PLOS ONE

RESEARCH ARTICLE

Utilizing machine learning with knockoff filtering to extract significant metabolites in Crohn's disease with a publicly available untargeted metabolomics dataset

Shoaib Bin Masud[1☯], Conor Jenkins[3☯], Erika Hussey[2], Seth Elkin-Frankston[2], Phillip Mach[3], Elizabeth Dhummakupt[3‡*], Shuchin Aeron[1‡*]

1 Department of Electrical and Computer Engineering, Tufts University, Medford, MA, United States of America, **2** DEVCOM Soldier Center, Natick, MA, United States of America, **3** DEVCOM Chemical Biological Center, Aberdeen Proving Ground, Aberdeen, MD, United States of America

☯ These authors contributed equally to this work.
‡ ED and SA also contributed equally to this work.
* elizabeth.s.dhummakupt.civ@mail.mil (ED); shuchin@ece.tufts.edu (SA)

Citation: Bin Masud S, Jenkins C, Hussey E, Elkin-Frankston S, Mach P, Dhummakupt E, et al. (2021) Utilizing machine learning with knockoff filtering to extract significant metabolites in Crohn's disease with a publicly available untargeted metabolomics dataset. PLoS ONE 16(7): e0255240. https://doi.org/10.1371/journal.pone.0255240

Editor: Clara Sousa, Universidade Catolica Portuguesa Escola Superior de Biotecnologia, PORTUGAL

Received: March 17, 2021

Accepted: July 12, 2021

Published: July 29, 2021

Data Availability Statement: All relevant data are within the manuscript and Supporting information.

Funding: This analysis technique was funded by DEVCOM Soldier Center under the Measuring and Advancing Soldier Tactical Readiness and Effectiveness program. This funding was awarded to ESD. EKH and SEF, members of the Soldier Center, provided minor preparation of the manuscript.

Abstract

Metabolomic data processing pipelines have been improving in recent years, allowing for greater feature extraction and identification. Lately, machine learning and robust statistical techniques to control false discoveries are being incorporated into metabolomic data analysis. In this paper, we introduce one such recently developed technique called aggregate knockoff filtering to untargeted metabolomic analysis. When applied to a publicly available dataset, aggregate knockoff filtering combined with typical p-value filtering improves the number of significantly changing metabolites by 25% when compared to conventional untargeted metabolomic data processing. By using this method, features that would normally not be extracted under standard processing would be brought to researchers' attention for further analysis.

Introduction

Inflammatory bowel disease (disease (IBD) is an umbrella term that describes conditions like ulcerative colitis (UC) and Crohn's disease (CD). IBD is referred to as a symptom cluster, where numerous pathologies result in a subset of symptoms, diagnosis methods, and treatments [1]. This class of disorders is generally characterized by diarrhea, rectal bleeding, abdominal pain, weight loss and fatigue [1]. Ulcerative colitis, CD, and additional disorders, like inflammatory bowel disease unclassified (IBDU) are generally diagnosed and classified on a spectrum of clinical and endoscopic criteria [2]. In recent years, due to the advancement in mass spectrometry in clinical medicine, biomarkers have been proposed to aid in the diagnosis of these inflammatory bowel diseases [3–5]. As a result of IBD is a being classified as a spectrum disorder, and thus there is a desire to observe a correlation associated with regulation of these biomarkers (in either an up or down manner) and severity of disorder.

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

Important biomarker identification in metabolomics that differentiate two or more groups, has been studied widely using univariate and multivariate statistical feature selection methods. Based on the knowledge of the feature distribution, both parametric as well as non-parametric univariate statistical techniques e.g., ANOVA, Student's t test, Kolmogorov-Smirnov test, Mann-Whitney U test, Kruskal-Wallis one way analysis of variance test [6–9] have been used to select significant metabolites. These univariate methods perform multiple hypothesis tests (one hypothesis per feature), and an additional correction method is required to adjust for multiple hypothesis testing. A typical correction method, called the Bonferroni correction [10] is very conservative and leads to a lot of false negatives, especially if the number of features is very large. Benjamini-Hochberg [11] proposed a less conservative approach that controls the proportion of false discoveries among the overall discoveries (rejection of null hypothesis) made. These univariate statistical methods are incapable considering the highly correlated structure of metabolomics data beforehand, thus increasing the probability of obtaining false positives and false negatives.

Recently machine learning methods [12–18] have been shown as important tools to identify significant biomarkers. Principal component analysis (PCA) [19], hierarchical clustering analysis (HCA) [20, 21], self-organizing maps (SOMs) [22, 23], partial least square-discriminant analysis (PLS-DA) [19] and Random Forest [24] are widely used multivariate machine learning methods in metabolomics study. Recently, Mendez [13] used a single hidden layer artificial neural network (ANN) to discover significant metabolites and hypothesized this approach was equivalent to PLS-DA. The advantage of these multivariate methods is that they can consider all the features simultaneously and, consequently, deal with the correlation among the metabolites. However, most of these methods do not have the inherent ability to compute valid p-values, thus are not able to guarantee the statistical significance for the selected features. Computation of p-value requires the knowledge of the distribution under the null, which is generally unknown and is highly dependent on the feature selection algorithm. Post selection inference techniques [25–27] can compute valid p-values for the chosen features after deciding upon a model but are only applicable in restricted settings.

Barber, Candes and authors [28, 29] introduced a seminal feature selection approach called "knockoff filtering" which has the capability of handling more general model selection approaches with provable control over false discovery rate (FDR). The basic idea behind knockoff filtering is to create dummy features that are conditionally independent of the responses and satisfy pairwise exchangeability with the original features. One then concatenates the original features and these dummy features called "knockoffs" and employ any regression and classification algorithm [24, 30] to generate feature importance scores. Barber [28] proposed a new statistic called "knockoff adjusted score" by comparing the original feature importance score and corresponding dummy feature importance score. Given FDR level, a data driven threshold is then generated based on these knockoff scores for the rejection of null hypothesis with provable FDR control. One of the drawbacks of this method is that it introduces randomness in the process of generating dummy variables that may lead to high variability in the outcome. To address this problem, Nguyen [31] proposed a technique called an "Aggregation of multiple Knockoffs" (AKO). This method generates multiple copies of knockoff features independently, produces an intermediate p-value from the knockoff adjusted score [28] for each feature across all copies and then performs quantile aggregation on the p-values. AKO selects significant features by applying Benjamini-Hochberg (BHq) [11] step up procedure on the quantile aggregated p-values.

