MIA: Multi-cohort Integrated Analysis for Biomarker Identification

Conference Paper · August 2018

DOI: 10.1145/3233547.3233605

4 authors, including:

Brian Marks
University of Nevada, Reno
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Hung Nguyen
University of Nevada, Reno
3 PUBLICATIONS **1** CITATION
SEE PROFILE

Tin Nguyen
University of Nevada, Reno
25 PUBLICATIONS **60** CITATIONS
SEE PROFILE

MIA: A Multi-cohort Integrated Analysis for biomarker identification

Brian Marks[1], **Nina Hees**[1], **Hung Nguyen**[1], **and Tin Nguyen**[1,*]

[1]Department of Computer Science and Engineering, University of Nevada, Reno, Nevada, USA
[*]tinn@unr.edu

Abstract: Advanced high-throughput technologies have produced vast amounts of biological data. Data integration is the key to obtain the power needed to pinpoint the biological mechanisms and biomarkers of the underlying disease. Two critical drawbacks of computational approaches for data integration is that they do not account for study bias, as well as the noisy nature of molecular data. This leads to unreliable and inconsistent results, i.e., the results change drastically when the input is slightly perturbed or when additional datasets are added to the analysis. Here we propose a multi-cohort integrated approach, named MIA, for biomarker identification that is robust to noise and study bias. We deploy a leave-one-out strategy to avoid the disproportionate influence of a single cohort. We also utilize techniques from both p-value-based and effect-size-based meta-analyses to ensure that the identified genes are significantly impacted. We compare MIA versus classical approaches (Fisher's, Stouffer's, maxP, minP, and the additive method) using 7 microarray and 4 RNASeq datasets. For each approach, we construct a disease signature using 3 datasets and then classify patients from 8 remaining datasets. MIA outperforms all existing approaches in terms of both the highest sensitivity and specificity by accurately distinguishing symptomatic patients from healthy controls.

1 Introduction

High-throughput technologies have generated vast amounts of genomic data with unprecedented rates. Large public repositories such as the Gene Expression Omnibus,[1] The Cancer Genome Atlas (`cancergenome.nih.gov`), and ArrayExpress,[2] store thousands of datasets, within which there are independent experimental series with similar patient cohorts and experimental design. Gene expression data, both microarray and RNA-Seq, are particularly prevalent in public databases, such that some disease conditions are represented by half a dozen studies or more.

However, real progress in understanding disease phenomena still lags far behind the gathering of data. Studies often fail to identify the true cause of phenomena, due to noise, study bias, or the subtlety of changes in biological signals between disease and healthy samples. Batch effects, patient heterogeneity, and disease complexity all complicate the integration of data from different sources.[3] Indeed, for the same disease, different studies produce different sets of differentially expressed (DE) genes.[4,5] It would be tremendously beneficial if all datasets associated with a given condition could be analyzed together, in order to overcome study bias and to increase sample size.

Meta-analysis of gene expression data has primarily been used for DE gene detection.[6] Rhodes et al.[7] were among the earliest to apply sophisticated meta-analysis methods for DE gene detection. In their work, p-values from multiple prostate cancer datasets were combined using Fisher's method.[8] Since then, other p-value based meta-analysis methods have been applied, such as Stouffer's method,[9] minP,[10] maxP,[11] weighted Fisher's method,[12] and latent variable approaches.[13] This p-value based sort of integration is one means by which meta-analyses are commonly performed. A recent literature review[6] revealed that p-value based meta-analysis for gene detection accounts for approximately twice as many studies as any other type of meta-analysis, and is favored for its simplicity and extensibility. One critical drawback of these p-value-based approaches is that they neglect the actual changes in gene expression, i.e. effect sizes. This results in a critical loss of information. While p-values are in part a function of effect size, it is also partly a function of sample size.[14] For example, with large sample size, a statistical test will almost always demonstrate a significant difference, unless the effect size is exactly zero, which is very unlikely in reality. Simply combining individual p-values would not be enough to correct such a problem. In addition, most methods for combining p-values are sensitive to outliers.

