Article

Bayesian Statistics Improves Biological Interpretability of Metabolomics Data from Human Cohorts

Christopher Brydges [1], Xiaoyu Che [2,3], Walter Ian Lipkin [2,4] and Oliver Fiehn [1,*]

[1]   West Coast Metabolomics Center, UC Davis, Davis, CA 95616, USA; christopherbrydges@gmail.com
[2]   Center for Infection and Immunity, Mailman School of Public Health of Columbia University,
      New York, NY 10032, USA; xc2273@cumc.columbia.edu (X.C.); wil2001@cumc.columbia.edu (W.I.L.)
[3]   Department of Biostatistics, Mailman School of Public Health of Columbia University,
      New York, NY 10032, USA
[4]   Vagelos College of Physicians and Surgeons of Columbia University, New York, NY 10032, USA
[*]   Correspondence: ofiehn@ucdavis.edu

Abstract: Univariate analyses of metabolomics data currently follow a frequentist approach, using *p*-values to reject a null hypothesis. We here propose the use of Bayesian statistics to quantify evidence supporting different hypotheses and discriminate between the null hypothesis versus the lack of statistical power. We used metabolomics data from three independent human cohorts that studied the plasma signatures of subjects with myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS). The data are publicly available, covering 84–197 subjects in each study with 562–888 identified metabolites of which 777 were common between the two studies and 93 were compounds reported in all three studies. We show how Bayesian statistics incorporates results from one study as "prior information" into the next study, thereby improving the overall assessment of the likelihood of finding specific differences between plasma metabolite levels. Using classic statistics and Benjamini–Hochberg FDR-corrections, Study 1 detected 18 metabolic differences and Study 2 detected no differences. Using Bayesian statistics on the same data, we found a high likelihood that 97 compounds were altered in concentration in Study 2, after using the results of Study 1 as the prior distributions. These findings included lower levels of peroxisome-produced ether-lipids, higher levels of long-chain unsaturated triacylglycerides, and the presence of exposome compounds that are explained by the difference in diet and medication between healthy subjects and ME/CFS patients. Although Study 3 reported only 92 compounds in common with the other two studies, these major differences were confirmed. We also found that prostaglandin F2alpha, a lipid mediator of physiological relevance, was reduced in ME/CFS patients across all three studies. The use of Bayesian statistics led to biological conclusions from metabolomic data that were not found through frequentist approaches. We propose that Bayesian statistics is highly useful for studies with similar research designs if similar metabolomic assays are used.

Keywords: Bayesian; *p*-value; metabolomics; ME/CFS; statistics

1. Introduction

Although the literature detailing the limitations and misconceptions of *p*-values has established that the way that many researchers employ statistics in their research is ritualistic and inappropriate [1,2] and impedes scientific progress [3,4], metabolomic researchers rarely acknowledge the pitfalls of the methods used for the assessment of statistical significance. Even the American Statistical Association has warned about misusing the *p*-value [5]. Cut-offs for significance testing can be easily subject to changing perspectives [6–8].

This discussion has not become generally accepted in the realms of omics sciences, including in metabolomics. While it is common for univariate analyses to be conducted with false discovery rate or familywise error rate corrections, researchers still only consider metabolites to be interesting if their *p*-values fall below an arbitrary threshold. However,

p-values only give the probability that the measured data fit an assumed null hypothesis. If a *p*-value is <0.05, the null hypothesis is commonly rejected. However, even if $p > 0.05$, a metabolite might still be different between the two study groups. Simply put, strict "null hypothesis significance testing" cannot distinguish whether there is a true null effect or whether the data are insensitive [9]. More importantly, a *p*-value cannot provide support for an alternative hypothesis. A classic *p*-value reports the probability of the data given the hypothesis—not the probability of the hypothesis, given the data [5]. These are not identical statements (i.e., p(data | H) $\neq$ p(H | data)) [10], and they do not answer the same questions. Strangely, a *p*-value does not provide a measure of the strength of evidence in favor of a hypothesis. One cannot rank *p*-values by being "more significant". For example, a *p*-value of 0.0001 is not stronger evidence in favor of the alternative hypothesis than a *p*-value of 0.049. Given a threshold of 0.05, they both reject the null hypothesis, but neither of them tests an alternative hypothesis. Furthermore, although reporting actual effect sizes (the differences in metabolite levels) is crucially important, a *p*-value does not give that information.

We here present Bayesian alternatives to complement or replace *p*-values and the analyses that produce them that we will hereafter be referred to as null hypothesis significance testing (NHST). We exemplify the power of Bayesian analyses on previously published metabolomics data to highlight the differences between the two statistical approaches. However, how is Bayesian statistics different? Unlike classic *p*-values, Bayesian statistics can be used to quantify the size of an effect, quantify the strength of evidence in favor of one hypothesis over another, and allow researchers to discriminate between an inconclusive finding and evidence in favor of the null hypothesis. We argue that this is exactly what researchers want to obtain. These analyses test competing models of hypotheses and the distribution of the observed data [11,12]. Bayesian statistics always starts with a prior distribution. This is an expectation of a range of possible effect sizes that could feasibly be observed. Put simply, this means that researchers should (or at least can) have some idea of the plausible size and/or direction of an effect that they are studying. For example, a statement such as "I think that compound X is most likely to be upregulated in cases compared to controls, with a fold-change of about 3" is an example of a prior distribution: this description can be modeled very easily so that the probability of downregulation is zero, and the most likely effect sizes are around a fold change = 3.

Bayesian statistics are used across a vast array of fields and domains [13], including everyday cognition and decision-making [14]. An example is hurricane forecasts. Figure 1 shows a fictional hurricane that may have formed in the Gulf of Mexico (top-left panel), with a cone of different colors displaying the path the hurricane is most likely to take. Given previous hurricanes and other data such as ocean currents, storm trackers already have an expectation, a "prior distribution". This is like a metabolomicist knowing the literature and having ideas about how the levels of metabolites might vary in diabetes mellitus. The colors of the cone correspond to the modeled prior distribution (bottom-left panel). As the hurricane progresses (Figure 1, top-right panel) and researchers incorporate the new data into an updated model, their expectations change about where the hurricane is most likely to go. The cone also narrows as the researchers become more confident about the path of the hurricane. This results in a narrower distribution (bottom-right panel). Hence, unlike classic *p*-values, Bayesian models take prior data into account and can be updated continuously.
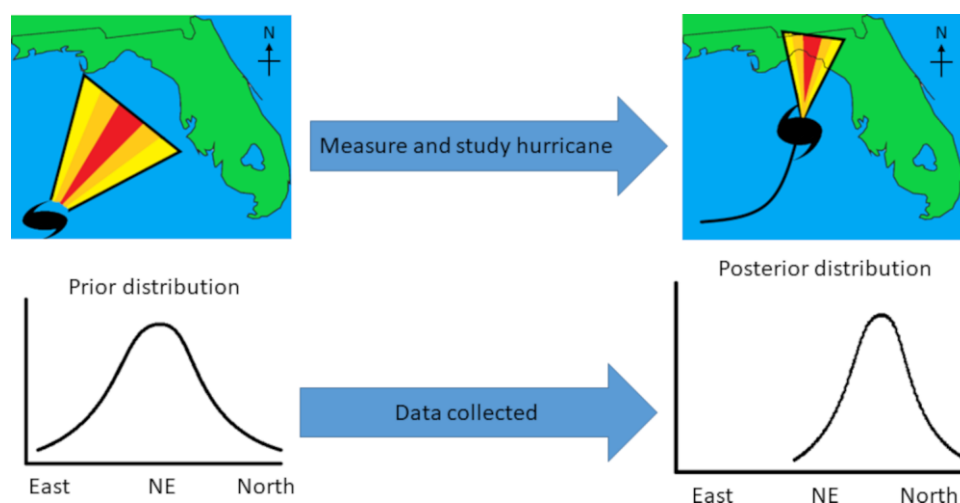
Figure 1. Hypothetical hurricane paths before (**top-left**) and after data have been collected (**top-right**). The cones represent the likely path of the hurricane (yellow = unlikely; red = most likely). Distributions (**bottom-left** and **bottom-right**) represent the cones plotted as probability distributions.

