BMC Bioinformatics

Methodology article

Statistical implications of pooling RNA samples for microarray experiments

Xuejun Peng*[1], Constance L Wood[1], Eric M Blalock[2], Kuey Chu Chen[2], Philip W Landfield[2] and Arnold J Stromberg[1]

Address: [1]Department of Statistics, University of Kentucky, Lexington, KY 40506, USA and [2]Department of Molecular and Biomedical Pharmacology, University of Kentucky, Lexington, KY 40536, USA

Email: Xuejun Peng* - peng@ms.uky.edu; Constance L Wood - cwood@uky.edu; Eric M Blalock - emblal@uky.edu; Kuey Chu Chen - kueyc@uky.edu; Philip W Landfield - pwland@uky.edu; Arnold J Stromberg - astro@ms.uky.edu

* Corresponding author

Published: 24 June 2003

BMC Bioinformatics 2003, **4**:26

Received: 02 December 2002
Accepted: 24 June 2003

This article is available from: http://www.biomedcentral.com/1471-2105/4/26

Abstract

Background: Microarray technology has become a very important tool for studying gene expression profiles under various conditions. Biologists often pool RNA samples extracted from different subjects onto a single microarray chip to help defray the cost of microarray experiments as well as to correct for the technical difficulty in getting sufficient RNA from a single subject. However, the statistical, technical and financial implications of pooling have not been explicitly investigated.

Results: Modeling the resulting gene expression from sample pooling as a mixture of individual responses, we derived expressions for the experimental error and provided both upper and lower bounds for its value in terms of the variability among individuals and the number of RNA samples pooled. Using "virtual" pooling of data from real experiments and computer simulations, we investigated the statistical properties of RNA sample pooling. Our study reveals that pooling biological samples appropriately is statistically valid and efficient for microarray experiments. Furthermore, optimal pooling design(s) can be found to meet statistical requirements while minimizing total cost.

Conclusions: Appropriate RNA pooling can provide equivalent power and improve efficiency and cost-effectiveness for microarray experiments with a modest increase in total number of subjects. Pooling schemes in terms of replicates of subjects and arrays can be compared before experiments are conducted.

Background

Researchers are increasingly realizing the importance of *true* biological replicates for assessing statistical confidence in microarray experiments [1–6], but replication is often hindered by financial or technical constraints. One problem is that large, prefabricated microarray chips can be relatively expensive, driving up the total cost of an experiment. (In fact, the cost of a subject is often lower.) Another obstacle to using replicates is that the biological tissues from which RNA is extracted can often be of such

small quantity by nature that it is technically difficult to get enough RNA sample from one subject for hybridization to one array [3]. Either or both of these problems have motivated biologists to pool RNA samples together before hybridization. Many research papers using this method have been published [3–5], but the statistical properties of pooling have not been explicitly addressed. There are two different approaches of sample pooling. One is dubbed "complete pooling", where all samples from one treatment group are pooled onto one chip and there is no replication of chips for one treatment. This approach does not provide an estimate of variability among chips and therefore can not be used for statistical analysis. The other approach is dubbed "sub-pooling", where subsets of samples are randomly selected and pooled onto one chip but there are still multiple chips within each group. Here we investigated the statistical properties and the technical and financial implications of the second, sub-pooling approach.

Results

Simulation study of statistical characteristics

First, simulated microarray data were used to investigate the statistical properties (see Methods for details). Figure 1 shows the simulated power curves (for two-sample t tests) with respect to different α levels (i.e. type I error rates), effect sizes (i.e. differences of the means divided by common standard deviation), and sample sizes. Some well-known statistical properties are shown here: power increases with effect size and/or sample size when other conditions are fixed (e.g. curve 8 vs. curve 5, or curve 4 vs. curve 1); but it decreases when α level is lowered with other conditions being fixed (e.g. curve 8 vs. curve 4). However, the main purpose of Figure 1 is to demonstrate the effect of pooling. Consider curves 5 to 8, all of which have α level fixed at 0.05. Without pooling, nine replicates per group (curve 8) give much higher power than three replicates per group (curve 5). With pooling, power curves are intermediate. Both curve 6 and curve 7 reveal the power of pooling nine samples onto three chips each. The difference between them is that equal pooling with replicates in the same pool contributing equally (curve 7) has better power than non-equal pooling (curve 6). See methods section for statistical justification. Notice that power curve 7 is not very far from power curve 8, with only one third the number of chips being used.