Here we propose a new approach that utilizes techniques from both classical p-value-based and modern effect-size-based meta-analyses to reliably identify genes that are significantly impacted from both perspectives: classical hypothesis testing and standardized mean difference. We also apply a robust leave-one-out technique to avoid disproportionate influence from a single cohort. We demonstrate the performance of the proposed approach using 640 Alzheimer's samples from 7 Affymetrix and 3 RNASeq datasets. We compare our new approach to 5 other approaches: Fisher's,[8] Stouffer's,[9] minP,[10] maxP,[11] and addCLT.[15–17] The framework outperforms existing approaches in identifying disease signatures which distinguish symptomatic individuals from healthy individuals with significant p-values.

We used 3 Affymetrix datasets to serve as our training sets and then tested the validity our gene set using the remaining datasets. Each of the meta-analysis techniques were given the same training set and testing set. For each of the statistical modeling methods, the testing set classification results were graphically represented on a receiver operating characteristic (ROC) curve. The area under the curve (AUC) of each

method was used to determine how effective the gene signatures were at classifying the different samples. The purpose of our proposed method is to determine a curated set of genes for a given dataset which can be used to classify a patient as prone to AD, or not. This can then be used as a means of making a preemptive diagnosis, which helps to improve the effectiveness of the management of AD symptoms.[18]

2 Methods

The pipeline of the proposed framework is shown in Figure 1. The input consists of n independent studies for the same disease. Each study can be represented as a matrix where columns are samples/patients and rows are genes/components. In each study, the samples are divided into two groups – disease and control. The goal of the framework is to identify a robust and consistent set of genes/components that can be used as biomarkers for future diagnostics. The output of the framework is a set of genes which has the potential to distinguish symptomatic individuals from healthy ones based on expression data alone.

The framework is divided into four distinct steps: (A) hypothesis testing and effect size estimation for each gene in each study, (B) meta-analysis in which the p-values and the effect sizes are combined from m studies, (C) feature selection based on the computed statistics, and (D) leave-one-out procedure to test for the stability of the obtained biomarkers.

In the first step (step A), we work with each dataset independently. For each gene, we compute the standardized mean difference (STD) between disease and control samples. Similar to fold-change, STD represents the difference in expression of the gene between two groups of samples. The difference is that STD is less sensitive to the scaling of each platform. We also perform a classical hypothesis test, using empirical Bayesian test,[19] to determine if the difference between two groups is observed by chance. After step (A), for each dataset, we have a list of STDs and p-values. For each gene, we have n p-values and n STDs – one p-value and one STD per dataset.

In step (B), we combine the p-values and STDs of each gene. Since these n p-values and n STDs are obtained from independent datasets, they can be combined using classical approaches. Here we use the additive approach,[15] which is based on the Central Limit Theorem,[20] to combine each gene's p-values into one single combined p-value for each individual gene. At the same time, we also combine the n independent STDs using the REstricted Maximum Likelihood (REML) algorithm.[21–23] When the REML stops, it outputs the central tendency of effect sizes for a given gene, as well as its standard error. We also compute a p-value from the estimated effect size and standard error. This p-value represents how reliable the effect size difference is.

In step (C), we select the genes whose summary statistics satisfy certain conditions. We first adjust the combined Bayesian p-values and the effect-size p-values using False Discovery Rate.[24] We then choose genes whose adjusted p-values are each less than 1%. After step (C), we have a list of genes that are significantly impacted by the underlying disease. To make the framework more robust, we also perform a leave-one-out procedure. In step (D), we perform n additional analyses. In each analysis, we remove one of the n studies and repeat the whole pipeline. Each of the n analyses outputs a set of impacted genes. We intersect the $n+1$ sets of genes to obtain a conservative set of genes which can be used as a biomarker for the disease. We will describe the details of each step in the following sections.

2.1 Hypothesis testing and meta-analysis

Here we use the empirical Bayesian test, provided by *limma* package,[19] to calculate the two-tailed p-values. We then convert these p-values into one-sided p-values. Depending on the hypothesis, researchers can choose to work with either the left or right-sided values. Assuming that we are interested in down-regulated genes, we would focus only on the left-sided p-values.

