PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Increasing reproducibility, robustness, and generalizability of biomarker selection from meta-analysis using Bayesian methodology

Laurynas Kalesinskas[1,2,3], **Sanjana Gupta**[1,2], **Purvesh Khatri**[1,2]*

1 Institute for Immunity, Transplantation and Infection, School of Medicine, Stanford University, Stanford, California, United States of America, **2** Center for Biomedical Informatics Research, Department of Medicine, Stanford University, Stanford, California, United States of America, **3** Department of Biomedical Data Science, School of Medicine, Stanford University, Stanford, California, United States of America

* pkhatri@stanford.edu

OPEN ACCESS

Citation: Kalesinskas L, Gupta S, Khatri P (2022) Increasing reproducibility, robustness, and generalizability of biomarker selection from meta-analysis using Bayesian methodology. PLoS Comput Biol 18(6): e1010260. https://doi.org/10.1371/journal.pcbi.1010260

Editor: Yuchao Jiang, University of North Carolina at Chapel Hill Gillings School of Global Public Health, UNITED STATES

Received: January 9, 2022

Accepted: May 29, 2022

Published: June 27, 2022

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1010260

Data Availability Statement: The bayesMetaIntegrator R package is publicly available for use at https://github.com/Khatri-Lab/

Abstract

A major limitation of gene expression biomarker studies is that they are not reproducible as they simply do not generalize to larger, real-world, heterogeneous populations. Frequentist multi-cohort gene expression meta-analysis has been frequently used as a solution to this problem to identify biomarkers that are truly differentially expressed. However, the frequentist meta-analysis framework has its limitations–it needs at least 4–5 datasets with hundreds of samples, is prone to confounding from outliers and relies on multiple-hypothesis corrected p-values. To address these shortcomings, we have created a Bayesian meta-analysis framework for the analysis of gene expression data. Using real-world data from three different diseases, we show that the Bayesian method is more robust to outliers, creates more informative estimates of between-study heterogeneity, reduces the number of false positive and false negative biomarkers and selects more generalizable biomarkers with less data. We have compared the Bayesian framework to a previously published frequentist framework and have developed a publicly available R package for use.

Author summary

There has long been a reproducibility crisis in medical research–driven by small, single-cohort studies with low-to-moderate statistical power. One of the reasons for this lack of generalizability is not accounting for heterogeneity representative of the real-world patient population. To address this issue, researchers have turned to meta-analysis–which allows for researchers to combine data from across previously published studies to generate an overall estimate of an effect, which has been used with gene expression data to create diagnostic and prognostic markers of disease. However, traditional meta-analysis techniques have limitations–they need at least 4–5 datasets with hundreds of samples and are prone to confounding from outliers in datasets. In this study, we create a new framework for gene expression meta-analysis using Bayesian statistics and show that it is more robust to outliers, creates more informative estimates of heterogeneity, reduces the

bayesMetaIntegrator. All data used in this study is publicly available – identifiers of which are found in S1–S3 Tables.]

Funding: PK is funded in part by the Bill and Melinda Gates Foundation (OPP1113682); the National Institute of Allergy and Infectious Diseases (NIAID) grants 1U19AI109662 and U19AI057229; Department of Defense contracts W81XWH-18-1-0253 and W81XWH1910235; and the Ralph & Marian Falk Medical Research Trust. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

amount of data required, and reduces the number of false positive and false negative biomarkers. We have compared the Bayesian framework to a previously published framework and have developed a publicly available R package for use.

Introduction

With the advent of high-throughput transcriptomics, researchers have been able to profile gene expression in millions of samples at low costs, opening many new avenues of research. However, single-cohort studies with low-to-moderate statistical power are one of the four horsemen of irreproducibility in biomedical research [1]. Recent studies have estimated that only 10–25% of biomedical studies are reproducible [2–4]. This is especially a problem in biomarker studies, where differentially expressed genes (DEGs) between subjects with disease of interest and controls rarely generalize to the real-world patient populations. One of the reasons for this lack of generalizability is not accounting for heterogeneity representative of the real-world patient population.

Broadly, there are three sources of heterogeneity in the real-world patient population: biological (age, sex, tissue, cell type), clinical (treatment, disease duration, comorbidities), and technical (experimental protocol, batch effects). Traditionally, in a single cohort analysis, these sources of heterogeneity reduce the statistical power, requiring a large number of samples. Therefore, single cohort studies strive to increase statistical power by limiting heterogeneity as much as possible. However, this reduction in heterogeneity in single cohort studies leads to reduced generalizability to heterogeneous, real-world patient populations. We have repeatedly shown that leveraging heterogeneity across independent cohorts using a frequentist meta-analysis approach can identify robust disease signatures that are diagnostic and prognostic, and have been translated into a point-of-care test for clinical use [5,6]. We have previously identified best practices for gene expression meta-analyses, such as the number of studies and samples needed [7].

Despite its repeated success in identifying robust disease signatures, the frequentist approach has its limitations. First, previous work has shown that approximately 4–5 datasets with about 250 samples are needed to perform a successful frequentist meta-analysis [7]. However, several diseases simply do not have enough samples or datasets publicly available for successful integration using this guideline. Second, the statistic used to estimate and summarize effect sizes (e.g., Cohen's *d*, Hedge's *g*) can be susceptible to outlier samples within a subset of studies, resulting in misleading effect size estimates. Finally, frequentist approaches rely on multiple hypotheses corrected p-values, which are shown to be substantially underestimated [7]. Bayesian meta-analysis approaches have the potential to overcome these limitations. For example, Bayesian estimation has previously been shown to be more outlier resistant than traditional hypothesis testing [8]. Importantly, unlike frequentist meta-analysis, adjusting for multiple comparisons is not required for Bayesian approaches and yields more efficient and reliable estimates of effect [9].

We compared a frequentist approach with a new framework utilizing Bayesian approximation supersedes the t-test (BEST) for the meta-analysis of transcriptome using multiple independent datasets from humans with different diseases [8]. Here, we show that using this Bayesian approach, we are able to: 1) select more generalizable and robust biomarkers with fewer datasets, 2) be robust to outliers, 3) create better estimates of between study heterogeneity for biomarker selection, and 4) reduce the number of false positives and false negative genes for classification. This framework has also been developed into an R package,

bayesMetaIntegrator, that is publicly available for use (https://github.com/Khatri-Lab/bayesMetaIntegrator).

Results

Bayesian meta-analysis is resistant to outliers and provides a better estimate of heterogeneity in gene expression meta-analysis

We investigated whether the two meta-analysis approaches, Bayesian (**Fig 1**) and frequentist, identified the same or different set of genes using four publicly available asthma bronchial epithelial cell gene expression datasets—differentiating samples from asthma patients and healthy controls (**S1 Table**) [10–12]. Although the summary effect sizes were highly correlated between approaches ($r = 0.94$, $p<2.2e\text{-}16$; **Fig 2A**), the Bayesian approach consistently estimated higher between-dataset heterogeneity, τ^2, than the frequentist approach (**Fig 2B**). While the frequentist approach found a large number of genes (26%) with no between-dataset heterogeneity, the Bayesian approach did not assign near-zero heterogeneity ($\tau^2 < 0.01$) to any genes. This difference in τ^2 between the two approaches is due to how it is estimated. In a frequentist meta-analysis, high within-study heterogeneity leads to wider confidence intervals, which in turn drowns out the between-study heterogeneity. In contrast, Bayesian meta-analysis uses a probabilistic distribution to represent τ^2 instead of a confidence interval, resulting in more conservative estimates of heterogeneity. The range of between-datasets heterogeneity was substantially higher for the Bayesian approach compared to the frequentist approach, further suggesting a Bayesian approach is more conservative than a frequentist approach.