Figure 2 shows that approximately equivalent power to non-pooling can be achieved if the number of gene chips is reduced but the number of samples is increased. The simulated power curves for two-sample t tests for different pooling schemes under the same type I error rate control are very similar. In this figure, n25c5 means that RNA samples from 25 subjects are randomly pooled onto 5 chips with 5 subjects contributing equally to each pool. Specifically, the power of (n25c5) ≈ power of (n24c6) ≈ power of (n22c11) ≈ power of (n21c7) ≈ power of (n20c20). Notice that power of (n25c5) ≈ power of (n20c20), which suggests that by randomly pooling RNA samples from 5 subjects onto one chip with equal contributions, we can decrease number of chips needed per group from 20 to 5 while number of subjects per group only needs to be increased from 20 to 25. As will be shown later, this may have great financial consequences.

"Virtual" pooling example using data from an Affymetrix microarray experiment

Blalock et al [2] investigated the correlation of gene expression with cognitive impairment in rats of different ages. Original data were from 29 Affymetrix microarrays. To further study the effect of pooling, we used 24 of the 29 arrays to do "virtual pooling". Namely, we first randomly selected 24 arrays and assigned them to two groups. P-values from two-sample t-tests on individual genes were recorded. Next, in each of the two groups, we "virtually" pooled the 12 samples onto 6, 4, 3 arrays respectively (assuming equal contribution in each pool). The statistical tests were then conducted on the "pooled" data with degrees of freedom being reduced to 10 (i.e. 2*(6-1)), 6 (i.e. 2*(4-1)), or 4 (i.e. 2*(3-1)), respectively. The P-values from the "pooled" data were then plotted against the original P-values. As shown in Figure 3, there is a good agreement of the P-values among the pooling schemes for the majority of the genes among the pooling and no pooling schemes. Table 1 shows that a moderate percentage of those genes that are found to be significant at a fixed α (0.05) level can be picked up by designs with pooling at the same α level. Multiple-testing adjustment was not considered here because it is a separate statistical issue. Note that "virtual" pooling serves as indirect evidence only.

Effect size considerations

Because we are testing hundreds or thousands of genes at the same time, it is not easy to determine the effect size needed for power or sample size estimation. Effect size can vary from experiment to experiment, from one kind of biological sample to another, as well as from gene to gene. If we set the goal of an experiment to detect all genes that are differentially expressed at any scale, then the sample size needed to do so would be prohibitively large. Therefore, we need to set up a realistic goal, say, to detect genes with effect size ≥ 0.5. For a two-sample t test, this means the true difference between the means is one half of the common standard deviation. Empirical data from many experiments suggest that this goal is often achievable. A pilot study is also desirable to obtain an estimate of approximate effect size for specific experiments and genes. To assist researchers in making such estimates, we include here some empirical results from several experiments