2. Materials and Methods

2.1. Metabolomics Datasets

Three datasets were reanalyzed for the current study. All three studies investigated metabolic differences between patients who suffer from myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) and matched healthy controls. In order to be consistent in our analyses, we only used the identified metabolites and ignored unknown features. The first data set by Nagy-Szakal et al. [15] is available at the Metabolomics Workbench repository [16] under Project ID PR000576 (DOI: http://doi.org/10.21228/M86X1F (accessed on 29 August 2023)). The second dataset by Che et al. [17] is available under Project DOI: http://doi.org/10.21228/M8PD9N (accessed on 29 August 2023), covering 106 cases and 91 control subjects and reporting 888 metabolites. In both studies, participant metadata (age, sex, body mass index (BMI), race/ethnicity, diagnosis of irritable bowel syndrome (IBS), geographic/clinical site, and season of sample) was also collected. Twenty subjects (ten ME/CFS patients, ten healthy controls) participated in both studies. For the comparison of the strengths of the Bayesian analyses, we also used a previously published dataset by Naviaux et al. [18] (Project DOI http://doi.org/10.21228/M82K58) with 45 cases and 39 controls (accessed on 29 August 2023). Naviaux et al. used a different metabolomic assay with 612 identified metabolites, for which we found only 92 compounds in common with the other two studies. For this study, participant metadata as given above were not available apart from disease status and sex.

2.2. Statistical Analyses

All analyses were conducted in R 4.1.2 (https://www.r-project.org/, accessed on 29 August 2023). Prior to the analyses, any compound not observed in at least 50% of samples was removed from the analyses. Any missing data were imputed with half-minimum values. In metabolomics, missing values are usually found when there is no signal to be detected, meaning a metabolite is below the limit of detection. Hence, we can exclude that missing data were "missing at random". In these (publicly available) datasets, the maximum proportion of missing data on any compound was 21%. Each compound was log-transformed for normality and then auto-scaled. Only compounds common to both the *Nagy-Szakal* and *Che* datasets were used in the analyses, which resulted in 632 compounds (551 identified) being analyzed. Common compounds were matched using International Chemical Identifier keys. For the third *Naviaux* dataset, 92 compounds were matched to the other two datasets, using RefMet annotations [19]. Classic univariate statistical analyses were performed by linear regression using base R functions, and Bayesian regression was

conducted using the rstanarm [20] and bayestestR [21] packages. We only used the posterior distribution for the metabolite (βp) to keep the demonstration simple and because the data for the Naviaux et al. study were collected and processed in a different metabolomics laboratory than the Nagy-Szakal et al. and Che et al. studies (hence, error terms and covariate effects may differ). However, modeling these parameters and passing the posterior distributions into the subsequent models would be justifiable in many cases. Additionally, Supplementary Table S1 provides the Bayes factors, credible interval medians, and 95% bounds to detail how the theorized βp-values agreed or differed across studies. These analyses were used to determine between-groups differences for each compound and were used instead of *t*-tests or Mann–Whitney *U*-tests due to the inclusion of covariates and having a sufficiently large sample size. All models included age, sex, BMI, race/ethnicity, IBS diagnosis, geographic/clinical site, and season of sample as covariates. The default prior distributions recommended by rstanarm were used to model the expected effect sizes for each compound in the *Nagy-Szakal* dataset. These defaults are considered to be "weakly informative" in that they provide some information on the expected magnitude of the effect based on the scales of the variables. However, they do not strongly affect the posterior distribution and help stabilize the computation, while still allowing for extreme effect sizes if warranted by the data [22,23]. Posterior distributions were created from four Markov Chain Monte Carlo chains of 2000 iterations each, with the first 1000 iterations in each chain used as burn-ins. The posterior distribution of each compound from the *Nagy-Szakal* data was then used as the prior distribution for the same compound in the *Che* dataset. Compounds were considered to be altered if the 95% credible interval did not overlap with zero. A negative posterior median was indicative of downregulation in the ME/CFS group, and a positive posterior median was indicative of upregulation. ChemRICH [24] was performed for set enrichment statistics, with the posterior median used as an estimate of the effect size, and the probability of direction was used as a Bayesian analog of the *p*-value [25].

3. Results

3.1. Classic Univariate Statistics Analyses

Figure 2 shows the results of the classic *p*-value (frequentist) analyses for the *Nagy-Szakal* and *Che* datasets. Of the 632 compounds common to both studies, only 18 were significantly different (FDR < 0.10) in the *Nagy-Szakal* data for which phosphatidylcholines (28:0, 30:0, 32:1, 32:2, 33:0, 34:1, 34:3, 34:4, 38:2, 38:6) were found to be downregulated and triacylglycerides (52:4, 54:6, 54:7, 56:5, 56:8) upregulated, in addition to lower levels of carnitine and tyrosine in the ME/CFS group. Interestingly, when analyzing the *Che* dataset by itself, not a single metabolite was found to be significantly different between groups, using regression analyses (Figure 2).

