Published in final edited form as:
Genet Epidemiol. 2010 September ; 34(6): 603–612. doi:10.1002/gepi.20517.

Pooled versus Individual Genotyping in a Breast Cancer Genome-wide Association Study

Ying Huang[1], **David A. Hinds**[2], **Lihong Qi**[3], and **Ross L. Prentice**[1,†]

[1]Fred Hutchinson Cancer Research Center Public Health Sciences, 1100 Fairview Avenue N., M2-B500, Seattle, WA 98109-1024

[2]Perlegen Sciences Inc., Mountain View, CA

[3]University of California Davis, School of Medicine, Department of Public Health Sciences, Division of Biostatistics, Davis, CA, 95616

SUMMARY

We examine the measurement properties of pooled DNA odds ratio estimates for 7357 single nucleotide polymorphisms (SNPs) genotyped in a genome-wide association study of postmenopausal breast cancer. This study involved DNA pools formed from 125 cases or 125 matched controls. Individual genotyping for these SNPs subsequently came available for a substantial majority of women included in seven pool pairs, providing the opportunity for a comparison of pooled DNA and individual odds ratio estimates and their variances. We find that the 'per minor allele' odds ratio estimates from the pooled DNA comparisons agree fairly well with those from individual genotyping. Furthermore, the log-odds ratio variance estimates support a pooled DNA measurement model that we previously described, though with somewhat greater extra-binomial variation than was hypothesized in project design. Implications for the role of pooled DNA comparisons in the future genetic epidemiology research agenda are discussed.

Keywords

Breast Cancer; Case-Control Studies; DNA pooling; GWAS; Odds Ratio; SNP

1. Introduction

Modern microarray technology has allowed genome-wide association studies (GWAS) among correlated individuals to emerge as a powerful tool for the detection of common disease susceptibility loci, even when the associations are weak. The early single nucleotide polymorphisms (SNP) platforms, however, entailed considerable cost. For example, a study of 500,000 SNPs for 1000 cases of a disease and 1000 controls, at a genotyping cost of $0.01 per SNP, projects a genotyping cost of $10 million. Hence, there was considerable interest in applying microarray technologies to pools formed from equimolar amounts of DNA from a possibly large number of cases or controls, and assessing evidence for association in a cost-efficient manner by comparison of SNP allele frequencies in case versus control pools (e.g., Bansal et al., 2002; Sham et al., 2002; Hinds et al., 2004). Compared to genotyping individual samples, pooled DNA comparisons may involve additional sources of variation due to pool construction (i.e., variation in DNA amounts among individuals contributing to the pool) or due to microarray measurement error (i.e.,

[†]*Corresponding author's address:* rprentic@fhcrc.org.

variation in estimation of luminescence intensities for the SNP alleles following hybridization of the target specimen with corresponding labeled probes). The magnitude of the additional variation may depend on various factors, including measurement platform and pool size (e.g., Barrett et al., 2002; Downes et al., 2004). Moreover, when pooled DNA allele frequencies are assessed, the minor and major alleles for a SNP may not be amplified to the same extent, resulting in a SNP-specific distortion that can bias allele frequency estimates and case versus control comparisons (e.g., Hoogendoorn et al., 2000; LeHellard et al., 2002; Moskvina et al., 2005). A common approach to tackling this issue was the separate estimation of SNP-specific distortion factors, with subsequent adjustment of allele frequency estimates (Hoogendoorn et al., 2000: Vissher and LeHellard, 2003; Moskvina et al., 2005).

Within the past few years SNP genotyping costs have dropped dramatically, to the point that individual genotyping costs for the study design mentioned above may now be in the vicinity of $500,000. In spite of these decreasing costs, the research community has retained interest in pooled DNA research strategies. For example, Abraham et al. (2008) report the ability to replicate known associations and to identify novel SNP associations for late-onset Alzheimers disease from pooled DNA comparisons. These authors write that individual 'GWA studies are expensive, generally restricting this type of work to groups or consortia with substantial funding for that purpose'. Similarly, Shifman et al. (2008) report a sex-specific SNP association with schizophrenia, based on a GWAS with DNA pooling at the first stage. Bossé et al. (2009) note the success of GWAS for complex diseases, and comment that the 'one persistent major hurdle is the cost of those studies'. These authors go on to confirm the ability of pooled DNA comparisons to replicate established SNP associations for type 2 diabetes, and to yield a suitably enriched set of SNPs for further evaluation in subsequent study stages.

Our Women's Health Initiative (WHI Study Group, 1998) carried out a GWAS of invasive breast cancer that involved about 360,000 SNPs and 8 pool pairs, with pools formed from equal amounts of DNA from 125 women who developed breast cancer during WHI Observational Study follow-up and from a corresponding 125 pair matched controls (Prentice and Qi, 2006). About 4000 highly ranked SNPs for breast cancer association, and about 6000 SNPs from other sources, primarily from the individual-level breast cancer GWAS in the National Cancer Institute's Cancer Genetic Markers of Susceptibility (CGEMS) program (e.g., Hunter et al., 2007), were subsequently genotyped for nearly 2000 breast cancer cases arising in the follow-up the WHI Clinical Trial cohort and for pair-matched controls. This study demonstrated associations between SNPs in intron 2 of the fibroblast growth factor receptor two (FGFR2) gene with breast cancer risk, including associations for SNPs selected from the pooled GWAS. It also yielded evidence of interaction of FGFR2 SNPs with the effect of hormonal and dietary interventions on breast cancer risk (Prentice et al., 2009; Prentice et al., 2010).

In designing the WHI GWAS, it was noted (Prentice and Qi, 2006) that the odds ratio was invariant to differential allelic amplification for a SNP, provided the distortion was common to cases and controls. Hence, planned association analyses were based on odds ratio estimates. The specification of equal sized pools led to a simple log-odds ratio estimator for each SNP, with a corresponding empirical variance estimator that incorporated allele frequency variation among study subjects, as well as measurement error related to both pool formation and array error. A statistical model was specified for the log-odds ratio variance estimate, and a model parameter was identified that controlled the measurement error variance and hence the comparative efficiency properties of the pooled estimator.

A substantial majority of the 1000 breast cancer cases and 1000 controls included in the WHI pooled GWAS were subsequently included in the first replication stage of the CGEMS

breast cancer GWAS, which included individual genotyping for a large number of overlapping SNPs.

