayed in three technical replicates, and each PSP subpool was assayed in two technical replicates for the Affymetrix 500K platform. The US AD cohort was assayed in two technical replicates on the Illumina 300K platform by combining replicate subpools for each cohort and by following the protocol for individual genotyping version 1.0 Illumina HumanHap 300K arrays (Illumina). For samples in the US AD cohort, individual-genotype data were available for all samples in the pool on the Affymetrix 500K platform.

For individual genotyping, SNPs were called using two genotyping calling algorithms, SNiPer-HD and BRLMM (Affymetrix). SNiPer-HD uses an expectation-maximization training-based algorithm, and BRLMM uses a modified robust linear model with Mahalanobis distance classifier (RLMM) algorithm.[22] Both algorithms provide superior calls over the standard dynamic modeling approach. However, SNiPer-HD uses only a subset of 380,000 SNPs with highly reliable calls. Only SNPs whose calls agreed in both BRLMM and SNiPer-HD were used for analysis, with ~99.8% of the reduced 380,000 SNP set in agreement. Predicted allelic frequencies were calculated using the *k*-correction method described by Craig et al.[23] Training for *k*-correction values resulted from separate individual-genotype data from ~900 Affymetrix 500K array sets by the same laboratory. Comparing allelic frequencies predicted by pooling and measured by individual genotyping, we experimentally determined that nine Affymetrix 500K arrays measure allelic frequency with an SD of 2.5% with the use of typical DNA. Importantly, that is the measurement error of one cohort, and not the measurement error associated with subtracting the difference between cases and controls. Different reports find different accuracies, which may be largely because of different qualities of starting DNA. In this study, we used typical DNA, which includes samples that may have been stored in a freezer for several months or several years as part of a repository. We expect that, if freshly isolated cell DNA is used, accuracy would be substantially higher and similar to the values reported by other groups. Thus, it is possible that some groups identify "better" results when high-quality starting material is used.

Simulated Pooling

Pooling was simulated in Matlab 7.0 (MathWorks) on the basis of experimental measurements of probe intensities from pooled and individual samples run on Affymetrix 500K GeneChip Mapping arrays. Specifically, we generated paired case-control data sets equivalent to those expected by individual genotyping and if one were to have measured allelic frequencies by pooling. Thus, the pooled data sets are the individual-genotype data sets in which noise consistent with pooling measurement error is introduced.

Simulated data for each case-control cohort was generated independently, under the assumptions of the number of chromosomes pooled (twice the number of individuals), the number of SNPs assayed, the LD between SNPs, and the minor-allele frequency (MAF) for each SNP. In addition to probed SNPs, the "associated causal variant" was simulated in the cases by indirect sampling from a neighboring variant, assumed to be in LD with the associated causal variant with $r^2 = 0.8$ and MAF of 10%. Specifically, individual-genotype data were generated by multiple random sampling of a binomial distribution (binornd function) with the use of 250,000 SNPs and under the assumption of no LD between SNPs. In figure 1*D*, larger SNP sets were generated under the assumption that 500,000, 750,000, and 1,000,000 SNPs were measured with two, three, and four SNPs, respectively, in complete LD.

A control data set and case data set were generated by pooling genotypes under an assumption of Hardy-Weinberg equilibrium. For both case and control data sets, pooled measurement noise was separately added by randomly sampling a normal distribution (normrnd function), under the assumption $\sigma = 2.5\%$ from the individual-genotype data sets. The simulated pooled error was approximately equal to that experimentally observed in nine Affymetrix 500K arrays. Both cases and controls were each treated as single pools, rather than subpools. In each simulation, the rank of the "associated SNP" to the causal variant was recorded for both simulated individual genotyping and simulated pooling. If multiple SNPs were in LD with the associated causal variant (as in fig. 1*D* and 1*E*), we took the best rank of these SNPs. The number of SNPs in LD with an associated causal variant never exceeded four SNPs.

In figure 1*E*, variable LD between all SNPs was added to the simulations. First, all SNPs were assigned an r^2 value to the preceding and following SNP. Values for r^2 were selected from a normal distribution in which 70% of the SNPs exceeded an r^2 of 0.8 and the average r^2 was 0.85. This distribution is roughly equivalent to that of the Affymetrix 500K on CEPH samples. SNP data were then sequentially constructed across 500,000 SNPs. Specifically, genotype data for the first SNP were generated by random sampling from a binomial distribution. Genotype data for all subsequent SNPs were then sequentially generated by adding genotype information from the neighboring SNPs and random sampling from a binomial distribution, according to the defined r^2.

