Bioinformatics, 38, 2022, i378–i385
https://doi.org/10.1093/bioinformatics/btac232
ISCB/ISMB 2022

High-sensitivity pattern discovery in large, paired multiomic datasets

Andrew R. Ghazi[1,2,3], Kathleen Sucipto[1], Ali Rahnavard[1,2], Eric A. Franzosa[1,2,3], Lauren J. McIver[1,2,3], Jason Lloyd-Price[1,2], Emma Schwager[1], George Weingart[1,3], Yo Sup Moon[1], Xochitl C. Morgan[4], Levi Waldron[5] and Curtis Huttenhower [1,2,3,6,*]

[1]Biostatistics Department, Harvard T. H. Chan School of Public Health, Boston, MA 02115, USA, [2]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA, [3]Harvard Chan Microbiome in Public Health Center, Harvard T. H. Chan School of Public Health, Boston, MA 02115, USA, [4]Department of Microbiology and Immunology, University of Otago, Dunedin 9016, New Zealand, [5]Department of Epidemiology and Biostatistics, City University of New York Graduate School of Public Health and Health Policy, New York City, NY 10035, USA and [6]Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health, Boston, MA 02115, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Modern biological screens yield enormous numbers of measurements, and identifying and interpreting statistically significant associations among features are essential. In experiments featuring multiple high-dimensional datasets collected from the same set of samples, it is useful to identify groups of associated features between the datasets in a way that provides high statistical power and false discovery rate (FDR) control.

Results: Here, we present a novel hierarchical framework, HAllA (Hierarchical All-against-All association testing), for structured association discovery between paired high-dimensional datasets. HAllA efficiently integrates hierarchical hypothesis testing with FDR correction to reveal significant linear and non-linear block-wise relationships among continuous and/or categorical data. We optimized and evaluated HAllA using heterogeneous synthetic datasets of known association structure, where HAllA outperformed all-against-all and other block-testing approaches across a range of common similarity measures. We then applied HAllA to a series of real-world multiomics datasets, revealing new associations between gene expression and host immune activity, the microbiome and host transcriptome, metabolomic profiling and human health phenotypes.

Availability and implementation: An open-source implementation of HAllA is freely available at http://huttenhower.sph.harvard.edu/halla along with documentation, demo datasets and a user group.

Contact: chuttenh@hsph.harvard.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Pattern discovery in high-dimensional, heterogeneous data is a long-standing problem in applied statistics (Bühlmann and Van De Geer, 2011; Johnstone and Titterington, 2009). It is challenging for several reasons, including the inherent tradeoffs between sensitivity and generality—that is, the ability and power to detect associations given varying assumptions about the functional form of the relationship (Altman and Bland, 1994). When applied in contexts such as high-throughput biology, these challenges are exacerbated by noisy, diverse and non-linear data. Examples include biospecimens drawn from large cohorts, in which each sample may be decorated with heterogeneous phenotypic variables (clinical features, diseases status, etc.) and multiple high-dimensional molecular measurements (microbial taxa, epigenetic markers, gene expression, etc.). In the biological sciences specifically, selecting a subset of associations for follow-up validation experiments can be a complex yet important decision point. A gap remains to efficiently identify related features in such data, while both maintaining sensitivity and controlling spurious association reporting.

All-against-all (AllA) approaches, which test all pairs of features and then correct for false discovery, scale well only in completely independent tests of moderate size (Bourgon *et al.*, 2010). Under other conditions, such feature-wise approaches can have limited statistical power due to testing many correlated hypotheses for individually weak associations (Rosenberg *et al.*, 2006). This has led to the development of a variety of (typically parametric) block-testing approaches, such as partial least squares (PLS; Abdi, 2010), canonical correlation analysis (CCA; Hardoon *et al.*, 2004), PLS discriminant analysis, sparse principal component analysis (SPCA; Zou *et al.*, 2006) and SPARSE-CCA (Lykou and Whittaker, 2010). These serve to detect associations between reduced-dimensional representations of large input datasets, but they are typically limited by one or more of (i) applicability only to continuous measurements with no missing values (or only categorical, not mixed; PLS, CCA, SPCA); (ii) a focus on the single, strongest axis of covariation between the datasets

(CCA); (iii) an assumption of linear covariation (CCA, SPCA, PLS); (iv) identifying complex combinations of feature loadings implicated in associations, rather than specific features [particularly in kernel methods such as Kernel PCA (Mika *et al.*, 1998)]; and (v) a lack of explicit control of the false discovery rate (FDR).

Recent advances have focused on nonparametric methods for identifying highly general (i.e. linear and non-linear) associations between individual pairs of features, sometimes relying on computational or permutation-based methods not readily accessible to early applied statisticians. These include, for example, distance correlation (dCor; Székely *et al.*, 2007), which measures (not necessarily linear) the dependency of two random variables with possibly different dimensions. The Chatterjee rank correlation (XICOR; Chatterjee, 2020) is another recently introduced similarity measure that uses rank differences to assess the degree to which one variable is a measurable function of another. While dCor and XICOR provide comparatively general methods to discover complex associations between individual pairs of features, when applied to many combinations of linear feature pairs with varying degrees of dependence, the resulting statistical power can fall below simpler traditional approaches after controlling FDR for multiple hypothesis tests (Kinney and Atwal, 2014).

In this work, we develop a hierarchical AllA association testing framework (HAllA) that identifies highly general association types in paired, high-dimensional and potentially heterogeneous datasets. HAllA preserves statistical power in the presence of collinearity by testing coherent clusters of variables in a hierarchical manner, while controlling overall FDR with hierarchical multiple hypothesis testing. HAllA discovers associations between blocks of features among paired datasets in a way that increases interpretability by grouping features according to their relatedness.

2 Materials and methods

In this section, we provide an overview of the HAllA algorithm and its component steps. Additional methods details, including pseudocode, are provided in the Supplementary material.

2.1 The HAllA algorithm

Hierarchical All-against-All Association testing (HAllA) identifies block associations between two potentially heterogeneous datasets coindexed along one axis (Fig. 1A). This coindexing is referred to as the 'samples' axis (columns) and the measurement axis as 'features' (rows). For a pair of datasets containing measurements that describe the same set of samples and a specified pairwise similarity measure, the HAllA algorithm proceeds by (i) optionally discretizing features to a uniform representation (if required by the similarity measure), (ii) finding the Benjamini–Hochberg (BH) FDR threshold, (iii) hierarchically clustering each dataset separately to generate two data hierarchies, (iv) coupling clusters of equivalent resolution between the two data hierarchies, (v) testing coupled clusters for statistically significant association in block format where the block passes a threshold for false-negative tolerance (FNT) and (vi) iteratively increasing resolution by descending through the pair of hierarchies according to which split results in a higher Gini score gain. The final groups of features are those that lead to the largest hypothesis blocks that pass the FNT threshold (Fig. 1C and Supplementary Fig. S2).