2.1.1 Combining p-values

Fisher's method is the most widely used method to combine independent p-values. Consider m individual null hypotheses H_{0i} ($i \in [1..n]$) of n independent studies. The null hypothesis for the Fisher's method[8] is H_0: H_{0i} is true for all $i \in [1..n]$. The alternative hypothesis is H_A: H_{0i} is false for at least one $i \in [0..n]$. Under the null hypothesis, all individual p-values are independently and uniformly distributed between zero and one. Fisher's method uses the log product of the p-values as the test statistic, which follows chi-squared distribution under the null: $X = -2\sum_{i=1}^{n} \ln(P_i) \sim \chi^2_{2n}$. This statistic is used to calculate the combined p-value, which represents how likely the individual p-values are obtained by chance. One disadvantage of Fisher's method is that if one of the individual p-values approaches zero, then the combined p-value approaches zero as well, regardless of other individual p-values. This eventually leads to an excessive false positive rate. Therefore, we propose to use the addition of the p-values, rather than their product.

Denote the sum of these p-values, $X = \sum_{i=1}^{n} P_i$ ($X \in [0,n]$), as the new random variable. X is known to follow the Irwin-Hall distribution[25,26] with the following probability density function:

$$f(x) = \frac{1}{(n-1)!} \sum_{i=0}^{\lfloor x \rfloor} (-1)^i \binom{n}{i} (x-i)^{n-1} \tag{1}$$

Unlike Fisher's method, the additive method is not sensitive to small individual p-values. However, for large values of n, Equation (1) involves some intensive computation due to a sum of a combinatorial and division by a factorial, the result of which can lead to an "arithmetic underflow", i.e. the result can be a number smaller than what a computer can actually store in memory.

To avoid arithmetic underflow, we change the random variable from the sum of the p-values to the average of the p-values.[15] For large values of n, we replace the additive method

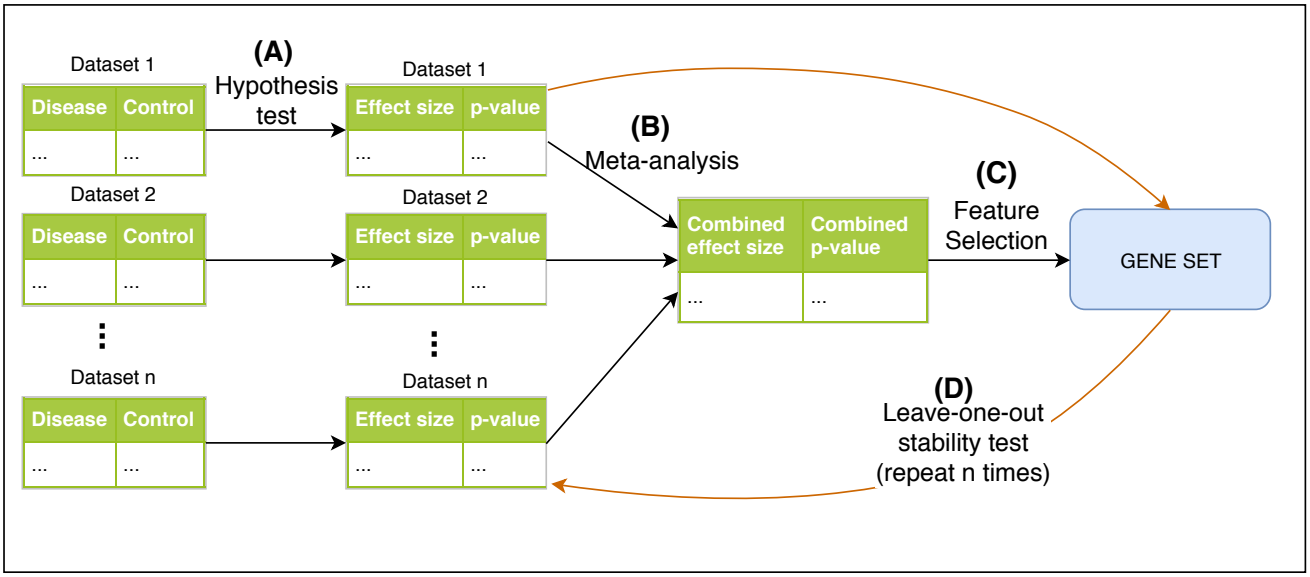

Figure 1. The proposed pipeline for robust meta-analysis.

with the Central Limit Theorem (CLT). The reason for the modification is that the additive method is accurate for small values of n, while the Central Limit Theorem is more accurate for large values of n. We select $n = 20$ as a conservative cut-off. In the rest of the manuscript, we refer to this method as addCLT.

