

# 1 High-sensitivity pattern discovery in large, paired

## 2 multi-omic datasets

3

4

5 Andrew R. Ghazi<sup>1,2,3</sup>, Kathleen Sucipto<sup>1</sup>, Gholamali Rahnavard<sup>1,2</sup>, Eric A. Franzosa<sup>1,2,3</sup>,  
6 Lauren J. McIver<sup>1,2,3</sup>, Jason Lloyd-Price<sup>1,2</sup>, Emma Schwager<sup>1</sup>, George Weingart<sup>1,3</sup>, Yo  
7 Sup Moon<sup>1</sup>, Xochitl C. Morgan<sup>4</sup>, Levi Waldron<sup>5</sup>, Curtis Huttenhower<sup>1,2,3,6\*</sup>

<sup>8</sup> <sup>9</sup> <sup>1</sup>Biostatistics Department, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA

10 <sup>2</sup>Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

11 <sup>3</sup>Harvard Chan Microbiome in Public Health Center, Harvard T. H. Chan School of Public  
12 Health, Boston, Massachusetts, USA

13     <sup>4</sup>Department of Microbiology and Immunology, University of Otago, Dunedin, New  
14     Zealand

15 <sup>5</sup>City University of New York Graduate School of Public Health and Health Policy,  
16 New York City, New York, USA

17 <sup>6</sup>Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of  
18 Public Health, Boston Massachusetts, USA

19 \*Corresponding author

20 E-mail: chuttenh@hsph.harvard.edu

21

## 22 **Abstract**

23 Modern biological screens yield enormous numbers of measurements, and identifying  
24 and interpreting statistically significant associations among features is essential. Here,  
25 we present a novel hierarchical framework, HAllA (Hierarchical All-against-All  
26 association testing), for structured association discovery between paired high-  
27 dimensional datasets. HAllA efficiently integrates hierarchical hypothesis testing with  
28 false discovery rate correction to reveal significant linear and non-linear block-wise  
29 relationships among continuous and/or categorical data. We optimized and evaluated  
30 HAllA using heterogeneous synthetic datasets of known association structure, where  
31 HAllA outperformed all-against-all and other block testing approaches across a range of  
32 common similarity measures. We then applied HAllA to a series of real-world multi-omics  
33 datasets, revealing new associations between gene expression and host immune  
34 activity, the microbiome and host transcriptome, metabolomic profiling, and human  
35 health phenotypes. An open-source implementation of HAllA is freely available at  
36 <http://huttenhower.sph.harvard.edu/halla> along with documentation, demo datasets, and  
37 a user group.

## 38 **Author Summary**

39 Modern scientific datasets increasingly include multiple measurements of many  
40 complementary data types. Here, we present HAllA, a method and implementation that  
41 overcomes the statistical challenges presented by data of this type by using feature  
42 similarity within each dataset to find statistically significant groups of features between  
43 them. We applied HAllA to simulated and real datasets, showing that HAllA  
44 outperformed existing procedures and identified compelling biological relationships.  
45 HAllA is widely applicable to diverse data structures and presents the user with grouped  
46 results that are easier to interpret than traditional methods.

47

## 48 Introduction

49 Pattern discovery in high-dimensional, heterogeneous data is a longstanding problem in  
50 applied statistics [1,2]. It is challenging for several reasons, including the inherent  
51 tradeoffs between sensitivity and generality - that is, the ability and power to detect  
52 associations given varying assumptions about the functional form of the relationship [3].  
53 When applied in contexts such as high-throughput biology, these challenges are  
54 exacerbated by noisy, diverse, and non-linear data. Examples include biospecimens  
55 drawn from large cohorts, in which each sample may be decorated with heterogeneous  
56 phenotypic variables (clinical features, diseases status, etc.) and multiple high-  
57 dimensional molecular measurements (microbial taxa, epigenetic markers, gene  
58 expression, etc.). In the biological sciences specifically, selecting a subset of  
59 associations for follow-up validation experiments can be a complex yet important  
60 decision point. A gap remains to efficiently identify related features in such data, while  
61 both maintaining sensitivity and controlling spurious association reporting.

62 All-against-all (AllA) approaches, which test all pairs of features and then correct for  
63 false discovery, scale well only in completely independent tests of moderate size [4].  
64 Under other conditions, such feature-wise approaches can have limited statistical power  
65 due to testing many correlated hypotheses for individually weak associations [5]. This  
66 has led to the development of a variety of (typically parametric) block-testing  
67 approaches, such as partial least squares (PLS) [6], canonical correlation analysis  
68 (CCA) [7], PLS discriminant analysis (PLS-DA), sparse principal component analysis  
69 (SPCA) [8], and SPARSE-CCA [9]. These serve to detect associations between  
70 reduced-dimensional representations of large input datasets, but they are typically  
71 limited by one or more of 1) applicability only to continuous measurements with no  
72 missing values (or only categorical, not mixed; PLS, CCA, SPCA); 2) a focus on the  
73 single, strongest axis of covariation between the datasets (CCA); 3) an assumption of  
74 linear covariation (CCA, SPCA, PLS); 4) identifying complex combinations of feature  
75 loadings implicated in associations, rather than specific features (particularly in kernel  
76 methods such as Kernel PCA [10]); and 5) a lack of explicit control of the false discovery  
77 rate (FDR).

78 Recent advances have focused on nonparametric methods for identifying highly general  
79 (i.e., linear and non-linear) associations between individual pairs of features, sometimes  
80 relying on computational or permutation-based methods not readily accessible to early  
81 applied statisticians. These include, for example, distance correlation (dCor) [11], which  
82 measures (not necessarily linear) dependency of two random variables with possibly  
83 different dimensions. The Chatterjee rank correlation (XICOR) [12] is another recently-  
84 introduced similarity measure that uses rank differences to assess the degree to which  
85 one variable is a measurable function of another. While dCor and XICOR provide  
86 comparatively general methods to discover complex associations between individual  
87 pairs of features, when applied to many combinations of linear feature pairs with varying  
88 degrees of dependence, the resulting statistical power can fall below simpler traditional  
89 approaches after controlling FDR for multiple hypothesis tests [13].

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

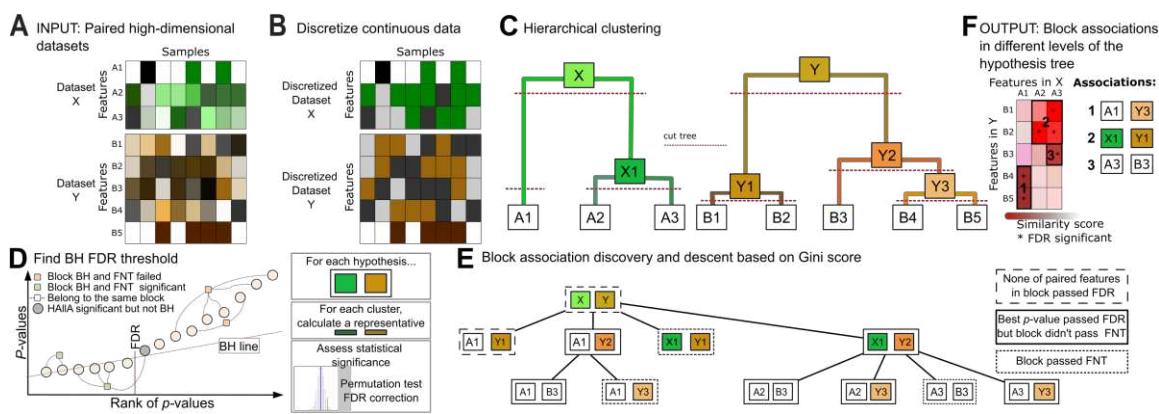
## 97 **Methods**

98 In this section, we provide an overview of the HALIA algorithm and its component steps.  
99 Additional methods details, including pseudocode, are provided in S1 Appendix.

### 100 **The HALIA Algorithm.**

101 Hierarchical All-against-All Association testing (HALIA) identifies block associations  
102 between two potentially heterogeneous datasets co-indexed along one axis (Fig 1A).  
103 This co-indexing is referred to as the "samples" axis (columns), and the measurement  
104 axis as "features" (rows). For a pair of datasets containing measurements that describe  
105 the same set of samples and a specified pairwise similarity measure, the HALIA algorithm

106 proceeds by 1) optionally discretizing features to a uniform representation (if required by  
 107 the similarity measure), 2) finding the Benjamini–Hochberg (BH) FDR threshold, 3)  
 108 hierarchically clustering each dataset separately to generate two data hierarchies, 4)  
 109 coupling clusters of equivalent resolution between the two data hierarchies, 5) testing  
 110 coupled clusters for statistically significant association in block format where the block  
 111 passes a threshold for false negative tolerance (FNT), and 6) iteratively increasing  
 112 resolution by descending through the pair of hierarchies according to which split results  
 113 in a higher Gini score gain. The final pair of hierarchies are those that lead to the largest  
 114 hypothesis blocks that pass the FNT threshold (Fig 1 and S1 Appendix).