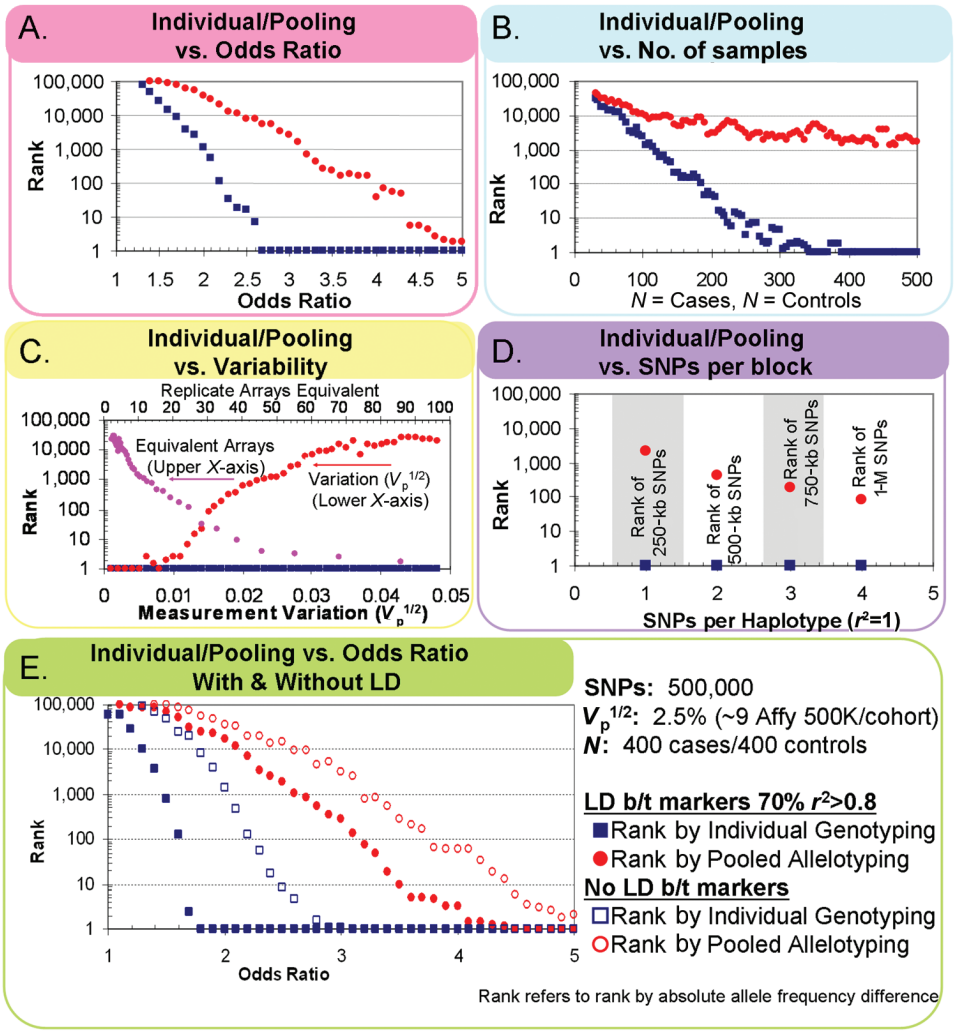

Figure 1. Simulations of the effectiveness of pooling. The expected rank of an associated SNP measured by pooling is compared with its rank ascertained by individual genotyping. *A,* Expected rank of the associated SNP measured by pooling (*red circles*), as the causal variant OR is increased from 1 to 5, compared with the expected rank ascertained by individual genotyping (*blue squares*). With lower rank, associated SNPs are more likely to be resolved. The assumptions are that the LD between the measured SNP and causal variant has an r^2 of 0.8, that inaccuracy due to pooling follows a normal distribution with $V_p^{1/2} = 2.5\%$ (approximately equivalent to nine Affymetrix 500K arrays with average-quality DNA), that 250,000 SNPs are genotyped with an average MAF of 25%, that 400 individuals are in each pooled cohort, and that there is only 1 SNP per haplotype (thus, all SNPs are completely independent). *B,* Same as panel A, but the OR is fixed at 3 and *N* varies from 100 to 400. *C,* Same as panel A, but with the assumption that the probe intensity measurement accuracy varies from 1% to 5% and the OR is fixed at 3. *D,* Same as panel A, but the OR is fixed at 3.0, the number of SNPs in each haplotype block increases from 1 to 4 (SNPs within a block have $r^2 = 1$), and the number of genotyped SNPs is increased from 250,000 to 1,000,000. In this case, the expected rank is the best rank of all SNPs (in the associated haplotype). *E,* Further examination of how LD allows for identification of functional variants with lower OR. Solid squares and circles represent simulation under a scenario in which 70% of the SNPs have a pairwise LD with $r^2 > 0.8$ (similar to Affymetrix [Affy] 500K), whereas outlined squares and circles are the scenario in which there is no LD between (b/t) SNPs. In all cases, these represent data simulated using multiple sampling of a binomial distribution that assumes normal distribution for pooled measurement and Hardy-Weinberg equilibrium. Each data point is the average rank of 75 simulations.

PSP Samples

Cases and controls were from a previously described case-control series, and institutional review board (IRB) approval was obtained for all human subjects.[24] Case patients chosen for pooling had a primary pathological diagnosis of PSP according to standard criteria ($n = 288$).[25,26] The case patients had a mean ($\pm$SD) age at death of 75 $\pm$ 7.6 years and were 51% male. Cognitively normal, age- and sex-matched controls were collected under the Normal and Pathological Aging Protocols at Mayo Clinic Scottsdale ($n = 344$). All patients and controls used in the study were white.[27,28]

AD Samples

Four white case-control cohorts were available for AD: three clinically characterized cohorts and one postmortem clinically and neuropathologically characterized cohort, as summarized in table 1. IRB approval was obtained for all human subjects. Both individual-genotype data and pooled-genotype data were available for the US postmortem cohort. DNA samples were extracted from brain tissue in 398 brain donors, who were at least 65 years of age at the time of their death. The donors included 242 patients who satisfied clinical and neuropathological criteria for the diagnosis of AD and 156 persons who did not meet neuropathological criteria for AD. All the brain donors were white.

For the German cohort, AD patients were recruited from the Department of Psychiatry, University of Bonn. Patients were diagnosed according to DSM-IV, which was supported by clinical examination, detailed structured interviews, neuropsychological testing, cognitive screening done by Mini-Mental State Examination, and neuroimaging studies. Healthy controls were recruited with the support of the local census bureau and the regional Board of Data Protection (Nordrhein-Westphalia, Germany), and diagnosis was done by structured interviews and neuropsychological testing. All patients and control subjects gave informed consent for participation in the study. The study protocol was approved by the Ethics Committee of the Faculty of Medicine at the University of Bonn.

For the Norwegian cohort, patients were recruited from the geriatric and neurological outpatient clinics at St. Olav's Hospital in Trondheim and from local nursing homes, as part of a study of the genetics of dementias in central Norway, as described elsewhere.[29] In brief, guidelines given in the *International Classification of Diseases* (ICD-10) were applied for diagnosing of dementia, with patients who received the diagnosis of AD fulfilling National Institute of Neurological and Communicative Diseases and Stroke/Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) criteria. Controls from the same geographic area were recruited from societies for retired people or were spouses of patients with dementia. All controls had subjective good memory and no first-degree relatives with dementia, and diagnosis was done using a brief interview. For the Greek cohort, patients with AD were recruited from the Department of Neurology, University of Thessaloniki. Patients fulfilled the NINCDS-ADRDA criteria for probable AD after clinical examination and neuropsychological testing. Healthy controls were the patients' spouses and were cognitively intact as assessed by neuropsychological examination.

Sudden Infant Death with Dysgenesis of the Testes Syndrome

For sudden infant death with dysgenesis of the testes syndrome (SIDDT [MIM 608800]), samples were genotyped and pooled as part of a separate study. Previously generated data were used to provide additional validation metrics for analysis procedures.[23,30]

GenePool Software

GenePool is written in C++ (gpextract) and C (gpanalyze). These programs can be run individually using command line Unix. GenePool can be downloaded from the GenePool Web site. The software is currently provided as a precompiled binary for X86-Linux, and as source code. Man pages for all executables are bundled in both source and binary distributions and are also available from the GenePool Web site in PDF and HTML formats for online viewing. The SIDDT10K data set for Affymetrix is provided for download.

Data Transformation

In the Affymetrix platform, each SNP is interrogated by 6–10 probe quartets, where each quartet contains a perfect match (PM) probe for the A allele, a PM probe for the B allele, a mismatch (MM) probe for the A allele, and an MM probe for the B allele. A relative allele score (RAS) is calculated for each quartet. We considered each $RAS_{1..10}$ to be an independent measure of allele frequency, where i refers to a quartet. RAS is equivalent to the ratio of A allele to A and B alleles for PM probes. That is, $RAS_{i=1..10} = PMA_i/(PMA_i + PMB_i)$. In the Illumina platform, each SNP is interrogated by a variable number of beads, with an average of 16 beads per SNP on an Illumina 300K HumanHap BeadChip. Unlike the Affymetrix platform, beads are assumed to have similar hybridization, and RAS is a one-dimensional vector ($i = 1$). For each bead, red and green channel data corresponding to the two SNP alleles are acquired and are stored in 10 text files within a BeadStation-specified data folder. Both channels undergo a simple normalization by dividing the overall mean intensity value for that channel, because of the observation that the green channel has overall greater intensity than that of the red channel. We recognize that future research efforts may lead to development of more-advanced global normalization methods, noting that calculation of RAS values provides for SNP-specific normalization. Lastly, if any single SNP is probed by fewer than five beads, this SNP is discarded because this SNP measurement will have high variability due to under sampling. Typically, fewer

than a few hundred SNPs were discarded because of this filter.

