

| | | |
|--|--|--------------------------|
| Manuscript Number: | GIGA-D-17-00337R2 | |
| Full Title: | A network-based conditional genetic association analysis of the human metabolome | |
| Article Type: | Technical Note | |
| Funding Information: | the European Union FP7 framework project Pain-Omics (602736) | Dr. Christian Gieger |
| | Ministry of Education and Science of the Russian Federation (the 5-100 Excellence Programme) | Mr. Sodbo Zh. Sharapov |
| | the Federal Agency of Scientific Organisations via the Institute of Cytology and Genetics (0324-2018-0017) | Dr. Yakov A. Tsepilov |
| | the Federal Agency of Scientific Organisations via the Institute of Cytology and Genetics (0324-2018-0017) | Prof. Yurii S. Aulchenko |
| Abstract: | <p>Background: Genome-wide association studies have identified hundreds of loci that influence a wide variety of complex human traits; however, little is known regarding the biological mechanism of action of these loci. The recent accumulation of functional genomics (“omics”), including metabolomics data, has created new opportunities for studying the functional role of specific changes in the genome. Functional genomic data are characterized by their high dimensionality, the presence of (strong) statistical dependency between traits, and—potentially—complex genetic control. Therefore, the analysis of such data requires specific statistical genetics methods.</p> <p>Results: To facilitate our understanding of the genetic control of omics phenotypes, we propose a trait-centered, network-based conditional genetic association (cGAS) approach for identifying the direct effects of genetic variants on omics-based traits. For each trait of interest, we selected from a biological network a set of other traits to be used as covariates in the cGAS. The network can be reconstructed either from biological pathway databases (a mechanistic approach) or directly from the data, using a Gaussian Graphical Model applied to the metabolome (a data-driven approach). We derived mathematical expressions which allow comparison of the power of univariate analyses with conditional genetic association analyses. We then tested our approach using data from a population-based KORA study (n=1784 subjects, 1.7 million SNPs) with measured data for 151 metabolites.</p> <p>Conclusions: We found that compared to single-trait analysis, performing a genetic association analysis that includes biologically relevant covariates can either gain or lose power, depending on specific pleiotropic scenarios, for which we provide empirical examples. In the context of analyzed metabolomics data, the mechanistic network approach had more power compared to the data-driven approach. Nevertheless, we believe that our analysis shows that neither a prior-knowledge-only approach nor a phenotypic-data-only approach is optimal, and we discuss possibilities for improvement.</p> | |
| Corresponding Author: | Yurii Aulchenko Institute of Cytology and Genetics SB RAS Novosibirsk, RUSSIAN FEDERATION | |
| Corresponding Author Secondary Information: | | |
| Corresponding Author's Institution: | Institute of Cytology and Genetics SB RAS | |
| Corresponding Author's Secondary Institution: | | |
| First Author: | Yakov A. Tsepilov, Ph.D. | |
| First Author Secondary Information: | | |