3.2. Bayesian Analyses

To demonstrate the advantages of Bayesian analyses over classic univariate analyses, we first analyzed the *Nagy-Szakal* data using weakly informative prior distributions. The resulting posterior distribution was then used as the input into the Bayesian analyses of the *Che* data. The Bayesian model comparison commonly refers to the calculation of Bayes factors (BFs). Bayes factors are ratios that quantify the probability of one hypothesis over another by estimating the strength of evidence [26]. Such tests do not determine if one hypothesis is true and the other is not, but instead, whether one hypothesis is more likely than an alternative hypothesis, given the observed data. Additionally, BF values are easily interpretable in terms of the strength of a finding. Bayesian analyses allow statements such as "the null hypothesis was five times more likely to be true than the alternative hypothesis, given the data". Classic *p*-values cannot support such statements. BFs are continuous estimates that range from approaching zero to approaching infinity. A BF = 1 indicates equal likelihoods of either hypothesis, given the data. Values further from 1 imply stronger evidence in favor of one hypothesis over the other. Jeffreys [27] provided arbitrary

guidelines to categorize these values as anecdotal, moderate, strong, or extreme evidence for or against one hypothesis over another. Based on these guidelines, BFs between $1/3$ and 3 are referred to as anecdotal evidence and imply that the data are not sensitive enough to conclusively state that one hypothesis is more likely than another. Studies with BFs between $1/3$ and 3 are typically underpowered, and more data would need to be collected. However, any general rule used to categorize BFs will not be appropriate for all research contexts. Extraordinary claims require extraordinary evidence: if geographic location or birth date would have been found to be associated with ME/CFS, one would certainly require more evidence than usual for such a claim.
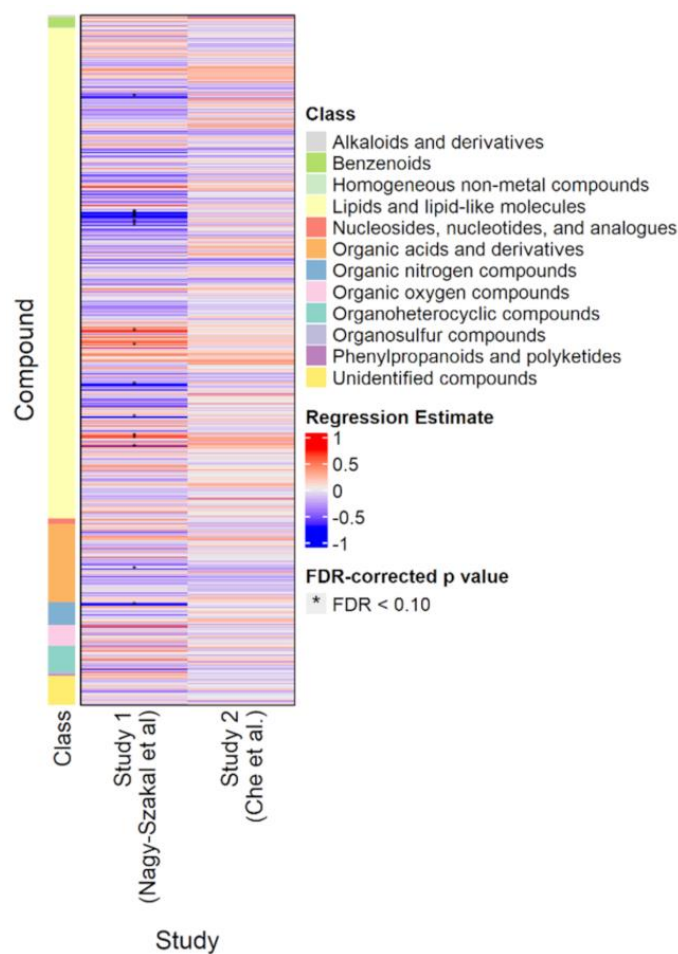

Figure 2. Heatmap of compounds analyzed in the *Nagy-Szakal* and *Che* datasets. Blue indicates downregulation; red indicates upregulation. Asterisks indicate FDR-adjusted *p*-values < 0.10.

Compounds with BFs > 10 closely match those that were reported as significantly different in the classic univariate analyses, for the simple reason that both methods tested whether metabolite levels were different between the case and control groups. For example, phosphatidylcholines (PCs) were still found to be downregulated by Bayesian statistics, and triacylglycerides (TGs) 54:7 were still upregulated in ME/CFS patients (Figure 3). For compounds with a BF > 3, see Table S1. However, unlike the classic statistics, Bayes' models also determine which compounds are sufficiently unlikely to be altered by ME/CFS (those with a BF < 1/3) and those that may be affected, but do not have sufficient statistical power to draw a conclusion (BFs between 1/3 and 3).
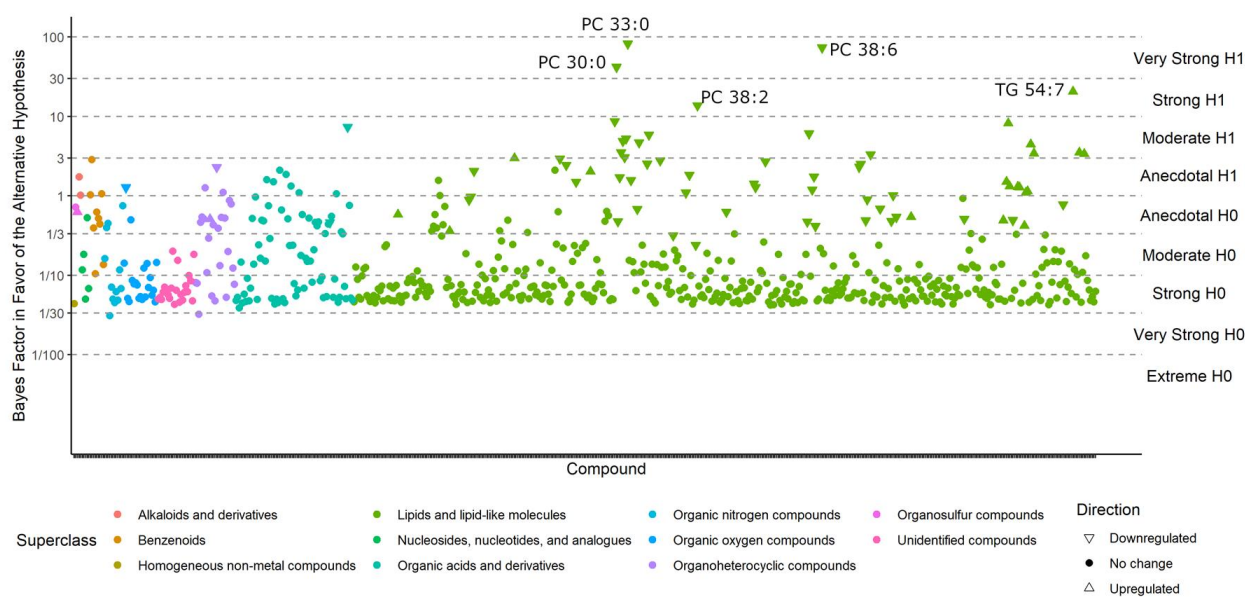
Figure 3. Bayes factors of all compounds in the *Nagy-Szakal* dataset. Compounds with a BF > 10 (i.e., strong evidence in favor of the alternative hypothesis) are labeled.

