Multimarker analysis and imputation of multiple platform pooling-based genome-wide association studies

Nils Homer1,2, Waibhav D. Tembe1, Szabolcs Szelinger1, Margot Redman1, Dietrich A. Stephan1, John V. Pearson1, Stanley F. Nelson2, and David Craig1

1Translational Genomics Research Institute (TGen), Phoenix, AZ 85004;
2University of California, Los Angeles 90095-7088

ABSTRACT

Summary: For many genome-wide association (GWA) studies individually genotyping one million or more SNPs provides a marginal increase in coverage at a substantial cost. Much of the information gained is redundant due to the correlation structure inherent in the human genome. Pooling based GWA studies could benefit significantly by utilizing this redundancy to reduce noise, improve the accuracy of the observations, and increase genomic coverage. We introduce a measure of correlation between individual genotyping and pooling, under the same framework that \( r^2 \) provides a measure of linkage disequilibrium between pairs of SNPs. We then report a new non-haplotype multimarker multi-loci method that leverages the correlation structure between SNPs in the human genome to increase the efficacy of pooling based GWA studies. We first give a theoretical framework and derivation of our multimarker method. Next, we evaluate simulations using this multimarker approach in comparison to single marker analysis. Finally, we experimentally evaluate our method using different pools of HapMap individuals on the Illumina 450S Duo, Illumina 550K, and Affymetrix 5.0 platforms for a combined total of 1,333,631 SNPs. Our results show that use of multimarker analysis reduces noise specific to pooling based studies, allows for efficient integration of multiple microarray platforms, and provides more accurate measures of significance than single marker analysis. Additionally, this approach can be extended to allow for imputing the association significance for SNPs not directly observed using neighboring SNPs in linkage disequilibrium. This multimarker method can now be used to cost-effectively complete pooling-based GWA studies with multiple platforms across over one million SNPs and to impute neighboring SNPs weighted for the loss of information due to pooling.

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Genome-wide association (GWA) studies have emerged as a new and powerful tool to detect genetic predisposition to complex diseases. Frequently, upwards of thousands of individuals are genotyped for several hundred thousand SNPs in order to find the single most significant SNP using a genotype or an allele based \( \chi^2 \) test. Considering the cost of such an experiment is several hundred thousand dollars with no guarantee of success, it is of high importance to identify cost-effective methods for completing GWA studies. Pooling genomic DNA and assaying on a few replicate arrays is such an approach, and has yielded new candidate associations in such an approach, and has yielded new candidate associations in situations where individual genotyping of samples was not possible [1-7].

Genotype multiplexing has reached the level of one million largely non-redundant SNPs for two separate technologies developed in parallel, Illumina and Affymetrix. For many populations, such as Caucasian or Asian, the additional information gained from genotyping 1,000,000 or more SNPs versus 500,000 SNPs is limited and largely redundant due to high correlation (linkage disequilibrium or LD) between neighboring SNPs. However, in the context of a pooling-based GWA study, this redundancy in coverage should theoretically allow for reduction in noise from the assay and substantial improvement in the performance of a pooling-based GWA study thus increasing the number of true associations found and reducing the number of false positives. Furthermore, one should be able to utilize multiple platforms within a single study in order to improve overall resolution and increase genomic coverage. To date, numerous papers have examined the efficacy of pooling-based GWA studies [8-11], including study design [8, 12-14], accounting for sources of errors [8, 10, 15, 16], and cost/power analysis[11, 14, 17-21]. Some of these studies have explored the use of multimarker statistics [17, 22, 23], though in large part the effectiveness of these approaches has not been explored under the context of a GWA study. Methods have been developed to leverage the correlation structure between SNPs with respect to haplotype analysis and individual genotyping[23, 24] but methods for pooling-based studies have been largely limited. Additionally, methods have been developed that are able to impute or estimate unobserved genotypes, increasing the power to detect association[25-27]. Nevertheless, their application to pooling-based studies where individual genotypes are unknown has not been fully explored.