2.2 Optionally discretizing input datasets

This step permits direct comparison of continuous and categorical features and further enables the application of highly general measures of association from information theory, such as mutual information (MI). This combination allows HAllA to detect significant (i) non-linear associations between paired continuous features (e.g. quadratic or sinusoidal relationships), (ii) differences in group means for paired continuous and categorical features and (iii) non-random associations between paired categorical features. HAllA's default discretization scheme divides continuous features into bins of equal size once at the start of processing. By default, the number of bins is the cube root of the sample size, which provides reasonable power at a variety of sample sizes and correlation levels (Supplementary Fig. S2). HAllA also removes features with low variance by applying a configurable frequency threshold (defaulting to 100%, meaning only features with no variability are removed) in order to reduce the number of unnecessary tests.

2.3 Hierarchical clustering and cluster coupling allow detection of associations between groups of features

Each dataset is subjected to average-linkage hierarchical clustering using the specified similarity measure (Spearman's rank correlation by default) within each dataset (Fig. 1A). Associations between datasets are tested in a top-down manner by pairing nodes of similar resolution between their respective data trees. More specifically, HAllA recursively builds a tree of hypotheses to test (the 'hypothesis tree'), beginning at the top of each dataset's tree, descending to a set of nodes within each data tree and then pairing each selected node from the first tree with each selected node of the second tree. At each step in the descent process, the choice of whether to descend within the X or Y hypothesis tree is made by comparing which split leads to a higher Gini score gain. In the case of ties, both descent steps are made. This procedure is repeated until termination, i.e. when the hypothesis block passes the FNT threshold or when the selected nodes represent single features in their respective data trees

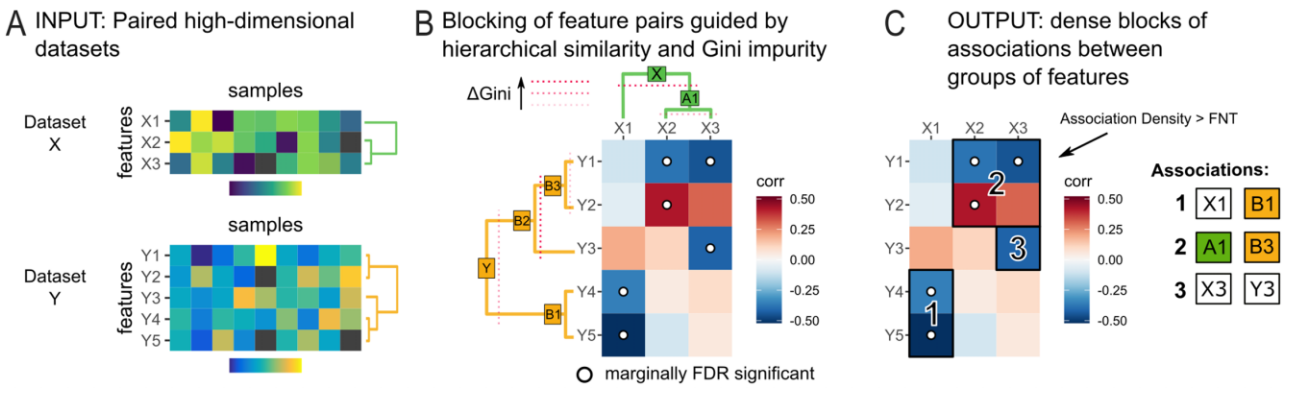

Fig. 1. Hierarchical all-against-all (HAllA) association testing. (**A**) HAllA provides a novel method for heterogeneous association discovery in high-dimensional data. Input data are represented in matrix form as features (rows) and samples (columns). Features within each dataset are hierarchically clustered using average-linkage and Spearman association as default methods. (**B**) Starting with the feature hierarchies and the full pairwise association matrix, HAllA descends through both trees rejecting putative blocks that fail the FNT threshold using Benjamini–Hochberg FDR threshold for pairwise associations within the block. When a block fails the FNT threshold, the decision of whether to cut the X or Y feature tree is guided by whichever cut yields a higher Gini impurity improvement. The process stops when a block is dense with marginal associations. (**C**) Significant associations are reported in a block-wise manner once the hierarchical descent step has terminated. In this example, the X3:Y2 pair is correctly included among the significant associations, where it would have been missed by an AllA approach

(Fig. 1B). Another way to visualize this process is by focusing on the all-by-all hypothesis matrix. The process begins by checking if the entire matrix passes the FNT threshold. If not, the matrix is recursively cut horizontally or vertically into smaller hypothesis blocks, with the position of each cut decided by each dataset's similarity tree and Gini score gain. The cutting process stops when the smaller hypothesis blocks pass the FNT threshold or have been reduced to one-by-one blocks (Fig. 1C).

The notion of identifying and testing hypotheses in a hierarchical manner was previously proposed by Yekutieli (2008). HAllA's hypothesis tree similarly groups more specific child hypotheses below a more general parent hypothesis. However, HAllA's approach differs fundamentally from the Yekutieli approach in that HAllA tests hierarchical hypotheses until a null hypothesis can be rejected; Yekutieli's method tests until the first failure to reject a null hypothesis. This results in HAllA maintaining greater power, while Yekutieli's method instead maintains greater specificity (Supplementary Fig. S1).

2.4 Determining the statistical significance of block associations

The method proceeds by testing the nodes in the hypothesis tree (each representing a pair of feature clusters, one from each dataset) for significant between-cluster associations. Each node in the hypothesis tree is evaluated using the following procedure: let $\mathcal{H}$ denote the null hypothesis that the two clusters of features are not related, and $\mathcal{H}_i$ be the null hypothesis of no association between two individual features within those clusters. Define R^i as the P-value of the association between an individual pair of features considered by $\mathcal{H}_i$. We then count all rejected $\mathcal{H}_i$ (i.e. $R^i \leq k_{BH}$), and all $\mathcal{H}_i$ that failed to reject, i.e. $R^i > k_{BH}$ where k_{BH} is the global BH FDR threshold. The block-wise FNT is provided by the user (default FNT = 0.2) and acts as the allowed fraction of paired associations that are expected to fail to reject despite being true associations. If the fraction of paired associations in a block with $R^i > k_{BH}$ is greater than or equal to FNT, we reject the entire block hypothesis $\mathcal{H}$.

If any hypothesis involved clusters rather than feature tips, and failed to reject, the procedure is repeated with new null hypotheses for associations between sub-clusters (Fig. 1B), as described in section 'Descending in sub-hypotheses of block hypotheses' in the Supplementary material. HAllA reports all significant associations between clusters of any size that pass the FNT threshold (Fig. 1C).

