BMC Bioinformatics

RESEARCH

Open Access

Single sample pathway analysis in metabolomics: performance evaluation and application

Cecilia Wieder[1], Rachel P. J. Lai[2] and Timothy M. D. Ebbels[1*]

*Correspondence:
t.ebbels@imperial.ac.uk

[1] Section of Bioinformatics, Division of Systems Medicine, Department of Metabolism, Digestion, and Reproduction, Faculty of Medicine, Imperial College London, London, UK
[2] Department of Infectious Disease, Faculty of Medicine, Imperial College London, London, UK

Abstract

Background: Single sample pathway analysis (ssPA) transforms molecular level omics data to the pathway level, enabling the discovery of patient-specific pathway signatures. Compared to conventional pathway analysis, ssPA overcomes the limitations by enabling multi-group comparisons, alongside facilitating numerous downstream analyses such as pathway-based machine learning. While in transcriptomics ssPA is a widely used technique, there is little literature evaluating its suitability for metabolomics. Here we provide a benchmark of established ssPA methods (ssGSEA, GSVA, SVD (PLAGE), and z-score) alongside the evaluation of two novel methods we propose: ssClustPA and kPCA, using semi-synthetic metabolomics data. We then demonstrate how ssPA can facilitate pathway-based interpretation of metabolomics data by performing a case-study on inflammatory bowel disease mass spectrometry data, using clustering to determine subtype-specific pathway signatures.

Results: While GSEA-based and z-score methods outperformed the others in terms of recall, clustering/dimensionality reduction-based methods provided higher precision at moderate-to-high effect sizes. A case study applying ssPA to inflammatory bowel disease data demonstrates how these methods yield a much richer depth of interpretation than conventional approaches, for example by clustering pathway scores to visualise a pathway-based patient subtype-specific correlation network. We also developed the sspa python package (freely available at https://pypi.org/project/sspa/), providing implementations of all the methods benchmarked in this study.

Conclusion: This work underscores the value ssPA methods can add to metabolomic studies and provides a useful reference for those wishing to apply ssPA methods to metabolomics data.

Keywords: Single-sample pathway analysis, Metabolomics pathway analysis, Enrichment analysis, Benchmarking, Simulation, Pathway visualisation

Introduction

As metabolomics continues to emerge as a powerful platform for profiling small molecules in various areas of research [1–3], interpretation of the results of such studies remains of paramount importance. This is particularly critical in untargeted

experiments, in which as many metabolites as possible are assayed in order to build a comprehensive metabolic profile of the phenotype under study [4]. Such a study does not usually begin with a specific hypothesis, but seeks to develop and refine hypotheses through downstream analysis of the data. A typical metabolomics analysis workflow usually involves identifying a subset of metabolites of interest in relation to the study objective, which can be achieved in various ways not limited to statistical association testing, multivariate statistical approaches, and machine learning (classification and regression) [5]. Once metabolites of interest are identified, the next step often involves placing these in a biological context. Pathway analysis (PA) is perhaps the most well-known computational approach for doing so, typically providing users with a list of pathways considered enriched in the condition of interest (e.g. disease vs. healthy) [6–9]. These pathways are curated using both manual and computational approaches and deposited in databases, representing sets of biochemical reactions that collectively perform a specific function [10].

Conventional PA approaches widely used across omics data types (genomics, transcriptomics, proteomics, and metabolomics, etc.) commonly focus on a two-group analysis, seeking to identify significantly impacted pathways between the study groups. These methods can be broadly classified into three main categories: over-representation analysis (ORA) [11], functional class scoring approaches such as GSEA [12], and network topology-based approaches [13]. While there exists a wealth of literature describing and evaluating these approaches in transcriptomics [6, 9, 13], the development and application of these methods is only at the early exploration phase in the metabolomics context [14–16]. Despite the popularity and success of these methods, there are several use cases in which they are unsuitable: i) studies in which there is more than a single contrast (or continuous outcome) to be analysed, and ii) when the aim is to estimate the importance of a pathway at an individual sample level, rather than across experimental groups as a whole.

Single-sample PA (ssPA) refers to methods used to compute a score representing the enrichment level of each pathway for each individual sample in a study [17]. Another way to conceptualise ssPA is that it can be used to transform individual molecular-level data to the pathway level (Fig. 1a). For example, the metabolite abundance matrix $X_{n\times m}$, with n rows representing samples and m columns representing metabolites, could be transformed to the matrix $Y_{n\times p}$ with rows representing identical samples but columns instead representing pathways (Fig. 1b). This metabolite to pathway-level transformation is achieved using ssPA algorithms, which utilise existing knowledge from pathway databases [18–21] to combine metabolite-level measurements into pathway scores. This principle of transforming a data matrix from the individual molecular measurements to the pathway space is generalisable to any type of omics data, including metabolomics. By calculating pathway scores for each sample within a dataset, the concept of ssPA overcomes the limitations of conventional PA methods, allowing for multi-group PA and enabling PA at the individual sample level.

Current ssPA methods can be broadly categorised into three main groups: dimensionality reduction (DR)-based [22, 23], GSEA-based [24, 25], and z-score-based [26]. The earliest example of ssPA can be attributed to Tomfohr et al. [22], who developed the PLAGE method for calculating pathway scores using singular value

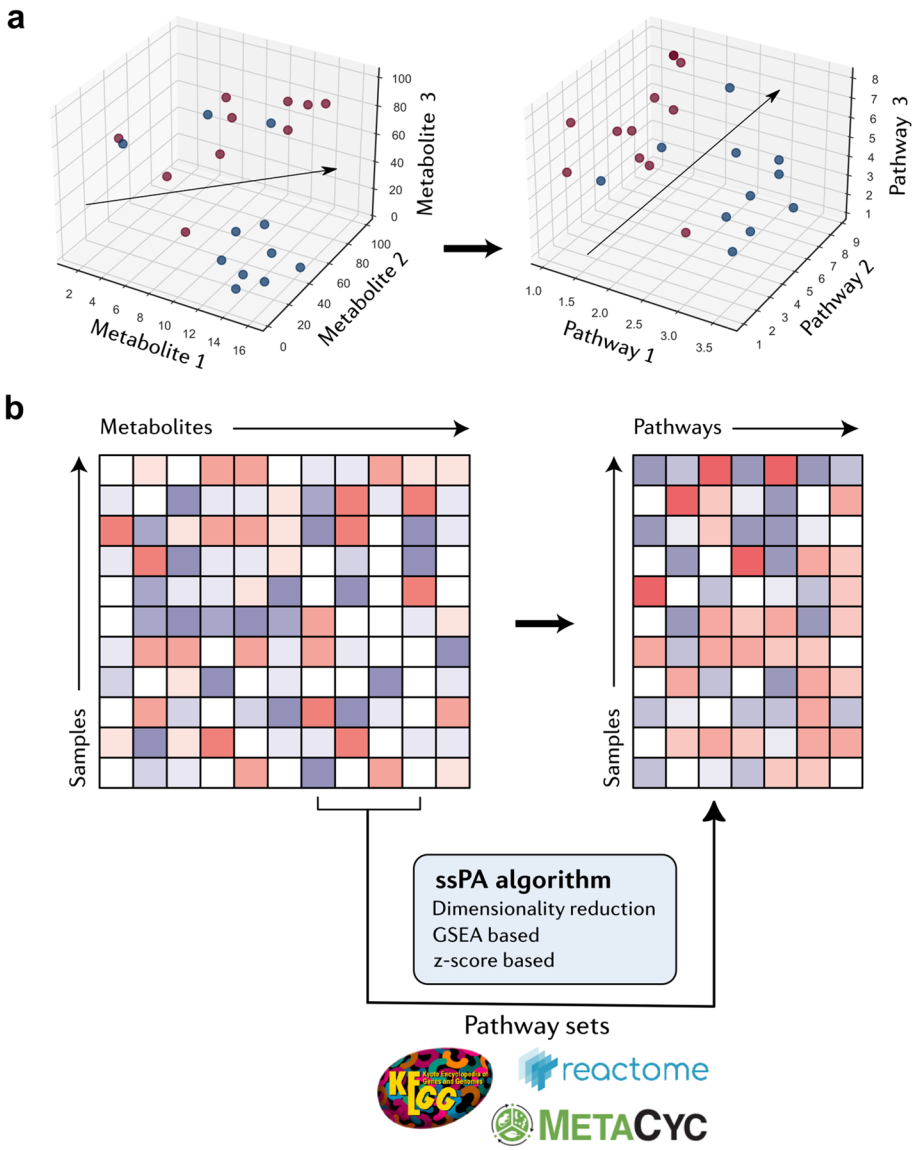

Fig. 1 Schematic representation of single sample pathway analysis. **a** Transformation of omics data from the metabolite space (left) to the pathway space (right). Data point colours represent sample groups. **b** Left represents original omics data matrix (metabolomics) $X_{n\times m}$, and right represents ssPA-transformed pathway level omics data matrix $Y_{n\times p}$. Both matrices contain rows representing the same samples, but ssPA transforms the columns of the original matrix representing individual molecular measurements (in this example metabolites) into pathways. An ssPA algorithm is used to perform the pathway transformation, taking the input metabolomics matrix (left) alongside a set of pathways as input. Pathway logos shown correspond to KEGG, Reactome, and MetaCyc databases, which are examples of where pathway sets can be obtained

decomposition (SVD). ssGSEA and GSVA are two similar ssPA methods based on the Kolmogorov–Smirnov like random walk statistic proposed by Subramanian et al. [12] for conventional GSEA. The z-score method developed by Lee et al. [26] calculates pathway scores based on a normalised z-score across all molecules in a pathway. Full details of these methods can be found in the original articles [22, 24–26].

By generating sample-wise pathway scores, ssPA enables numerous downstream analyses to be performed that cannot be achieved using conventional PA approaches.

At the most basic level, multi-group comparisons can be made using the pathway scores, for example using statistical association testing to determine pathways which differ significantly between groups [25]. This can be useful in studies where there are more than two treatment groups or disease subtypes. Another prominent example of the utility of ssPA methods is that they enable application of machine learning or multivariate statistical methods to pathway level data [27]. Patients can be classified based on their pathway scores, rather than metabolite-level measurements, which in some cases has been demonstrated to improve classification performance [26]. Predictive models built in the pathway space have also shown to be more robust to noise than those constructed from individual molecular measurements [28]. Pathway scores further enable a number of pathway-based visualisation options, such as plotting the scores of two pathways against each other to discriminate between disease subtypes [29], or using them to generate hierarchical clustering heatmaps [22]. An emerging use-case of ssPA is in multi-omics data integration, in which pathway scores calculated for each omics layer can be combined and analysed concurrently [30].

Although most ssPA methods are applicable to most omics datatypes, this will not necessarily equate to their consistent performance across omics. Indeed, the most well-known ssPA methods have all been developed for transcriptomic data analysis. Other omics datatypes such as metabolomics differ greatly in composition and statistical properties from transcriptomic data. The most obvious difference lies in the number of metabolites versus the number of genes/transcripts profiled in metabolomics and transcriptomics respectively, which has a direct effect on pathway coverage (the proportion of entities in a given pathway that have been assayed). Another key difference is the uncertainty in metabolite identification in metabolomics which remains one of the field's greatest challenges [31], an issue that is far less prominent in transcriptomics and proteomics. There is a growing body of literature focused on benchmarking PA methods across various omics datatypes [13, 15, 17, 32–37], but to date there have been no studies investigating the efficacy of ssPA methods applied to metabolomics datasets.