In this paper, we validate the use of aggregate knockoff filtering [31] for metabolomics, in particular untargeted metabolomics. We used a publicly available dataset "Longitudinal Metabolomics of the Human Microbiome in Inflammatory Bowel Disease" [32]. The study was

involved in the NIH Integrative Human Microbiome Project, in which 546 samples were analyzed utilizing four different chromatographic methods to completely profile the metabolome of each sample. The methods relied on high resolution/accurate mass methods of acquisition and 597 features were annotated with confirmation by standard. We will demonstrate that aggregate knockoff filtering discovers additional biomarkers of interest, specifically when non-IBD versus CD were compared with existing methods while ensuring the control over false discovery rate. We will consider the percentage of missing values as a threshold to remove metabolites in the preprocessing step as a hyperparameter and find the best threshold that guarantees the maximum discovery of significant biomarkers related to IBD.

Materials and methods

Dataset

The data is available at the NIH Common Fund's National Metabolomics Data Repository (NMDR) website, the Metabolomics Workbench https://www.metabolomicsworkbench.org under the Project ID: PR000639. The data can be accessed directly via the project DOI: 10.21228/M82T15. This work is supported by NIH grant, U2C-DK119886. Study design, instrumental methods, equipment, collection, sample preparation and other relevant study data are located within the reference cited. Of note was the comprehensive chromatography analysis utilizing four different conditions e.g. C18 Reverse-Phase negative mode acquisition, C8 Reverse-Phase positive mode acquisition, Hydrophilic interaction chromatography (HILIC) negative mode acquisition, HILIC positive mode acquisition. 546 samples were collected under each mode of chromatography condition but the number of named metabolites varies (Table 1).

Data preprocessing

Several preprocessing steps were applied to each of the four datasets. Missing values are common in mass-spectrometry (MS) based metabolomics data. Since too many missing entries will cause difficulties for subsequent analysis, handling missing values is important. To address this problem, we first applied a threshold-based prefilter on each dataset to keep or remove a particular metabolite based on this threshold. In general, we kept only those metabolites that have nonzero value in at least t% of the total number of samples. This procedure is widely known as t% rule [33]. We picked threshold t from the set $T = \{0, 60, 70, 80, 100\}$ that includes two extreme values 0, 100. In case of $t = 0$ we did not apply any thresholding to remove metabolites from the data. On the other hand, for $t = 100$, metabolites having at least one missing value are removed. Though these two extreme thresholds are not widely used as standards to deal with moderately large metabolomics dataset, we studied these extreme cases in order to assess the effect of this preprocessing step on the performance of our proposed method. We then apply K- Nearest Neighbor (KNN) missing value imputation [34] technique, that works based on the principle described in [35]. It is worth mentioning that for $t = 100$, missing value imputation step was not required. Before applying the knockoff filtering, the imputed data was standardized to zero mean and unit variance. Going forward we will define t as the missing value imputation threshold. The detailed workflow can be seen in Fig 1.

Table 1. Number of named metabolites collected under different modes.

Mode	C18 Negative	C8 Positive	HILIC Negative	HILIC Positive
Named metabolites	91	213	115	177

https://doi.org/10.1371/journal.pone.0255240.t001

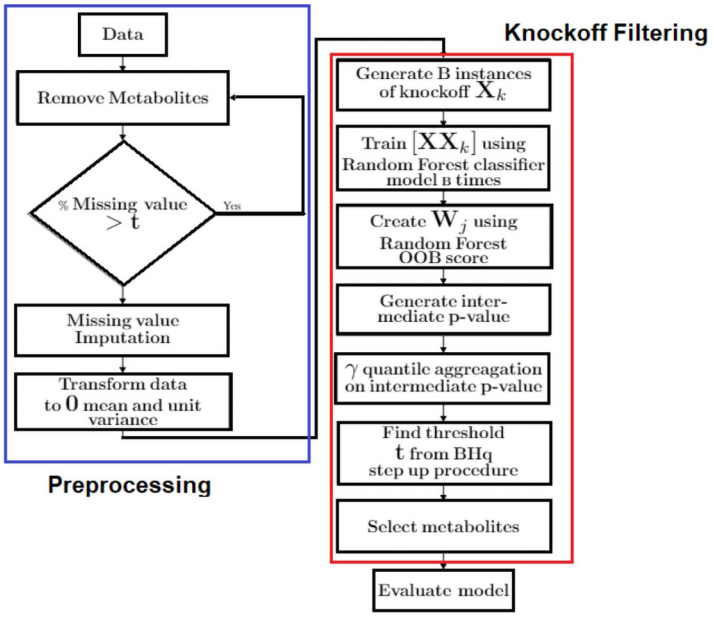

Fig 1. Workflow of the whole process. Blue box represents the data preprocessing step whereas blue box denotes the steps of knockoff filtering.

https://doi.org/10.1371/journal.pone.0255240.g001

Aggregate knockoff filtering on preprocessed data

We applied Aggregate Knockoff (AKO) filtering [31] on the preprocessed data. The filtering process begins by generating the knockoff of the original data matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, where n and p represent the number of samples and number of metabolites respectively. Knockoff data $\mathbf{X}_K \in \mathbb{R}^{n\times p}$ are generated by sampling from the conditional distribution $X_K|X \sim \mathcal{N}(\mu, \mathbf{V})$ without looking at the response vector $\mathbf{y} \in \mathbb{R}^n$. We approximated μ (mean) and $\mathbf{V}$ (covariance) using the regression formulas stated in [29] assuming original data distribution is Gaussian. As mentioned in [31], we generated B instances of Knockoff copies $\{\mathbf{X}_K^b\}_{b=1}^B$ independently. We set $B = 15$. Each knockoff $\mathbf{X}_K^b$ and original data matrix $\mathbf{X}$ were columnwise concatenated into an augmented data $[\mathbf{X}\mathbf{X}_K^b] \in \mathbb{R}^{n\times 2p}$ having twice the number of features compared to the original data matrix. Note that each dataset has three different classes e.g., CD, UC, non-IBD and response variable y_i is assigned to either 0, 1, or 2 based on the group each sample belongs to for $i = 1, \ldots, n$. We applied Random Forest classifier [12] on the augmented data to generate feature importance scores. We set the number of features that are randomly selected at each node to the square root of the number of input features. Another hyperparameter for the random forest algorithm, the number of trees was set to 1000, which we obtained using cross validation. We used the absolute mean decrease of accuracy in Out-Of-Bag (OOB) samples with random permutation of features as feature importance Z_j for $j = 1, \ldots, 2p$. OOB score is defined as the impact of each feature on the classification accuracy when removed from the input data during training. We generated the knockoff adjusted scores W_j by taking the difference between the absolute of the original feature importance score and absolute of corresponding knockoff feature importance score for $j = 1, \ldots, p$. A large positive W_j ensures that variable j truly belongs to the model. We created an intermediate p- value π_j as defined in AKO [31] for $j = 1, \ldots, p$ from the knockoff adjusted score. In brief, for B independent draws of knockoff variables we obtained the corresponding B sets of knockoff adjusted score, from which we computed p values $\pi_j^{(b)}$, for all $j = 1, \ldots, p$ and $b = 1, \ldots, B$. Then we performed γ-quantile aggregation introduced in [36] for each variable in parallel to get a new statistic $\bar{\pi}_j$ for $j = 1, \ldots, p$. We chose γ to