2.5 Visualizing outputs

Once the analysis is complete, the results are visualized in a 'HAllAgram' (Fig. 4). This heatmap visualizes the relatedness and strength of association between pairs of features in the two datasets. Features are ordered along each axis according to their position in the hierarchical tree so that clusters of significant features can be boxed into contiguous units. Marginally associated pairs are dotted, and each hypothesis block is labeled with the rank of its association strength. Features not associated with any block are not plotted by default. For analysis results where large numbers of blocks are detected, only the strongest blocks are shown (30 by default), with potentially incomplete, lower-ranked blocks boxed in gray. Together, this set of plotting techniques allows users to visually understand the related sets of hypotheses that HAllA has detected. Other plotting utilities are also included with the method's current implementation, such as a clustermap that displays the entire association tree in the margins for both datasets, as well as a diagnostic plot that displays the input data associated with individual hypothesis blocks.

3 Results

3.1 Halla increases power while controlling FDR to report block-wise associations

When applied to paired datasets with no significantly related blocks of features, HAllA's descent algorithm reduces to AllA direct pairwise feature testing. In such circumstances, HAllA is expected to perform similarly to AllA. However, when there are sets of correlated features within one dataset that are correlated with another set of features in the other, HAllA will report the association block. Notably, we expect this behavior to be common in multiomics data, where we see large clusters of molecular features (e.g. coexpressed genes in a pathway).

To evaluate these claims, we applied HAllA and AllA to paired, synthetic datasets generated with the data simulator function in the HAllA software. These datasets contained prespecified block

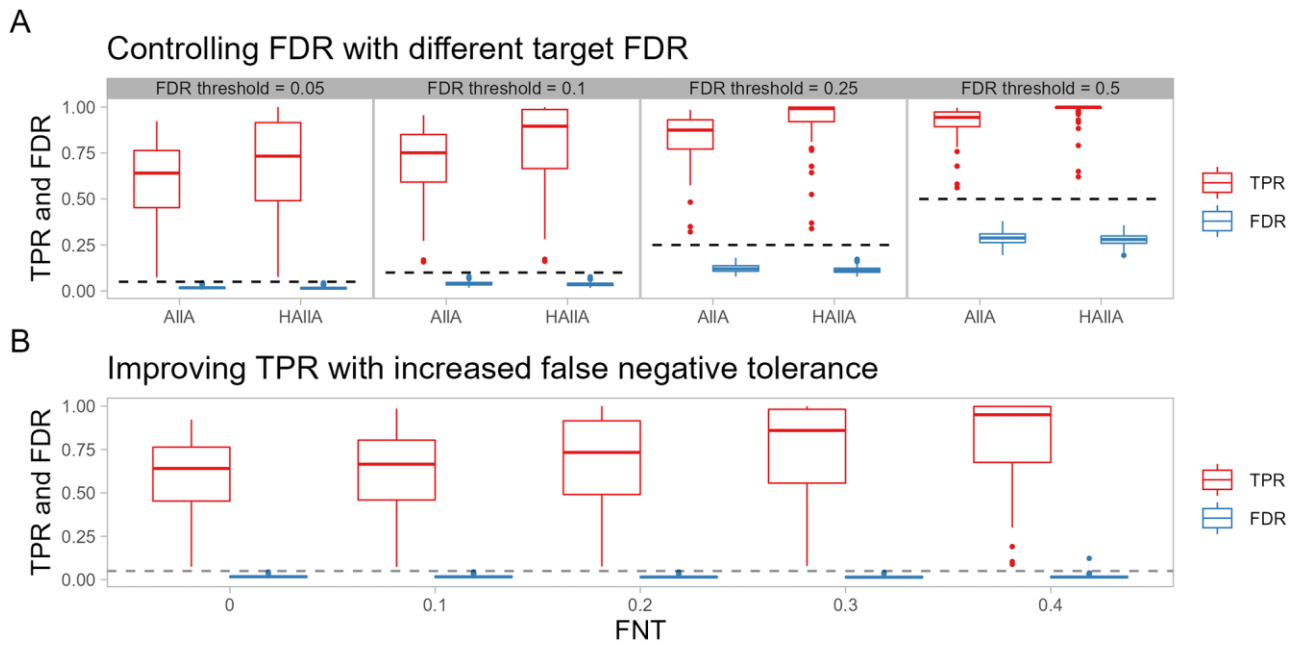

Fig. 2. HAllA improves statistical power while controlling the FDR. Fifty paired, synthetic datasets with 200 features and 50 samples containing clusters with linear block associations were analyzed. (**A**) With FNT = 0.2, HAllA maintains the simulated FDR below the target (here (0.05, 0.1, 0.25 and 0.5), with associated tradeoffs in statistical power. In addition, HAllA is consistently better powered than AllA association testing across this range of target FDR values. Dashed lines parallel to the x-axis indicate the target FDR value in each comparison. (**B**) By increasing the FNT, HAllA can improve the true positive rate with a comparatively minor increase in FDR

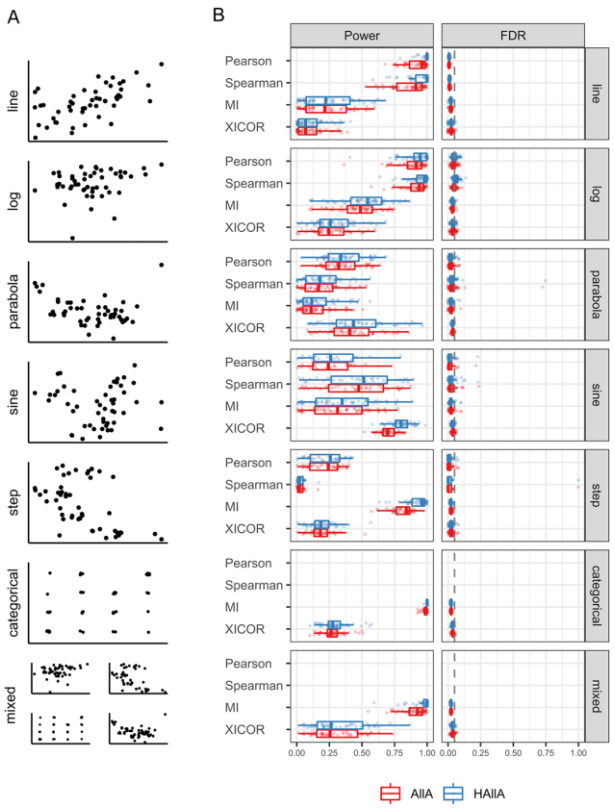

Fig. 3. HAllA discovers block-structured associations while controlling FDR. (**A**) For a variety of feature linkage relationships, we simulated 50 independent paired datasets, each containing 200 features, 50 samples and clusters of correlated features. We then evaluated the ability of hierarchical versus AllA testing to recover these associations using a variety of similarity metrics. (**B**) Performance was evaluated by comparing power and FDRs. Our hierarchical AllA approach improved sensitivity relative to naive AllA approaches at a comparable FDR. Similarity metrics that do not accept categorical data have not been evaluated in the categorical or mixed association type. Other similarity metrics included in HAllA (dCor, NMI) were not applied in these simulations because their reliance on permutation tests made them too slow for simulations of this size (i.e. with many repeated iterations), although they are typically practical in individual real-world datasets

associations, which allowed us to evaluate the statistical and computational performance of these two methods (Figs 2 and 3). Dimension-reducing methods such as CCA and PCA were not included because of their inability to dichotomize individual pairs of features as associated or not associated *post hoc*. With a constant target FNT in associated blocks of 0.2, HAllA controls FDR, reports association in block form and improves power on average by 7–11% (Fig. 2A) across varied FDR thresholds. HAllA also consistently boosts the true positive rate relative to AllA using different target FNT values in associated blocks (Fig. 2B).