In addition, Bayes' analyses can be used to estimate the magnitude of an effect of interest. It is based on the posterior distribution, the updated prior distribution after the data have been collected and incorporated into the model. The posterior distribution provides an estimate of the credibility of every possible regression coefficient between ME/CFS diagnosis and metabolite levels after accounting for the researcher's hypotheses before data collection and after observing the current data. The mode of the distribution is the most-likely estimate of the true regression coefficient. The variability of a Bayesian effect size estimate is based on "credible intervals". Credible intervals are a range of values within which the true effect falls within specified bounds [27]. When a 95% credible interval does not overlap with zero, the probability of an effect being zero is <5%. Conversely, if the credible interval overlaps with zero, it could be considered as a Bayesian analog of a significance test [25]. Notably, the similar-sounding "confidence intervals" are different because those values would provide an interval within which the true effect size would fall at 95% confidence if the same study were repeated 100 times [3].

Figure 4 shows the compounds that have 95% credible intervals that do not overlap with zero and have a BF > 3. Again, phosphatidylcholines were found to be downregulated along with tyrosine and one phosphatidylethanolamine (PE) ether-lipid, PE (p-38:6).

A range of triacylglycerides were found to be upregulated in ME/CFS cases, as also observed with classic univariate statistics. The most-interesting difference in using Bayesian statistics was found when using the *Che* dataset with a prior distribution for each compound taken from the results of the *Nagy-Szakal* dataset as given in the Materials and Methods Section, while the Bayes factors, credible interval medians, and 95% bounds are given in Supplementary Table S1. Figure 5 shows the BFs of each compound, and Figure 6 shows the compounds that have 95% credible intervals that do not overlap with zero. In contrast with the classic univariate *p*-values (Figure 2), 98 compounds were found to be altered in the Bayesian analyses. Importantly, the results were very consistent with those observed from the *Nagy-Szakal* dataset: specific phosphatidylcholines were still found to be downregulated, whereas specific triacylglycerides upregulated. Yet, a range of other compounds were now found to be differentially regulated as well: with Bayesian statistics informed by prior research (here, the *Nagy-Szakal* data), the *Che* data now showed that the branched chain amino acid leucine and aromatic amino acids tyrosine and phenylalanine were downregulated in ME/CFS cases. Additional lipid species were now also found to be downregulated such as specific lysophosphatidylcholines, phosphatidylcholines,

and plasmalogens. Notably, specific diacylglycerides were found at higher plasma levels in ME/CFS patients, as well as specific pharmaceutical drugs such as gabapentin and p-acetamidophenol (acetaminophen), both used as pain medications, in addition to pantothenic acid (vitamin B5), a vitamin often taken as a dietary supplement. A range of food compounds was found at lower levels in ME/CFS patients indicating less use of specific foods, such as caffeine, theobromine, and trigonelline (coffee biomarkers) and piperine (found in pepper).
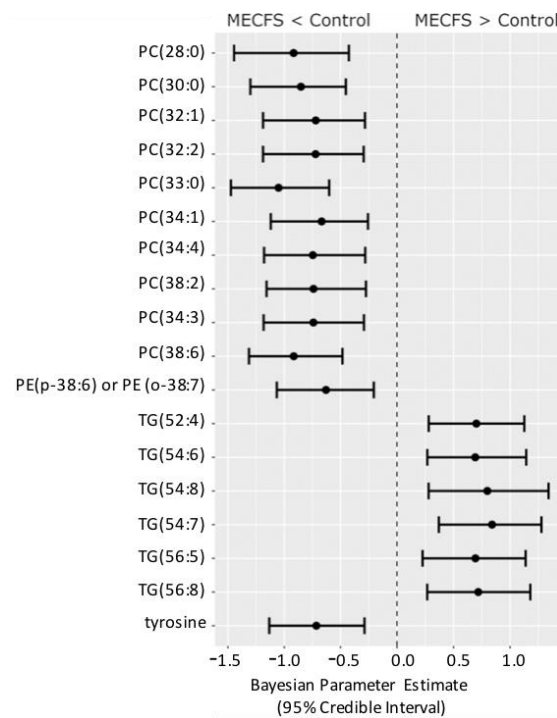
Figure 4. Forest plot of compounds with Bayesian 95% credible intervals not overlapping with zero and a BF > 3 in the *Nagy-Szakal* dataset. Points in each bar represent the posterior distribution median, and bars represent 95% credible interval bounds.
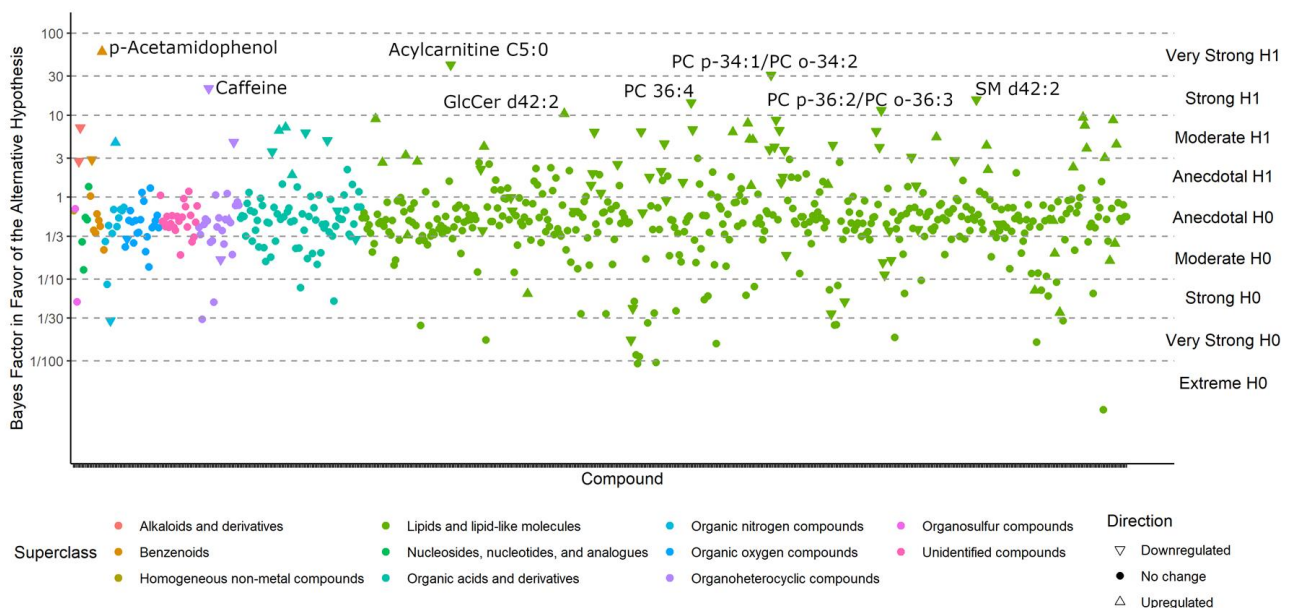