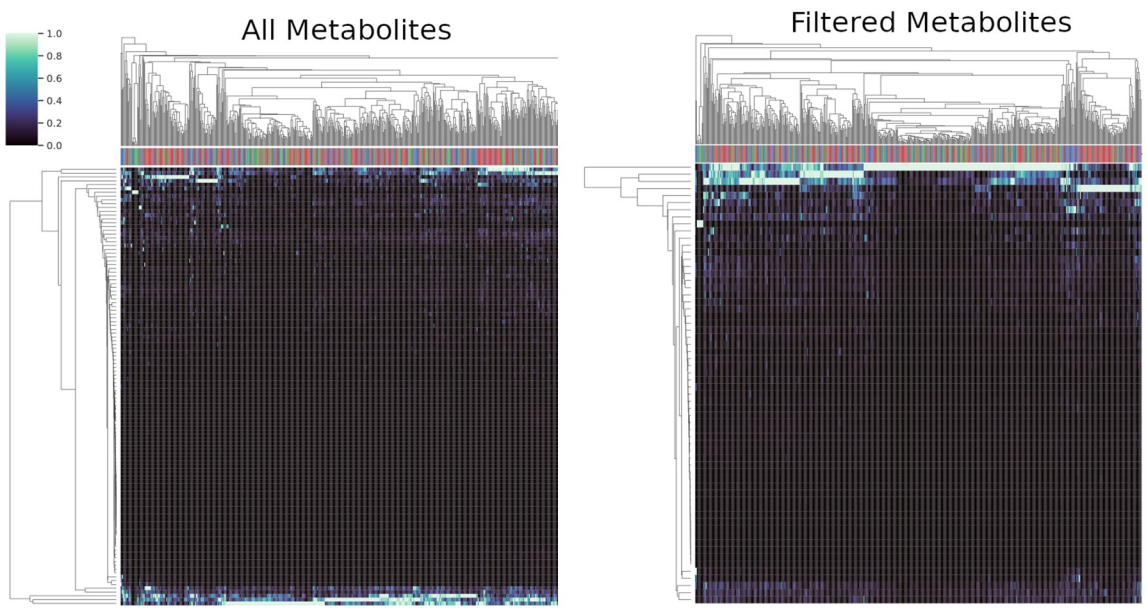

Fig 2. Clustermaps of the metabolites identified in the HILIC Positive acquisition group. All identified metabolites (left) vs filtering insignificantly changing metabolites with respect to sample group (right) are clustered respectively by Euclidean distance of expression levels (Each individual metabolite expression profile normalized to 1 across individuals) (x-axis) and colored by disease factor (Red for CD, Green for UC and Blue for non-IBD).

https://doi.org/10.1371/journal.pone.0255240.g002

0.5. After obtaining a list of p-values, we followed Benjamini-Hochberg step-up procedure [11] to select significant features given an FDR control level $\alpha = 0.05$. (We refer the reader to S1 Appendix and references therein for detailed Knockoff filtering method).

Results and discussion

Clustering of metabolite expression levels (Fig 2) performs well in differentiating families of metabolites. This level of analysis observes CD, UC, and non-IBD samples and groups them by the expression level and trends (i.e., up- or down regulation) through the study. Due to the richness of the results from untargeted metabolomics studies, it is not feasible to look at all of the metabolites and derive scientifically accurate conclusions. A further simplification of datasets often utilize a p-value cutoff. By which, this probability calculation evaluates the occurrence of extreme results and their likelihood of reoccurrence of extreme results in support of a null hypothesis. Many researchers utilize this p-value cutoff technique, to reduce datasets to more probable groupings, resulting in more manageable datasets. While compounds of similar metabolite families group well along the right y-axis (e.g., fatty acid type molecules), overall no truly observable trends or groupings allow for greater differentiation potential concerning the non-IBD and CD samples. Unfortunately, this method often does not distill the information to a manageable level. It should also be noted that when an additional p-value cutoff is applied to the data from Fig 2, certain cholates have been dropped. The loss of those compounds is irrelevant because levels of cholates have been shown to be significantly decreased in patients with IBD compared to non-IBD patients [38]. P-value cutoffs can lead to incorrect biomarker identification, wasted computational expenses and expertise. This shows the necessity of using methods to narrow the researchers' focus to fewer metabolites quickly, for initial review.

By utilizing the Aggregate Knockoff filtering technique, we can enrich the results by extracting out metabolites that truly are significant. We have identified different numbers of

Table 2. Each cell in the table represents the number of selected metabolites under different modes of data collection and threshold.

Threshold (*t*%)	C18 negative	C8 positive	HILIC negative	HILIC positive
0%	20	9	21	23
60%	33	4	32	3
70%	35	6	28	12
80%	38	6	35	13
100%	23	3	23	27

https://doi.org/10.1371/journal.pone.0255240.t002

metabolites as significant based on selected thresholds for keeping metabolites in accordance with the *t*% rule (Table 2).