To this end, we develop and evaluate a new method of analysis and demonstrate the effectiveness of this approach using 1,333,631 SNPs combined from the Affymetrix 5.0, Illumina 550K, and Illumina 450S Duo arrays. Our approach leverages the correlation structure between SNPs to reduce Type I and Type II errors, combines multiple platforms, and increases the accuracy and power of pooling-based GWA studies. We introduce the concept of pooling correlation coefficient, the square root of \( r_{p}^2 \), analogous to \( r^2 \) in individual genotyping, where \( r_{p}^2 \) is the amount of information recovered from an allele-based test of association from individual genotype data by using a pooling based method. Our multimarker test statistic utilizes both the pooling correlation coefficient and the correlation or LD between neighboring SNPs to combine data from multiple neighboring SNPs and from multiple platforms (Affymetrix and Illumina) within a single pooling-based GWA study. Therefore, we more accurately determine the significance of association for each SNP as well as giving greater coverage of the human genome. Additionally, our method lends itself to imputation, where an unobserved SNP is given a significance value based on directly observed neighboring SNPs in LD. Combining the Illumina 450S Duo and Illumina 550K v3 platforms we are able to accurately impute 748,348 additional SNPs from the HapMap that are not present on any of the three platforms. Nevertheless, in genomic regions of high LD the number of proxies for an imputed SNP can exceed 50 SNPs but typically have only a few (less than 5) proxies. Therefore, as microarray technologies are able to probe more SNPs, our method will have more observations to reduce noise further increasing the accuracy of the observations.
2 METHODS

2.1 Experimental
Two main pools (A and B) were pooled for a total of 2 pools (see Supplementary Table 1). Both pools (cohorts) were run on duplicates Illumina 550K v3 arrays, Illumina 450S Duo arrays, and Affymetrix 5.0 arrays respectively (see Supplementary Methods).

2.2 Derivation of a pooling-based test-statistic.
We assume Hardy-Weinberg equilibrium. Let the probability of having allele A (or the population frequency) be $p_A$, so that $(p_A)^2 + 2p_Aq_A + (q_A)^2 = 1$. We choose to represent variables with a ‘+’ as belonging to the cases and variables with a ‘−’ as belonging to the controls. For individual genotyping, suppose we observe $p_+$ as belonging to the cases and $p_−$ as belonging to the controls. For individual genotyping, suppose we observe $p_+$ as belonging to the cases and $p_−$ as belonging to the controls. Then from individual genotyping the frequency or probability of allele A in the cases is $p_+ = N_+/(N_+ + N_-)$ and in the controls is $p_- = N_-/(N_+ + N_-)$ where $A$ is the other allelic variant. We assume in practice that $p_+ ≈ p_A$ since typically $p_A$ is not known. To test for association we used a two-sample test of proportions, which is equivalent to a t-test under HWE, as shown in Equation 1 where $T_A$ is the test-statistic.

$$T_A = \frac{p_A^+ - p_A^-}{\sqrt{2p_Aq_A/N_A}}$$

Under the null hypothesis, we have the expected value $E(p_+^A - p_-^A) = 0$, and the variance $V(p_+^A - p_-^A) = p_A(1-p_A)/N_A + p_-(1-p_-)/N_-$. If we approximate $V(p_+^A - p_-^A)$ with $2p_Aq_A/N_A$, then $T_A$ is expected to follow the normal distribution under HWE. In a pooling-based estimate of allele frequency, we do not observe the allele counts but instead indirectly observe an allele frequency for each pool by measuring pooled amplified genomic DNA, labeled with a fluorophore, and hybridized to an oligonucleotide probe, though not in that order. Typically, a predicted allelic frequency is calculated based on the observed relative probe intensity of the oligonucleotide probes interrogating both SNP alleles. Here we are more concerned with predicting allele frequency differences than accurately predicting the allele frequencies themselves as will become evident by defining our pooling test statistic below. We define $\hat{p}_A^+$, $\hat{p}_A^-$, and $\hat{p}_A = 1 - \hat{q}_A$ as the respective measured frequencies for the A allele in the case, control, and combined populations through pooling. We consequently define an analogous test-statistic for our measurement of pooled DNA:

$$\hat{T}_A = \frac{\hat{p}_A^+ - \hat{p}_A^-}{\sqrt{2\hat{p}_A\hat{q}_A/M}}$$

Here we have that $2\hat{e}/M$ is the variance of $N_A$ alleles with $M$ replicate measurements, where $\hat{e}$ is the measurement variance. In order to simplify our discussion in later sections, we denote the total variance from sample mean with a defined set of individuals as $\hat{V}_s = \hat{V}_p + \hat{V}_m$, where $\hat{V}_p = 2\hat{p}_A\hat{q}_A/N_A$ and $\hat{V}_m = 2\hat{e}/M$. There potentially exist other sources of bias and variance, including systematic biases to the measured values for $\hat{p}_A$, additional source of variances from the arrays, use of multiple sub-pools, and experimental variance. Previous studies have investigated the relative source of variation in pooled experiments and have shown that the variance from the measured arrays is significantly larger than all other sources of variances [8, 19].

2.3 Derivation of a pooling-based quality control statistic ($R^2_{T_A}$ or $r^2_p$).
To mathematically investigate a relationship between individual genotyping and pooling based tests, we introduce a measure of correlation between individual genotyping and pooling under the same framework that $r^2$ provides a measure of linkage disequilibrium between pairs of SNPs. Briefly, we compute the sum of squared deviations to determine the correlation $R^2_{T_A\hat{p}_A}$ between the pooling test statistic $\hat{T}_A$ and a modified individual genotyping test statistic that has a shifted mean $\hat{p}_A$. The shift in the mean comes from the introduction of errors due to pooling. Next, we repeat the calculation to determine the correlation $R^2_{\hat{p}_A\hat{T}_A}$ between the individual genotyping test statistic with a shifted mean $\hat{p}_A$ and the standard individual genotyping test statistic $T_A$. Finally, we combine these correlations to obtain a theoretical pooling correlation coefficient $R^2_{\hat{T}_A\hat{T}_A}$, which is simply the correlation between the pooling test statistic $\hat{T}_A$ and the individual genotyping test statistic $T_A$. A detailed derivation can be found in the Supplementary Methods, where we derive the relationship:

$$R^2_{\hat{T}_A\hat{T}_A} = \left(\frac{V_s}{V_s + V_p}\right)$$

We can therefore view pooling-based experiments according to their theoretical pooling correlation coefficient $R^2_{\hat{T}_A\hat{T}_A}$. This value could be used as a measure for the ability of a SNP to resolve allelic associations and also allows us to correlate our test-statistics with individual genotyping, critical for development of a multimarker statistic. An alternative viewpoint is that the pooling correlation gives us a measure of the loss of power due to pooling when compared to individual genotyping. For clarity and since it holds a similar theoretical basis as the term $r^2$ for linkage disequilibrium, we similarly refer to this value as $r^2_p$ in the discussion sections.

2.4 Development of a multimarker test statistic.
To develop a multimarker test statistic for pooling-based GWA studies, we use the previously derived pooling correlation coefficient, the square root of $r^2_p$, and the measured linkage disequilibrium between two different SNPs, measured as $r^2$ or the coefficient of determination between a typed and an un-typed marker. From indirect association we know that the power of ob-
serving an association at marker \( A \) for a causal mutation at marker \( B \) is simply scaled by the correlation between SNP \( A \) and SNP \( B \) [28]. Combining this correlation with our pooling correlation we create a multimarker test statistic that combines the information from neighboring SNPs to give more accurate and meaningful association values.