We evaluated many different forms of feature association, including linear, quadratic, logarithmic, sinusoidal, stepwise, parabolic and mixed (combined discrete and continuous) data. We compared HAllA and AllA across these association types using a variety of similarity measures, including XICOR, MI, Spearman correlation and Pearson correlation. Across datasets and similarity measures, HAllA consistently detected more built-in associations (had better average power by as much as 10%) than AllA while controlling FDR at the same prespecified level (Fig. 3B). Each similarity measure exhibited various strengths and weaknesses across evaluations depending on data type. As expected, for mixed and categorical data, MI is appropriate, and for monotonic associations in continuous data, Spearman correlation performs well. XICOR is applicable to both continuous and discrete outcomes and performs well on difficult non-linear association types. However, it is rarely the most statistically powerful option, and its interpretation is limited to measuring the association of features in Y as a measurable function of features in X and not vice versa. A similar power analysis that used a fixed association structure with varying correlation strength led to similar conclusions (Supplementary Fig. S3). Supplementary Figure S5 provides a flowchart to assist in deciding which similarity measures are appropriate depending on the structure of the input data. Together these results show that the HAllA approach increases statistical power while maintaining the FDR across a wide variety of association structures under simulation.

3.2 Halla identifies novel fatty acid-xenobiotic metabolism associations in PPARα-deficient mice

PPARα is a nuclear receptor that regulates transcription of genes related to lipid metabolism in the liver (Martin *et al.*, 2007). These genes show high fatty acid catabolism rates, which can in turn affect hepatic fat storage and lipoprotein metabolism. We used HAllA to examine associations between 120 hepatic transcript levels and 21 liver lipid levels in a previously published dataset (González *et al.*, 2008; Fig. 4). These data were originally collected from 40 wild-type and peroxisome proliferator-activated receptor-α (PPARα)-deficient mice (Martin *et al.*, 2007). HAllA recovered 109 block associations comprising 225 pairwise associations at target FDR of 0.05 (chosen to match the previous study). HAllA's results included all associations that were previously reported using CCA, including a key relationship between fatty acids and the xenobiotic metabolism genes Cyp3a11 and Car1 (MGI:88268).

We further identified several novel associations, including a link between polyunsaturated fatty acids eicosatrienoic acid (C20:3n6) and arachidonic acid (C20:4n6; Selvaraju *et al.*, 2012) with a group of transcripts including Mcad (Acadm, MGI:87867). This gene (C-4 to C-12 straight chain acyl-Coenzyme A dehydrogenase) encodes one of the main catalysts of the beta-oxidation process used for the degradation of these fatty acids. Genes Car1 (MGI:88268) and Acot11 [MGI:1913736; a carbonic anhydrase and lipid transfer protein, respectively (Hunt *et al.*, 2000; Lynch *et al.*, 1995)] fell in the same cluster with C20.3n.6 and C20.4n.6, which would suggest a trafficking and transport relationship between these genes and fatty acids.

3.3 Associating microbes with metabolites in the infant gut microbiome

In a prior study, Kostic *et al.* (2015) examined the development of the human gut microbiome in a prospective, longitudinally sampled cohort of 33 Finnish and Estonian infants at high risk for type 1 diabetes. Stool samples and clinical metadata (e.g. breastfeeding status, diet and appearance of allergies) were collected monthly. Subjects' stool samples were analyzed using (i) 16S rRNA amplicon sequencing (to profile gut microbiome composition) and (ii) targeted mass spectrometry (to profile host and microbial metabolites). The dataset included 103 samples from 19 individuals, each with paired metabolomics and 16S rRNA gene sequencing data. We applied HAllA to identify associations between the residuals of microbial and metabolite abundances after correcting for longitudinal trends and subject-specific random effects using a linear mixed-effects model (Skrondal and Rabe-Hesketh, 2004; DiabImmune dataset, Supplementary material).

HAllA recovered 44 microbial/metabolite cluster associations between 13 microbial genera and 44 metabolites using the same $q < 0.05$ threshold as in the original study (Fig. 5A). These encompassed 57 pairwise associations, using Spearman correlation as the measure of pairwise feature similarity (as both data types are continuous). Using pairwise, AllA testing, 56 associations were significant at the same threshold.

Our results again replicate all significant associations from the previous study's CCA, and most of the associations from the original pairwise association analysis of the previous paper. HAllA also found additional associations, including a novel association between Prevotella and inosine (Spearman coefficient = -0.439, FDR Q-value = 0.0053), which could be explained by a mechanism where increased levels of urotoxins in the body from inosine decreased the

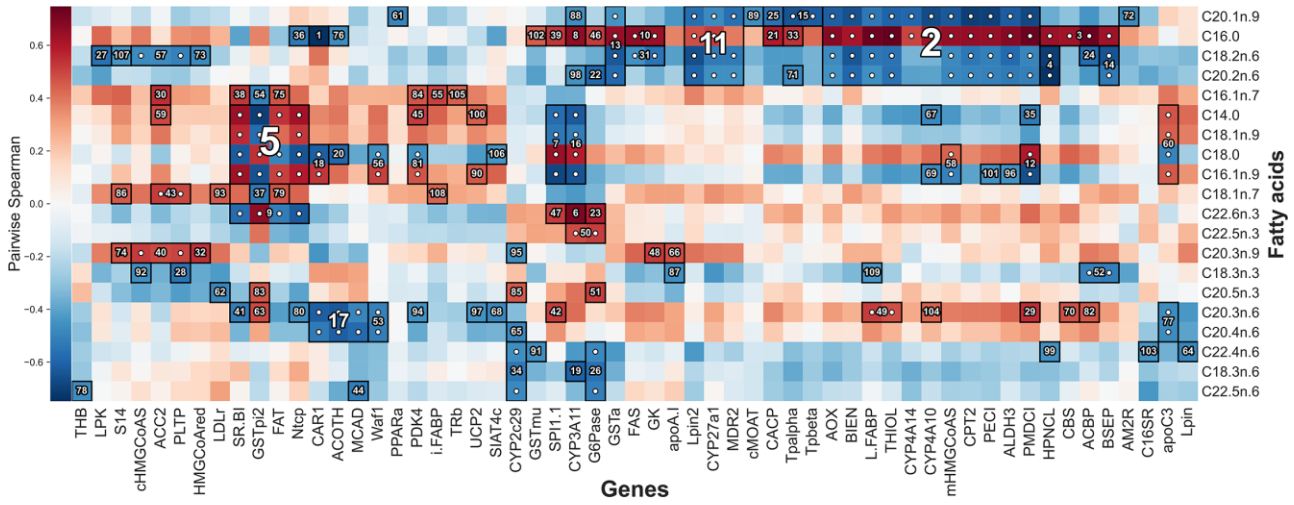

Fig. 4. Association of fatty acids with host transcriptional activity in murine liver. We applied HAllA to paired data comprising 120 hepatic transcript levels and 21 liver lipid levels in a set of 40 previously profiled mice (Martin *et al.*, 2007). In this 'HAllAgram' visualization of results, block associations are numbered in descending order of significance, with each numbered block corresponding to a group of coexpressed transcripts related to a group of co-occurring lipids. A white dot indicates the marginal significance of a particular pair of features. A total of 109 block associations achieved significance at FDR 0.05, matching the previous study's threshold based on canonical correlation (González *et al.*, 2008). HAllA's associations were a strict superset of those found earlier by CCA. Spearman correlation was used as a similarity metric