115  
 116 **Figure 1. Hierarchical all-against-all (HALIA) association testing.** (A) HALIA provides  
 117 a novel method for heterogeneous association discovery in high dimensional data. Input  
 118 data are represented in matrix form as features (rows) and samples (columns). (B) Data  
 119 are discretized to provide a unified representation of heterogeneous feature types. This  
 120 step is skipped for similarity metrics that requires continuous data (e.g. Spearman). (C)  
 121 Features within each data set are hierarchically clustered using average linkage and  
 122 Spearman association as default methods. (D) Reject block-wise null hypotheses that  
 123 pass the false negative tolerance (FNT) threshold using Benjamini-Hochberg FDR  
 124 threshold for pair-wise associations within the block. (E) Block format hypotheses are  
 125 built by pairing clusters between two datasets at equivalent relative homogeneity. Each  
 126 hypothesis node has two data clusters whose descendants are used for the next level of  
 127 hypothesis testing. In hypothesis testing, the FNT threshold is used to determine which  
 128 clusters are significantly associated between the two datasets. (F) Significant  
 129 associations are reported after controlling the FDR for each hypothesis set in the  
 130 descending approach using hypothesis tree-oriented structure.

131 **Optionally discretizing input datasets.**

132 This step permits direct comparison of continuous and categorical features (Fig 1B) and  
 133 further enables the application of highly general measures of association from

134 information theory, such as mutual information (MI). This combination allows HALIA to  
135 detect significant 1) non-linear associations between paired continuous features (e.g.,  
136 quadratic or sinusoidal relationships), 2) differences in group means for paired  
137 continuous and categorical features, and 3) non-random associations between paired  
138 categorical features. HALIA's default discretization scheme divides continuous features  
139 into bins of equal size once at the start of processing. By default, the number of bins is  
140 the cube root of the sample size, which provides reasonable power at a variety of  
141 sample sizes and correlation levels (Fig 1 in S1 Appendix). HALIA also removes features  
142 with low variance by applying a configurable frequency threshold (defaulting to 100%,  
143 meaning only features with no variability are removed) in order to reduce the number of  
144 unnecessary tests.

145 **Hierarchical clustering and cluster coupling allow detection of associations**  
146 **between groups of features.**

147 Each dataset is subjected to average-linkage hierarchical clustering using the specified  
148 similarity measure (Spearman's rank correlation by default) within each dataset (Fig 1C).  
149 Associations between datasets are tested in a top-down manner by pairing nodes of  
150 similar resolution between their respective data trees. More specifically, HALIA  
151 recursively builds a tree of hypotheses to test (the "hypothesis tree"), beginning at the  
152 top of each dataset's tree, descending to a set of nodes within each data tree, and then  
153 pairing each selected node from the first tree with each selected node of the second  
154 tree. At each step in the descent process, the choice of whether to descend within the X  
155 or Y hypothesis tree is made by comparing which split leads to a higher Gini score gain.  
156 In the case of ties, both descent steps are made. This procedure is repeated until  
157 termination, i.e. when the hypothesis block passes the FNT threshold or when the  
158 selected nodes represent single features in their respective data trees (Fig 1E). Another  
159 way to visualize this process is by focusing on the all-by-all hypothesis matrix (Fig 1F,  
160 left). The process begins by checking if the entire matrix passes the FNT threshold. If  
161 not, the matrix is recursively cut horizontally or vertically into smaller hypothesis blocks,  
162 with the position of each cut decided by each dataset's similarity tree and Gini score  
163 gain. The cutting process stops when the smaller hypothesis blocks pass the FNT  
164 threshold or have been reduced to one-by-one blocks.

165 The notion of identifying and testing hypotheses in a hierarchical manner was previously  
166 proposed by Yekutieli [14]. HAlIA's hypothesis tree similarly groups more specific child  
167 hypotheses below a more general parent hypothesis. However, HAlIA's approach differs  
168 fundamentally from the Yekutieli approach in that HAlIA tests hierarchical hypotheses  
169 until a null hypothesis can be rejected; Yekutieli's method tests until the first failure to  
170 reject a null hypothesis. This results in HAlIA maintaining greater power, while Yekutieli's  
171 method instead maintains greater specificity.

172 **Determining the statistical significance of block associations.**

173 The method proceeds by testing the nodes in the hypothesis tree (each representing a  
174 pair of feature clusters, one from each dataset) for significant between-cluster  
175 associations. Each node in the hypothesis tree is evaluated using the following  
176 procedure: let  $\mathcal{H}$  denote the null hypothesis that the two clusters of features are not  
177 related, and  $\mathcal{H}_i$  be the null hypothesis of no association between two individual features  
178 within those clusters. Define  $R^i$  as the p-value of the association between an individual  
179 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  
180  $\mathcal{H}_i$  that failed to reject, i.e.  $R^i > k_{BH}$  where  $k_{BH}$  is the global BH FDR threshold. The  
181 blockwise FNT is provided by the user (default FNT = 0.2) and acts as the allowed  
182 fraction of paired associations which are expected to fail to reject despite being true  
183 associations. If the fraction of paired associations in a block with  $R^i > k_{BH}$  is greater than  
184 or equal to FNT, we reject the entire block hypothesis  $\mathcal{H}$ .

185 If any hypothesis involved clusters rather than feature tips, and failed to reject, the  
186 procedure is repeated with new null hypotheses for associations between sub-clusters  
187 (Fig 1E), as described in section "Descending in sub-hypotheses of block hypotheses" in  
188 S1 Appendix. HAlIA reports all significant associations between clusters of any size that  
189 pass the FNT threshold (Fig 1F).

190 **Visualizing outputs**

191 Once the analysis is complete, the results are visualized in a "HAlIAgram" (Fig 4). This  
192 comprises a heatmap visualizing the relatedness and strength of association between

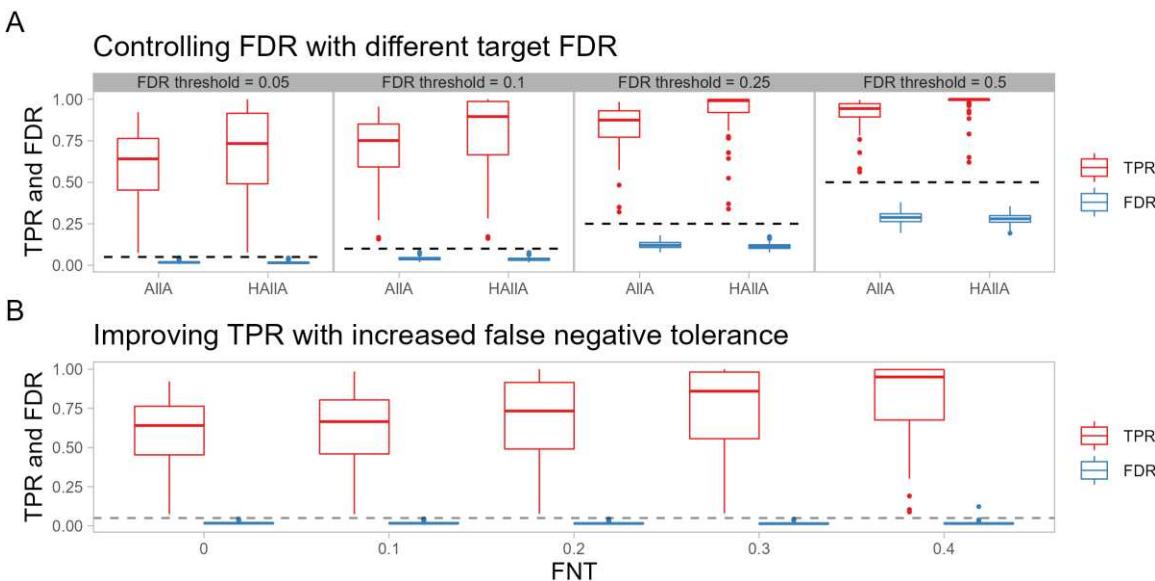
193 pairs of features in the two datasets. Features are ordered along each axis according to  
194 their position in the hierarchical tree so that clusters of significant features can be boxed  
195 into contiguous units. Marginally associated pairs are dotted, and each hypothesis block  
196 is labelled with the rank of its association strength. Features not associated with any  
197 block are not plotted by default. For analysis results where large numbers of blocks are  
198 detected, only the strongest blocks are shown (30 by default), with potentially-  
199 incomplete, lower-ranked blocks boxed in grey. Together, this set of plotting techniques  
200 allows users to visually understand the related sets of hypotheses that HALIA has  
201 detected. Other plotting utilities are also included with the method's current  
202 implementation, such as a clustermap that displays the entire association tree in the  
203 margins for both datasets, as well as a diagnostic plot that displays the input data  
204 associated with individual hypothesis blocks.

