Optimal Design of Genetic Studies of Gene Expression With Two-Color Microarrays in Outbred Crosses

Alex C. Lam,^{*,†,1} Jingyuan Fu,[‡] Ritsert C. Jansen,[‡] Chris S. Haley^{*} and Dirk-Jan de Koning^{*}

^{*}Roslin Institute (Edinburgh) and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Roslin Biocentre, Roslin, Midlothian EH25 9PS, United Kingdom, [†]Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom and [‡]Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, 9751 NN, Haren, The Netherlands

Manuscript received April 14, 2008

Accepted for publication September 2, 2008

ABSTRACT

Combining global gene-expression profiling and genetic analysis of natural allelic variation (genetical genomics) has great potential in dissecting the genetic pathways underlying complex phenotypes. Efficient use of microarrays is paramount in experimental design as the cost of conducting this type of study is high. For those organisms where recombinant inbred lines are available for mapping, the “distant pair design” maximizes the number of informative contrasts over all marker loci. Here, we describe an extension of this design, named the “optimal pair design,” for use with F_2 crosses between outbred lines. The performance of this design is investigated by simulation and compared to several other two-color microarray designs. We show that, for a given number of microarrays, the optimal pair design outperforms all other designs considered for detection of expression quantitative trait loci (eQTL) with additive effects by linkage analysis. We also discuss the suitability of this design for outbred crosses in organisms with large genomes and for detection of dominance.

GENETIC analysis of variation in gene expression, also known as genetical genomics (JANSEN and NAP 2001), has great potential for dissecting the mechanisms underlying complex phenotypes (MEHRABIAN *et al.* 2005; SCHATZ *et al.* 2005). Although variation in transcript abundance is often in response to external environmental factors, part of the between-individual variation in expression of a substantial number of genes can be explained by DNA polymorphisms (JIN *et al.* 2001). To date, the vast majority of published studies in this research area have been conducted in model organisms such as yeast (BREM *et al.* 2002), flowering plant (KEURENTJES *et al.* 2007), nematode worm (LI *et al.* 2006), mouse (SCHATZ *et al.* 2003; BYSTRYKH *et al.* 2005), and rat (HUBNER *et al.* 2005). There are also a number of studies that focused on human populations (MONKS *et al.* 2004; MORLEY *et al.* 2004; STRANGER *et al.* 2005). Efforts in mapping expression quantitative trait loci (eQTL) have provided strong evidence for candidate gene selection in studies of complex phenotypes such as hypertension (HUBNER *et al.* 2005) and childhood asthma (DIXON *et al.* 2007).

Like in any QTL study, appropriate sample size is essential for adequate power in eQTL detection. Although many of the published studies have provided very interesting insights into the properties of genetic loci that regulate gene-expression phenotypes, the small

sample sizes of the early studies meant they have limited power to detect eQTL of small to moderate effects (DE KONING and HALEY 2005). In many cases, there is no shortage of animals or cell lines for a genetical genomics approach as the genetic materials have already been collected for concurrent large-scale studies. Therefore, the major factor that restricts sample sizes tends to be the high cost of the associated technologies, particularly the cost of microarrays. To address this issue, significant improvement in the usage of microarray resources for genetical genomics has been proposed in a number of articles. JIN *et al.* (2004) presented an algorithm for “selective phenotyping” in which a subsample was chosen from the entire sample set for maximum genotypic dissimilarity as a way to reduce the amount of phenotyping without sacrificing sensitivity in QTL detection. In a different article, PIEPHO (2005) discussed the optimal allocation of samples to cDNA microarrays for detecting heterosis. BUENO FILHO *et al.* (2006) covered a range of optimal microarray designs, from studying the genotypic effect of a single locus to models that include both fixed treatment and random polygenic effects. ROSA *et al.* (2006) provided a comprehensive review on microarray design for eQTL mapping. FU and JANSEN (2006) proposed a more general approach called the distant pair design, which combines optimal allocation by hybridizing most dissimilar samples and selective genotyping when the population resource is large. They used recombinant inbred lines (RILs) to demonstrate the power of this approach. In this article, the research

¹Corresponding author: Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, United Kingdom. E-mail: alex.lam@roslin.ed.ac.uk

emphasis is on further developing this experimental design to other populations, like the outbred F_2 crosses.