abundance of intolerant Prevotella. HAllA also reports novel associations between fecal bile acids lithocholate and lithocholic acid and genera Faecalibacterium and Veillonella (Spearman coefficients = 0.36, -0.39; Q-values = 0.026, 0.015, respectively). Faecalibacterium is Gram-positive anaerobic bacteria genera from order Clostridiales, while Veillonella are Gram-negative anaerobic cocci. Relationships between these genera and global bile acid levels (with matching correlation signs) have been previously indicated by several studies, particularly in cirrhosis (Kakiyama *et al.*, 2013). These data thus demonstrate HAllA's potential benefits relative to pairwise or omnibus (e.g. CCA) testing by simultaneously providing both greater interpretability and power.

3.4 Associating the gut microbiome with host transcription in ulcerative colitis

We next applied HAllA to data combining (i) 16S rRNA amplicon sequencing of the human gut microbiome and (ii) Affymetrix microarray screens of ileal RNA expression across 204 individuals in a cohort of ileal pouch-anal anastomosis (IPAA) patients (Morgan *et al.*, 2012). In the original multivariate analysis of these data (Morgan *et al.*, 2015), microbial operational taxonomic unit (OTU) abundances were decomposed into principal components (PCs), and PCs accounting for up to 50% of the variance in the datasets were compared by AllA testing (an example of PC regression). While this approach enables well-powered comparisons of large numbers of features, the features are embedded as loadings in PCs, which complicates biological interpretation of the resulting associations.

HAllA identified 327 block associations in these microbial and gene expression data using an FDR threshold of 0.05 and an FNT of 0.1 (Fig. 5B). Total relationships encompassed 125 OTUs, 187 transcripts and the equivalent of 368 pairwise associations. The original study focused on the 9th principal component (PC9) of the dataset due to its linking of a group of IL12/complement pathways to members of the microbiome, using an FDR threshold of 0.25. Of HAllA's reported microbe–transcript associations when run with the same threshold, 20 genes were drawn from the 26 transcripts whose largest loading was in PC9. HAllA's findings support a surprising result of the original study: although PC9 represented only 1% of the transcriptional variation in these samples, it captured most associations between transcription and the microbiome during pouchitis. These results also agree with a previous re-analysis of these data (Zhan *et al.*, 2017) assessing global covariation between gut microbial and transcriptional structure, which called out three pathways (interleukin-12, inflammatory and inflammatory bowel disease genes) that overlap heavily with HAllA's block results (e.g. 28 out of 51 tested genes in the Kyoto Encyclopedia of Genes and Genomes (KEGG) TRP channel mediator pathway and 34 of 61 tested genes in the KEGG IBD pathway were significantly associated with microbial species).

Expanding on these previous associations, HAllA found a group of facultative anaerobes (mainly streptococci) to be positively associated with the expression of the genes WDR49 and SERPINI2. WDR49 is a WD repeat-containing protein upregulated in alveolar macrophages, a cell type specifically responsible for nasopharyngeal pathogen uptake (Patel *et al.*, 2017). This association suggests this protein may also be involved in recognition of bacteria in the gut environment. Another novel association in HAllA's results linked a group of Bifidobacterium OTUs with FABP1, a member of the long-chain fatty acid-binding protein family involved both in lipid sensing and metabolic regulation of energy harvest (Furuhashi and Hotamisligil, 2008). This positive relationship has also been observed in mice (Patterson *et al.*, 2017). Finally, during intestinal inflammation and bleeding, host-microbial iron competition is a limiting factor in subsets of microbial growth (Werner *et al.*, 2011), which may be responsible for the significant negative association identified between the siderophore-rich genus Blautia and SLC40A1, a human intestinal epithelial iron ion transmembrane transporter (Donovan *et al.*, 2005).

3.5 HAllA's applicability to heterogeneous datasets

We finally applied HAllA to identify associations between mixed clinical metadata and RNA expression in the breast cancer cohort of the Cancer Genome Atlas (TCGA; Weinstein *et al.*, 2013) available from the LinkedOmics R package (Vasaikar *et al.*, 2018), focusing on highly expressed yet variable transcripts (Supplementary Fig. S4). HAllA identified 483 significant (Q-value < 0.1) metadata-RNA associations within 261 blocks, including clusters of transcripts associated with tumor purity, PAM50 subtype and ER Status. Notably, the transcripts occupying the block associated with PAM50 subtype include CA12, GABRP, NAT1 and TBC1D9, which have been previously proposed as predictor genes for breast cancer mortality, recurrence (Andres *et al.*, 2013) and drug response (Pogue-Geile *et al.*, 2013). Coupled with the results of the preceding applications, these results speak to the generality of HAllA's association discovery power across large, heterogeneous datasets.

In order to demonstrate the usefulness of alternative similarity measures like XICOR, we decided to look for non-linear functional relationships between RNA and protein expression in the breast cancer cohort of the Cancer Genome Atlas (TCGA; Weinstein *et al.*,