205 **Results**

206 **HALIA increases power while controlling FDR to report blockwise associations**

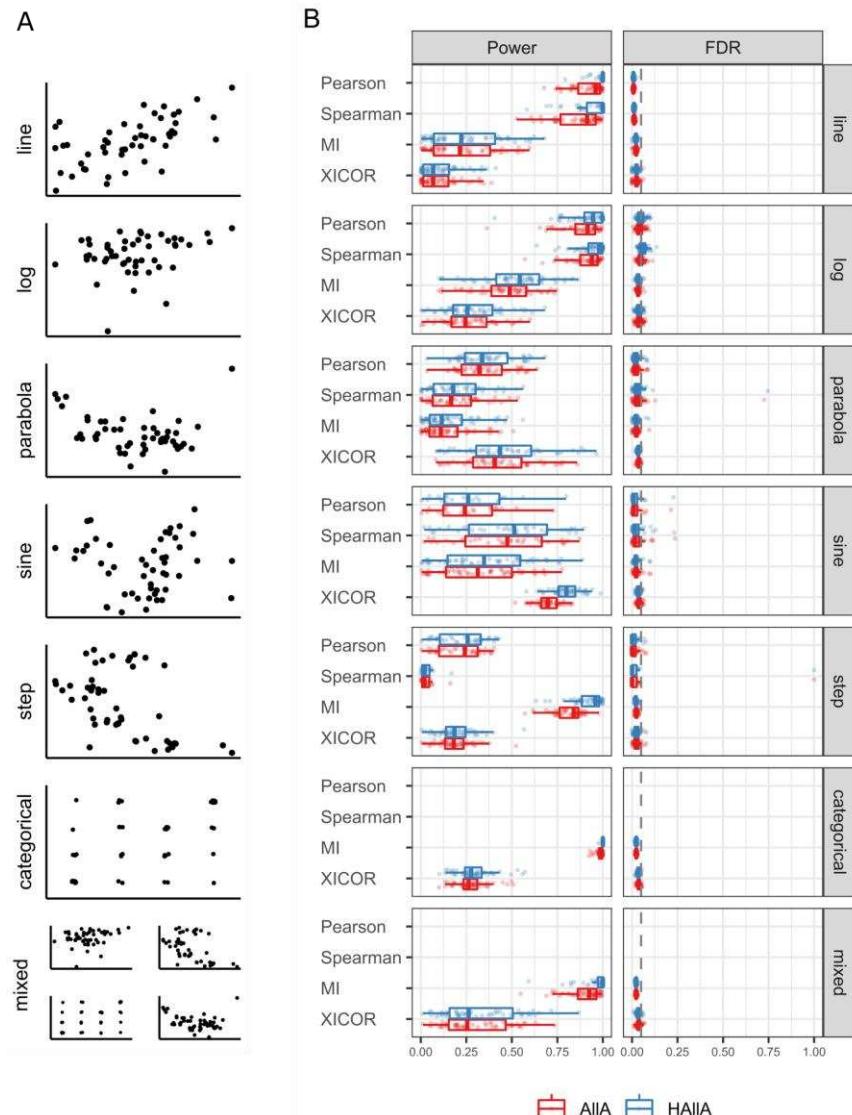
207 When applied to paired datasets with no significantly related blocks of features, HALIA's  
208 descent algorithm reduces to all-against-all (AlIA) direct pairwise feature testing. In such  
209 circumstances, HALIA is expected to perform similarly to AlIA. However, when there are  
210 sets of correlated variables within one dataset that are correlated with another set of  
211 variables in the other, HALIA will report the block-wise associations. Notably, we expect  
212 this behavior to be common in multi-omics data, where we see large clusters of  
213 molecular features (e.g. co-expressed genes in a metabolic pathway).

214 To evaluate these claims, we applied HALIA and AlIA to paired, synthetic datasets  
215 generated with the data simulator function in the HALIA software. These datasets  
216 contained pre-specified block associations, which allowed us to evaluate the statistical  
217 and computational performance of these two methods (Fig 2 and Fig 3). With a constant  
218 target FNT in associated blocks of 0.2, HALIA controls FDR, reports association in block  
219 form, and improves power on average by 7-11% (Fig 2A) across varied FDR thresholds.  
220 HALIA also consistently boosts the true positive rate relative to AlIA using different target  
221 FNT values in associated blocks (Fig 2B).



222  
223  
224  
225  
226  
227  
228  
229  
230

**Figure 2. HAIIA improves statistical power while controlling the FDR.** 50 paired, synthetic datasets with 200 features and 50 samples containing clusters with linear block associations were analyzed. A) with FNT = 0.2, HAIIA maintains the simulated FDR below the target (here (0.05, 0.1, 0.25, and 0.5), with associated trade-offs in statistical power. In addition, HAIIA is consistently better powered than all-against-all (AIIA) 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, HAIIA can improve the true positive rate with a comparatively minor increase in FDR.



231

232 **Figure 3. HAIIA discovers block-structured associations while controlling false**  
233 **discovery rate.** For a variety of feature linkage relationships, we simulated 50  
234 independent paired datasets, each containing 200 features, 50 samples, and clusters of  
235 correlated features. We then evaluated the ability of hierarchical versus all-against-all  
236 testing to recover these associations using a variety of similarity metrics. Performance  
237 was evaluated by comparing power and false discovery rates. Our hierarchical all-  
238 against-all approach improved sensitivity relative to naive all-against-all approaches at a  
239 comparable false discovery rate. Similarity metrics that don't accept categorical data  
240 have not been evaluated in the categorical or mixed association type. Other similarity  
241 metrics included in HAIIA (dCor, NMI) were not applied in these simulations because  
242 their reliance on permutation tests made them too slow for simulations of this size (i.e.  
243 with many repeated iterations), although they are typically practical in individual real-  
244 world datasets.

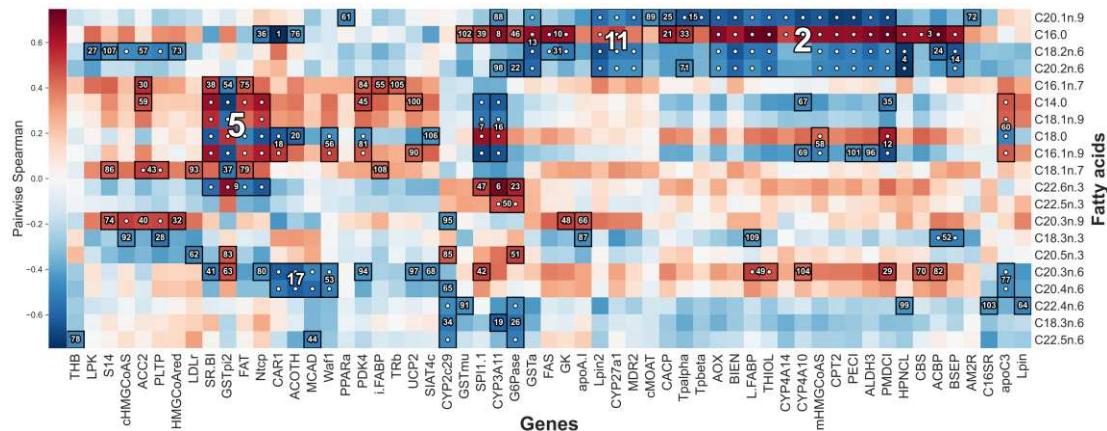
245

246 We evaluated many different forms of feature association, including linear, quadratic,  
247 logarithmic, sinusoidal, stepwise, parabolic, and mixed (combined discrete and  
248 continuous) data. We compared HAlIA and AlIA across these association types using a  
249 variety of similarity measures, including XICOR, mutual information (MI), Spearman  
250 correlation, and Pearson correlation. Across datasets and similarity measures, HAlIA  
251 consistently detected more built-in associations (had better average power by as much  
252 as 10%) than AlIA while controlling FDR at the same pre-specified level (Fig 3B). Each  
253 similarity measure exhibited various strengths and weaknesses across evaluations  
254 depending on data type. As expected, for mixed and categorical data, MI is appropriate,  
255 and for monotonic associations in continuous data, Spearman correlation performs well.  
256 XICOR is applicable to both continuous and discrete outcomes and performs well on  
257 difficult nonlinear association types. However, it is rarely the most statistically powerful  
258 option, and its interpretation is limited to measuring the association of features in Y as a  
259 measurable function of features in X and not vice versa. A similar power analysis that  
260 used a fixed association structure with varying correlation strength led to similar  
261 conclusions (Fig 2 in S1 Appendix). Together these results show that the HAlIA  
262 approach increases statistical power while maintaining the FDR across a wide variety of  
263 association structures under simulation.

264 **HAlIA identifies novel fatty acid-xenobiotic metabolism associations in PPAR $\alpha$ -  
265 deficient mice**

266 PPAR $\alpha$  is a nuclear receptor that regulates transcription of genes related to lipid  
267 metabolism in the liver [15]. These genes show high fatty acid catabolism rates, which  
268 can in turn affect hepatic fat storage and lipoprotein metabolism. We used HAlIA to  
269 examine associations between 120 hepatic transcript levels and 21 liver lipid levels in a  
270 previously published dataset [16] (Fig 4). These data were originally collected from 40  
271 wild type and peroxisome proliferator-activated receptor- $\alpha$  (PPAR $\alpha$ )-deficient mice [15].  
272 HAlIA recovered 109 block associations comprising 225 pairwise associations at target  
273 FDR of 0.05 (chosen to match the previous study). HAlIA's results included all  
274 associations that were previously reported using canonical correlation analysis, including

275 a key relationship between fatty acids and the xenobiotic metabolism genes Cyp3a11  
276 and Car1(MGI:88268).



277  
278 **Figure 4. Association of fatty acids with host transcriptional activity in murine**  
279 **liver.** We applied HAIIA to paired data comprising 120 hepatic transcript levels and 21  
280 liver lipid levels in a set of 40 previously profiled mice [15]. In this “HAIIGram”  
281 visualization of results, block associations are numbered in descending order of  
282 significance, with each numbered block corresponding to a group of co-expressed  
283 transcripts related to a group of co-occurring lipids. A white dot indicates marginal  
284 significance of a particular pair of features. A total of 109 block associations achieved  
285 significance at FDR 0.05, matching the previous study’s threshold based on canonical  
286 correlation [16] (detailed in S1 Table). HAIIA’s associations were a strict superset of  
287 those found earlier by CCA. Spearman correlation was used as a similarity metric.