In Fu and Jansen's study, the authors chose to explore two-color microarray technologies over single-color platforms because for the same number of slides, two-color microarrays can potentially generate twice as much hybridization data as single-color arrays. In addition, two-color microarray platforms remain the only choice for many research projects because commercial single-color microarrays such as Affymetrix GeneChip are available for only a handful of species. Furthermore, two-color microarrays offer greater flexibility in experimental design in which pairs of samples can be selected to cohybridize deliberately, which enhances performance. Fu and Jansen proposed the "distant pair design," which outperforms the conventional designs, namely the common reference and the loop designs, when a panel of RILs is used. Moreover, for the expression profiling of a given number of biological samples, the distant pair design would require only half as many slides as the common reference or the loop design in a two-color microarray.

The distant pair design presents an effective direct-pairing strategy that increases the ratio of within- over between-slides genotypic dissimilarity (ROSA *et al.* 2006). However, it is not clear how to apply this rationale to populations in which more than two genotypes are possible for each marker locus. Although some insight was provided on how best to detect overdominance, or heterosis, which involved subjects with hybrid genotypes (PIEPHO 2005), this strategy for sample allocation is not designed for mapping gene-expression variation to any specific loci. BUENO FILHO *et al.* (2006) addressed the problem with three possible genotypes at a single locus and gave a generalization of the solution to multiple loci. Their method may be more tractable when marker genotypes can be treated as fixed effects for contrast, which is true for inbred line crosses. For researchers studying genetics of outbred species, mapping resources like inbred strains or RILs are often not feasible. By contrast, F_2 intercrosses between two genetically divergent outbred populations are much more readily available. A major complication arising in outbred crosses is due to the fact that there are common sets of alleles segregating in both of the founder populations. Hence, it is often the case that marker genotypes in the F_2 generation would not be fully informative for the origins of lineage at any given locus. This uncertainty obscures how one can define genotypic dissimilarity for the purpose of pair assignment in distant pair design. Furthermore, researchers face the issue regarding large genome sizes. It is expected that when genome size increases, finding distant pairs will become more and more difficult. FU and JANSEN (2006) have shown that in RILs a small advantage is achievable with large genomes. However, whether this advantage is also present in an F_2 design remains uncertain. This question is directly

relevant to researchers who are interested in studying the genetics of gene expression in nonmodel organisms. Therefore, the usefulness of the distant pair design for genetical genomic studies in outbred F_2 crosses warrants investigation.

In this article, we propose an extension to the distant pair design by adapting the least-squares QTL mapping framework (HALEY *et al.* 1994). Here we refer to this extension as the "optimal pair design." We also assess the performance of this design in the presence of dominance. Moreover, we consider the impact of genome size on power and discuss the usefulness of our extension of the distant pair design for eQTL studies in outbred experimental crosses.

MATERIALS AND METHODS

QTL analysis: The method for mapping QTL follows the least-squares approach (HALEY *et al.* 1994). Briefly, the line origins at fixed intervals (*e.g.*, 1 cM) along the genome for the individuals in the F_2 generation are expressed as lineage probabilities, conditional on the marker genotype. This can be done by considering all possible line-origin combinations based on the parental and grandparental genotypes and has been implemented in the online software "QTL Express" (SEATON *et al.* 2002). Assuming that founder lines are fixed for alternative QTL alleles, the lineage probabilities can be used to predict the putative QTL genotypes. Phenotypic values are then regressed onto genetic coefficients calculated for a putative QTL at a fixed position. The genetic coefficients for additive and dominance effects are derived from the conditional probabilities: the additive coefficient (denoted x_a) is the difference of the probabilities for the homozygous line origins, and the dominance coefficient (denoted x_d) is the sum of the probabilities for the heterozygous line origins. An F -ratio test statistic can be used to test the null model (without QTL fitted) against the full model (with QTL fitted) and determine the significance of the presence of QTL. For full details on the derivation of line-origin probabilities and regression-based QTL mapping, see HALEY *et al.* (1994).