The purpose of the present research, therefore, is twofold: we begin by performing a benchmark of widely used and established ssPA methods using semi-synthetic metabolomics datasets, after which we demonstrate how ssPA methods can aid in the interpretation of metabolomics data by performing an application case study using publicly available inflammatory bowel disease (IBD) data. Specifically, in the benchmarking portion of this manuscript, we compare the performance of four well-known ssPA methods, namely SVD (PLAGE) [22], ssGSEA [24], GSVA [25], and z-score[26] as well as two ssPA methods we propose, based on k-means clustering (ssClustPA) and kernel principal component analysis (kPCA). We also introduce the sspa Python package which provides user-friendly implementations of all methods benchmarked, alongside a tutorial, specifically designed for application to metabolomics data. In the second part of this manuscript we demonstrate using experimental IBD data how ssPA generates pathway scores which can be used to cluster and discriminate samples based on their IBD subtype, facilitating interpretation of the data at the pathway-level. To summarise, this is the first study to benchmark ssPA methods on metabolomics

data, offering important insights into their suitability and potential applications in metabolomics research.

Methods

Datasets used

In the benchmarking section of this work, the Su et al. [38] COVID19 metabolomics mass spectrometry (MS) dataset has been used, serving as a basis for semi-synthetic simulated data creation. This data is part of a multi-omics study of COVID severity, and contains patients from two groups: COVID (of varying levels of severity) (n = 130) and non-COVID (n = 133), from which blood plasma samples were obtained during the first week of infection after diagnosis.

For the application section, we used inflammatory bowel disease (IBD) MS metabolomics data from Lloyd-Price et al. [39]. This data derives from a longitudinal multi-omics study of IBD, where metabolomics analysis was performed on stool samples collected over a 1-year period. We used samples obtained across all timepoints (weeks 0–52). The study contains two experimental groups: IBD (Crohn's disease (CD, n = 265) and ulcerative colitis (UC, n = 146)) and non-IBD (n = 135).

Both datasets are publicly available (see Availability of data and materials for repository identifiers) and are derived from untargeted mass spectrometry, see Table 1 for further details. Datasets were post-processed in the same manner: missing value imputation using iterative SVD [40], probabilistic quotient normalisation, log_2 transformation, and standardisation of each variable ($\mu = 0$ and $\sigma = 1$). Metabolite names were mapped to ChEBI identifiers using the MetaboAnalyst [41] identifier conversion tool (https://www.metaboanalyst.ca/MetaboAnalyst/upload/ConvertView.xhtml).

Pathway definitions and non-redundant pathway set

Reactome pathways (release 76) were downloaded from https://reactome.org/download-data. The ChEBI2Reactome_All_Levels.txt file was used and filtered for Homo sapiens pathways only. The metabolite coverage of the Reactome human pathways using the COVID dataset ranged from 2 to 39 metabolites per pathway. KEGG human pathways

Table 1 Summary of datasets used

Dataset	Samples	Sample source	Assay	Number of metabolites mapping to ChEBI	Number of Reactome pathways accessible[a]
Su et al. [38]	COVID19 (n = 130), Non-COVID19 (n = 133)	Blood plasma	UHPLC/MS/MS (Metabolon)	335	255
Lloyd-Price et al. [39]	CD (n = 265), UC (n = 146), Non-IBD (n = 135)	Stool	LC–MS HILIC-pos, LC–MS HILIC-neg, LC–MS C18-neg, LC–MS C8-pos	329	228

[a] Number of Reactome pathways accessible corresponds to the number of pathways in each dataset which contained at least two metabolites assayed in the metabolomics data

(release 101) were downloaded using the KEGG REST API (https://www.kegg.jp/kegg/rest/keggapi.html).

We describe a pathway by the set of metabolites annotated to it: $p = \{m_1, m_2, \ldots m_L\}$ for a pathway of size L. As part of the benchmarking procedure a non-redundant set of Reactome pathways $N = \{p_1, p_2, ..., p_N\}$ was created along with the associated set of all metabolites in these pathways $M = \{m_1, m_2, ...m_M\} = p_1 \cup p_2 \cup ... \cup p_N$. N and M were obtained by iterating through the list of pathways P (k=255 with at least 2 metabolites present profiled in the COVID dataset, in original order) and adding pathway p_i to N if there was no overlap with the current non-redundant set, $|p_i \cap M| = 0$. The resulting set contained a total of 18 pathways with no overlapping metabolites, with coverage ranging from 2 to 6 metabolites per pathway. We note that other non-redundant pathway sets are possible but do not expect results to be strongly dependent on the exact set used.

Creation of semi-synthetic metabolomics data

The simulations in this work are based on semi-synthetic datasets created using the COVID dataset. The use of a "permutation and spike" procedure allows the original signal in the dataset to be erased and replaced with known pathway signals, while preserving the underlying distributions (both joint and marginal) and heterogeneity of the experimental data. This is advantageous compared to generating fully synthetic data such as by random sampling from a Gaussian distribution, as it preserves the complex biological relationships occurring in omics data, which will influence method performance. A limitation of our approach is that the underlying distribution of a particular dataset may differ from that of other datasets. Despite this limitation, we believe this approach reflects more accurately the level of complexity found in real data compared to fully-synthetic simulation approaches.

The semi-synthetic metabolomics data generation process is illustrated schematically in Fig. 2 and outlined as follows.

Let $X_{i,j}$ denote the original log_2-transformed data matrix composed of n samples and m metabolites. X contains two groups of samples, $i \in A$, representing the control group and $i \in B$, representing the disease group. For simplicity we have simulated a two-group design, but this simulation setup can be extended to multiple groups, continuous outcomes or more complex experimental designs. To remove the original signal in the data, the class labels of each sample A or B were randomly shuffled at each realisation of the simulated data (Fig. 2a).

A pathway collection is required, consisting of $P = \{p_1, p_2, ..., p_i\}$ Reactome or KEGG pathways and the associated set of all metabolites in these pathways

(See figure on next page.)

Fig. 2 Schematic outlining the benchmarking process used in this work. **a** Sample group labels are randomly permuted. **b** An artificial signal is added (α) to all metabolites in M_k (set of all metabolites within the k = 3 randomly selected pathways (E)) only to samples in group B. **c** Single-sample pathway analysis is performed on the semi-synthetic data matrix generated in step b) using the various methods benchmarked. **d** Independent two-sample t-tests are performed for each pathway (p) analysed using ssPA, to provide *p*-values for the significance of the different in pathway score means between samples in group A and group B. **e** Benjamini–Hochberg adjusted q-values for each pathway p are used to compute performance metrics, taking account of the level of overlap between the pathway p and M_k, as well as whether p is a member of the set E (enriched pathways). Considering the q-value and overlap threshold, each pathway can be classed as a true positive, false positive, false negative, or true negative, and used to compute the performance metrics: precision, recall, and area under the curve (AUC)

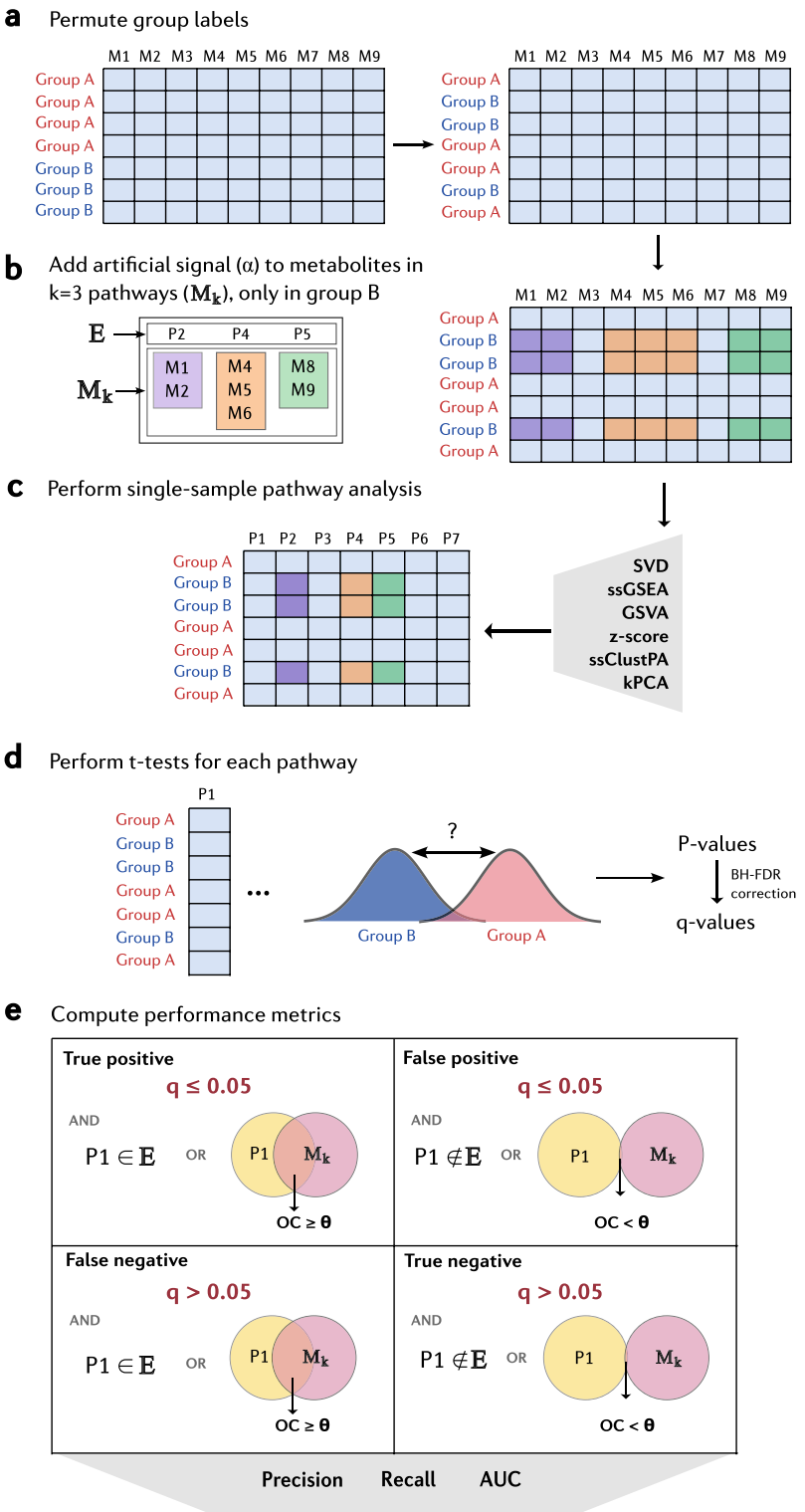

Fig. 2 (See legend on previous page.)