The maximum number of metabolites discovered was achieved at various missing value imputation thresholds for different datasets. The reason behind obtaining different thresholds for different dataset is that each data was collected under a certain condition. Therefore, the quality of the data varies so as the missing value percentage. As an example, the maximum number of metabolites for the C18 negative dataset is observed when the threshold is 80%; however, for the HILIC positive data, the threshold is 100%. The missing value imputation level appeared to have an effect on which metabolites are selected as of interest, therefore in order to obtain the largest coverage of metabolites selected by the knockoff filtering methodology, different missing value imputation levels (0%, 60%, 70%, 80%and100%) were employed and results aggregated under each respective chromatography group (Table 2). The knockoff filtering appears to improve with smaller initial metabolite groups (HILIC Negative = 115 Metabolites, C18 Negative = 91 Metabolites vs. HILIC Positive = 177, C8 Positive = 213). C18 Negative had 15 metabolites consistently identified among each missing value imputation level but optimal coverage at 80% (Fig 3). This was not consistent in the HILIC Positive group as one missing value imputation level (100%) contains a large majority of unique metabolites vs the other levels. This shows the need to leverage various missing value imputation levels to not exclude potentially important metabolites.

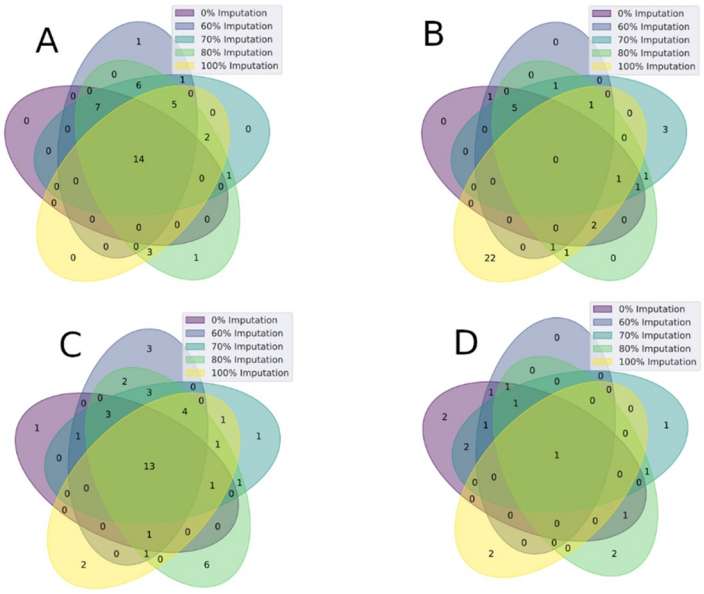

Fig 3. Venn overlaps of the metabolites identified by knockoff filtering of the C18 Negative (A), HILIC Positive (B), HILIC Negative (C) and the C8 Positive (D) chromatography groups.

https://doi.org/10.1371/journal.pone.0255240.g003

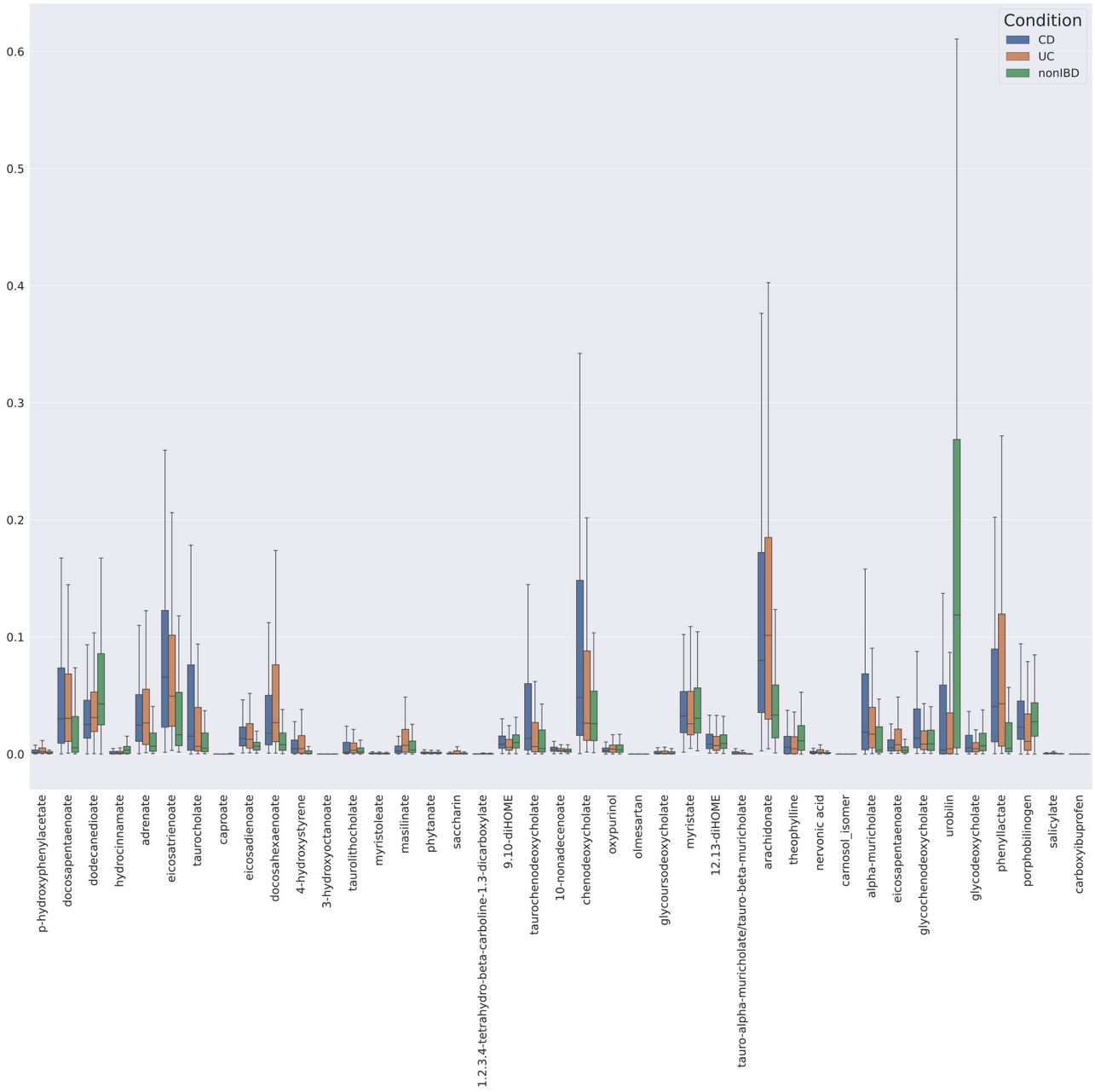

Fig 4. Boxplot of selected metabolites from C18 negative dataset.