Evaluation of Test Statistics

Multiple test statistics were evaluated for their effectiveness in ranking SNPs. Effective ranking was determined using pools composed of samples that previously had been individually genotyped on the same Affymetrix 500K platform. Ranking SNPs was done by difference in allelic frequency, not by *P* value calculation. The test statistics evaluated were silhouette scores,[31] trace criteria, a multivariate *t* statistic, determinant criteria, principal-components–based linear classifiers, bolstered linear classifiers, a centroid Dunn index, and a Hausdorff Dunn index.[32] Implementation of each test statistic is available on request as supplemental data, although not all methods were incorporated into GenePool. Ranking by silhouette score was consistently found to be the most effective method, as measured by correct identification of the greatest number of top 10, 100, 500, and 1,000 SNPs by individual genotyping with the use of the top 10, 100, 500, and 1,000 SNPs for the test statistic on data from pooled arrays. This method of evaluation was used because it is similar to how a research group might proceed in a pooling-based GWA. Additional metrics were used, including identifying the greatest number of top 0.1%, 1%, and 5% of SNPs, with similar results. We further investigated different distance measures for the silhouette statistic, finding that Manhattan distance generally outperforms Euclidean distance, since the Euclidean distance measure does not preserve directionality.

Comparison of Allele Frequencies by Pooling with Individual-Genotype Data

Pooling accuracy was investigated by calculating the predicted allelic frequency by use of the *k*-correction method. A median difference of 2.8% was observed. Other methods are emerging,[33] but we used the approach applied by Craig et al.,[23] which uses heterozygote and homozygote data building from the original employment of the *k*-correction factor.[6] For training of values, individual-genotype data were used for SNPs for which the 900 training samples showed at least five each of the three possible call states (AA, AB, and BB). Only SNPs with >90% of samples called and arrays with >85% call rates were used.

Multimarker Statistics

Multimarker statistical methods for pooled data are rapidly evolving and are an ongoing area of research.[1,11,34] We have implemented a sliding-window statistic of mean or median rank across the genome for a fixed window size. This method allows prioritization of regions that have several neighboring high-ranking SNPs.[1] As the number of genotyped SNPs on a commercially available platform exceeds 1 million, multimarker statistics will become an increasingly important noise-reduction approach. However, accomplishing this goal will require yet-to-be-developed methods for merging data from separate platforms, such as use of the *k*-correction factor.

Results

Practical Challenges of Identifying Associations by Use of a Pooling-Based GWA Study Design

One of the least appreciated aspects of pooling-based GWA studies is that allelic differences between cohorts must be measured and sorted for hundreds of thousands of SNPs. We examine this aspect by anecdotal example, then by theory, and again by simulation. After these sections, we describe development of software and analysis tools for conducting of pooling-based GWA studies. Finally, we report experimental validation of these methods on complex disorders with published associations.

Example

In an anecdotal example, we pool 400 individuals as part of a case cohort and 400 individuals as part of control cohort. We then measure 250,000 SNPs for differences in allelic frequencies between cohorts. By multiple sampling of 250,000 SNPs for 400 cases and 400 controls, we can reasonably expect that at least one SNP will have a 15% difference between cohorts. The second best SNP may have a difference of 14.9%, the third may have a difference of 14.8%, and so forth. However, the accuracy of our predicted differences (by use of measurements of pooled data) is typically only within 1%–3% of the true allelic frequency.[1,35] Consequently, if our predicted allelic frequency is off by 2%, the predicted rank may move from 1st of 250,000 to 800th of 250,000, or vice versa. Since there may be only a few truly associated SNPs and 250,000 SNPs are being measured, false-positive and false-negative results due to measurement error are a major concern. Realistically, an inaccuracy of 1%–3% could lower the rank of a truly associated SNP below the threshhold for inclusion in phase II of a GWA study.

Theory

We now review theoretical considerations related to controlling the variance and to the design of case-control pooling-based studies. We note that more-extensive reviews of DNA pooling theory are available elsewhere and that we primarily develop concepts directly relating to commercially available high-density SNP genotyping microarrays.[11,21,36,37]

Controlling measurement variance is essential to identification of true associations in the context of multiple sampling of ⩾250,000 SNPs. In its most simplistic form, the total variance (V_T) for an allelotyped SNP in a cohort of pooled individuals is the sum of variance arising from sampling a limited number of individuals (the sampling variance [V_s]) and the experimental variance observed

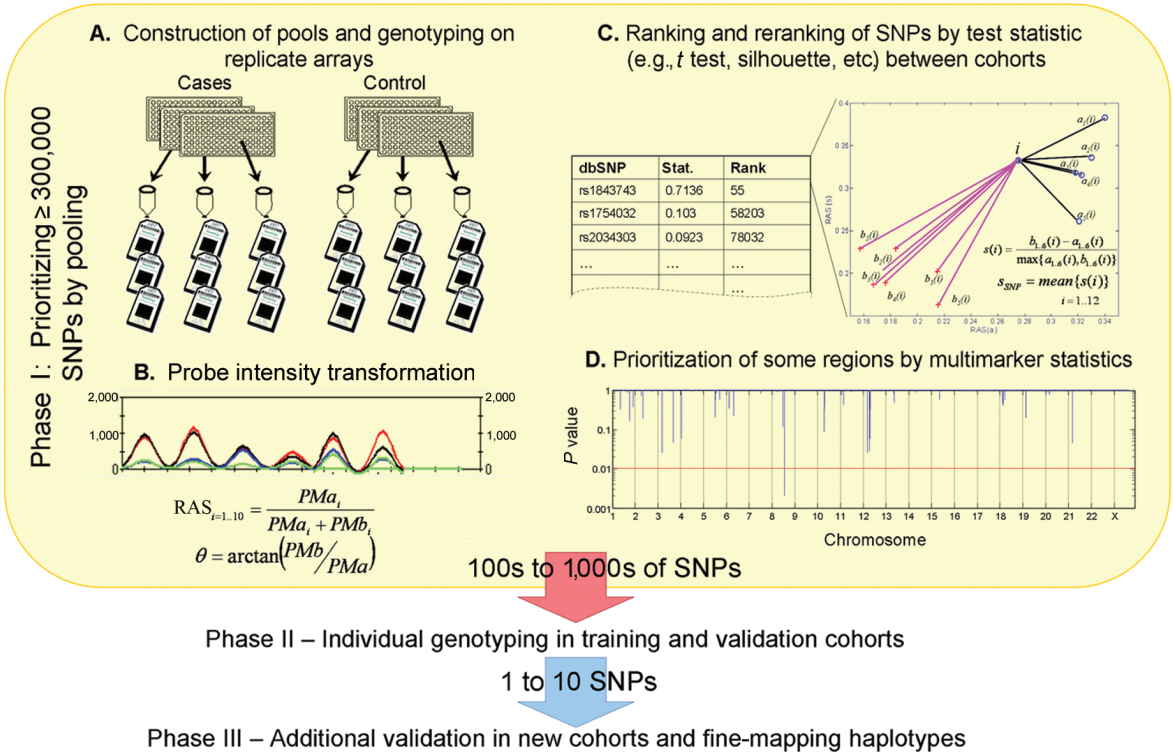

Figure 2. Stages of phase I of a pooling-based GWA. *A,* In stage 1, DNA is quality checked, quantified, and placed into subpools. The number of technical replicates and subpools will be study dependent, although theoretical considerations are evaluated in the discussion. *B,* In stage 2, paired probe-intensity data from both alleles is transformed and/or normalized into a singular value, generally correlating with allelic frequency. RAS values are one type of normalization, generally referring to the proportion of overall signal arising from one allele. *C,* In stage 3, each set of transformed-value scores are evaluated and are ranked by a test statistic (Stat.) to provide a measure of association between cohorts. *D,* In stage 4, multimarker statistics are calculated, leveraging LD between SNPs to reduce noise or to incorporate external information, such as haplotype data.[8,23,34]

from allelotyping the pooled DNA (the pooling variance [V_p]):