It has been previously shown that the test statistics of two neighboring SNPs \( A \) and \( B \) are equivalent when scaled by the correlation \( r^2_{AB} \) between the two SNPs [28]. We give a formal derivation in the Supplementary Methods. Now, suppose we have a causal mutation in SNP \( A \), and a set \( S_A \) of other SNPs in LD with \( A \). Let \( \tilde{T}_A \) be the test statistic for the true genotypes but with a shifted mean as above, let \( \tilde{T}_A \) be the pooling test statistic, and let \( \hat{T}_A \) be the multimarker test statistic. Then we propose the following test statistic:

\[
\hat{T}_A = \frac{\hat{T}_B / r^2_{pA} + \sum_{B \in S_A} 2 \sqrt{\frac{r^2_{AB}}{p_B}}}{1 / p_A + \sum_{B \in S_A} \frac{1}{\sqrt{p_B}}}.
\]  

(4)

where \( r^2_{AB} \) is the coefficient of determination (or linkage disequilibrium) between SNP \( A \) and SNP \( B \) and \( \frac{1}{p_B} \) is the square of the pooling correlation for SNP \( B \). Essentially, using the square root of \( r^2_{pB} \) and the square root of \( r^2_{AB} \), we transform multiple indirect observations of SNP \( A \) into equivalent measurements and take the weighted average of those observations. Note that if we assume \( \varepsilon^2 \approx \varepsilon^2 \) then:

\[
\hat{T}_B = \frac{\hat{T}_B / r^2_{pA} + \sum_{B \in S_A} 2 \sqrt{\frac{r^2_{AB}}{p_B}}}{1 / p_A + \sum_{B \in S_A} \frac{1}{\sqrt{p_B}}} = r^2_{pA}.
\]  

(5)

Otherwise, we have:

\[
\hat{T}_B = \frac{\hat{T}_B / r^2_{pA} + \sum_{B \in S_A} 2 \sqrt{\frac{r^2_{AB}}{p_B}}}{1 / p_A + \sum_{B \in S_A} \frac{1}{\sqrt{p_B}}} = r^2_{pA}.
\]  

(6)

To compute \( \hat{T}_B \), we estimate \( V_p = 2p_Aq_A/N_A \) and \( V_p = 2\varepsilon^2/M \). We use the approximation \( p_A \approx p_A \) for computing \( V_p \). To estimate \( \varepsilon^2 \) we simply sum the variance from each cohort, where the variance from a cohort is simply the sum of variances between the microarrays in that cohort, with the variances within each microarray. In practice the number of individuals within each cohort (or pool) may not be equal, which is adjusted for by substituting \( N_i \) for \( 2N_A^+N_A^+/(N_A^++N_A^-) \).

### 2.5 Imputation using the multimarker test statistic.

Imputing the significance of association for SNPs that are not directly observed is achieved by using the derived multimarker test statistic. For a given unobserved SNP \( A \), we simply have the set \( S_A \) of other observed SNPs in LD with \( A \) excluding the (unobserved) multimarker test statistic for SNP \( A \) \( \hat{T}_A \). The set of SNPs \( S_A \) simply act as proxies for SNP \( A \). The main advantage to using the multimarker test statistic is that we have multiple proxies from which we measure significance. Modifying Equation (4) we obtain the following multimarker for SNP \( A \):

\[
\hat{T}_A = \frac{\sum_{B \in S_A} 2 \sqrt{\frac{r^2_{AB}}{p_B}}}{1 / p_A + \sum_{B \in S_A} \frac{1}{\sqrt{p_B}}}.
\]  

(7)

Intuitively, as the size of \( S_A \) increases so does the accuracy of the multimarker since we then have more than one proxy for the given SNP. The variance may increase as well but is determined by the accuracy of the pooling correlation and linkage disequilibrium estimates.

### 2.6 Combining multiple platforms using the multimarker test statistic.

The multimarker test statistic can also be used to combine data from multiple SNP microarray platforms, even when the platforms contain common SNPs. To combine the data we first calculate the pooling test statistic and pooling correlation for each SNP and each platform separately. Let \( S_A \) be the set of SNPs in LD with SNP \( A \). Let \( B_i \) be the observation for SNP \( B \) in \( S_A \) on the \( i \)th platform. Then from Equation (4) the pooling test statistic is simply:

\[
\hat{T}_A = \frac{\sum_{B \in S_A} 2 \sqrt{\frac{r^2_{AB}}{p_B}}}{1 / p_A + \sum_{B \in S_A} \frac{1}{\sqrt{p_B}}}.
\]  