$M = \{m_1, m_2, ...m_M\} = p_1 \cup p_2 \cup ... \cup p_i$. Following the sample class label permutation, $k = 3$ pathways $E = \{p_1, p_2, ..., p_k\}$ were chosen at random from P to be "enriched", corresponding to metabolites $M_k = p_1 \cup p_2 \cup ... \cup p_k$. The abundance values samples in

group B were altered (by adding a constant α to the original log space values) for metabolites in set M_k (Fig. 2b). Note this represents a multiplicative (fold change) effect in the original non-log space. The values for group A were left unchanged. The value of α represents the strength of the enrichment of the metabolites within p_i. In our experiments we enriched $k = 3$ pathways. The simulated data matrix Y can thus be expressed as:

$$Y_{i,j} = X_{i,j}, \quad i \in A \tag{1}$$

$$Y_{i,j} = X_{i,j} + \alpha, \quad i \in B \tag{2}$$

where $j \in M_k$ and $\alpha \in [0, 1]$.

ssPA implementation details

Details of the Python/R packages used for each of the methods benchmarked in this work are given in Additional file 1: Table S1. The ssPA package is available to download from the Python Package Index at https://pypi.org/project/sspa/, and the source code is freely available at https://github.com/cwieder/py-ssPA.

ssClustPA method

Additional file 1: Fig. S1 shows an overview of ssClustPA and kPCA methods. The ssClustPA method is appropriate when we expect two clusters in the data (e.g. case and control). ssClustPA projects each sample onto the vector separating the two clusters, and is outlined as follows:

1. For each of the P pathways $p_k = \{m_1, m_2, ..., m_p\}$, create a pathway matrix Z_k composed of the abundance data corresponding to the pathway, $Z_{ijk} = X_{ij}$ where $m_j \in p_k$. Here there are M_k metabolites in pathway k.
2. On each pathway matrix Z_k, perform k-means clustering based on Euclidean distance with k = 2 clusters.
3. Obtain the coordinates of the cluster centroids c_1&c_2 in the m-dimensional space defined by Z_k.
4. For each pathway obtain the unit column vector between the two cluster centroids $u = c_1 - c_2$. The dimension of u will correspond to the number of metabolites in the dataset mapping to that pathway, M_k.
5. Project the pathway matrix Z_k (dimension n × M_k) onto u (dimension M_k × 1). The projected values can be used as pathway scores A_k (Eq. (3)).

$$A_k = Z_k u \tag{3}$$

6. Repeat steps 4–5 for each pathway matrix Z_k and concatenate the resulting vectors A_k horizontally to produce the pathway score matrix $A_{n \times P}$.

The projection of each pathway matrix Z_k onto the vector u will capture the alignment of the samples with the clustering in the pathway space with more accuracy than simply using Euclidean distances to a single cluster centroid. The latter would be unable to distinguish a sample datapoint's proximity to the first cluster centroid relative to the second centroid. The vector u represents the direction between the two cluster centroids,

capturing the main difference between the two clusters. By projecting the pathway matrix onto u we capture variation related to the difference between clusters, and ignore variation which is orthogonal (uncorrelated) to this difference, thus providing a score which is more relevant to the contrast studied.

kPCA method

The kPCA method is outlined as follows:

1. For each of the P pathways $p_k = \{m_1, m_2, ..., m_p\}$, create a pathway matrix Z_k composed of the abundance data corresponding to the pathway, $Z_{ijk} = X_{ij}$ where $m_j \in p_k$.
2. For each pathway matrix Z_k, perform kernel PCA [42] with a radial basis function (RBF) kernel. The kernel width parameter γ is set to a default of $\frac{1}{n}$.
3. The scores of the first principal component (PC1) can directly be used as scores A_k.
4. Repeat step 2 for each pathway matrix Z_k and concatenate the resulting vectors A_k horizontally to produce the pathway score matrix $A_{n \times P}$.

Benchmarking details

Pathway overlap

Overlap between pathway metabolites was calculated using the Szymkiewicz-Simpson Overlap Coefficient (OC) (Eq. 4). An OC of 0 indicates there is no overlap between set A and set B, whereas an overlap of 1 indicates that the smaller set is a subset of the larger set.

$$OC(A, B) = \frac{|A \cap B|}{min(|A|, |B|)} \tag{4}$$

We used the OC as an alternative to the Jaccard Index (JI) as it is more sensitive to overlapping metabolites, i.e., the OC will return a value of 1 when the metabolites in a pathway are all present in a larger pathway (irrespective of the size of the larger pathway), whereas the JI will only equal 1 if two sets are identical.

Performance metrics

As outlined in the section 'Creation of semi-synthetic metabolomics data', in each of the simulations in this work, $k=3$ randomly selected pathways $E = \{p_1, p_2, p_3\}$ were defined as enriched (Fig. 2b). In the effect size simulations, all metabolites M_k in the set E had a constant $\alpha \in [0, 1]$ added to the value of those metabolites in group B, representing the disease class. Adding a constant of 1 to a metabolite in the log space changes its abundance value by a fold change (FC) of 2 ($log_2 FC = 1$). In the signal strength simulation, varying percentages $s \in [0, 100]$ of randomly selected metabolites in E had a constant $\alpha = 1$ added to them, again only in group B.

Pathway scoring using ssPA was performed on the simulated data abundance matrix (Fig. 2c), and a series of independent two-sample t-tests between simulated disease and control groups were used to determine the significance of each pathway (Fig. 2d). The Benjamini Hochberg false discovery rate correction was applied and pathways with q ≤

0.05 were considered significantly enriched. Of these pathways, those that were members of E or those with an OC $\geq \theta$ between a pathway p_i and the set M_k (the set of all metabolites in E) were considered positives (Fig. 2e, Eq. 5). We use the OC to define positives, as pathways sufficiently overlapping with the artificially enriched ones can also be considered to be truly enriched. We tested two OC thresholds of $\theta \in \{0.25, 0.5\}$. Pathways with q > 0.05 were considered true negatives if the pathway is not a member of E and has OC $< \theta$ (Eq. 5). Those pathways with q > 0.05 that are members of E or have OC $\geq \theta$ are considered false negatives (Eq. 5).

$$\begin{aligned} &\text{For a given pathway } p: \\ &p \text{ is a true positive (TP) if}[p \in E \text{ or } \mathrm{OC}(M_k, p) \geq \theta], \text{ and } q \leq 0.05 \\ &p \text{ is a true negative (TN) if}[p \notin E \text{ and } \mathrm{OC}(M_k, p) < \theta], \text{ and } q > 0.05 \\ &p \text{ is a false positive (FP) if}[p \notin E \text{ and } \mathrm{OC}(M_k, p) < \theta], \text{ and } q \leq 0.05 \\ &p \text{ is a false negative (FN) if}[p \in E \text{ or } \mathrm{OC}(M_k, p) \geq \theta], \text{ and } q > 0.05 \end{aligned} \tag{5}$$

The sklearn.metrics functions were used to calculate recall, precision, and area under the receiver-operating curve (AUC). All simulations were repeated 200 times, with randomly selected enriched pathways and semi-synthetic data randomly permuted at each realisation.

Method runtime profiling

Runtime profiling was repeated 10 times for each method and average results are reported. The Python libraries cProfiler and line-profiler were used to determine the wall time of each method.

IBD application

The kPCA method was used to demonstrate ssPA applied to the IBD metabolomics dataset.

Hierarchical clustering

kPCA pathway scores were standardised prior to clustering ($\mu = 0$ and $\sigma = 1$). Hierarchical clustering was performed using the scipy cluster.hierarchy function with Euclidean distance and Ward linkage parameters. The maxclust parameter was set to 2 in order to cut the tree at a depth of 2 branches. The adjusted Rand index [43]was calculated using the metrics.adjusted_rand_score function, with the original and predicted cluster sample labels as input.

Pathway score correlation network

kPCA pathway scores derived from the IBD dataset were ranked by t-test p-values (testing for differences between IBD and non-IBD groups) and the top 50 pathways were used to produce a hierarchical clustering as detailed above. Pathway IDs from the resulting clusters were extracted and used to build the network. A Spearman rank correlation matrix was produced from the pathway scores. The NetworkX Python package was used to create a pathway-pathway network, with nodes representing pathways and edges representing correlation between the pathway scores. Cytoscape was used to visualise the network using an edge-weighted spring embedded layout.

The COVID pathway correlation network was created in the same way as the IBD network, using the top 30 pathways (ranked by t-test p-values) for visual clarity. Each node is coloured by the average pathway score of samples in each COVID WHO severity level (0, 1–2, 3–4, 5–7).

Results

Performance evaluation of ssPA methods

We first carried out a comprehensive benchmarking of ssPA methods applied to semi-synthetic metabolomics datasets generated based on the Su et al. COVID dataset [38], see Methods. The Reactome pathway database was used as the main source of pathways throughout this work, with key simulations reproduced using the KEGG database. Where applicable, we also compared ssPA results to those obtainable using conventional PA methods ORA and GSEA.

Outline of simulation procedure

In order to calculate various performance metrics (i.e. precision and recall), it is essential to know the identity of truly perturbed “enriched" pathways. Here we use the terms “positively enriched" or “negatively enriched" to refer to pathways that are significantly perturbed with respect to the control group, depending on the directionality of the effect. In experimental datasets it remains challenging to distinguish between true and false positive enriched pathways, so instead we used semi-synthetic data to accomplish this task. As detailed in the methods, we removed the original signal from an experimental untargeted metabolomics dataset of COVID patients and replaced it with artificial known pathway signals (enriched pathways) in one of two study groups. Using this simulation procedure, we can vary the strength of the enrichment of a pathway, the number of differentially abundant metabolites in the pathway, and the coverage level.

ssClustPA and kPCA: pathway scoring approaches based on unsupervised learning

We propose two novel ssPA methods based on unsupervised machine learning concepts for generating pathway scores (see Methods for further details). ssClustPA makes use of k-means clustering and for each pathway projects the original datapoints onto the unit vector between two cluster centroids to yield pathway scores. ssClustPA is best suited to two-group study designs due to the two-cluster centroid constraint; for designs with two or more study groups, we developed the kPCA ssPA method. The kPCA ssPA method uses kernel PCA [42] with a radial basis function kernel to model the distribution of data for the pathway and is particularly advantageous when the underlying manifold of the data is non-linear. The first principal component scores are used as pathway scores. Both methods are applicable to any omics datatype.

The importance of pathway overlap in method evaluation

Biological pathways represented as lists of molecules can be seen as an arbitrary way of partitioning a network and defining precise pathway boundaries remains a challenge in pathway curation. It is therefore expected that all pathway databases will contain some degree of redundancy, meaning that not all molecules will be unique to a single pathway. If a pathway p is significantly enriched, it is likely that other pathways that contain

overlapping molecules with p will also be considered enriched to varying degrees. When performing benchmarking on redundant pathway sets, one must therefore decide the level of pathway overlap that constitutes a true positive pathway.

A simple example of the effect of pathway overlap is demonstrated in Fig. 2. Here we performed a simulation in which a single randomly selected pathway "R-HSA-1483206" (Glycerophospholipid biosynthesis) was enriched at effect size $\alpha = 1$. Using the pathway scores derived from each ssPA method, a series of independent two-sided t-tests were performed for each of the 255 Reactome pathways, testing for significant differences in mean pathway scores between the two study groups. The resulting pathway p-values are compared to the overlap coefficient (OC) of each pathway with pathway R-HSA-1483206 in Fig. 3. We observe that regardless of the ssPA method, there is a clear correlation between pathway overlap and p-value, with pathways with higher overlap tending to have lower p-values. Importantly, if an arbitrary p-value threshold e.g. $p \leq 0.05$ was used to select significant pathways in this manner, R-HSA-1483206 and a number of additional pathways would be considered enriched. It is important to note this effect is not only observed with ssPA methods, but also in conventional methods such as GSEA (Fig. 3).