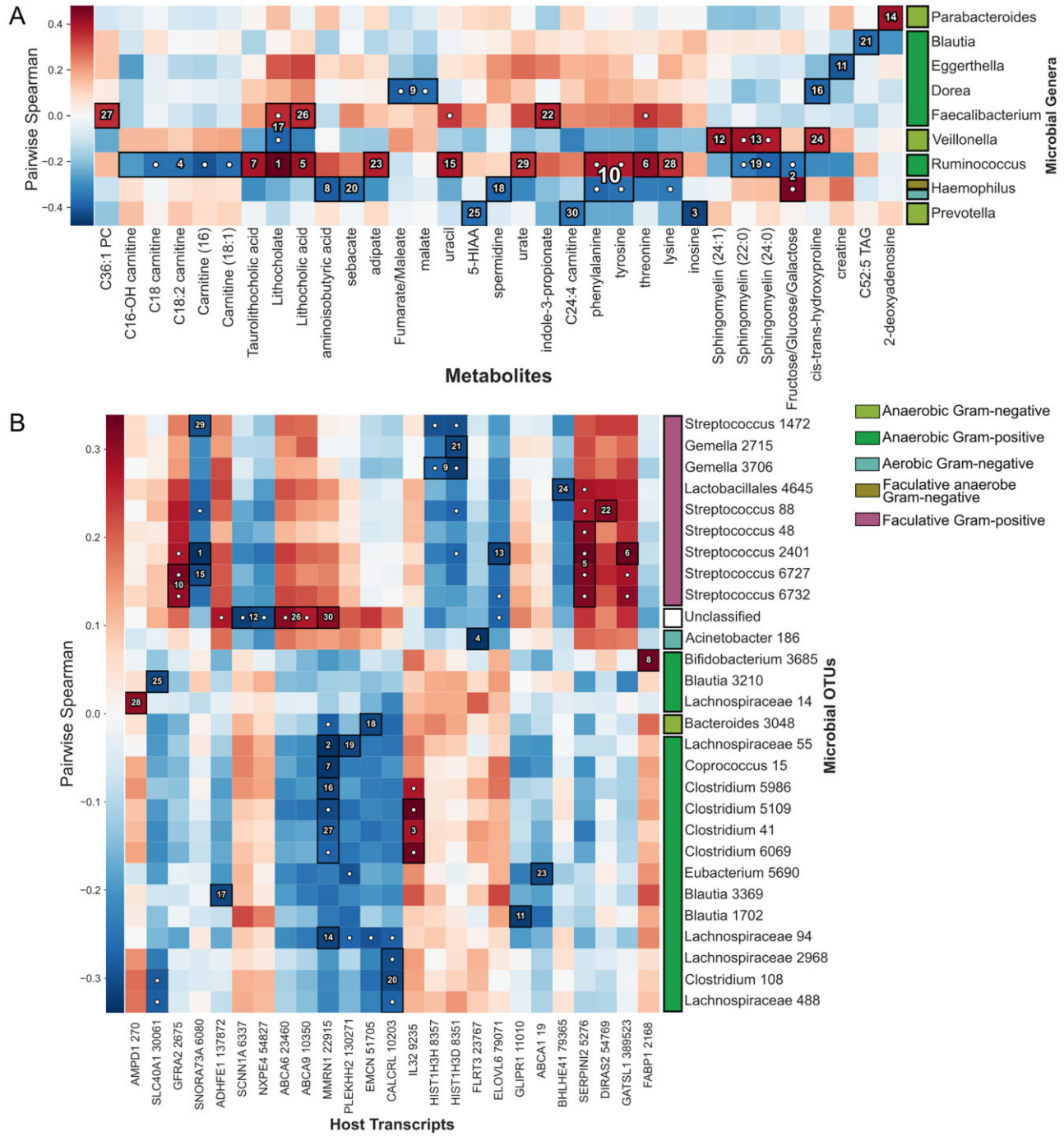

Fig. 5. HAllAgram for block-wise associations. (**A**) Using HAllA to associate multiomic data for the analysis of metabolome–microbiome interactions. We used HAllA to associate paired stool metabolomic and 16S rRNA gene sequencing data from the DIABIMMUNE (Kostic *et al.*, 2015) cohort, in which infants were recruited at birth and sampled monthly for the first 3 years of life. The data comprise 104 samples and describe the abundance of 20 genera and 284 labeled metabolites. A white dot indicates the marginal significance of a particular pair of features. Here, we show the 30 strongest associations ranked by *P*-value (target FDR = 0.05). (**B**) Relating host transcriptome and microbial taxa in inflammatory bowel disease (IBD) patients. We applied HAllA to identify associations between the human gut microbiome and transcriptome in 204 patients receiving IPAA surgeries (Morgan *et al.*, 2012). Block associations are numbered in descending order of significance based on best *P*-values in each block with each numbered block corresponding to a group of coexpressed transcripts related to a group of co-occurring microbial taxa (OTUs)

2013). We applied HAllA to these data using both Spearman and XICOR as similarity measures, then examined the significant associations that came out with the latter but not the former. Among these, we noticed three associations between RNA expression of transcription factor FOXC1 and protein expression of CCNE2, PIK3CA and SRSF1 (FDR Q-value $= 9.3 \times 10^{-7}$, 3.9×10^{-5}, 0.015, respectively), which showed compelling U-shaped relationships (Fig. 6). When compared with PAM50 clinical subtypes, these relationships emerge as a result of two features of the originating tumors. First, the different PAM50 subtypes vary in average FOXC1 expression (i.e. average position on the x-axis). Secondly, the effect of FOXC1 on the expression of each protein appears to vary between the subtypes, with the opposite sign in the basal subtype. There are individually well-established links between subtype and FOXC1, CCNE2 and PIK3CA (Caldon *et al.*, 2012; Elian *et al.*, 2018; López *et al.*, 2010). However, the varying relationship of each protein with FOXC1 by subtype has seemingly gone unnoticed in the literature, presumably due to the marginally non-linear shape of

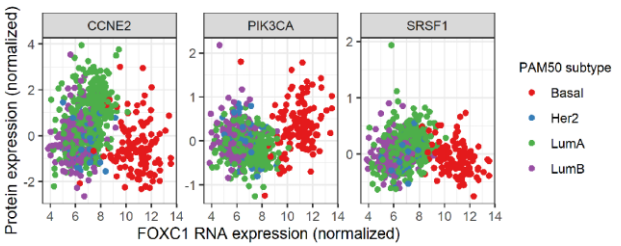

Fig. 6. Non-linear relationships detected between RNA and protein expression in a breast cancer cohort. By using an association metric sensitive to non-linear relationships (XICOR), HAllA detects U-shaped relationships between FOXC1 RNA expression and the protein expression of three genes. Overlaying the PAM50 subtype reveals that the U-shapes seem to emerge from a varying response to increased FOXC1 RNA expression by subtype. This effect seems to have gone unnoticed in the literature, thus demonstrating the ease with which HAllA can aid in the discovery of complicated relationships that might be missed otherwise

the overall relationship. While further study of the clinical importance of these relationships is warranted, these findings demonstrate the ease of well-powered, flexible, non-linear association discovery with HAllA.

4 Discussion

In this work, we proposed and validated HAllA, a novel statistical method to find associations between multiomic datasets. HAllA addresses several important methodological challenges in the analysis of high-dimensional datasets. It is applicable to data that are heterogeneous both within and between experiments, and it maintains statistical power using a novel hierarchical association testing and FDR control procedure. In this method, groups of correlated tests are modeled as blocks, ultimately reporting associations within blocks and between block representatives from multiple data types and experiments. This permits both great flexibility in the types of measurements to which it is applied and ease of interpretation of the resulting significant associations.