2.2 Effect size and standard error

2.2.1 Standardized mean difference.

Similar to a commonly used fold-change, standardized mean difference (STD) represents the difference in expression values between two groups. Since the datasets are obtained from different platforms, the gene expression values are scaled differently in every dataset. Therefore, it is more reasonable to use standardized mean difference (SMD) as the metric to measure effect sizes, rather than raw mean difference. In this work, we use *Hedge's g*[27] as the metric to measure expression change between control and disease samples.

2.2.2 Estimated effect size and p-values.

The central tendency of effect sizes for the gene can be calculated either using a fixed-effects model or a random-effects model. A fixed-effects model would assume that there is one true effect size which underlies all of the studies in the analysis. However, this assumption is implausible since it cannot account for heterogeneity between studies.[21,28] In contrast, the random-effects model allows for variability of the true effect. For example, the effect size might be higher (or lower) in studies where the participants are older, or have a healthier lifestyle. The random-effects model assumes that each effect size estimate can be decomposed into two variance components by a two stage hierarchical process.[21,29] The first variance represents the variability of the effect size across studies, and the second variance represents the sampling error within each study.

Consider one specific gene and denote $y_1, y_2, \ldots, y_m$ as *Hedge's g* values computed for m studies. We can write the random-effects model as $y_i = \mu + N(0,\sigma^2) + N(0,\sigma_{\varepsilon_i}^2)$, where μ is the central tendency of the effect size, $N(0,\sigma^2)$ represents the error term by which the effect size in the i^{th} study differs from the central tendency μ, and $N(0,\sigma_{\varepsilon_i}^2)$ represents the sampling error. The overall effect size μ of the gene and its standard error σ are estimated iteratively, as described by our references.[21–23] The algorithm stops when further iteration does not change the values of μ and σ.

After the REML algorithm stops, we compute the z-score using the formula $z = \frac{\mu}{\sigma}$ and then calculate the left- and right-tailed p-values of observing such a z-score. The obtained μ and p-values (ep_l and ep_r where ep stands for "effect size p-value") represent the overall expression change of the gene and how reliable the estimated effect size is.

2.3 Leave One Out

The leave-one-out (LOO) method of analysis[30] is used to ensure that no single dataset has a disproportionate effect on the results of the analysis. We use LOO as an intrinsic part of our method in order to ensure that the set of genes we ultimately reach is refined, consistent, and robust against outliers.[16] In the scope of our study, the LOO analysis consists of n steps, where n is the number of datasets being used in the study. Consider the datasets as being numbered, 1 to n, and each step in the LOO analysis is numbered in the same way. If you are on step i of the LOO analysis, you would omit dataset i from the analysis for that iteration, then reintroduce it in the next. In total, we conduct $n+1$ analyses: n LOO and 1 analysis where all of the datasets are present. The $n+1$ outputted gene sets are intersected to obtain one final gene set.

Table 1. The 11 datasets used in our data analysis include 7 Affymetrix and 4 sequencing (RNASeq) datasets. The three datasets highlighted in green were used for training while the 8 datasets highlighted in blue were used for testing.

Dataset	#C	#D	Tissue
GSE5281	74	87	Entorhinal cortex, medial temporal gyrus, posterior cingulate, superior frontal gyrus, hippocampus, primary visual cortex
GSE36980	47	32	Frontal cortex, temporal cortex, and hippocampus
GSE48350	173	80	Entorhinal cortex, post-central gyrus, hippocampus, and superior frontal gyrus
GSE1297	9	22	Hippocampus
GSE4757	10	10	Entorhinal cortex
GSE16759	4	4	Parietal lobe
GSE39420	7	14	Posterior cingulate area
GSE53695	6	9	Drsolateral prefrontal cortex
GSE53697	6	9	Dorsolateral prefrontal cortex
GSE57152	16	8	Superior temporalis gyrus
GSE104704	16	11	Lateral temporal lobe

3 Results

Here we demonstrate the performance of the proposed method using 11 independent studies related to Alzheimer's disease. Alzheimer's disease (AD) is a neurodegenerative condition, and is one of the most common forms of dementia. Currently, the best method for diagnosing patients with Alzheimer's Disease is by assessing their symptoms. This often leads to late diagnoses, and mis-diagnoses, which reduces the potential for effective symptom management and treatment. Our proposed means for developing an effective early detection method is the application of meta-analysis techniques for use in gene expression data analysis from different regions of the brain. Comparing the gene expression data from samples that are diagnosed with Alzheimer's against that of healthy samples assists us in making informed decisions about which genes significantly contribute to the development of AD.