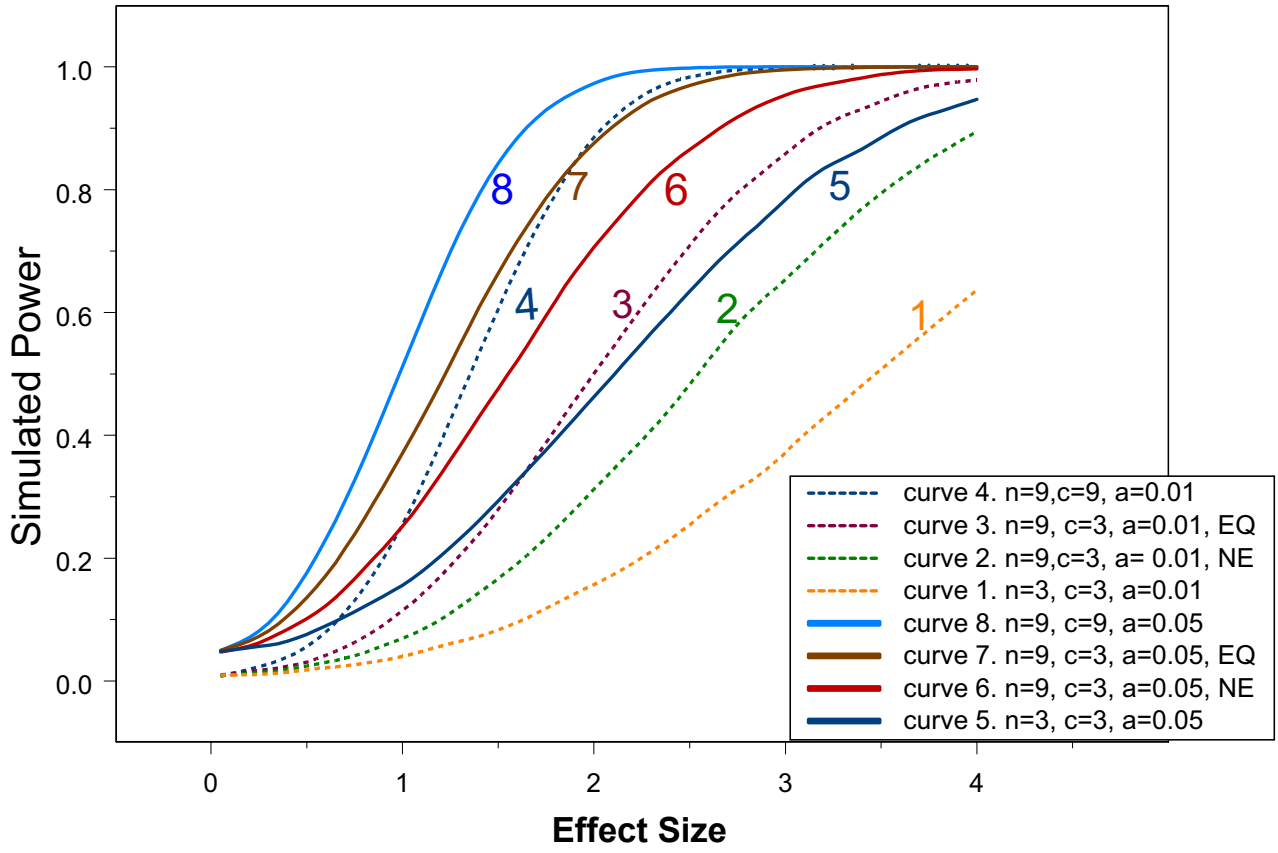

Figure 1
Relationships among type I error rate, sample size, effect size, and power with or without pooling. Note: *n* is the number of biological replicates per treatment group; *c* is the number of gene chips per group; *a* is the type I error rate; *EQ* means that samples are pooled with equal contribution; *NE* means samples do not contribute equally when pooled together (weights assigned randomly to each chip: 0.7, 0.2, and 0.1).

conducted in our core facility, which may provide a rough idea regarding the range of effect sizes that might be realistically expected with one type of microarray – the Affymetrix oligonucleotide GeneChip® array. Of course, the reader should be cautious in relying on these numbers, as they are estimates based on pilot studies from only one microarray core facility.

Financial implications of sample pooling and optimal pooling design

Since a microarray chip often costs more than a biological subject and we have more than one pooling design to achieve the same statistical control, it is possible for us to search for the most cost effective pooling design while maintaining the desired statistical properties. For a given combination of maximal type I error rate, minimal power, and the estimated or expected effect size to be detected, there are multiple pooling designs that may satisfy these conditions. We can compare the total cost of chips and subjects for each design and choose the one with minimal total cost, provided it is technical feasible. While this is a simplified economic framework (because it does not consider other cost in the experiment), it manifests the major point of the financial efficiency with pooling. An example function written in R language used to automate the searching and comparing process is attached as an additional file.