Due to this overlap effect, and to simplify the procedure, we first performed benchmarking on a small subset of non-redundant pathways from the Reactome pathway database (k=18). The key motivation for using this non-redundant pathway set was to introduce readers to our benchmarking strategy prior to taking into account the effect of pathway overlap. Furthermore, the utility of non-redundant pathway sets has been previously demonstrated [44, 45] and the results of benchmarking on such sets may be of interest in themselves to some readers. We then repeated the performance evaluation

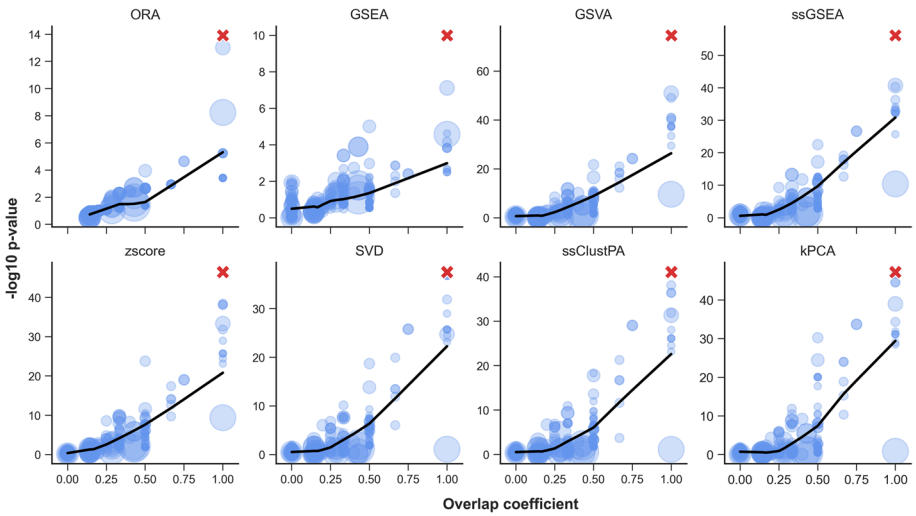

Fig. 3 The relationship between pathway overlap and significance. The artificially enriched pathway (R-HSA-1483206) is represented by a red cross. Overlap coefficient values of each pathway with R-HSA-1483206 are shown on the x-axis. t-test p-values of pathway scores (testing for significant difference in mean pathway score between disease and control group) for each pathway are shown on the y-axis (-log10 scale). Note, for GSEA the p-values are calculated using the original GSEA permutation procedure, and for ORA, p-values are calculated using the Fisher's exact test. A LOWESS regression line is shown in black. Point size corresponds to the coverage of each pathway (i.e. the number of metabolites in the pathway which were present in the dataset)

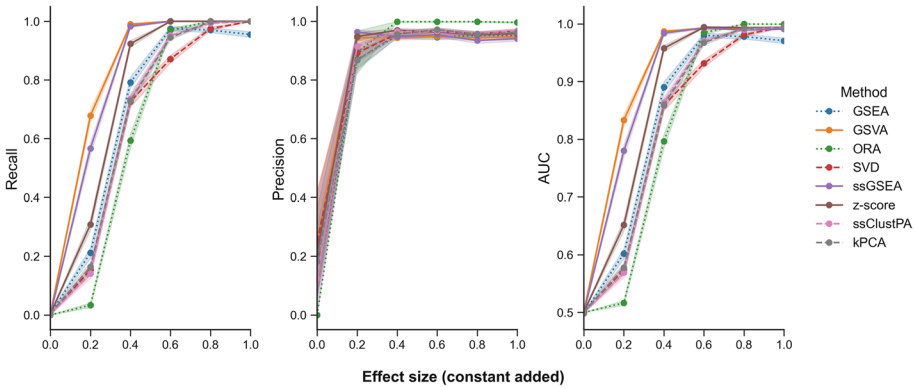

Fig. 4 Performance of ssPA methods using a non-redundant pathway set based on 3 randomly selected enriched pathways. For each panel Recall, Precision, and AUC, the x-axis represents the effect size of the simulated pathway enrichment. Points represent the mean performance metric over 200 iterations. Note for AUC, y-axis values range from 0.5 to 1. Shaded intervals indicate the standard error of the mean. Dotted lines represent conventional pathway analysis methods, and dashed lines represent clustering/dimensionality-reduction based methods

on the full set of redundant Reactome pathways (k=255 with sufficient coverage in the COVID dataset), which although containing overlapping pathways, is a more realistic and complex scenario. To summarise, we make use of the level of overlap between the enriched pathways and the rest of the pathways to define the set of true positive enriched pathways (see Methods for formal definition), which aims to take into account the arbitrary separation of the metabolic network into pathway sets.

Performance evaluation of ssPA methods

In this section we evaluated the performance of the SVD (PLAGE), ssGSEA, GSVA, z-score, ssClustPA, and kPCA methods using semi-synthetic data. These are categorised as GSEA-based (ssGSEA and GSVA) and dimensionality reduction (DR)/clustering-based (SVD, ssClustPA, and kPCA). We also included comparison to non-ssPA methods ORA and GSEA where possible.

Beginning with the non-redundant pathway set, in which each of the pathways contains unique metabolites, three random pathways were artificially enriched, and performance metrics were computed and averaged over 200 iterations (Fig. 4) across a range of increasing effect sizes. The GSEA-based methods GSVA and ssGSEA appear to outperform the other methods in terms of recall and AUC at low to moderate effect sizes (0.2–0.6). At higher effect sizes of 0.8–1 (corresponding to $FC \approx 2$), all methods appear to perform equally well with average recall, precision and AUC values > 0.9.

We then used the full set of 255 Reactome pathways (containing redundancy) with the same simulation setup to evaluate performance (Fig. 5). Again, three random pathways are artificially enriched. As mentioned previously, when calculating performance metrics using overlapping pathways, one must define a threshold for true positive pathways. In this case, we pooled all metabolites from the simulated enriched pathways together and calculated an OC between this set and each pathway tested. Figure 5a shows the performance metrics where a positive corresponds to OC $\theta \geq 0.25$, whereas Fig. 5b shows the performance metrics where a positive corresponds to OC $\theta \geq 0.5$. For

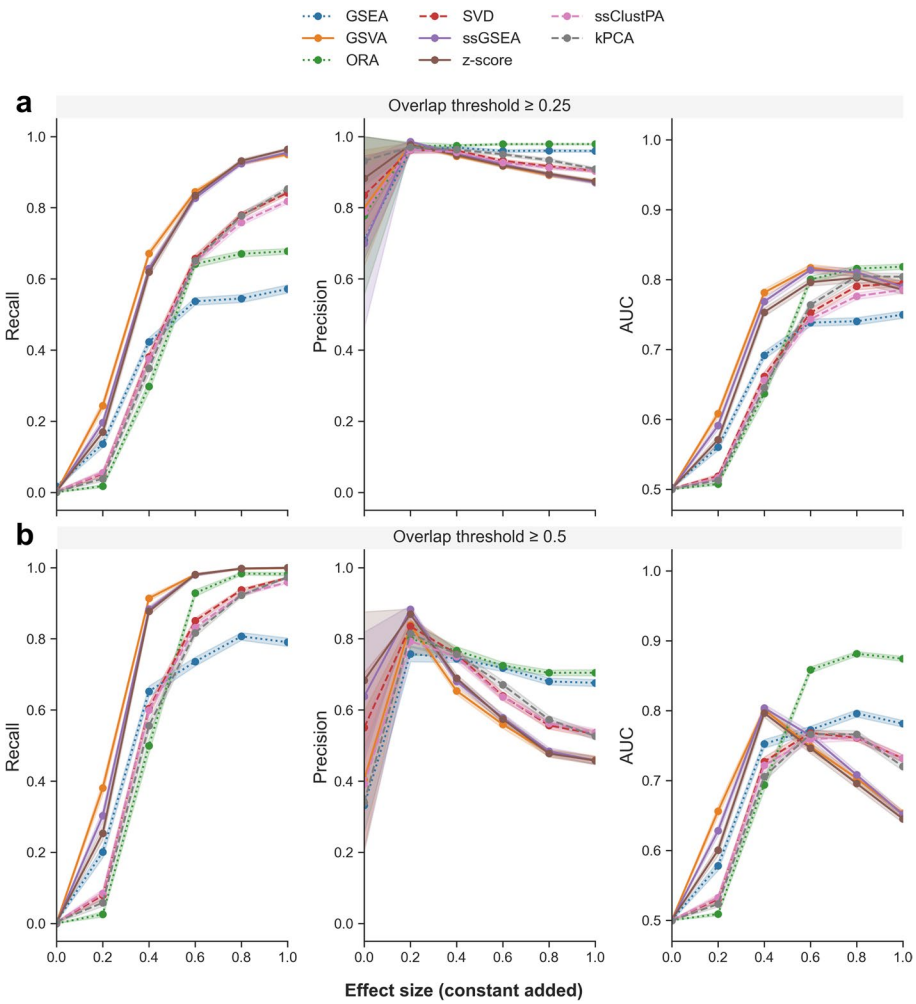

Fig. 5 Performance of ssPA methods on the full pathway set (including redundancy) based on 3 randomly enriched pathways. **a** Top panel OC ≥ 0.25, **b** bottom panel OC ≥ 0.5. All metabolites in enriched pathways have identical effect size. Points show performance metrics averaged across 200 iterations. Shaded intervals represent average SEM. Dotted lines represent conventional pathway analysis methods, and dashed lines represent clustering/dimensionality-reduction based methods

all effect sizes in Fig, 5, we computed pairwise Mann–Whitney U tests across methods (Bonferroni corrected) to determine the statistical significance of the differences in performance metrics (Additional file 2). In both Figs. 5a and 5b, the GSEA-based methods (ssGSEA and GSVA) and z-score have the highest recall across all effect sizes ($p < 0.05$). In terms of precision, regardless of the overlap threshold, conventional pathway analysis methods ORA and GSEA outperform ssPA methods at moderate-high effect sizes (0.6–1.0) ($p < 0.05$). Following this, DR/clustering-based methods (kPCA, ssClustPA, and SVD respectively) provide the highest precision values amongst the ssPA methods, with GSEA-based methods ssGSEA and GSVA as well as the z-score yielding the lowest precision values ($p < 0.05$ at effect sizes 0.8–1 for OC $\theta > 0.25$, $p < 0.05$ at effect sizes 0.4–1 for OC $\theta > 0.5$). Similar observations can be made in terms of AUC at an overlap threshold of ≥ 0.5 (Fig. 5b). The GSEA-based methods and z-score yield the highest AUC at lower effect sizes (0.2–0.4), but at higher effect sizes (0.8–1), ORA and GSEA

outperform all ssPA methods, followed by the clustering/DR-based methods, and finally GSEA/z-score-based methods which have the lowest AUC values ($p < 0.05$). At an overlap threshold ≥ 0.25 (Fig. 5a), these trends are more subtle in terms of AUC, but as effect sizes become larger there is a shift in performance from GSEA-based/z-score methods to clustering/DR-based methods.

GSEA-based methods, as well as the z-score, are the poorest performers in terms of precision. The drop in precision from effect size 0.2 onwards is likely due to an increase in the number of false positives resulting from the stronger effect size which do not reach the overlap threshold to be considered true positives. There is a larger probability that overlapping pathways are detected as significantly enriched. This drop in precision is less evident when the OC for identifying true positive pathways is lower, such as in Fig. 5a, as a higher proportion of the overlapping pathways are considered true positives, resulting in fewer false positives. Compared to conventional PA methods, ssPA methods generally have improved recall, particularly at larger effect sizes. However, conventional methods tend to have a higher precision than ssPA methods, regardless of effect size. At higher effect sizes, ORA generally appears to have higher power to detect significant pathways than GSEA. Additionally, we reproduced very similar results using a different semi-synthetic dataset based on IBD data from Lloyd-Price et al. [39] (Additional file 1: Fig. S2). The sole difference was that kPCA had higher precision than all other methods at the higher OC threshold, regardless of effect size.

