European Journal of Human Genetics (2007) 15, 501–504

www.nature.com/ejhg

SHORT REPORT

Most pooling variation in array-based DNA pooling is attributable to array error rather than pool construction error

Stuart Macgregor*,[1]

[1]*Genetic Epidemiology, Queensland Institute of Medical Research, Brisbane, Australia*

Genome-wide association (GWA) approaches are important in complex disease gene mapping studies but are often prohibitively expensive. Array-based DNA pooling has been shown to offer substantial cost savings compared with individual genotyping. This reduced cost potentially brings well-powered GWA studies well within the reach of most laboratories. The main factor, which affects the efficiency of pooling compared with individual genotyping is the magnitude of the pooling error variance. By examining variation between and within pools it is shown that most of the error associated with pooling is attributable to array variation not pooling construction variation (assuming the pools are not small and the pools are accurately constructed). With Affymetrix *Hind*III 50K arrays used here the array-specific variance is seven times the pooling construction variance. This has important implications for optimal study design for array-based pooling. Given carefully constructed pools, resources should be allocated to increasing the number of arrays per sample rather than to constructing multiple pools.
European Journal of Human Genetics (2007) **15,** 501–504. doi:10.1038/sj.ejhg.5201768; published online 31 January 2007

Keywords: DNA pooling; pooled DNA; microarray; genome-wide association

Introduction

Genome-wide association (GWA) is a popular technique for disease gene mapping of complex traits. The availability of microarrays has made GWA technically possible but it is prohibitively costly for many researchers. A cost efficient alternative to individual genotyping is DNA pooling,[1] an approach recently extended to use arrays.[2–4] With array-based pooling, well-powered GWA studies can be conducted at vastly reduced cost, bringing them well within the reach of most laboratories.[2] The primary factor which affects the efficiency of pooling compared with individual genotyping is the magnitude of the pooling variance. Appreciation of the sources of variation is critical to the efficient allocation of resources in terms of the number of arrays and the number of pools used.

Previously, Macgregor *et al*[2] presented pooling data using Affymetrix arrays but did not address the composition of the pooling variance. Here is shown that by examining variation between and within pools, it is possible to partition the variation into a component attributable to error on the arrays (ie, 'technical' error) and a component owing to errors in pooling construction. This demonstrates that most of the error in pooling is attributable to variation on the arrays and that the error introduced when pool are carefully constructed is of substantially less importance. For optimal efficiency, resources should be allocated in increasing the number of arrays per pool rather than constructing multiple pools.

*Correspondence: Dr S Macgregor, Genetic Epidemiology, Queensland Institute of Medical Research, Herston Road, Brisbane 4029, Australia. Tel: +61 7 3845 3563; Fax: +61 7 3362 0101;
E-mail: stuart.macgregor@qimr.edu.au
Received 4 October 2006; revised 16 November 2006; accepted 17 November 2006; published online 31 January 2007

Materials and methods

Data

Full details of the data used are given elsewhere.[2,5] In brief, genomic DNA was extracted (using the same method

throughout) from peripheral venous blood samples collected in the period 1997–2003. Two DNA pools (case and control) of 384 individuals were constructed by mixing equal amounts of adjusted DNA samples. Three Affymetrix Genechip *Hind*III arrays (56494 SNPs) were applied to each pool.

Statistical methods

Sources of error with pooling With pooling there are a number of sources of error. The sample frequency estimate, $\tilde{p}_a$, from pooled data can be written (cf. appendix 1 in Macgregor *et al*[2])

$$\begin{aligned}\tilde{p}_a &= \hat{p}_a + e_{pool\_array} + e_{pool\_construction} \\ &= p_a + e_b + e_{pool\_array} + e_{pool\_construction}\end{aligned}$$

where p_a is the true population frequency, $\hat{p}_a$ is the estimate of the frequency in that sample (this does not equal true population frequency, p_a, because of binomial sampling error), e_b is the binomial sampling error, e_{pool_array} is the error associated with estimating the frequency from the pool on an array and $e_{pool\_construction}$ is the error associated with creating a pool.

Different estimates of pooling variance

Estimates of pooling variance using a single sample There are two methods for estimating the array variance from a single sample; the first method is simplest to outline and applies straightforwardly to the case where there are two array measures from same pool. The second method is given subsequently. With case pool sample estimates $\tilde{p}_{ai}$ (for controls replace a with u) on array i ($i = 1,2$)

$$\tilde{p}_{ai} = \hat{p}_a + e_{pool\_array\_i}$$

where $\hat{p}_a$ is the true frequency in that set of cases. The variance of the difference is

$$\begin{aligned}\mathrm{var}(\tilde{p}_{a1} - \tilde{p}_{a2}) &= \mathrm{var}(e_{pool\_array\_1} - e_{pool\_array\_2}) \\ &= 2 \times \mathrm{var}(e_{pool\_array})\end{aligned}$$