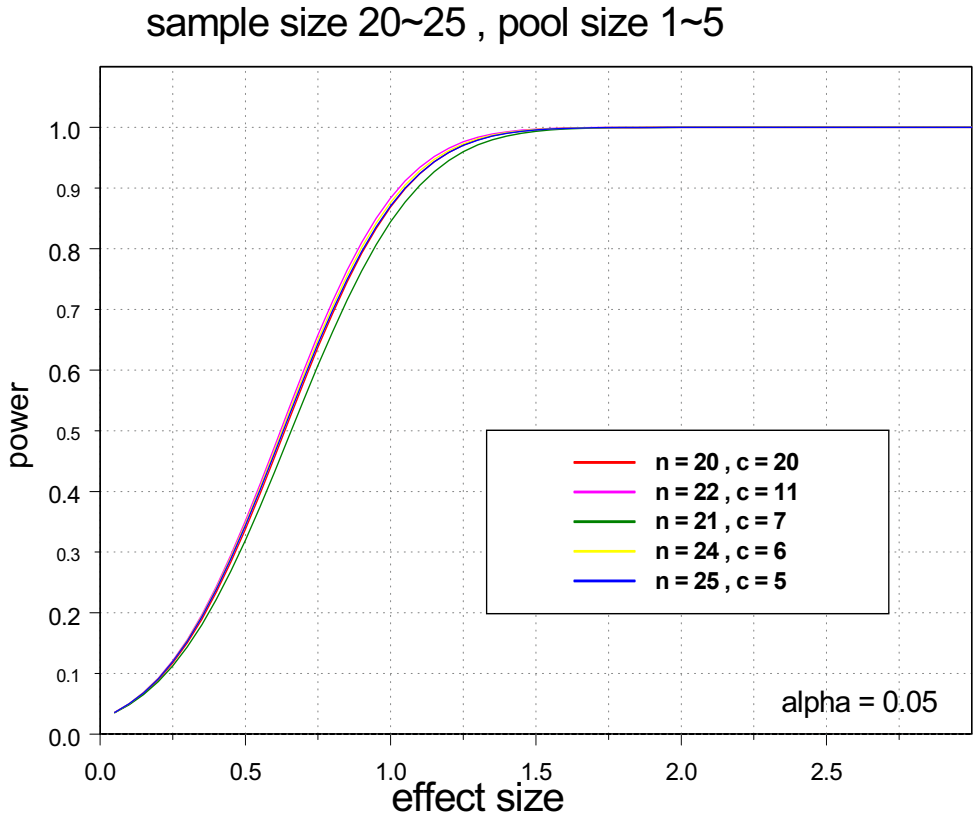

Figure 2
Approximately equivalent power curves under different pooling schemes. Power curves generated for two-sample t tests. Equal pooling assumed. Legend: *n* is the number of subjects per treatment group; *c* is the number of arrays per group. The five pooling schemes with different choices of number of subjects and number of arrays have approximately equivalent power curves when type I error rate is controlled at 0.05.

Discussion

Assessing variability on a per gene basis is an increasingly important aspect of microarray analysis. In the present paper we demonstrate that variance could be estimated accurately with different pooling schemes, and that the chip cost (and therefore experimental cost) can be dramatically affected by decisions regarding the pooling scheme employed. Many researchers are aware that replicates of biological samples are needed to assess experimental error. Even though thousands of genes are interrogated simultaneously on each chip, estimates of variance based on these measures do not of course include biological variance for any single gene in question. In this paper, we addressed the question of how many replicates are needed to achieve adequate power while considering an efficient technical procedure that is applicable in microarray experiments.

Since the specific design and data processing of microarrays can cause some confusion about what constitutes true replicates, it is important to distinguish technical replicates from biological replicates. In the Affymetrix microarray, 11 to 20 probe pairs are used to measure expression level of a probe set. It is important to note that these duplicated probe pairs are different measurement units rather than experimental units. The true biological replicates are the arrays (provided that they do not measure the same biological sample). An estimate of variability among the probe pairs reveals the precision of the measurement at the probe level but not the variability among biological samples. The latter is usually what is needed when