Interestingly, despite the high correlation between summary effect sizes, within-dataset effect size correlations ranged widely from 0.87 to 0.95. Closer examination found that both approaches differed substantially in effect size estimates for a subset of genes. For example, in GSE64913, a subset of genes has an effect size of 0 when using the Bayesian approach, but a non-zero effect size when using a frequentist meta-analysis (**Fig 2B**). Therefore, we investigated whether these differences in effect size estimates and between-dataset heterogeneity will lead to identification of different set of differentially expressed genes (DEGs).

We used false discovery rate (FDR) for the frequentist approach and Bayesian probability for the Bayesian approach as the measures of statistical significance. Low correlation between both measures ($r = 0.37$, $p<2.2e\text{-}16$) suggested they identified different sets of DEGs. As the number of significant genes by either approach increased, the number of genes identified as

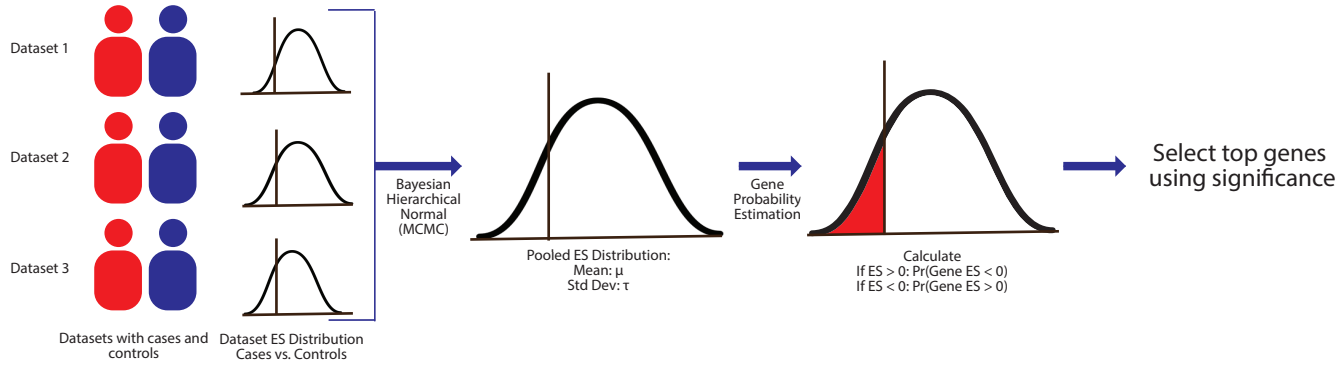

Fig 1. Bayesian gene expression schematic. We first use the BEST framework to estimate the posterior distribution of effect size between cases and controls for each gene in each dataset. Then we combine the distributions from independent studies using a gaussian hierarchical model, estimating both the pooled effect size and between-study heterogeneity in the process. Following, we estimate the probability of a gene being upregulated or downregulated based on the pooled effect size distribution.

https://doi.org/10.1371/journal.pcbi.1010260.g001

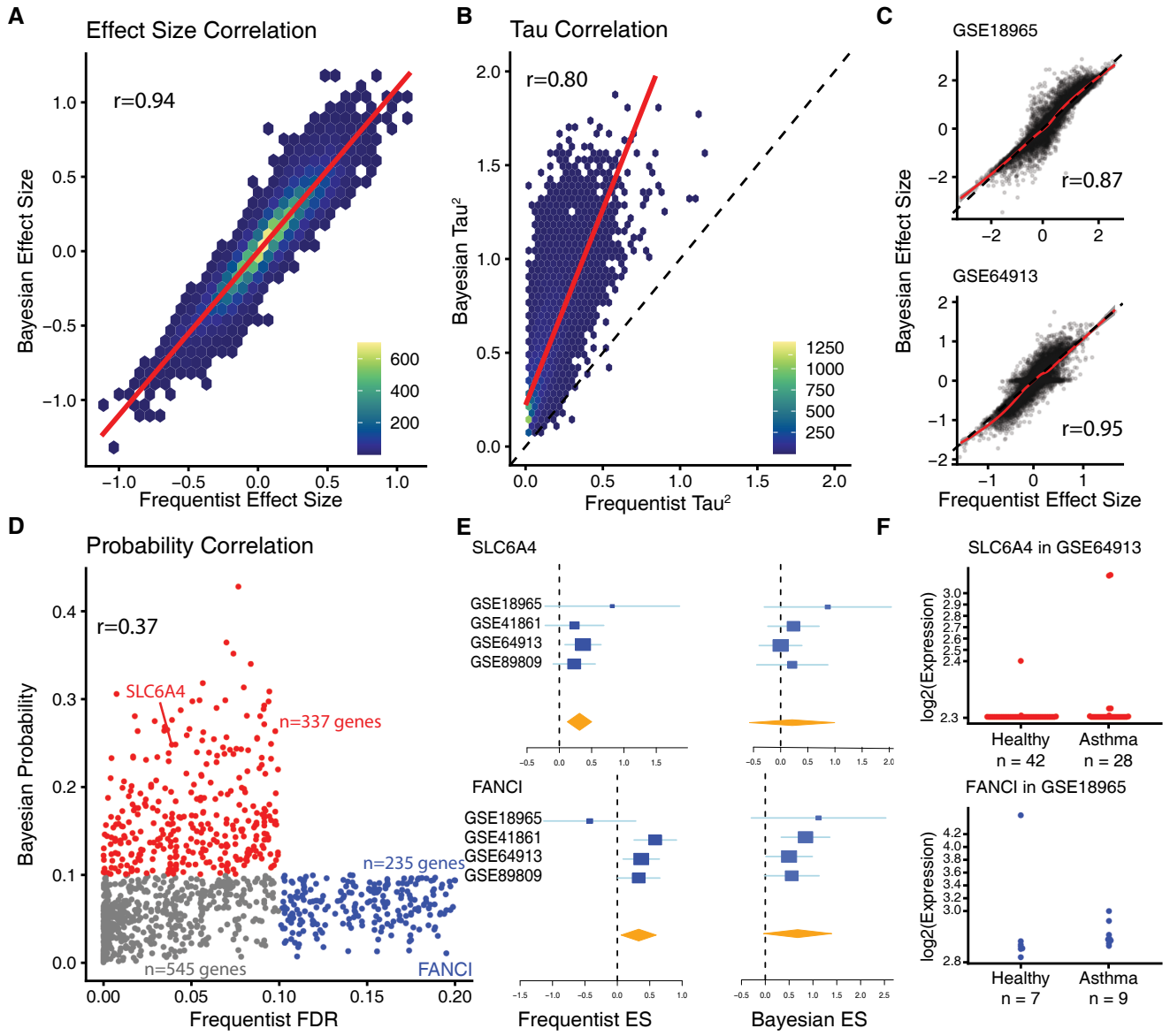

Fig 2. Asthma Meta-Analysis Comparison. (A) Comparison of pooled effect sizes by the frequentist and Bayesian meta-analysis for all genes. Pearson correlation (r) = 0.94. (B) Comparison of between-dataset heterogeneity, τ^2, by the frequentist and Bayesian meta-analysis for all genes. Pearson correlation (r) = 0.80. (C) Comparison of within-dataset effect size estimates by the frequentist and Bayesian meta-analysis for all genes. (D) Comparison between false discovery rate from the frequentist meta-analysis and probabilities from the Bayesian meta-analysis. (E) The x axes represent standardized mean difference between patients with asthma and healthy controls, computed as Hedges' *g*, in $\log_2$ scale. The size of the blue rectangles is proportional to the standard error of mean difference in the study. Whiskers represent the 95% confidence interval. The orange diamonds represent overall, combined mean difference (summary effect size) for a given gene. Width of the orange diamonds represents the 95% confidence interval of overall mean difference. (F) Expression of SLC6A4 in GSE64913 and FANCI in GSE18965 in patients with asthma and healthy controls.

https://doi.org/10.1371/journal.pcbi.1010260.g002

significant by both methods also increased (S3 Fig). Out of 1,117 DEGs that were statistically significant by either approach, 545 genes (48.8%) were statistically significant by both approaches (FDR < 10% and Bayesian probability < 0.1). The remaining 572 genes were significant when using either the frequentist approach (235 genes, 21%) or the Bayesian approach (337 genes, 30.2%) (Fig 2D).