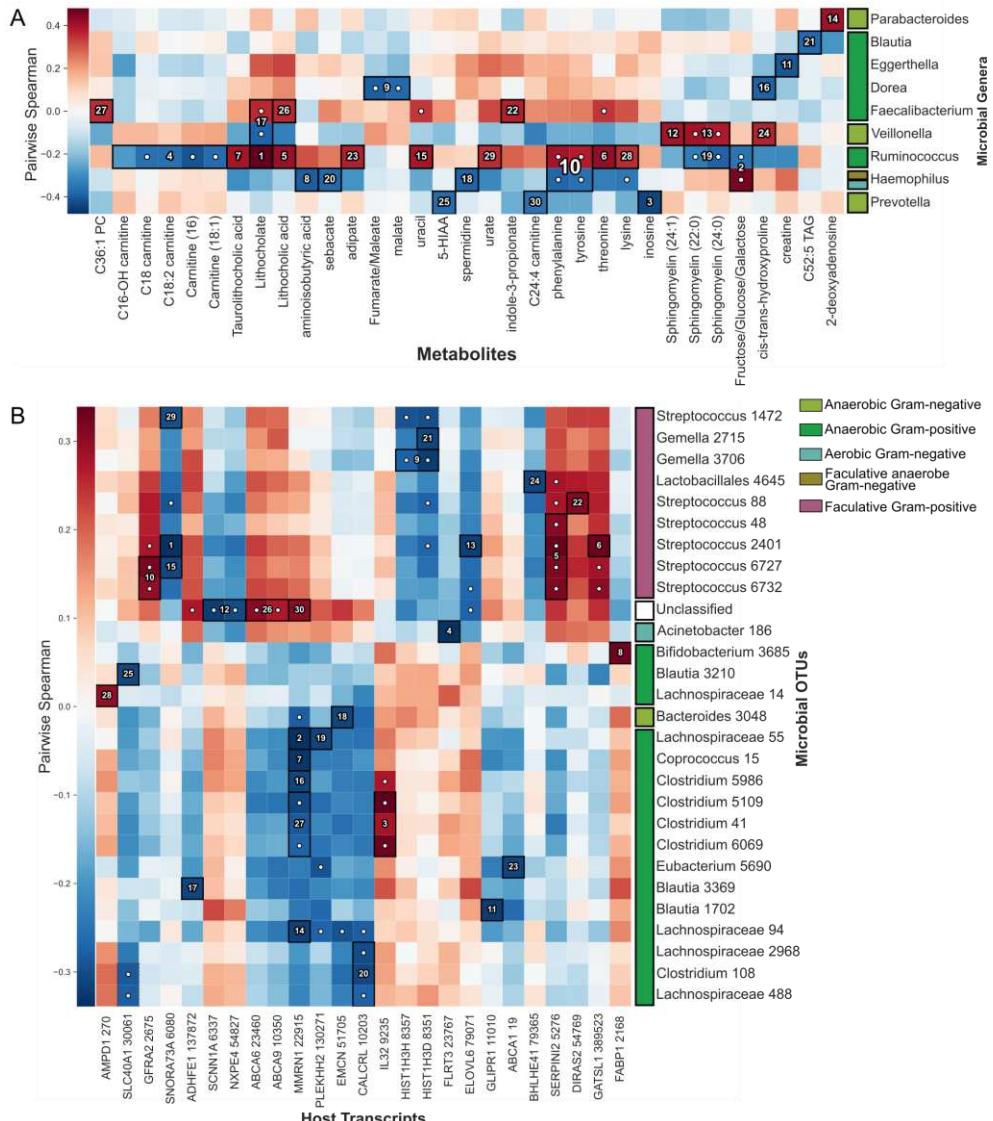
288 We further identified several novel associations, including a link between  
289 polyunsaturated fatty acids eicosatrienoic acid (C20:3n6) and arachidonic acid  
290 (C20:4n6) [17] with a group of transcripts including *Mcad* (*Acadm*, MGI:87867). This  
291 gene (C-4 to C-12 straight chain acyl-Coenzyme A dehydrogenase) encodes one of the  
292 main catalysts of the beta-oxidation process used for degradation of these fatty acids.  
293 Genes *Car1* (MGI:88268) and *Acot11* (MGI:1913736) (a carbonic anhydrase and lipid  
294 transfer protein, respectively [18-19]) fell in the same cluster with C20.3n.6 and  
295 C20.4n.6, which would suggest a trafficking and transport relationship between these  
296 genes and fatty acids.

## 297 **Associating microbes with metabolites in the infant gut microbiome**

298 In a prior study, Kostic and colleagues examined the development of the human gut  
299 microbiome in a prospective, longitudinally sampled cohort of 33 Finnish and Estonian

300 infants at high risk for type-1 diabetes [20]. Stool samples and clinical metadata (e.g.  
301 breastfeeding status, diet, and appearance of allergies) were collected monthly.  
302 Subjects' stool samples were subsequently analyzed using 1) 16S rRNA amplicon  
303 sequencing (to profile gut microbiome composition) and 2) targeted mass spectrometry  
304 (to profile host and microbial metabolites). The dataset included 103 samples from 19  
305 individuals, each with paired metabolomics and 16S rRNA gene sequencing data. We  
306 applied HALIA to identify associations between the residuals of microbial and metabolite  
307 abundances after correcting for longitudinal trends and subject specific random effects  
308 using a linear mixed effects model [21] (S1 Appendix).

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



314

315 **Figure 5. HAIIGram for block-wise associations. a) Using HAIIA to associate**

316 multi-omic data for the analysis of metabolome-microbiome interactions. We used

317 HAIIA to associate paired stool metabolomic and 16S rRNA gene sequencing data from

318 the DIABIMMUNE [20] cohort, in which infants were recruited at birth and sampled

319 monthly for the first three years of life. The data comprise 104 samples and describes

320 the abundance of 20 genera and 284 labeled metabolites. Here, we show the 30

321 strongest associations ranked by *p*-value (target FDR=0.05). **b) Relating host**

322 transcriptome and microbial taxa in IBD patients. We applied HAIIA to identify

323 associations between the human gut microbiome and transcriptome in 204 patients

324 receiving ileal pouch-anal anastomosis (IPAA) surgeries [23]. Block associations are

325 numbered in descending order of significance based on best *p*-values in each block with

326 each numbered block corresponding to a group of co-expressed transcripts related to a

327 group of co-occurring microbial taxa (operational taxonomic units, OTUs).

14

328 Our results again replicate all significant associations from the previous study's  
329 canonical correlation analysis (CCA), and most of the associations from the original  
330 pairwise association analysis of the previous paper. HALIA also found additional  
331 associations, including a novel association between *Prevotella* and inosine (Spearman  
332 coefficient = -0.439, FDR Q-value = 0.0053), which could be explained by a mechanism  
333 where increased levels of urotoxins in the body from inosine decreased the abundance  
334 of intolerant *Prevotella*. HALIA also reports novel associations between fecal bile acids  
335 lithocholate and lithocholic acid and genera *Faecalibacterium* and *Veillonella* (Spearman  
336 coefficients = 0.36, -0.39; Q-values = 0.026, 0.015, respectively). *Faecalibacterium* is  
337 Gram-positive anaerobic bacteria genera from order Clostridiales, while *Veillonella* are  
338 Gram-negative anaerobic cocci. Relationships between these genera and global bile  
339 acid levels (with matching correlation signs) has been previously indicated by several  
340 studies, particularly in cirrhosis [22]. These data thus demonstrate HALIA's potential  
341 benefits relative to pairwise or omnibus (e.g. CCA) testing by simultaneously providing  
342 both greater interpretability and power.

### 343 **Associating the gut microbiome with host transcription in ulcerative colitis**

344 We next applied HALIA to data combining 1) 16S rRNA amplicon sequencing of the  
345 human gut microbiome and 2) Affymetrix microarray screens of ileal RNA expression  
346 across 204 individuals in a cohort of ileal pouch-anal anastomosis (IPAA) patients [23].  
347 In the original multivariate analysis of these data [24], microbial operational taxonomic  
348 unit (OTU) abundances were decomposed into principal components (PCs), and PCs  
349 accounting for up to 50% of the variance in the datasets were compared by all-against-  
350 all testing (an example of PC regression). While this approach enables well-powered  
351 comparisons of large numbers of features, the features are embedded as loadings in  
352 PCs, which complicates biological interpretation of the resulting associations.

353 HALIA identified 327 block associations in these microbial and gene expression data  
354 using an FDR threshold of 0.05 and a FNT of 0.1 (Fig 5B and S2 Table). Total  
355 relationships encompassed 125 OTUs, 187 transcripts, and the equivalent of 368  
356 pairwise associations. The original study focused on the 9<sup>th</sup> principal component (PC9)  
357 of the dataset due to its linking of a group of IL12/complement pathways to members of

358 the microbiome, using an FDR threshold of 0.25. Of HALIA's reported microbe-transcript  
359 associations when run with the same threshold, 20 genes were drawn from the 26  
360 transcripts whose largest loading was in PC9. HALIA's findings support a surprising result  
361 of the original study: although PC9 represented only 1% of the transcriptional variation in  
362 these samples, it captured most associations between transcription and the microbiome  
363 during pouchitis. These results also agree with a previous re-analysis of these data [25]  
364 assessing global covariation between gut microbial and transcriptional structure, which  
365 called out three pathways (interleukin-12, inflammatory, and inflammatory bowel disease  
366 genes) that overlap heavily with HALIA's block results (e.g. 28 out of 51 tested genes in  
367 the KEGG TRP channel mediator pathway and 34 of 61 tested genes in the KEGG IBD  
368 pathway were significantly associated with microbial species).