Finally, we ran the same set of simulations using the KEGG human pathway database rather than the Reactome database (Additional file 1: Fig. S3). Despite there being subtle changes in the performance metric results, which are to be expected due to varying pathway composition and size, the ranking of the methods across metrics remained consistent with that observed in the Reactome simulations.

The effect of pathway signal strength on ssPA performance

In the previous simulations, we varied the effect size of the pathway enrichment, with the abundance of all metabolites in the pathway modified to the same extent. In a real-world scenario, it is highly unlikely that all metabolites in a differentially active pathway would have the same effect size. We therefore performed another simulation in which the abundance of varying proportions of randomly selected metabolites in a pathway was modified while keeping the effect size at a constant of 1.0. In this scenario, we consider pathways with $q \leq 0.05$ and $OC \geq 0.5$ true positives and randomly enriched 3 pathways in each iteration.

As expected the performance of all methods improves as the signal strength increases (Fig. 6), aside from a small drop in precision which is caused by the increase in signal strength rendering overlapping pathways more significant, as highlighted in the previous simulation (Fig. 5). In terms of recall, all ssPA methods perform very similarly across signal strength sizes, and clearly outperform the conventional PA methods ORA and GSEA. In contrast, the conventional methods slightly outperform the rest of the ssPA methods in terms of precision, of which the dimensionality-reduction based methods (SVD, ssClustPA, and kPCA) have higher precision than GSEA/z-score-based methods. Overall, when taking into account varying levels of signal strength, and that not all

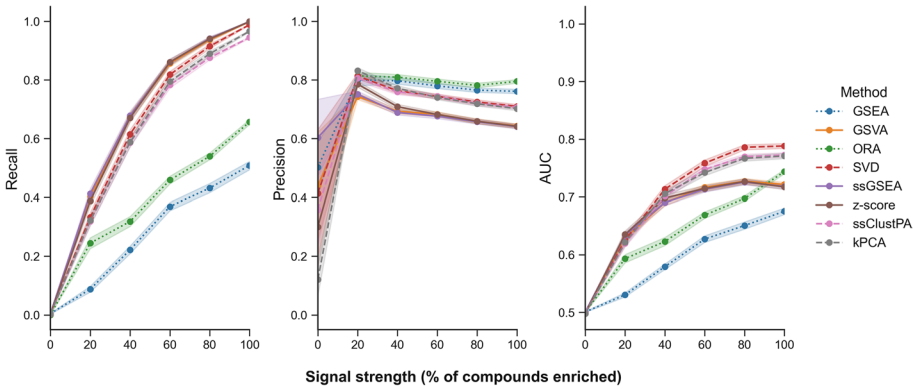

Fig. 6 Effect of varying signal strength on ssPA method performance. Points show mean performance over 200 iterations with 3 randomly enriched pathways on the full Reactome pathway set (true positive pathways are those with an OC ≥ 0.5). Shaded intervals represent average SEM. Dotted lines represent conventional pathway analysis methods, and dashed lines represent clustering/dimensionality-reduction based methods

molecules in an enriched pathway are likely to be significantly differentially abundant, clustering/DR-based methods appear to offer the best overall performance in terms of ssPA, particularly at moderate to high effect sizes (FC $\approx \geq 1.5$).

ssPA method ability to rank enriched pathways highly

The ability of the ssPA methods to rank enriched pathways highly was investigated, again using the semi-synthetic COVID data. Briefly, t-tests were performed on the ssPA scores for each pathway (testing for a difference between mean scores in disease and control groups) and the resulting *p*-values were used to rank them in ascending order. The full set of 255 Reactome pathways was used. In order to account for the fact that ORA generally tests fewer pathways than the other methods (as it only calculates *p*-values for pathways with at least one differentially abundant metabolite), the ranks were normalised by the total number of pathways tested using each method.

In concordance with the performance metrics calculated in the previous section, at lower effect sizes, the GSEA/z-score-based methods rank the truly enriched pathways the highest (Additional file 1: Fig. S4). At higher effect sizes (0.8–1.0), there is very little difference in the rankings of the enriched pathways by all ssPA methods, with all methods being able to rank the 3 enriched pathways within the top 10% of pathways.

The effect of pathway coverage on method performance

It is well established that metabolomics assays generally profile a smaller fraction of the total metabolome than the fraction of the genome/transcriptome captured by sequencing technologies. We compared the level of Reactome human pathway coverage in metabolomics and transcriptomics datasets from the same study [39] (Fig. 7), and as expected there were far fewer metabolites mapping to pathways than transcripts. This well-known issue motivated us to examine how the ssPA methods performed in response to varying levels of pathway coverage. Two Reactome human pathways were selected as exemplars ('SLC-mediated transmembrane transport' and 'Biological oxidations'), for their high coverage in the original COVID dataset (39 and 26 metabolites respectively) and different metabolite composition (OC = 0.27). For each pathway we

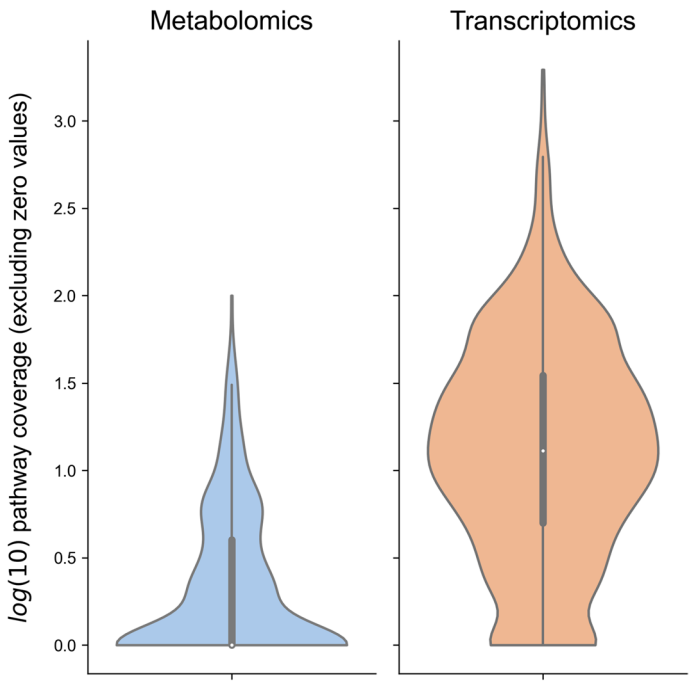

Fig. 7 Comparison of Reactome human pathway coverage in metabolomics versus transcriptomics data from the same study (IBD data, Lloyd-Price et al. [39]). Violin plots show $\log_{10}$ transformed distributions of the number of metabolites/genes mapping to each Reactome pathway

randomly selected varying proportions of metabolites in each pathway to remove from the data. Doing so, we simulate reduced coverage in these pathways, from 100% relative coverage (all of the original metabolites present) to 20% relative coverage (80% original metabolites deleted).

We calculated performance metrics as in the previous simulations, keeping the minimum overlap coefficient threshold to declare true positive pathways fixed at OC = 0.5, and demonstrating two effect size scenarios (low effect at $\alpha = 0.25$ and moderate effect at $\alpha = 0.75$). For both pathways (Additional file 1: Figs S5 and S6), we observed similar rankings in PA methods across all three performance metrics as in the effect size simulations (Fig. 5). As expected, both recall and precision gradually increased as relative coverage of the pathway increased. While AUC appears to decline with increased coverage (particularly at the higher effect size), which may seem counterintuitive, this can be attributed to an increase in the false positive rate. As coverage increases, more pathways share an overlap with the simulated enriched pathway and therefore attain significant *p*-values, but do not have a high enough OC to be considered true positives, hence resulting in a rise in false positives and a decreased AUC. Importantly, we found that higher coverage generally improves method performance, but does not have profound effects on the rankings of method performance, which recapitulate those observed in the previous simulations investigating effect size and signal strength.

Method runtimes

Each ssPA method was run 10 times on a laptop with standard hardware and 16 GB of RAM using the COVID dataset, which contained 263 samples (rows) and 335

metabolites (columns), as well as Reactome pathways with sufficient coverage in the dataset (k=255). In terms of runtime, the fastest method was the z-score, followed by SVD (Additional file 1: Table S2).

Application of ssPA in metabolomics: a case-study of inflammatory bowel disease

We then focused on the application and interpretation of ssPA in typical metabolomics data, using real experimental datasets. In order to simplify our presentation, we demonstrated the use of a single method, kPCA, to showcase one of our newly proposed methods, although the following application is generalisable to any ssPA method.

In order to demonstrate some of the potential use cases of ssPA methods applied to metabolomics data, we selected a different dataset to that used in the benchmarking portion of this work and made use of ssPA to help detect and interpret complex biological patterns within the data. The untargeted metabolomics data used in this case study was derived from a multi-omics study of inflammatory bowel disease (IBD) by Lloyd-Price et al. [39]. It is important to note that the group termed “non-IBD" are not necessarily healthy individuals and may still exhibit symptoms of IBD. Further details of this dataset can be found in the Methods and the original article [39].

To obtain an exploratory perspective on the strength of the biological signals in the dataset, PCA was used to visualise the data at both the metabolite and pathway level. We transformed the metabolite-level data to pathway scores using the kPCA method described earlier and compared the PCA results obtained. No strong separation between the sample classes was observed using either approach, with most samples appearing amalgamated together in a central cluster (Additional file 1: Fig. S7). This is consistent with findings from the original work, in which intra-individual variation in the IBD metabolome was greater than inter-individual variation [39]. However, the percentage variance explained in the first two principal components of the data was higher using pathway scores than individual metabolites (Additional file 1: Fig. S7).

In order to more rigorously quantify the clustering performance achieved using metabolites (only those present in pathways) vs. pathway scores, we ranked each of the metabolites/pathways using BH-FDR corrected t-test q-values (testing for differences between the IBD and non-IBD groups) and performed clustering after progressively adding each entity significant at $q \leq 0.05$ (Fig. 8). Hierarchical clustering was used to partition the samples into groups based on the two main branches of the tree, which were compared to the original sample labels (IBD or non-IBD) using the Adjusted Rand Index (ARI) [43]. Note that negative ARI values can be obtained if the RI is less than expected by chance. This heuristic method suggested that for this particular dataset, using pathway scores for clustering generally achieves higher ARI across a range of different cut-off thresholds, as well as improved robustness to the threshold used for clustering, than by using metabolites alone. Consequently, we made use of the clustering applied to the top 50 sets of pathway scores to identify clusters of pathways that could discriminate between the IBD sample classes. Two distinct pathway clusters were evident (Additional file 1: Fig. S8), which we visualised using Cytoscape in Fig. 9.

The networks shown in Fig. 9 represent two distinct groups of pathways that discriminate between IBD and non-IBD samples. An edge-weighted spring-embedded layout was applied to the network to visually group nodes by pathway score Spearman correlation.