Importantly, for genes that were significant by either approach but not both, we found that the Bayesian approach is more robust to outlier samples within a single dataset than the frequentist approach. For example, *SLC6A4* was significant by the frequentist approach (FDR<4%), but not by the Bayesian approach (p = 0.24). Comparison of effect sizes from both approaches showed that for GSE64913, the frequentist approach estimated statistically significant non-zero effect size, whereas the Bayesian approach estimated non-significant effect size (Fig 2E). Further analysis showed that although 59 out of 70 samples in GSE64913 had identical expression values for *SLC6A4*, irrespective of estimate of statistically significant effect size by the frequentist approach was driven by only 2 out of 70 samples (Fig 2F). In contrast, the Bayesian approach correctly estimated near-zero effect size due to its reliance on parameter estimation and sampling, and was not confounded by a small number of outliers. This observation demonstrated that the Bayesian approach reduced false positives.

In contrast, another gene, *FANCI*, was statistically significant by the Bayesian approach (p = 0.01), but not by the frequentist approach (FDR = 19%). Although both approaches estimated the effect size for *FANCI* in the same direction, the difference in statistical significance was due to the difference in estimated effect size in a single dataset (GSE18965) (Fig 2E). The frequentist approach estimated negative effect size for *FANCI* in GSE18965, which was driven by a single healthy control sample (Fig 2F), which in turn led to higher between-study heterogeneity for the gene and the summary effect size for the gene being statistically insignificant. In contrast, the Bayesian approach was not confounded by a single sample, correctly estimated its effect size as positive in GSE18965, and identified it as being statistically significant overall. Collectively, these results show that the Bayesian meta-analysis approach is robust to outliers, which in turn reduces false positives and false negatives.

Comparing Bayesian and frequentist meta-analysis methods

A key difference between the frequentist and Bayesian meta-analysis approaches is how between-study heterogeneity affects estimates of summary effect size and its statistical significance. To investigate the effect of between-study heterogeneity, we simulated data across 5 independent studies (**Methods**) such that either the effect size, its variance, or both changed for one study (Fig 3). Increasing the effect size for one study without changing its variance estimate, which simulated adding a study with high certainty of a strong positive effect, statistical significance reduced for the frequentist approach (i.e., higher FDR), but increased for the Bayesian approach (i.e., lower probability), although the between-study heterogeneity increased for both approaches (Fig 3B). When we increased only variance or both effect size and variance, the summary heterogeneity and statistical significance decreased for both approaches (Fig 3C and 3D). These results, combined with robustness of the Bayesian approach to outliers within a dataset, are desired characteristics for identifying generalizable signal across heterogeneous datasets.

Comparing the predictive performance of Bayesian and frequentist meta-analysis

Given that the Bayesian approach is robust to outlier samples within a study and better estimates between-study heterogeneity, we investigated whether it would identify DEGs that are more generalizable to unseen data than the frequentist approach. To investigate this, we applied both meta-analysis approaches to transcriptome profiles from patients with cardiomyopathy (14 datasets, 1039 samples) [13–23] or tuberculosis (27 datasets, 3069 samples) [24–45]. For both diseases, we identified most differentially expressed genes by successively increasing the number of datasets and compared their discriminatory power in unused

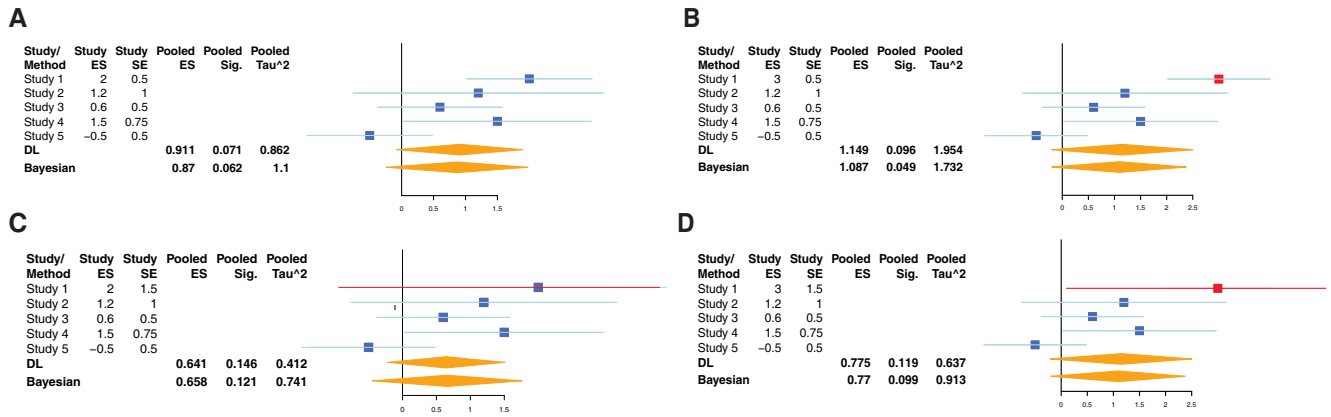

Fig 3. Simulated Meta-Analysis. A) Using simulated data in five studies, we explored the effect of changing the effect size and variance of a single study on each model's pooled effect size, heterogeneity (τ^2) and significance. Here, we show the baseline meta-analysis. We picked a scenario where both Bayesian and frequentist are borderline significant. B) Increasing effect size. We increased the effect size of a single study–creating high certainty of a strong positive effect of that study. The Bayesian model became more significant, whereas the frequentist model became less significant. Both increased in effect size and τ^2. C) Increasing variability. We increased the variability of a single study. We found that both methods behave similarly–with effect size, significance and τ^2 decreasing. D) Increasing variability and effect size. We increased both the variability and effect size of a single study. We found that both methods behave similarly–with effect size, significance and τ^2 decreasing.

https://doi.org/10.1371/journal.pcbi.1010260.g003

datasets (Fig 4A and **Methods**). For both diseases, we randomly selected *N* datasets 100 times and applied both meta-analysis approaches to each set of randomly selected *N* datasets. For each iteration, we selected DEGs with absolute summary effect size > 0.6 and used the top 10 genes with the lowest FDR or Bayesian probability. We used difference of geometric means of over- and under-expressed genes as a classifier in unseen datasets to distinguish cases from healthy controls. We chose difference of geometric mean as a classification model because such a classifier has been repeatedly demonstrated to be generalizable and has been translated in a point-of-care test [5,6,46,47]. When we varied the effect size threshold (0.4 to 1.1) and the number of selected genes (10 to 200), while keeping the number of datasets used for analysis constant at 4, the genes selected using the Bayesian approach consistently led to higher AUC (S2 Fig).