https://doi.org/10.1371/journal.pone.0255240.g004

There are noteworthy metabolites identified that have been shown in literature to be affected in IBD populations in the C18 chromatography group (Fig 4). Arachidonate (arachidonic acid) has been shown dysregulated in IBD patients, with decreasing fold change [32]. While not a direct essential fatty acid, there is some debate regarding linoleic acid and its conversion to arachidonate to account for a deficiency in the aforementioned [37]. The Bacteria-Protease-Mucus-Barrier hypothesis suspects that saccharin may dysregulate gut bacteria and inactivate key digestive proteases [38]. Docosapentaenoate was found to be downregulated in patients [39]. Eicosatrienoate (eicosatrienoicac acid) showed dysregulation in Crohn's Disease [40]. There were 212 detected enrichments in cholate bile acids, including glycine and

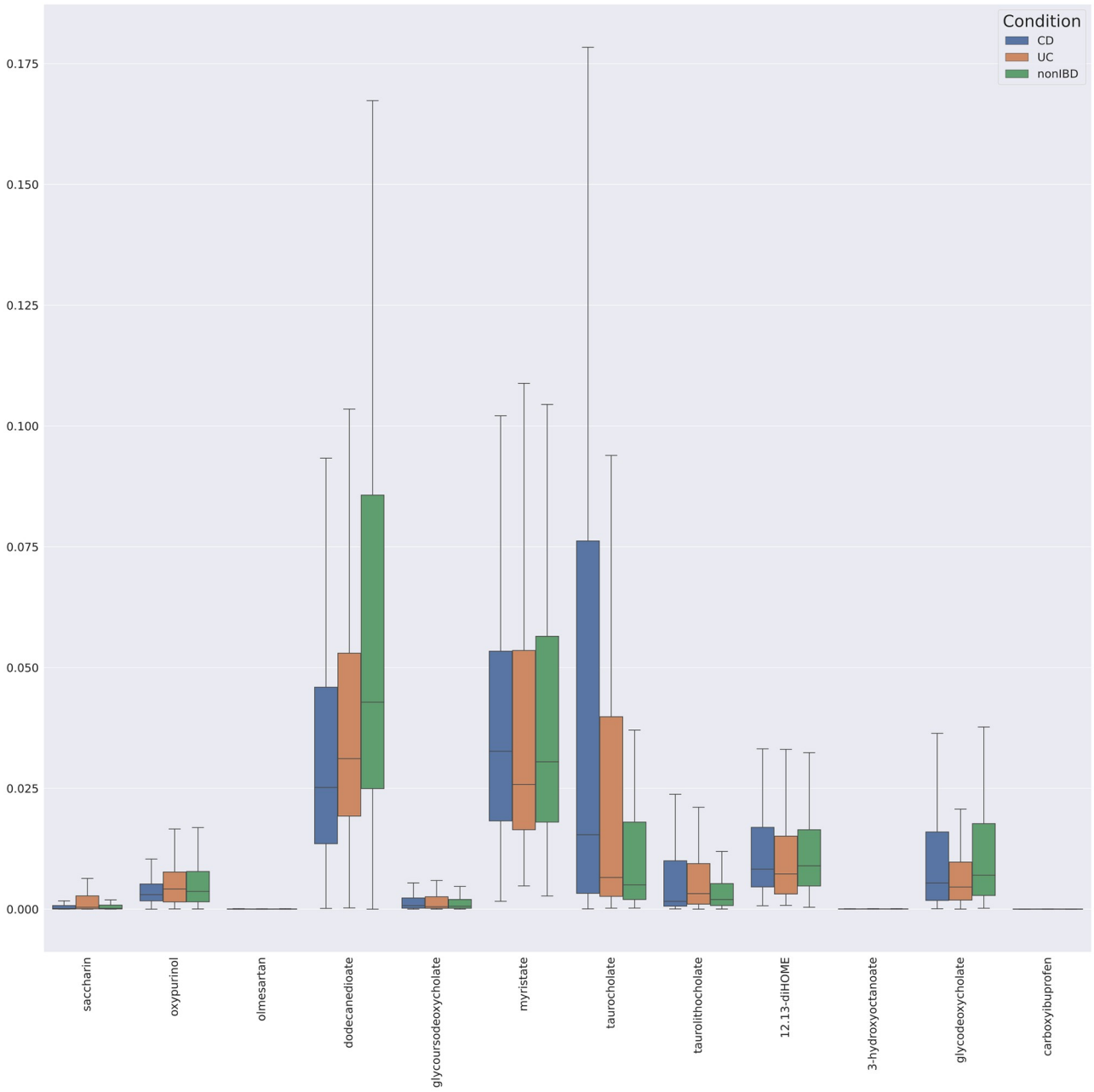

Fig 5. Boxplot of metabolites identified as important by knockoff filtering but do not pass a p-value filter. CD expression shown in blue, UC in Orange and non-IBD in Green.

https://doi.org/10.1371/journal.pone.0255240.g005

taurine conjugates [32, 41]. Numerous salicylates have been associated with IBD, including oral dosing for prevention of IBD relapse [42, 43]. Other fatty acids, like eicosadienoate, have been tracked as a possible biomarker [44]. Comprehensively, many of the other metabolites listed are also found in literature, however numerous have been tracked in one study [32]. "S2 Table" details the metabolites that were identified by running each of the datasets that were identified by running each of the datasets through aggregate knockoff filtering.

The Knockoff filtering method can also extract metabolites that do not pass the *p*-value threshold (Fig 5). These identified metabolites have also been associated with IBD-focused

research. Elevated concentration of 12,13-diHOME (12,13-dihydroxy-9Z-octadecenoic acid) impedes immune tolerance in fecal material [45]. Additional cholates were identified from this knockoff filtering method, including additional taurine conjugates [32]. Carboxylates, such as 1,2,3,4-tetrahydro-beta-carboline-1,3-dicarboxylate, were elevated, and these metabolites significantly correlate with disease prediction [46]. Dodecanedioate was also identified as important [47]. With respect to UC subjects, the model revealed variations in the occurrence of dicarboxylic acids, such as undecanedioate, dodecanedioate and sebacate, which are proposed to regulate mitochondrial fatty acid oxidation and to be involved in IBD-related liver dysfunctions.