369 Expanding on these previous associations, HALIA found a group of facultative anaerobes  
370 (mainly streptococci) to be positively associated with expression of the genes WDR49  
371 and SERPINI2. WDR49 is a WD repeat-containing protein upregulated in alveolar  
372 macrophages, a cell type specifically responsible for nasopharyngeal pathogen uptake  
373 [26]. This association suggests this protein may also be involved in recognition of  
374 bacteria in the gut environment. Another novel association in HALIA's results linked a  
375 group of *Bifidobacterium* OTUs with FABP1, a member of the long-chain fatty acid  
376 binding protein family involved both in lipid sensing and metabolic regulation of energy  
377 harvest [27]. This positive relationship has also been observed in mice [28]. Finally,  
378 during intestinal inflammation and bleeding, host-microbial iron competition is a limiting  
379 factor in subsets of microbial growth [29], which may be responsible for the significant  
380 negative association identified between the siderophore-rich genus *Blautia* and  
381 SLC40A1, a human intestinal epithelial iron ion transmembrane transporter [30].

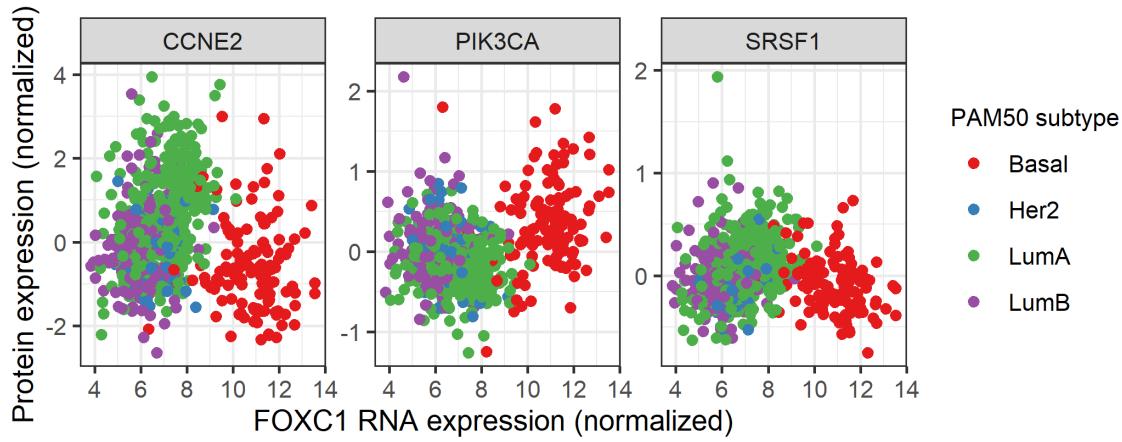
382

### 383 **HALIA's applicability to heterogeneous datasets**

384 We finally applied HALIA to identify associations between mixed clinical metadata and  
385 RNA expression in the breast cancer cohort of the Cancer Genome Atlas (TCGA) [31]  
386 available from the LinkedOmics R package [32], focusing on highly expressed yet  
387 variable transcripts (Fig 3 in S1 Appendix). HALIA identified 483 significant (Q-value <  
16

388 0.1) metadata-RNA associations within 261 blocks, including clusters of transcripts  
389 associated with tumor purity, PAM50 subtype, and ER Status. Notably, the transcripts  
390 occupying the block associated with PAM50 subtype include CA12, GABRP, NAT1, and  
391 TBC1D9, which have been previously proposed as predictor genes for breast cancer  
392 mortality, recurrence [33], and drug response [34]. Coupled with the results of the  
393 preceding applications, these results speak to the generality of HALIA's association-  
394 discovery power across large, heterogeneous datasets.

395 In order to demonstrate the usefulness of alternative similarity measures like XICOR, we  
396 decided to look for non-linear functional relationships between RNA and protein  
397 expression in the breast cancer cohort of the Cancer Genome Atlas (TCGA) [31]. We  
398 applied HALIA to this data using both Spearman and XICOR as similarity measures, then  
399 examined the significant associations that came out with the latter but not the former.  
400 Among these we noticed three associations between RNA expression of transcription  
401 factor FOXC1 and protein expression of CCNE2, PIK3CA, and SRSF1 (FDR Q-value =  
402  $9.3 \times 10^{-7}$ ,  $3.9 \times 10^{-5}$ , 0.015, respectively) which showed compelling U-shaped relationships  
403 (Fig 6). When compared with PAM50 clinical subtypes, these relationships emerge as a  
404 result of two features of the originating tumors. First, the different PAM50 subtypes vary  
405 in average FOXC1 expression (i.e. average position on the x-axis). Secondly, the effect  
406 of FOXC1 on the expression of each protein appears to vary between the subtypes, with  
407 the opposite sign in the basal subtype. There are individually well-established links  
408 between subtype and FOXC1, CCNE2, and PIK3CA [35-37]. However, the varying  
409 relationship of each protein with FOXC1 by subtype has seemingly gone unnoticed in  
410 the literature, presumably due to the marginally non-linear shape of the overall  
411 relationship. While further study of the clinical importance of these relationships is  
412 warranted, these findings demonstrate the ease of well-powered, flexible, nonlinear  
413 association discovery with HALIA.



414

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

423 **Discussion**

424 In this work, we proposed and validated HALIA, a novel statistical method to find  
425 associations between multi-omic datasets. HALIA addresses several important  
426 methodological challenges in the analysis of high-dimensional datasets. It is applicable  
427 to data that are heterogeneous both within and between experiments, and it maintains  
428 statistical power using a novel hierarchical association testing and FDR control  
429 procedure. In this method, groups of correlated tests are modeled as blocks, ultimately  
430 reporting associations within blocks and between block representatives from multiple  
431 data types and experiments. This permits both great flexibility in the types of

432 measurements to which it is applied and ease of interpretation of the resulting significant  
433 associations.

434 Class prediction approaches are commonly used to model relationships between high-  
435 dimensional datasets with variables measured using shared observational units. For  
436 example, Partial Least Squares [38] and its close relative Canonical Correlation Analysis  
437 [39] identify latent variables in one dataset that are maximally correlated to latent  
438 variables in the other dataset. These methods, and robust and penalized varieties [40-  
439 41], can identify blocks of variables that are correlated within one dataset and in turn  
440 with another block of correlated variables in another dataset. They do not, however,  
441 control for family-wise error or FDR, and so are most suitable for prediction or  
442 exploratory, visual, and descriptive analysis. With these methods, inference on the  
443 existence of associations between the variables of two datasets against null hypotheses  
444 of independence still relies on univariate hypothesis tests (and possibly dimension  
445 reduction or clustering) and is performed subsequently in a separate step. The FDR for  
446 the potentially large number of tests can be controlled by the Benjamini and Hochberg  
447 method [42], which has been adapted for dependent tests [43] and hierarchically  
448 organized tests [44] that are continued until non-significance. The approach described  
449 here thus aims to combine the best features of these different existing approaches,  
450 yielding clustering of potentially heterogeneous variable types within each dataset with  
451 hierarchical testing and control of FDR.

452 While these approaches are frequentist, Bayesian models are also used to improve  
453 power and share information among feature blocks [45-48]. While such methods are  
454 extremely powerful within their target domains, they are typically intended for  
455 incorporation of specific prior knowledge, such as graph structure [44, 49], phylogeny  
456 [50], or pathway-based functional roles [51]. They can also be computationally  
457 expensive in cases where many or long simulation chains are required for convergence  
458 [52]. HAIIA's nonparametric frequentist approach will likely result in reduced power  
459 relative to such models within the domains for which they are designed, but with  
460 substantially reduced computational cost and without the need to specify model  
461 relationships and priors in each new application domain. Like most statistical tradeoffs,  
462 HAIIA's generality as a tool for association discovery thus comes at a cost in specific

463 circumstances where it is desirable to instead utilize prior knowledge and known data  
464 structure.

465 A limitation of the current method is that it can only look for associations between two  
466 datasets at a time. While the method can be applied to multiple pairs of joint datasets  
467 manually, this becomes combinatorially prohibitive in particularly thorough studies where  
468 a large number of high-dimensional data types are available (e.g. studies which collect  
469 genetics, gene expression, epigenetics, microbial profiles, metabolites, and metadata  
470 from each sample). In circumstances such as these, repeated application of HALIA  
471 across each pair of datasets would no longer properly control FDR. A potential extension  
472 would be to incorporate multivariate testing directly as an association measure, e.g.  
473 block PERMANOVA [53-54] or Procrustes analysis [55], to lower the combinatorial  
474 burden by performing inference on sets of features rather than individual feature pairs.  
475 Second, the model does not share information between blocks, as would be the case in  
476 a fully multivariate test [53] or a hierarchical Bayesian model [48]. Cases in which data  
477 do include such multi-layered nonindependence structure may indeed be better handled  
478 in a Bayesian framework. Finally, and relatedly, it is not straightforward to incorporate  
479 any type of prior knowledge into the HALIA framework, again because of HALIA's intention  
480 for wide applicability. Pre-filtering can be used, as in several of our own examples, but  
481 this can be either beneficial or detrimental depending on context [56-57].

482 Future work could also provide several refinements to the method, in addition to  
483 addressing these limitations. Currently, for example, known but undesirable confounders  
484 must be separately regressed out prior to using HALIA, and the method run on the  
485 resulting residuals instead of raw data. Integrating such covariate adjustment would be  
486 possible in future versions of the method's implementation. Perhaps most importantly, it  
487 may be possible to place tighter theoretical bounds on the block-wise and global FDR  
488 control beyond what is provided by HALIA's adaptation of the Benjamini-Hochberg [42]  
489 and Benjamini-Yekutieli methods [58]. This would also suggest a theoretical framework  
490 within which to characterize the amount and types of non-independence best handled by  
491 hierarchical block association testing. Ultimately, tradeoffs must be made between  
492 power and generality [59]. However, we aim for HALIA to provide a happy medium,

493 capable of serving as an easy-to-use first pass analysis in a wide range of multi-omics  
494 data types.