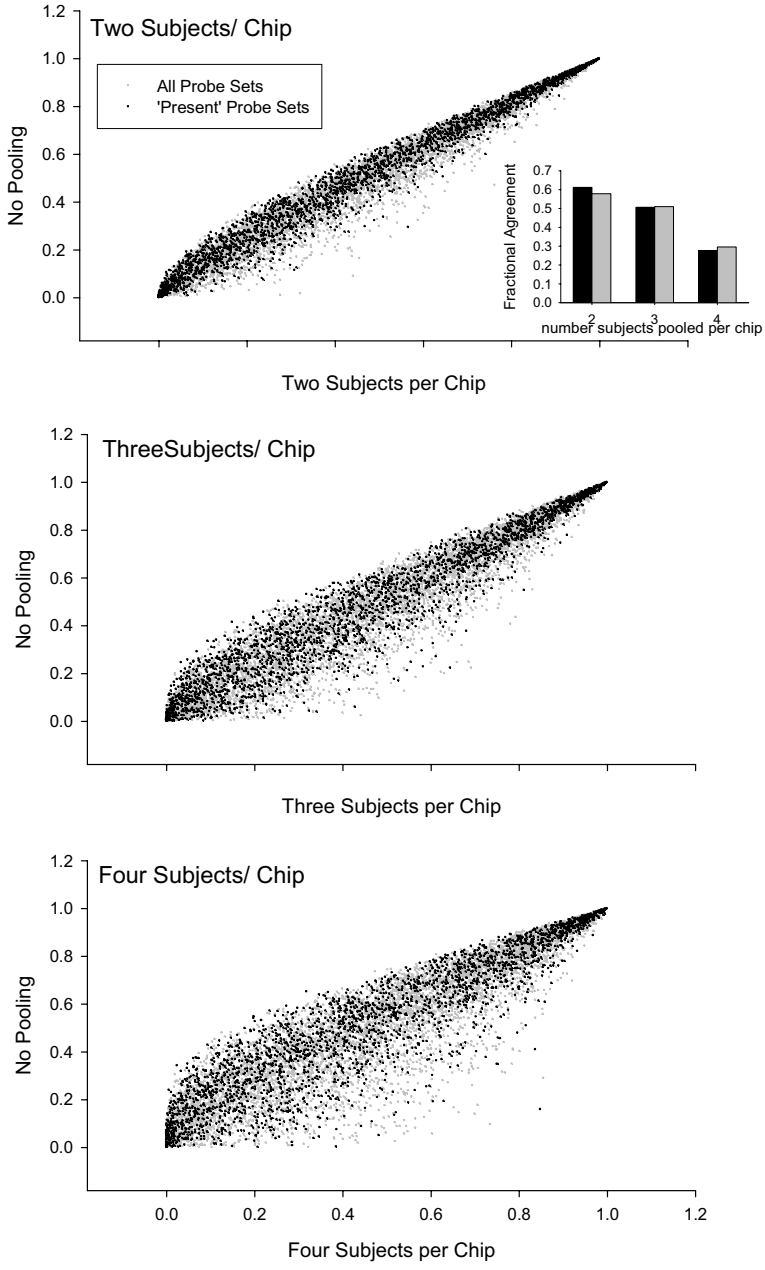

Figure 3
Scatter plots of the P-values with different "virtual" pooling schemes. On Y-axis are the P-values from two-sample t-tests for 8799 genes on the RGU34A gene chip with no pooling (12 subjects, 12 arrays per group). On X axis are the P-values from two-sample t-tests for pool size 2 (12 subjects, 6 arrays per group), pool size 3 (12 subjects, 4 arrays per group), and pool size 4 (12 subjects, 3 arrays per group), respectively.

Table 1: Agreement of significant genes between "virtual" pooling and no pooling with data from one real experiment. Note: total number of "genes" on the chip = 8799, α = 0.05. *Pool size*=number of subjects per chip (# subjects per group/ # chips per group)

Pool size	# of subjects per group	# of arrays per group	# of significant genes	% agreement between pooling and no pooling
1	12	12	228	
2	12	6	152	67%
3	12	4	108	47%
4	12	3	111	49%

conducting statistical hypothesis testing in order to make inferences about differential gene expression under various treatment or environmental conditions. Confusion between technical duplicates and biological replicates can sometimes lead to misconceptions in conducting and interpreting statistical tests. Based on the intensity readings on the probe pairs, Affymetrix Microarray Suite Version 5.0 gives a "change P-value" for the comparison analysis of each probe set ("gene") between a baseline chip and an experiment chip (MAS V 5.0 manual) based on Wilcoxon signed rank test. According to the manual, the change P-value indicates the significance of the difference between the experimental chip and the baseline chip. However, it should be noted that a significant difference of gene expression between a *single* chip from the experimental group and a *single* chip from the baseline group can not provide statistical evidence to show that the two groups are different. Often the goal of comparing a treatment group to a control group is to detect the true difference of the population means, and the "change P-value" is not the appropriate P-value for that goal.

To overcome the difficulty of estimating the variance of microarray data with few or no biological replicates, so-called "cross-gene models", "global error models" or "local error models" have been proposed [7,8]. According to such models, genes with similar expression levels from the same chip can be "borrowed" to construct a pseudo sample as the basis to estimate the "global" or "local" errors for each individual gene. This is very appealing because "Statistical group comparisons can now be done on experiments without replicates by using the global error model"[8]. However, these algorithms are based on two implicit assumptions that are difficult to support:

(1). The measurement of the expression of a gene on one single chip can be used to estimate the *true* population mean expression of that gene. This assumption is often false because an observed intensity can be far from the *true* mean with moderate or high probability. For example, even if gene expression measurements follow a perfect normal distribution, the chance that a measurement falls beyond one standard deviation from the mean on either side is as large as 32%.

(2). Genes with similar measurement values "share" the same variance within the treatment group. This assumption may be reasonably accurate for some genes (e.g. those with low expression intensities), but empirical observations show that it is clearly not true for all genes, especially for those with high intensity.

The cross gene error model references for support the two component error model proposed by Rocke et al [9] and Durbin et al [10] but those papers made it very clear that true replication could not be circumvented, especially to estimate variances for genes with high expression. (For additional discussion on different sources of error in microarray experiments, see [10–13].)

Except under very restrictive models, duplicate observations per chip do not provide valid estimates of variability among subjects. Hence, multiple chips per treatment group, each measuring an RNA sample from a separate biological subject or pooled groups of subjects, are required for statistical analysis. However, this approach often entails financial and/or technical difficulties. To address these issues, we tested RNA sample pooling and showed that it provides an efficient alternative solution. Because arrays are usually more expensive than subjects, sample pooling frequently may help defray the total cost of an experiment. The larger the difference between the cost of a single array and that of a single subject, the more the sub-pooling strategy will save. Simulated microarray data and "virtual pooling" of actual data utilized here as well as statistical theory suggest that the underlying principles of this proposal are sound. Note that even with pooling, we still show a requirement for reasonable replication of (pooled) arrays because there is no other way to assess biological variation.

A limitation of our current study is that we have yet to conduct a definitive experiment in which the same biological samples are compared with or without pooling. Although this has not been done, we nevertheless have indirect supporting evidences from multiple actual microarray experiments. After analyzing data from over 50 research projects (*with replications*) done in our facility, we have consistently seen smaller within-group variability

(and more genes with significant differences) in experiments using appropriate RNA pooling strategies than those with approximately the same number of arrays that did not employ pooling. This observation implies that experiments with pooling have greater power than those without pooling for a fixed number of chips, when conditions are comparable (i.e. similar experimental conditions, statistical methods and multiple testing correction procedures, etc.). In addition, a recent study [5] found high correlation between the intensities from the calculated pool and those from the actual pool using the same samples, (although in that study the pooling effect on reducing variance was not explicitly addressed).