Upon performing an enrichment analysis on the metabolites uniquely identified by the algorithm for disease signature utilizing Metaboanalyst [48]. Unclassified IBD, Ulcerative Colitis, and Crohn's Disease are all significantly enriched (Fig 6). This provides a secondary validation that the metabolites selected from the algorithm have a role in IBD.

Knockoff filtering is shown to be a great tool to discover important metabolites that were not identified by the p-value cutoff method, which is widely used in many existing metabolomics processing tools e.g., Metaboanalyst. However, the performance of the knockoff filtering method is highly dependent on the variable selection algorithms that are used to generate feature importance scores which, in turn is sensitive to data preprocessing steps e.g., pre-filtering, missing value imputation technique. One of the drawbacks of the proposed method is that we generated the second-order knockoffs by approximating mean and covariance assuming data distribution is Gaussian. In cases where this assumption is not satisfied, the proposed method will generate poor quality knockoffs and consequently lead to poor performance. This knockoff method also suffers from false discovery vs. power tradeoff like all the existing works that only control the FDR. Future study can be conducted with recently developed generative model-based knockoff generation techniques [49–51], as well as in the direction of increasing the power of detection while keeping the FDR below a significant level.

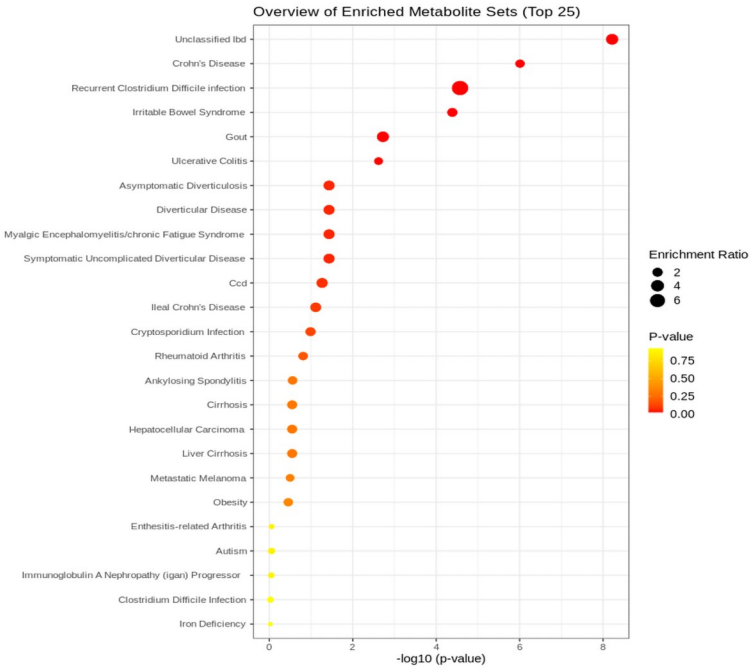

Fig 6. Metaboanalyst enrichment mapping of unique features identified by algorithm.

https://doi.org/10.1371/journal.pone.0255240.g006

Conclusion

Utilizing knockoff filtering in combination with more traditional techniques (i.e., p-value cut-off) improves researchers' abilities to sift through the large amounts of data that are generated in metabolomic experiments. The combination of aggregate knockoff filtering and p-value cut-offs allows for more rapid secondary validation and additional hypothesis generation than taking the time tracing down the dead-end leads. Aggregate Knockoff filtering technique also produces metabolites that simple p-value filtering misses. These metabolites have been implicated in having a roll in CD/IBD and would otherwise go unseen if not for the Knockoff filtering method. Aggregate knockoff filtering method also ensures the statistical significance of the selected metabolites which may not be guaranteed in case of many traditional machine learning techniques. In conclusion, this paper introduces the knockoff filtering technique to the metabolomics community which is shown to be a better tool to identify metabolites with statistical guarantee.

Supporting information

S1 Fig. Box plot of the metabolites in the C8 positive, HILIC negative and HILIC positive. Contains metabolites that are selected by the knockoff filtering algorithm but do not pass a *p*-value filter of 0.05 for three different datasets.
(PNG)

S1 Table. Identified metabolites from the knockoff filtering methodology utilizing aggregate repeated missing value imputation. This table contains the union of metabolites coming from the sets of identified metabolites using different thresholds for each dataset.
(XLSX)

S2 Table. Metaboanalyst output of an enrichment analysis of the metabolites that were identified by the algorithm but are not identified by a p-value selection. This table contains enrichment of the metabolites that only passed through the knockoff filtering.
(XLSX)

S1 Appendix. Supplementary notes. In the notes we provide some background theory on Model-X Knockoff filtering and Aggregate knockoff filtering method.
(TEX)

S1 Code. Python code.
(TXT)

Author Contributions

Conceptualization: Conor Jenkins, Shuchin Aeron.

Data curation: Shoaib Bin Masud.

Formal analysis: Shoaib Bin Masud, Conor Jenkins.

Funding acquisition: Erika Hussey, Elizabeth Dhummakupt.

Methodology: Shoaib Bin Masud, Conor Jenkins.

Project administration: Elizabeth Dhummakupt, Shuchin Aeron.

Resources: Elizabeth Dhummakupt.

Software: Shoaib Bin Masud.

Supervision: Seth Elkin-Frankston, Elizabeth Dhummakupt, Shuchin Aeron.

Validation: Shoaib Bin Masud, Conor Jenkins.

Visualization: Shoaib Bin Masud, Conor Jenkins.

Writing – original draft: Shoaib Bin Masud, Conor Jenkins, Phillip Mach, Elizabeth Dhummakupt, Shuchin Aeron.

Writing – review & editing: Erika Hussey, Seth Elkin-Frankston, Phillip Mach, Elizabeth Dhummakupt, Shuchin Aeron.