$$V_T = V_s + V_p \ . \qquad (1)$$

Because of differences in allelic frequency and assay performance, a wide spectrum of values for V_T is observed among the thousands of SNPs on a microarray.[38] Reducing V_T increases power and can be accomplished by reducing either V_p or V_s. For a given SNP, V_s is reduced by increasing the number of individuals, since $V_s = f(1-f)(2N)^{-1}$, where $N = x \times y$, or the number of total individuals per cohort; x is the number of individuals per subpool; y is the number of subpools; and f is the SNP MAF. Importantly, V_s does not change by splitting the cohorts into numerous independent subpools. Conversely, V_p arises from numerous known and unknown factors, such as pool construction, assay, hybridization, and chip quality. Although each of these can be reduced by careful experimentation, V_p is most obviously reduced by use of more replicates—for example, $V_p = \sigma_p^2/m$, where m is the number of replicates and σ_p^2 is the final sum of variances due to pooled SNP allelotyping. Characteristic of current commercial SNP genotyping arrays, the major portion of cost and variance for a pooling-based GWA occurs near or before the final step of array hybridization. Consequently, previously described approaches to creating technical replicates of the more expensive preceding step (e.g., PCR) to recover power lost to pooling are not cost-effective for commercial SNP genotyping microarrays.[11] Since pooling design choices of subpools and technical replicates (as in fig. 2*A*) do not directly influence power in SNP microarray pooling studies, we return to the advantages of both these strategies in the "Discussion" section.

By pooling, one loses the power to detect associations, and this can be intuitively described as N^*, or the equivalent sample size remaining after pooling of N individuals. As defined by Barratt et al.,[11] the relative sample size (RSS) is the proportion of N^* (effective sample size) to N (pooled sample size), or RSS $= N^*/N$. RSS can also be derived using V_s and V_p. The total variance (V_T) is described by equation (1), where $V_s = f(1-f)(2N)^{-1}$. Conversely, by individual genotyping, V_T is dependent only on sample variance—that is, $V_T = f(1-f)(2N^*)^{-1}$, where N^* is the unknown ef-

fective sample size that will result in the equivalent V_T value observed when N samples are pooled. Equating V_T for both, we find $V_s + V_p = f(1-f)(2N^*)^{-1}$, where $V_s = f(1-f)(2N)^{-1}$. This can be simplified to two equivalent equations:

$$N^* = N \times V_s/(V_s + V_P) \quad (2)$$

and

$$\mathrm{RSS} = V_s/(V_s + V_P)\ . \quad (3)$$

Although this particular derivation of N^* and RSS provides us significant insight into the amount of power lost to pooled allelotyping, it does not entirely allow us to determine the resolution of a pooling-based GWA. SNPs on genotyping arrays are not independent, but rather correlate from LD. Analytical solutions for assessing resolution of a pooling-based GWA are not clear in the context of variable genomewide LD[21,37]; thus, we now move to simulated data sets.

Simulation

To more accurately assess whether sufficient information remains after introduction of measurement variance to identify the most substantial associations, or "low-hanging fruit," we analyzed simulated data sets. These simulations establish whether pooling can be effective, determine which variables most influence its effectiveness, and establish expectations for pooling-based GWA studies. They also allow us to heuristically test experimental variables that do not have clear analytical solutions, such as the impact of variable genomewide LD. Data sets were simulated by first generating "true" genotype data sets and "pooled" data sets, as follows (and described in greater detail in the "Methods" section): a true genotype set was constructed by multiple sampling of a binomial distribution for 250,000 SNPs for 400 cases and 400 controls (800 chromosomes each) and by combining these samples into simulated pools under an assumption of Hardy-Weinberg equilibrium. A single "associated" SNP was simulated in LD ($r^2 = 0.8$) with a causal variant at a defined OR. Measurement error was randomly added to each SNP, following a normal distribution and $V_p^{1/2} = 2.5\%$. The number of SNPs, the number of samples, and pooled variance were similar to those used later in experimental validation.

Inspecting figure 1*A,* we see that, in these simulations, one should be able to detect associations with an OR > 3.5 by taking the top 0.5% (or 1,250 SNPs) to phase II. However, this scenario represents a conservative calculation, because we have assumed all SNPs are independent of one another, whereas on both the Affymetrix 500K and Illumina 300K, ~70% of SNPs are in LD with one another at an $r^2 \geq 0.8$.[39] In figure 1*C,* we fix OR at 3.0 and show that as the number of arrays are doubled from 9 to 18, we can resolve genetic associations with OR > 3.0 using the best 100 SNPs by allelic frequency difference. In figure 1*B,* we see that increasing the number of samples increases resolution, although there appear to be diminishing returns as more individuals are added. This occurs as the sampling variance (V_s) decreases to levels at or below the pooled variance (V_p). Finally, in figure 1*D* and 1*E,* we investigate the effect of adding more SNPs that are in LD with one another. In figure 1*D,* we progressively increase the total number of SNPs from 250,000 to 1,000,000, in 250,000 increments. However, we assume new SNPs are in full LD with their neighbors, creating 250,000 perfect haplotype blocks. Effectively, at 250,000 SNPs, there is 1 SNP per block; at 500,000, there are 2 SNPs in complete LD per block; at 750,000, there are 3 SNPs per block; and, at 1 million, there are 4 SNPs per block. As before, one haplotype block (of 250,000 SNPs) is biased by a true association ($r^2 = 0.8$) indirectly probed by all SNPs on that block at a defined OR. Although this is undeniably an idealized scenario, it does allow us to demonstrate the point that, even though we are genotyping more SNPs, we are more likely to identify a SNP in the correct haplotype block within the top 1,000 SNPs, because of the redundancy of information content. Finally, in figure 1*E,* we move away from the idealized scenarios to simulations that are very similar to many ongoing GWA studies. In these simulations, we add variable LD across the genome, following a normal distribution, such that that 70% of SNPs are in LD with $r^2 \geq 0.8$ and the mean r^2 is 0.85. In effect, this scenario is very similar to the overall distribution of pairwise LD observed for SNPs on the Affymetrix 500K platform in the CEPH population.[40] Likewise, the error we have introduced by pooling ($V_s^{1/2} = 2.5\%$) is similar to having nine Affymetrix 500K replicates per cohort. From figure 1*E,* we show that, in simulation, we can easily identify associations with an OR between 2.5 and 3.0 by taking the top 1,250 SNPs to phase II of a GWA study. The above values, such as sample size and replicate arrays, will be dependent on the study and were chosen because they closely aligned with our experimental validation in later sections. We explore alternate study designs in the "Discussion" section.

Conclusions of Simulated Pooling

From these sets of simulations, four findings are immediately clear: (1) one should be able to detect associations with an OR of at least 2.5, and lower OR associations may be detectable depending on design; (2) probe precision (i.e., variance) greatly influences power; (3) there is a point beyond which increasing the pool size becomes less efficient (as V_s decreases to V_p); and (4) increasing the number of probed SNPs increases the ability to detect association. This last finding is particularly intriguing, since there are >1,000,000 nonredundant SNPs on the combined Affymetrix 500K, Affymetrix 100K, Illumina HumanHap 550K, and Illumina 100K panels.