and $\mathrm{var}(e_{pool\_array})$ is estimated using

$$\mathrm{var}(e_{pool\_array}) = \mathrm{var}(\tilde{p}_{a1} - \tilde{p}_{a2})/2$$

where $\mathrm{var}(\tilde{p}_{a1} - \tilde{p}_{a2})$ is obtained by calculating the average of the squared differences between $\tilde{p}_{a1}$ and $\tilde{p}_{a2}$ across the full set of SNPs on the array. $\mathrm{var}(e_{pool\_array})$ is assumed constant across SNPs. When there are more than two arrays, multiple pairings of array measures are possible and the best estimate of $\mathrm{var}(e_{pool\_array})$ is the average over all pairs.

An alternative method, which applies immediately to the case where there are more than two arrays per pool, is to fit an analysis of variance to the set of $\tilde{p}_{ai}$ values. This second method gives similar results to the first method on the data used here (three arrays per pool).

In Macgregor *et al*[2] the three arrays (per case or control pool) were taken together and a quality control (QC) step applied. This step discarded SNPs with <8 probe measurements available across the three arrays. Here the arrays are considered separately and a per-array QC step implemented; this involved discarding SNPs with <2 probe measurements on the array under study.

Estimates of pooling variance using cases and controls Macgregor *et al*[2] describe a method that estimates the pooling variance from the cases and controls (summarized in appendix in supplementary online material). Unlike the case described above for estimating the pooling variance using a single sample, when cases and controls are used there is an additional component of variation owing to random (binomial) sampling. This sampling is explicitly accounted for the method described by Macgregor *et al*.[2] In this case, the two possible sources of pooling error are confounded and it is only possible to estimate a single variance (containing both the array pooling variance and the pool construction variance); this is henceforth referred to as $\mathrm{var}(e_{pool\_total})$.

To allow a suitable comparison with the estimates of pooling variance from a single pool, the estimate of $\mathrm{var}(e_{pool\_total})$ was calculated by considering each of the nine possible pairwise comparison between the case and control pools (ie, case pool array 1 *vs* control pool array 1, case pool array 1 *vs* control pool array 2, …). The overall estimate of $\mathrm{var}(e_{pool\_total})$ was then averaged over all pairs. The same QC step that was applied to the single sample analysis was used. The estimate of $\mathrm{var}(e_{pool\_total})$ will not equal the pooling variance estimate reported in Macgregor *et al*[2] (which used the same data as used here but calculated the pool variance on all three arrays) because in that case the estimate of $\mathrm{var}(e_{pool\_total})$ was a compound of the array-specific error (which is three times smaller with three arrays than with one array) and the pooling construction error (which is unaffected by the number of arrays). Furthermore, as above, a slightly different QC step was used when all three arrays were taken together.

Pooling construction variance estimates $\mathrm{var}(e_{pool\_construction})$ cannot be calculated directly from these data. However, as there are separate estimates of $\mathrm{var}(e_{pool\_array})$ (from single pools) and $\mathrm{var}(e_{pool\_total})$ (from case–control differences), $\mathrm{var}(e_{pool\_construction})$ can be estimated by subtraction

$$\mathrm{var}(e_{pool\_construction}) = \mathrm{var}(e_{pool\_total}) - \mathrm{var}(e_{pool\_array})$$

An alternative estimate of $\mathrm{var}(e_{pool\_construction})$ can also be calculated from the two possible estimates of $\mathrm{var}(e_{pool\_total})$. The first estimate, denoted $\mathrm{var}(e_{pool\_total\_arrays\_pairwise})$, from the average of the nine pairwise combinations given above yields an estimate of $\mathrm{var}(e_{pool\_array}) + \mathrm{var}(e_{pool\_construction})$. The second estimate, denoted $\mathrm{var}(e_{pool\_total\_3\_arrays})$, from the three case pool arrays together *vs* the three control pool array together yields an estimate of $\mathrm{var}(e_{pool\_array})$/

$3 + \text{var}(e_{\text{pool\_construction}})$ (this was what was calculated in Macgregor *et al*[2]). Re-arranging the previous two equations (solving the system of equations) yields

$$\text{var}(e_{\text{pool\_construction}}) = 0.5 \times \{3 \times \text{var}(e_{\text{pool\_total\_3\_arrays}}) - \text{var}(e_{\text{pool\_total\_arrays\_pairwise}})\}$$

Calculations were carried out using *R*.[6]

Results

The estimates of var($e_{\text{pool\_array}}$) were 0.00118 and 0.00133 for control and case pools, respectively. The overall estimate of var($e_{\text{pool\_array}}$) over both pools was 0.00126. The estimate of var($e_{\text{pool\_total}}$) was 0.00144 (average over all nine possible pairs). Subtracting the estimate of var($e_{\text{pool\_array}}$) from var($e_{\text{pool\_total}}$) gives an estimate of var($e_{\text{pool\_construction}}$) of 0.00018. In terms of variance explained, this suggests that only 12.5% of the variance in pooling is due to pooling construction. In all 87.5% of the variance is due to array variation.