First, we compared area under the receiver operating characteristic (AUROC) curves in the unused datasets in each iteration as a proxy for identifying generalizable gene signatures for both meta-analysis approaches. For both diseases, irrespective of the number of datasets used, the DEGs identified using the Bayesian approach had consistently higher AUROC in unseen datasets than using the frequentist approach. When using 2 out of 27 datasets to identify DEGs for tuberculosis, there was no significant difference in AUROC between the two approaches, which suggests that $N = 2$ does not represent the heterogeneity across the other 27 datasets. Interestingly, for both diseases, the median AUROCs for the Bayesian approach using 3 datasets was always equal or greater than the median AUROCs for the frequentist approach using substantially larger number of datasets (9 datasets for cardiomyopathy, 14 datasets for tuberculosis). This result demonstrated that the DEGs identified using the Bayesian approach are more generalizable to previously unseen data than those identified using the frequentist approach. Further, it also suggests that using as few as 3 datasets may be sufficient to identify robust gene signatures of a disease using the Bayesian meta-analysis approach.

Next, we investigated between-study heterogeneity, τ^2, and effect sizes for the DEGs identified by the two approaches and whether those differences explained more generalizability for the Bayesian approach. For both diseases, median τ^2 decreased with the increasing number of datasets (Fig 4C). The DEGs identified by the Bayesian approach had significantly lower τ^2

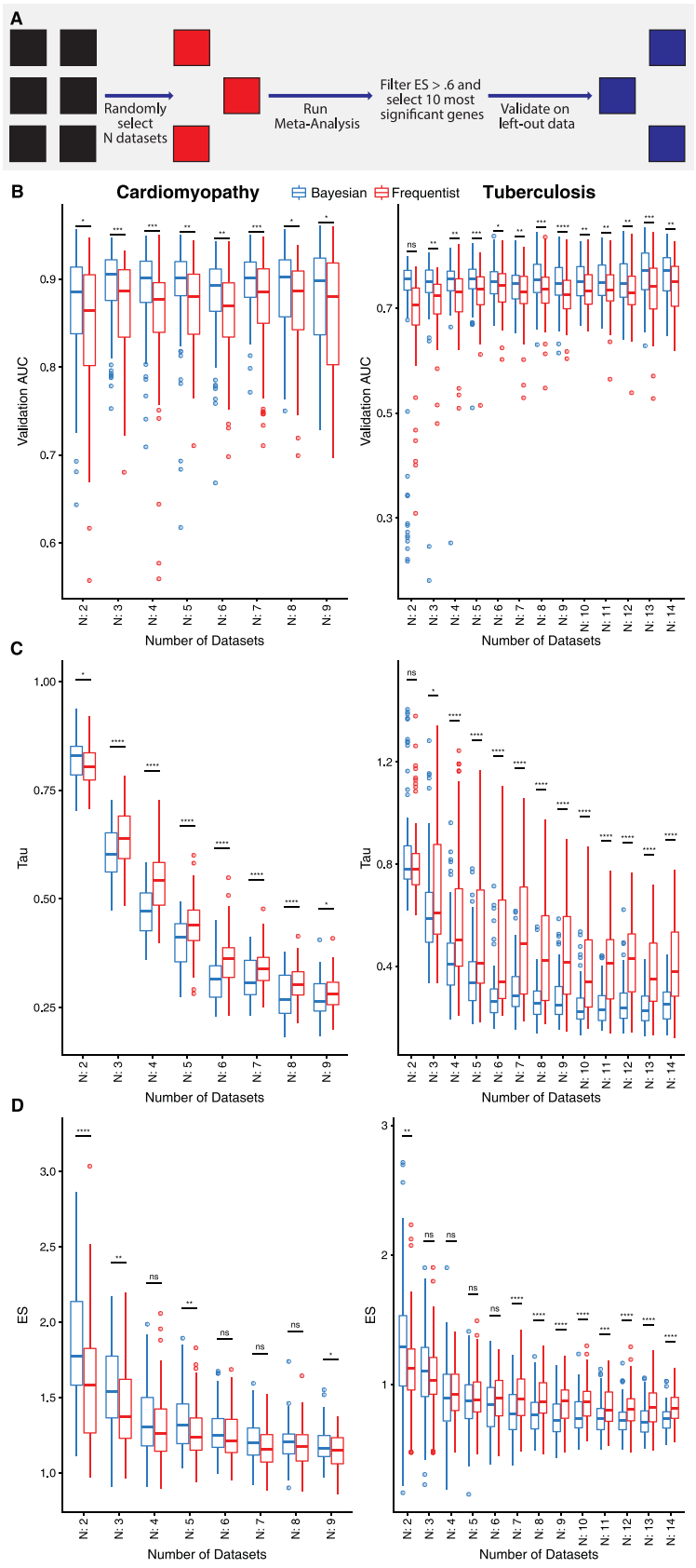
A
Randomly select N datasets
Run Meta-Analysis
Filter ES > .6 and select 10 most significant genes
Validate on left-out data
B
Cardiomyopathy
Bayesian
Frequentist
Tuberculosis
Validation AUC
Number of Datasets
C
Tau
Number of Datasets
D
ES
Number of Datasets

Fig 4. Comparing gene signatures from frequentist and Bayesian meta-analysis. A) Selection method—for both diseases, we randomly selected *N* datasets 100 times. For each iteration, we selected DEGs with absolute pooled effect size > 0.6 and selected 10 genes with the smallest FDR or Bayesian probability. The signature is then created using a geometric mean signature score and validated on completely left-out datasets. B) Average AUC–the Bayesian method picked gene signatures that are more generalizable—achieving higher AUCs for both diseases on left-out datasets. The Bayesian method also requires less datasets to achieve higher classifier performance. C) Average τ–we find that the between study heterogeneity (τ) is significantly lower for the gene signatures selected by the Bayesian method compared to the frequentist, suggesting the Bayesian method is selecting genes that have lower heterogeneity. Due to the frequentist method calculating τ as 0 for a large percentage of genes, Bayesian tau is displayed for all genes. D) Mean pooled effect size–we do not see consistent trends in the average effect sizes of the gene signatures, except for the tuberculosis analysis where the frequentist method tended to pick higher effect size genes with larger Ns. $^{****} < 0.0001$, $^{***} < 0.001$, $^{**} < 0.01$, $^{*} < 0.05$, ns: > 0.05.

https://doi.org/10.1371/journal.pcbi.1010260.g004

than those identified by the frequentist approach. Interestingly, for the frequentist approach as the number of datasets increased, variability in τ^2 increased for tuberculosis analysis but not cardiomyopathy analysis. However, there was no consistent difference in the summary effect sizes of the genes selected by either approach. For example, for cardiomyopathy datasets, the difference in summary effect sizes of DEGs were mostly non-significant, whereas for tuberculosis datasets, there was no statistically significant difference in summary effect sizes, when using a smaller number of datasets ($N \leq 6$), but the frequentist approach selected genes with significantly higher summary effect sizes for a larger number of datasets ($N > 6$). Interestingly, for tuberculosis analysis, higher summary effect size and variability in τ^2 with the increased number of datasets suggest that the DEGs may be affected by outlier datasets, which in turn reduces their generalizability to unseen datasets with reduced AUROCs.