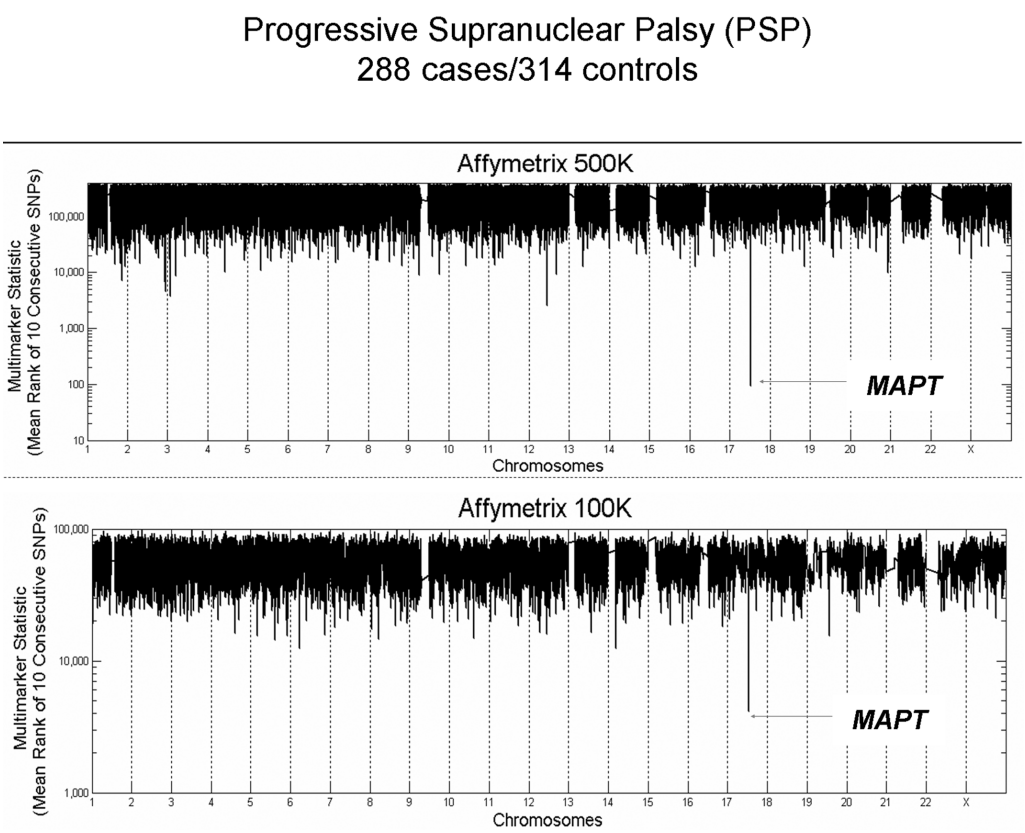

Figure 3. Example of multimarker statistics applied to a pooling-based GWA. PSP is a complex rare disorder with one consistently confirmed genetic risk factor, the *MAPT* H1 haplotype, with an allelic OR of 3.3. By use of a sliding-window calculation of mean rank across adjacent SNPs, the best-ranked window sits directly over *MAPT* when genotyped on the Affymetrix 100K and 500K platforms, and this was true for all window sizes between 3 and 31. In practice, as the window size is increased, the effectiveness of sliding-window calculations using a fixed window size will decrease, since larger window sizes will include random SNPs beyond the underlying associated haplotype block.

Development of Analysis Tools

Given that pooling should theoretically allow for detection of major associations, lack of analysis tools is a major limiting factor in completion of pooling-based GWA studies. Indeed, only a few software tools (e.g., Pooled DNA Analyzer[34]) currently address pooled microarray data. Consequently, we developed GenePool, an analysis software package for pooling-based GWA studies that has been validated on several data sets with previously published associations. GenePool and its source code can be downloaded from the Web and is free for noncommercial use. Detailed descriptions of the analysis approaches are provided in the "Methods" section. We summarize the principles of analysis in figure 2 into three stages: (1) data transformation and normalization, (2) SNP ranking by a test statistic, and (3) multimarker statistics. At each stage, there are a wide variety of analysis methods that could be applied, and only a relatively small selection has currently been implemented in GenePool. Consequently, an effort has been made to compartmentalize GenePool's code so that additional algorithms can easily be added, to extend its functionality, by individual groups.

Stage 1 of analysis is raw-data transformation. SNP microarray data begins with probe intensity data for both alleles. Effective data transformation ensures that additional information is not lost. For example, Affymetrix uses 6 to 10 probe pairs for each SNP, and each pair interrogates the SNP at a different position within the probe oligo sequence. Consequently, each of the probe pairs has unique hybridization properties. In the case of Affymetrix data, we find it most effective to treat each probe pair as an independent measure and to discard mismatch data. We transform each probe set pair into an RAS value, which is the ratio of the signal from the A allele PM probe to the sum of A and B allele PM probes. Independent transformation avoids introduction of unnecessary noise, by averaging independent hybridization events or adding in mismatch variance. Mismatch data are discarded because baseline subtraction by use of mismatch data unnecessarily adds the noise associated with the probe into the composite score,[41] and future Affymetrix SNP genotyping

arrays may not contain mismatch probes. Data transformation will be platform specific and, perhaps, study specific. For example, use of Illumina HumanHap arrays requires normalization of two-channel data, where the intensity of the green channel is substantially greater than that of the red channel. Alternative transformations, such as $\theta = \arctan(B/A)$, are also possible and may be desirable for some research groups. Lastly, the different hybridization properties of each SNP can be corrected for by calculation of the *k*-correction factor for each SNP that fits individual-genotype data to 0% (minor-allele homozygotes), 50% (heterozygotes), and 100% (major-allele homozygotes) pooled equivalents.[6,7,23] In the current GenePool, a custom *k*-correction file may be loaded. However, we note that, even without this file, significant associations can be easily found, as demonstrated in the "Experimental Validation of Case-Control GWA Studies" section and as suggested by other groups.[20]

Stage 2 of analysis is ranking by test statistic. From stage 1, we will have one or more RAS (or comparable transformation) values for each SNP for both cases and controls. For a given SNP, some RAS values may be highly informative, whereas others may be less informative. Also, a SNP probed with different sequences (e.g., quartets 1..6 with different offsets) may have highly precise $RAS_{1..6}$ values that exhibit different absolute values, because of differential hybridization with each probe quartet. With independent measures of the same SNPs, multidimensional test statistics are likely to be more appropriate. We evaluated Dunn index, linear classifier, principal-components analysis, trace criteria, and silhouette scores, using data from an AD case-control study (280 cases and 169 controls) in which both pooled data and individual-genotype data were available (Affymetrix 500K). Test statistics were evaluated by their ability to correctly rank SNPs that individual genotyping had shown to have the most significant allele frequency differences between cases and controls. Overall, we found that silhouette scores, previously employed for SNP probe intensity analysis,[31] performed the best at ranking the SNPs. As we elaborate further in the "Discussion" section, other test statistics that derive from the underlying variance will be implemented as validated. However, in the context of ranking >250,000 markers, the risk of statistical artifacts with untested (even if theoretically sound) test statistics requires an extensive validation process.

Stage 3 of analysis is multimarker analysis. Whereas stage 2 analyzed markers independently, a multimarker statistic may allow us to smooth out measurement noise and identify disorders with lower OR by leveraging LD between SNPs. A sliding-window statistic of mean rank by a test statistic is currently implemented for a fixed window size across all windows throughout the genome. This method has previously proved effective, although it is conservative because of the fixed window size and does not take into account haplotype block structure or the genomic or recombination distances between adjacent SNPs.[11] Multimarker statistics that leverage haplotype information or LD data to reduce noise are clearly a major area of active research[8,11,34] and will be implemented with future releases. However, given the highly variable LD across the genome, future validation data sets will need to be generated to evaluate the effectiveness of these approaches and to identify any potential for statistical artifact. Examples of future validation data sets may include pooled parents-child trios for which individual-genotype data are available and phase can at least be partially calculated.