Here we compare odds ratio estimates from pooled genotyping with those from individual genotyping, among subjects genotyped in the first stage of WHI GWAS and in CGEMS. Our objectives are to: (i) study the correlation of estimates between pooled DNA and individual genotyping; (ii) assess the validity of the log-odds ratio variance framework specified in Prentice and Qi (2006) and evaluate the magnitude of additional measurement error due to pooling, and (iii) compare the pooling method and individual genotyping methods for the identification of associations with established breast cancer susceptibility SNPs, based on data from these WHI studies.

2. Method

2.1 Study Cohort and Genotyping

In the first stage of the WHI GWAS, 1000 breast cancer cases from the WHI Observational Study (WHI,1998) (OS) were matched one-to-one to OS controls on baseline age, enrollment date, race/ethnicity, and hysterectomy status. Genotyping was then conducted in eight case pools and eight matched control pools, each of size 125, using Perlegen's 360,000 tag-SNP set. In the CGEMS first replication stage, 24,909 SNPs were individually genotyped in 4,547 cases and 4,434 controls, among which 2395 cases and 2410 controls were selected from WHI Observational Study cohort. Overall there were 7357 SNPs genotyped in both cohorts for 1493 subjects. Table 1 provides the number of individuals genotyped in CGEMS among each case or control pool from WHI GWAS stage 1. The analysis in this paper is based on the first 7 case pools and their matching control pools where a substantial majority of the 125 subjects (who contribute to pooled DNA) were individually genotyped. Comparisons will be carried out for 7235 SNPs for which at least one pool pair has allele frequency estimates passing Perlegens pooled DNA quality control criteria. The numbers of individually genotyped study subjects range from 94 to115 across the 14 pools. The eighth pool pair in the WHI GWAS was comprised primarily of minority women, while CGEMS restricted their genotyping to white women.

Individual DNA in CGEMS was genotyped using the Human Hap500 Infinium Assay (Illumina) array. The details were reported in Hunter et al. (2007). Pooled DNA in WHI GWAS was genotyped using high-density oligonucleotide arrays. Details about genotyping and algorithms for determining pooled allele frequency estimates were reported in Hinds et al. (2004). Two quality control metrics: conformance and signal to background ratio, were used for SNPs in pooled data. Conformance is defined as the fraction that perfect-match feature was brighter compared to mismatch feature, and signal to background ratio is calculated from intensity measurement. Only SNPs with conformance greater than 0.9 and signal to background ratio greater than 1.5 were included.

2.2 Hypothesis Testing

We consider the setting where cases and their matching controls are each divided into m pools of size n. Let i indicate case-control status, $i = 1$ for cases and $i = 2$ for controls. Let j be the pool indicator, $j = 1, \ldots, m$. For a particular SNP of interest, let

- p_i be the minor allele frequency among cases (i=1) or controls (i=2);
- $\hat{p}_{ij}$ be the minor allele frequency estimates obtained from the j^{th} case ($i = 1$) or control ($i = 2$) pool;
- $\tilde{p}_{ij}$ be the empirical minor allele frequency estimate based on individual level data for those subjects involved in the j^{th} case ($i = 1$) or control ($i = 2$) pool;

- $Y_{ij} = \log\{\tilde{p}_{ij}/(1 - \hat{p}_{ij})\}$ be the logit (log odds) estimated from the pooled DNA;
- $Y_{ij}^{\star} = \log\{\tilde{p}_{ij}/(1 - \tilde{p}_{ij})\}$ be the logit estimate based on individual level data.

Note that $\hat{p}_{ij}$ and Y_{ij} may be biased by differential allelic amplification, but not the corresponding pooled data log odds ratio estimator. Following Prentice and Qi (2006), let $\hat{\beta}_j = Y_{1j} - Y_{2j}$ be the log-odds ratio estimate for the j^{th} case-control pool pair, and let $\widehat{\beta} = \sum_{j=1}^{m} \widehat{\beta}_j / m$ be the average log-odds ratio. The empirical variance estimate of $\hat{\beta}$ can be calculated as $s^2 = \sum_{j=1}^{m} \left(\widehat{\beta}_j - \widehat{\beta}\right)^2 / \{m(m-1)\}$. The test statistic $\hat{\beta}/s$ is then compared to a t_{m-1} distribution for testing of the null hypothesis of no association between the SNP and the breast cancer risk. To obtain the odds ratio for a SNP based on individual level data, a logistic regression model was applied, adjusting for breast cancer risk factors including (log transformed) Gail 5-yr breast cancer risk score, previous hormone use (indicators for < 5, 5 − 10, and > 10 years for both estrogen and estrogen plus progestin), and (log transformed) body mass index. Variables used for matching controls to cases in control selection are also included in the regression model. In addition, eigenvectors from the first four principal components from correlation analysis of the CGEMS replication study genotype data are included to account for any effect of population stratification. We compare pooled versus individual genotyping with respect to log odds within each case or control pool, log odds ratio for each case and matching control pool pair, mean log odds ratio averaged overall all pools, and test statistics for testing equal minor allele frequencies between cases and controls.

2.3 Assessing Pooling Error

Following Prentice and Qi (2006), we assume log odds based on pooled data can be written as the log odds based on individual level data plus an independent pooling error, i.e., $Y_i = Y_i^{\star} + \varepsilon_i$. This leads to $\mathrm{var} Y_i = V_i + \sigma_i^2$, where

$$V_i = \mathrm{var} Y_i^{\star} \simeq \{2np_i(1 - p_i)\}^{-1} \tag{1}$$

from binomial sampling theory under an additive logit model. Moreover, we assume that $\sigma_i^2 = \mathrm{var}\varepsilon_i$, the additional variance that arises from the use of pooled DNA, takes the form $\Delta^2\{p_i(1 - p_i)\}^{-1}$.

We estimate Δ^2 by estimating variance of $Y_i - Y_i^{\star}$ empirically based on the 7 case or control pools, and then multiplying the result by $\bar{\tilde{p}}_{i.}(1 - \bar{\tilde{p}}_{i.})$, where $\bar{\tilde{p}}_{i.} = \sum_{j=1}^{m} \tilde{p}_{ij}/m$ is a consistent estimate of p_i.

We test equality of Δ^2 between cases and controls using a permutation test. For each SNP, the ratio of Δ^2 estimate for cases relative to that for controls is calculated and compared with its null distribution, which is generated by permuting the case and control labels of the 7 case-control pool pairs.