Figure 5. Bayes factors of all compounds in the *Che* dataset. Compounds with BF > 10 (i.e., strong evidence in favor of the alternative hypothesis) are labeled.
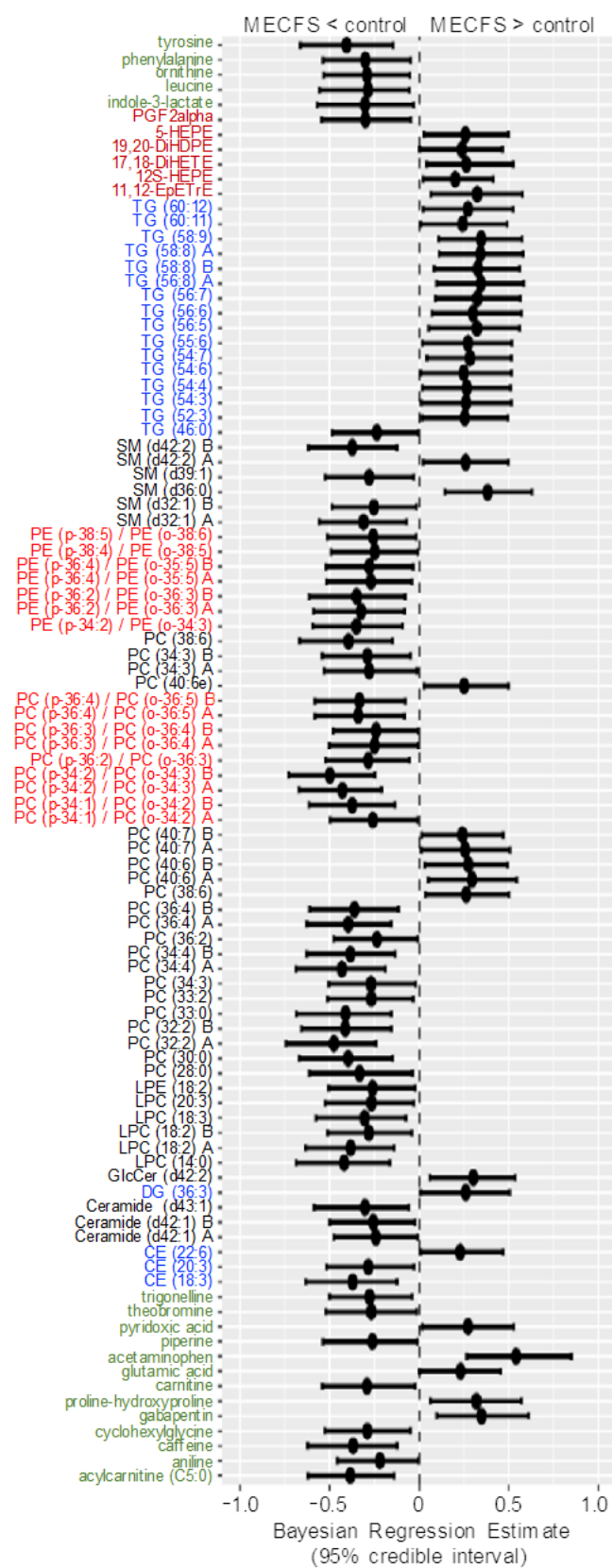
Figure 6. Forest plot of compounds with Bayesian 95% credible intervals not overlapping with zero in the *Che* dataset. Points in each bar represent the posterior distribution median, and bars represent 95% credible interval bounds. Green: hydrophilic compounds. Blue: neutral lipids. Black: polar lipids. Bright red: polar ether lipids. Dark red: oxylipins.

There is one more type of question that metabolomics investigators, and their clinical partners, would ask: How are these findings connected? One way to answer such questions is to map metabolites to their biochemical pathways of synthesis and degradation. However, metabolite levels in human blood do not only correspond to pathways in cells, but surely also to the differences between organs and, of course, dietary patterns and exposures. We, therefore, tested for the significance that specific groups of compounds were found to be over-enriched, beyond what would be expected at random. For this type of set enrichment statistics [28,29], we grouped the metabolites by similarity in chemical structure, using the Kolmogorov–Smirnov test for significance. Using the results from Bayesian probabilities as the input into chemical enrichment statistics [24], we found very strong evidence of very specific differential regulation of whole groups of compounds (Figure 7). For example, unsaturated ceramides and unsaturated lysophosphatidylcholines were found to be downregulated, but not their saturated counterparts. Sphingomyelins and unsaturated phosphatidycholines were significantly associated with ME/CFS, but among these compound classes, some members were found to be upregulated and others downregulated. We also found that only unsaturated triacylglycerides were upregulated in ME/CFS patients, but not saturated triacylglycerides. Such a finding points to a specific biochemical mechanism instead of simple explanations such as differences in the number or type of lipid-carrying lipoprotein particles. A further indication for this mechanism was a profound downregulation of unsaturated phospholipid ethers and plasmalogens (Figures 6 and 7), which are exclusively produced by peroxisomes and which, hence, might be involved in the etiology of the disease [17]. Importantly, the results of the Bayesian analysis further strengthened the biological interpretation of peroxisome damage as an important factor underlying ME/CFS: very long chain polyunsaturated triacylglycerides were found at increased levels in ME/CFS subjects (Figure 6), pointing to a lack of oxidation in peroxisomes that exclusively perform this reaction (not mitochondria). Peroxisome damage is further supported by the specific increase in phosphatidylcholines with polyunsaturated very-long-chain fatty acids (PC 40:6 and PC40:7), whereas medium- and long-chain PC lipids were found at decreased levels in the ME-CFS Bayesian statistics (PCs with 38 fewer carbons). More detailed analyses, including for covariates, are given in Che et al. [17].
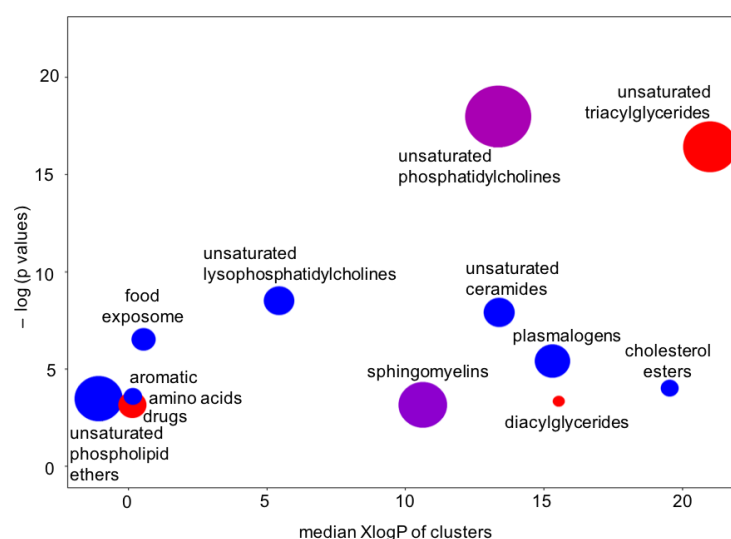

Figure 7. Chemical set enrichment statistics using the results of the Bayesian analyses conducted on the *Che* dataset. Bubble sizes indicate the number of compounds belonging to chemical groups. Bubble colors indicate the direction of effects (red = all compounds upregulated in the ME/CFS group; blue = all compounds downregulated in the ME/CFS group; purple = mixed effects).