RNA pooling may also have adverse consequences, and can be inappropriate in some cases. Pooling should not be used if inferences are needed for single subjects. For example, when the goal of the research is to correlate gene expression with some other variables measured at the subject level [2] or to identify gene profiles that help classify individual subjects and predict their membership in groups (e.g. cancer patients vs. normal patients). Pooling will also prevent later analysis of the data on variables that may have been ignored initially. As an example, suppose that we pooled all samples from the same gender and compared the differential expression of some genes between the two genders. This would make it impossible to subsequently analyze differential gene expression related to aging effects since samples from different ages would have been pooled together. The researcher must decide at the outset whether the potential loss of information is outweighed by the increased statistical power and cost-efficiency of the design. Similarly, pooling will prevent users from finding differences in expression that might divide only one set of samples. For example, if the goal of the experiment is to find genes that help differentiate subtypes of colon cancers rather than between colon cancer and healthy tissues, then it is inappropriate to pool cancer subtypes.

As pointed out by one reviewer, sample pooling is able to reflect group-specific variance, but assumes that residual individual variance is not influenced by the group variable. While this assumption may be a reasonable first approximation, it does not allow for the possibility of a cross-product relationship in which a group-specific variable (for example, toxic exposure) might lead to informative sub-states that exhibit separate groups of coordinately regulated responsive genes as a function of the extent to which the individual responds. Thus, there may also be loss of information regarding possible cross-product relationships by pooling. The experimenter should be aware of this caveat before deciding to pool samples.

There are also a few technical concerns with pooling. Theoretically, the more samples pooled, the greater the improvement in power. Practically, however, researchers may be reluctant to pool more than five samples to one chip due to technical limitations, and here we restricted our comparisons among pool sizes to no more than five. Further, we recommend pooling RNA samples instead of tissue or cell samples, because the variability at the tissue level is usually larger than at the RNA level. However, as long as equal contribution is assured, pooling at the tissue level should not make a big difference.

Finally, although we used only Affymetrix oligonucleotide data models as examples in this paper, the same principles should be readily applicable to two color cDNA arrays as well. In general, pooling should have similar advantages for more complex experimental designs (such as factorial designs, time course designs, etc.), when inferences are being made at the group level.

Conclusions

Appropriately designed RNA sample pooling can provide adequate statistical power, and improve efficiency and cost-effectiveness, for many types of microarray experiments when inferences are made at the group level. However, researchers have to consider the pros and cons of pooling for their experimental objectives. Designing optimal pooling schemes to achieve statistical control with minimal total cost is readily possible before the experiments are conducted.

Methods

Mixtures of Individual Gene Expressions

Gene expression levels for a single gene for n subjects will be denoted as $X_1,...,X_n$. If the RNA samples from p subjects are randomly selected and pooled, then we can model the resulting gene expression level as $X^* = \sum_{i=1}^{p} w_i X_i$, where $w_1,...,w_p$ are the mixing coefficients. These coefficients may be random. If the original gene expression levels, $X_1,...,X_n$, are identically and independently distributed random variables with mean μ and variance σ^2 and the mixing coefficients are independent of the individual gene expression levels, then we have

$$\sigma^2 / p \le Var\,(X^*) \le \sigma^2. \quad (1)$$

This result follows immediately from the fact that

$$E(X^*) = E(\sum_{i=1}^{p} w_i E(X_i)) = E(\mu \sum_{i=1}^{p} w_i) = \mu$$

and

Table 2: Observed effect sizes of data from some experiments with Affymetrix microarrays.

Study	Subject	# of arrays per group	Genome	% of genes with effect size ≥ 0.5
1	cell line	9	RGU 34	30.3
2	rat	10	RGU 34	18.1
3	rat	16	RGU 34	21.5
4	mouse	6	MGU 74	16.5
5	human	14	HGU133A	10.6
6	human	25	HGU133A	38.9

$$Var(X^*) = Var(\sum_{i=1}^{p} w_i X_i) = E(\sum_{i=1}^{p} w_i^2)\sigma^2.$$

Since $p^{-1} \le \sum_{i=1}^{p} w_i^2 \le 1$, (1) follows. Note that the lower bound is achieved when the mixing coefficients are equal; i.e., uniform mixing. This shows that RNA pooling reduces variance and the minimal variance of the mixture is achieved with equal pooling.