Class prediction approaches are commonly used to model relationships between high-dimensional datasets with variables measured using shared observational units. For example, PLS (Chin, 1998) and its close relative CCA (Sun *et al.*, 2009) identify latent variables in one dataset that are maximally correlated to latent variables in the other dataset. These methods, and robust and penalized varieties (Hubert and Branden, 2003; Witten *et al.*, 2009), can identify blocks of variables that are correlated within one dataset and in turn with another block of correlated variables in another dataset. They do not, however, control for family-wise error or FDR, and so are most suitable for prediction or exploratory, visual and descriptive analysis. With these methods, inference on the existence of associations between the variables of two datasets against null hypotheses of independence still relies on univariate hypothesis tests (and possibly dimension reduction or clustering) and is performed subsequently in a separate step. The FDR for the potentially large number of tests can be controlled by the Benjamini and Hochberg method (Benjamini and Hochberg, 1995), which has been adapted for dependent tests (Yekutieli and Benjamini, 1999) and hierarchically organized tests (Winkler, 1967) that are continued until non-significance. The approach described here thus aims to combine the best features of these different existing approaches, yielding clustering of potentially heterogeneous variable types within each dataset with hierarchical testing and control of FDR.

While these approaches are frequentist, Bayesian models are also used to improve power and share information among feature blocks (Cantor *et al.*, 2010; Lewinger *et al.*, 2007; Mourad *et al.*, 2010, 2011). While such methods are extremely powerful within their target domains, they are typically intended for the incorporation of specific prior knowledge, such as graph structure (Ben-Gal, 2008; Winkler, 1967), phylogeny (Ronquist and Huelsenbeck, 2003) or pathway-based functional roles (Huson *et al.*, 2011). They can also be computationally expensive in cases where many or long simulation chains are required for convergence (Huelsenbeck *et al.*, 2001). HAllA's non-parametric frequentist approach will likely result in reduced power relative to such models within the domains for which they are designed, but with substantially reduced computational cost and without the need to specify model relationships and priors in each new application domain. Like most statistical tradeoffs, HAllA's generality as a tool for association discovery thus comes at a cost in specific circumstances where it is desirable to instead utilize prior knowledge and known data structure.

A limitation of the current method is that it can only look for associations between two datasets at a time. While the method can be applied to multiple pairs of joint datasets manually, this becomes combinatorially prohibitive in particularly thorough studies where a large number of high-dimensional data types are available (e.g. studies that collect genetics, gene expression, epigenetics, microbial profiles, metabolites and metadata from each sample). In circumstances such as these, repeated application of HAllA across each pair of datasets would no longer properly control FDR. A potential extension would be to incorporate multivariate testing directly as an association measure, e.g. block PERMANOVA (Anderson, 2001; McArdle and Anderson, 2001) or Procrustes analysis (Goodall, 1991), to lower the combinatorial burden by performing inference on sets of features rather than individual feature pairs. Second, the model does not share information between blocks, as would be the case in a fully multivariate test (Anderson, 2001) or a hierarchical Bayesian model (Mourad *et al.*, 2011). Cases in which data do include such multilayered non-independence structure may indeed be better handled in a Bayesian framework. Finally, and relatedly, it is not straightforward to incorporate any type of prior knowledge into the HAllA framework, again because of HAllA's intention for wide applicability. Prefiltering can be used, as in several of our own examples, but this can be either beneficial or detrimental depending on context (Fan *et al.*, 2009; Waldron *et al.*, 2011).

Future work could also provide several refinements to the method, in addition to addressing these limitations. Currently, for example, known but undesirable confounders must be separately regressed out prior to using HAllA, and the method run on the resulting residuals instead of raw data. Integrating such covariate adjustment would be possible in future versions of the method's implementation. Perhaps most importantly, it may be possible to place tighter theoretical bounds on the block-wise and global FDR control beyond what is provided by HAllA's adaptation of the Benjamini–Hochberg (Benjamini and Hochberg, 1995) and Benjamini–Yekutieli methods (Benjamini and Yekutieli, 2001). This would also suggest a theoretical framework within which to characterize the amount and types of non-independence best handled by hierarchical block association testing. Ultimately, tradeoffs must be made between power and generality (Simon *et al.*, 2014). However, we aim for HAllA to provide a happy medium, capable of serving as an easy-to-use first pass analysis in a wide range of multiomics data types.

Acknowledgements

We thank Alex Kostic, Tommi Vatanen and Vincent Carey for assistance obtaining and curating datasets for the applications section; Hera Vlamakis, Hector Corrada Bravo, William Shannon, A. Brantley Hall, Himel Mallick, Siyuan Ma and Susan Holmes for helpful discussions, suggestions and feedback.

Data availability

The datasets analyzed in this work were derived from publicly available sources. See the accompanying reference for each dataset for availability.

Funding

This study was supported by Army Research Office grant W911NF-11-1-0429, National Science Foundation DBI-1053486 and National Institutes of Health U54DE023798 to C.H.

Conflict of Interest: none declared.

References

Abdi,H. (2010) Partial least squares regression and projection on latent structure regression (PLS regression). *WIREs Comp. Stat.* **2**, 97–106.

Altman,D.G. and Bland,J.M. (1994) Diagnostic tests. 1: sensitivity and specificity. *BMJ* **308**, 1552.

Anderson,M.J. (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecol.* **26**, 32–46.

Andres,S.A. *et al.* (2013) Interrogating differences in expression of targeted gene sets to predict breast cancer outcome. *BMC Cancer* **13**, 1–8.

Ben-Gal,I. (2008) Bayesian networks. In: Ruggeri,F. et al. (eds.) *Encyclopedia of Statistics in Quality and Reliability*. Vol. 1, John Wiley & Sons, Ltd.https://doi.org/10.1002/9780470061572.eqr089.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300.

Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 1165–1188.

Bourgon,R. *et al.* (2010) Independent filtering increases detection power for high-throughput experiments. *Proc. Natl. Acad. Sci. U S A* **107**, 9546–9551.

Bühlmann,P. and Van De Geer,S. (2011) *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media, Berlin, Germany. https://doi.org/10.1007/978-3-642-20192-9.

Caldon,C.E. *et al.* (2012) Cyclin E2 overexpression is associated with endocrine resistance but not insensitivity to CDK2 inhibition in human breast cancer cells. *Mol. Cancer Ther.* **11**, 1488–1499.

Cantor,R.M. *et al.* (2010) Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *Am. J. Hum. Genet.* **86**, 6–22.

Chatterjee,S. (2020) A new coefficient of correlation. *J. Am. Stat. Assoc.* **116**, 1–39. https://doi.org/10.1080/01621459.2020.1758115.

Chin,W.W. (1998) The partial least squares approach to structural equation modeling. In: Marcoulides,G.A. (ed.) *Modern Methods for Business Research. Lawrence Erlbaum Associates, Mahwah, New Jersey*, pp. 295–336.

Donovan,A. *et al.* (2005) The iron exporter ferroportin/Slc40a1 is essential for iron homeostasis. *Cell Metab.* **1**, 191–200.

Elian,F.A. *et al.* (2018) FOXC1, the new player in the cancer sandbox. *Oncotarget* **9**, 8165–8178.