In combination, therefore, Bayesian analyses unequivocally found evidence that reflects known behaviors in ME/CFS patients (such as avoidance of specific foods, but

increases in the use of pharmaceuticals), in addition to biochemically and physiologically interpretable findings that would have been completely overlooked and ignored by classic univariate statistics.

Yet, Bayesian statistics can go even further. To illustrate the usefulness of Bayesian analyses, we extended our investigations by taking a third ME/CFS dataset into account, the *Naviaux* study; see Figure 8. Compounds were matched across all three datasets using RefMet [19], yielding a common core of 92 analyzed compounds that were observed in all three studies. The posterior distributions from the *Che* data were used as the prior distributions in the *Naviaux* study. Bayesian statistics yielded credible intervals for 14 compounds to be different between ME/CFS subjects and that were common for all three studies. Similar to the combined Bayes analysis of just the *Nagy-Szagal* and *Che* studies (Figure 6), we found specific phosphatidylcholines to be downregulated along with the aromatic amino acids phenylalanine and tyrosine, while vitamin levels were elevated in ME/CFS patients (pantothenic acid and 4-pyridoxic acid). Interestingly, eicosapentaenoic acid (a very-long-chain polyunsaturated fatty acid, C20:5) was now found at increased levels, further supporting the peroxisome damage mechanism. Similarly, PGF2-α (an oxylipin product from polyunsaturated fatty acids and physiologically active mediator) was found to be regulated in the opposite direction, positing interesting new physiological and biochemical hypotheses on the etiology of ME/CFS that are now grounded in three independent datasets. Results such as these also show the need to further standardize metabolomics data acquisitions that would accomplish progress in direct data integrations by Bayesian statistics, as demonstrated for this ME/CFS data integration study.
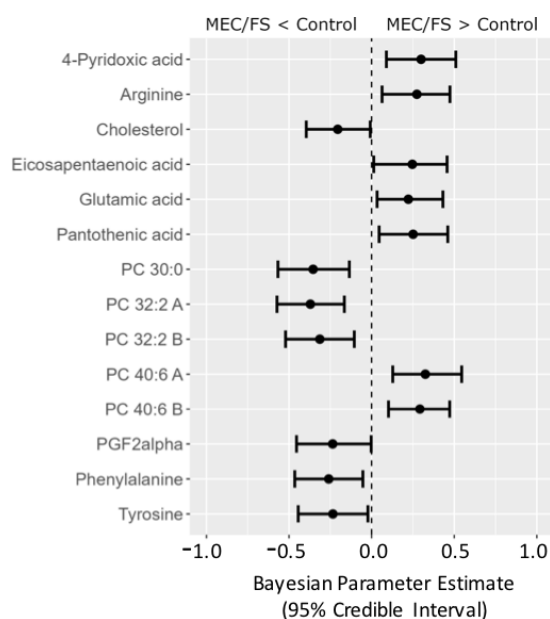

Figure 8. Forest plot of compounds with Bayesian 95% credible intervals not overlapping with zero in the *Naviaux* dataset. Points in each bar represent the posterior distribution median, and bars represent 95% credible interval bounds.

4. Discussion

Science should progress over time in its insights into specific phenomena. Bayesian statistics allow researchers to incorporate previous knowledge into their models. Hence, it is surprising that Bayesian approaches are rarely used in Omics research, and even less in metabolomics. Here, we exemplified the suitability of Bayesian approaches on three datasets that investigate the metabolic profile of ME/CFS [15,17,18]. There is little known about ME/CFS with respect to the origins of this complex disease, disease progression, possible treatment options, chances or timing of remission, or why women are more affected than men [30]. These studies were chosen precisely because the overall effects are

not well studied and because they investigated the same disease complex. The similarity in the study designs, including similar numbers of subjects, allowed us to conduct Bayesian analyses on the first dataset, then incorporate the knowledge gleaned from those results into the Bayesian analyses conducted on the second and then the third dataset. However, as a caveat, we must state that there are likely differences in the ME/CFS cohorts that are unknown to us or to the different study investigators, because the disease etiology cannot be very sharply defined. It is very possible that the disease symptoms that are today summarized as ME/CFS may be better categorized into subtypes in the future. Hence, by random chance, there may have been differences in the ME/CFS subjects (or the matched healthy controls) that could have led to differences in the metabolic imprints in the plasma metabolome data studied here. While all three studies used here emphasized the importance of complex lipids in plasma metabolomic profiles, the use of Bayesian analyses allowed us to refine the original ideas and metabolic pathways involved. For example, Naviaux et al. emphasized the importance of sphingolipids as having the "largest disturbances in the chemical signature of CFS" [18], explaining 44–50% of the metabolic impact in men and women [18]. With the results by Nagy-Szagal [15] and Che [17], published later, we can now rule out these molecules as being of high importance when integrating the studies. Instead, we were able to use prior data such as the ceramides and sphingomyelins as hypotheses to be tested, but changing distributions and, therefore, altering the posterior distributions, the effect size estimates, and the strengths of the statements using subsequent datasets. Indeed, using Bayesian statistics and incorporating previous results into subsequent analyses, we found evidence of peroxisomal dysfunction, differences in diet, and PGF2 alpha in the *Che* dataset that we would not have otherwise seen if we only used traditional statistical analyses. Similar to other statistical approaches, Bayesian analyses can only weigh between different hypotheses, but cannot ultimately provide absolute statements. For example, as Bayesian analyses provide strong evidence of the involvement of peroxisomes in the pathology of ME/CFS, we cannot rule out that mitochondria or the endoplasmic reticulum are also involved through the oxidation and modification of complex lipids. Similarly, our analyses focused on patient plasma, which almost always precludes definitive conclusions on the timing and involvement of specific organs. Future research may use animal models and the possible timing of events to study routes from potential initial causes (such as viral infections that lead to immune hyper-responses) to secondary defects (such as peroxisomal damages in the liver and brain), which then may end in differences of physiologically active lipid mediators such as PGF2alpha, which may lead to lower blood flow to brain regions, causing pain and brain fog, as reported by ME/CFS patients.

Although the analyzed datasets we used were from metabolomics backgrounds, such tests can be applied to any field of quantitative hypothesis-driven research, where classic univariate (*p*-value-driven) analyses are traditionally used. For both Bayesian and classic frequentist statistical analyses, missing values and their imputation are a cause of concern. At least in our datasets, missing values are not missing at random. Missing at random would imply a failure of reliably detecting signals in raw-data-processing software. We can exclude this problem due to back-filling computation from raw data in MS-DIAL and manual inspections of all targeted data. Hence, while this report does not discuss the best methods to impute levels for missing values, using half -minimum values is routine in metabolomics [31–33].