In the context of a pair design in two-color microarrays, the gene-expression phenotypes can be expressed either in ratios or in signals of the separate channels. In this article we chose ratios over signals as the phenotypes because the use of ratios can minimize the risk of bias as a result of spot or array effects (WIT and MCCLURE 2004). FU and JANSEN (2006) argued that there is a negligible difference in the final results between ratios and signals, provided that the distributional assumptions for the array and spot effects used in the signal-based analysis are correct. The log ratio of the red channel intensity to the green channel intensity of a probe is equivalent to the difference of the two signal intensities in logarithmic scale. To utilize such phenotypes in the Haley-Knott least-squares framework, the linear regression model can be written as

$$\Delta y_i = \mu + \Delta x_{a_i} a + \Delta x_{d_i} d + e_i, \quad (1)$$

where Δy_i is the difference by subtracting the log signal of the green channel from that of the red channel for the i th microarray ($i = 1, \dots, n$); μ is the overall mean; Δx_{a_i} is the difference of the additive coefficients by subtracting x_a of the individual assigned to the green channel from x_a of the individual assigned to the red channel for the i th microarray; Δx_{d_i} is the coefficient difference for dominance x_d ; a and d are

the additive and dominance parameters, respectively; and e_i is the residual error. In matrix form, the expression can be simplified as $Y = Xb + e$, where $b = (\mu, a, d)^T$.

Finding optimal pairs: We used the same definition for the optimal design as in the original publication on the distant pair design (FU and JANSEN 2006), which is the minimum for the sum of the variances of $\hat{b}$ in the matrix form of our model. Following the A-optimality criterion (WIT and McCURE 2004), this is equivalent to minimizing the trace of $(X'X)^{-1}$. For our regression model in (1), the matrix X consists of a column of 1's for the mean μ , a column of Δx_{ai} coefficients for the additive parameter, and, if dominance is included in the model, a third column of Δx_{di} coefficients. To reach the optimal pairing design over all positions in the genome, we search for the minimum of the sum over all marker loci the trace of $(X'X)^{-1}$. Genetic coefficients at marker loci only are used for optimization to keep the computation tractable. Because the genetic coefficients between marker intervals are derived from the markers, we do not anticipate our optimization method to be different from using the coefficients at every centimorgan.

The simulated annealing technique (KIRKPATRICK *et al.* 1983) was used to find a pairing configuration that is optimal or close to optimal according to the definition above. Consult FU and JANSEN (2006) for details of the procedures. The implementation of finding optimal pairs was accomplished using the R statistical computer program (R DEVELOPMENT CORE TEAM 2007).

Power assessment via simulations: We studied three different genome sizes: 100, 1000, and 2000 cM; and for each genome size we simulated 100 replicates of F_2 intercrosses. First, F_1 individuals were generated by randomly mating 20 F_0 sires from founder line one to 80 F_0 dams from founder line two (4 dams per sire), each having 5 offspring. Then, another 400 offspring were generated in the F_2 generation by randomly mating 20 F_1 sires to 80 F_1 dams (5 progenies per mating). Marker data were simulated for all samples, with 11 evenly spaced markers per chromosome of 100 cM in length. Four alleles were simulated for every marker segregating at equal frequencies in both founder lines, with marker genotypes in Hardy-Weinberg equilibrium. A single biallelic QTL that is fixed for alternative alleles in the founder lines was simulated on the first chromosome at 46 cM. For this QTL, we simulated two alternative settings: (a) an additive QTL without dominance, where the homozygous genotypic value $a = 0.5$ and the heterozygous genotypic value $d = 0$, and (b) a QTL with complete dominance, where $a = 0.5$ and $d = 0.5$. Polygenic background effects were modeled as 10 unlinked biallelic loci, each with an additive effect of 0.25 and segregating at a frequency of 0.5 in both founder lines, as described in ALFONSO and HALEY (1998). To mimic the nongenetic factors affecting the gene-expression phenotype and technical errors of microarrays, we added an environmental component sampled from a normal distribution with a variance of 0.5 to the simulated phenotype. The narrow-sense heritability (h^2) is 0.47 for the trait and 0.20 for the main QTL on the first chromosome.