2.4 Comparing Pooling vs Individual Genotyping for Established Breast Susceptibility SNPs

We compare results of pooled and individual odds ratio estimates for established breast cancer SNPs from the literature. Based on Akaike information criterion (AIC) from logistic regression, we classify established SNPs according to genetic model (additive, dominant,

recessive) using individual-level data, and examine the correspondence between per allele odds ratio estimates from individual and pooled genotyping, for SNPs classified under each of the three genetic models.

3. Results

For the 7357 SNPs measured in both the WHI GWAS and CGEMS replication study, Table 2 presents the mean and standard deviation of their minor allele frequency 'estimates', for each case or control pool. The pool-specific allele frequency for individual level data was estimated by the mean allele frequency of subjects included in a particular pool. The minor allele frequency estimates presented based on pooled data are those that ignore the unequal amplification issue, leading to SNP-specific bias that is evident in the mean allele frequency estimate across SNPs. On average, the minor allele frequency estimates and their variability for the set of SNPs considered appear to be larger based on pooled data compared to individual level data. This pattern is consistent between cases and controls and across different study pools.

Also shown in Table 2 are mean and standard deviation of the log odds ratio estimates based on the pooled and individual genotyped data. It appears that the overall bias in allele frequency estimates has little impact on log odds ratio estimation. Also, as expected, estimates based on pooled data tend to have somewhat greater spread than do those based on individual level data.

Within each matched case-control pool, we investigated the relationship between log odds ratio estimates from the two genotyping methods using a scatterplot (Figure 1). Overall the two types of estimates seem to agree fairly well with each other. The added noise due to pooling, nevertheless, is also apparent from the elliptical shape evident in the plot. The Pearson correlation between the two estimates based on the 7357 SNPs ranges from 0.51 to 0.63 across the 7 matched case-control pools.

Figure 2(a) displays a scatterplot of pooled versus individual estimate for log odds ratio averaged across all pools. The log odds ratio on the horizontal coordinate was obtained by fitting an ordinary logistic regression model to the individual level data adjusting for other risk factors as explained in Section 2.2. The log odds ratio on the vertical coordinate was the average of $\text{logit}(\tilde{p}_1) - \text{logit}(\tilde{p}_2)$ over the 7 matched case-control pools. Again, the estimate from pooling method seems to be a noisy but approximately unbiased version of the corresponding estimates from individual genotyping. Similar association between the two methods can be observed regarding the test statistics (Z-value) for the hypothesis of equal minor allele frequencies between cases and controls (Figure 2(b)).

Estimation of the additional pooling error was based on 6601 SNPs with complete values of Y_i and $Y_i^\star$ in all 7 case or control pools. The empirical variance estimate for $Y_i - Y_i^\star$ is plotted against the estimate of $\{p_i(1-p_i)\}^{-1}$ for cases and controls separately in Figures 3(a)(b). The relationship between $\{p_i(1-p_i)\}^{-1}$ and $\text{var}(Y_i - Y_i^\star)$ does not appear to deviate substantially from linear as can be viewed by the loess smooth curves. Figures 3(c) and (d) show the scatterplot of Δ^2 estimate, the coefficient related to extra noise due to pooling, versus minor allele frequency estimate for cases and controls separately. The corresponding loess curves are fairly flat, supporting the invariance of Δ^2 to minor allele frequency as hypothesized in Prentice and Qi (2006). Figures 4(a) and(b) display histograms of Δ estimates in cases and controls. The average of Δ are 0.057 and 0.062 based on cases and controls respectively, after removing extreme outliers (outside of inter-quartile range by three times the length of inter-quantile range). The first three quartiles of Δ estimate are 0.041, 0.054, and are 0.070 for cases, and 0.046, 0.058, and 0.074 for controls. Equality of Δ^2 between cases and

controls was tested based on 50,000 permutations. The pooling error does not seem to differ between cases and controls, as suggested by the uniformly distributed p-values between 0 and 1 (Figure 4(c)).

The imperfect match of study subjects between the pooled data and the individual level data could potentially reduce the correlation between the (pool-specific) individual and pooled logodds estimates. We conducted simulation studies to assess the subsequent impact on Δ estimate, for minor allele frequency p varying from 0.05 to 0.5 and Δ varying from 0.001 to 0.07, and for cases and controls separately. Specifically, for each case or control pool, we generate independent binary allele data of size 2×125 with probability p, assuming $2 \times n_i$ alleles are randomly selected from this pool for individual genotyping, where n_i is the actual number of subjects in the CGEMS data. Figure 5 presents plots of average Δ estimate based on 5,000 Monte-Carlo simulations versus true Δ. Having fewer subjects for individual genotyping appears to result in some inflation of Δ estimate, especially for small Δ. For relatively large Δ estimate as observed in our analysis, the inflation is relatively small. The magnitude of the minor allele frequency has minimal impact on this trend. Corresponding to average Δ estimate of 0.04–0.07, the actual Δ value falls into the range of 0.03–0.068 for cases and 0.02–0.064 for controls. Corresponding to an average Δ estimate of 0.06, the actual Δ value is around 0.055–0.057 for cases and 0.050–0.052 for controls.

Figure 6 shows relative efficiency of pooled genotyping vs individual genotyping for Δ ranging from 0 to 0.06 and various pool sizes, calculated as the variance of log odds estimate based on individual level data divided by the variance of log odds estimate based on the pooled data, i.e.,

$$\frac{\mathrm{var}(Y_i^\star)}{\mathrm{var}(Y_i)} = \frac{1/(2n)}{1/(2n)+\Delta^2},$$

where n is the pool size. One can observe a loss of efficiency with increasing Δ, particularly for large pools. For example, with $n = 125$, the loss of efficiency is 2.4% for $\Delta = 0.01$ and 9.1% for $\Delta = 0.02$, but a substantial 47.4% for $\Delta = 0.06$.