Statistical analyses of metabolomics data have typically followed a predefined routine of conducting a series of univariate tests, such as *t*-tests/ANOVAs or their non-parametric equivalents [34–36]. For effect size considerations, the current practice is even more dismal: fold changes are often (but not always) reported, but even these are limited as they do not take the within-compound variation into account [37]. Bayesian analyses are advantageous as they can consider both the likelihood of a hypothesis being true and an estimation of the effect size at the same time. Researchers should consider reporting *p*-values, effect sizes—whether fold changes or a standardized measure of effect size such as Cohen's

d—and Bayesian results, so that readers can gain greater insight from the data [38]. Entirely Bayesian analyses, which report Bayes factors, posterior estimates, and credible intervals, may also be suitable when designing a study.

Although Bayesian model comparisons can be conducted on any set of competing models [39], we here focused on simple case–control-style analyses. More-advanced Bayesian analyses can be employed to answer increasingly complex and specific research questions and do not have to be restricted to linear relationships [40]. Researchers who add these analytical skills to their statistical toolbox will be able to elicit richer, more-detailed information from their data than others who simply report a *p*-value.

There are some limitations of Bayesian statistics. First, the potentially subjective nature of selecting a prior distribution is a common criticism [39], as researchers may attempt multiple analyses to find a favored result using different prior distribution parameters. Therefore, researchers should aim to be transparent with their choice and justification for their priors and their statistical analyses. Choosing a suitable prior distribution can also be challenging, particularly for researchers unfamiliar with Bayesian analyses [41]. Researchers should consider the following issues when deciding upon a prior distribution: The expected effect size and its potential variability, whether the hypothesis is one- or two-tailed, and the researcher's confidence in observing a relatively specific effect size [42]. Using one dataset to inform priors for another dataset is a good idea if and only if the parameters describe the same population. A mismatch in the populations that are compared or biased sampling from the population in the first study would lead to worse inferences using a Bayesian approach. Using a mouse plasma dataset to inform a prior of a human plasma cohort would be an example of an extreme bias. However, even age-, sex-, and BMI-matched human cohorts must be treated carefully, because lifestyles including diets, disease histories, and genetic backgrounds might not be comparable. The limitations of BFs also include their interpretation. BFs only provide relative, and not absolute, evidence for a hypothesis: a statement such as "the BF of 1/50 proves the null hypothesis" is incorrect. Related to the previous comments regarding choosing a suitable prior distribution, if a prior distribution is not appropriate, the resulting BF is likely to be biased against it, thereby inaccurately estimating the strength of evidence.

5. Conclusions

Relying on *p*-values alone may only provide a limited perspective of the research findings. Bayesian analyses provide an alternative to traditional statistical analyses by enriching the information extracted from the data. The validity of the research findings is the foundation of the scientific evidence that contributes to translational research and evidence-based practice. As demonstrated in this paper, Bayesian analyses are no more difficult to understand and interpret than traditional analyses. Researchers are encouraged to incorporate the results of previous research into their current studies through the use of Bayesian statistics, thereby increasing the robustness of the results reported to inform future research and increase field knowledge.

References

1. Gigerenzer, G. Mindless statistics. *J. Socio-Econ.* **2004**, *33*, 587–606. [CrossRef]
2. Gigerenzer, G. Statistical rituals: The replication delusion and how we got there. *Adv. Methods Pract. Psychol. Sci.* **2018**, *1*, 198–218. [CrossRef]
3. Greenland, S.; Senn, S.J.; Rothman, K.J.; Carlin, J.B.; Poole, C.; Goodman, S.N.; Altman, D.G. Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *Eur. J. Epidemiol.* **2016**, *31*, 337–350. [CrossRef] [PubMed]
4. Wagenmakers, E.J. A practical solution to the pervasive problems of p values. *Psychon. Bull. Rev.* **2007**, *14*, 779–804. [CrossRef] [PubMed]
5. Wasserstein, R.L.; Lazar, N.A. The ASA statement on p-values: Context, process, and purpose. *Am. Stat.* **2016**, *70*, 129–133. [CrossRef]
6. Benjamin, D.J.; Berger, J.O.; Johannesson, M.; Nosek, B.A.; Wagenmakers, E.J.; Berk, R.; Bollen, K.A.; Brembs, B.; Brown, L.; Camerer, C.; et al. Redefine statistical significance. *Nat. Hum. Behav.* **2018**, *2*, 6–10. [CrossRef] [PubMed]
7. Lakens, D.; Adolfi, F.G.; Albers, C.J.; Anvari, F.; Apps, M.A.; Argamon, S.E.; Baguley, T.; Becker, R.B.; Benning, S.D.; Bradford, D.E.; et al. Justify your alpha. *Nat. Hum. Behav.* **2018**, *2*, 168–171. [CrossRef]
8. McShane, B.B.; Gal, D.; Gelman, A.; Robert, C.; Tackett, J.L. Abandon statistical significance. *Am. Stat.* **2019**, *73*, 235–245. [CrossRef]
9. Dienes, Z. Bayesian versus orthodox statistics: Which side are you on? *Perspect. Psychol. Sci.* **2011**, *6*, 274–290. [CrossRef]
10. Krueger, J.I.; Heck, P.R. The heuristic value of p in inductive statistical inference. *Front. Psychol.* **2017**, *9*, 908. [CrossRef]
11. Goodman, S.N. Toward evidence-based medical statistics. 2: The Bayes factor. *Ann. Intern. Med.* **1999**, *130*, 1005–1013. [CrossRef] [PubMed]
12. Wagenmakers, E.J.; Marsman, M.; Jamil, T.; Ly, A.; Verhagen, J.; Love, J.; Selker, R.; Gronau, Q.F.; Šmíra, M.; Epskamp, S.; et al. Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychon. Bull. Rev.* **2018**, *25*, 35–57. [CrossRef] [PubMed]
13. Silver, N. *The Signal and the Noise: Why So Many Predictions Fail-but Some Don't*; Penguin: New York, NY, USA, 2012.
14. Kruschke, J.K.; Liddell, T.M. Bayesian data analysis for newcomers. *Psychon. Bull. Rev.* **2018**, *25*, 155–177. [CrossRef] [PubMed]
15. Nagy-Szakal, D.; Barupal, D.K.; Lee, B.; Che, X.; Williams, B.L.; Kahn, E.J.; Ukaigwe, J.E.; Bateman, L.; Klimas, N.G.; Komaroff, A.L.; et al. Insights into myalgic encephalomyelitis/chronic fatigue syndrome phenotypes through comprehensive metabolomics. *Sci. Rep.* **2018**, *8*, 10056. [CrossRef] [PubMed]
16. Sud, M.; Fahy, E.; Cotter, D.; Azam, K.; Vadivelu, I.; Burant, C.; Edison, A.; Fiehn, O.; Higashi, R.; Nair, K.S.; et al. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* **2016**, *44*, D463–D470. [CrossRef] [PubMed]
17. Che, X.; Brydges, C.R.; Yu, Y.; Price, A.; Joshi, S.; Roy, A.; Lee, B.; Barupal, D.K.; Cheng, A.; Palmer, D.M.; et al. Metabolomic Evidence for Peroxisomal Dysfunction in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome. *Int. J. Mol. Sci.* **2022**, *23*, 7906. [CrossRef] [PubMed]
18. Naviaux, R.K.; Naviaux, J.C.; Li, K.; Bright, A.T.; Alaynick, W.A.; Wang, L.; Baxter, A.; Nathan, N.; Anderson, W.; Gordon, E. Metabolic features of chronic fatigue syndrome. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, E5472–E5480. [CrossRef] [PubMed]
19. Fahy, E.; Subramaniam, S. RefMet: A reference nomenclature for metabolomics. *Nat. Methods* **2020**, *17*, 1173–1174. [CrossRef]
20. Goodrich, B.; Gabry, J.; Ali, I.; Brilleman, S. Rstanarm: Bayesian Applied Regression Modeling via Stan. R Package Version. 2020. Available online: https://mc-stan.org/rstanarm/ (accessed on 29 August 2023).
21. Makowski, D.; Ben-Shachar, M.S.; Lüdecke, D. bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *J. Open Source Softw.* **2019**, *4*, 1541. [CrossRef]
22. Gelman, A.; Jakulin, A.; Pittau, M.G.; Su, Y.S. A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.* **2008**, *2*, 1360–1383. [CrossRef]
23. Muth, C.; Oravecz, Z.; Gabry, J. User-friendly Bayesian regression modeling: A tutorial with rstanarm and shinystan. *Quant. Methods Psychol.* **2018**, *14*, 99–119. [CrossRef]