To assess the performance of the optimal pair design under the least-squares framework, we scanned in 1-cM steps for the most significant P -values obtained in the marker interval that contains the QTL (between 40 and 50 cM on the first chromosome) under four scenarios. These four scenarios are summarized in Table 1 and are described as follows: first, all 400 F_2 subjects and their individual phenotypic measurements were analyzed. Conceptually this is equivalent to the common reference design that includes all F_2 individuals. Second, 200 F_2 subjects were randomly selected, together with their individual phenotypic measurements. This scenario also represents the common reference design, but a smaller budget

TABLE 1
Summary of the four scenarios investigated in the power study

Scenario abbreviations	Description	No. of F_2 subjects profiled	No. of slides required
all.data	Individual phenotypic values are available for all subjects	400	400
half.data	Same as all.data except that 50% of the subjects are selected	200	200
ran.pair	Pairs are assigned randomly	400	200
opt.pair	Pairs are assigned according to the outcome of simulated annealing	400	200

limits the profiling of gene expression to fewer individuals than in the first scenario. Due to the random sampling nature of this scenario, for each simulated population replicate we repeated the random sampling 100 times and scanned for the most significant P -value in the QTL-containing interval as above. Then the median P -value was selected to represent the performance under this scenario for the given population replicate. Third, we randomly paired up all 400 F_2 subjects and analyzed the data with regression model (1). Under this scenario, we also repeated the process 100 times per simulated population replicate and proceeded to obtain the P -value in the same way as in the second scenario. Finally, we paired up all 400 F_2 subjects, using the optimal pair design. We abbreviate these four scenarios above as "all.data," "half.data," "ran.pair," and "opt.pair," respectively, for reference in the rest of this article. For both "additive only" and "additive and dominance" QTL settings, the data were analyzed under those four scenarios.

Alternative marker allele frequencies and population sizes: In the simulations above the marker allele frequencies are equal over all four alleles in both founder lines. This represents a suboptimal scenario in which the marker genotypes in the F_2 generation are expected to have limited information for the line origins. For the genome size of 2000 cM, we also simulated the "best-case scenario" in which each founder line has two unique alleles; *i.e.*, two of the four alleles are segregating within each founder line, with no common alleles shared by both lines. Such an intercross is equivalent to an F_2 cross between two inbred lines. These two sets of marker allele frequencies would enable us to determine a below-average range and the upper bound for the performance of the optimal pair design. In addition, we performed further simulations in which we fixed the number of microarrays being used to 400 and evaluated an F_2 population size of 1000. We compared the performance of the optimal pair design and the common reference design when expression profiling of every individual in the sample population is not possible.

RESULTS

Additive effect: We studied the power for detecting additive QTL under the four scenarios. For the results of opt.pair presented in this section, we minimized the variance of the additive effect in the regression model by

FIGURE 1.—Performance for detecting additive QTL effects under various scenarios. (A) Genome size of 100 cM, a single chromosome; (B) genome size of 2000 cM, 20 chromosomes. The horizontal dotted line shows the significant level of $P = 1 \times 10^{-5}$. The simulations are sorted in ascending order of the $-\log_{10}P$ on the x -axis.

simulated annealing. Figure 1 shows the minus log-transformed P -values (sorted in ascending order) for the four scenarios. The scenario with the highest proportion of the largest minus log-transformed P -values can be considered as the most powerful design. For a single chromosome (Figure 1A), the most significant P -values can be found under the all.data scenario. But for the opt.pair scenario, under which only 200 microarrays would be required, the power to detect the QTL is remarkably close to that under the all.data scenario. Under the half.data and ran.pair scenarios, likewise, only 200 microarrays would be required, but the power is much reduced compared to both all.data and opt.pair. Incidentally, the performances of half.data and ran.pair are almost identical; hence most of the data points for these two designs are overlapping in Figure 1.

Table 2 summarizes the performance under the four scenarios by the mean $-\log_{10}P$ and shows the effect of genome size on the power for detecting QTL. The mean $-\log_{10}P$ across different genome sizes under the all.data, half.data, and ran.pair scenarios shows little deviation. However, the mean $-\log_{10}P$ under the opt.pair scenario follows a notable downward trend when the genome size increased. At the genome size of 2000 cM (Figure 1B) all.data performs best of the four scenarios. But more importantly, the opt.pair scenario is the most powerful of the designs that require 200 microarrays.

We analyzed the simulations of the F_2 cross with fully informative markers for the genome size of 2000 cM and found that the power increased slightly under all four scenarios (Table 2). The increase in power is expected because line origins can be inferred with certainty. It is important to note that the difference in the power between the suboptimal and the best-case scenario for the marker allele frequencies is small. This indicates that our power assessment using equal marker allele frequencies in the simulations is robust and representative of real outbred F_2 intercrosses, of which the marker allele frequencies in the founder lines are in between those two extremes.