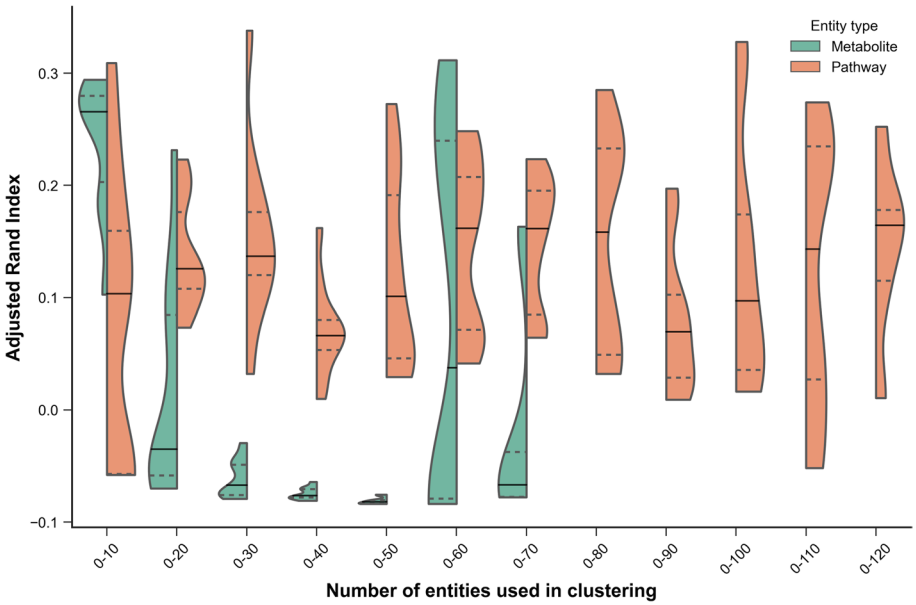

Fig. 8 Clustering performance at the metabolite and pathway levels (kPCA method) for different selection thresholds. Clustering performance is quantified using the Adjusted Rand Index (y-axis). The number of entities used to perform the clustering are shown on the x-axis, and only entities significant at $q \leq 0.05$ were included in the analysis. A total of 69 metabolites and 116 pathways were significant and are shown in the violin plots. Entities were ranked by q-value and only those within each selection threshold are used for each comparison. Violin plots show the distribution of ARI values achieved using the cumulative thresholds (e.g. the top 10 pathways). Solid lines represent the median ARI, whereas dashed lines represent the upper and lower quartiles. The violin plot range is truncated to the range of the data

The network nodes are coloured by the average pathway score across the samples in each subtype (from top to bottom: non-IBD, CD, UC). Using the kPCA approach (or clustering/DR-based approaches in general), it is not possible to make a direct link between the sign of the pathway scores obtained and the direction of the pathway enrichment. However, GSEA-based ssPA approaches can be used to determine this, if desired. GSVA was used to determine that cluster 1 of Fig. 9 represents pathways whose metabolites are upregulated in IBD relative to the non-IBD group, whereas cluster 2 represents pathways with metabolites depleted in IBD relative to the non-IBD group.

IBD encompasses a group of diseases broadly characterised by chronic inflammation of the gut, and is widely attributed to a complex interplay of immune dysregulation, dysbiosis of the gut microbiome, and genetic and environmental factors. A sub-cluster of immune-related pathways can be seen in cluster 1 (Fig. 9a), highlighting the role of phospholipids and Fc-gamma receptors in phagocytosis, which is concordant with findings in current literature [46, 47]. Changes in phospholipid metabolism are strongly implicated in IBD pathology, particularly in degradation of the intestinal epithelial mucus layer [47]. Another pathway associated with maintaining the integrity of the gut mucosal barrier highlighted in this cluster is 'Metabolism of polyamines' (Fig. 9b), which corroborates the work of Weiss et al. [48], who found increased levels of the polyamine spermidine alongside decreased levels of spermine in IBD patients. One major benefit of ssPA is that it enables comparison of pathway scores across multiple study groups. Using the 'Metabolism of polyamines' pathway as an example, we note that enrichment of the pathway

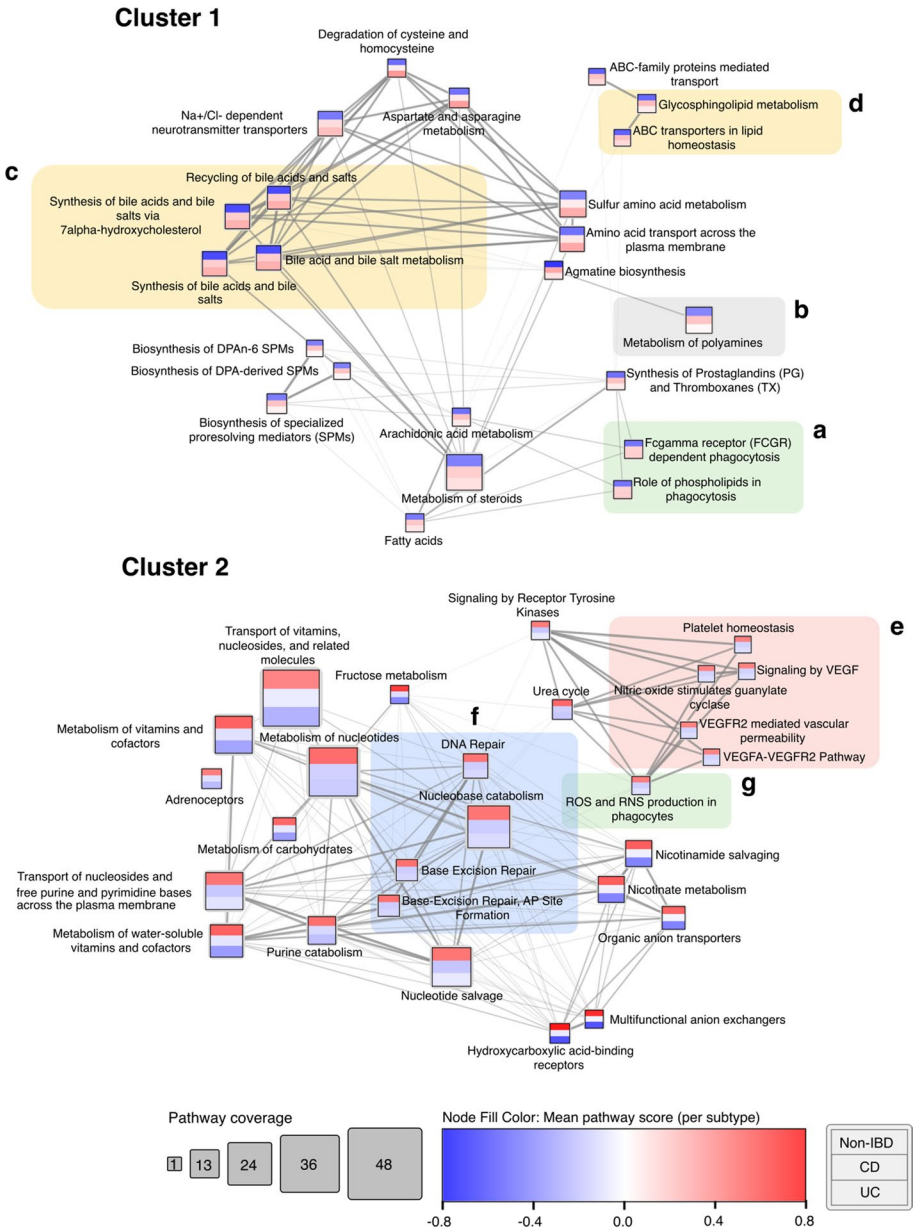

Fig. 9 Pathway cluster network derived using hierarchical clustering on IBD pathway scores (kPCA, top 50 pathways). Coloured bands in each node represent mean pathway score within each subtype (from top to bottom: non-IBD, CD, UC). Edge weight represents Spearman correlation between pathway scores. Only edges with Spearman's rank correlation coefficient (ρ) ≥ 0.4 are displayed. Shaded sub-clusters depict pathways consistent with current IBD literature and are discussed in the text. Shading of subclusters corresponds to Reactome pathway parent category

is greater in CD patients than it is in UC patients, relative to non-IBD patients. A large sub-cluster of differentially active pathways relates to bile acid and salt metabolism (Fig. 9c), consistent with alterations in primary (e.g. cholate) and secondary bile acid metabolism in IBD observed by Lloyd-Price et al. [39], likely attributable to changes in gut microbial composition, as microbes directly modulate bile acids [49]. Other notable pathways enriched in cluster 1 (particularly in CD patients) include 'Glycosphingolipid

metabolism' and 'ABC transporters in lipid homeostasis' (Fig. 9d), which are congruous with the known dysregulation of lipid metabolism in IBD coupled with inflammation [50–52].

In the second cluster, which consists of pathways whose metabolites are depleted in IBD, a prominent sub-cluster containing pathways related to platelet homeostasis, VEGF, and nitric oxide (NO) signalling is visible (Fig. 9e). It is widely recognised that platelet abnormalities are associated with IBD, alongside complications such as microvascular thrombosis and thromboembolism [53]. The reduction in NO signalling is consistent with findings of reduced NO-mediated vasodilation found in IBD patients [54], which could further contribute to thrombosis. This cluster of pathways focused on VEGF signalling is one example that was not detected as significantly enriched using ORA, but was detected amongst the top pathways using ssPA (see Additional file 1: Table S3 for ORA results).

Another large sub-cluster within cluster 2 focuses on DNA repair processes, including base-excision repair (Fig. 9f). This subcluster can be linked to another significant pathway, 'ROS and RNS production in phagocytes' (Fig. 9g), as reactive oxygen species (ROS) contribute to DNA damage, which in turn induce DNA repair mechanisms [55, 56]. The negative enrichment of these pathways may allude to potential defects or reduction in DNA repair processes in IBD.

We also trained a random forest (RF) classifier on the IBD dataset, which based on fivefold repeated stratified cross-validation achieved comparable AUC using kPCA scores ($\mu = 0.861$, $\sigma = 0.042$) as did the classifier trained on the pathway-annotated metabolites ($\mu = 0.865$, $\sigma = 0.040$). Pathway importances were calculated by permuting each of the features individually, and ranked by the mean decrease in AUC. Many of the top 50 pathways ranked by the RF classifier are shared with those in the networks in Fig. 9 (see Additional file 1: Table S4).

Using the same approach, a pathway-based correlation network was created using the COVID dataset (Additional file 1: Fig. S9). The use of ssPA shows a clear correlation between the WHO status of the samples, corresponding to COVID severity, and the enrichment level of the top 30 pathways shown in the network. Both case studies highlight the value of ssPA methods in quantifying pathway enrichment at an individual sample level, enabling direct comparison of multiple sample sub-groups, as well as in the case of the IBD network, identifying a cluster of enriched pathways related to VEGF/NO signalling that were not identified by the conventional method ORA. Taken together, these case studies demonstrate a small subset of the downstream analyses possible using pathway scores, but highlight the benefits in interpretability and predictive robustness they can achieve.

Discussion

Single-sample PA methods have continued to gain popularity in recent literature as a way to examine pathway signatures at an individual sample level [29, 30, 57, 58]. Unlike conventional PA approaches, which usually compare two groups of samples, ssPA approaches enable researchers to dissect the heterogeneity of complex diseases or response to treatments at an individual level, facilitating advances in precision medicine. Within the transcriptomics field, continuous advances to ssPA methodologies are being

proposed, alongside demonstrated applications on gene-expression data [29, 57–59]. Despite the importance of metabolic pathways in disease and drug-response, there exists very little literature surveying the applicability of ssPA approaches to metabolomics data [60]. The present study was designed to address this gap by providing a critical evaluation of the most widely used ssPA methods and their performance, robustness, and applicability to metabolomics data.