Fan,J. *et al.* (2009) Ultrahigh dimensional feature selection: beyond the linear model. *J. Mach. Learn. Res.* **10**, 2013–2038.

Furuhashi,M. and Hotamisligil,G.S. (2008) Fatty acid-binding proteins: role in metabolic diseases and potential as drug targets. *Nat. Rev. Drug Discov.* **7**, 489–503.

González,I. *et al.* (2008) CCA: an R package to extend canonical correlation analysis. *J. Stat. Soft.* **23**, 1–4.

Goodall,C. (1991) Procrustes methods in the statistical analysis of shape. *J. R. Stat. Soc. Ser. B Methodol.* **53**, 285–321.

Hardoon,D.R. *et al.* (2004) Canonical correlation analysis: an overview with application to learning methods. *Neural Comput.* **16**, 2639–2664.

Hubert,M. and Branden,K.V. (2003) Robust methods for partial least squares regression. *J. Chemometrics* **17**, 537–549.

Huelsenbeck,J.P. *et al.* (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**, 2310–2314.

Hunt,M.C. *et al.* (2000) Acyl-CoA thioesterases belong to a novel gene family of peroxisome proliferator-regulated enzymes involved in lipid metabolism. *Cell Biochem. Biophys.* **32**, 317–324.

Huson,D.H. *et al.* (2011) Integrative analysis of environmental sequences using MEGAN4. *Genome Res.* **21**, 1552–1560.

Johnstone,I.M. and Titterington,D.M (2009) Statistical challenges of high-dimensional data. *Philos. Trans. Royal Soc.* **367**, 4237–4253. https://doi.org/10.1098/rsta.2009.0159.

Kakiyama,G. *et al.* (2013) Modulation of the fecal bile acid profile by gut microbiota in cirrhosis. *J. Hepatol.* **58**, 949–955.

Kinney,J.B. and Atwal,G.S. (2014) Equitability, mutual information, and the maximal information coefficient. *Proc. Natl. Acad. Sci. U S A* **111**, 3354–3359.

Kostic,A.D. *et al.*; DIABIMMUNE Study Group. (2015) The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host Microbe* **17**, 260–273.

Lewinger,J.P. *et al.* (2007) Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genet. Epidemiol.* **31**, 871–882.

López-Knowles,E. *et al.* (2010) PI3K pathway activation in breast cancer is associated with the basal-like phenotype and cancer-specific mortality. *Int. J. Cancer* **126**, 1121–1131.

Lykou,A. and Whittaker,J. (2010) Sparse CCA using a lasso with positivity constraints. *Comput. Stat. Data Anal.* **54**, 3144–3157.

Lynch,C.J. *et al.* (1995) Role of hepatic carbonic anhydrase in de novo lipogenesis. *Biochem. J.* **310**, 197–202.

Martin,P.G. *et al.* (2007) Novel aspects of PPARα-mediated regulation of lipid and xenobiotic metabolism revealed through a nutrigenomic study. *Hepatology* **45**, 767–777.

McArdle,B.H. and Anderson,M.J. (2001) Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* **82**, 290–297.

Mika,S. *et al.* (1998) Kernel PCA and de-noising in feature spaces. *In NIPS* **11**, 536–542.

Morgan,X.C. *et al.* (2012) Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* **13**, R79.

Morgan,X.C. *et al.* (2015) Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. *Genome Biol.* **16**, 1–5.

Mourad,R. *et al.* (2010) Learning hierarchical Bayesian networks for genome-wide association studies. In: *Proceedings of COMPSTAT'2010*. Physica-Verlag HD, pp. 549–556. https://doi.org/10.1007/978-3-7908-2604-3_56.

Mourad,R. *et al.* (2011) A hierarchical Bayesian network approach for linkage disequilibrium modeling and data-dimensionality reduction prior to genome-wide association studies. *BMC Bioinformatics* **12**, 16–20.

Patel,V.I. *et al.* (2017) Transcriptional classification and functional characterization of human airway macrophage and dendritic cell subsets. *J. Immunol.* **198**, 1183–1201.

Patterson,E. *et al.* (2017) Bifidobacterium breve with α-linolenic acid alters the composition, distribution and transcription factor activity associated with metabolism and absorption of fat. *Sci. Rep.* **7**, 43300–43302.

Pogue-Geile,K.L. *et al.* (2013) Predicting degree of benefit from adjuvant trastuzumab in NSABP trial B-31. *J. Natl. Cancer Inst.* **105**, 1782–1788.

Ronquist,F. and Huelsenbeck,J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574.

Rosenberg,P.S. *et al.* (2006) Multiple hypothesis testing strategies for genetic case–control association studies. *Stat. Med.* **25**, 3134–3149.

Selvaraju,S. *et al.* (2012) Evaluation of maize grain and polyunsaturated fatty acid (PUFA) as energy sources for breeding rams based on hormonal, sperm functional parameters and fertility. *Reprod. Fertil. Dev.* **24**, 669–678.

Simon,N. and Tibshirani,R. (2014) Comment on 'Detecting novel associations in large data sets by Reshef Et Al, Science Dec 16, 2011. *Science*. https://doi.org/10.48550/arXiv.1401.7645.

Skrondal,A. and Rabe-Hesketh,S. (2004) *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. CRC Press, Berlin, Germany. https://doi.org/10.1201/9780203489437.

Sun,L. *et al.* (2009) On the equivalence between canonical correlation analysis and orthonormalized partial least squares. In: *IJCAI. Vol.* **9**, Morgan Kaufmann Publishers Inc., San Francisco, CA, United States, 1230–1235.

Székely,G.J. *et al.* (2007) Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35**, 2769–2794.

Vasaikar,S.V. *et al.* (2018) LinkedOmics: analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Res.* **46**, D956–D963.

Waldron,L. *et al.* (2011) Optimized application of penalized regression methods to diverse genomic data. *Bioinformatics* **27**, 3399–3406.

Weinstein,J.N. *et al.*; Cancer Genome Atlas Research Network. (2013) The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**, 1113–1120.

Werner,T. *et al.* (2011) Depletion of luminal iron alters the gut microbiota and prevents Crohn's disease-like ileitis. *Gut* **60**, 325–333.

Winkler,R.L. (1967) The assessment of prior distributions in Bayesian analysis. *J. Am. Stat. Assoc.* **62**, 776–800.

Witten,D.M. *et al.* (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**, 515–534.

Yekutieli,D. (2008) Hierarchical false discovery rate–controlling methodology. *J. Am. Stat. Assoc.* **103**, 309–316.

Yekutieli,D. and Benjamini,Y. (1999) Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Stat. Plann. Inference* **82**, 171–196.

Zhan,X. *et al.* (2017) A fast small-sample kernel independence test for microbiome community-level association analysis. *Biometrics* **73**, 1453–1463.

Zou,H. *et al.* (2006) Sparse principal component analysis. *J. Comput. Graphical Stat.* **15**, 265–286.