For the 12 established breast cancer susceptibility SNPs (Easton et al., 2007; Hunter et al., 2007; Ahmed et al., 2009; Thomas et al., 2009) included in the study set, we compared the results based on pooled and individual genotyping. For each SNP, logistic regression models assuming additive, dominant, or recessive genetic effect are applied to the individual level data, and the SNP is classified into the one of the three genetic models with minimum AIC. For each of the 12 SNPs, Table 3 presents the per allele odds ratio estimate, and the p-value for testing SNP association with breast cancer, assuming an additive allele effects on the log odds. Results are presented for pooled data, individual level data, as well as based on literature. In general, regardless of the underlying genetic model into which the SNP is classified, the two genotyping methods lead to fairly comparable per-allele odds ratio estimates, which are similar to those reported in the literature. Genotyping individual level DNA is more efficient in identifying those established SNPs, demonstrated by the smaller p-value for 10 out the 12 SNPs. Using 0.05 as a significance cut-off for p-value, the individual genotyping selects three of the establish SNPs while the pooling method identifies two. P-values of the likelihood ratio test with 2df based on individual level data are also presented. The test is significant at 0.05 level for one SNP classified as dominant (rs981782), which was not significant by either the individual or pooled data test under the assumption of an additive allele effect.

An advantage of the WHI GWAS pooling study is the construction of multiple pools from different subjects, which allows estimation of the empirical variance of log odds ratio for each SNP separately. While this provides robustness that is highly desirable in GWAS, variance estimate based on 7 case-control pools might be unstable. To address this issue we explored an alternative variance estimate using additional pooled data from WHI-perlegen GWAS. Specifically, WHI GWAS collected data on 1000 cases and their matching controls for two other diseases as well in addition to breast cancer: coronary heart disease (CHD) and stroke. For each disease type, 8 case pool and 8 control pool were constructed with size 125 each. CHD and stroke studies each has 6 matched case-control pooled comprised mostly of Caucasian Women. Restricting analysis to those pools, we derived variance estimate for the log-odds ratio by adding the three empirical variance estimates from the three diseases in the WHI GWAS, thereby yielding a variance estimate with up to 16 df. The corresponding pooling Z-value is highly correlated with the pooling Z-value using variance estimate from breast cancer case-control pools alone, with a correlation of 0.97. Figure 7 shows scatterplot of the modified Z-values with Z-values from individual data. The correspondence between Z-values based on pooled data and Z-values based on individual data appears similar whether the variance of log-odds ratio is estimated solely from breast cancer pools (Figure 2(b)) or from all three diseases (Figure 7). The correlation between Z-values varies from 0.65 in Figure 2(b) to 0.66 in Figure 7. Furthermore, to assess the impact of the increase in degree of freedom on identifying an “true” association, in Table 3, for the 12 established breast cancer susceptibility SNPs we added their p-values using pooled data with variance estimated from all three diseases together. Compared with the analysis based on breast cancer pools alone, significance of p-values for those established SNPs appears in general to be more compatible with that using individual data. All three SNPs identified by individual genotyping at significance level 0.05 are picked up by the pooling method, suggesting a power gain with the increase in pool numbers.

4. Discussion

In this manuscript we compared a pooled with an individual genotyping method for the identification of disease susceptible SNPs based on case-control studies among highly overlapping set of study subjects between two breast cancer studies. Particularly, we evaluated the log odds ratio estimators proposed by Prentice and Qi (2006) by comparison with corresponding estimators based on a traditional logistic regression analysis of highly reliable individual genotype data. Unlike other estimators based on absolute allele frequency differences, this pooled DNA odds ratio estimator does not require estimation of a SNP-specific distortion factor, which can only be obtained from individual level data. Thus this method is easier to implement than are other available pooled genotyping estimation procedures. The close correspondence between the pooled and individual odds ratio estimators and test statistics supports the pooling technique as a cost efficient approach for the initial phase of GWAS. It is also clear that a considerably larger sample size may be needed, depending in part on pool sizes, to overcome the additional noise in these odds ratio estimators due to pooling.

In the framework of Prentice and Qi (2006), we investigated a pooling measurement error coefficient Δ, the magnitude of which reflects the additional noise due to pooling. Earlier simulations entertained Δ value of 0.01 and 0.02, obtained based on a small set of 16 SNPs (Mohlke et al., 2002). In this study we were able to obtain a better view of the distribution of Δ based on a much larger sets of SNPs. Our estimate of Δ appears to be robust to minor allele frequency estimate, supporting the framework in Prentice and Qi (2006). However, the estimate of Δ from the actual dataset appears to be larger than was hypothesized at the design stage. Knowledge regarding pooling error is important to study design and to an assessment of the cost efficiency gained through pooling. For example, Zhao and Wang

(2009) argue that to achieve optimal cost-efficiency, smaller pool size should be used with larger pooling error. On the other hand, the genotyping costs for a study will be approximately inversely proportional to pool size, so there is incentive to retain a pool size as large as practical without incurring undue efficiency loss, relative to a study with individual genotyping. MacGregor et al (2007, 2008) argues that most variation that arises from pooling is due to array measurement error, rather than pool formation. Abraham et al. (2008) follow this perspective in using single pools of Alzheimer's disease cases and controls, which they hybridize multiple times. Shifman et al. (2008) and Bossé et al. (2009) followed a similar strategy. If pool construction errors are negligible, then q independent hybridizations from a pool should lead to a reduction in the additional variance due to pooling by a factor of q, resulting in an efficient study design.

Note that our estimate of Δ is a combination of array measurement error and pool construction error at the log-odds scale. If one ignores the unequal allelic amplification issue, then the allele frequency estimate from pooled data has asymptotic variance given by $p(1-p)(1/(2n)+\Delta^2)$ based on the delta method. Consider a Δ value of 0.06, for allele frequency p varying from 0.1 to 0.5, the pooling error variance for allele frequency estimates using the Perlegen high-density oligonucleotide array falls into the range of 0.00068–0.0019, similar in magnitude to the pooling error reported in Macgregor (2007) using Affymatrix HindIII array.

Craig et al. (2009) note that 'many groups researching diseases in developing nations, or traits of perceived lesser clinical significance, have been unable to pursue GWAS methodology due to the high budgets required'. They go on to write that even pooled DNA strategies typically involve laborious DNA extraction for individual cases and controls, and sophisticated DNA quantitation procedures for pool construction. Hence, they proposed a research strategy involving pooling equal volumes of whole blood prior to DNA extraction, with 'potential to reduce GWAS costs by several orders of magnitude'. Since this strategy may entail considerable variation in the amount of DNA contributed by pool members, it may then be desirable to include smaller or intermediate sized pools to control pool construction measurement error influences, even if there are multiple hybridizations per pool.

