Partition Resampling and Extrapolation Averaging: Approximation Methods for Quantifying Gene Expression in Large Numbers of Short Oligonucleotide Arrays

Darlene R. Goldstein
École Polytechnique Fédérale de Lausanne (EPFL)
Institut de mathématiques
Bâtiment MA, Station 8
CH-1015 Lausanne, Switzerland
Darlene.Goldstein@epfl.ch

14 June 2006

Contents

Chapter 1

Introduction

Microarray technologies measure mRNA abundance for thousands of sequences (or 'genes') in parallel. The high throughput nature of microarrays has contributed to their rise in importance for studying the molecular basis of fundamental biological processes and complex disease traits. They are now regularly used in a variety of biological and medical studies.

Several different types of microarrays are available. Studies of gene expression using high-density short oligonucleotide arrays (or 'chips'), such as those made by Affymetrix or NimbleGen have become standard in a variety of biological contexts. Examples include plant and animal studies as well as clinical research, particularly in cancer. Of the expression measures that have been proposed to quantify expression in these arrays, multi-chip-based measures have been shown to perform well (Bolstad *et al.*, 2003). As gene expression studies increase in size, however, utilizing multi-chip expression is more challenging in terms of computing memory requirements and time.

A strategic alternative to exact multi-chip quantification on a large chip set is to approximate expression values based on subsets of chips. This paper introduces extrapolation and resampling methods for approximate quantification of expression in large studies. An examination of the properties indicates that these methods can perform well compared to exact quantification. The focus is on short oligonucleotide chips, but the same ideas apply equally well to any array type for which expression is quantified using an entire set of arrays, rather than for only a single array at a time.

Chapter 2

Gene expression quantification

2.1 Expression measures

Affymetrix GeneChip ® arrays contain several (usually 11 – 20) 25-mer oligonucleotides used to measure the abundance of a given target sequence, the perfect match (PM) probes, as well as an equal number of negative controls, the mismatch (MM) probes. The set of probes for a given target sequence is called a probe set. A single fluorescently labeled sample is hybridized to the array which is then scanned with a laser, yielding absolute measures of fluorescence intensity. The intensities are indicative of the amounts of mRNAs containing the target sequence in the sample, and thus provide a means of quantifying levels of gene expression. Conversion of probe level signal intensities to an expression measure can be viewed as a multistep process comprising background correction, normalization and probe set summarization.

There exist several methods for converting the raw signal intensities to measures of gene expression. Some methods work on chips singly, but many quantify expression on multiple chips together as a set. Those currently in common use include: MAS 5/GCOS (Affymetrix, 2001); the Li-Wong Model-Based Expression Index (MBEI), implemented in the software dChip (Li and Wong, 2001); and the Robust Multichip Average (RMA) (Irizarry *et al.*, 2003a) and variant gcRMA (Wu and Irizarry, 2005b), implemented respectively in the `affy` (Irizarry *et al.*, 2006) and `gcrma` (Wu and Irizarry, 2005a) packages of the BioConductor Project (Gentleman *et al.*, 2004). A relatively new algorithm produced by Affymetrix is the probe logarithmic

intensity error method (PLIER) (Affymetrix, 2005). For comprehensive information on these and other expression measures, as well as a comparison of methods, see `http://affycomp.biostat.jhsph.edu/` (Cope *et al.*, 2004; Irizarry *et al.*, 2005). It is easily seen that no method performs best under every circumstance, but that a few methods stand out as providing a reasonable balance between bias and variance.

2.2 Method advantages and drawbacks in large studies

In very large studies, consisting of hundreds or even thousands of chips, the choice of expression measure involves consideration of not only the performance properties of the method but also computational issues.

Single chip measures, such as MAS 5, are computationally fast and require no additional RAM for quantification of multiple chips. Once a target scaling value has been chosen, expression may be quantified on individual chips without waiting for the complete set.

A problem with MAS 5 as an expression measure, though, is that the variance is not stable for low expressed genes. This variance inflation results in an increase in false positive differential expression calls (Cope *et al.*, 2004; Irizarry *et al.*, 2005). Using a variance stabilization procedure in addition to MAS 5 improves this aspect, but then quantification is no longer strictly a single chip method and the benefits of single chip methods are thus reduced.

In calibration-type comparison studies with 'known' truth, RMA has been demonstrated to provide an improved measure of expression over several other measures, and has since gained in popularity as a measure of expression (Irizarry *et al.*, 2003a,b; Bolstad *et al.*, 2003). The variant gcRMA (Wu and Irizarry, 2005b) is also becoming more commonly used.

However, even with recent algorithmic improvements, for many users on typical machines the available RAM limits the number of chips that may be quantified using current implementations of RMA and gcRMA. The desirability of using multi-chip methods on large sets of chips, combined with the problems of hardware and software limitations, calls for a fresh approach to gene expression quantification.

Chapter 3

Subset strategies for large studies

In large studies, computational difficulties may preclude gene expression quantification by multi-chip methods. The major obstacle is the amount of RAM required for the quantification algorithm: if the user's machine does not have sufficient RAM for the chosen method, it simply is not possible to obtain gene expression measures for the chip set. More efficient implementations will raise the number of chips that can be quantified on a machine with a given amount of RAM, but some limit on the number of chips may be reached with even the most efficient algorithm. The number of chips which can be simultaneously quantified depends not only on machine specifications and algorithm but also on type of chip.

Table 3.1 gives some indication of the number of chips which can be quantified together on one machine. This study is a modification of one available at `http://www.stat.berkeley.edu/~bolstad/ComputeRMAFAQ/size.html`, which assessed quantification of varying numbers of the HG U95Av2 chip on machines with 1 GB of RAM. Here, increasing numbers of chips for two chip types, HG U95Av2 (12,625 probe sets) and HG U133A (22,283 probe sets), are quantified using `justRMA` (Irizarry *et al.*, 2006) on a machine with the Windows XP Professional operating system, a Pentium M 760 2.0 GHz processor and 2 GB of RAM. On this computer, a maximum of 425 – 450 HG U95Av2 chips or 300 – 325 HG U133A chips could be quantified together. There are already studies larger than this in progress. As well, newer generations of chips tend to include more probes, decreasing the number of chips that can be quantified together. Finally, as this machine may be

better equipped than the 'typical' analyst's desktop, these estimates may be optimistic for many users.

Table 3.1: *Time (in seconds) to compute RMA values using justRMA with 2 GB RAM. – = not done; X = failed due to memory limitations.*

# chips	HG U95Av2	HG U133A
100	253.21	196.67
200	322.15	399.57
300	619.85	628.97
325	–	X
400	650.75	X
425	713.43	X
450	X	X

Using a computer with a larger amount of RAM, as well as an operating system with efficient memory use, raises the effective number of chips that can be quantified as a set. However, access to very high-end machines is outside the reach (and budget) of many analysts, who require immediate solutions to the problem of large chip set quantification.

One work-around that has been suggested is to use a subset of chips as a basis for multi-chip quantification of the entire set. There are several possibilities for how this may be carried out. The aim is to produce a $p \times n$ matrix of expression measures, one for each probe set ($i = 1, \ldots, p$) in each sample ($j = 1, \ldots, n$).

The methods are illustrated and compared on the *ALL* dataset, publicly available from St. Jude Children's Hospital, using the expression measure RMA.

The RMA expression measure is based on a log scale linear additive model (Irizarry *et al.*, 2003a). The $\log_2$ of background-corrected, quantile-normalized PM intensities can be written as the sum of $\log_2$ chip expression value e_i and $\log_2$ probe affinity a_j (plus random error ε_{ij}). In the notation of Irizarry *et al.* (2003a), $T(PM_{ij}) = e_i + a_j + \varepsilon_{ij}$, for chip $i = 1, \ldots, I$, probe $j = 1, \ldots, J$, and where T is the transformation that background corrects, normalizes and logs the original PM intensities.

It should be emphasized that the methods described here do not depend on RMA; they apply equally well to any multi-chip expression measure. All

analyses reported here were coded in the R (2.3.0) statistical programming environment (R Development Core Team, 2006) along with the BioConductor (release 1.8) packages (Gentleman *et al.*, 2004) `affy` (Irizarry *et al.*, 2006) and `multtest` (Pollard *et al.*, 2004, 2005).

3.1 Dataset description

The data consist of 335 Affymetrix HG-U95Av2 chips (12,625 probe sets) hybridized as part of a study of pediatric acute lymphoblastic leukemia ALL (Yeoh *et al.*, 2002). The samples comprise 9 types of ALL along with some Normal samples. The data are available at `http://www.stjuderesearch.org/data/ALL1/`.

Although not a massive sample size, there are still enough arrays to elude full chip set quantification on many machines. The dataset is useful as an illustration because the number of chips is large enough to demonstrate the utility of the method, while at the same time sufficiently small that RMA expression values can be computed by the full multi-chip method on better machines. The RMA values computed on all chips together provide a useful baseline for comparison, henceforth referred to as the 'true' values.

3.2 Subset strategy: Extrapolation

In the *Extrapolation* strategy, a 'fitting subset' is selected from the the full chip set for fitting the multi-chip model. The size of this subset should be chosen so that all chips can be quantified together (*i.e.* should not exceed the number of chips that the machine can accommodate). The set should also be sufficiently large and representative that model parameters may be well estimated. For example, in the ALL dataset the fitting subset might contain 50 chips; the remaining $335 - -50 = 285$ chips comprise the 'extrapolated' subset. A representative sample may be obtained by stratified sampling of the original chips, so that the fitting set contains the different types in roughly the same proportions as the full dataset.

The model fitting results in expression measures of each probe set for each sample in the fitting subset. The estimated model is then applied to the remaining chips to yield expression measures on all probe sets for samples in the extrapolated subset. The Extrapolation strategy is described here and

depicted in Figure 3.1(a).

Obtaining RMA values requires background correction, normalization and probe set summary via the model. Background correction is a one chip at a time operation, and therefore does not require subsetting of chips. Each chip is therefore background-corrected with the default RMA background correction (Irizarry *et al.*, 2006). Quantile normalization is a multi-chip operation. The extrapolation strategy computes the normalizing transform on the fitting subset, and applies it to the extrapolated subset. First, the fitting subset is quantile normalized (Bolstad *et al.*, 2003). Then, for each extrapolated chip, the background-corrected PM intensities are ranked and the probes are assigned the corresponding normalizing intensity determined from the fitting subset. Finally, the RMA model is estimated on the fitting subset. Assuming that the probe effects a_j are constant across chips, the chip effect (expression value) may be estimated each chip as follows: (1) for each probe $j = 1, \ldots, J$ in a given probe set on a single chip i, compute the residual $r_{ij} = T(PM_{ij}) - \hat{a}_j$, where $\hat{a}_j$ is estimated by median polish; (2) the median over j of the r_{ij} gives an estimate of the expression value on chip i for that probe set. (Estimates other than the median of the residuals may instead be used in step (2) above. For example, a two-stage weighted least squares estimate of expression has been suggested (Collin, 2004).) Operations (1) and (2) are carried out for each probe set on each chip, resulting in RMA values for the chip in the extrapolation subset.

Extrapolation has the advantage that expression can be quantified before all samples have been collected, thereby allowing for preliminary analyses in the case of large studies taking place over a long period of time. In addition, chips do not require requantification as more samples arrive (as is the case for a full multi-chip method). However, if this strategy is used before all chips are available, representativeness of the fitting set to the full set cannot be assured.

Extrapolation also has some appeal as a step toward 'context independence'. That is, expression measures obtained by extrapolation are not dependent on which particular chips are in the extrapolated set. Thus, any chip in the extrapolated set would report the same expression values regardless of which other chips are analyzed with it. Expression does of course depend on the chips in the fitting set, but in some applications (*e.g.* pharmaceutical studies) a set of reference standards may exist.

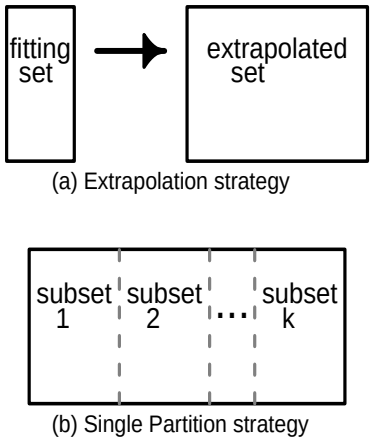

Figure 3.1: *Representations of Extrapolation and Single Partition strategies.*

3.3 Subset strategy: Single Partition

A slight variation of the Extrapolation strategy involves partitioning the entire chip set into a single set of subsets of similar size, or *Single Partition*, shown in Figure 3.1(b). Again, subset size should be such that the chips within a subset may be quantified with a multi-chip method. Ideally, each of the separate subsets would also be representative of the full set. As above, this may be achieved by stratification.

Separately for each subset, expression measures are obtained on all probe sets via a multi-chip method (*e.g.* RMA) for each sample contained in the subset. For example, the *ALL* dataset may be partitioned into 7 subsets each of size about 50. The full gene expression matrix is obtained by simply rejoining the individual subsets.

3.4 Problems with Extrapolation and Single Partition

Extrapolation and Single Partition strategies are straightforwardly simple and generally fast to compute. The problem of insufficient RAM is avoided by choosing the fitting or partition subset size to be smaller than the maximum number of chips the machine can simultaneously process.

One adverse property of the Extrapolation strategy is that the fitting

subset characteristics are 'locked in', then propagated to the extrapolated subset. This aspect is problematic if the fitting subset is not representative of the full set, or if it is flawed in some other, perhaps unknown, way. The Single Partition strategy has this problem as well although to a lesser degree: there is 'lock in' but no propagation, as each subset is quantified separately from the others.

Ideally, the expression values obtained by a suitable subsetting strategy would match those produced by multi-chip quantification of the full chip set. However, expression values depend on the specific chips contained in the subsets. Both the Extrapolation and Single Partition strategies exhibit some sensitivity to the choice of subset.

Figure 3.2 illustrates the variability in measured expression across probe sets for a single chip quantified in different partitions. A partition was generated by dividing the full chip set into 7 subsets of size approximately 50 (6 subsets of size 48 and 1 of size 47), either at random (panels (a) – (c)) or with stratification based on subtype (panels (d) – (f)). Single Partition expression measures were computed as described above. The process was repeated several times, yielding additional sets of Single Partition expression values.

The plots compare for a single chip the RMA values ($\log_2$ scale) computed from 6 different Single Partitions to the true values. In each of the 6 subplots, the difference between the partition value and the true value is plotted against the true value. If a partition produced the true expression values, then the points would lie on the horizontal line centered at 0 (dashed gray line).

It is readily seen that Single Partition values deviate from the true values, sometimes markedly. Variability across partitions of expression values for the same chip can be seen by comparing the subplots. It is also seen that there can be substantial bias within Single Partitions. Figures 3.2(a) – (e) show pronounced bias, whereas Figure 3.2(f) shows relatively little bias. The patterns are similar for stratified and unstratified partitions but there is typically less variability with stratification, occasionally drastically less (Figure 3.2(f)).

The issues of variability and bias with both the Extrapolation and Single Partition strategies are sufficiently serious to discourage their widespread use. However, with modification based on averaging the strategies become more viable.

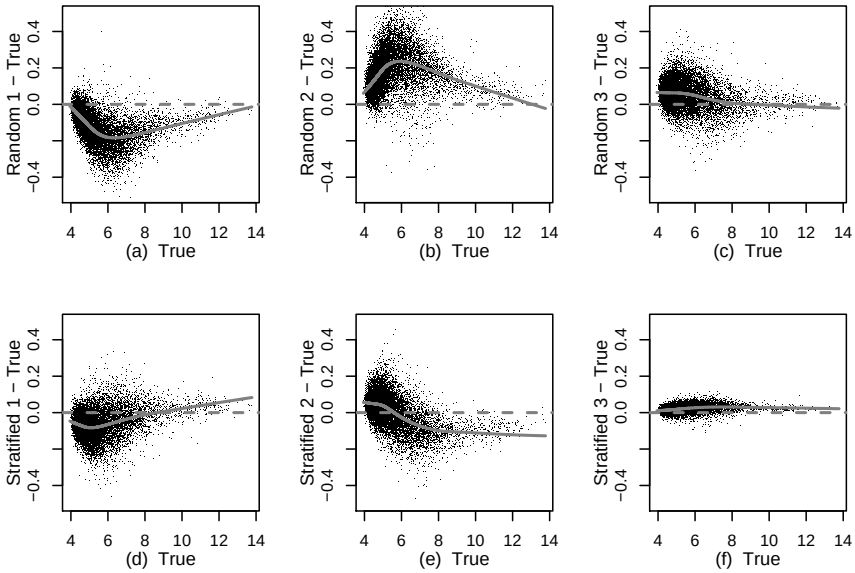

Figure 3.2: *Difference between Single Partition RMA values and True RMA values vs. True for a single chip from 6 different Single Partitions. Each point represents a probe set. Solid line is a loess fit.*

3.5 Subset strategy: Extrapolation Averaging

The potentially poor performance of extrapolation may be alleviated by drawing on the power of averaging. To diminish 'lock-in', we may perform the subsetting and extrapolation step multiple times and average the resulting expression measures, a strategy we refer to as *Extrapolation Averaging* (EA). Thus, we would expect that a few unfortunate fitting subsets should not have a strong adverse impact on the final expression measures, which are averaged from extrapolations from multiple instances of fitting sets. This strategy will be most practical if the majority of chips for the complete study are already available (and not, for example, in the early stages).

3.6 Subset strategy: Partition Resampling

Various partitions are possible for any given (full) chip set and subset size. The Single Partition strategy selects, randomly or deliberately, only one of the many possible partitions as a basis for computing expression values. We can instead take advantage of the power of averaging with an alternative strategy that will be referred to as *Partition Resampling* (PR).

The total number of possible partitions of a large set will be very large, and infeasible to enumerate and use in computation of expression. We may, however, sample a subset of the possible partitions as a basis for expression quantification. Partition Resampling applies the Single Partition strategy on multiple randomly generated partitions, then averages the resulting expression matrices across partitions to produce its gene expression matrix.

Chapter 4

Partition Resampling and true values

Conceptually, the EA and PR strategies work in a similar fashion. Expression is based on an average of expression values for the given chip based on different subsets. The strategies differ in detail though: EA uses one model within a dataset, while PR uses different models for the different subsets within a dataset.

PR is very simple to implement and automate, as it only requires an implementation of the desired quantification algorithm and a random number generator. For EA, the quantification algorithm needs to be reorganized so that the multi-chip aspects may be reduced to single chip operations. Memory management can also be more problematic.

In terms of performance, PR and EA appear to behave broadly similarly. Thus, detailed results are only shown for PR. To avoid an overly optimistic assessment, results here are based on random rather than stratified samples. Appropriate stratification generally provides faster convergence to the true expression values.

4.1 Expression values

The initial examination compares expression values from full data (true values) to those from PR for varying number of resampled partitions and subsets of varying size. For each combination of subset size and number of resampled partitions, PR-RMA values are obtained for each probe set on each chip. As

a reminder, each probe set entry in the PR-RMA expression matrix for a chip contains the average of the RMA values from the resampled partitions.

Results presented here are for partition subset size 48 (with one subset of size 47), and number of resampled partitions equal to 1 (*i.e.* Single Partition), 5, 10, 20, 50 and 100. Results are illustrated for one chip, which typifies the findings from the set; the same gross trend occurs for chips throughout the entire *ALL* dataset.

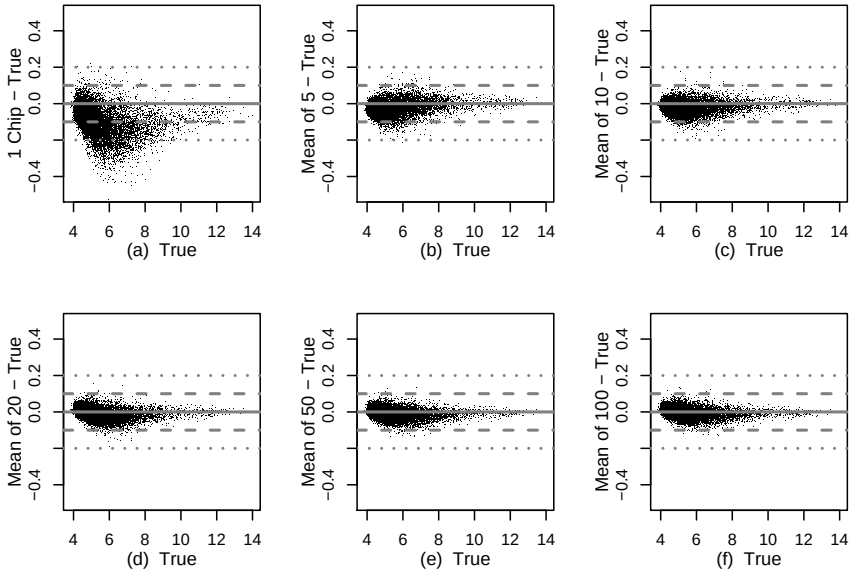

Figure 4.1: *Difference between PR-RMA values and True RMA values vs. True for a single chip for varying number of resamples. Each point represents the pair of values for a probe set.*

Figure 4.1 displays the difference between PR-RMA values and true values versus true values for the 6 resampling values. The subpanels are plotted on the same scale and include reference lines to facilitate comparison. The decrease in the deviation from true values between a single partition (panel (a)) and the mean of even as few as 5 resamplings is striking. The variability is further reduced with increased resampling, although at a decreasing rate. Ignoring the slight dependence between chips within the same partition subset induced by the finite chip set size, we may consider $\sqrt{n}$ as a benchmark for the decrease in variability of the mean (here n is the number of resampled

partitions). The observed narrowing of the cloud of points appears roughly consistent with this rate.

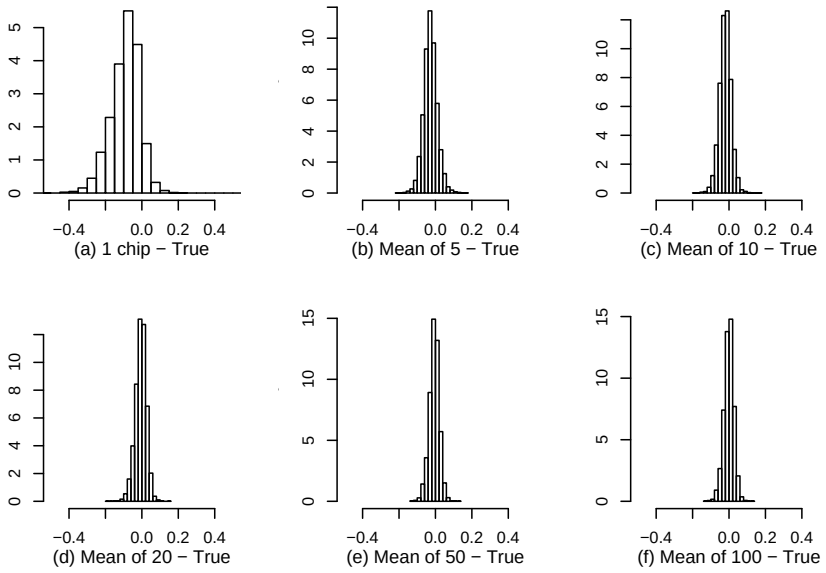

Figure 4.2: *Histograms of differences between PR-RMA and True values across probe sets of a single chip with varying number of resamples.*

In addition to lower variability, there is also an apparent 'central tendency' behavior of PR-RMA values with increasing number of resampled partitions (Figure 4.2). Expression values that are approximately unbiased should appear as a histogram roughly symmetric around 0. The marked asymmetry in the Single Partition (panel (a)) decreases with additional resampling (panels (b) – (f)).

4.2 Test statistic and p-value comparison

We have seen that quantification by the subset approximations considered here results in expression values which vary somewhat from the values that would be obtained by full multi-chip computation. However, what may well be of greater interest is the extent to which subsequent inference based on the approximate values is affected. If the conclusions drawn from the data are the

same for both true and approximate expression values, then the variability of individual expression values due to approximation is of little import.

There are several types of inference that might be made in a gene expression study. These include identification of differentially expressed genes, ranking of genes warranting further examination, choice of genes for building a classifier, and identification of novel subtypes. As an example, we consider here the problem of identifying genes differentially expressed between ALL types.

There are many possible test statistics to use for identifying differentially expressed genes. As this is not a study on the performance of such test statistics but rather an examination into the performance of the PR-RMA approximation compared to exact full RMA, a simple to compute criterion with acceptable operating characteristics suffices here. We consider 12 two-sample t-tests: subtype versus normal (9 different subtypes), and 3 other tests with different sample sizes (large versus large, small versus small, large versus small). Because rather similar patterns occurred for all tests, results are shown here for only one.

We are unable to examine true and false positive identifications of differential expression by PR-RMA, as the true status is unknown. However, we are able to compare t-statistics and corresponding nominal, unadjusted p-values obtained from PR-RMA with the 'true' RMA values obtained on the full data. In this way, we can see whether the same inference regarding differential expression would be made by both the approximate and exact methods.

The full data two-sample t-statistic is computed in the standard way based on full RMA values. The PR-RMA t-statistic is similarly obtained, but is instead based on the PR-RMA gene expression matrix. It should be noted that the PR-RMA t (and corresponding p-value) is *not* obtained by averaging the individual partition t-values across partitions. Rather, the PR-RMA expression matrix is an average across partitions; the PR-RMA t is based on these (averaged across partitions) expression values.

Figure 4.3 shows the comparison between the PR-RMA based t-statistic and the 'true' (full data) t for the t-test comparing subtype $T - ALL$ to Normal. If the PR-RMA based t were exactly equal to the full data RMA t, all points would lie on the horizontal line at 0. Agreement clearly increases with the number of resamples.

Since inference is often based on (rankings of) p-values, it is useful to look at p-value agreement as well. As above, note here that the PR-RMA

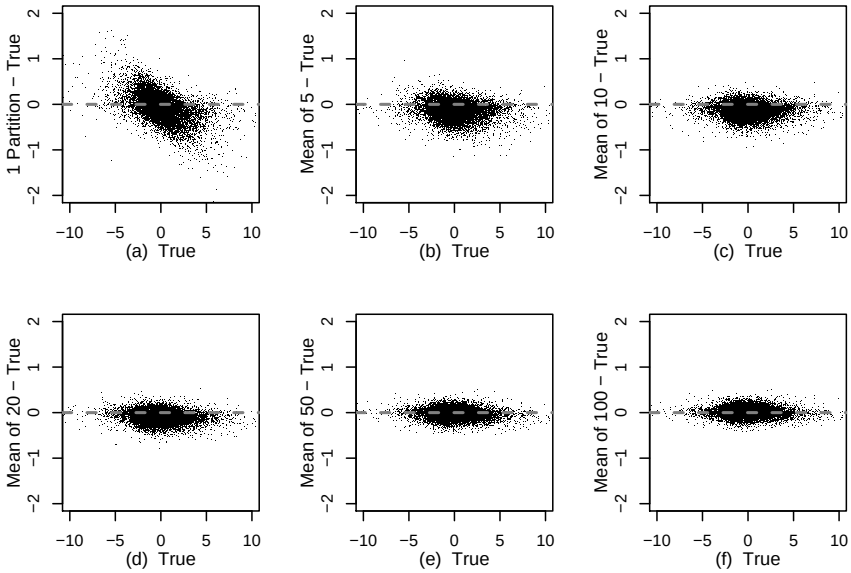

Figure 4.3: *Difference between PR-RMA t-statistics and True t vs. True t for the test subtype $T - ALL$ vs. Normal for varying number of resamples. Each point represents a probe set.*

p-value is *not* an average of p-values; it is the p-value corresponding to the PR-RMA t-statistic. A specific example of the general trend of $(-\log_{10})$ p-value agreement observed in the 12 tests is shown in Figures 4.4 (nominal, unadjusted p-value) and 4.5 (p-value after Bonferroni adjustment; other multiplicity adjustments give very similar results).

In Figure 4.4, the probe sets corresponding to the points to the right of the vertical line at 2 and above the horizontal line at 2 are those which are found significant (nominally at level $\alpha = 0.01$) with exact full data quantification, but not in the approximation — these probe sets are the ones that would be identified as differentially expressed with full data RMA, but are missed with PR-RMA (false negatives). Similarly, probe sets corresponding to points in the region to the left of the vertical line at 2 (not significant at $\alpha = 0.01$ with full data RMA) and above the horizontal line at 2 are false positives.

False negative and false positive numbers at two thresholds are indicated on each subplot of Figure 4.4, and for a threshold of $-\log_{10}(.05) \approx 1.3$ in Figure 4.5. For example, in panel (a) (Single Partition) there are 320 false negatives and 34 false positives at a $-\log_{10}$ threshold of 2 ($\alpha = 0.01$), and

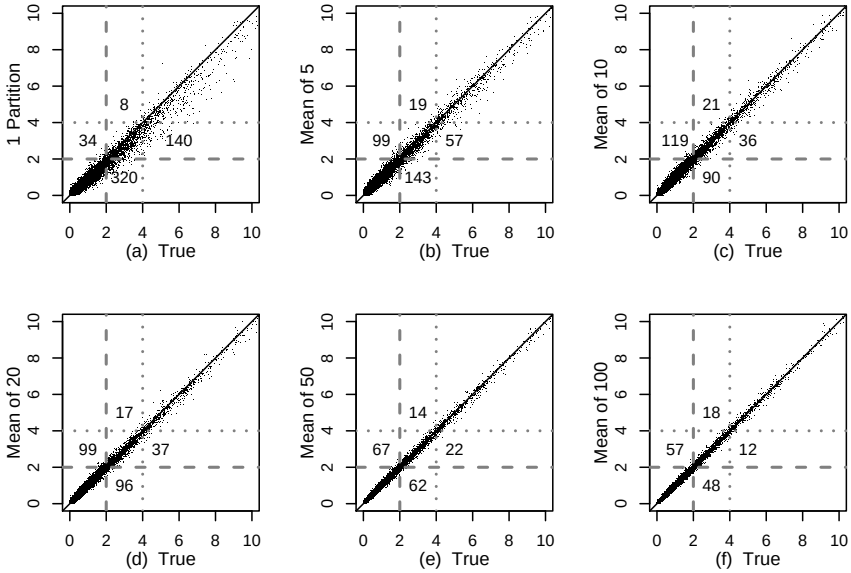

Figure 4.4: $-\log_{10}$ *PR-RMA p-value vs.* $-\log_{10}$ *True p for the test subtype* $T-ALL$ *vs. Normal for varying number of resamples. Numbers indicate how many points are in the false negative and false positive regions for 2 thresholds. p-values are nominal, unadjusted values. Each point represents a probe set.*

140 false negatives and 8 false positives at a $-\log_{10}$ threshold of 4 ($\alpha = 0.0001$).

Increasing the number of partitions for a given subset size tends to reduce the total number of false negative and false positive results. In addition, the false negative and false positive rates tend to become less mismatched with an increasing number of resampled partitions. This finding indicates that there is decreased bias with increased resampling, with points falling above or below the diagonal due mainly to sampling variability.

Table 4.1 summarizes different error rates for this test for unadjusted and adjusted p-values. The rates are defined for a given significance threshold as follows: α = number of PR significant genes but not true/number of true nonsignficant genes; FDR = number of PR significant genes but not true/number of PR significant; β = number of PR nonsignificant genes/number of true significant genes. For example, for a Single Partition with a cutoff of 0.0001, of the 12,625 probe sets there are 668 true significant genes, 140 of which are also PR significant, 8 PR significant genes which are not true significant,

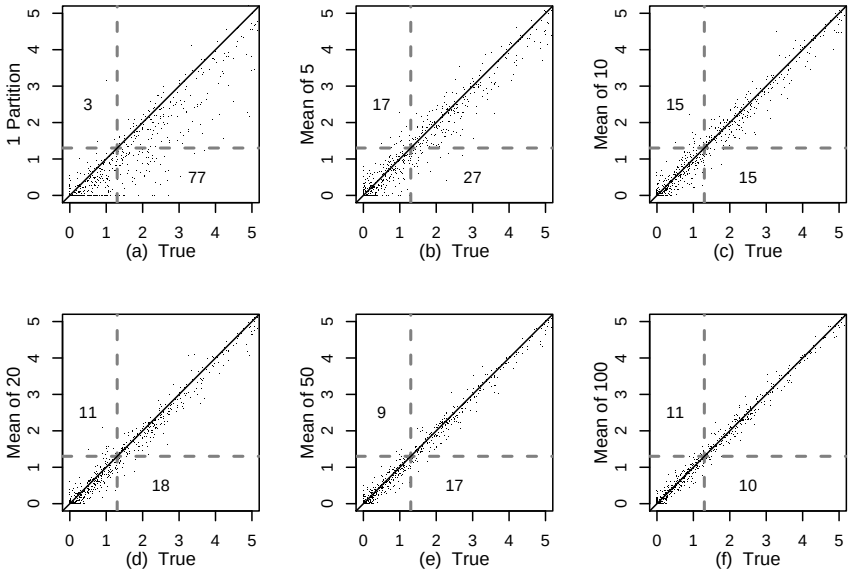

Figure 4.5: $-\log_{10}$ *PR-RMA p-value vs.* $-\log_{10}$ *True p for the test subtype* $T-ALL$ *vs. Normal for varying number of resamples. Numbers indicate how many points are in the false negative and false positive regions for a threshold of* $p = 0.05$*. p-values are Bonferroni-adjusted. Each point represents a probe set.*

and a total of 536 PR significant genes. The corresponding rates are $\alpha = 8/(12{,}625 - 668) = 0.00067$, FDR $= 8/536 = 0.015$, and $\beta = 140/668 = 0.21$.

In this instance the Single Partition had slightly lower α and FDR values than PR. However, these rates are subject to random fluctuation so that for a different Single Partition they could instead turn out to be higher. In addition, the lower false positives are at the cost of a greatly increased false negative rate (reduced power). We therefore cannot rely on Single Partition to provide smaller false positive/discovery rates or a reasonable tradeoff between false positive and false negative results.

In all cases, the false negatives and false positives tend to occur quite close to the threshold. Thus while agreement is not perfect at the threshold, we can be reasonably confident of agreement for probe sets at the top of the differential expression list.

Table 4.1: *Error rates for A: unadjusted p-values (cutoff = 0.0001; # True sig. = 668); B: Bonferroni-adjusted p-values (cutoff = 0.05; # True sig. = 353)*

# partitions	# PR sig.	α	FDR	β
A: for nominal, unadjusted p-values				
1	536	0.00067	0.015	0.21
5	630	0.0016	0.030	0.085
10	653	0.0018	0.032	0.054
20	648	0.0014	0.026	0.055
50	660	0.0012	0.021	0.033
100	674	0.0015	0.027	0.018
B: for Bonferroni-adjusted p-values				
1	279	0.00024	0.011	0.22
5	343	0.0014	0.050	0.076
10	353	0.0012	0.042	0.042
20	346	0.00090	0.032	0.051
50	346	0.00073	0.026	0.048
100	355	0.00090	0.031	0.028

Chapter 5

Conclusion

Exact multi-chip expression quantification on full chip sets is not always feasible in large studies. Currently in progress are several studies large enough to prohibit exact calculations (*e.g.* for RMA). The strategies introduced here provide useful approximations to the exact value based on the full chip set.

Although there are situations for which (single) Extrapolation may be the most attractive strategy, the averaging strategies (Partition Resampling and Extrapolation Averaging) behave more favorably in general.

PR and EA enjoy the reduced variability obtained through averaging along with an apparent bias reduction. A common criticism of resampling methods is that, based on a single sample, they should not be used to generalize to a larger population. This objection is less relevant here, as interest resides mainly in approximating the full chip set (empirical) 'truth'; that is, there is no larger set of chips for which it is desired to infer expression values.

Here, we have considered fixed numbers of resamplings. However, by adoption of a suitable convergence criterion, the methods can be readily modified to allow the procedure to stop automatically once 'enough' resamplings are selected. Examples of possible stopping criteria include correlation or variability between previous and current values of expression across the dataset.

The main user-supplied ingredients to PR are the within partition subset size and stopping rule (number of resampled partitions or convergence criterion). No comprehensive numerical or theoretical study has been made on this aspect. As a rough guide, a subset size of around 50 – 100 seems workable, depending on chip type, chosen expression measure and machine capabilities. Given the closeness of results for 50 and 100 resamples, 50 re-

sampled partitions may be sufficient in many instances to produce acceptable expression values; more may be desirable if the fully quantifiable subset size is small. Further study along with widespread adoption of the method should produce more insight into properties and tradeoffs so that these guidelines may be suitably refined.

Both resampling and within partition computing are inherently parallel operations, not dependent on other resamples or within partition subsets. Thus, PR is readily parallelizable, bringing gains in speed to multi-chip expression quantification in large studies.

Ideally, improvements in algorithms for exact computation would lessen the need for approximation strategies. However, an additional benefit of using a resampling strategy is that it can provide an estimate of expression measure standard error, not readily obtained otherwise. Such an estimate may prove useful in sensitivity and robustness studies.

PR is a readily applicable general tool that provides an immediate powerful and practical solution to the problem of multi-chip gene expression quantification for arbitrarily large sample sizes. EA requires more attention to the details of the quantification algorithm (*e.g.* RMA or gcRMA). The favorable properties of PR and EA recommend either as a method of choice when exact, full chip set methods are computationally infeasible. Software implementing Partition Resampling and Extrapolation Averaging is under development as an R package for the BioConductor project.

Bibliography

Affymetrix (2001) *Affymetrix: Microarray Suite User's Guide, version 5.0.* Santa Clara, CA.

Affymetrix (2005) *Technical Note: Guide to Probe Logarithmic Intensity Error (PLIER) Estimation.* Santa Clara, CA.

Bolstad, B. M., Irizarry, R. A., Astrand, M. and Speed, T. P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics* **19**, 185–193.

Collin, F. (2004) *Analysis of oligonucleotide data with a view to data quality assessment.* Ph.D. thesis, Department of Statistics, University of California, Berkeley.

Cope, L. M., Irizarry, R. A., Jaffee, H. A., Wu, Z. and Speed, T. P. (2004) A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics* **20**, 323–331.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H. and Zhang, J. (2004) BioConductor: Open software development for computational biology and bioinformatics. *Genome Biology* **5**, R80.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B. and Speed, T. P. (2003a) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research* **31**, e15.

Irizarry, R. A., Gautier, L., Bolstad, B. M. and Miller, C. (2006) *affy: Methods for Affymetrix Oligonucleotide Arrays.* R package version 1.10.0.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. and Speed, T. P. (2003b) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264.

Irizarry, R. A., Wu, Z. and Jaffee, H. A. (2005) Comparison of Affymetrix GeneChip expression measures. Technical report, Johns Hopkins University, Dept. of Biostatistics Working Papers.

Li, C. and Wong, W. H. (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences USA* **98**, 31–36.

Pollard, K. S., Dudoit, S. and van der Laan, M. J. (2005) Multiple testing procedures: the multtest package and applications to genomics. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, eds R. C. Gentleman, V. J. Carey, W. Huber, R. Irizarry and S. Dudoit, pp. 249–271. New York: Springer-Verlag.

Pollard, K. S., Ge, Y. and Dudoit, S. (2004) *multtest: Resampling-based multiple hypothesis testing*. R package version 1.5.2.

R Development Core Team (2006) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Wu, Z. and Irizarry, R. (2005a) *gcrma: Background Adjustment Using Sequence Information*. R package version 2.4.1.

Wu, Z. and Irizarry, R. A. (2005b) Stochastic models inspired by hybridization theory for short oligonucleotide arrays. *Journal of Computational Biology* **12**, 882–893.

Yeoh, E. J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R., Behm, F. G., Raimondi, S. C., Relling, M. V., Patel, A., Cheng, C., Campana, D., Wilkins, D., Zhou, X., Li, J., Liu, H., Pui, C. H., Evans, W. E., Naeve, C., Wong, L. and Downing, J. R. (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* **1**, 133–143.