Additive and dominance effects: For the dominant QTL, two levels of analysis were carried out: (a) QTL detection by comparing the full model (additive and dominance) to the null model and (b) detection of dominance effect by comparing the full model to the reduced model (additive only). In the simulated annealing step of optimal pairing, the dominance coefficients were included as the third column in the matrix X in the linear model (see MATERIALS AND METHODS).

With a single-chromosome (100 cM) genome, the power to detect QTL under the opt.pair scenario is clearly lower (Figure 2A, left) than that under all.data. It can be seen in Table 3 that the mean $-\log_{10}P$ under all.data is $\sim 50\%$ greater than that under opt.pair. But opt.pair is still more powerful than both half.data and ran.pair. By contrast, our results (Figure 2A, right) show that opt.pair and all.data are similarly powerful for detecting dominance effects and are superior to both half.data and ran.pair in a small genome.

Tables 3 and 4 show that genome sizes again have little effect on power under all.data, half.data, and ran.pair. However, the increase in genome size affects optimal pairing more severely here than when no dominance effect has been simulated. At the genome size of 2000 cM, opt.pair is only marginally more powerful in detecting the QTL than half.data and ran.pair (Figure 2B, left). The power for detecting the dominance effect is more drastically affected and opt.pair performs similarly to half.data and ran.pair (Figure 2B, right). We repeated the optimization solely on the additive parameter with the genome size of 2000 cM. The power of detecting QTL was found to have improved slightly, while there was little change in the power for detecting dominance

TABLE 2

Summary of P -values (mean and standard deviation on $-\log_{10}$ scale) at the main QTL position for additive QTL detection under the four scenarios, where only an additive effect was simulated

Genome size (cM)	No. of chromosomes	all.data	half.data	ran.pair	opt.pair
100	1	11.9 (2.9)	6.4 (1.5)	6.3 (1.5)	11.0 (2.7)
1000	10	12.3 (2.6)	6.6 (1.3)	6.6 (1.4)	9.2 (2.4)
2000	20	12.1 (2.9)	6.5 (1.5)	6.4 (1.5)	8.3 (2.4)
2000 ^a	20	12.9 (2.9)	6.9 (1.5)	6.8 (1.5)	8.9 (2.4)

Standard deviations are in parentheses.

^a Fully informative markers.

(results not shown). Therefore, in the presence of dominance effects, the advantage in the performance of the optimal pair design in detecting QTL is reduced. Including dominance in the optimization has a negative impact on the optimal pair design, especially for large genome sizes, when QTL detection is the primary objective.

Fixed number of microarrays with a large F_2 sample size: In previous simulations, we observed that all.data, which required 400 microarrays, was more powerful in detecting the additive QTL effect than using 200 microarrays under the opt.pair scenario. Here, we studied the power of these two designs conditioned on a total of 400 microarrays. With an F_2 sample size of 1000, neither design can profile all the individuals with 400 microarrays. Under the optimal pair design, 400 pairs were deliberately selected to give the minimum

variance for the estimated additive genetic parameter. On the other hand, only 400 individuals (randomly selected from 1000 individuals) could be profiled using the common reference design. Given the equal number of microarrays being used, our results in Figure 3 show that the optimal pair design outperforms the common reference design.

DISCUSSION

The distant pair design enables the mapping of eQTL in an efficient and effective manner using recombinant inbred lines. For researchers studying genetics of many outbred species, however, the creation of recombinant inbred lines is impractical. Here we explore whether eQTL studies of natural species would benefit from the

FIGURE 2.—Performance for detecting QTL additive and dominance (left) and dominance effects (right) under various scenarios. (A) Genome size of 100 cM, a single chromosome; (B) genome size of 2000 cM, 20 chromosomes. The horizontal dotted line shows the significant level of $P = 1 \times 10^{-5}$. The simulations are sorted in ascending order of the $-\log_{10}P$ on the x-axis.