Experimental Validation of Case-Control GWA Studies of Disorders with Known Associations

As shown in table 1, we experimentally studied three diseases with previously published associations by using pooling-based GWA and our GenePool analysis tools: (1) AD, both antemortem-diagnosed cases and postmortem-diagnosed cases, on the Affymetrix 500K and Illumina 300K platforms; (2) PSP, with the 500K and 100K platforms; and (3) SIDDT, with the 10K platform.[23] Each of the three studies tests a different factor influencing the ability of pooling to detect genetic associations. AD has a common variant, *APOE-ε4,* with a strong OR but is only interrogated marginally ($r^2 = 0.57$) by one SNP on the Affymetrix 500K platform and by two SNPs on the Illumina 300K platform. We generated data for one paired cohort diagnosed postmortem where diagnosis was certain (OR for *APOE-ε4* is 8.13) and three cohorts with clinically diagnosed probable AD (OR for *APOE-ε4* is ~3–8).[42,43] In PSP, the frequency of the extended microtubule-associated protein tau (*MAPT*) H1 haplotype is significantly increased in PSP cases with an OR of 3.3 in our cohort.[24] The *MAPT* H1 haplotype is covered by ~168 SNPs on the Affymetrix 500K platform and 38 SNPs on the 100K platform, because of extension of the H1 and H2 haplotypes by hundreds of kilobases due to a locus inversion, which limits recombination across this region.[44] Studying this disease tests the assertion that interrogation of multiple SNPs in LD will allow us to better detect associations. SIDDT is not a complex disorder but has desirable features. It has been studied previously,[23] it has a known genetic basis, and it has multiple SNPs in LD.

Single SNP rankings for AD, PSP, and SIDDT are shown in table 1. In cohorts with both antemortem- and postmortem-diagnosed AD, the SNP closest to *APOE-ε4* is ranked in the top 100 of 500,568 SNPs and is ranked 6th in the postmortem-diagnosed cohort for the Affymetrix 500K platform. This is a highly encouraging result, because the *APOE* region is characterized by weak LD and is covered by only one SNP on the Affymetrix 500K platform, *rs4420638,* which is ~14 kb away from the *APOE-ε4* haplotype block, with an r^2 of 0.57. Although ranked slightly lower (18 and 125) on the Illumina 300K platform, two SNPs identify the *APOE-ε4* haplotype on the same pooled samples. In PSP, on both the 100K and 500K plat-

forms, the *MAPT* region was ranked in the top two SNPs. Multimarker statistics (fig. 3) showed the *MAPT* locus to be the most significant locus. For *SIDDT,* SNPs associated with the *TSPYL* locus ranked it in the top 10 regions.

On the basis of these data sets, our experimental results confirm the expectations of the theoretical simulations and validate the analysis procedures used in GenePool. Furthermore, it is clear that pooling-based GWA studies that use existing SNP genotyping technology can identify major genetic associations with disease.

Discussion

In this article, we demonstrate, both experimentally and theoretically, that pooling-based GWA studies are effective at identifying major genetic contributions to disease. Furthermore, we discuss a general framework for pooling-based GWA studies and offer GenePool as a software tool for the analysis of pooling-based GWA studies.

In figure 2 and in the "Results" section, we investigated study design of a pooling-based GWA. Frequently asked questions about study design are (1) Should a cohort be broken down into subpools or should one have fewer large pools? (2) How many technical replicates should one conduct per pool? and (3) What is the resolution of a particular design? The answers to these questions will largely depend on the study population, the funding resources, and the characteristics of the hereditable disorder or physiological trait.

Addressing the first two questions, on the basis of our own retrospective analysis of experimental data and the theoretical framework presented earlier, we suggest study designs with multiple subpools containing equal number of samples, each run in triplicate technical replicates (as illustrated in fig. 2). First, technical replicates allow one to measure variance arising from pooled allelotyping (V_p), to eliminate poorly performing SNPs, and to provide quality-control metrics for identifying failed assays. Second, although the use of subpools does not recover power without the use of more overall microarrays, one can approximate the total variance (V_T) for each cohort. Knowing V_T is theoretically attractive because one can calculate the intuitive test statistic, $T_p = \Delta\text{RAS} \times (V_T^1 + V_T^2)^{-1/2}$, directly from probe intensity differences, whereas accounting for sampling variance depends on allelic frequency and copy-number changes. For the test statistic T_p, $\Delta\text{RAS} = |\text{RAS}^1 - \text{RAS}^2|$ and V_T^1 and V_T^2 are the total variances between cohorts 1 and 2, respectively. As an example, X-chromosome SNPs are sampled less often than are autosomal SNPs, because XY males have only one X chromosome and two of each autosome, resulting in a bias for large allelic frequency differences on the X chromosomes compared with the autosomes. Since V_T includes V_s, one accounts for these biases. Although the T_p statistic has theoretical potential, in the context of ≥250,000 SNP measurements, its use leads to biased ranking, with preferential selection of SNPs with underestimated values for V_T as a result of low relative sampling. However, one could use this statistic to reorder the top few hundred or few thousand SNPs ranked from a stage 1 analysis that used an empirically selected test statistic as currently implemented in the GenePool program.

The optimal number of arrays to be used in the pooling portion of a GWA is dependent on a complex number of variables not entirely amenable to easy solution. Zhao and colleagues have analytically derived formulas for studying some of these variables in the absence of LD.[21,37] In the present study, we used simulation to include the effect of variable LD and probe variance. Although use of nine replicate arrays was effective at identifying the *APOE-ε4* association in AD or the *MAPT* association in PSP, more replicate arrays may be required to detect associations with smaller genetic relative risk. Given the potential complexities of applying the above-described methods, we now present a simple approach for approximating the effectiveness of a pooling-based GWA, given a set number of arrays. First, RSS was shown to be equal to the proportion of sampling variance to total variance (eq. [3]). Intuitively, RSS is equivalent to the percent of the original sample remaining after pooling. All these variables are straightforward to calculate, since $V_s = f(1-f) \times (2N)^{-1}$ and $V_p = 2\sigma_p^2 \times m^{-1}$. Although we have seen σ_p^2 decrease with Affymetrix optimization of the 500K assay, σ_p^2 approximates at 5×10^{-3} in these and other studies.[35] Other defined variables are *m,* for the number of arrays; *f,* for allele frequency (~0.25); and *N,* for the number of pooled samples. Apparent from these formulas, as more microarrays are added, more power is recovered, having a significant effect when $V_s = V_p$ and RSS = 50%. With 400 cases and 400 controls, this occurs at 21 replicate arrays per cohort. Additionally, by calculating N^*, where $N^* = N \times V_s/(V_s + V_p)$ (eq. [2]) or $N^* = N \times \text{RSS}$, one can use widely available power calculators, such as Quanto,[45,46] to approximate the power or resolution of a pooling-based GWA study. Although it was not experimentally conducted here, one could reasonably have power to detect 80% of genetic associations with ORs >2.0 by using 20 replicate Affymetrix 500K arrays per cohort and 400 individuals per cohort.