## 495 Acknowledgements

496 We thank Alex Kostic, Tommi Vatanen, and Vincent Carey for assistance obtaining and  
497 curating datasets for the applications section; Hera Vlamakis, Hector Corrada Bravo,  
498 William Shannon, A. Brantley Hall, Himel Mallick, Siyuan Ma, and Susan Holmes for  
499 helpful discussions, suggestions, and feedback. This study was supported by Army  
500 Research Office grant W911NF-11-1-0429, NSF DBI-1053486, and NIH U54DE023798  
501 to Curtis Huttenhower.

## 502 Author Contributions

503 G.R., E.F., L.W., and C.H. conceived the method; G.R., K.S., and A.G. implemented the  
504 software; G.R., K.S., A.G., and L.M. tested and packaged the software. G.R., A.G., and  
505 E.F. evaluated the performance; G.R., G.W., A.G., and L.M. provide online documents  
506 and software. G.R., J.L.-P., Y.M., A.G., and X.M. prepared synthetic data and  
507 applications. G.R., E.F., A.G. and C.H. wrote the manuscript. All authors discussed the  
508 results and commented on the paper.

## 509 References

- 510 1. Bühlmann P, Van De Geer S. Statistics for high-dimensional data:  
511 methods, theory and applications. Springer Science & Business Media;  
512 2011 Jun 8. doi: 10.1007/978-3-642-20192-9
- 513 2. Johnstone IM, Titterington DM. Statistical challenges of high-dimensional  
514 data. doi: 10.1098/rsta.2009.0159
- 515 3. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. BMJ:  
516 British Medical Journal. 1994 Jun 11;308(6943):1552. doi:  
517 10.1136/bmj.308.6943.1552

518 4. Bourgon R, Gentleman R, Huber W. Independent filtering increases  
519 detection power for high-throughput experiments. *Proceedings of the*  
520 *National Academy of Sciences*. 2010 May 25;107(21):9546-51. doi:  
521 10.1073/pnas.0914005107

522 5. Rosenberg PS, Che A, Chen BE. Multiple hypothesis testing strategies for  
523 genetic case-control association studies. *Statistics in medicine*. 2006 Sep  
524 30;25(18):3134-49. doi: 10.1002/sim.2407

525 6. Abdi H. Partial least squares regression and projection on latent structure  
526 regression (PLS Regression). *Wiley interdisciplinary reviews: computational statistics*. 2010 Jan;2(1):97-106. doi: 10.1002/wics.51

528 7. Hardoon DR, Szedmak S, Shawe-Taylor J. Canonical correlation analysis:  
529 An overview with application to learning methods. *Neural computation*.  
530 2004 Dec 1;16(12):2639-64. doi: 10.1162/0899766042321814

531 8. Zou H, Hastie T, Tibshirani R. Sparse principal component analysis.  
532 *Journal of computational and graphical statistics*. 2006 Jun 1;15(2):265-  
533 86. doi: 10.1198/106186006X113430

534 9. Lykou A, Whittaker J. Sparse CCA using a Lasso with positivity  
535 constraints. *Computational Statistics & Data Analysis*. 2010 Dec  
536 1;54(12):3144-57. doi: 10.1016/j.csda.2009.08.002

537 10. Mika S, Schölkopf B, Smola AJ, Müller KR, Scholz M, Rätsch G. Kernel  
538 PCA and De-noising in feature spaces. InNIPS 1998 Dec 1 (Vol. 11, pp.  
539 536-542).

540 11. Székely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by  
541 correlation of distances. *The Annals of Statistics*. 2007;35(6):2769-94. doi:  
542 10.1214/009053607000000505

543 12. Chatterjee S. A New Coefficient of Correlation. *J Am Stat Assoc*. 2020;1–  
544 39. doi: 10.1080/01621459.2020.1758115

545 13. Kinney JB, Atwal GS. Equitability, mutual information, and the maximal  
546 information coefficient. *Proceedings of the National Academy of Sciences*.  
547 2014 Mar 4;111(9):3354-9. doi: 10.1073/pnas.1309933111

548 14. Yekutieli D. Hierarchical false discovery rate-controlling methodology.  
549 *Journal of the American Statistical Association*. 2008 Mar 1;103(481):309–  
550 16. doi: 10.1198/016214507000001373

551 15. Martin PG, Guillou H, Lasserre F, Déjean S, Lan A, Pascussi JM,  
552 SanCristobal M, Legrand P, Besse P, Pineau T. Novel aspects of PPAR $\alpha$ -  
553 mediated regulation of lipid and xenobiotic metabolism revealed through a  
554 nutrigenomic study. *Hepatology*. 2007 Mar;45(3):767-77. doi:  
555 10.1002/hep.21510

556 16. González I, Déjean S, Martin P, Baccini A. CCA: An R package to extend  
557 canonical correlation analysis. *Journal of Statistical Software*.  
558 2008;23(12):1-4. doi: 10.18637/jss.v023.i12

559 17. Selvaraju S, Raju P, Rao SB, Raghavendra S, Nandi S, Dineshkumar D,  
560 Thayakumar A, Parthipan S, Ravindra JP. Evaluation of maize grain and  
561 polyunsaturated fatty acid (PUFA) as energy sources for breeding rams  
562 based on hormonal, sperm functional parameters and fertility.



585 23. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, Reyes  
586 JA, Shah SA, LeLeiko N, Snapper SB, Bousvaros A. Dysfunction of the  
587 intestinal microbiome in inflammatory bowel disease and treatment.  
588 *Genome biology*. 2012 Sep;13(9):1-8. doi: 10.1186/gb-2012-13-9-r79

589 24. Morgan XC, Kabakchiev B, Waldron L, Tyler AD, Tickle TL, Milgrom R,  
590 Stempak JM, Gevers D, Xavier RJ, Silverberg MS, Huttenhower C.  
591 Associations between host gene expression, the mucosal microbiome,  
592 and clinical outcome in the pelvic pouch of patients with inflammatory  
593 bowel disease. *Genome biology*. 2015 Dec;16(1):1-5. doi:  
594 10.1186/s13059-015-0637-x

595 25. Zhan X, Plantinga A, Zhao N, Wu MC. A fast small-sample kernel  
596 independence test for microbiome community-level association analysis.  
597 *Biometrics*. 2017 Dec;73(4):1453-63. doi: 10.1111/biom.12684

598 26. Patel VI, Booth JL, Duggan ES, Cate S, White VL, Hutchings D, Kovats S,  
599 Burian DM, Dozmorov M, Metcalf JP. Transcriptional classification and  
600 functional characterization of human airway macrophage and dendritic cell  
601 subsets. *The Journal of Immunology*. 2017 Feb 1;198(3):1183-201. doi:  
602 10.4049/jimmunol.1600777

603 27. Furuhashi M, Hotamisligil GS. Fatty acid-binding proteins: role in  
604 metabolic diseases and potential as drug targets. *Nature reviews Drug*  
605 *discovery*. 2008 Jun;7(6):489-503. doi: 10.1038/nrd2589

606 28. Patterson E, Wall R, Lisai S, Ross RP, Dinan TG, Cryan JF, Fitzgerald  
607 GF, Banni S, Quigley EM, Shanahan F, Stanton C. *Bifidobacterium breve*  
608 with  $\alpha$ -linolenic acid alters the composition, distribution and transcription

609 factor activity associated with metabolism and absorption of fat. *Scientific*  
610 *reports*. 2017 Mar 7;7(1):1-2. doi: 10.1038/srep43300

611 29. Werner T, Wagner SJ, Martínez I, Walter J, Chang JS, Clavel T, Kisling S,  
612 Schuemann K, Haller D. Depletion of luminal iron alters the gut microbiota  
613 and prevents Crohn's disease-like ileitis. *Gut*. 2011 Mar 1;60(3):325-33.  
614 doi: 10.1136/gut.2010.216929

615 30. Donovan A, Lima CA, Pinkus JL, Pinkus GS, Zon LI, Robine S, Andrews  
616 NC. The iron exporter ferroportin/Slc40a1 is essential for iron  
617 homeostasis. *Cell metabolism*. 2005 Mar 1;1(3):191-200. doi:  
618 10.1016/j.cmet.2005.01.003

619 31. Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K,  
620 Shmulevich I, Sander C, Stuart JM. The cancer genome atlas pan-cancer  
621 analysis project. *Nature genetics*. 2013 Oct;45(10):1113-20. doi:  
622 10.1038/ng.2764

623 32. Vasaikar SV, Straub P, Wang J, Zhang B. LinkedOmics: analyzing multi-  
624 omics data within and across 32 cancer types. *Nucleic acids research*.  
625 2018 Jan 4;46(D1):D956-63. doi: 10.1093/nar/gkx1090

626 33. Andres SA, Brock GN, Wittliff JL. Interrogating differences in expression of  
627 targeted gene sets to predict breast cancer outcome. *BMC cancer*. 2013  
628 Dec;13(1):1-8. doi: 10.1186/1471-2407-13-326

629 34. Pogue-Geile KL, Kim C, Jeong JH, Tanaka N, Bandos H, Gavin PG,  
630 Fumagalli D, Goldstein LC, Sneige N, Burandt E, Taniyama Y. Predicting  
631 degree of benefit from adjuvant trastuzumab in NSABP trial B-31. *Journal*

632 of the National Cancer Institute. 2013 Dec 4;105(23):1782-8. doi:  
633 10.1093/jnci/djt321

634 35. Elian FA, Yan E, Walter MA. FOXC1, the new player in the cancer  
635 sandbox. *Oncotarget*. 2018 Jan 30;9(8):8165. doi:  
636 10.18632/oncotarget.22742

637 36. Caldon CE, Sergio CM, Kang J, Muthukaruppan A, Boersma MN, Stone  
638 A, Barraclough J, Lee CS, Black MA, Miller LD, Gee JM. Cyclin E2  
639 overexpression is associated with endocrine resistance but not  
640 insensitivity to CDK2 inhibition in human breast cancer cells. *Molecular*  
641 *cancer therapeutics*. 2012 Jul 1;11(7):1488-99. doi: 10.1158/1535-  
642 7163.MCT-11-0963

643 37. López-Knowles E, O'Toole SA, McNeil CM, Millar EK, Qiu MR, Crea P,  
644 Daly RJ, Musgrove EA, Sutherland RL. PI3K pathway activation in breast  
645 cancer is associated with the basal-like phenotype and cancer-specific  
646 mortality. *International journal of cancer*. 2010 Mar 1;126(5):1121-31. doi:  
647 10.1002/ijc.24831

648 38. Chin WW. The partial least squares approach to structural equation  
649 modeling. *Modern methods for business research*. 1998 Jan 1;295(2):295-  
650 336.