TABLE 3

Summary of *P*-values (mean and standard deviation in $-\log_{10}$ scale) at the main QTL position for QTL detection (additive plus dominance model *vs.* null model) under the four scenarios, where both additive and dominance effects were simulated

Genome size (cM)	No. of chromosomes	all.data	half.data	ran.pair	opt.pair
100	1	15.2 (3.2)	7.8 (1.6)	7.8 (1.7)	10.5 (2.6)
1000	10	15.5 (3.3)	8.0 (1.7)	7.9 (1.7)	9.6 (2.3)
2000	20	15.4 (3.7)	7.9 (1.8)	7.9 (1.8)	8.9 (2.5)

Standard deviations are in parentheses.

same design principles used in distant pairing. We show that the optimal pair design, an extension of the distant pair design for outbred lines crosses, can indeed improve the efficiency of the use of microarrays and increase the statistical power for detecting eQTL, even for studying organisms with large genome sizes.

Under the linear regression framework, the greatest power is achieved by having the regression coefficients in equal proportions near the top and bottom extremes. For the regression model proposed for the optimal pair design in this article, this would be achieved by pairing up individuals who have large genetic coefficients with opposite signs. However, in a line cross such as the F_2 , it is inevitable that not every pair would result in a regression coefficient that is near one extreme or the other. Furthermore, when the number of independent loci increases (increase in chromosome length and number of chromosomes), the optimal pair assignment for one locus will usually not be optimal for the other loci. The optimal pair assignment over the whole genome is therefore suboptimal in the perspective of a single locus, *i.e.*, fewer regression coefficients around the extremes. We can therefore expect the performance of distant pairing to degrade to the same level as that of random pairing eventually as the genome size continues to increase.

Clear benefits in detecting additive effects: We show that when there are few loci to consider, such as in a small genome, the power of detecting additive effects with the optimal pair design is similar to using a common reference design that consumes twice the number

of microarrays. With near-optimal pairing for individual loci (achievable when there are small numbers of effectively independent loci), the efficiency of the optimal pair design is very attractive. Moreover, the common reference design with only half the sample size (*i.e.*, the same number of microarrays) performs significantly worse. This highlights the problem of small sample size leading to reduction in power in complex trait analysis. Extremely similar performances were observed for random pairing and the common reference design with 50% of the samples. The difference in performance between these two designs could have been more marked if our simulations had explicitly modeled the possible difference in the biological sampling variance between a pair design and a common reference design.

As expected, the performance of the optimal pair design drops when the genome size increases. Nevertheless, it is very promising that in a large genome the optimal pair design still notably outperforms designs that use the same number of microarray slides. Furthermore, as shown by the excellent performance in smaller genomes, it is evident that the optimal pair design would be beneficial for a focused study of one or more candidate regions within a large genome. The power can be maximized for genomic regions for which the researchers have the most interest, while the power in the rest of the genome would be at least as good as the random pair design. In addition, we show that with the number of microarrays used being equal, the optimal pair design always gives the highest statistical power of the approaches compared. Our comparison to the common

TABLE 4

Summary of *P*-values (mean and standard deviation on $-\log_{10}$ scale) at the main QTL position for dominance detection (additive plus dominance model *vs.* additive model) under the four scenarios, where both additive and dominance effects were simulated

Genome size (cM)	No. of chromosomes	all.data	half.data	ran.pair	opt.pair
100	1	5.8 (2.4)	3.3 (1.2)	3.2 (1.2)	5.4 (2.1)
1000	10	5.7 (1.7)	3.1 (0.9)	3.1 (0.9)	3.8 (1.5)
2000	20	5.8 (2.4)	3.2 (1.2)	3.2 (1.2)	3.7 (1.9)

Standard deviations are in parentheses.

FIGURE 3.—Comparison of the performance for QTL detection under the common reference design and the optimal pair design when the number of arrays is fixed at 400, the genome size is 2000 cM, and the F_2 sample size is 1000. The horizontal dotted line shows the significant level of $P = 1 \times 10^{-5}$. The simulations are sorted in ascending order of the $-\log_{10}P$ on the x-axis.

reference design was made to random selection of individuals. Although selective phenotyping (JIN *et al.* 2004) can improve the efficiency of the common reference design, the optimal pair design would allow more subjects to be assayed as well as maximize the genotypic dissimilarity. Therefore, for outbred species that possess large genomes, the optimal pair design can provide both efficient use of the microarray resource and good power for the detection of eQTL with additive effects.