24. Barupal, D.K.; Fiehn, O. Chemical Similarity Enrichment Analysis (ChemRICH) as alternative to biochemical pathway mapping for metabolomic datasets. *Sci. Rep.* **2017**, *7*, 14567. [CrossRef]

25. Makowski, D.; Ben-Shachar, M.S.; Chen, S.H.; Lüdecke, D. Indices of effect existence and significance in the Bayesian framework. *Front. Psychol.* **2019**, *10*, 2767. [CrossRef] [PubMed]

26. Lee, M.D.; Wagenmakers, E.J. *Bayesian Cognitive Modeling: A Practical Course*; Cambridge University Press: New York, NY, USA, 2014.

27. Jeffreys, H. *The Theory of Probability*; Oxford University Press: Oxford, UK, 1961.

28. Ackermann, M.; Strimmer, K. A general modular framework for gene set enrichment analysis. *BMC Bioinform.* **2009**, *10*, 47. [CrossRef] [PubMed]

29. Hung, J.H.; Yang, T.H.; Hu, Z.; Weng, Z.; DeLisi, C. Gene set enrichment analysis: Performance evaluation and usage guidelines. *Brief. Bioinform.* **2012**, *13*, 281–291. [CrossRef]

30. Canavan, C.; West, J.; Card, T. The epidemiology of irritable bowel syndrome. *Clin. Epidemiol.* **2014**, *6*, 71.

31. Prince, N.; Stav, M.; Cote, M.; Chu, S.H.; Vyas, C.M.; Okereke, O.I.; Palacios, N.; Litonjua, A.A.; Vokonas, P.; Sparrow, D.; et al. Metabolomics and Self-Reported Depression, Anxiety, and Phobic Symptoms in the VA Normative Aging Study. *Metabolites* **2023**, *13*, 851. [CrossRef] [PubMed]

32. Panyard, D.J.; McKetney, J.; Deming, Y.K.; Morrow, A.R.; Ennis, G.E.; Jonaitis, E.M.; Van Hulle, C.A.; Yang, C.; Sung, Y.J.; Ali, M.; et al. Large-scale proteome and metabolome analysis of CSF implicates altered glucose and carbon metabolism and succinylcarnitine in Alzheimer's disease. *Alzheimer's Dement.* 2023; *in press*. [CrossRef]

33. Hoel, F.; Hoel, A.; Pettersen, I.K.; Rekeland, I.G.; Risa, K.; Alme, K.; Sørland, K.; Fosså, A.; Lien, K.; Herder, I.; et al. A map of metabolic phenotypes in patients with myalgic encephalomyelitis/chronic fatigue syndrome. *JCI Insight* **2021**, *6*, e149217. [CrossRef]

34. Bartel, J.; Krumsiek, J.; Theis, F.J. Statistical methods for the analysis of high-throughput metabolomics data. *Comput. Struct. Biotechnol. J.* **2013**, *4*, e201301009. [CrossRef]

35. Lamichhane, S.; Sen, P.; Dickens, A.M.; Hyötyläinen, T.; Orešič, M. An overview of metabolomics data analysis: Current tools and future perspectives. *Compr. Anal. Chem.* **2018**, *82*, 387–413.

36. Saccenti, E.; Hoefsloot, H.C.; Smilde, A.K.; Westerhuis, J.A.; Hendriks, M.M. Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics* **2014**, *10*, 361–374. [CrossRef]

37. Livera, A.M.; Olshansky, M.; Speed, T.P. Statistical analysis of metabolomics data. In *Metabolomics Tools for Natural Product Discovery*; Humana Press: Totowa, NJ, USA, 2013; pp. 291–307.

38. Lakens, D.; McLatchie, N.; Isager, P.M.; Scheel, A.M.; Dienes, Z. Improving inferences about null effects with Bayes factors and equivalence tests. *J. Gerontol. Ser. B* **2020**, *75*, 45–57. [CrossRef] [PubMed]

39. Etz, A.; Haaf, J.M.; Rouder, J.N.; Vandekerckhove, J. Bayesian inference and testing any hypothesis you can specify. *Adv. Methods Pract. Psychol. Sci.* **2018**, *1*, 281–295. [CrossRef]

40. Haaf, J.M.; Rouder, J.N. Developing constraint in Bayesian mixed models. *Psychol. Methods* **2017**, *22*, 779. [CrossRef] [PubMed]

41. Morey, R.D.; Hoekstra, R.; Rouder, J.N.; Lee, M.D.; Wagenmakers, E.J. The fallacy of placing confidence in confidence intervals. *Psychon. Bull. Rev.* **2016**, *23*, 103–123. [CrossRef] [PubMed]

42. Dienes, Z.; Mclatchie, N. Four reasons to prefer Bayesian analyses over significance testing. *Psychon. Bull. Rev.* **2018**, *25*, 207–218. [CrossRef]