Using semi-synthetic metabolomics data, our benchmarking procedure evaluated several properties of ssPA methods: recall, precision, AUC, and the ability to rank enriched pathways highly. By varying simulation parameters such as effect size and signal strength, we were able to ascertain how the methods performed under these more realistic scenarios. Applied to transcriptomics data, GSVA is often a top performer, particularly in gene-set recall [25, 29]. This finding is consistent with our results, where GSVA was found to have higher recall than all other ssPA methods across lower effect sizes, and at higher effect sizes had similar recall to ssGSEA. GSEA-based methods may provide more power at lower effect sizes as they calculate scores as a function of the metabolites inside and outside of the pathway, testing a competitive null hypothesis, whereas all other methods benchmarked (besides z-score) calculate scores based only on the metabolites within a pathway itself. When effect sizes become larger however, the recall of GSVA was found to be very similar to that of ssGSEA and z-score. In contrast, when identifying correctly the enriched pathways (precision), GSVA was one of the lowest performing methods, as opposed to clustering and DR-based methods ssClustPA, kPCA and SVD. In general, it can be inferred that GSEA/z-score-based methods offer higher recall (most evident at lower effect sizes), whereas clustering/DR-based methods offer improvements in precision, particularly at higher effect sizes, for metabolomics data. In general, we suggest the use of clustering/DR-based methods for datasets with moderate-high effect sizes, and GSVA for datasets with lower effect sizes.

We offer two novel methods for ssPA: ssClustPA and kPCA, which can theoretically be applied to any omics datatype. Leveraging fundamental machine learning concepts of clustering and kernels respectively, both ssClustPA and kPCA have been designed to make use of these unsupervised approaches to discriminate between samples at the pathway level. ssClustPA is particularly advantageous when there is clear structure to the underlying data, exploiting this to provide more precise pathway scores. The RBF kernel used in the kPCA method allows complex non-linear structure within the data to be modelled and projected into a subspace where datapoints become more linearly separable, and hence more likely to provide precise pathway scores, which would otherwise not be feasible using linear separation methods such as PCA or SVD. Applied to metabolomics data, these methods have been shown to offer advances in precision compared to established ssPA methods. Using metabolomics data from IBD samples [39] we employed kPCA to generate pathway scores, and created an IBD subtype-specific pathway network using this data.

Although the objective of this work was not to ascertain whether metabolite or pathway level data yields better performance in downstream analyses, a key question we were able to partially address was whether pathway-level data exhibits greater predictive robustness than its metabolite-level counterpart [28]. Using the IBD dataset, improved clustering performance was observed when using kPCA pathway scores as opposed to

individual metabolites. Not only was the maximum ARI achievable higher using the pathway scores, but the range of ARI scores obtained using a variety of different thresholds used to select which entities to include in the clustering was consistently higher using pathway scores than metabolites. This observation supports the hypothesis that in some cases pathway scores can provide improved robustness to noise, one example of which is the number of pathways used as features in downstream analyses.

We have undertaken the first benchmark into ssPA methods for metabolomics data. The insights gained from this study may provide guidance to practitioners in the field for selecting an appropriate ssPA method, as well as highlighting the broad range of applications for downstream analyses using pathway scores. The semi-synthetic simulation approach we have outlined is not only applicable to metabolomics data, but can be generalised to evaluate a multitude of PA methods across various omics datatypes. Although we used the Reactome and KEGG pathway databases within this work, our approach is easily extensible to other databases e.g. WikiPathways [61]or MetaCyc [21] and we do not expect our results to be highly dependent on the pathway database used. Furthermore, it is common knowledge that the results of metabolomics analysis vary greatly in response to the sample source, assay(s) used, etc. Pathway coverage as well as assay chemical bias remain amongst the key challenges in metabolomics, though advancements in metabolic profiling technologies will undoubtedly ameliorate these in the coming years. Although it was out of the scope of the current work to test ssPA methods on all possible assay types and sample matrices, we found the performance rankings of all methods tested remained robust regardless of the sample source (plasma vs. stool), as well as to changes in pathway coverage.

We aimed to include a range of ssPA methods in our benchmark that are well established in the literature, suitable for use on metabolomics data, while also having programmatic implementations available, however we acknowledge there are additional important ssPA algorithms that have not been included but are highly relevant to the present work [23, 29, 62–64]. Pathifier [23] is one such approach that uses nonlinear principal curve modelling applied to gene expression data to calculate pathway scores by computing the distance from the centroid of control samples to case samples along the curve. The kPCA ssPA approach we developed uses a similar nonlinear dimensionality-reduction based approach, but requires fewer computational resources (wall time and memory) than principal curve modelling. iPath [63] is similar to Pathifier in that it is developed for ssPA analysis of transcriptomic case–control cancer datasets, but instead computes z-scores based on a 'transcriptomic homeostasis' expression profile derived from normal samples which is then used to rank samples as input to an ssGSEA-like algorithm to generate ssPA scores. SLEA [62] also uses z-scores to produce ssPA scores, comparing the mean expression profile of samples within a pathway to that of a null distribution derived from randomised pathway definitions. Furthermore, a class of methods we have not included in this work are topology-based ssPA methods [13]. Further work is required to determine how such methods compare to non-topology-based alternatives, and whether the inclusion of topological information improves performance notwithstanding the low coverage of pathways observed in metabolomics data. Overall, ssPA methods show a great deal of promise both for analysing and interpreting metabolomics data, in addition to the prospect of integrating metabolomics with other omics datatypes.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-022-05005-1.

Additional file 1: Table S1. Method implementation details. **Fig. S1.** Overview of novel methods ssClustPA and kPCA for a single pathway example using simulated data based on the COVID dataset. **Fig. S2.** Performance of ssPA methods using the full pathway set (including redundancy) based on 3 randomly enriched pathways, derived using semi-synthetic data based on the IBD dataset. **Fig. S3.** Performance of ssPA methods using the full KEGG human pathway set (including redundancy) based on 3 randomly enriched pathways, derived using semi-synthetic data based on the COVID dataset. **Fig. S4.** ssPA method ability to rank highly the 3 randomly enriched pathways using the full set of Reactome pathways. **Fig. S5.** ssPA method performance in response to varying levels of pathway coverage. **Fig. S6.** ssPA method performance in response to varying levels of pathway coverage. **Fig. S7.** PCA scatter plots and density plots of PC1 scores obtained using the IBD data at the metabolite (upper panels) and pathway level using kPCA (lower panels). **Fig. S8.** Clustered heatmap of IBD data transformed to pathway scores using the kPCA method. **Fig. S9.** Pathway clusters derived using hierarchical clustering on COVID dataset transformed to pathway scores using kPCA (top 30 pathways). **Table S2.** Runtimes of ssPA methods, alongside GSEA* for comparison to conventional PA methods (average across 10 iterations). **Table S3.** Over-representation analysis results from IBD dataset. **Table S4.** Top 50 features in random forest model based on IBD data.

Additional file 2. Excel spreadsheet containing results of statistical testing of benchmarking results. Mann Whitney U tests were performed for each pairwise combination of methods tested in the effect size simulation (corresponds to Fig. 5 in the main text). *P*-values were adjusted using Bonferroni FWER correction.

Acknowledgements
The authors gratefully acknowledge Fabien Jourdan, Nathalie Poupin, Clément Frainay, and Juliette Cooke for insightful and stimulating discussions regarding this work and for critical reading of the manuscript. We also thank Nathalie Poupin for extensive testing of the sspa Python package.

Author contributions
CW, RPJL, and TE designed the study. CW performed the data analysis and developed the sspa Python package. RPJL and TE supervised the project. CW produced the initial manuscript draft. All authors edited and approved the manuscript.

Funding
This research was funded in whole, or in part, by the Wellcome Trust [222837/Z/21/Z]. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. CW is supported by a Wellcome Trust PhD Studentship [222837/Z/21/Z]. TE gratefully acknowledges partial support from UKRI BBSRC grants BB/T007974/1 and BB/W002345/1, NIH grant 1 R01 HL133932-01. RPJL is funded by the MRC (MR/R008922/1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability
The COVID19 dataset from [38] is available to download from https://data.mendeley.com/datasets/tzydswhhb5/5 (Additional file 1: Table S1). The IBD dataset from [39] is available to download from https://www.hmpdacc.org/ihmp or at MetabolomicsWorkbench (PR000639).

Code availability
Python scripts used to produce the results of this work are publicly available at https://github.com/cwieder/sspa-in-metabolomics. The sspa Python package is available to download from https://pypi.org/project/sspa/ and the source code as well as documentation is available at https://github.com/cwieder/py-ssPA.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 22 July 2022 Accepted: 25 October 2022
Published online: 14 November 2022

References

1. Salem MA, de Souza LP, Serag A, Fernie AR, Farag MA, Ezzat SM, et al. Metabolomics in the context of plant natural products research: from sample preparation to metabolite analysis. Metabolites. 2020. https://doi.org/10.3390/METABO10010037.
2. Odom JD, Sutton VR. Metabolomics in clinical practice: improving diagnosis and informing management. Clin Chem. 2021;67:1606–17. https://doi.org/10.1093/CLINCHEM/HVAB184.