Relative Efficiency

As far as the number of chips needed is concerned, the relative efficiency of pooling n subject to n/p chips with uniform mixing can be computed as:

$$RE = \frac{(n+1)(n/p+3)}{(n/p+1)(n+3)} \frac{S^2_{pooling}}{S^2_{nopooling}} = \frac{(1+1/n)(1+3p/n)}{(1+p/n)(1+3/n)} \frac{S^2_{pooling}}{S^2_{nopooling}} \longrightarrow \frac{\sigma^2_{pooling}}{\sigma^2_{nopooling}} = \frac{1}{p}$$

as $n \to \infty$ and $p/n \to 0$. Hence, when p samples are equally mixed, we need approximately $1/p$ of the original number of chips to achieve similar power with pooling when both n is large and p/n is small. This is a large sample result and does not take into account the loss of degrees of freedom for the test. Next, we consider the effects of both aspects on statistical power.

Power Equivalence

Although pooling RNA samples with uniform mixing substantially reduces the variability among chips within a group, this may not result in an increase in power while testing differences among groups with a fixed number of subjects. Here we investigate the power of the two-sample t-test. Let $X_1,...,X_n$ be identically and independently distributed with $N(\mu_1, \sigma_1^2)$ and $Y_1,...,Y_n$ be identically and independently distributed with $N(\mu_2, \sigma_2^2)$. For simplicity, let $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Suppose we use two-sample t-test to test H_0: $\mu_1 = \mu_2$ against H_a: $\mu_1 < \mu_2$. Without pooling, the t statistic follows the central t distribution $f(t, df_1)$ with $df_1 = 2(n - 1)$ under H_0; while under H_a, it follows a non-central t with df_1 and non-centrality parameter $\delta = \frac{\mu_2 - \mu_1}{\sqrt{2\sigma^2/n}}$, which is inversely related to the standard deviation σ. For a fixed [illegible] and a critical value t_0 under H_0 such that $\int_{t_0}^{\infty} f(t, df_1)dt = \alpha$. To compute the power, we only need to find the rejection region corresponding to t_0 under the alternative non-central t distribution $g(t, df_1, \delta)$: $power = \int_{t_0}^{\infty} g(t, df_1, \delta)dt$. Now when the n samples are randomly assigned into pools of size p with equal mixing, power is decreased because the degrees of freedom are reduced to $df_2 = 2(n/p - 1)$, which makes the tails of the central t distribution under H_0 heavier and thus the critical value t_0^* will be further from 0, while the non-centrality parameter $\delta^* = \frac{\mu_2 - \mu_1}{\sqrt{2(\sigma^2/p)/(n/p)}} = \frac{\mu_2 - \mu_1}{\sqrt{2\sigma^2/n}} = \delta$ unchanged: (since the variance is reduced to σ^2/p and the number of chips reduced to n/p). However, if n + k subjects with pool size p are compared to n subjects without pooling, then the power change is not monotonic. The loss of power due to reduction of degrees of freedom will be partially compensated by gain of power due to increase of effect size. This can be seen by noting that $\delta^* = \frac{\mu_2 - \mu_1}{\sqrt{2(\sigma^2/p)/((n+k)/p)}} = \frac{\mu_2 - \mu_1}{\sqrt{2\sigma^2/n}} \frac{\sqrt{n}}{\sqrt{n+k}}$, which is [illegible] than $\delta = \frac{\mu_2 - \mu_1}{\sqrt{2\sigma^2/n}}$. This shifts the alternative distribution further from the null distribution and thus increases power. Hence we have heuristically shown that the power change is not monotonic when decreasing replication of chips while simultaneously increasing number of subjects. This implies that power can be preserved with the right choice of (subject) sample size and pool size. Many statistical software packages, such as R, S Plus or SAS, have built-in functions for both central and non-central t distributions and thus we can write programs to evaluate the power change or equivalent power curves under different pooling schemes. For two sided two-sample t test or experiments with more than 2 treatment groups, we can use the central and non-central F distributions rather than t distributions. The power equivalence should be very similar. Readers not very familiar with the above statistical concepts are referred to [14–16].

Simulation study

To investigate the above properties, we generated random samples based on two-treatment design and then compared the performance of different pooling schemes. We simulated microarray data as follows: randomly generate a data matrix of 5000 rows and 2n columns, with each

Table 3: Comparison of different pooling schemes and total cost using model data. Several pooling designs that can achieve power at least 0.8 while controlling type I error rate at 0.01 for an effect size of 1.0 are shown. Assuming a microarray chip costs $1000 and a subject costs $300, the total cost for each design is also computed and the optimal design with the minimal total cost is underlined. A function written in R (a free statistical software downloadable at http://www.r-project.org) to perform the above search automatically is attached as additional file.