Pooled DNA comparisons can partially control for confounding factors by matching cases and controls on race/ethnicity and other factors. Moreover, leads from pooled DNA GWAS, as with individually genotyped GWAS, will nearly always require individual-level replication in independent cases and controls where issues of population stratification and confounding can be addressed in a customary fashion.

When estimating allele odds ratio using logistic regression anlaysis of individual level data, we presented results from models adjusting for common risk factors besides factors used to match controls to cases, which is common practice in GWAS studies. Considering the fact that these risk factors cannot be accounted for in pooled data, we also examined logistic regression models with adjustment for matching factors only. The impact of adjusting for additional risk factors turned out to be minimal in the comparison between pooled and individual level data.

A limitation of this analysis arises from the imperfect match of study subjects between the pooled data and the individual level data, which likely somewhat reduces the correlation between the pool-specific individual and pooled log-odds ratio estimates, and leads to somewhat inflated estimates of Δ. The inclusion rate ranges from 79% to 92% in case pools and 75% to 81% in control pools. There is no apparent reason, nevertheless, to believe that there is a systematic difference between subjects included and not included in the subsets of

each pool for which individual genotyping data is available. Also, based on simulations mimicking the study setting, we were able to entertain the impact of having somewhat fewer subjects individually genotyped on the estimation of Δ and found it to be modest.

In summary, our analyses indicate a good correspondence between odds ratio estimates from individual genotyping, and those from genotyping pools formed from equal amounts of DNA from 125 individuals. The role of pooled DNA comparisons in the translation of modern genotyping capabilities to the assessment of genetic aspects for a wide variety of diseases and traits has yet to be established. The notion that reliable GWAS comparisons may be able to be conducted at such a modest cost that such studies become practical for single investigators drawing on cohort study resources is sufficiently intriguing and important to encourage the further development of pooling strategies and related study designs.

Acknowledgments

This work was partially supported by National Institute of Health awards CA53996 and HHSN268200764314C.

REFERENCES

Ahmed S, Thomas G, Ghoussaini M, et al. Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. Nat Genet 2009;41:585–590. [PubMed: 19330027]

Abraham R, Moskvina V, Sims R, Hollingworth P, Morgan A, Georgieva L, Dowzell K, Cichon S, Hillmer AM, O'Donovan MC. A genome-wide association study for late-onset Alzheimer's disease using DNA pooling. BMC Med Genomics 2008;1:44. [PubMed: 18823527]

Bansal A, van der Boorn D, Kammerer S, Honisch C, Adam G, Cantor CR, Kleyn P, Braun A. Association testing by DNA pooling: An effective initial screen. PNAS 2002;99:16871–16874. [PubMed: 12475937]

Barratt BJ, Payne F, Rance HE, Nutland S, Todd JA, Clayton DG. Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. Ann. Hum. Genet 2002;66:393–405. [PubMed: 12485472]

Bossé Y, Bacot F, Montpetit A, Rung J, Qu H, Engert JC, Polychronakos C, Hudson TJ, Froguel P, Sladek R, Desrosiers M. Identification of susceptibility genes for complex diseases using pooling-based genome-wide association scans. Hum Genet 2009;125:305–318. [PubMed: 19184112]

Benjamin EJ, Dupuis J, Larson MG, Lunetta KL, Booth SL, Govindaraju DR, Kathiresan S, Keaney JFJ, Keyes MJ, Lin L, Meigs JB, Robins SJ, Rong J, Schnabel R, Vita JA, Wang TJ, Wilson PW, Wolf PA, Vasan RS. Genome-wide association with select biomarker traits in the Framingham Heart Study. BMC Med Genet 2007;8 Suppl 1:S11. [PubMed: 17903293]

Craig JE, Hewitt AW, McMellon AE, Henders AK, Ma L, Wallace L, Sharma S, Burdon KP, Visscher PM, Montgomery GW, MacGregor S. Rapid inexpensive genome-wide association using pooled whole blood. Genome Research 2009;19:2075–2080. [PubMed: 19801603]

Downes K, Barratt BJ, Akan P, Bumpstead SJ, Taylor SD, Clayton DG, Deloukas P. SNP allele frequency estimation in DNA pools and variance components analysis. Biotechniques 2004;36(5): 840–845. [PubMed: 15152604]

Easton DF, Pooley KA, Dunning AM, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 2007;447:1087–1093. [PubMed: 17529967]

Hinds DA, Seymour AB, Durham K, Banerjee P, Ballinger DG, Milos PM, Cox DR, Thompson JF, Frazer KA. Application of pooled genotyping to scan candidate regions for association with HDL cholesterol levels. Human Genomics 2004;1(6):421–434. [PubMed: 15606997]

Hunter DJ, Kraft P, Jacobs KB, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat. Genet 2007;39:870–874. [PubMed: 17529973]

Hoogendoorn B, Norton N, Kirov G, Williams N, Hamshere ML, Spurlock G, Austin J, Stephens MK, Buckland PR, Owen MJ, ODonovan MC. Cheap, accurate and rapid allele frequency estimation of

single nucleotide polymorphisms by primer extension and DHPLC in DNA pools. Human Genetics 2000;107:488–493. [PubMed: 11140947]

Le Hellard S, Ballereau SJ, Visscher PM, Torrance HS, Pinson J, Morris SW, Thomson ML, Semple CAM, Muir WJ, Blackwood DHR, Porteous DJ, Evans KL. SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semiautomated method for data storage. Nucleic Acids Res 2002;30(74):1–10. [PubMed: 11752241]

MacGregor S. Most pooling variation in array-based DNA pooling is attributable to array error rather than pool construction error. Eur J Hum Genet 2007;15:501–504. [PubMed: 17264871]

MacGregor S, Zhao Z, Henders A, Martin NG, Montgomery GW, Visscher PM. Highly cost-efficient genome-wide association studies using DNA pools and dense SNP arrays. Nucleic Acids Research 2008;36(6):e35. [PubMed: 18276640]

Mohlke KL, Erdos MR, Scott LJ, Fingerlin TE, Jackson AU, Silander K, Pablo H, Boehnke M, Francis SC. High-throughput screening for evidence of association by using mass spectrometry genotyping on DNA pools. PNAS 2002;99(26):16928–16933. [PubMed: 12482934]

Moskvina V, Norton N, Williams N, Holmans P, Owen M, ODonovan M. Stream-lined analysis of pooled genotype data in SNP-based association studies. Genetic Epidemiology 2005;28:273–282. [PubMed: 15700279]