Of course, as multiple genotyping platforms simultaneously emerge, one can leverage nonredundant SNP content. Between Affymetrix 100K and 500K platforms and Illumina 550K and 100K platforms, there are >1 million nonredundant SNPs being genotyped. With this density, redundant SNP content increases resolution of a pooling-based GWA by decreasing measurement noise on highly correlated neighboring SNPs in LD and insures that there are very few gaps in overall genomewide SNP coverage. Merging platforms is likely possible by use of platform-specific *k*-correction factors, and thus substantial utility exists in emerging databases that catalog allele-specific preferential amplification and/or hybridization for genotyping array platforms.[47,48] Furthermore, as many individual-genotyping GWA studies commence with one

platform that covers only 70%–80% of the genome,[39] pooling-based GWA studies on an alternate platform provide a low-cost mechanism to insure that no major associations are missed in the remaining portions of the genome. Similarly, pooling may present an opportunity to assess critical nonsynonymous SNPs more directly (e.g., by use of a 20,000 panel) than is accomplished by individual genotyping using Illumina 550K or Affymetrix 500K microarrays.

Because of the power requirements, substantial funding is required for completion of phase I of a GWA study. This may be one of the major reasons why only a few discoveries have been made using high-density SNP genotyping technology. Interestingly, the few discoveries made using this design have often been associations of large genetic effect. One of the more prominent discoveries was association of a common variant in complement H variant with age-related macular degeneration (AMD), with an OR between 2.5 and 5.[49] These types of findings are within reach and perhaps could be discovered by a pooling-based GWA.

It is unknown why many other diseases, disorders, or complex traits have common variants of large effect that remain undiscovered. AMD was not identified through multiple linkage studies and was only under the shoulder of a major linkage peak but was easily identified by a GWA study.[49] Without a strong understanding of gene function, we may not be able to pick the correct gene by using a hypothesis-driven approach. Pooling-based GWA studies are a viable alternative for screening disorders for common variations of large effect without onerous funding requirements. Use of 20 replicate arrays requires <$5,000 yet could lead to a major discovery. The alternative to individual genotyping requires 10-fold more funding, something that may be out of reach for many orphan diseases, or diseases specific to isolated populations. We have demonstrated a theoretical framework by which these questions may be investigated.

In summary, we could detect published genetic associations in a monogenic disease, a rare complex disease, and a common complex disease by using a pooling-based GWA. We anticipate major improvements in the methods and analysis tools over the next few years, with integration of pooled data from multiple platforms using combined densities of >1 million genotyped SNPs. Data analysis is a major obstacle for completion of these studies, and we have created a software tool, GenePool, for this purpose and have made its source code available for development of new approaches. We expect these tools to provide a mechanism for rare and common disorders and traits in a variety of specific populations to be rapidly and cost effectively investigated using a hypothesis-free GWA study design.

Acknowledgments

Funding of GenePool, D.A.S., D.W.C., and J.V.P. was through a National Institutes of Health ENDGAME grant U01-HL086528-01. Additional funding was provided to D.W.C. and W.D.T. by the Stardust Foundation, to D.A.S. by grant 1U24NS043571, and to E.M.R. by National Institute on Aging grant P30 AG19610, National Institute of Mental Health grant RO1 MH057899, and the state of Arizona. We thank TGen IT support staff for assistance during development of GenePool. We thank Kevin Brown for fruitful discussions. We thank David Duggan for the use of an Illumina BeadStation and Kathleen Kennedy for valuable guidance through the course of the assaying.

Web Resources

The URLs for data presented herein are as follows:

Affymetrix, http://www.affymetrix.com/
GenePool, http://genepool.tgen.org/
Illumina, http://www.illumina.com/
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for AD, PSP, and SIDDT)

References

1. Johnson C, Drgon T, Liu QR, Walther D, Edenberg H, Rice J, Foroud T, Uhl GR (2006) Pooled association genome scanning for alcohol dependence using 104,268 SNPs: validation and use to identify alcoholism vulnerability loci in unrelated individuals from the collaborative study on the genetics of alcoholism. Am J Med Genet B Neuropsychiatr Genet 141:844–853
2. Liu QR, Drgon T, Walther D, Johnson C, Poleskaya O, Hess J, Uhl GR (2005) Pooled association genome scanning: validation and use to identify addiction vulnerability loci in two samples. Proc Natl Acad Sci USA 102:11864–11869
3. Craig DW, Stephan DA (2005) Applications of whole-genome high-density SNP genotyping. Expert Rev Mol Diagn 5:159–170
4. Butcher LM, Meaburn E, Dale PS, Sham P, Schalkwyk LC, Craig IW, Plomin R (2005) Association analysis of mild mental impairment using DNA pooling to screen 432 brain-expressed single-nucleotide polymorphisms. Mol Psychiatry 10:384–392
5. Butcher LM, Meaburn E, Knight J, Sham PC, Schalkwyk LC, Craig IW, Plomin R (2005) SNPs, microarrays and pooled DNA: identification of four loci associated with mild mental impairment in a sample of 6000 children. Hum Mol Genet 14:1315–1325
6. Hoogendoorn B, Norton N, Kirov G, Williams N, Hamshere ML, Spurlock G, Austin J, Stephens MK, Buckland PR, Owen MJ, et al (2000) Cheap, accurate and rapid allele frequency estimation of single nucleotide polymorphisms by primer extension and DHPLC in DNA pools. Hum Genet 107:488–493
7. Le Hellard S, Ballereau SJ, Visscher PM, Torrance HS, Pinson J, Morris SW, Thomson ML, Semple CA, Muir WJ, Blackwood DH, et al (2002) SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage and analysis. Nucleic Acids Res 30: e74
8. Hinds DA, Seymour AB, Durham LK, Banerjee P, Ballinger DG, Milos PM, Cox DR, Thompson JF, Frazer KA (2004) Application of pooled genotyping to scan candidate regions for association with HDL cholesterol levels. Hum Genomics 1: 421–434
9. Norton N, Williams NM, O'Donovan MC, Owen MJ (2004)