| | |
|--|--|
| Order of Authors: | <p>Yakov A. Tsepilov, Ph.D.</p> <p>Sodbo Zh. Sharapov</p> <p>Olga O. Zaytseva, Ph.D.</p> <p>Jan Krumsek, Ph.D.</p> <p>Cornelia Prehn, Ph.D.</p> <p>Jerzy Adamski, Ph.D.</p> <p>Gabi Kastenmüller, Ph.D.</p> <p>Rui Wang-Sattler, Ph.D.</p> <p>Konstantin Strauch, Ph.D.</p> <p>Christian Gieger, Ph.D.</p> <p>Yurii S. Aulchenko, Ph.D.</p> |
| Order of Authors Secondary Information: | |
| Response to Reviewers: | <p>Dear Hans,</p> <p>We would like to thank you and reviewers again for comments and suggestions. Please find attached our point-by-point answers below. We hope that you and the reviewers will find the revised manuscript suitable for publication.</p> <p>While addressing the comments we have noticed that we have used SNP rs17112944 when comparing the components of GGM- and BN-cGAS (LL 315-323). However, in other parts of the manuscript we consider this SNP to be false-positive and thus exclude it from other computations. We now excluded this SNP from computations, making our story more coherent. These changes did not affect our conclusions.</p> <p>Yours Sincerely. also on behalf of other authors, prof. Yurii Aulchenko and dr. Yakov Tsepilov</p> <p>Reviewer #1 report:</p> <p>1) For Supplementary Table 1 - the ARCHIVE tab should be removed from the file, the row freeze should include the 11th SNP (row 17) in Supplementary Table 1A, the heading should be edited for Supplementary Table 1C to emphasize that it is specifically for 1C. It is also unclear to me where the p-values come from in Table S1C - they are not exactly the same as the uGAS p-values, but are very similar. I also think Figure 2 should reference this table rather than Table S1B.</p> <p>Thank you for these suggestions. We made all necessary corrections in text and in Supplementary Tables. In tables ST1A-B and ST2 the freeze now includes only two first columns and the header. The reason of minor differences in p-values was the application of genomic control in S1A,B and S2 (see also comment #5).</p> <p>2) Thank you for clarifying that the Wald test was used, assuming a chi-squared distribution. The reason I had mentioned the t-test is that it is generally more common to use the t-test for individual parameters in either simple or multivariable regression, as chi-squared tests rely on asymptotic properties (of the number of samples going to infinity) and are more commonly used in generalized linear regression models. The choice to use this test may thus be worth a short discussion.</p> <p>We have used the Wald test statistic because this allowed us to analytically express the log-ratio between conditional and univariate tests (equations 1 and 2). We believe this is valid approximation because the typical number of samples in genetic association studies is thousands or orders of magnitude more, which we now mention on lines LL 168-170.</p> <p>3) For Tables 1 and 2, it would be great if the noise/pleiotropic components were in different columns (easier to look at and potentially use in downstream analyses)</p> |

| | |
|--|---|
| | <p>Thank you for this comment. Done.</p> <p>4) Thank you for expanding the description of the model and changing the notation in the Results section. I think a few small edits should be made here: - on line 161, rho_cg is the correlation between the covariate and the genotype, not the covariate and the trait - on lines 165-167, should include "estimated" before "residual variance" and "partial correlation"</p> <p>Corrected.</p> <p>5.1) Thank you for making everything available through Code Ocean! The figures reproduce very well, however it would be helpful if all the main and supplementary tables were also available in the same format as they are in the manuscript.</p> <p>Thank you for these suggestions. We now have implemented the full pipeline starting from clumping to the final supplementary tables. It should be noted, that some steps (like comparison with Tsepilov et.al., 2015 and Draisma et.al., 2016) were done manually and are not implemented in CodeOcean.</p> <p>5.2) I also wasn't sure where the uGAS output was to be found.</p> <p>The uGAS output could be found here: CodeOcean workspace/Data/uGWAS Thanks to your comment, we now updated the README file to make the navigation more clear. Note that uGAS is denoted as uGWAS in CodeOcean.</p> <p>5.3) Additionally, I found some small discrepancies with the cGAS p-values, for example in Table 2, the p-value for the association of rs2286963 with C9 is 7.41E-73 whereas in the Code Ocean result it is 1.53E-73.</p> <p>The reason for these discrepancies in p-values was the application of genomic control in S1A, S1B, S2. To all supplementary tables we have added new columns that list p-values before and after genomic control.</p> |
| Additional Information: | |
| Question | Response |
| Are you submitting this manuscript to a special series or article collection? | No |
| <p>Experimental design and statistics</p> <p>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends.</p> <p>Have you included all the information requested in your manuscript?</p> | Yes |
| Resources | Yes |

| | |
|---|------------|
| <p>A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?</p> | |
| <p>Availability of data and materials</p> <p>All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.</p> <p>Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?</p> | <p>Yes</p> |

A network-based conditional genetic association analysis of the human metabolome

Y.A. Tsepilov^{1,2}, S.Z. Sharapov², O.O. Zaytseva^{1,2}, J. Krumsek³, C. Prehn⁴, J. Adamski^{4,5,6}, G.
Kastenmüller⁷, R. Wang-Sattler^{6,8,9}, K. Strauch^{10,11}, C. Gieger^{6,8,9}, Y.S. Aulchenko^{1,2,12*}

1 Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

2 Novosibirsk State University, Novosibirsk, Russia

3 Institute of Computational Biology, Helmholtz Center Munich - German Research Center
for Environmental Health, Neuherberg, Germany