Complications due to dominance effects: How does dominance affect the performance of this design? We evaluate the optimal pair design that optimizes for both additive and dominance effects simultaneously: the conclusion is that by including the dominance parameter, the design becomes less optimized for detecting the main (additive) effect. Although over a small genome, the optimal pair design can offer a moderate power advantage for detecting QTL and dominance effects over no optimization, the performance is affected severely in that the power for detecting both the main and the dominance effect degrade to almost the same levels as random pairing with a large genome. Our results agree with other studies (PIEPHO 2005; BUENO FILHO *et al.* 2006) that finding a design that is optimal for detecting both additive and dominance effects cannot be achieved. They have shown that optimizing for detecting dominance effects would decrease power for detecting additive effects. Therefore, when one has to make a choice between additive and dominance effects for optimization, the question relates directly to the goal of the experiment. If the goal is to scan across the whole genome for linked loci to gene-expression phenotypes, we argue that one could consider focusing

on the additive parameter alone for the optimization. After all, the ultimate interest is to detect QTL. In most cases QTL are expected to have an additive component, even in cases where dominance is present. Optimizing for dominance effects should be considered only if there is strong *a priori* evidence for overdominance in the QTL of interest in a candidate gene study.

Final remarks: We conclude that our extension of the distant pair design, the optimal pair design, can be applied efficiently to outbred line crosses for genetical genomic studies. Having stated that, we acknowledge that in an experimental design for genetical genomics, there is no “one-size fits-all” solution. The most powerful and efficient design will depend on the population structure, the marker density, the chosen method of analysis, the numbers of treatments, and the parameter of interest. BUENO FILHO *et al.* (2006) proposed different designs for multiple genotypes, epistasis, and multiple treatments. In human or other natural populations, the Haseman–Elston method (HASEMAN and ELSTON 1972) can be applied to sib-pair analysis, in which case, the most effective use of microarray resources to conduct an eQTL linkage analysis would be to profile the expression of a pair of sibs on the same array. This is because the trait squared differences between two sibs are the dependent variable used in this method; these quantities are obtained most accurately when sibs are paired up on the same array.

It is also worth considering the implication of the use of high-density single-nucleotide polymorphism (SNP) genotyping on the optimal pair design described in this article. High-density SNP genotyping is most widely used in association studies in natural human populations rather than in line crosses of the animals discussed above. As linkage disequilibrium spans relatively short distances in human populations, the effective number of independent loci is much higher than what we have modeled in our line-cross simulations. This effect is equivalent to increasing the genome size and is likely to have a negative impact on the performance of the optimal pair design than what can be expected in outbred line crosses. Eventually, the distant pairing strategy might become almost equivalent to a pairing strategy based on relationships, in which less-related individuals should be paired for each hybridization (ROSA *et al.* 2006; BUENO FILHO *et al.* 2006). Theoretically, the optimal pair design should always be preferred; since the variance of the estimate of the parameter is minimized, its performance should be at least as good as the common reference design. However, other factors, such as technical simplicity and flexibility in the choice of statistical methods, might shift the balance in favor of the common reference design when the performance advantage in using the optimal pair design becomes less marked. Therefore, it is imperative to consider each experiment and the question of interest on a case-by-case basis. Nevertheless, our results suggest that the

efficient design principles outlined by FU and JANSEN (2006) can be applied to a wider context than RILs. With larger eQTL experiments becoming more affordable, we can expect to discover more loci with moderate to small effects. Such attainment will ultimately lead to greater advances in our understanding of the molecular basis of complex traits.

We thank two anonymous referees for their helpful suggestions. The R code for optimizing pairing configuration can be obtained by request from the corresponding author. This research was funded by the Biotechnology and Biological Sciences Research Council (BBSRC). A.C.L. is grateful for support from the BBSRC (grant no. BBSF200512735), the Genesis Faraday Partnership, and Genus plc.