Prentice RL, Huang Y, Hinds DA, Peters U, Cox DR, Beilharz E, Chlebowski RT, Rossouw JE, Caan B, Ballinger DG. Variation in the FGFR2 gene and the effects of postmenopausal hormone therapy on invasive breast cancer. Cancer Epidemiol Biomarkers Prev 2009;18(11):3079–3085. [Epub 2009 Oct 27]. [PubMed: 19861516]

Prentice RL, Huang Y, Hinds DA, Peters U, Cox DR, Beilharz E, Chlebowski RT, Rossouw JE, Caan B, Ballinger DG. Variation in the FGFR2 gene and the effect of a low-fat dietary pattern on invasive breast cancer. Cancer Epidemiol Biomarkers Prev 2010;19(1):74–79. [PubMed: 20056625]

Prentice RL, Qi L. Aspects of the design and analysis of high-dimensional SNP studies for diseased risk estimation. Biostatistics 2006;7(3):339–354. [PubMed: 16443924]

Sham P, Bader JS, Craig I, ODonovan M, Owen M. DNA pooling: a tool for large-scale associations studies. Nat Rev Genet 2002;3:862–871. [PubMed: 12415316]

Shifman S, Johanesson M, Bronstein M, Chen SX, Collier DA, Craddock NJ, Kendler KS, Li T, O'Donovan M, O'Neill FA, Owen MJ, Walsh D, Weinberger DR, Sun C, Flint J, Darvasi A. Genome-wide association identifies a common variant in the reelin gene that increases the risk of schizophrenia only in women. PLoS Genet 2008;4(2):e28. [PubMed: 18282107]

Thomas G, Jacobs KB, Kraft P, et al. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1(*RAD51L1*). Nat Genet 2009;41:579–584. [PubMed: 19330030]

Visscher PM, Le Hellard S. Simple method to analyze SNP-based association studies using DNA pools. Genetic Epidemiology 2003;24:291–296. [PubMed: 12687646]

The Women's Health Initiative Study Group. Design of the Womens Health Initiative Clinical Trial and Observational Study. Control Clin Trials 1998;19(1):61–109. [PubMed: 9492970]

Zhao Y, Wang S. Optimal DNA pooling-based two-stage designs in case-control association studies. Hum Hered 2009;67:46–56. [PubMed: 18931509]

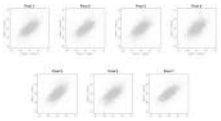

Figure 1.
Scatterplot of $\text{logit}(\tilde{p}_1)-\text{logit}(\tilde{p}_2)$ estimated from individual genotyping and $\text{logit}(\hat{p}_1)-\text{logit}(\hat{p}_2)$ from pooled data within each matched case-control pool.

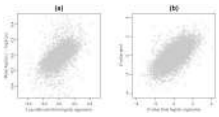

Figure 2.
Scatterplots of average log odds ratio estimate (a) and Z-value (b) based on pooled data versus those based on individual level data.

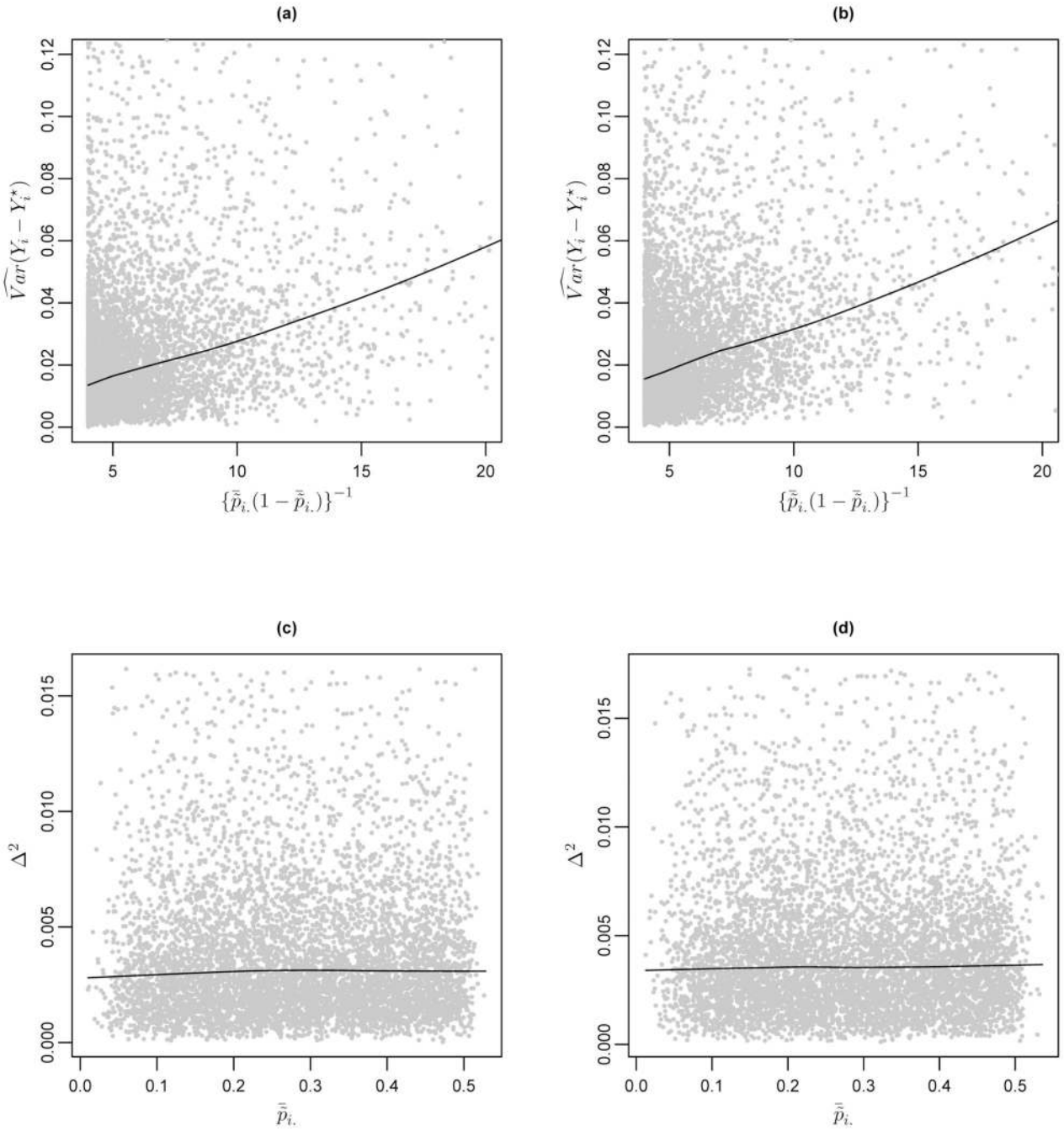

Figure 3.