Table 1 shows the details of each dataset, such as the number of control and disease samples, tissues, and platforms. The 7 microarray datasets were downloaded from the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/). The Accession IDs of the 9 datasets are: GSE5281,[31] GSE36980, GSE48350, GSE1297, GSE4757, GSE16759, GSE39420. Pre-processing was performed on each dataset using the *threestep* function from the package *affyPLM version 1.38.0*.[32] The parameters used for the *threestep* function are: robust multi-array analysis (RMA) background adjustment, quantile normalization, and median polish summarization. The 7 datasets contain a total of 324 control samples, and 249 disease samples. We also obtained 4 RNA-Seq datasets from GEO. The accession IDs of the 4 RNA-Seq datasets are: GSE53695, GSE53697, GSE57152, and GSE104704. For data preprocessing, we used the *salmon*[33] package to align raw reads and quantify transcript expression. The expression value of a gene is calculated as the sum of its transcripts. We also used the *scater*[34] package to remove outliers from each of the datasets. The 4 RNA-Seq datasets contain a total of 44 control samples and 37 disease samples.

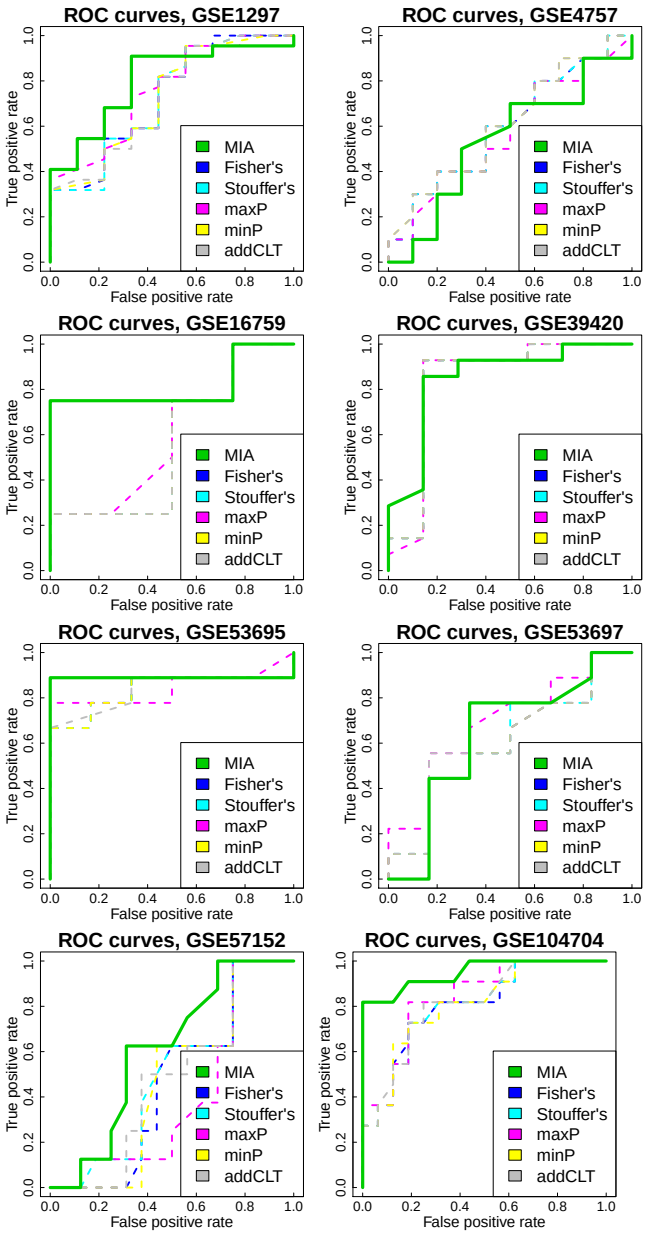

Figure 2. The Receiver Operating Characteristic (ROC) curves for each of the 8 testing datasets using six approaches: MIA, Fisher's, Stouffer's, maxP, minP, and addCLT.