(8)

If SNP \( A \) is not directly observed we can impute SNP \( A \) from observations on multiple platforms with the following test statistic:

\[
\tilde{T}_A = \frac{\sum_{B \in S_A} 2 \sqrt{\frac{r^2_{AB}}{p_B}}}{1 / p_A + \sum_{B \in S_A} \frac{1}{\sqrt{p_B}}}.
\]  

(9)

### 3 RESULTS

To experimentally evaluate the efficacy of our multimarker method we used the HapMap dataset to compare individual genotyping and pooling under an example GWA study. From the HapMap project we are able to retrieve the genotypes for the CEU population. We randomly split CEU trios into two separate pools,
consisting of 41 individuals in pool A and 47 individuals in pool B to create a model genome-wide association study whereby the genotypes for each individual were certain. Due to sample quality, we excluded one individual from a given trio from each pool. Both pools were run on duplicates Illumina 550K v3 arrays, Illumina 450S Duo arrays, and Affymetrix 5.0 arrays respectively. For each microarray we removed the lowest 1% of raw intensity values and normalized the microarray by dividing by the mean channel intensity.

We were able to probe 504,604 SNPs on the Illumina 550K v3 arrays, with 487,723 (~96.6%) of those SNPs having associated genotypes in the HapMap dataset. On the Illumina 450S Duo arrays we were able to probe 510,506 SNPs, with 493,495 (96.7%) of those SNPs having associated genotypes in the HapMap dataset. Finally, we were able to probe 440,729 SNPs on the Affymetrix 5.0 arrays, with 427,254 (~96.9%) of those SNPs having associated genotype information in the HapMap dataset. There were two replicates for each pool and platform.

To evaluate our multimarker method we used all SNPs on the arrays filtering those that could not generate pooling test statistic due to errors or insufficient data. Nevertheless, it has been found that there is an enrichment of false-positives due to genotyping error among the most significant SNPs when individually genotyping [29]. In order to accurately and fairly assess our approach, it is necessary to remove these false positives due to genotyping error. In other words, simply because a SNP is identified as the single most associated SNP by individual genotyping, this does not mean that this result is not due to a calling problem, copy number variant, or an assay problem. While there is no perfect method to screen out SNPs that give rise to false positives, the most accepted approach is a series of filters. Thus, only SNPs passing the following filters were used in successive order for evaluation:

1. All SNPs that had an individual genotyping minor allele frequency greater than 0.05.
2. All SNPs that had less than or equal to two no calls in both the case and control pools respectively with HapMap genotypes.
3. All SNPs that when tested for Hardy-Weinberg equilibrium with a $\chi^2$ test had a p-value greater than or equal to 0.01 across cohorts.
4. Only autosomal SNPs were used.
5. All SNPs that had at least one other SNP in LD with value of $R^2$ $\geq$ 0.8.
6. All SNPs that had genotype data in the HapMap.

We define the true rank of a SNP to be the rank of the SNP according to the Fisher’s exact p-value from individual genotype data. Additionally we define the top X truly associated SNPs as those SNPs are in the top X inclusive when ranked. We adopt these filters because the remaining SNPs allow us to better assess the performance of our method. A detailed explanation of these filters can be found in the Supplementary Results. From these filters, we are left with 139,202 SNPs (~29.1%) for the Illumina 550 K v3 arrays, 87,678 SNPs (~27.6%) for the Illumina 450S Duo arrays, and 194,074 SNPs (~44.0%) for the Affymetrix 5.0 arrays, with the overwhelming majority filtered by the fifth criteria in all three cases.

For the analysis of Illumina 550K data alone, using previous individual genotype data we were able to correct for preferential amplification during the PCR process for the Illumina arrays. This was done through a traditional k-correction factor [30, 31]. This type of correction can significantly reduce biases in alleles between true and observed allelic frequency. Nevertheless, for both the Illumina 450S Duo analysis and the Affymetrix 5.0 analysis there was no previous individual genotype data associated with the versions of arrays used. Additionally, when combining platforms k-correction was not used. It is also interesting to note that because the HapMap CEU individuals are composed of trios, the number of independent chromosomes per trio is four instead of six for unrelated individuals, which may cause our variance estimates to be less accurate than when using unrelated individuals.