To further investigate the differences between gene sets created using Bayesian or frequentist meta-analysis, we performed an overall meta-analysis with all of the Asthma, Cardiomyopathy and Tuberculosis datasets. Calculating the Jaccard similarity of the gene sets, we found that for the smallest meta-analysis, asthma, the Jaccard similarity between the models was low for the top 500 genes, only reaching 0.3. For the larger meta-analyses, Cardiomyopathy and Tuberculosis, we found that the Jaccard similarity increased up to 0.5 (**S3 Fig**). In each case, the Jaccard similarity increased until reaching a plateau. We also compared the gene sets using pathway analysis using ReactomePA [48] (**S4 Fig**). The pathways represented were similar in both analysis types for all three analyses, which suggested that although Bayesian meta-analysis identified more generalizable genes that have higher discriminatory power, it is still identifying the same biological pathways as frequentist meta-analysis.

Collectively, our results show that compared to the frequentist approach, the Bayesian approach for meta-analysis identifies genes with lower between-study heterogeneity and comparable summary effect sizes, and is robust to outlier samples, which in turn leads to more generalizable classifier for unseen datasets. Our results also suggest that the Bayesian approach requires lower number of datasets to identify generalizable DEGs compared to the frequentist approach.

Discussion

We performed three gene expression meta-analyses to compare the Bayesian and frequentist meta-analysis approaches. Using dozens of publicly available gene expression studies, we found that Bayesian approach tends to identify differentially expressed genes that have lower between-dataset heterogeneity and higher discriminatory power, which leads to more generalizable classifiers. Importantly, we found that the Bayesian approach consistently required lower number of datasets than the frequentist approach.

Several factors contribute to drive these effects for the Bayesian meta-analysis approach. First, our analysis showed that the Bayesian approach is resistant to outliers due to the *t*-

distribution underlying the estimation of the effect size distribution per dataset. Second, the Bayesian approach uses probabilistic distributions to represent effect size in each dataset as opposed to confidence intervals that the frequentist approach relies on. When the number of datasets used for meta-analysis is small and the within-dataset variability is higher, confidence intervals tend to be wider, leading the frequentist approach to estimate the between-study heterogeneity as zero or very low for a large number of genes the large within dataset heterogeneity. In contrast, the use of probabilistic distribution leads to conservative estimates of between-study heterogeneity. For biomarker discovery, this is preferable, for we seek to find the biomarkers that have the smallest between-study heterogeneity across all datasets in our analysis. Finally, the p-values in the frequentist approach represent the probability of observing data under a hypothesis of no effect and must be multiple hypothesis adjusted, whereas Bayesian probability represents the posterior belief of the difference between groups and require no multiple hypothesis adjustment [49].

Although the Bayesian meta-analysis tends to perform better than the frequentist method at finding consistently differentially expressed genes across studies with low heterogeneity, one area in which it would be less advantageous is the unsupervised identification of subgroups within patient populations. In this case, we would want to select for genes that have high effect sizes when compared to controls, but also have heterogeneity to separate between cases. For this task, the frequentist method would likely be more effective than the Bayesian method described in this study. However, priors and probability calculations can be adjusted, providing Bayesian meta-analysis the flexibility to succeed in many different scenarios. A limitation of our study is that we used minimally informative priors for all estimation in order to produce the most accurate estimates of effect size and heterogeneity. However, depending on the context, these priors could be changed. For example, in the case of finding biomarkers for diagnostic use, the prior for τ^2 could be changed from a uniform to a monotonically increasing function. This would in turn create a form of regularization, pushing τ^2 to be estimated as larger and increasing the effect of heterogeneity on Bayesian probability estimate. For subgroup identification and clustering of patients, one could similarly adjust priors and probability estimates to select genes that have high effect size and moderate-to-high within and/or between study heterogeneity. This shows the true flexibility and potential adaptability of the Bayesian framework for different uses.

Methods

Dataset selection

We used publicly available transcriptome data from the NCBI GEO for three diseases: (1) 223 samples across 4 datasets from healthy controls and patients with asthma (**S1 Table**) [10–12], (2) 1039 samples across 14 datasets from healthy controls and patients with cardiomyopathy (**S2 Table**) [13–23], and (3) 3069 samples across 27 datasets from healthy controls and patients with tuberculosis (TB) (**S3 Table**) [24–45]. Each dataset was appropriately normalized and log2 transformed, if not already in log scale. We removed genes that were not present in at least half of the datasets for a given disease.

Frequentist meta-analysis

We used the frequentist meta-analysis implemented in MetaIntegrator, which uses random effects inverse variant model, for comparison with the Bayesian method [50]. Briefly, MetaIntegrator computes a Hedge's *g* as an effect size for each gene in each dataset. The effect sizes are combined using random effects inverse variance model the DerSimonian-Laird method, and the corresponding p-value is estimated using a standard normal distribution, which is

corrected for multiple hypotheses testing using the Benjamini-Hochberg FDR adjustment [51]. Following these calculations, the top genes are selected by using FDR and effect size thresholds. We used difference between geometric mean of over-expressed genes to that of under-expressed genes as a classifier because it has been repeatedly shown to be more generalizable across datasets [52] and has also been translated in a point-of-care test [5,6,46,47].

Bayesian meta-analysis–dataset effect size calculation

The first step of the Bayesian meta-analysis pipeline involves creating an effect size distribution for each case and control for each gene in each dataset (**Fig 1**). We used the BEST [8] framework with default parameters and priors for this purpose. The BEST framework estimates the posterior distribution of effect size between cases and controls for each gene in each dataset by assuming that the data is independently distributed and comes from a *t* distribution with different mean (μ) and standard deviation (σ) parameters for each group. Then, we combine the distributions from independent studies using a gaussian hierarchical model, estimating both the pooled effect size and between-study heterogeneity in the process. Overall normality parameter (ν) that denotes the size of tails of the *t* distribution and the level of normality. Overall, the BEST framework estimates 5 parameters: μ_1, μ_2, σ_1, σ_2 and ν using minimally informative priors. μ_1 and μ_2 are the population means of cases and controls and are parameterized with a wide normal prior with a large standard deviation. σ_1 and σ_2 are the population standard deviations of cases and controls and are parameterized with a broad uniform. The normality parameter, ν, has a broad, shifted exponential prior. Following the parameter estimation, we calculated the effect size as a standardized mean difference:

$$\frac{(\mu_1 - \mu_2)}{\sqrt{(\sigma_1 + \sigma_2)/2}}$$

For this study, we ran all individual datasets with 2000 steps with 400 for model burn-in with 3 chains to ensure convergence, which is then calculated using $\hat{r}$ [53]. We removed any genes with an $\hat{r}$ greater than 1.1 from the dataset.

Bayesian meta-analysis–pooling step

Following the dataset effect size distribution estimation, a pooling step is performed to estimate an overall pooled effect size using a hierarchical model. The effect size distribution from each dataset for each gene is assumed to be normally distributed with mean μ_i and σ_i. To calculate the pooled effect size, we use each of the calculated dataset effect size distributions and assume that they are sampled from an overall, pooled normal distribution represented as $Normal(\mu_{pooled}, \tau^2)$. Both μ and τ are parameterized with minimally informative priors: μ_{pooled} as $Normal(0, 3)$ and τ as $Uniform(0, 2)$ and parameters are estimated using Gibbs sampling [54]. We chose priors for effect size and between-study heterogeneity using 122 previous gene expression meta-analyses (**S1A** and **S1B Fig**) [50]. Sensitivity analysis of the priors found that when we varied the parameters of the Normal priors for effect size, the Bayesian probabilities for differential expression remained concordant (**S1C** and **S1D Fig**). For between-study heterogeneity, we found that using Uniform(0,1) as prior had lower Bayesian probability estimates (**S1E Fig**), whereas using Uniform(0,2) or Uniform(0,3) as prior had highly concordant Bayesian probability estimates (**S1F Fig**). Hence, the final gene rankings and posterior probabilities did not change by widening the priors further.