LITERATURE CITED

- ALFONSO, L., and C. S. HALEY, 1998 Power of different F-2 schemes for QTL detection in livestock. *Anim. Sci.* **66**: 1–8.
- BREM, R. B., G. YVERT, R. CLINTON and L. KRUGLYAK, 2002 Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**: 752–755.
- BUENO FILHO, J. S., S. G. GILMOUR and G. J. ROSA, 2006 Design of microarray experiments for genetical genomics studies. *Genetics* **174**: 945–957.
- BYSTRYKH, L., E. WEERSING, B. DONTJE, S. SUTTON, M. T. PLETCHER *et al.*, 2005 Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat. Genet.* **37**: 225–232.
- DE KONING, D. J., and C. S. HALEY, 2005 Genetical genomics in humans and model organisms. *Trends Genet.* **21**: 377–381.
- DIXON, A. L., L. LIANG, M. F. MOFFATT, W. CHEN, S. HEATH *et al.*, 2007 A genome-wide association study of global gene expression. *Nat. Genet.* **39**: 1202–1207.
- FU, J. Y., and R. C. JANSEN, 2006 Optimal design and analysis of genetic studies on gene expression. *Genetics* **172**: 1993–1999.
- HALEY, C. S., S. A. KNOTT and J. M. ELSN, 1994 Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics* **136**: 1195–1207.
- HASEMAN, J. K., and R. C. ELSTON, 1972 The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* **2**: 3–19.
- HUBNER, N., C. A. WALLACE, H. ZIMDAHL, E. PETRETTO, H. SCHULZ *et al.*, 2005 Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat. Genet.* **37**: 243–253.
- JANSEN, R. C., and J. P. NAP, 2001 Genetical genomics: the added value from segregation. *Trends Genet.* **17**: 388–391.
- JIN, C., H. LAN, A. D. ATTIE, G. A. CHURCHILL, D. BULUTUGLO *et al.*, 2004 Selective phenotyping for increased efficiency in genetic mapping studies. *Genetics* **168**: 2285–2293.
- JIN, W., R. M. RILEY, R. D. WOLFINGER, K. P. WHITE, G. PASSADOR-GURGEL *et al.*, 2001 The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nat. Genet.* **29**: 389–395.
- KEURENTJES, J. J., J. FU, I. R. TERPSTRA, J. M. GARCIA, G. VAN DEN ACKERVEN *et al.*, 2007 Regulatory network construction in *Arabidopsis* by using genome-wide gene expression quantitative trait loci. *Proc. Natl. Acad. Sci. USA* **104**: 1708–1713.
- KIRKPATRICK, S., C. D. GELATT, JR. and M. P. VECCHI, 1983 Optimization by simulated annealing. *Science* **220**: 671–680.
- LI, Y., O. A. ALVAREZ, E. W. GUTTELING, M. TIJSTERMAN, J. FU *et al.*, 2006 Mapping determinants of gene expression plasticity by genetical genomics in *C. elegans*. *PLoS Genet.* **2**: e222.
- MEHRABIAN, M., H. ALLAYEE, J. STOCKTON, P. Y. LUM, T. A. DRAKE *et al.*, 2005 Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. *Nat. Genet.* **37**: 1224–1233.
- MONKS, S. A., A. LEONARDSON, H. ZHU, P. CUNDIFF, P. PIETRUSIAK *et al.*, 2004 Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.* **75**: 1094–1105.
- MORLEY, M., C. M. MOLONY, T. M. WEBER, J. L. DEVLIN, K. G. EWENS *et al.*, 2004 Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**: 743–747.
- PIEPHO, H. P., 2005 Optimal allocation in designs for assessing heterosis from cDNA gene expression data. *Genetics* **171**: 359–364.
- R DEVELOPMENT CORE TEAM, 2007 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.
- ROSA, G. J., N. DE LEON and A. J. ROSA, 2006 Review of microarray experimental design strategies for genetical genomics studies. *Physiol. Genomics* **28**: 15–23.
- SCHADT, E. E., S. A. MONKS, T. A. DRAKE, A. J. LUSIS, N. CHE *et al.*, 2003 Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**: 297–302.
- SCHADT, E. E., J. LAMB, X. YANG, J. ZHU, S. EDWARDS *et al.*, 2005 An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* **37**: 710–717.
- SEATON, G., C. S. HALEY, S. A. KNOTT, M. KEARSEY and P. M. VISSCHER, 2002 QTL Express: mapping quantitative trait loci in simple and complex pedigrees. *Bioinformatics* **18**: 339–340.
- STRANGER, B. E., M. S. FORREST, A. G. CLARK, M. J. MINICHIELLO, S. DEUTSCH *et al.*, 2005 Genome-wide associations of gene expression variation in humans. *PLoS Genet.* **1**: e78.
- WIT, E., and J. MCCLURE, 2004 *Statistics for Microarrays: Design, Analysis and Inference*. John Wiley & Sons, Chichester, UK.

Communicating editor: R. W. DOERGE