651 39. Sun L, Ji S, Yu S, Ye J. On the Equivalence between Canonical  
652 Correlation Analysis and Orthonormalized Partial Least Squares. In: *IJCAI*  
653 2009 Jul 11 (Vol. 9, pp. 1230-1235).

654 40. Hubert M, Branden KV. Robust methods for partial least squares  
655 regression. *Journal of Chemometrics: A Journal of the Chemometrics*  
656 Society. 2003 Oct;17(10):537-49. doi: 10.1002/cem.822

657 41. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with  
658 applications to sparse principal components and canonical correlation  
659 analysis. *Biostatistics*. 2009 Jul 1;10(3):515-34. doi:  
660 10.1093/biostatistics/kxp008

661 42. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical  
662 and powerful approach to multiple testing. *Journal of the Royal statistical  
663 society: series B (Methodological)*. 1995 Jan;57(1):289-300. doi:  
664 10.1111/j.2517-6161.1995.tb02031.x

665 43. Yekutieli D, Benjamini Y. Resampling-based false discovery rate  
666 controlling multiple test procedures for correlated test statistics. *Journal of  
667 Statistical Planning and Inference*. 1999 Dec 1;82(1-2):171-96. doi:  
668 10.1016/S0378-3758(99)00041-5

669 44. Winkler RL. The assessment of prior distributions in Bayesian analysis.  
670 *Journal of the American Statistical association*. 1967 Sep 1;62(319):776-  
671 800. doi: 10.1080/01621459.1967.10500894

672 45. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: a review  
673 of statistical methods and recommendations for their application. *The*  
674 *American Journal of Human Genetics*. 2010 Jan 8;86(1):6-22. doi:  
675 10.1016/j.ajhg.2009.11.017

676 46. Lewinger JP, Conti DV, Baurley JW, Triche TJ, Thomas DC. Hierarchical  
677 Bayes prioritization of marker associations from a genome-wide  
678 association scan for further investigation. *Genetic Epidemiology: The*  
679 *Official Publication of the International Genetic Epidemiology Society.*  
680 2007 Dec;31(8):871-82. doi: 10.1002/gepi.20248

681 47. Mourad R, Sinoquet C, Leray P. Learning hierarchical Bayesian networks  
682 for genome-wide association studies. InProceedings of COMPSTAT'2010  
683 2010 (pp. 549-556). Physica-Verlag HD. doi: 10.1007/978-3-7908-2604-  
684 3\_56

685 48. Mourad R, Sinoquet C, Leray P. A hierarchical Bayesian network  
686 approach for linkage disequilibrium modeling and data-dimensionality  
687 reduction prior to genome-wide association studies. *BMC bioinformatics.*  
688 2011 Dec;12(1):1-20. doi: 10.1186/1471-2105-12-16

689 49. Ben-Gal I. Bayesian networks. *Encyclopedia of statistics in quality and*  
690 *reliability.* 2008 Mar 15;1. doi: 10.1002/9780470061572.eqr089

691 50. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference  
692 under mixed models. *Bioinformatics.* 2003 Aug 12;19(12):1572-4. doi:  
693 10.1093/bioinformatics/btg180

694 51. Huson DH, Mitra S, Ruscheweyh HJ, Weber N, Schuster SC. Integrative  
695 analysis of environmental sequences using MEGAN4. *Genome research.*  
696 2011 Sep 1;21(9):1552-60. doi: 10.1101/gr.120618.111

697 52. Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. Bayesian inference  
698 of phylogeny and its impact on evolutionary biology. *science*. 2001 Dec  
699 14;294(5550):2310-4. doi: 10.1126/science.1065889

700 53. Anderson MJ. A new method for non-parametric multivariate analysis of  
701 variance. *Austral ecology*. 2001 Feb;26(1):32-46. doi: 10.1111/j.1442-  
702 9993.2001.01070.pp.x

703 54. McArdle BH, Anderson MJ. Fitting multivariate models to community data:  
704 a comment on distance-based redundancy analysis. *Ecology*. 2001 Jan  
705 1;82(1):290-7. doi: 10.1890/0012-  
706 9658(2001)082[0290:FMMTCD]2.0.CO;2

707 55. Goodall C. Procrustes methods in the statistical analysis of shape. *Journal*  
708 *of the Royal Statistical Society: Series B (Methodological)*. 1991  
709 Jan;53(2):285-321. doi: 10.1111/j.2517-6161.1991.tb01825.x

710 56. Fan J, Samworth R, Wu Y. Ultrahigh dimensional feature selection:  
711 beyond the linear model. *The Journal of Machine Learning Research*.  
712 2009 Dec 1;10:2013-38.

713 57. Waldron L, Pintilie M, Tsao MS, Shepherd FA, Huttenhower C, Jurisica I.  
714 Optimized application of penalized regression methods to diverse genomic  
715 data. *Bioinformatics*. 2011 Dec 15;27(24):3399-406. doi:  
716 10.1093/bioinformatics/btr591

717 58. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple  
718 testing under dependency. *Annals of statistics*. 2001 Aug 1:1165-88. doi:  
719 10.1214/aos/1013699998

720 59. Simon N, Tibshirani R. Comment on " Detecting Novel Associations In  
721 Large Data Sets" by Reshef Et Al, Science Dec 16, 2011. arXiv preprint  
722 arXiv:1401.7645. 2014 Jan 29.

723 **Supporting information**

724 **S1 Appendix. Supplementary methods and evaluation.**

725 **S1 Table. HALIA results on data from PPAR $\alpha$ -deficient mice.** Significant HALIA  
726 results with FDR threshold  $q = 0.05$  for fatty acid-transcript associations in  
727 PPAR $\alpha$ -deficient mice [15].

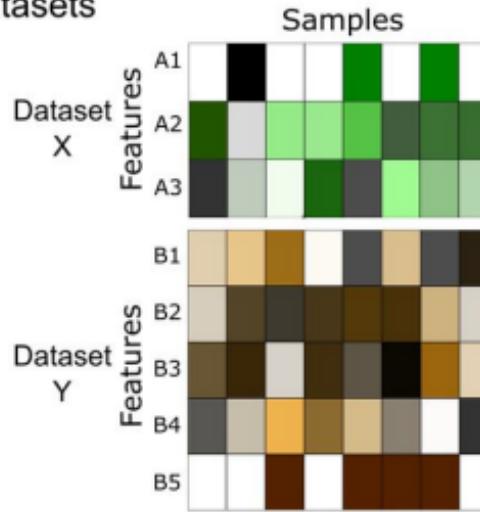
728 **S2 Table. HALIA results on microbe-gene relationships.** Significant HALIA  
729 results with FDR threshold  $q = 0.1$ , Spearman correlation as similarity metric, and  
730 medoid as the decomposition method for microbial and gene expression profiling  
731 data [23]. Reported associations encompassed 427 OTUs, 1,991 transcripts, and  
732 the equivalent of 8,382 pairwise associations.

733 **S3 Table. HALIA results microbe-metabolite relationships.** Significant HALIA  
734 results with FDR threshold  $q = 0.25$ , Spearman correlation as the similarity  
735 metric, and medoid as decomposition method, for the DiabImmune cohort data  
736 from [21]. These include 20 microbial genera and 284 metabolites of 103  
737 samples.