We divide the 11 datasets into a training set and a testing set. The training set consists of 3 datasets (GSE5281, GSE36980, GSE48350) while the testing set consists of 8 datasets (GSE1297, GSE4757, GSE16759, GSE39420, GSE53695, GSE57152, GSE104704). Here we use six different methods to identify the set of biomarker genes which can be used to classify the samples in the testing datasets. The six methods are: MIA, Fisher's,[8] Stouffer's,[9] maxP,[11] minP,[10] and the addCLT method.[15,35] The file classical p-value-based methods (Fisher's, Stouffer's, maxP, minP, addCLT) are available in the package BLMA.[36]

The pipeline of MIA is described above (see Section 2 and Figure 1). For all methods, including MIA, we calculate the p-values and log fold-changes using the empirical Bayesian approach provided by the limma package.[19] Since we are interested in genes which are down-regulated as a result of Alzheimer's disease, we only focus on the left-sided p-values. In addition, we narrow the analysis to only genes which are consistently down-regulated in all 3 training datasets. We use Fisher's method to combine the independent p-values of each gene. After this, each gene has a combined p-value. We then correct the p-values using False Discovery Rate (FDR).[24] Finally we define the biomarker genes as the genes that are consistently down-regulated in every single training dataset and have an FDR-adjusted p-value smaller than 1%. This defines the set of biomarker genes for Fisher's method. In addition to Fisher's method, we also used 4 other methods to combine p-values: Stouffer's,[9] maxP,[11] minP,[10] and the addCLT method. These 5 approaches only differ in the way that the p-values are combined.

Each of the six methods identified a set of biomarker genes. MIA identified 31 impacted genes while Fisher's, Stouffer's, maxP, minP, and addCLT identified 4406, 4394, 745, 4282, and 3511, respectively. In the next step, we used each of these sets to classify the samples of the 8 testing datasets. For each of the datasets in the testing set, we calculated the median expression of the given biomarker genes for each sample and then use that median value to classify the samples.

Figure 2 shows the Receiver Operating Characteristic (ROC) curves using six different meta-analysis approaches. In each panel, the horizontal axis represents the False Positive Rate (FPR) and the vertical axis represents the True Positive Rate (TPR). Since the identified biomarkers are down-regulated genes, a sample which has a computed metric lower than the threshold would be considered as being positive (a disease sample). For each dataset, we first set the threshold below the range of expression values to get zero positives – 0% TPR, and 0% FPR. We then gradually increase the threshold to get more positives. Simultaneously, the TPR and FPR also increase. When the threshold is set high enough, all of the samples become negatives. With this, TPR and FPR reach 100%. Overall, the ROC curves for MIA are above the ROC curves of other methods, indicating that MIA has higher TPR and lower FPR compared to other methods.

For an ideal classification, there is a threshold where all of the samples are correctly classified, with 100% TPR and 0%FPR. At this threshold, the ROC curve jumps from [0,0] to [0,1], making the area under the curve (AUC) to be 1. In principle, the AUC values are used to access the performance of the classification methods. The box-plot of the AUC values is shown in Figure 3, and the values for each dataset are shown in table 2. For each row, cells highlighted in green have the highest AUC value. For six out of eight datasets tested, MIA outperforms other methods by having the highest AUC values. In addition, the mean and median AUC values obtained by MIA are higher than those obtained by the five other methods.

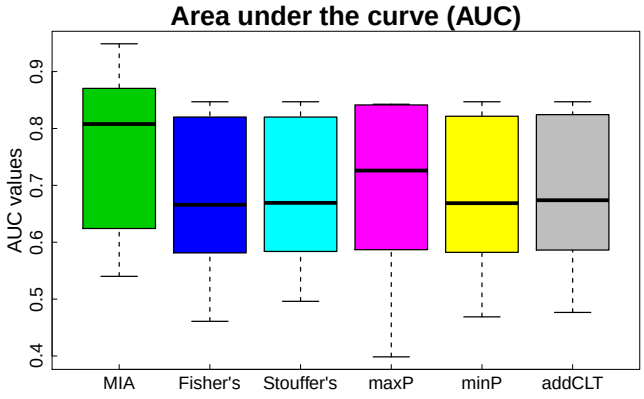

Figure 3. Box-plot of AUC values obtained from 8 datasets using six different methods. The AUC values obtained from MIA are higher than those obtained from other methods.