DNA pooling as a tool for large-scale association studies in complex traits. Ann Med 36:146–152
10. Sham P, Bader JS, Craig I, O'Donovan M, Owen M (2002) DNA pooling: a tool for large-scale association studies. Nat Rev Genet 3:862–871
11. Barratt BJ, Payne F, Rance HE, Nutland S, Todd JA, Clayton DG (2002) Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. Ann Hum Genet 66:393–405
12. Bansal A, van den Boom D, Kammerer S, Honisch C, Adam G, Cantor CR, Kleyn P, Braun A (2002) Association testing by DNA pooling: an effective initial screen. Proc Natl Acad Sci USA 99:16871–16874
13. Buetow KH, Edmonson M, MacDonald R, Clifford R, Yip P, Kelley J, Little DP, Strausberg R, Koester H, Cantor CR, et al (2001) 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 98:581–584
14. Jurinke C, van den Boom D, Cantor CR, Koster H (2001) Automated genotyping using the DNA MassArray technology. Methods Mol Biol 170:103–116
15. Jurinke C, van den Boom D, Cantor CR, Koster H (2002) The use of MassARRAY technology for high throughput genotyping. Adv Biochem Eng Biotechnol 77:57–74
16. Kammerer S, Burns-Hamuro LL, Ma Y, Hamon SC, Canaves JM, Shi MM, Nelson MR, Sing CF, Cantor CR, Taylor SS, et al (2003) Amino acid variant in the kinase binding domain of dual-specific A kinase-anchoring protein 2: a disease susceptibility polymorphism. Proc Natl Acad Sci USA 100:4066–4071
17. Nelson MR, Marnellos G, Kammerer S, Hoyal CR, Shi MM, Cantor CR, Braun A (2004) Large-scale validation of single nucleotide polymorphisms in gene regions. Genome Res 14: 1664–1668
18. Tang K, Oeth P, Kammerer S, Denissenko MF, Ekblom J, Jurinke C, van den Boom D, Braun A, Cantor CR (2004) Mining disease susceptibility genes through SNP analyses and expression profiling using MALDI-TOF mass spectrometry. J Proteome Res 3:218–227
19. Meaburn E, Butcher LM, Liu L, Fernandes C, Hansen V, Al-Chalabi A, Plomin R, Craig I, Schalkwyk LC (2005) Genotyping DNA pools on microarrays: tackling the QTL problem of large samples and large numbers of SNPs. BMC Genomics 6: 52
20. Macgregor S, Visscher PM, Montgomery G (2006) Analysis of pooled DNA samples on high density arrays without prior knowledge of differential hybridization rates. Nucleic Acids Res 34:e55
21. Zuo Y, Zou G, Zhao H (2006) Two-stage designs in case-control association analysis. Genetics 173:1747–1760
22. Rabbee N, Speed TP (2006) A genotype calling algorithm for affymetrix SNP arrays. Bioinformatics 22:7–12
23. Craig DW, Huentelman MJ, Hu-Lince D, Zismann VL, Kruer MC, Lee AM, Puffenberger EG, Pearson JM, Stephan DA (2005) Identification of disease causing loci using an array-based genotyping approach on pooled DNA. BMC Genomics 6:138
24. Rademakers R, Melquist S, Cruts M, Theuns J, Del-Favero J, Poorkaj P, Baker M, Sleegers K, Crook R, De Pooter T, et al (2005) High-density SNP haplotyping suggests altered regulation of tau gene expression in progressive supranuclear palsy. Hum Mol Genet 14:3281–3292
25. Hauw JJ, Daniel SE, Dickson D, Horoupian DS, Jellinger K, Lantos PL, McKee A, Tabaton M, Litvan I (1994) Preliminary NINDS neuropathologic criteria for Steele-Richardson-Olszewski syndrome (progressive supranuclear palsy). Neurology 44:2015–2019
26. Litvan I, Hauw JJ, Bartko JJ, Lantos PL, Daniel SE, Horoupian DS, McKee A, Dickson D, Bancher C, Tabaton M, et al (1996) Validity and reliability of the preliminary NINDS neuropathologic criteria for progressive supranuclear palsy and related disorders. J Neuropathol Exp Neurol 55:97–105
27. Caselli RJ, Osborne D, Reiman EM, Hentz JG, Barbieri CJ, Saunders AM, Hardy J, Graff-Radford NR, Hall GR, Alexander GE (2001) Preclinical cognitive decline in late middle-aged asymptomatic apolipoprotein E-e4/4 homozygotes: a replication study. J Neurol Sci 189:93–98
28. Caselli RJ, Hentz JG, Osborne D, Graff-Radford NR, Barbieri CJ, Alexander GE, Hall GR, Reiman EM, Hardy J, Saunders AM (2002) Apolipoprotein E and intellectual achievement. J Am Geriatr Soc 50:49–54
29. Toft M, Sando SB, Melquist S, Ross OA, White LR, Aasly JO, Farrer MJ (2005) *LRRK2* mutations are not common in Alzheimer's disease. Mech Ageing Dev 126:1201–1205
30. Puffenberger EG, Hu-Lince D, Parod JM, Craig DW, Dobrin SE, Conway AR, Donarum EA, Strauss KA, Dunckley T, Cardenas JF, et al (2004) Mapping of sudden infant death with dysgenesis of the testes syndrome (SIDDT) by a SNP genome scan and identification of TSPYL loss of function. Proc Natl Acad Sci USA 101:11689–11694
31. Lovmar L, Ahlford A, Jonsson M, Syvanen AC (2005) Silhouette scores for assessment of SNP genotype clusters. BMC Genomics 6:35
32. Francisco A, Bolshakova N (2002) Clustering genomic expression data: design and evaluation principles. In: Berrar D, Dubitzky W, Granzow M (eds) A practical approach to microarray data analysis. Kwuwer Academic Publishers, Boston, Dordrecht, and London
33. Brohede J, Dunne R, McKay JD, Hannan GN (2005) PPC: an algorithm for accurate estimation of SNP allele frequencies in small equimolar pools of DNA using data from high density microarrays. Nucleic Acids Res 33:e142
34. Yang HC, Pan CC, Lin CY, Fann CS (2006) PDA: pooled DNA analyzer. BMC Bioinformatics 7:233
35. Meaburn E, Butcher LM, Schalkwyk LC, Plomin R (2006) Genotyping pooled DNA using 100K SNP microarrays: a step towards genomewide association scans. Nucleic Acids Res 34: e27
36. Moskvina V, Norton N, Williams N, Holmans P, Owen M, O'Donovan M (2005) Streamlined analysis of pooled genotype data in SNP-based association studies. Genet Epidemiol 28:273–282
37. Zou G, Zhao H (2004) The impacts of errors in individual genotyping and DNA pooling on association studies. Genet Epidemiol 26:1–10
38. Huentelman MJ, Craig DW, Shieh AD, Corneveaux JJ, Hu-Lince D, Pearson JV, Stephan DA (2005) SNiPer: improved SNP genotype calling for Affymetrix 10K GeneChip microarray data. BMC Genomics 6:149
39. Pe'er I, de Bakker PI, Maller J, Yelensky R, Altshuler D, Daly MJ (2006) Evaluating and improving power in whole-genome

association studies using fixed marker sets. Nat Genet 38:663–667
40. Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P (2005) A haplotype map of the human genome. Nature 437:1299–1320
41. Millenaar FF, Okyere J, May ST, van Zanten M, Voesenek LA, Peeters AJ (2006) How to decide? Different methods of calculating gene expression from short oligonucleotide array data will give different results. BMC Bioinformatics 7:137
42. Bennett DA, Wilson RS, Schneider JA, Evans DA, Aggarwal NT, Arnold SE, Cochran EJ, Berry-Kravis E, Bienias JL (2003) Apolipoprotein E ϵ4 allele, AD pathology, and the clinical expression of Alzheimer's disease. Neurology 60:246–252
43. Qiu C, Kivipelto M, Aguero-Torres H, Winblad B, Fratiglioni L (2004) Risk and protective effects of the APOE gene towards Alzheimer's disease in the Kungsholmen project: variation by age and sex. J Neurol Neurosurg Psychiatry 75:828–833
44. Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, Jonasdottir A, Ingason A, Gudnadottir VG, et al (2005) A common inversion under selection in Europeans. Nat Genet 37:129–137
45. Skol AD, Scott LJ, Abecasis GR, Boehnke M (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38:209–213
46. Gauderman WJ, Morrison JM (2006) Quanto 1.1: a computer program for power and sample size calculations for genetic epidemiology studies (http://hydra.usc.edu/gxe)
47. Yang HC, Liang YJ, Huang MC, Li LH, Lin CH, Wu JY, Chen YT, Fann CS (2006) A genome-wide study of preferential amplification/hybridization in microarray-based pooled DNA experiments. Nucleic Acids Res 34:e106
48. Simpson CL, Knight J, Butcher LM, Hansen VK, Meaburn E, Schalkwyk LC, Craig IW, Powell JF, Sham PC, Al-Chalabi A (2005) A central resource for accurate allele frequency estimation from pooled DNA genotyped on DNA microarrays. Nucleic Acids Res 33:e25
49. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, et al (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308:385–389