Number of chips per group	pool size	power	Total cost
7	5	0.84	**35000**
8	5	0.91	40000
8	4	0.83	35200
9	4	0.89	39600
10	3	0.82	38000
11	3	0.87	41800
14	2	0.82	44800
26	1	0.82	67600

row represents one gene and each column represents one subject. The first n subjects are random samples from a normal distribution with mean μ_1 and standard deviation σ; the last n subjects are random samples from a normal distribution with mean μ_2 and standard deviation σ, where $\mu_2 = \mu_1 + \delta\sqrt{2}\sigma$. For each row, μ_1 was generated from a uniform distribution U(0, 30000) and standard deviation σ is set to be 0.2 * μ_1. The effect size δ was specified for every data matrix.

Simulation of power curves without pooling: for δ from 0 to 4.0, *n* from 3 to 30, we generated data matrices and then performed two-sample t tests on each row between the first *n* observations and the last *n* observations. For each specified δ, *n* and α level (0.05 or 0.01), we repeated the above simulation 1000 times and recorded the proportion of rejections, $R_{\alpha, \delta, n}$, for each iteration. We then calculated the averaged $R_{\alpha, \delta, n}$ as the simulated power.

Simulation of power curves with pooling: similar to the above with the only difference being that the two-sample t tests were performed on first *m* pools and last *m* pools, where each of the *m* pools is the average of *p* subjects randomly selected from the original *n* subjects. Note that *n* = *m***p* and we used sampling without replacement. This is to simulate the scenario of equal pooling.

Authors' Contributions

XP carried out the study. CLW and AJS supervised the study. EMB, KC, and PWL contributed discussions. All authors contributed to writing the manuscript. All authors read and approved the final manuscript.

Acknowledgement

Aspects of this study were supported by AG-10836 and AG-04542 from NIA to PWL and NIH-1P20RR16481-01 and NSF-EPS-0132295 to AJS. The authors thank the reviewers and editors for their helpful comments. We also thank Ms Donna Wall for her excellent technical support.

References

1. Lee MLT, Kuo FC, Whitmore GA and Sklar J: **Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations** *Proc Nat Acad Sci* 2000, **97:**9834-9839.
2. Blalock EM, Chen KC, Sharrow K, Foster TC and Landfield PW: **Gene microarray analyses of hippocampal aging: statistical profiling reveals novel expression programs correlated with cognitive impairment** *Journal of Neuroscience* 2003, **23:**3807-3819.
3. Pletcher SD, Macdonald SJ, Marguerie R, Certa U, Stearens SC, Goldstein DB and Partridge L: **Genome-wide transcript Profiles in aging and calorically restricted drosophila melanogaster** *Current Biology* 2002, **12:**712-723.
4. Miller RA, Galecki A and Shmookler-Reis RJ: **Interpretation, design, and analysis of gene array expression experiments** *J of Gerontol A Biol Sci Med Sci* 2001, **56:**B52-57.
5. Agrawal D, Chen T, Irby R, Quackenbush J, Chambers AF, Szabo M, Cantor A, Coppola D and Yeatman TJ: **Osteopontin identified as a lead marker of colon cancer progression, using pooled sample expression profiling** *J Natl Cancer Inst* 2002, **94(7):**513-521.
6. Pan W, Lin J and Le CT: **How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach** *Genome Biol* 2002, **3(5):**research0022.
7. **GeneSpring:** [http://www.silicongenetics.com]
8. **Clontech:** . http://www.clontech.com/techinfo/manuals/AtlasNavHelp/GlobalErrorModel.shtml, http://www.clontech.com/techinfo/manuals
9. Rocke DM and Lorenzato S: **A two-component model for measurement error in analytical chemistry** *Technometrics* 1995, **37(2):**176-185.
10. Durbin BP, Hardin JS, Hawkins DM and Rocke DM: **A variance-stabilizing transformation for gene-expression microarray data** *Bioinformatics* 2002, **18(Suppl 1):**S105-S110.
11. Nadon R and Shoemaker J: **Statistical issues with microarrays: processing and analysis** *Trends Genet* 2002, **18(5):**265-271.
12. Bakay M, Chen YW, Borup R, Zhao P, Nagaraju K and Hoffman E: **Sources of variability and effect of experimental approach on expression profiling data interpretation** *BMC Bioinformatics* 2002, **3:**4.
13. Welle S, Brooks AI and Thornton CA: **Computational methods for reducing variance with Affymetrix microarrays** *BMC Bioinformatics* 2002, **3:**23.
14. Casella G and Berger RL: **Statistical Inference** 2nd Edition. *Duxbury Press*; 2002.
15. Snedecor GW and Cochran WG: **Statistical Methods** 8th Edition.*Iowa University Press*; 1989.
16. Douglas C: **Montgomery: Design and Analysis of Experiments** 5th Edition.*John Willy & Sons*; 2000.