4 Institute of Experimental Genetics, Genome Analysis Center, Helmholtz Center Munich -
German Research Center for Environmental Health, Neuherberg, Germany

5 Institute of Experimental Genetics, Life and Food Science Center Weihenstephan, Technical
University of Munich, Freising-Weihenstephan, Germany

6 German Center for Diabetes Research, Neuherberg, Germany

7 Institute of Bioinformatics and Systems Biology, Helmholtz Center Munich - German
Research Center for Environmental Health, Neuherberg, Germany

8 Research Unit of Molecular Epidemiology, Helmholtz Center Munich - German Research
Center for Environmental Health, Neuherberg, Germany

9 Institute of Epidemiology II, Helmholtz Center Munich - German Research Center for
Environmental Health, Neuherberg, Germany

10 Institute of Genetic Epidemiology, Helmholtz Center Munich - German Research Center
for Environmental Health, Neuherberg, Germany

11 Institute of Medical Informatics, Biometry and Epidemiology, Chair of Genetic
Epidemiology, Ludwig Maximilian University of Munich, Munich, Germany

12 PolyOmica, 's-Hertogenbosch, The Netherlands

* Correspondence to

Yurii S. Aulchenko

Institute of Cytology and Genetics SB RAS, 630090 Novosibirsk, Russia

yurii@bionet.nsc.ru

*Keywords: genome-wide association study; multivariate model; metabolomics; conditional
analysis; pleiotropy*

38 **Abstract**

1 39 **Background:** Genome-wide association studies have identified hundreds of loci that influence a
2 wide variety of complex human traits; however, little is known regarding the biological mechanism
3 of action of these loci. The recent accumulation of functional genomics (“omics”), including
4 metabolomics data, has created new opportunities for studying the functional role of specific
5 changes in the genome. Functional genomic data are characterized by their high dimensionality,
6 the presence of (strong) statistical dependency between traits, and—potentially—complex genetic
7 control. Therefore, the analysis of such data requires specific statistical genetics methods.
8

9 46 **Results:** To facilitate our understanding of the genetic control of omics phenotypes, we propose a
10 trait-centered, network-based conditional genetic association (cGAS) approach for identifying the
11 direct effects of genetic variants on omics-based traits. For each trait of interest, we selected from
12 a biological network a set of other traits to be used as covariates in the cGAS. The network can be
13 reconstructed either from biological pathway databases (a mechanistic approach) or directly from
14 the data, using a Gaussian Graphical Model applied to the metabolome (a data-driven approach).
15 We derived mathematical expressions which allow comparison of the power of univariate analyses
16 with conditional genetic association analyses. We then tested our approach using data from a
17 population-based KORA study (n=1784 subjects, 1.7 million SNPs) with measured data for 151
18 metabolites.
19

20 56 **Conclusions:** We found that compared to single-trait analysis, performing a genetic association
21 analysis that includes biologically relevant covariates can either gain or lose power, depending on
22 specific pleiotropic scenarios, for which we provide empirical examples. In the context of analyzed
23 metabolomics data, the mechanistic network approach had more power compared to the data-
24 driven approach. Nevertheless, we believe that our analysis shows that neither a prior-knowledge-
25 only approach nor a phenotypic-data-only approach is optimal, and we discuss possibilities for
26 improvement.
27

28 63

64 **Short abstract**

1 65 We propose a trait-centric network-based conditional approach for performing a genetic
2
3 66 association analysis of multivariate omics phenotypes. This approach can incorporate existing
4
5 67 biological knowledge regarding biological pathways obtained from external sources and is
6
7 68 designed to specifically test for direct genetic effects. We applied this approach to existing
8
9 69 metabolomics data and found that it may have more power by having increased accuracy of genetic
10
11 70 effect estimates in the presence of specific “counterintuitive” pleiotropic scenarios in which locus-
12
13 71 specific genetically induced and residual covariance are opposite, but it may lose power when
14
15 72 genetically induced and residual covariance have a concordant sign. We provide empirical
16
17 73 examples of different pleiotropic scenarios that we observed in metabolomics, and we discuss
18
19 74 possible additional applications for this approach.
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