For this study, all pooling steps were run with 5000 steps with 1000 for burn-in with 3 chains. The convergence parameter ($\hat{r}$) is calculated using the chains and any genes with an $\hat{r}$

greater than 1.1 are removed. Using the hierarchical model structure, once the pooled distribution is estimated, we adjusted the individual dataset effect size distributions based on the summary distribution, akin to a random effects model in frequentist meta-analysis. To calculate statistical significance, we calculate the probability that a certain gene is upregulated or downregulated, calculating $Pr(\mu_{pooled} < 0)$ or $Pr(\mu_{pooled} > 0)$, respectively. This is done with a standard cumulative density function for a normal distribution.

Simulated data for comparison of frequentist and Bayesian meta-analysis

Random study data was used to compare the Bayesian and frequentist methods. The frequentist random-effects meta-analysis was run using the metagen function from the meta package [55] in the R using default parameters. The Bayesian model was run as with the parameters described above, with 50000 steps with 10000 for burn-in.

Comparison of AUCs, effect sizes, and heterogeneity between frequentist and Bayesian meta-analysis

To compare the methods, we used the Tuberculosis and Cardiomyopathy cohorts, as defined above. We removed genes that were not present in at least half of the datasets. For both diseases, we randomly selected *N* datasets 100 times (or 91 times in the case of Cardiomyopathy: N = 2) and applied both meta-analysis approaches Bayesian (bayesMetaIntegrator) and frequentist (MetaIntegrator) to each set of randomly selected *N* datasets. For each iteration, we filtered to differentially expressed genes with absolute summary effect size > 0.6 and used the top 10 genes with the smallest FDR or Bayesian probability. We used difference of geometric means of over- and under-expressed genes as a classifier in unseen datasets to distinguish cases from healthy controls, as defined below. For the tau comparison, we report the Bayesian estimate of tau for each gene due to frequentist meta-analysis reporting large numbers of genes as 0. For pathway analysis, the ReactomePA package [48] was used to perform pathway analysis on all three diseases, using the most significant genes for both methods– 1000 genes for Cardiomyopathy and Tuberculosis and 500 genes for Asthma. A p-value cutoff of 0.2 was used for Cardiomyopathy, 0.2 for Asthma and 0.05 for Tuberculosis.

Supporting information

S1 Fig. A) Using 122 previous gene expression meta-analyses we observed the pooled effect size and tau to determine our initial priors. Our priors, picked to be minimally informative, are shown in red. B) Sensitivity analysis of effect size prior using Asthma data. C) Sensitivity analysis of tau prior using Asthma data.
(EPS)

S2 Fig. A) Using N = 4 datasets and top 10 statistically significant genes, we examined the effect of effect size thresholds using the AUC performance on all other left-out datasets. We find that the Bayesian model consistently outperforms the frequentist at effect sizes < 1. B) Using N = 4 datasets and effect size threshold of .6, we examined how the number of genes in a signature affect performance. We find that no matter how many genes were used in the signature, from 10–200, the Bayesian model consistently outperforms the frequentist.
(EPS)

S3 Fig. Jaccard similarity of the top N genes, by statistical significance, between the Bayesian and frequentist meta-analysis methods, which is generally low.
(EPS)

S4 Fig. Pathway analysis using Reactome PA—using the most significant genes for both methods– 1000 genes for Cardiomyopathy and Tuberculosis and 500 genes for Asthma. A p-value cutoff of .2 was used for Cardiomyopathy, .2 for Asthma and .05 for Tuberculosis. BMA_only denotes genes only significant in Bayesian, but not frequentist. FMA_only denotes genes only significant in frequentist, but not Bayesian.
(EPS)

S1 Table. Datasets used for meta-analysis of asthma.
(XLSX)

S2 Table. Datasets used for meta-analysis of cardiomyopathy.
(XLSX)

S3 Table. Datasets used for meta-analysis of tuberculosis.
(XLSX)

Acknowledgments

We would like to acknowledge Michele Donato and Ian Lee, for their contributions and work testing the method and package. We would like to thank the researchers that have contributed the datasets used within this study.

Author Contributions

Conceptualization: Laurynas Kalesinskas, Purvesh Khatri.

Data curation: Laurynas Kalesinskas, Sanjana Gupta.

Formal analysis: Laurynas Kalesinskas, Purvesh Khatri.

Funding acquisition: Purvesh Khatri.

Investigation: Laurynas Kalesinskas, Sanjana Gupta, Purvesh Khatri.

Methodology: Laurynas Kalesinskas.

Project administration: Purvesh Khatri.

Resources: Purvesh Khatri.

Software: Laurynas Kalesinskas, Sanjana Gupta.

Supervision: Purvesh Khatri.

Validation: Laurynas Kalesinskas.

Visualization: Laurynas Kalesinskas.

Writing – original draft: Laurynas Kalesinskas, Purvesh Khatri.

Writing – review & editing: Laurynas Kalesinskas, Purvesh Khatri.

References

1. Bishop D. Rein in the four horsemen of irreproducibility. Apr 2019: 435–435.
2. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nature Reviews Drug Discovery. 2011; 1–2. https://doi.org/10.1038/nrd3439-c1 PMID: 21892149
3. Ioannidis JPA. Why Most Published Research Findings Are False. PLoS Medicine. 2005; 2: e124. https://doi.org/10.1371/journal.pmed.0020124 PMID: 16060722