Estimated variance of $Y_i - Y_i^\star$ versus estimates of $p_i(1 - p_i)$ for cases (a) and controls (b). Estimates of Δ^2 versus estimates of p_i for cases (c) and controls (d). Here $\bar{\tilde{p}}_{i.} = \sum_{j=1}^{m} \tilde{p}_{ij}/m$.

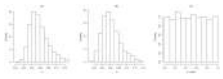

Figure 4.
Histogram of Δ value for cases (a) and controls (b), and histogram of the p-value for testing equal Δ^2 between cases and controls (c).

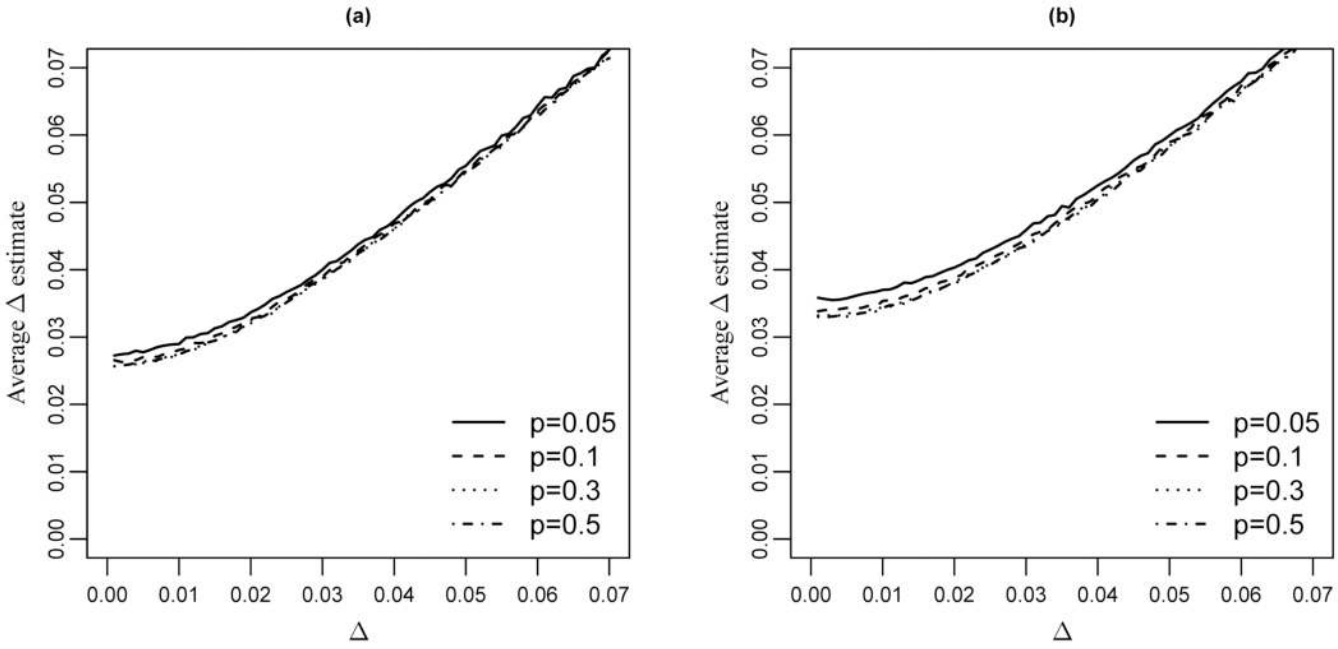

Figure 5.
Average Δ estimate versus 'true' Δ based on simulation studies for cases (a) and controls (b).

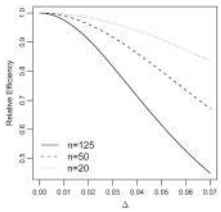

Figure 6.
Relative efficiency of pooled genotyping versus individual genotyping.