References

1. Whitehead WE, Engel BT, Schuster MM. Irritable bowel syndrome. Digestive diseases and sciences. 1980; 25(6):404–413. https://doi.org/10.1007/BF01395503 PMID: 7379673
2. Thurgate LE, Lemberg DA, Day AS, Leach ST. An overview of inflammatory bowel disease unclassified in children. Inflammatory Intestinal Diseases. 2019; 4(3):97–103. https://doi.org/10.1159/000501519 PMID: 31559261
3. Bennike T, Birkelund S, Stensballe A, Andersen V. Biomarkers in inflammatory bowel diseases: current status and proteomics identification strategies. World Journal of Gastroenterology: WJG. 2014; 20(12):3231. https://doi.org/10.3748/wjg.v20.i12.3231 PMID: 24696607
4. Iskandar HN, Ciorba MA. Biomarkers in inflammatory bowel disease: current practices and recent advances. Translational Research. 2012; 159(4):313–325. https://doi.org/10.1016/j.trsl.2012.01.001 PMID: 22424434
5. Nanni P, Parisi D, Roda G, Casale M, Belluzzi A, Roda E, et al. Serum protein profiling in patients with inflammatory bowel diseases using selective solid-phase bulk extraction, matrix-assisted laser desorption/ionization time-of-flight mass spectrometry and chemometric data analysis. Rapid Communications in Mass Spectrometry: An International Journal Devoted to the Rapid Dissemination of Up-to-the-Minute Research in Mass Spectrometry. 2007; 21(24):4142–4148. https://doi.org/10.1002/rcm.3323 PMID: 18022963
6. Broadhurst DI, Kell DB. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. Metabolomics. 2006; 2(4):171–196. https://doi.org/10.1007/s11306-006-0037-z
7. Vinaixa M, Samino S, Saez I, Duran J, Guinovart JJ, Yanes O. A guideline to univariate statistical analysis for LC/MS-based untargeted metabolomics-derived data. Metabolites. 2012; 2(4):775–795. https://doi.org/10.3390/metabo2040775 PMID: 24957762
8. Albaum SP, Hahne H, Otto A, Haußmann U, Becher D, Poetsch A, et al. A guide through the computational analysis of isotope-labeled mass spectrometry-based quantitative proteomics data: an application study. Proteome science. 2011; 9(1):30. https://doi.org/10.1186/1477-5956-9-30 PMID: 21663690
9. Suzuki K, Nosyreva E, Hunt KW, Kavalali ET, Monteggia LM. Effects of a ketamine metabolite on synaptic NMDAR function. Nature. 2017; 546(7659):E1–E3. https://doi.org/10.1038/nature22084 PMID: 28640258
10. Cross AJ, Moore SC, Boca S, Huang WY, Xiong X, Stolzenberg-Solomon R, et al. A prospective study of serum metabolites and colorectal cancer risk. Cancer. 2014; 120(19):3049–3057. https://doi.org/10.1002/cncr.28799 PMID: 24894841
11. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological). 1995; 57(1):289–300.
12. Trainor PJ, DeFilippis AP, Rai SN. Evaluation of classifier performance for multiclass phenotype discrimination in untargeted metabolomics. Metabolites. 2017; 7(2):30. https://doi.org/10.3390/metabo7020030 PMID: 28635678
13. Mendez KM, Broadhurst DI, Reinke SN. Migrating from partial least squares discriminant analysis to artificial neural networks: a comparison of functionally equivalent visualisation and feature contribution tools using jupyter notebooks. Metabolomics. 2020; 16(2):17. https://doi.org/10.1007/s11306-020-1640-0 PMID: 31965332
14. Antonelli J, Claggett BL, Henglin M, Kim A, Ovsak G, Kim N, et al. Statistical workflow for feature selection in human metabolomics data. Metabolites. 2019; 9(7):143. https://doi.org/10.3390/metabo9070143 PMID: 31336989
15. Turck CW, Mak TD, Goudarzi M, Salek RM, Cheema AK. The ABRF Metabolomics Research Group 2016 Exploratory Study: Investigation of Data Analysis Methods for Untargeted Metabolomics. Metabolites. 2020; 10(4):128. https://doi.org/10.3390/metabo10040128

16. Bünger R, Mallet RT. Metabolomics and ROC Analysis: A Promising Approach for Sepsis Diagnosis. Critical care medicine. 2016; 44(9):1784. https://doi.org/10.1097/CCM.0000000000001795 PMID: 27525998

17. Worley B, Powers R. Multivariate analysis in metabolomics. Current Metabolomics. 2013; 1(1):92–107. https://doi.org/10.2174/2213235X11301010092 PMID: 26078916

18. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. Journal of The Royal Society Interface. 2018; 15(141):20170387. https://doi.org/10.1098/rsif.2017.0387 PMID: 29618526

19. Treutler H, Tsugawa H, Porzel A, Gorzolka K, Tissier A, Neumann S, et al. Discovering regulated metabolite families in untargeted metabolomics studies. Analytical chemistry. 2016; 88(16):8082–8090. https://doi.org/10.1021/acs.analchem.6b01569 PMID: 27452369

20. Sreekumar A, Poisson LM, Rajendiran TM, Khan AP, Cao Q, Yu J, et al. Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. Nature. 2009; 457(7231):910–914. https://doi.org/10.1038/nature07762 PMID: 19212411

21. Kreft H, Jetz W. Global patterns and determinants of vascular plant diversity. Proceedings of the National Academy of Sciences. 2007; 104(14):5925–5930. https://doi.org/10.1073/pnas.0608361104 PMID: 17379667

22. Goodwin CR, Sherrod SD, Marasco CC, Bachmann BO, Schramm-Sapyta N, Wikswo JP, et al. Phenotypic mapping of metabolic profiles using self-organizing maps of high-dimensional mass spectrometry data. Analytical chemistry. 2014; 86(13):6563–6571. https://doi.org/10.1021/ac5010794 PMID: 24856386

23. Mäkinen VP, Soininen P, Forsblom C, Parkkonen M, Ingman P, Kaski K, et al. 1H NMR metabonomics approach to the disease continuum of diabetic complications and premature death. Molecular systems biology. 2008; 4(1):167. https://doi.org/10.1038/msb4100205 PMID: 18277383

24. Kokla M, Virtanen J, Kolehmainen M, Paananen J, Hanhineva K. Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: a comparative study. BMC bioinformatics. 2019; 20(1):1–11. https://doi.org/10.1186/s12859-019-3110-0 PMID: 31601178

25. Berk R, Brown L, Buja A, Zhang K, Zhao L, et al. Valid post-selection inference. The Annals of Statistics. 2013; 41(2):802–837. https://doi.org/10.1214/12-AOS1077

26. Lee JD, Sun DL, Sun Y, Taylor JE, et al. Exact post-selection inference, with application to the lasso. The Annals of Statistics. 2016; 44(3):907–927. https://doi.org/10.1214/15-AOS1371

27. Tibshirani RJ, Taylor J, Lockhart R, Tibshirani R. Exact post-selection inference for sequential regression procedures. Journal of the American Statistical Association. 2016; 111(514):600–620. https://doi.org/10.1080/01621459.2015.1108848

28. Barber RF, Candès EJ, et al. Controlling the false discovery rate via knockoffs. The Annals of Statistics. 2015; 43(5):2055–2085. https://doi.org/10.1214/15-AOS1337

29. Candes E, Fan Y, Janson L, Lv J. Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection. arXiv preprint arXiv:161002351. 2016.

30. Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2011; 73(3):273–282. https://doi.org/10.1111/j.1467-9868.2011.00771.x

31. Nguyen BT, Chevalier JA, Thirion B, Arlot S. Aggregation of Multiple Knockoffs. arXiv preprint arXiv:200209269. 2020.

32. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 2019; 569(7758):655–662. https://doi.org/10.1038/s41586-019-1237-9 PMID: 31142855

33. Bijlsma S, Bobeldijk I, Verheij ER, Ramaker R, Kochhar S, Macdonald IA, et al. Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. Analytical chemistry. 2006; 78(2):567–574. https://doi.org/10.1021/ac051495j PMID: 16408941

34. Armitage EG, Godzien J, Alonso-Herranz V, López-Gonzálvez Á, Barbas C. Missing value imputation strategies for metabolomics data. Electrophoresis. 2015; 36(24):3050–3060. https://doi.org/10.1002/elps.201500352 PMID: 26376450

35. Gromski PS, Xu Y, Kotze HL, Correa E, Ellis DI, Armitage EG, et al. Influence of missing values substitutes on multivariate analysis of metabolomics data. Metabolites. 2014; 4(2):433–452. https://doi.org/10.3390/metabo4020433 PMID: 24957035

36. Meinshausen N, Meier L, Bühlmann P. P-values for high-dimensional regression. Journal of the American Statistical Association. 2009; 104(488):1671–1681. https://doi.org/10.1198/jasa.2009.tm08647

37. Rett BS, Whelan J. Increasing dietary linoleic acid does not increase tissue arachidonic acid content in adults consuming Western-type diets: a systematic review. Nutrition & metabolism. 2011; 8(1):36. https://doi.org/10.1186/1743-7075-8-36

38. Qin X. Etiology of inflammatory bowel disease: a unified hypothesis. World journal of gastroenterology: WJG. 2012; 18(15):1708. https://doi.org/10.3748/wjg.v18.i15.1708 PMID: 22553395

39. Solakivi T, Kaukinen K, Kunnas T, Lehtimäki T, Mäki M, Nikkari ST. Serum fatty acid profile in subjects with irritable bowel syndrome. Scandinavian journal of gastroenterology. 2011; 46(3):299–303. https://doi.org/10.3109/00365521.2010.533380 PMID: 21073373

40. Kuroki F, Iida M, Matsumoto T, Aoyagi K, Kanamoto K, Fujishima M. Serum n3 polyunsaturated fatty acids are depleted in Crohn's disease. Digestive diseases and sciences. 1997; 42(6):1137–1141. https://doi.org/10.1023/A:1018873217192 PMID: 9201073

41. Tiratterra E, Franco P, Porru E, Katsanos KH, Christodoulou DK, Roda G. Role of bile acids in inflammatory bowel disease. Annals of gastroenterology. 2018; 31(3):266. https://doi.org/10.20524/aog.2018.0239 PMID: 29720851

42. Travis S, Jewell D. Salicylates for inflammatory bowel disease. Baillière's clinical gastroenterology. 1994; 8(2):203–231. https://doi.org/10.1016/0950-3528(94)90002-7 PMID: 7949456

43. Franchis RD, Omodei P, Ranzi T, Brignola C, Rocca R, Prada A, et al. Controlled trial of oral 5-aminosalicylic acid for the prevention of early relapse in Crohn's disease. Alimentary pharmacology & therapeutics. 1997; 11(5):845–852. https://doi.org/10.1046/j.1365-2036.1997.00212.x PMID: 9354191

44. Sitkin S, Pokrotnieks J. Alterations in polyunsaturated fatty acid metabolism and reduced serum eicosadienoic acid level in ulcerative colitis: is there a place for metabolomic fatty acid biomarkers in IBD? Digestive diseases and sciences. 2018; 63(9):2480–2481. https://doi.org/10.1007/s10620-018-5182-5 PMID: 29987625

45. Levan SR, Stamnes KA, Lin DL, Panzer AR, Fukui E, McCauley K, et al. Elevated faecal 12, 13-diHOME concentration in neonates at high risk for asthma is produced by gut bacteria and impedes immune tolerance. Nature microbiology. 2019; 4(11):1851–1861. https://doi.org/10.1038/s41564-019-0498-2 PMID: 31332384

46. Volkova A, Ruggles KV. Predictive Metagenomic Analysis of Autoimmune Disease Identifies Robust Autoimmunity and Disease Specific Signatures. bioRxiv. 2019; p. 779967.

47. Lee T, Clavel T, Smirnov K, Schmidt A, Lagkouvardos I, Walker A, et al. Oral versus intravenous iron replacement therapy distinctly alters the gut microbiota and metabolome in patients with IBD. Gut. 2017; 66(5):863–871. https://doi.org/10.1136/gutjnl-2015-309940 PMID: 26848182

48. Chong J, Wishart DS, Xia J. Using MetaboAnalyst 4.0 for comprehensive and integrative metabolomics data analysis. Current protocols in bioinformatics. 2019; 68(1):e86. https://doi.org/10.1002/cpbi.86 PMID: 31756036

49. Romano Y, Sesia M, Candès E. Deep knockoffs. Journal of the American Statistical Association. 2020; 115(532):1861–1872. https://doi.org/10.1080/01621459.2019.1660174

50. Liu Y, Zheng C. Auto-encoding knockoff generator for FDR controlled variable selection. arXiv preprint arXiv:180910765. 2018.

51. Lu Y, Fan Y, Lv J, Noble WS. DeepPINK: reproducible feature selection in deep neural networks. In: Advances in Neural Information Processing Systems; 2018. p. 8676–8686.