### 3.1 Analysis improvement by a multimarker statistic.

A comparison between single marker and multimarker analysis of a pooled GWA study is shown in Figure 1 and Supplementary Figures 1-2 under several scenarios. Supplementary Figure 1 and 2 show the multimarker analysis for Affymetrix 5.0 arrays and Illumina arrays respectively, and considers the trade-off between restricting our analysis to a single platform (alone) and combining platforms (combined). To analyze the Illumina platform we combined the data from Illumina 550K v3 arrays and the Illumina 450S Duo arrays for a total of 1,015,110 SNPs before filtering, and 309,688 SNPs (30.5%) after filtering. Figure 1 shows a combined multimarker analysis when data is merged from the three different microarrays. When combining the Illumina 550K v3 arrays, Illumina 450S Duo arrays, and Affymetrix 5.0 arrays, the total number of SNPs before filtering was 1,333,631 SNPs and 560,202 SNPs (~42.0%) after filtering. We completed the same analysis within graphs A-E of each figure. There are various methods by which one could evaluate performance of a multimarker statistic, and this choice is largely arbitrary. We observe that our test statistic presented may not follow a chi-square distribution and therefore we choose a rank-based evaluation (Spearman rank correlation). We also wish to perform an evaluation on how a researcher would use the data and so we focus on the number of true associations identified from individual genotyping that would be carried forward in a two-stage design. We define the analysis of our initial pooling test statistic as single marker analysis (SA) and our multimarker test statistic as multimarker analysis (MM).

#### 3.1.1 Evaluation Metric 1 – Identification of the most associated SNPs within a 2 stage design (Graphs A and B).

Within a GWA study typically the first objective of the researcher is to identify those SNPs exhibiting the largest change in allelic association. Typically and especially with two-stage GWA designs, a somewhat arbitrary number of SNPs are taken forward for individual genotyping in order to accurately calculate significance, reducing the dataset from 500K+ to a few hundred or few thousand. Suppose we wish to carry forward as little as 100 and at most 5,000 SNPs for validation. Therefore, it is important to consider what percentage of the true associated SNPs that are observed in the set of SNPs carried forward. For this analysis we consider the top 100 truly associated SNPs. In graph A in Figure 1 and Supplementary Figures 1 and 2, we plot for a given observed rank threshold (the number of SNPs to be carried forward), the percentage of SNPs that were observed to be in the observed rank threshold ($x$-axis) and were a top 100 truly associated SNP. We plot this percentage for both the Affymetrix and Illumina platforms respectively, and also for both the single marker and multimarker analysis, respectively. In graph B we simply look at the difference, or improvement, in the percentages from observing the single marker ranks to observing the multimarker ranks.
3.1.2 Evaluation Metric 2 – Rank Correlation (Graphs C and D). Another measure of significance is the correlation between the true Ranks and our observed single marker or multimarker ranks. In graph C in Figure 1 and Supplementary Figures 1 and 2, we plot the Spearman rank correlation between the true and observed ranks considering the SNPs within the true rank threshold (x-axis). We see in graph D the improvement in correlation between the single marker ranks and the multimarker ranks.

3.1.3 Evaluation Metric 3 – Identification of top SNPs and directionality of change (Graph E). In graph E we plot the percentage of SNPs that fall within one of two criteria: either the SNP was both observed in the top 100 and within the true rank threshold (x-axis) or the SNP was moved in the correct direction by the multimarker analysis. A SNP moves in the correct direction if the multimarker rank is closer to the true rank than the single marker rank. Our main goal is to improve the correspondence between the true and observed ranks and thus an improvement in the observed ranks should be found.