The pooling variance estimate from Macgregor *et al*[2] was 0.00058, based on three arrays. By contrasting this estimate with the one obtained from the nine possible pairwise combinations of case–control, an alternative estimate of var($e_{\text{pool\_construction}}$) is 0.00015. In this case a slightly different QC step is applied so this may account for the slight difference between this estimate and the one in the previous paragraph.

Discussion

The success of array-based pooling depends upon reducing the overall pooling error and the results here suggest that the majority of this error arises as a result of array-specific variability. To reduce the array-specific variance several arrays should be used per pool. Based on the variance seen in the data used here, up to seven Affymetrix arrays could have been used per pool before the pooling construction variance would have become larger than the array-specific variance. In some previous array-based pooling studies,[4,7] smaller numbers of individuals ($N = 10–20$) were placed in each pool. This contrasts with the large number ($N = 384$) used here. The work presented here suggests that, as the pooling error is largely array-specific error, using larger numbers of arrays on smaller numbers of pools (with more individuals per pool) will be more effective than smaller numbers of arrays on larger numbers of pools. As discussed in Macgregor *et al*,[2] the overall optimal study design will varying depending on the size of the overall pooling variance relative to the binomial sampling variance.

The estimates of var($e_{\text{pool\_construction}}$) were relatively small but replication of this result in other pools will be important. For the experiment described here, pools were carefully constructed following estimation of DNA concentrations in a step down procedure to achieve final DNA concentrations of 25 ng/μl ($\pm$0.55) before pooling.[5] It is difficult to know from a single data set how much variability there will be in the estimate of var($e_{\text{pool_construction}}$) and the overall levels of pooling construction variance will likely vary across laboratories. As the estimate of var($e_{\text{pool_construction}}$) calculated here was based on a limited number of arrays, the confidence interval on the estimate of var($e_{\text{pool_construction}}$) may not be particularly narrow.

In the above analysis the focus was on array variation being the source of technical variation. There are a number of technical steps necessary to produce data from pools and it is likely that both PCR variation and hybridization variation contribute to the overall technical variation. An experiment, which recycled the reaction product for multiple hybridizations would allow partition of the technical variation.

A number of assumptions were made in the analysis (see also Macgregor *et al*[2] for further coverage). Firstly, all SNPs were assumed to be unassociated with disease; this will hold for virtually all SNPs. Secondly, the pooling variance was assumed to be constant across SNPs on the array; no strong evidence was found for systematic variation, particularly for SNPs with allele frequencies in the range of primary interest (0.1–0.9). Finally, unequal amplification of alleles was assumed to not affect results; the focus was on the difference in allele frequencies (between case/control or between arrays 1 and 2 on a given pool, and so on) so this is unlikely to be an issue.

Acknowledgements

Thanks to Peter M Visscher and Grant Montgomery for helpful discussions on this topic. Zhen Zhen Zhao and the QIMR Molecular and Genetic Epidemiology Laboratories provided expert assistance in collection and preparation of the DNA pools. Sue Treloar's pioneering work enabled the establishment of the QIMR Endometriosis study. The study and sample collections were partly supported by Grants 339430, 339446 and 389892 from the National Health and Medical Research Council and by the Cooperative Research Centre for the Discovery of Genes for Common Human Diseases established and supported by the Australian Government's Cooperative Research Centre's Program.

References

1 Sham P, Bader JS, Craig I, O'Donovan M, Owen M: DNA pooling: a tool for large-scale association studies. *Nat Rev Genet* 2002; **3**: 862–871.
2 Macgregor S, Visscher PM, Montgomery G: Analysis of pooled DNA samples on high density arrays without prior knowledge of differential hybridization rates. *Nucleic Acids Res* 2006; **34**: e55.
3 Kirov G, Nikolov I, Georgieva L, Moskvina V, Owen MJ, O'Donovan MC: Pooled DNA genotyping on Affymetrix SNP genotyping arrays. *BMC Genomics* 2006; **7**: 27.
4 Liu QR, Drgon T, Walther D *et al*: Pooled association genome scanning: validation and use to identify addiction vulnerability loci in two samples. *Proc Natl Acad Sci USA* 2005; **102**: 11864–11869.
5 Zhao ZZ, Nyholt DR, James MR, Mayne R, Treloar SA, Montgomery GW: A comparison of DNA pools constructed following whole

genome amplification for two-stage SNP genotyping designs. *Twin Res Hum Genet* 2005; **8**: 353–361.
6 R Development Core Team: *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2004. ISBN 3-900051-00-3.
7 Brohede J, Dunne R, McKay JD, Hannan GN: 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* 2005; **33**: e142.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)