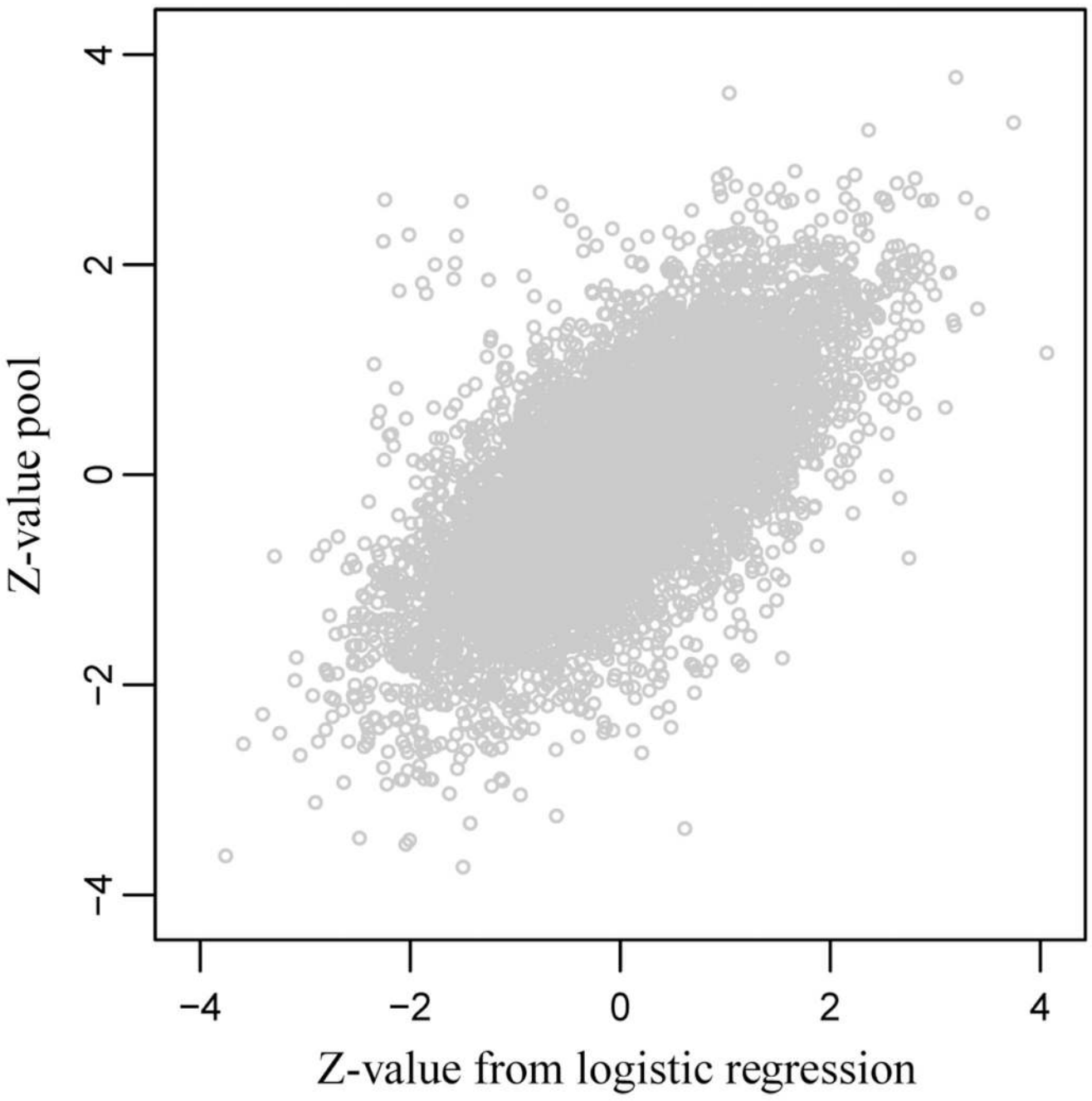

Figure 7.
Scatterplot of Z-value based on pooled data versus those based on individual level data. Here variance for log-odds ratio estimate from pooled data is based on case-control pools from all three diseases (breast cancer, CHD, and stroke).

Table 1

Number of cases and controls individually genotyped within each case or control pool.

Pool	1	2	3	4	5	6	7	8
Case	105	104	110	105	115	99	106	14
Control	96	98	98	94	101	95	96	7

Table 2

Summary measure of minor allele frequency estimates based on pooled and individual genotyped DNA for each case and control pool, and summary measure of log-odds ratio estimates based on pooled and individual genotyped DNA for each case and control pool pair.

	Case MAF				Control MAF			
	Pooled		Individual		Pooled		Individual	
Pool	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	0.340	0.165	0.280	0.130	0.340	0.165	0.279	0.129
2	0.340	0.166	0.280	0.129	0.341	0.165	0.280	0.129
3	0.344	0.165	0.279	0.129	0.345	0.158	0.280	0.130
4	0.339	0.162	0.279	0.129	0.338	0.164	0.279	0.130
5	0.339	0.167	0.280	0.129	0.337	0.168	0.279	0.130
6	0.339	0.167	0.279	0.129	0.344	0.163	0.280	0.130
7	0.344	0.157	0.280	0.129	0.343	0.158	0.280	0.129

	Log odds ratio			
	Pooled		Individual	
Pool	Mean	SD	mean	SD
1	−0.005	0.293	0.001	0.261
2	−0.009	0.269	0.003	0.254
3	−0.015	0.285	−0.002	0.257
4	0.007	0.297	0.005	0.259
5	0.012	0.289	0.009	0.249
6	−0.032	0.277	−0.005	0.261
7	0.001	0.244	−0.0005	0.259

Table 3

Comparative odds ratio estimators from pooled and individual genotyping for SNPs having an established breast cancer association.

rs#★	Allele★	Chr	MAF★	OR.E†	OR.I†(95% CI)	OR.P† (95% CI)	p.2df$	p.I$	p.P$	p.P3$
Additive										
rs13281615	G/A	8q24	0.41	1.08	1.06 (0.91,1.24)	1.00 (0.88,1.14)	0.730	0.430	0.977	0.976
rs2981582	A/G	10q26	0.39	1.26	1.23 (1.05,1.44)	1.16 (0.93,1.44)	0.027	0.009	0.153	0.045
rs3803662	A/G	16q12	0.27	1.28	1.27 (1.07,1.51)	1.26 (1.04,1.52)	0.020	0.006	0.024	0.007
rs6504950	A/G	17q22	0.27	0.95	1.06 (0.89,1.26)	0.95 (0.69,1.32)	0.809	0.517	0.734	0.691
Dominanant										
rs12443621	G/A	16q12	0.49	1.11	1.14 (0.97,1.33)	1.11 (0.95,1.30)	0.196	0.110	0.166	0.261
rs4973768	T/C	3p24	0.48	1.11	0.96 (0.83,1.12)	0.91 (0.73,1.13)	0.131	0.644	0.331	0.297
rs981782	C/A	5p12	0.46	0.96	0.93 (0.80,1.08)	0.98 (0.88,1.10)	0.021	0.345	0.716	0.766
Recessive										
rs2107425	T/C	11p15	0.30	0.96	1.09 (0.93,1.29)	1.06 (0.96,1.17)	0.268	0.299	0.205	0.295
rs3750817	T/C	10q26	0.41	0.78	0.86 (0.74,1.00)	0.92 (0.82,1.04)	0.083	0.052	0.135	0.130
rs3817198	C/T	11p15	0.33	1.07	1.06 (0.90,1.24)	1.03 (0.93,1.15)	0.215	0.472	0.514	0.596
rs4666451	A/G	2p	0.40	0.97	0.85 (0.73,0.99)	0.78 (0.62,0.97)	0.056	0.037	0.033	0.004
rs889312	C/A	5q11	0.29	1.13	1.09 (0.93,1.29)	1.07 (0.88,1.31)	0.224	0.296	0.421	0.352

★rs#: dbSNP rs number;

★Allele: Minor/Major Allele;

★MAF: minor allele frequency in the study population

†odds ratio per copy of minor allele assuming additive effect, based on literature (OR.E), based on individual genotyping (OR.I), based on pooled DNA (OR.P)

$p-value testing significant effect of a genotype, based on 2df test using individual data (p.2df), based on trend test using individual data (p.I), based on pooled DNA (p.P: variance of log odds ratiois estimated from breast cancer pools alone; p.P3: variance of log odds ratio is estimated from breast cancer, CHD, and stroke pools together.)