4. Design preclinical studies for reproducibility. Nat Biomed Eng. 2018; 2: 789–790. https://doi.org/10.1038/s41551-018-0322-y PMID: 31015618
5. Sutherland JS, Spuy G van der, Gindeh A, Thuong NT, Namuganga AR, Owolabi O, et al. Diagnostic accuracy of the Cepheid 3-gene host response fingerstick blood test in a prospective, multi-site study: interim results. Clin Infect Dis. 2021; ciab839–. https://doi.org/10.1093/cid/ciab839 PMID: 34550342
6. Sweeney TE, Braviak L, Tato CM, Khatri P. Genome-wide expression for diagnosis of pulmonary tuberculosis: a multicohort analysis. The Lancet Respiratory Medicine. 2016; 4: 213–224. https://doi.org/10.1016/S2213-2600(16)00048-5 PMID: 26907218
7. Sweeney TE, Haynes WA, Vallania F, Ioannidis JP, Khatri P. Methods to increase reproducibility in differential gene expression via meta-analysis. Nucleic Acids Res. 2017; 45: e1–e1. https://doi.org/10.1093/nar/gkw797 PMID: 27634930
8. Kruschke JK. Bayesian Estimation Supersedes the t Test. J Exp Psychology Gen. 2013; 142: 573–603. https://doi.org/10.1037/a0029146 PMID: 22774788
9. Gelman A, Hill J, Yajima M. Why We (Usually) Don't Have to Worry About Multiple Comparisons. J Res Educ Eff. 2012; 5: 189–211. https://doi.org/10.1080/19345747.2011.618213
10. Kicic A, Hallstrand TS, Sutanto EN, Stevens PT, Kobor MS, Taplin C, et al. Decreased Fibronectin Production Significantly Contributes to Dysregulated Repair of Asthmatic Epithelium. Am J Resp Crit Care. 2010; 181: 889–898. https://doi.org/10.1164/rccm.200907-1071OC PMID: 20110557
11. Singhania A, Rupani H, Jayasekera N, Lumb S, Hales P, Gozzard N, et al. Altered Epithelial Gene Expression in Peripheral Airways of Severe Asthma. Plos One. 2017; 12: e0168680. https://doi.org/10.1371/journal.pone.0168680 PMID: 28045928
12. Singhania A, Wallington JC, Smith CG, Horowitz D, Staples KJ, Howarth PH, et al. Multitissue Transcriptomics Delineates the Diversity of Airway T Cell Functions in Asthma. Am J Resp Cell Mol. 2018; 58: 261–270. https://doi.org/10.1165/rcmb.2017-0162OC PMID: 28933920
13. Ameling S, Herda LR, Hammer E, Steil L, Teumer A, Trimpert C, et al. Myocardial gene expression profiles and cardiodepressant autoantibodies predict response of patients with dilated cardiomyopathy to immunoadsorption therapy. Eur Heart J. 2013; 34: 666–675. https://doi.org/10.1093/eurheartj/ehs330 PMID: 23100283
14. Gaertner A, Schwientek P, Ellinghaus P, Summer H, Golz S, Kassner A, et al. Myocardial transcriptome analysis of human arrhythmogenic right ventricular cardiomyopathy. Physiol Genomics. 2012; 44: 99–109. https://doi.org/10.1152/physiolgenomics.00094.2011 PMID: 22085907
15. Liu Y, Morley M, Brandimarto J, Hannenhalli S, Hu Y, Ashley EA, et al. RNA-Seq identifies novel myocardial gene expression signatures of heart failure. Genomics. 2015; 105: 83–89. https://doi.org/10.1016/j.ygeno.2014.12.002 PMID: 25528681
16. Hannenhalli S, Putt ME, Gilmore JM, Wang J, Parmacek MS, Epstein JA, et al. Transcriptional Genomics Associates FOX Transcription Factors With Human Heart Failure. Circulation. 2006; 114: 1269–1276. https://doi.org/10.1161/CIRCULATIONAHA.106.632430 PMID: 16952980
17. Akat KM, Moore-McGriff D, Morozov P, Brown M, Gogakos T, Rosa JCD, et al. Comparative RNA-sequencing analysis of myocardial and circulating small RNAs in human heart failure and their utility as biomarkers. Proc National Acad Sci. 2014; 111: 11151–11156. https://doi.org/10.1073/pnas.1401724111 PMID: 25012294
18. Koczor CA, Lee EK, Torres RA, Boyd A, Vega JD, Uppal K, et al. Detection of differentially methylated gene promoters in failing and nonfailing human left ventricle myocardium using computation analysis. Physiol Genomics. 2013; 45: 597–605. https://doi.org/10.1152/physiolgenomics.00013.2013 PMID: 23695888
19. Molina-Navarro MM, Roselló-Lletí E, Ortega A, Tarazón E, Otero M, Martínez-Dolz L, et al. Differential Gene Expression of Cardiac Ion Channels in Human Dilated Cardiomyopathy. Plos One. 2013; 8: e79792. https://doi.org/10.1371/journal.pone.0079792 PMID: 24339868
20. Kittleson MM, Minhas KM, Irizarry RA, Ye SQ, Edness G, Breton E, et al. Gene expression analysis of ischemic and nonischemic cardiomyopathy: shared and distinct genes in the development of heart failure. Physiol Genomics. 2005; 21: 299–307. https://doi.org/10.1152/physiolgenomics.00255.2004 PMID: 15769906
21. Wittchen F, Suckau L, Witt H, Skurk C, Lassner D, Fechner H, et al. Genomic expression profiling of human inflammatory cardiomyopathy (DCMi) suggests novel therapeutic targets. J Mol Medicine Berlin Ger. 2007; 85: 257–271. https://doi.org/10.1007/s00109-006-0122-9 PMID: 17106732
22. Schwientek P, Ellinghaus P, Steppan S, D'Urso D, Seewald M, Kassner A, et al. Global gene expression analysis in nonfailing and failing myocardium pre- and postpulsatile and nonpulsatile ventricular assist device support. Physiol Genomics. 2010; 42: 397–405. https://doi.org/10.1152/physiolgenomics.00030.2010 PMID: 20460602

23. Barth AS, Kuner R, Buness A, Ruschhaupt M, Merk S, Zwermann L, et al. Identification of a Common Gene Expression Signature in Dilated Cardiomyopathy Across Independent Microarray Studies. J Am Coll Cardiol. 2006; 48: 1610–1617. https://doi.org/10.1016/j.jacc.2006.07.026 PMID: 17045896

24. Verhagen LM, Zomer A, Maes M, Villalba JA, Nogal BD, Eleveld M, et al. A predictive signature gene set for discriminating active from latent tuberculosis in Warao Amerindian children. BMC Genomics. 2013; 14: 74. https://doi.org/10.1186/1471-2164-14-74 PMID: 23375113

25. Berry MPR, Graham CM, McNab FW, Xu Z, Bloch SAA, Oni T, et al. An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. Nature. 2010; 466: 973–977. https://doi.org/10.1038/nature09247 PMID: 20725040

26. Walter ND, Miller MA, Vasquez J, Weiner M, Chapman A, Engle M, et al. Blood Transcriptional Biomarkers for Active Tuberculosis among Patients in the United States: a Case-Control Study with Systematic Cross-Classifier Evaluation. Land GA, editor. Journal Of Clinical Microbiology. 2016; 54: 274–282. https://doi.org/10.1128/JCM.01990-15 PMID: 26582831

27. Maertzdorf J, Weiner J, Mollenkopf H-J, Network Tb, Bauer T, Prasse A, et al. Common patterns and disease-related signatures in tuberculosis and sarcoidosis. Proc Natl Acad Sci U S A. 2012; 109: 7853–7858. https://doi.org/10.1073/pnas.1121072109 PMID: 22547807

28. Maertzdorf J, McEwen G, Weiner J, Tian S, Lader E, Schriek U, et al. Concise gene signature for point-of-care classification of tuberculosis. EMBO Molecular Medicine. 2016; 8: 86–95. https://doi.org/10.15252/emmm.201505790 PMID: 26682570

29. Bloom CI, Graham CM, Berry MPR, Wilkinson KA, Oni T, Rozakeas F, et al. Detectable changes in the blood transcriptome are present after two weeks of antituberculosis therapy. PLoS ONE. 2012; 7: e46191. https://doi.org/10.1371/journal.pone.0046191 PMID: 23056259

30. Kaforou M, Wright VJ, Oni T, French N, Anderson ST, Bangani N, et al. Detection of Tuberculosis in HIV-Infected and -Uninfected African Adults Using Whole Blood RNA Expression Signatures: A Case-Control Study. Cattamanchi A, editor. PLoS Medicine. 2013; 10: e1001538. https://doi.org/10.1371/journal.pmed.1001538 PMID: 24167453

31. Anderson ST, Kaforou M, Brent AJ, Wright VJ, Banwell CM, Chagaluka G, et al. Diagnosis of childhood tuberculosis and host RNA expression in Africa. N Engl J Med. 2014; 370: 1712–1723. https://doi.org/10.1056/NEJMoa1303657 PMID: 24785206