We clearly see in Figure 1 and Supplementary Figures 1 and 2 that the multimarker rank improves on the single marker under all scenarios. As an example, (see Supplementary Table 2) consider the single marker and multimarker ranks for the top 100 truly associated SNPs on the Illumina platform when considering just the 450S Duo and 550K platforms (alone) and when considering all three microarray types (combined). We notice that the improvement is greater on the Affymetrix platform, which is expected since the there is greater noise and the number of probes per SNP is fewer than on the Illumina platform. The improvement is significant since using the multimarker method (see graph B) we potentially increase the number of the top 100 truly associated SNPs carried forward by five to thirty-five percent depending on the number of observed SNPs to be carried forward for validation. Furthermore, if we combine the information from all platforms, we can include 100% of the top 100 truly associated SNPs by carrying forward the top 2,500 observed SNPs. We can include 90% of the top 100 truly associated SNPs by carrying forward the top 1,000 observed SNPs. We see the correlation between the true ranks and the observed ranks is higher in the multimarker analysis (graphs C and D), and we verify this in graph E by seeing the directionality of that change. Additionally, in analyzing Supplementary Figures 1 and 2, we see there is an improvement in combining both the Affymetrix and Illumina data versus considering them separately, clearly suggesting that our method improves further when data from multiple platforms are combined.

3.2 Simulation.

We performed a simulation of a pooling study using pools composed by random sampling individuals from the 1958 Control Cohort of the Wellcome Trust dataset[32]. Ignoring duplicates, relatives and other data anomalies left a total of 1,423 individuals. The genotype calls for these individuals were provided from the WTCCC and were previously genotype on the Affymetrix 500K platform. Using this dataset, we simulated the Affymetrix 5.0 arrays by using four probes per SNP and by adding a mean zero error with variance 0.006 to the value of each probe. The probe variance of 0.006 matches our observed probe variance for Affymetrix 5.0 arrays. Our simulated study design consisted of pools of one hundred cases and one hundred controls with four replicate arrays for each cohort using a total number of 500,567 SNPs. Similar to our experimental analysis, we assumed a correlation structure to that of the HapMap and used the correlation structure (LD) found in the HapMap as input to our method. We evaluated the simulation results using the same metrics used to evaluate our experimental results. Unlike in the experimental results, the correlation structure used in the simulations was not directly trained from the WTCCC data but instead from the HapMap. Nevertheless, the results showed almost the exact same improvements as the empirical results, including a noticeable increase in spearman rank correlation for our multimarker method over the single marker method (Figures omitted). In particular, we found an increase of five to thirty-five percent of true associated SNPs would be carried forward for validation if we were to go from 100 to 5,000 SNPs for validation, which is precisely the result observed experimentally.

3.3 Imputation.

To test the efficacy of our imputation method we performed the same experimental analysis described above, including a list of SNPs to be imputed. We imputed only SNPs from the HapMap
that had an LD value of at least 0.8 with an observed SNP on one of the two Illumina platforms (450S Duo and 550K v3). Additionally, we used the same filters previously stated and only used observed SNPs to impute if the observed SNPs had a true rank from genotypes as described above. In Supplementary Figure 3, we see Evaluation Metrics 1 and 2 in graphs (A and C). We compare four methods, a baseline where we perform no imputation as in the previous analyses and three methods where the minimum number of proxies required for imputation is one, two and three respectively. For at least one, two and three minimum required proxies we imputed 748,348, 544,041 and 424,314 SNPs respectively. As the minimum number of proxies required are increased, our imputation method performs slightly better and approaches the results from if we directly measure the given SNP (see Supplementary Figure 3 graphs A and C). This is expected since the SNPs imputed are in strong LD with our unobserved SNP and the number of indirect observations increases as the minimum number of proxies required increases. For some imputed SNPs in stretches of high LD, we see over a hundred proxies, opening the possibility to a great deal of information to be recovered as well as a problem of overfitting. Nevertheless, the imputation method achieves a high rank correlation between the true ranks and imputed ranks, and maintains a large number of truly associated SNPs within the Rank Threshold.