Table 2. AUC values obtained from 8 testing datasets using 6 different methods: MIA, Fisher's, Stouffer's, maxP, minP, and the addCLT method. Cells highlighted in green have the highest AUC value for the corresponding dataset. For 6 out of 8 datasets tested, MIA have higher AUC values than the rest. MIA also has the highest mean and median AUC values.

Dataset	MIA	Fisher	Stouffer	maxP	minP	addCLT
GSE1297	0.803	0.729	0.727	0.757	0.727	0.727
GSE4757	0.540	0.600	0.605	0.580	0.610	0.610
GSE16759	0.812	0.562	0.562	0.593	0.562	0.562
GSE39420	0.852	0.846	0.846	0.841	0.846	0.846
GSE53695	0.888	0.833	0.833	0.842	0.833	0.833
GSE53697	0.638	0.601	0.611	0.694	0.601	0.620
GSE57152	0.609	0.460	0.496	0.398	0.468	0.476
GSE104704	0.948	0.806	0.806	0.840	0.833	0.833
Mean	0.762	0.680	0.686	0.694	0.682	0.686
Median	0.808	0.666	0.669	0.726	0.669	0.674

Overall, the median AUC value for MIA is 0.81.This indicates that the genes obtained from the proposed approach are important in identifying patients with Alzheimer's disease.

4 Conclusions and Discussion

In this paper we present a meta-analysis method called MIA to identify a set of genes that can be used to classify individual samples as disease or control. We utilized techniques from both p-value-based and effect-size-based meta-analyses to obtain a robust biomarker. The obtained genes are significant from a classical hypothesis testing perspective, as well as have the effect sizes that are outside the bounds of standard error.

We validated our method using 11 independent Alzheimer's datasets obtained from different research laboratories with significantly different high-throughput platforms. The proposed approach was compared against five classical meta-analysis approaches. Our approach identified a tight set of genes (31) while other methods identified relatively large biomarker sets with thousands of differentially expressed genes. MIA outperforms existing approaches by having the highest median and mean AUC values. The identified biomarkers are able

to accurately distinguish Alzheimer's patients from healthy controls.

MIA implements several tried and true statistical modeling methods and algorithms in conjunction with one another to produce it's outcome. Each step provides its own unique contribution to the final result, and serves a particular purpose. For instance, we chose the REML algorithm in place of other effect-size combination methods because of its affinity toward modeling real-world data. We expect that the applications of MIA are to be generally geared toward drawing inference from datasets which come from significantly different sources, which therefore enables it to benefit from the flexibility of algorithms such as REML. We also chose to implement the LOO strategy, specifically to reduce the likelihood of false positives, which is nearly ubiquitous across modern meta-analysis techniques. The power of MIA is rooted in how it integrates multiple techniques in such a way that the strengths of each mitigate the shortcomings of others.