32. Tientcheu LD, Maertzdorf J, Weiner J, Adetifa IM, Mollenkopf H-J, Sutherland JS, et al. Differential transcriptomic and metabolic profiles of M. africanum- and M. tuberculosis-infected patients after, but not before, drug treatment. Genes and Immunity. 2015; 16: 347–355. https://doi.org/10.1038/gene.2015.21 PMID: 26043170

33. Cliff JM, Lee J-S, Constantinou N, Cho J-E, Clark TG, Ronacher K, et al. Distinct phases of blood gene expression pattern through tuberculosis treatment reflect modulation of the humoral immune response. The Journal of infectious diseases. 2013; 207: 18–29. https://doi.org/10.1093/infdis/jis499 PMID: 22872737

34. Leong S, Zhao Y, Ribeiro-Rodrigues R, Jones-López EC, Acuña-Villaorduña C, Rodrigues PM, et al. Cross-validation of existing signatures and derivation of a novel 29-gene transcriptomic signature predictive of progression to TB in a Brazilian cohort of household contacts of pulmonary TB. Tuberculosis. 2020; 120: 101898. https://doi.org/10.1016/j.tube.2020.101898 PMID: 32090859

35. Maertzdorf J, Ota M, Repsilber D, Mollenkopf HJ, Weiner J, Hill PC, et al. Functional correlations of pathogenesis-driven gene expression signatures in tuberculosis. PLoS ONE. 2011; 6: e26938. https://doi.org/10.1371/journal.pone.0026938 PMID: 22046420

36. Lee S-W, Wu LS-H, Huang G-M, Huang K-Y, Lee T-Y, Weng JT-Y. Gene expression profiling identifies candidate biomarkers for active and latent tuberculosis. Bmc Bioinformatics. 2016; 17: S3. https://doi.org/10.1186/s12859-015-0848-x PMID: 26818387

37. Ottenhoff THM, Dass RH, Yang N, Zhang MM, Wong HEE, Sahiratmadja E, et al. Genome-wide expression profiling identifies type 1 interferon response pathways in active tuberculosis. PLoS ONE. 2012; 7: e45839. https://doi.org/10.1371/journal.pone.0045839 PMID: 23029268

38. Banchereau R, Jordan-Villegas A, Ardura M, Mejias A, Baldwin N, Xu H, et al. Host Immune Transcriptional Profiles Reflect the Variability in Clinical Disease Manifestations in Patients with Staphylococcus aureus Infections. Plos One. 2012; 7: e34390. https://doi.org/10.1371/journal.pone.0034390 PMID: 22496797

39. Cai Y, Yang Q, Tang Y, Zhang M, Liu H, Zhang G, et al. Increased complement C1q level marks active disease in human tuberculosis. PLoS ONE. 2014; 9: e92340. https://doi.org/10.1371/journal.pone.0092340 PMID: 24647646

40. Marais S, Lai RPJ, Wilkinson KA, Meintjes G, O'Garra A, Wilkinson RJ. Inflammasome Activation Underlying Central Nervous System Deterioration in HIV-Associated Tuberculosis. J Infect Dis. 2017; 215: 677–686. https://doi.org/10.1093/infdis/jiw561 PMID: 27932622

41. Hu X, Liao S, Bai H, Gupta S, Zhou Y, Zhou J, et al. Long Noncoding RNA and Predictive Model To Improve Diagnosis of Clinically Diagnosed Pulmonary Tuberculosis. J Clin Microbiol. 2020; 58: e01973–19. https://doi.org/10.1128/JCM.01973-19 PMID: 32295893

42. Blankley S, Graham CM, Turner J, Berry MPR, Bloom CI, Xu Z, et al. The Transcriptional Signature of Active Tuberculosis Reflects Symptom Status in Extra-Pulmonary and Pulmonary Tuberculosis. Plos One. 2016; 11: e0162220. https://doi.org/10.1371/journal.pone.0162220 PMID: 27706152

43. Bloom CI, Graham CM, Berry MPR, Rozakeas F, Redford PS, Wang Y, et al. Transcriptional Blood Signatures Distinguish Pulmonary Tuberculosis, Pulmonary Sarcoidosis, Pneumonias and Lung Cancers. Plos One. 2013; 8: e70630. https://doi.org/10.1371/journal.pone.0070630 PMID: 23940611

44. Araujo LS de, Vaas LAI, Ribeiro-Alves M, Geffers R, Mello FCQ, Almeida AS de, et al. Transcriptomic Biomarkers for Tuberculosis: Evaluation of DOCK9. EPHA4, and NPC2 mRNA Expression in Peripheral Blood. Front Microbiol. 2016; 7: 1586. https://doi.org/10.3389/fmicb.2016.01586 PMID: 27826286

45. Esmail H, Lai RP, Lesosky M, Wilkinson KA, Graham CM, Horswell S, et al. Complement pathway gene activation and rising circulating immune complexes characterize early disease in HIV-associated tuberculosis. Proc National Acad Sci. 2018; 115: E964–E973. https://doi.org/10.1073/pnas.1711853115 PMID: 29339504

46. Södersten E, Ongarello S, Mantsoki A, Wyss R, Persing DH, Banderby S, et al. Diagnostic Accuracy Study of a Novel Blood-Based Assay for Identification of Tuberculosis in People Living with HIV. J Clin Microbiol. 2020; 59. https://doi.org/10.1128/jcm.01643-20 PMID: 33298607

47. Moreira FMF, Verma R, Santos PCP dos, Leite A, Santos A da S, Araujo RCP de, et al. Blood-based host biomarker diagnostics in active case finding for pulmonary tuberculosis: A diagnostic case-control study. Eclinicalmedicine. 2021; 100776. https://doi.org/10.1016/j.eclinm.2021.100776 PMID: 33842866

48. Fabregat A, Sidiropoulos K, Viteri G, Forner O, Marin-Garcia P, Arnau V, et al. Reactome pathway analysis: a high-performance in-memory approach. Bmc Bioinformatics. 2017; 18: 142. https://doi.org/10.1186/s12859-017-1559-2 PMID: 28249561

49. Kruschke JK, Liddell TM. The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. Psychon B Rev. 2018; 25: 178–206. https://doi.org/10.3758/s13423-016-1221-4 PMID: 28176294

50. Haynes WA, Vallania F, Liu C, Bongen E, Tomczak A, Andres-Terre M, et al. Empowering Multi-Cohort Gene Expression Analysis to Increase Reproducibility. Pacific Symposium on Biocomputing. 2016; 22: 144–153. https://doi.org/10.1101/071514

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

52. Mayhew MB, Buturovic L, Luethy R, Midic U, Moore AR, Roque JA, et al. A generalizable 29-mRNA neural-network classifier for acute bacterial and viral infections. Nature communications. 2020; 1–10. https://doi.org/10.1038/s41467-020-14975-w

53. Gelman A, Rubin DB. Inference from Iterative Simulation Using Multiple Sequences. Stat Sci. 1992; 7. https://doi.org/10.1214/ss/1177011136

54. Casella G, George EI. Explaining the Gibbs Sampler. Am Statistician. 2012; 46: 167–174. https://doi.org/10.1080/00031305.1992.10475878

55. Schwarzer G, Carpenter JR, Rücker G. Meta-Analysis with R. R. 2015; 217–236. https://doi.org/10.1007/978-3-319-21416-0_9