75 **Background**

1 76 Genome-wide association studies (GWAS) are a highly popular method for identifying alleles that
2
3 77 affect complex traits in humans, including the risk of common diseases. In the past decade,
4
5 78 GWASs have enabled the identification of thousands of loci, significantly increasing our
6
7 79 understanding of the genetic basis underlying the control of complex human traits [1]. On the other
8
9 80 hand, this has had only a limited impact on the development of biomarkers and therapeutic agents;
10
11 81 in most cases, any association found using GWAS approach can only serve as a starting point for
12
13 82 future research, rather than providing a direct answer to the question of the genetic region's precise
14
15 83 biological function. The recent accumulation of functional genomics (or "omics" for short) data—
16
17 84 including information regarding the levels of gene expression (the transcriptome), metabolites (the
18
19 85 metabolome), proteins (the proteome), and glycosylation (the glycome)—can provide new insight
20
21 86 into the functional role of specific changes in the genome [2,3].

22 87 Metabolomics is an emerging field that has been studied extensively in the past decade. A
23
24 88 number of GWASs of metabolites have been performed using various platforms [4–8], revealing
25
26 89 literally dozens of loci associated with variations in various lipid species, amino acids, and other
27
28 90 small molecules. Linking the variants that underlie these variations in metabolomics with various
29
30 91 diseases can provide functional insight into the many disease-related associations that were
31
32 92 reported in previous studies, including cardiovascular and kidney disease, type 2 diabetes, cancer,
33
34 93 gout, venous thromboembolism, and Crohn's disease [5].

35 94 However, analyzing metabolomics data requires specialized statistical methods due to their
36
37 95 characteristically high dimensionality and the presence of statistical dependencies that reflect
38
39 96 biological relationships between different variables. Conventional univariate GWAS (uGAS)
40
41 97 approaches ignore any possible dependencies between different omics traits, which can confound
42
43 98 the biological interpretation of the results and may lead to a loss of statistical power. On the other
44
45 99 hand, utilizing multivariate phenotype information increases the statistical power of the association
46
47 100 tests compared to univariate analysis [9–12]. Despite a large number of methodological studies,
48
49 101 however, only a few empirical multivariate GWASs have been published using data for humans.
50
51 102 We recently demonstrated [13] that using a multivariate analysis can substantially increase the
52
53 103 power of locus identification in the context of human *N*-glycomics; indeed, not only did our
54
55 104 multivariate analysis double the number of loci identified in the analysis sample, but also all five
56
57 105 novel loci were strongly replicated. With respect to metabolomics, Inouye et al. [6] performed a
58
59 106 multivariate GWAS on 130 metabolites (grouped in 11 sets) measured in approximately 6600
60
61 107 individuals. They found that multivariate analysis doubled the number of loci detected in this
62
63 108 sample; seven of these additional loci discovered were novel loci that had not been identified
64
65

109 previously in other GWAS analyses of related traits. While no replication of novel loci was
110 performed by Inouye et al., we compared the authors' results with a recently published univariate
111 GWAS of metabolomics derived from a cohort containing nearly 25,000 individuals [8]. We found
112 that three of the seven SNPs reported by Inouye et al. have a p -value $< 5 \times 10^{-11}$ for at least one
113 metabolite (i.e., are significant at the genome-wide level after Bonferroni correction for 130
114 analyses). These findings provide empirical evidence supporting the value of using multivariate
115 methods to analyze the genomics of metabolic traits, at least in the context of locus discovery.

116 It should be noted that these multivariate methods and tests were developed by statistical
117 geneticists to specifically increase the power of gene identification. In such "gene-centric" tests,
118 the model that includes the effects of genotype on multiple traits is contrasted with the null model
119 in which the gene has no effect on any trait analyzed. Although useful and powerful for genetic
120 mapping, this approach may have limited interpretability in a context in which one is interested in
121 the genetic control and biology of specific trait or a subset of traits (the "trait-centered" view).
122 Several statistical methods have been suggested to address the question of which specific traits are
123 affected in an analyzed ensemble (see for example [10,14]). One such method is based on
124 conditional analysis [15], in which a "target trait" is analyzed as a genotype-dependent variable
125 and related traits are included in the regression model as covariates. