Journal of Advanced Statistics, Vol. 2, No. 2, June 2017
https://dx.doi.org/10.22606/jas.2017.22003

Weighted Meta-Analysis for Small Sample Microarray Studies

Dallas Joder[1] and Nusrat Jahan[2*]

[1]Data Science Division, InferenSys Inc., Virginia, United States
[2]Department of Mathematics and Statistics, James Madison University, Virginia, United States
Email: jahannx@jmu.edu

Abstract An abundance of data from microarray studies are available in publicly-accessible databases. Most of these studies are conducted by university based research labs. It is not uncommon for such studies to run only three or four replicates for each experimental condition tested. With this low sample size and the high variability and multiple testing problems inherent to microarray technology, it is difficult to draw statistically significant conclusions from any one such study. Meta-analysis could improve this situation by combining evidence from related studies to increase statistical power. In this work we discussed several meta-analysis methods for small sample gene expression studies. We compared the performances of the traditional Fisher's log-sum and Stouffer's-Z meta-analysis methods, as well as three weighted variants of Stouffer's method. Higher false discovery rates were observed for the traditional methods compared to the weighted methods.

Keywords: Weighted meta-analysis, microarray, meta-p value, false discovery rate, integration driven discoveries, integration driven revisions, *Salmonella*.

1 Introduction

Microarrays are a high throughput method for quantifying gene expression to study the biological relationships between genes. Molecular biologists use this technology to identify the differences in gene expression induced by changes in experimental variables. Microarrays produce enormous amounts of data, but information content per gene is limited by the nature of the technology. The huge volume of information with limited sample size, produced by microarrays, poses a challenge for statistical analysis and interpretation.

Meta-analysis is a statistical approach that can be used to integrate results from independent but related microarray studies. This has the potential to increase both the statistical power and generalizability of single-study analyses. This approach is extremely useful, because microarray experiments are difficult and costly to replicate, and because results are often ambiguous due to the high potential for variability and the large number of genes examined [1]. Thus, it is often more practical to meta-analyze available results, rather than design a new experiment with sufficient power to tease up elusive gene functions.

Meta-analysis has its own challenges when applied to microarray data, because of differences in study protocols, platforms, normalization techniques, and analysis methods. These issues have been addressed in various microarray meta-analysis studies [2], [3], [4], [5], [6]. However, those papers focus primarily on microarray experiments in humans or animal models, where there is a relatively large amount of replication within and between studies. In contrast, microarray experiments in prokaryotic organisms are rarely independently replicated without substantial variations, such as using a different organism or introducing additional factors. Financial considerations also often limit the scope and data quality of these experiments. Small sample sizes, combined with abundant false positives from testing thousands of genes for differential expression, make it difficult to conclude that any gene is differentially expressed with statistical significance [7]. Therefore, small sample microarray, such as prokaryotic research could potentially benefit even more from meta-analysis than research in eukaryotes. However, the implementation is also much more challenging.

Our research explores methods for conducting meta-analysis on small sample microarray studies, using a combination of simulation and analysis of real data sets from the Gene Expression Omnibus (GEO), a data repository for microarray findings at the National Center for Biotechnology Information (NCBI). We focused on the human-disease-causing bacterium *Salmonella enterica* as a representative model for

meta-analysis of small sample prokaryotic microarray data. Microarray experiments with this organism are ample in the GEO database, but most studies have only two or three replicates. On top of this, very few of these studies investigate similar gene regulatory conditions. So scope of meta-analysis is limited. We focused our meta-analysis on the little understood *Salmonella* transcriptional modifier gene *stpA*. The *stpA* gene has been linked circumstantially to *Salmonella* virulence by virtue of its location in the *Samonella* pathogenicity island [8].

Two broad approaches to meta-analysis have been explored for microarray data—confidence based methods and effect size methods. Confidence methods attempt to combine confidence data, such as p-values, between studies [2]. Effect size methods, in contrast, attempt to find consensus values for measurements of treatment expression effects [3]. Confidence based meta-analysis has been criticized for not taking into account available information on the magnitude of the differences between conditions [5]. However, poor correlation at the expression-value level among different microarray platforms raises concerns about meta-analysis based on effect size methods [9], [10]. In contrast, Ghosh (2003) [10] found the t statistic to be less dependent on the platform technology. This is especially relevant for prokaryotic proteomics, where sample sizes may be too limited to accurately estimate the differential expression by using permutation tests [11], [12]. Sample size restrictions also exclude use of other differential analysis methods such as significance analysis [13], empirical Bayes methods [14], etc.

In this study, the traditional confidence methods, Fisher's log-sum and the unweighted Stouffer's-Z, are compared with three weighted versions of Stouffer's-Z method. The weight functions are inverse square root of variance (ISV), inverse variance (IV), and inverse coefficient of variation (ICV). These weight functions address data quality differences between studies by adjusting the amount of influence each study has on the result. The ISV and IV weightings penalize studies with high variability, IV more strongly so than ISV. The ICV weighting also penalizes studies for weak expression, as well as higher variability.

We performed meta-analysis on *Salmonella stpA* microarray data using these five confidence based methods. An especially stringent alpha level (0.001) was used to help reduce the false positive error rate in the p-value findings. The traditional approaches for controlling false discovery rate do not work well when sample size is small, a stringent alpha level provides better control of false discoveries [15], [16]. We used our findings and the characteristics of the *Salmonella stpA* datasets to construct a model of generalized microarray significance data. This allowed us to run meta-analyses on simulated small sample data to compare the sensitivity, specificity, and false discovery rate of our five meta-analysis methods.

2 Methods

Two classic confidence based meta-analysis methods are Fisher's log-sum and Stouffer's-Z. Fisher's method combines differential expression t test p-values across studies, to estimate a meta-p-value for each gene [2], [17]. For a meta-analysis of k independent studies, this technique calculates the summary statistic:

$$\chi_i^2 = -2\sum_{j=1}^{k}\ln(p_{i,j}), \tag{1}$$

where $p_{i,j}$ is the p-value of gene i from study j. In the above equation, χ_i^2 is the summary statistic corresponding to gene i. The meta-p-value for a summary statistic may be calculated from the χ^2 distribution with $2k$ degrees of freedom [17]. A meta-p-value falling below our stringent significance-level ($\alpha = 0.001$) indicates there is sufficient combined evidence from the studies to conclude that the treatment has a significant effect on that gene's expression.

Fisher's method is inherently one sided: the p-values it analyzes must come from a one-sided t test, with either an alternative hypothesis of up regulation or an alternative hypothesis of down regulation. The meta-p-values generated will correspond to the alternative hypothesis used. The full set of differentially expressed genes can be obtained by running separate meta-analyses for each of the two potential alternative hypotheses.

The second confidence method is known as Stouffer's Z-score method. It is similar to Fisher's method, except that it combines Z-scores associated with the p-values from different studies, instead of directly

combining p-values [18], [19], [20]. One-sided p-values can be easily converted to Z-scores as quantiles of the standard normal distribution. Stouffer's method combines the Z-scores to calculate a consensus meta-Z-score, Z_i, using the equation:

$$Z_i = \frac{\sum_{j=1}^{k} Z_{i,j}}{\sqrt{k}}, \tag{2}$$

where $Z_{i,j}$ is the Z-score of gene i from study j, and k is the number of studies being combined. The meta-Z-score, Z_i, produces a meta-p-value from the standard normal distribution, which can be used to interpret the significance of the regulatory effect.

Stouffer's method also easily accommodates the incorporation of weight functions to reward or penalize studies of different data quality, using the equation:

$$Z_i = \frac{\sum_{j=1}^{k} w_{i,j} Z_{i,j}}{\sqrt{\sum_{j=1}^{k} w_{i,j}^2}}, \tag{3}$$

where $Z_{i,j}$ is the Z-score of gene i from study j, and $w_{i,j}$ is a weighting factor that reflects the study quality. Note that this equation reduces to equation 2 when all study weights are equal.

Microarray meta-analysis is plagued by cross-platform and cross-laboratory variations, even when studies are performed under the same experimental conditions. Slight variations in DNA concentration, hybridization temperature, dye quality, contamination, and other factors can induce enormous systemic variation in gene expression values obtained from otherwise identical studies [21]. The amount of systemic and random variation in a study reflects the data quality of that study. In an ideal meta-analysis, data with low variability should have a bigger impact than data with higher variability. Weightings can be applied on a gene-by-gene basis to capture both gene-specific interstudy variation and microarray-wide variation [22].

Therefore, in addition to comparing Fisher's method and the traditional (unweighted) version of Stouffer's method, we included gene-specific ISV, IV, and ICV weighted implementations of Stouffer's method. The weight function formulas for these variants are:

$$w_{i,j} = \begin{cases} (s_{i,j})^{-1}, \text{ ISV;} \\ (s_{i,j}^2)^{-1}, \text{ IV;} \\ \dfrac{\bar{x}_{i,j}}{s_{i,j}}, \quad \text{ICV.} \end{cases} \tag{4}$$

Experimentally, $\bar{x}_{i,j}$ is the average strength of the expression signals, across both the mutant and wildtype conditions, for gene i in study j. The $s_{i,j}$ estimates the noise within that signal. Thus, the ISV and IV weightings penalize studies with larger amounts of noise, IV more strongly than ISV, and the ICV weighting penalizes studies with a poor signal-to-noise ratio.

We coded a library of functions in R to perform these types of meta-analysis. We then used these functions to analyze three real microarray datasets involving *stpA*. We also conducted simulations to more directly compare performance of the analysis methods and weighting functions.

2.1 Meta-Analysis of *stpA* Microarrays

We used the following inclusion criteria to select NCBI GEO data for meta-analysis. First, inclusion was limited to gene expression microarray studies for which mutation of the *stpA* gene was among the independent variables. Experiments with organisms other than *S. enterica Typhimurium* were excluded. Finally, data sets with less than three arrays were also excluded from the meta-analysis.

We identified two composite studies conducted by Lucchini et al. [23], that met all of these criteria. One of these studies, GSE18424, compared transcription rates of wildtype *S. enterica Typhimurium* and a derivative *stpA* knockout strain, at the early log, mid log, late log, and stationary growth stages in the lifecycle of a culture. This study was comprised of four experiments. The other study, GSE18428, measured the expression effects of *stpA* deletion in *S. enterica Typhimurium* and its interaction with deletion of a

known complementary regulator of related genes, called *RpoS*. This study had two experiments. Findings from both studies are published in [23]. They demonstrated that the *stpA* protein is not produced in substantial quantities in the wildtype during the early log or stationary stages [23]. They also concluded that the *RpoS* knockout counteracted some of the gene expression effects of *stpA* deletion [23]. Lucchini et al. findings suggest that the early log, stationary stage, and the double knockout with *RpoS* experiments are not biologically relevant to our meta-analysis focus. For these reasons, we included only the remaining three experiments in our meta-analysis. These experiments are described in Table 1.

Table 1. Experiments included in meta-analysis

Experiment	Name	Wildtype Samples	Mutant Samples
1	GSE18424 (mid log)	4	4
2	GSE18424 (late log)	4	4
3	GSE18428	3	3

All three datasets contained the log base 2 of sample intensity divided by genomic reference intensity. These were quantile normalized [24]. We identified 4,355 genes common to every microarray used in the three experiments, and restricted our analysis to these genes. Unequal-variance t test was used for detecting differences between the *stpA* knockout and wildtype conditions within each study. Because of small sample size and unequal variances between the conditions we refrained from using permutation based tests [11]. At an alpha level of 0.001, the p-values identified 92 genes from experiment 1 as significantly differentially expressed. Experiment 2 identified 139 genes, and experiment 3 identified 45 genes. Out of these 276 significant discoveries, only 30 were shared discoveries and remaining 246 were unique discoveries. The resulting p-values from the three experiments were used as the inputs for the meta-analysis functions. Figure 1 (A) shows a boxplot of the p-values by experiment.

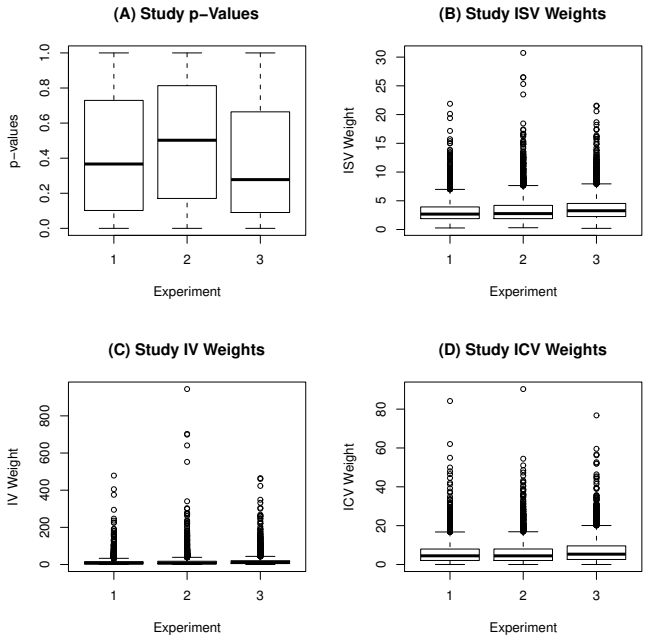

Figure 1. Box plots for gene p-values, and 3 different weight functions for Stouffer's method for each experiment.

The ISV, IV, and ICV weight functions for the weighted Stouffer's method, are designed to address the systematic variations occurring among studies. Within a study, both the wildtype and stpA knockout arrays will get equally affected by the study specific attributes. The weight functions are computed using the intensity log-ratios for each gene, from both wildtype and mutant data. The weight values for each gene were then constructed according to Equation 4. Boxplots of the ISV, IV, and ICV weights for each experiment are shown in Figure 1 (B) , (C), and (D). From Figure 1 (A), Experiment 2 appears to have an approximately symmetric p-value distribution, an indication of less systematic variation. Experiments 1 and 3 have right skewed p-value distributions, implying the presence of systematic variation. This finding is supported by the weight function box plots. The differing vertical scales of these plots indicate that the IV weight penalties are the most severe, and the ISV weights are the most liberal. The ISV (B) and IV (C) plots indicate that, both weight functions assign greater weight to experiment 2 among the 3 experiments. Experiment 1 and 3 are penalized for poor data quality by having smaller weights. The ICV (D) plot also has the same penalty pattern, but in lesser degree.

Our meta-analyses was structured as described by Equations 1 and 3. A gene was deemed to be differentially expressed if it's meta-p-value was less than or equal to our standard alpha level of 0.001. The significant genes from meta-analysis were compared to the 276 significant genes identified from the three individual studies in section 3.1.

2.2 Simulations

To estimate the sensitivity and specificity of the proposed meta-analysis methods, we conducted a simulation study with randomized significance data for a hypothetical genome of 10,000 genes. Each gene was characterized by a true differential expression status, p_0, similar to a lower tail p-value. The p_0-values were randomly sampled from uniform distribution, $U[0,1]$. A variable fraction of the genome, π, was made truly down-regulated, and the remaining fraction, $1-\pi$, was made non-differentially expressed. The p_0-values for the original $\pi \times 10{,}000$ genes were randomly sampled from $U[0,1]$ in such a way that these p_0-values have bounds $[0,\ 0.001]$, with 0.001 being our critical significance level. This conferred a condition of true down regulation for those genes. The p_0 for the remaining $(1-\pi) \times 10{,}000$ genes were randomly selected with bounds $(0.001,\ 1]$, to represent true non-differential expression. The combined set of p_0-values was considered as the "true" set of significance data for the hypothetical genome. Meta-analysis sensitivity, specificity, and false discovery rates were estimated in the context of this "true" set. The p_0-values were perturbed with random noise to produce three different studies' worth of unique p-value data for the meta-analysis.

To produce distinct p-values for each gene, we computed $\zeta_i = \phi^{-1}(p_{0,i})$, where ϕ is the standard normal cumulative distribution function. The simulated p-value for the ith gene in jth study, here denoted by $p^*_{i,j}$, was obtained by adding simulated noise to the ζ_i,

$$p^*_{i,j} = \phi(\zeta_i + \epsilon_{i,j}), \qquad i = 1,2,...,10{,}000 \ \text{ and } \ j = 1,2,3 \tag{5}$$

where $\epsilon_{i,j}$ acts as random error specific to each study. The $\epsilon_{i,j}$, was generated from a normal distribution with mean zero and standard deviation σ_s. The σ_s represents the magnitude of interstudy variation, and should be highly specific to the set of studies included in the meta-analysis. In our simulation we attempted to replicate the interstudy variability conditions of the Lucchini data. The σ_s was estimated by the average gene standard deviation of study Z-scores, corresponding to the confidence data from Lucchini's datasets. This convoluted method ensured that the three simulated p-values for each gene were similar to the corresponding true significance value p_0 but also allowed for interstudy variation, as expected in an actual meta-analysis [25]. Gene specific dependence between studies, investigating similar biological conditions, is expected. So if a specific gene is significant in one study, it is also expected to have similar significance in the other studies.

For Stouffer's weighted methods, we needed to generate weights within each run of each simulation. For this, we wanted to replicate the study variability conditions of the Lucchini experiments. We used a method to generate randomized weights based on the empirical distributions of $\bar{x}$ and s from all three experiments. The empirical distribution of $\bar{x}$ was divided into l equal data subsets, $v_1 < v_2 < \cdots < v_l$ such that, $v_1 < min\ x_i \leq max\ x_i < v_l$. The v_is represented the boundaries between the consecutive data subsets. In the first step of simulating $\bar{x}$, a data subset was randomly selected from all l equally

likely subsets. Once subset i was selected, $\bar{x}_i^*$ was randomly selected from a uniform distribution with bounds $(v_i,\ v_{i+1}]$. The fraction of $\bar{x}^*$ selected from the subsets of tail areas is equivalent to the proportion of tail areas of the empirical distribution of the Lucchini experiments. The same method was repeated for simulating s^* using the empirical distribution of s from the Lucchini experiments. These $\bar{x}^*$ and s^* were used to calculate ISV, IV, and ICV weights according to Equation 4.

To confirm that this weight generation process produces weights with similar behaviour to the Lucchini experiments, we performed a test run of the simulations, with 4,355 genes, for direct comparison with the real data. The plots of ISV and ICV for all 4,355 common genes from the Lucchini experiments against their study specific Z-scores are shown in Figure 2 (A) and (B). (Note that IV is simply the square of ISV and so is not shown.) The simulation ISV and ICV weights are plotted against Z-score in Figure 2 (C) and (D), respectively. The simulated weights satisfactorily replicated the characteristics of the Lucchini dataset weights as discussed above. Each simulated dataset was meta-analyzed by the five methods. We performed

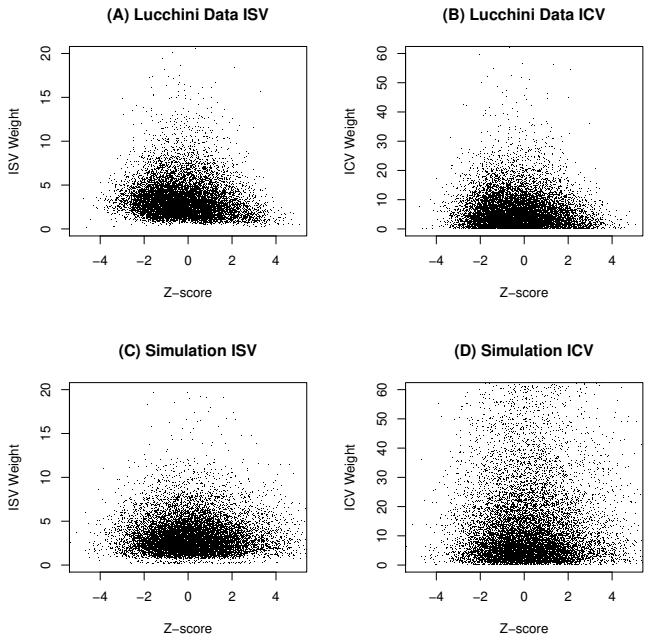

Figure 2. Plots of simulation and Lucchini et al. data IV and ICV weights for comparison. The simulation Z-scores (ζ) in these plots used $\pi = 0.05$, as an approximation for the findings of [23].

1,000 simulation runs. In each run, three studies were generated by adding random noise to a new set of 10,000 "true" significance data (p_0-values). The meta-analysis results were obtained by averaging the sensitivity, specificity, and false discovery rate over all runs. Finally this whole process was repeated for seven different proportions of true significant genes: $\pi =$ 0.005, 0.01, 0.02, 0.05, 0.1, 0.15, and 0.2.

3 Results and Discussion

3.1 Data Analysis

The three experiments obtained from [23], had 4,355 genes in common. Five methods of meta-analysis were performed on these genes. For all methods, gene significance was determined based on a stringent alpha level of 0.001. The number of genes found to be significantly up or down regulated by each method, and the number of integration driven (ID) discoveries and revisions, are exhibited in Table 2. Integration driven discoveries are genes found significant by the meta-analysis, but not by any individual study. They

are indicative of increased sensitivity attained by combining study results. Integration driven revisions are genes found significant by one or more studies independently, but not by the meta-analysis, indicating the meta-analysis provided increased specificity [5]. Figures 3 and 4 show the meta-p-value distribution of the significant genes.

Table 2. Meta-analyses results of the Lucchini *stpA* studies

No. of Genes	Fisher's	Stouffer's	ISV	IV	ICV
Up Regulated	180 (4.13 %)	186 (4.27 %)	170 (3.90 %)	146 (3.35 %)	147 (3.38 %)
Down Regulated	126 (2.89 %)	131 (3.01 %)	112 (2.57 %)	101 (2.32 %)	102 (2.34 %)
Total	306	317	282	247	249
ID Discoveries	169	179	159	131	131
ID Revisions	104	108	123	130	128

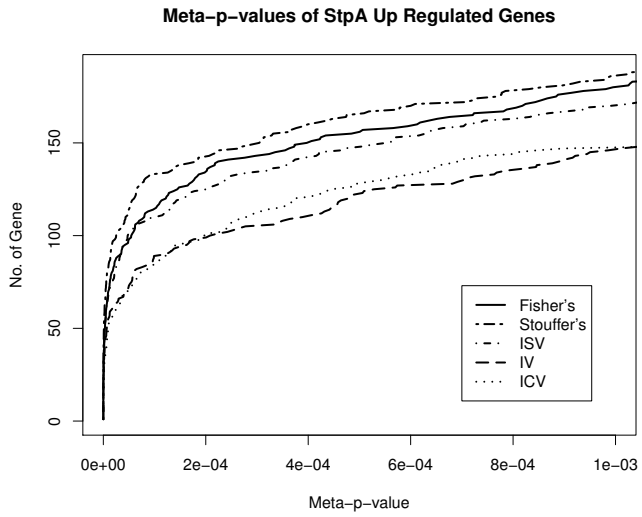

Figure 3. Meta-p-values of significantly up regulated genes from the Lucchini dataset meta-analyses.

There is a high concentration of extremely significant meta-p-values for all methods, suggesting a high proportion of these genes may be true discoveries. (Compare to Figure 5 from the simulations.) Down regulated meta-p-values are scattered fairly uniformly between 0 and 0.001, in contrast to Figure 3. This suggests these genes may contain a large proportion of false discoveries. Table 2 and Figures 3 and 4 show that all methods found more genes significantly up regulated than down regulated. These findings agree with the conclusion of [23], that *stpA* serves primarily as a repressor gene, causing genes to be over-expressed when it is knocked out. Stouffer's unweighted method, followed by Fisher's method identified the most genes, while the IV and ICV methods found the least. Figures 3 and 4 confirm that the weighted methods had smaller meta-p-values than their traditional counterparts. The weighted methods also had higher rates of integration driven revision and lower integration driven discovery. This is likely because the weighting adds an additional component of quality control, producing more stringent criteria for significance. The more aggressive IV and ICV methods had a greater tendency toward revision than the more lenient ISV method.

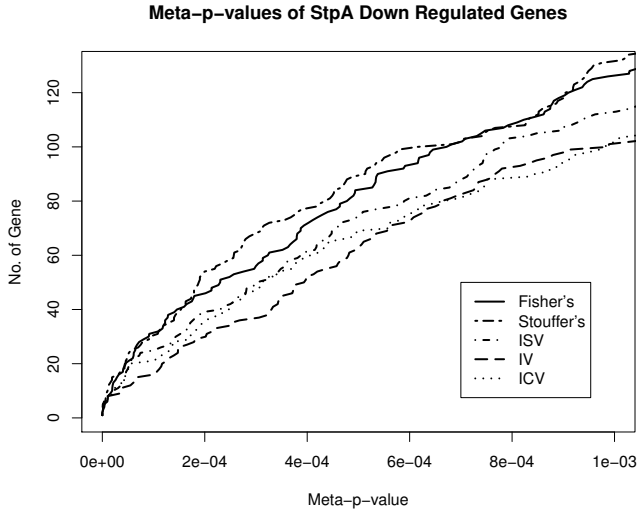

Figure 4. Meta-p-values of significantly down regulated genes from the Lucchini dataset meta-analyses.

The figures and the table show that the ICV and IV methods may be more conservative than the other methods. If this conservative tendency is associated with reduced false discovery, it could be advantageous for proteomics research. Our simulation results provide a more thorough investigation of this possibility (see Section 3.2).

To confirm that our results were biologically plausible, we performed a cursory literature search on the top 100 most significant genes, according to each method—137 unique genes in total. Of these genes, we found that 81 had already been singled out in [23], based on having greater than two-fold differential expression. The *stpA* gene itself was identified by every method at rank 7 or higher. Of the remaining 136 genes examined, we found 19 that were previously identified in the literature with significant roles in *Salmonella* virulence. We also found 16 genes that were members of six known metabolic operons. Additionally, there were 47 genes in sixteen groups for which their clustering of consecutive locus codes suggested they might be operons as well. We consider these genes, identified in our literature search, as plausible "true discoveries." Among these biologically plausible discoveries, we found 6, or 8.2%, that were not identified by any individual experiment from t test results or from having greater than two-fold differential expression. They were only detected by the meta-analysis.

These findings demonstrate that the techniques of confidence based meta-analysis can be applied to publicly available microarray data to improve the confidence of biological conclusions. All methods produced a large number of integration driven discoveries and revisions, improving upon the individual study results. A preliminary literature review showed that the set of genes identified was biologically consistent with current theories on *stpA*, and that a substantial fraction of these were integration driven discoveries rather than being identified by any individual study.

3.2 Simulation Analysis

For the simulation study, an "original" set of significance data (p_0-values) for a hypothetical 10,000 genes was generated with a specified proportion of the genes (π) exhibiting true differential expression. Three experiments, each containing study p-values for 10,000 genes were simulated by adding random error to the p_0-values of the original set. The random error was a Gaussian zero-mean process with standard deviation $\sigma_s = 1.0090$, which was the interstudy variability estimated from the Lucchini et al. datasets.

The weight functions ISV, IV, and ICV were also simulated using the characteristics of the Lucchini datasets (see Figure 2) by uniform random sampling from the $l = 100$ subsets of the empirical distributions of $\bar{x}$ and s. The entire simulation process was repeated for 7 different proportions of true significance (π = 0.005, 0.01, 0.02, 0.05, 0.1, 0.15, and 0.2), and each simulation consisted of 1,000 runs. The false

discovery rate (FDR), sensitivity, and specificity were computed by comparison of meta-p-values for each method to the original p_0-values, with respect to the alpha level.

For a unique insight into this simulation process, confidence distribution plots, similar to Figures 3 and 4, are presented in Figures 5 and 6. These show meta-p-values for a single run of the simulation at $\pi = 0.05$. Figure 5 shows the 500 simulated genes that were assigned to be differentially expressed, and Figure 6 shows the 600 most significant non-differential genes (based on p_0-values). False discoveries and missed discoveries can be observed directly in these plots. The plots also show the "true" p_0-values for the genes. From Figure 5, the IV and ICV methods identified 465 and 466 genes as significant, respectively. The ISV

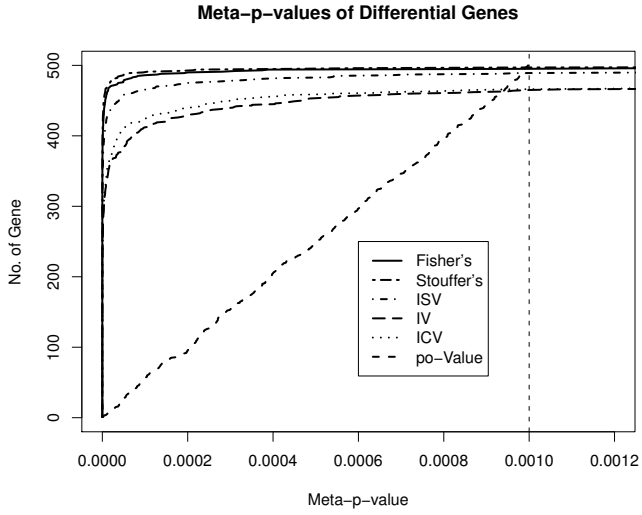

Figure 5. Meta-p-values and p_0-values of 500 differentially expressed genes ($\pi = 0.05$, $\alpha = 0.001$).

method found 489 genes and Fisher's, and Stouffer's methods found 495 and 497 genes, respectively. The p_0-values of these genes are also plotted. All methods cause tremendous inflation of the genes' significance.

From Figure 6, IV and ICV methods made 411 and 404 false discoveries, respectively, whereas ISV made 504. Fisher's method 546 and Stouffer's method made 590 false discoveries. Weighted methods, especially ICV and IV, produce many fewer false discoveries than the other methods.

In Figure 5, and especially Figure 6, the meta-analysis methods are shown to have greatly exaggerated the significance of the genes over their "true" p_0 significance. For the differential genes this contributed to a high rate of successful detection (Figure 5), but for the non-differential genes it produced a very high proportion of false discoveries (Figure 6). In the weighted methods the meta-p-value inflation was lower compared to the traditional methods. The IV and ICV in particular trended closer to their true p_0-values, suggesting they promote higher accuracy by virtue of being more conservative. These conclusions are reinforced by the aggregate sensitivity, specificity, and FDR of the full simulation.

Averaged sensitivity and specificity results across all 1,000 simulation runs are presented for $\pi = 0.005$, 0.02, 0.05, and 0.2 in Table 3. Sensitivity and specificity for each method remained nearly constant across all proportions of π. This is reasonable, as sensitivity and specificity estimate the correct classification rates of true positives and true negatives, respectively. Neither our model for simulating study confidence data, nor the meta-analysis methods themselves, account for any interactions between genes. Thus, the only distinction between simulations with different values of π is the number of true differentially or non-differentially expressed genes available for estimating sensitivity and specificity, respectively.

On average, Fisher's method and the unweighted Stouffer's method exhibited sensitivity of 99.2% and 99.5%, respectively. The weighted methods had slightly less sensitivity, with ISV at 97.5%, ICV at 92.9%, and IV, the least sensitive, at 92.0 percent. For specificity, the trend was reversed. Stouffer's method and

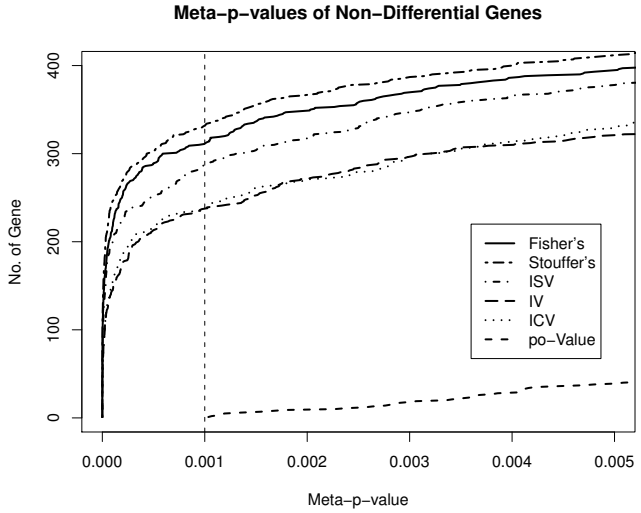

Figure 6. Meta-p-values and p_0-values of the 600 most significant non-differential genes for $\pi = 0.05$ and $\alpha = 0.001$.

Table 3. Meta-analysis sensitivity and specificity from simulation

	Fisher's	Stouffer's	ISV	IV	ICV
		Sensitivity			
$\pi = 0.005$	99.26 %	99.51 %	97.49 %	91.97 %	93.16 %
$\pi = 0.02$	99.18 %	99.46 %	97.51 %	91.85 %	92.82 %
$\pi = 0.05$	99.17 %	99.47 %	97.50 %	91.94 %	92.93 %
$\pi = 0.2$	99.17 %	99.46 %	97.52 %	91.96 %	92.88 %
		Specificity			
$\pi = 0.005$	94.51 %	93.94 %	94.99 %	96.11 %	96.01 %
$\pi = 0.02$	94.52 %	93.94 %	94.99 %	96.12 %	96.01 %
$\pi = 0.05$	94.51 %	93.94 %	94.99 %	96.11 %	96.01 %
$\pi = 0.2$	94.50 %	93.93 %	94.98 %	96.11 %	96.02 %

Fisher's method had the lowest specificity at 94.5% and 93.9%, respectively. ISV had 95.0% specificity. IV and ICV both performed well, exhibiting 96% average specificity.

Microarray analysis is an exploratory method. Any discoveries must be verified by more robust experimental methods such as western-blot. While it is desirable to have a large number of genes identified as potential targets, if that number is too large it will inflate the false discovery rate and waste resources in verification. This is especially a problem for prokaryote research, where funding and materials may already be limited. With greater than 99% sensitivity in our simulations, we consider the traditional Fisher's and Stouffer's meta-analysis methods too liberal for microarray analysis. Trading some sensitivity for specificity would be preferable.

The weighted methods we propose do exactly that by taking account of interstudy variation, which produces more conservative conclusions. The inverse square root of variance (ISV) is the least aggressive weighting and has the least pronounced effect. It offered only a modest gain of 1.1% specificity over the unweighted Stouffer's-Z in exchange for a 1.9% reduction in sensitivity. Inverse variance (IV) offered a larger specificity gain of 2.2%, but at the hefty cost of exhibiting 7.5% less sensitivity. The ICV method, however, offered nearly the same gain in specificity (2.1%) for a smaller loss in sensitivity (6.6%).

These tradeoffs are advantageous, because the number of differentially expressed genes in real microarray experiments tends to be small. Thus, small improvements in specificity can eliminate a large number of false discoveries, while a comparatively large loss of sensitivity would sacrifice only a few true discoveries.

For instance, in our 10,000 gene simulation, at $\pi = 0.05$, a 2% increase in specificity would eliminate 190 false discoveries, while a 7% decrease in sensitivity would only cause the loss of 35 true discoveries.

An ideal method in this context is one that minimizes the false discovery rate. This reduces waste of resources from pursuing biological verification for false discoveries. Averaged false discovery rates are presented for selected values of π in Table 4. Figure 7 plots the averaged FDR values against π for each of the five meta-analysis methods. The FDR is strongly dependent upon the value of π, and is largest when π is close to zero. Differences in the FDR among the five methods are consistent for all levels of π, but are more pronounced for values of π above 0.02.

Table 4. Meta-analysis method FDR from simulation

	Fisher's	Stouffer's	ISV	IV	ICV
		FDR			
$\pi = 0.005$	91.67 %	92.38 %	91.09 %	89.37 %	89.49 %
$\pi = 0.02$	73.04 %	74.90 %	71.56 %	67.44 %	67.79 %
$\pi = 0.05$	51.27 %	53.64 %	49.39 %	44.54 %	44.90 %
$\pi = 0.2$	18.14 %	19.62 %	17.07 %	14.48 %	14.63 %

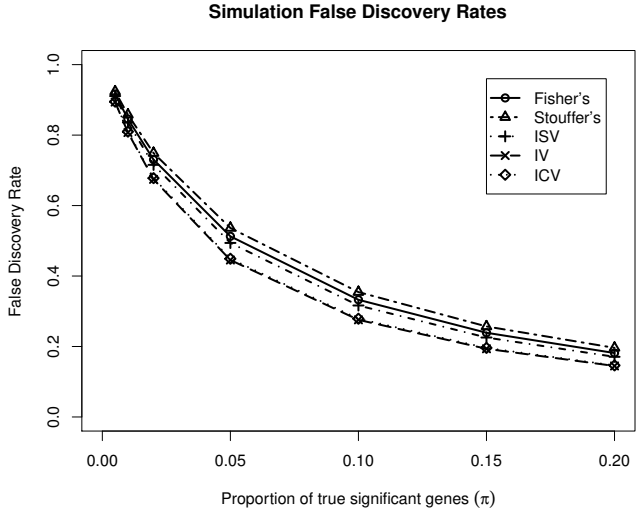

Figure 7. Average false discovery rate (at $\alpha = 0.001$) over 1,000 runs, is plotted against different values of π.

The weighted methods perform considerably better than the Fisher's method and the un-weighted version of Stouffer's method, as the value of π increases (Figure 7) . In particular, the IV and ICV methods outperform all other methods by a large margin. The improvement is approximately 5 to 8% on average, compared to the traditional methods, except when π is very small. For all values of π, the IV and ICV methods provided the lowest false discovery rates. Results for the ISV method were intermediate between those for IV and ICV and those for the traditional methods. Stouffer's method had the highest FDR of all the methods, and Fisher's method had the second highest. These results favor use of the ICV or IV methods when resources for biological verification are limited.