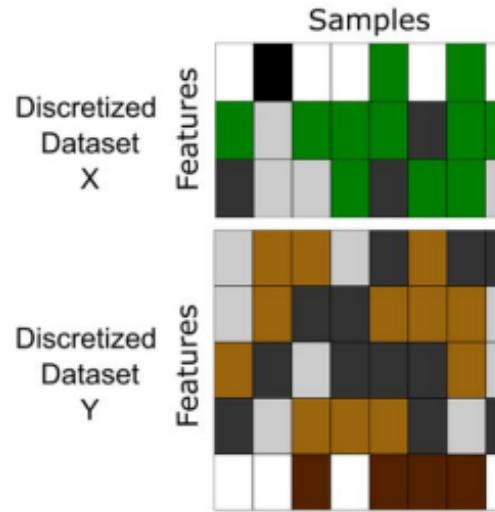
738 **Conflict of Interest**

739 The authors declare that they have no conflict of interest.

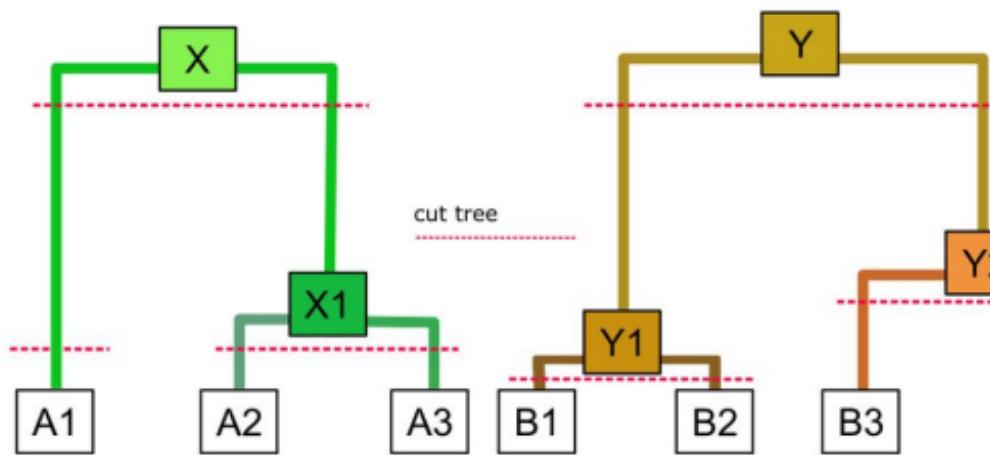
**A** INPUT: Paired high-dimensional datasets



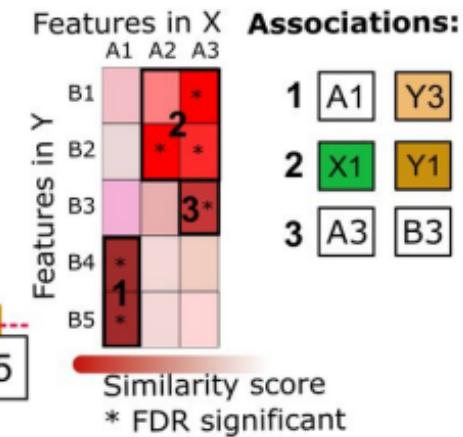
**B** Discretize continuous data



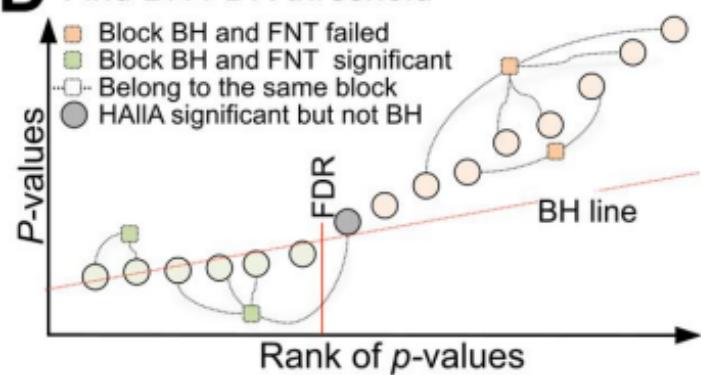
**C** Hierarchical clustering



**F** OUTPUT: Block associations in different levels of the hypothesis tree



**D** Find BH FDR threshold



For each hypothesis...

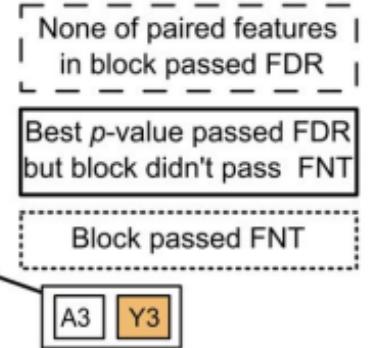
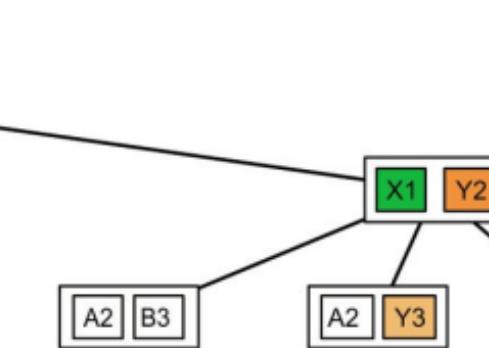
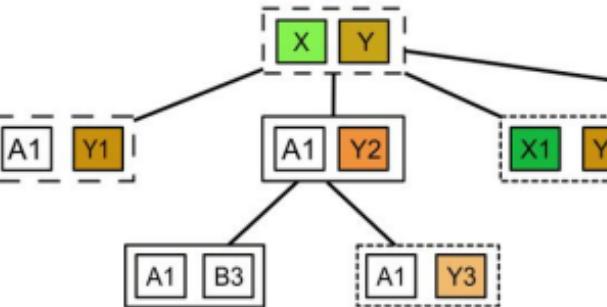


For each cluster, calculate a representative

Assess statistical significance

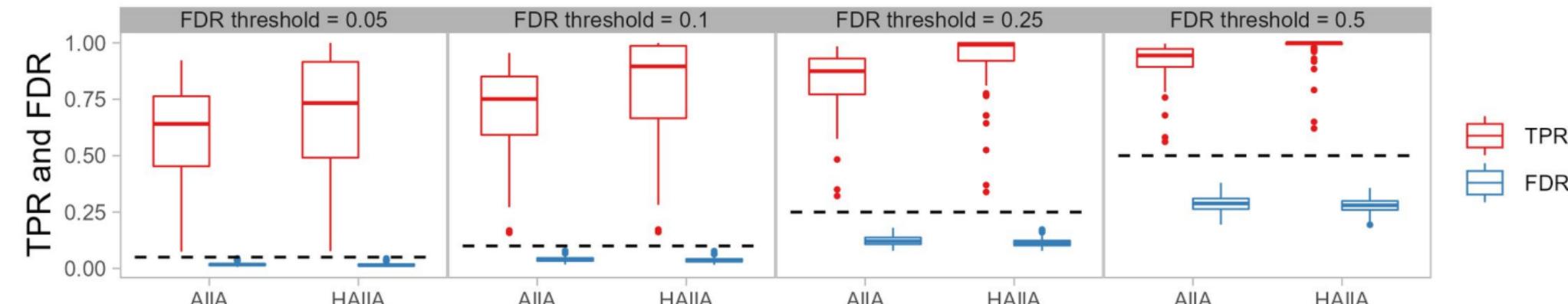
Permutation test FDR correction

**E** Block association discovery and descent based on Gini score



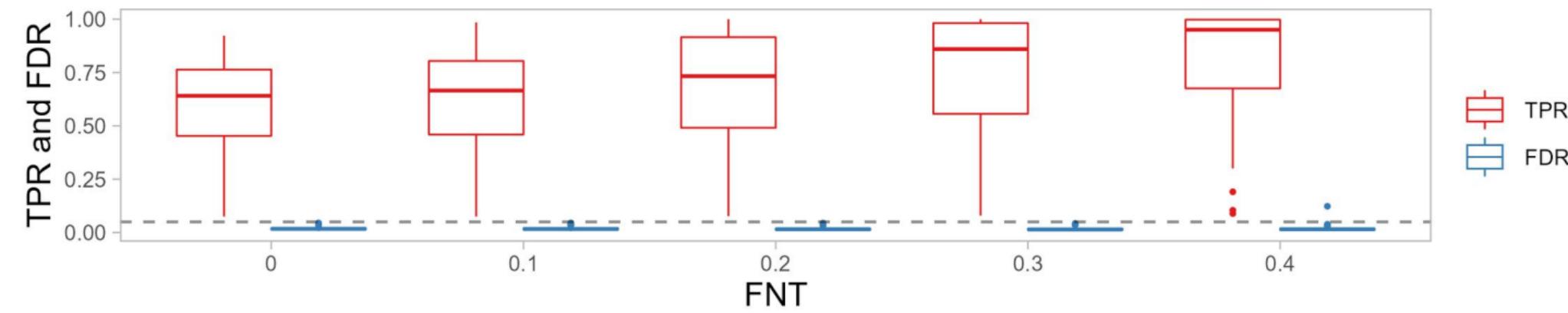
A

## Controlling FDR with different target FDR

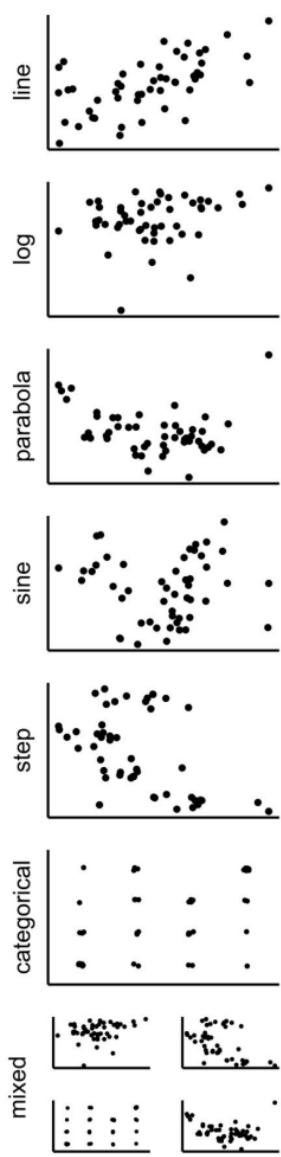


B

## Improving TPR with increased false negative tolerance



A



B