4 DISCUSSION

In this article, we developed theoretically, and demonstrated through simulation and experimentally a multimarker analysis method that improves the power of pooling-based genome-wide association studies. We first formalized a model for pooling-based studies with errors and gave a basic description of a suitable test statistic for both individual genotyping and pooling-based studies. We then tested this model using experimental data on multiple platforms, including the Illumina 550 v3, Illumina 450S Duo and Affymetrix 5.0 microarrays validating our results using simulations.

Perhaps more importantly, we demonstrate that an approach for combining Affymetrix and Illumina data is feasible and improves the assessment of the association significance noticeably. Combining platforms increases our genomic coverage as well as giving more measurements for those SNPs where the platforms intersect, with the number of SNPs measured in our experiments after combining platforms totaling 1,333,631 SNPs. Potentially, the number of truly associated SNPs or true positives selected for a second stage of validation can be increased by five to thirty-five percent when the number of SNPs to be carried forward is under 5,000 SNPs. Additionally, 100% of the truly associated SNPs are carried forward if the observed top 2,500 SNPs are chosen for validation. This percentage reduces to 90% if the observed top 1,000 SNPs are chosen for validation. In our analysis, we examined only those SNPs with at least one other SNP in LD. Through this criterion, we are able to show that using the LD information improves the accuracy of assessing the significance of SNPs and will continue to improve as denser microarray technologies become available. The resulting increase in the number of SNPs measured will give rise to more pair wise correlations between SNPs and because our method takes advantage of this increase in LD it will perform better as a result. Additionally, this new ability permits new considerations when designing a pooling-based study, namely that if arrays from one platform are significantly noisier than another platform, we could run more replicates on the noisier platform to compensate for the difference in noise. It is clear from our theoretical framework that increasing the number of replicates and increasing the number of probes per SNP reduces noise associated with pooling. In this case individual genotype data is not available, using the presented multimarker method could be used to improve the results when only allele counts or allele frequencies are present, thereby extending the utility of this method.

Finally, a novel application of our method is to impute the significance of association for unobserved SNPs. When an unobserved SNP is imputed, we essentially do not gain any more information since we are not gathering more observations. However, when an unobserved SNP has more than one proxy, we are able to increase the accuracy of imputation. This type of imputation is a good tool to evaluate associations within a specific region since we can use the imputed SNP A as a bridge between neighboring SNPs of SNP A. Nevertheless, as the number of SNPs measured increases, so will the number of proxies thereby increasing the accuracy of our method when assessing strength of association for an imputed SNP.

With emerging technologies from Affymetrix and Illumina having greater than one million SNPs, we hope to gain considerable power from the increased LD when combining data from both platforms. Additionally, our method uses linkage disequilibrium without considering the underlying haplotype structure. It is feasible to adapt current haplotype-based methods such as WHAP [24] to also increase power in pooling-based studies[17]. A strength of this LD based approach is that a correlation measure (r²p) is derived that describes the information content lost by pooling when compared to individual genotyping. This measure is similar to the power lost in individual genotyping when we do not observe directly the causal SNP. Consequently, it is theoretically straightforward to combine both measures in a hybrid multimarker statistic. This method will be particularly powerful where the phased data is not possible or not reliable. Frequently, this may be the case when HapMap populations are not used, or are seen as underpowered in lieu of larger LD databases derived from case-control association studies. Regardless of approach, the increasing densities present on a number of SNP microarray platforms the accuracy and utility of our method will only improve bringing wider adoption of pooling-based GWA to give a cost-effective alternative to individual genotyping with minimal loss in power.

ACKNOWLEDGEMENTS

We wish to provide acknowledgement of funding from NIH (1I24MS-43581), the Stardust foundation (DWC, WT), and the University of California Systemwide Biotechnology Research & Education Program GREAT Training Grant 2007-10 (NH).

This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk.

REFERENCES

1. Hanson, R.L., et al., A Potential Locus For End-Stage Renal Disease in Type 2 Diabetes Identified by a Pooling-Based Genome-Wide Association Study. Diabetes, 2007. In Press, Accepted Dec.06.