4 Conclusion

Microarray research has the potential to benefit tremendously from the use of meta-analysis, due to the preponderance of accessible studies with low individual sample size. The success of meta-analysis depends on the quality of the study data to be combined, so it is very important to pro-actively account for interstudy variability. We discussed three weighted confidence-based meta-analysis methods to address interstudy variation. The inverse coefficient of variation and inverse variance weight functions appear most promising. The IV method penalizes high noise, while the ICV method penalizes low signal-to-noise ratios. Both increase specificity at the cost of decreased sensitivity, but the ICV method produced a smaller sacrifice in sensitivity than the IV method. Both methods produce far fewer false discoveries than the traditional methods, and thereby reduce costs for subsequent biological verification of preliminary microarray findings.

References

1. R. Sasik, C. Woelk, and J. Corbiel, "Microarray truths and consequences." *Journal of Molecular Endocrinolgy*, vol. 33, pp. 1–9, 2004.
2. D. Rhodes, T. Barrette, M. Rubin, D. Ghosh, and A. Chinnaiyan, "Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer." *Cancer Research*, vol. 62, pp. 4427–4433, 2002.
3. J. Choi, U. Yu, S. Kim, and O. Yoo, "Combining multiple microarray studies and modeling interstudy variation." *Bioinformatics*, vol. 19, pp. 84–90, 2003.
4. i. G. Parmigian, E. Garrett-Mayer, R. Anbazhagan, and E. Gabrielson, "A cross-study comparison of gene expression studies for the molecular classification of lung cancer." *Clinical Cancer Research*, vol. 10, pp. 2922–2927, 2004.
5. J. Stevens and R. Doerge, "Combining multiple microarray studies and modeling interstudy variation." *BMC Bioinformatics*, vol. 6, pp. 1–19, 2006.
6. P. Hu, C. Greenwood, and J. Beyene, "Combining multiple microarray studies and modeling interstudy variation." *Information Systems Frontiers*, vol. 8, pp. 9–20, 2006.
7. J. Storey and R. Tibshirani, "Combining multiple microarray studies and modeling interstudy variation." *Proceedings of the National Academy of Sciences*, vol. 100, pp. 9440–9445, 2003.
8. N. Arricau, D. Hermant, H. Waxin, and M. Popoff, "Molecular characterization of the salmonella typhi stpa protein that is related to both yersinia yope cytotoxin and yoph tyrosine phosphatase." *Research in Microbiology*, vol. 148, pp. 21–26, 1997.
9. W. Kuo, T. Jenssen, A. Butte, L. Ohno-Machado, and I. Kohane, "Analysis of matched mrna measurements from two different microarray technologies." *Bioinformatics*, vol. 18, pp. 405–412, 2002.
10. D. Ghosh, T. Barrette, D. Rhodes, and A. Chinnaiyan, "Analysis of matched mrna measurements from two different microarray technologies." *Functional & Integrative Genomics*, vol. 3, pp. 180–188, 2003.
11. Y. Huang, H. Xu, V. Calian, and J. Hsu, "To permute or not to permute." *Bioinformatics*, vol. 22, pp. 2244–2248, 2006.
12. T. Roy, "The effect of heteroscedasticity and outliers on the permutation t-test." *ournal of Statistical Computation and Simulation*, vol. 72, pp. 23–26, 2002.
13. V. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response." *Proceedings of the National Academy of Sciences*, vol. 98, pp. 5116–5121, 2001.
14. G. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." *Statistical Applications in Genetics and Molecular Biology*, vol. 3, pp. 1–25, 2004.
15. R. Wetzels, D. Matzke, M. Lee, J. Rouder, J. Iverson, and E. Wagenmakerset, "Statistical evidence in experimental psychology: An empirical comparison using 855 t tests." *Perspectives on Psychological Science*, vol. 6, pp. 291–298, 2011.
16. V. Johnson, "Revised standards for statistical evidence." *Proceedings of the National Academy of Sciences*, vol. 110, pp. 19 313–19 317, 2013.
17. F. Mosteller and R. Fisher, "Questions and answers." *The American Statistician*, vol. 2, pp. 30–31, 1948.
18. S. Stouffer, "Questions and answers." *The American Soldier*, vol. 1, 1949.
19. R. Rosenthal, "Meta-analysis: a review." *Psychosomatic Medicine*, vol. 53, pp. 247–271, 1991.
20. L. Hedges, H. Cooper, and B. Bushman, "Testing the null hypothesis in meta-analysis: a comparison of combined probability and confidence interval procedures." *Psychological Bulletin*, vol. 111, pp. 188–194, 1992.
21. P. Cahan, "Meta-analysis of microarray results: Challenges, opportunities, and recommendations for standardization." *Gene*, vol. 401, pp. 12–18, 2007.

22. G. Alves and Y. Yu, "Accuracy evaluation of the unified p-value from combining correlated p-values." *PLoS One*, vol. 3, 2014.
23. S. Lucchini, "The h-ns-like protein stpa represses the rpos (sigma 38) regulon during exponential growth of salmonella typhimurium." *Molecular Microbiology*, vol. 74, pp. 1169–1186, 2009.
24. B. Bolstad, R. Irizarry, M. Astrand, and T. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." *Bioinformatics*, vol. 19, pp. 185–193, 2003.
25. S. Pyne, B. Futcher, and S. Skiena, "Meta-analysis based on control of false discovery rate: combining yeast chip-chip datasets." *Bioinformatics*, vol. 22, pp. 2516–2522, 2006.