References

1. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Research* **41**, D991–D995 (2013).
2. Rustici, G. *et al.* ArrayExpress update–trends in database growth and links to data analysis tools. *Nucleic Acids Research* **41**, D987–D990 (2013).
3. Nguyen, T., Mitrea, C. & Draghici, S. Network-based approaches for pathway level analysis. *Current Protocols in Bioinformatics* **61**, 8–25 (2018).
4. Tan, P. K., Downey, T. J., Spitznagel Jr, E. L., Xu, P., Fu, D., Dimitrov, D. S., Lempicki, R. A., Raaka, B. M. & Cam, M. C. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Research* **31**, 5676–5684 (2003).
5. Ein-Dor, L., Zuk, O. & Domany, E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences of the Unites States of America* **103**, 5923–5928 (2006).
6. Tseng, G. C., Ghosh, D. & Feingold, E. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Research* **40**, 3785–3799 (2012).
7. Rhodes, D. R., Barrette, T. R., Rubin, M. A., Ghosh, D. & Chinnaiyan, A. M. Meta-analysis of microarrays interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Research* **62**, 4427–4433 (2002).
8. Fisher, R. A. *Statistical methods for research workers* (Oliver & Boyd, Edinburgh, 1925).
9. Stouffer, S., Suchman, E., DeVinney, L., Star, S. & Williams, J., RM. *The American Soldier: Adjustment during army life*, vol. 1 (Princeton University Press, Princeton, 1949).
10. Tippett, L. H. C. *The methods of statistics* (Williams & Norgate, London, 1931).
11. Wilkinson, B. A statistical consideration in psychological research. *Psychological Bulletin* **48**, 156 (1951).
12. Li, J. & Tseng, G. C. An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *The Annals of Applied Statistics* **5**, 994–1019 (2011).
13. Choi, H., Shen, R., Chinnaiyan, A. M. & Ghosh, D. A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments. *BMC Bioinformatics* **8**, 364 (2007).
14. Sullivan, G. M. & Feinn, R. Using effect size-or why the p value is not enough. *Journal of Graduate Medical Education* **4**, 279–282 (2012).
15. Nguyen, T., Tagett, R., Donato, M., Mitrea, C. & Draghici, S. A novel bi-level meta-analysis approach-applied to biological pathway analysis. *Bioinformatics* **32**, 409–416 (2016).
16. Nguyen, T., Mitrea, C., Tagett, R. & Drăghici, S. DANUBE: Data-driven meta-ANalysis using UnBiased Empirical distributions - applied to biological pathway analysis. *Proceedings of the IEEE* **105**, 496–515 (2017).
17. Nguyen, T., Diaz, D., Tagett, R. & Draghici, S. Overcoming the matched-sample bottleneck: an orthogonal approach to integrate omic data. *Nature Scientific Reports* **6**, 29251 (2016).
18. Kim, D. H., Yeo, S. H., Park, J.-M., Choi, J. Y., Lee, T.-H., Park, S. Y., Ock, M. S., Eo, J., Kim, H.-S. & Cha, H.-J. Genetic markers for diagnosis and pathogenesis of alzheimer's disease. *Gene* **545**, 185–193 (2014).
19. Smyth, G. K. Limma: linear models for microarray data. In Gentleman, R., Carey, V., Dudoit, S., Irizarry, R. & Huber, W. (eds.) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, 397–420 (Springer, New York, 2005).
20. Kallenberg, O. *Foundations of modern probability* (Springer-Verlag, New York, 2002).
21. Viechtbauer, W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics* **30**, 261–293 (2005).
22. Harville, D. A. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72**, 320–338 (1977).
23. Corbeil, R. R. & Searle, S. R. Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics* **18**, 31–38 (1976).

24. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of The Royal Statistical Society B* **57**, 289–300 (1995).

25. Hall, P. The distribution of means for samples of size n drawn from a population in which the variate takes values between 0 and 1, all such values being equally probable. *Biometrika* **19**, 240–244 (1927).

26. Irwin, J. O. On the frequency distribution of the means of samples from a population having any law of frequency with finite moments, with special reference to Pearson's Type II. *Biometrika* **19**, 225–239 (1927).

27. Hedges, L. V. & Olkin, I. *Statistical method for meta-analysis* (Academic Press, London, 2014).

28. Milliken, G. A. & Johnson, D. E. *Analysis of messy data volume 1: designed experiments*, vol. 1 (Chapman & Hall/CRC, London, 2009).

29. Goldstein, H. *Multilevel statistical models*, vol. 922 (John Wiley & Sons, New York, 2011).

30. Cawley, G. C. & Talbot, N. L. Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. *Pattern Recognition* **36**, 2585–2592 (2003).

31. Liang, W. S. *et al.* Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons. *Proceedings of the National Academy of Sciences* **105**, 4441–4446 (2008).

32. Bolstad, B. M. *Low-level analysis of high-density oligonucleotide array data: background, normalization and summarization*. Ph.D. thesis, University of California (2004).

33. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14**, 417 (2017).

34. McCarthy, D. J., Campbell, K. R., Lun, A. T. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell rna-seq data in r. *Bioinformatics* **33**, 1179–1186 (2017).

35. Nguyen, T., Diaz, D. & Draghici, S. TOMAS: A novel TOpology-aware Meta-Analysis approach applied to System biology. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 13–22 (ACM, New York, 2016).

36. Nguyen, T. & Draghici, S. *BLMA: A package for bi-level meta-analysis*. Bioconductor (2017). R package.