3. Reinke SN, Chaleckis R, Wheelock CE. Metabolomics in pulmonary medicine: extracting the most from your data. Eur Respir J. 2022;60:2200102. https://doi.org/10.1183/13993003.00102-2022.
4. Johnson CH, Ivanisevic J, Siuzdak G. Metabolomics: beyond biomarkers and towards mechanisms. Nat Rev Mol Cell Biol. 2016;17:451–9. https://doi.org/10.1038/nrm.2016.25.
5. Blaise BJ, Correia GDS, Haggart GA, Surowiec I, Sands C, Lewis MR, et al. Statistical analysis in metabolic phenotyping. Nat Protoc. 2021;2021:1–28. https://doi.org/10.1038/s41596-021-00579-1.
6. Nguyen TM, Shafi A, Nguyen T, Draghici S. Identifying significantly impacted pathways: a comprehensive review and assessment. Genome Biol. 2019. https://doi.org/10.1186/s13059-019-1790-4.
7. la Ferlita A, Alaimo S, Ferro A, Pulvirenti A. Pathway analysis for cancer research and precision oncology applications. Adv Exp Med Biol. 2022;1361:143–61. https://doi.org/10.1007/978-3-030-91836-1_8/TABLES/6.
8. García-Campos MA, Espinal-Enríquez J, Hernández-Lemus E. Pathway analysis: state of the art. Front Physiol. 2015. https://doi.org/10.3389/fphys.2015.00383.
9. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: Current approaches and outstanding challenges. Ouzounis CA, editor. PLoS Comput Biol. 2012. https://doi.org/10.1371/journal.pcbi.1002375.
10. Labena AA, Gao YZ, Dong C, Hua H, Guo FB. Metabolic pathway databases and model repositories. Quant Biol. 2018. https://doi.org/10.1007/s40484-017-0108-3.
11. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nat Genet. 1999;22:281–5. https://doi.org/10.1038/10343.
12. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50. https://doi.org/10.1073/pnas.0506580102.
13. Ihnatova I, Popovici V, Budinska E. A critical comparison of topology-based pathway analysis methods. PLoS ONE. 2018;13: e0191154. https://doi.org/10.1371/JOURNAL.PONE.0191154.
14. Marco-Ramell A, Palau-Rodriguez M, Alay A, Tulipani S, Urpi-Sarda M, Sanchez-Pla A, et al. Evaluation and comparison of bioinformatic tools for the enrichment analysis of metabolomics data. BMC Bioinform. 2018;19:1. https://doi.org/10.1186/s12859-017-2006-0.
15. Wieder C, Frainay C, Poupin N, Rodríguez-Mier P, Vinson F, Cooke J, et al. Pathway analysis in metabolomics: recommendations for the use of over-representation analysis. Patil KR, editor. PLoS Comput Biol. 2021;17:e1009105. https://doi.org/10.1371/journal.pcbi.1009105.
16. Karnovsky A, Li S. Pathway analysis for targeted and untargeted metabolomics. In: Li S, editor. Methods in molecular biology. New York, NY: Humana Press Inc.; 2020. p. 387–400. https://doi.org/10.1007/978-1-0716-0239-3_19.
17. Zhang Y, Ma Y, Huang Y, Zhang Y, Jiang Q, Zhou M, et al. Benchmarking algorithms for pathway activity transformation of single-cell RNA-seq data. Comput Struct Biotechnol J. 2020;18:2953–61. https://doi.org/10.1016/j.csbj.2020.10.007.
18. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45:D353. https://doi.org/10.1093/NAR/GKW1092.
19. Jewison T, Su Y, Disfany FM, Liang Y, Knox C, Maclejewski A, et al. SMPDB 2.0: big improvements to the Small Molecule Pathway Database. Nucleic Acids Res. 2014. https://doi.org/10.1093/NAR/GKT1067.
20. Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, et al. The reactome pathway knowledgebase 2022. Nucleic Acids Res. 2021. https://doi.org/10.1093/NAR/GKAB1028.
21. Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res. 2014;42:D459–71. https://doi.org/10.1093/nar/gkt1103.
22. Tomfohr J, Lu J, Kepler TB. Pathway level analysis of gene expression using singular value decomposition. BMC Bioinform. 2005;6:225. https://doi.org/10.1186/1471-2105-6-225.
23. Drier Y, Sheffer M, Domany E. Pathway-based personalized analysis of cancer. Proc Natl Acad Sci U S A. 2013;110:6388–93. https://doi.org/10.1073/pnas.1219651110.
24. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature. 2009;462:108–12. https://doi.org/10.1038/nature08460.
25. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinform. 2013;14:7. https://doi.org/10.1186/1471-2105-14-7.
26. Lee E, Chuang HY, Kim JW, Ideker T, Lee D. Inferring pathway activity toward precise disease classification. Tucker-Kellogg G, editor. PLoS Comput Biol. 2008;4:e1000217. https://doi.org/10.1371/journal.pcbi.1000217.
27. Wang X, Sun Z, Zimmermann MT, Bugrim A, Kocher J-P. Predict drug sensitivity of cancer cells with pathway activity inference. BMC Med Genom. 2019;12:5–13. https://doi.org/10.1186/S12920-018-0449-4.
28. Segura-Lepe MP, Keun HC, Ebbels TMD. Predictive modelling using pathway scores: robustness and significance of pathway collections. BMC Bioinform. 2019;20:543. https://doi.org/10.1186/s12859-019-3163-0.
29. Foroutan M, Bhuva DD, Lyu R, Horan K, Cursons J, Davis MJ. Single sample scoring of molecular phenotypes. BMC Bioinform. 2018;19:404. https://doi.org/10.1186/s12859-018-2435-4.
30. Meng C, Basunia A, Peters B, Gholami AM, Kuster B, Culhane AC. MOGSA: integrative single sample gene-set analysis of multiple omics data. Mol Cell Proteom. 2019;18:S153–68. https://doi.org/10.1074/mcp.TIR118.001251.
31. Chaleckis R, Meister I, Zhang P, Wheelock CE. Challenges, progress and promises of metabolite annotation for LC–MS-based metabolomics. Curr Opin Biotechnol. 2019;55:44–50. https://doi.org/10.1016/J.COPBIO.2018.07.010.
32. Geistlinger L, Csaba G, Santarelli M, Ramos M, Schiffer L, Turaga N, et al. Toward a gold standard for benchmarking gene set enrichment analysis. Brief Bioinform. 2021;22:545–56. https://doi.org/10.1093/bib/bbz158.
33. Nguyen TM, Shafi A, Nguyen T, Draghici S. Identifying significantly impacted pathways: a comprehensive review and assessment. Genome Biol. 2019;20:203. https://doi.org/10.1186/s13059-019-1790-4.
34. Evangelou M, Rendon A, Ouwehand WH, Wernisch L, Dudbridge F. Comparison of methods for competitive tests of pathway analysis. PLoS ONE. 2012;7: e41018. https://doi.org/10.1371/JOURNAL.PONE.0041018.
35. Alhamdoosh M, Ng M, Wilson NJ, Sheridan JM, Huynh H, Wilson MJ, et al. Combining multiple tools outperforms individual methods in gene set enrichment analyses. Bioinformatics. 2017;33:414–24. https://doi.org/10.1093/BIOINFORMATICS/BTW623.

36. Hung JH, Yang TH, Hu Z, Weng Z, DeLisi C. Gene set enrichment analysis: performance evaluation and usage guidelines. Brief Bioinform. 2012;13:281–91. https://doi.org/10.1093/BIB/BBR049.
37. Väremo L, Nielsen J, Nookaew I. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. Nucleic Acids Res. 2013;41:4378–91. https://doi.org/10.1093/NAR/GKT111.
38. Su Y, Chen D, Yuan D, Lausted C, Choi J, Dai CL, et al. Multi-omics resolves a sharp disease-state shift between mild and moderate COVID-19. Cell. 2020;183:1479-1495.e20. https://doi.org/10.1016/j.cell.2020.10.037.
39. 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;569:655–62. https://doi.org/10.1038/s41586-019-1237-9.
40. Wei R, Wang J, Su M, Jia E, Chen S, Chen T, et al. Missing value imputation approach for mass spectrometry-based metabolomics data. Sci Rep. 2018;8:1–10. https://doi.org/10.1038/s41598-017-19120-0.
41. Pang Z, Chong J, Zhou G, de Lima Morais DA, Chang L, Barrette M, et al. MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. Nucleic Acids Res. 2021. https://doi.org/10.1093/nar/gkab382.
42. Schölkopf B, Smola A, Müller KR. Kernel principal component analysis. In: Gerstner W, Germond A, Hasler M, Nicoud JD, editors. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics). Berlin: Springer; 1997. p. 583–8. https://doi.org/10.1007/bfb0020217.
43. Hubert L, Arabie P. Comparing partitions. J Classif. 1985;2:193–218. https://doi.org/10.1007/BF01908075.
44. Stoney RA, Schwartz JM, Robertson DL, Nenadic G. Using set theory to reduce redundancy in pathway sets. BMC Bioinform. 2018;19:386. https://doi.org/10.1186/s12859-018-2355-3.
45. Bayerlová M, Jung K, Kramer F, Klemm F, Bleckmann A, Beißbarth T. Comparative study on gene set and pathway topology-based enrichment methods. BMC Bioinform. 2015;16:1–15. https://doi.org/10.1186/s12859-015-0751-5.
46. Castro-Dopico T, Clatworthy MR. IgG and Fcγ receptors in intestinal immunity and inflammation. Front Immunol. 2019;10:805. https://doi.org/10.3389/FIMMU.2019.00805/BIBTEX.
47. Boldyreva LV, Morozova MV, Saydakova SS, Kozhevnikova EN. Fat of the gut: epithelial phospholipids in inflammatory bowel diseases. Int J Mol Sci. 2021;22:11682. https://doi.org/10.3390/IJMS222111682.
48. Weiss TS, Herfarth H, Obermeier F, Ouart J, Vogl D, Schölmerich J, et al. Intracellular polyamine levels of intestinal epithelial cells in inflammatory bowel disease. Inflamm Bowel Dis. 2004;10:529–35. https://doi.org/10.1097/00054725-200409000-00006.
49. Guzior DV, Quinn RA. Review: microbial transformations of human bile acids. Microbiome. 2021. https://doi.org/10.1186/s40168-021-01101-1.
50. Hubler MJ, Kennedy AJ. Role of lipids in the metabolism and activation of immune cells. J Nutr Biochem. 2016;34:1. https://doi.org/10.1016/J.JNUTBIO.2015.11.002.
51. Fan F, Mundra PA, Fang L, Galvin A, Moore XL, Weir JM, et al. Lipidomic profiling in inflammatory bowel disease: comparison between ulcerative colitis and Crohn's disease. Inflamm Bowel Dis. 2015;21:1511–8. https://doi.org/10.1097/MIB.0000000000000394.
52. Abdel Hadi L, di Vito C, Riboni L. Fostering inflammatory bowel disease: sphingolipid strategies to join forces. Mediat Inflamm. 2016. https://doi.org/10.1155/2016/3827684.
53. Giannotta M, Tapete G, Emmi G, Silvestri E, Milla M. Thrombosis in inflammatory bowel diseases: what's the link? Thromb J. 2015;13:1–9. https://doi.org/10.1186/S12959-015-0044-2/FIGURES/2.
54. Hatoum OA, Binion DG, Otterson MF, Gutterman DD. Acquired microvascular dysfunction in inflammatory bowel disease: loss of nitric oxide-mediated vasodilation. Gastroenterology. 2003;125:58–69. https://doi.org/10.1016/S0016-5085(03)00699-1.
55. Aviello G, Knaus UG. ROS in gastrointestinal inflammation: rescue or sabotage? Br J Pharmacol. 2017;174:1704. https://doi.org/10.1111/BPH.13428.
56. Pereira C, Grácio D, Teixeira JP, Magro F. Oxidative stress and DNA damage: implications in inflammatory bowel disease. Inflamm Bowel Dis. 2015;21:2403–17. https://doi.org/10.1097/MIB.0000000000000506.
57. Liu C, Lehtonen R, Hautaniemi S. PerPAS: topology-based single sample pathway analysis method. IEEE/ACM Trans Comput Biol Bioinform. 2018;15:1022–7. https://doi.org/10.1109/TCBB.2017.2679745.
58. Li X, Li M, Zheng R, Chen X, Xiang J, Wu F-X, et al. Evaluation of pathway activation for a single sample toward inflammatory bowel disease classification. Front Genet. 2020;10:1401. https://doi.org/10.3389/fgene.2019.01401.
59. Yi M, Nissley DV, McCormick F, Stephens RM. ssGSEA score-based Ras dependency indexes derived from gene expression data reveal potential Ras addiction mechanisms with possible clinical implications. Sci Rep. 2020;10:1–16. https://doi.org/10.1038/s41598-020-66986-8.
60. McLuskey K, Wandy J, Vincent I, van der Hooft JJJ, Rogers S, Burgess K, et al. Ranking metabolite sets by their activity levels. Metabolites. 2021;11:103. https://doi.org/10.3390/metabo11020103.
61. Martens M, Ammar A, Riutta A, Waagmeester A, Slenter DN, Hanspers K, et al. WikiPathways: connecting communities. Nucleic Acids Res. 2021;49:D613–21. https://doi.org/10.1093/NAR/GKAA1024.
62. Gundem G, Lopez-Bigas N. Sample-level enrichment analysis unravels shared stress phenotypes among multiple cancer types. Genome Med. 2012;4:28. https://doi.org/10.1186/GM327.
63. Su K, Yu Q, Shen R, Sun S-Y, Moreno CS, Li X, et al. Pan-cancer analysis of pathway-based gene expression pattern at the individual level reveals biomarkers of clinical prognosis. Cell Rep Methods. 2021;1: 100050. https://doi.org/10.1016/j.crmeth.2021.100050.
64. Schubert M, Klinger B, Klünemann M, Sieber A, Uhlitz F, Sauer S, et al. Perturbation-response genes reveal signaling footprints in cancer gene expression. Nat Commun. 2018. https://doi.org/10.1038/S41467-017-02391-6.

Publisher's Note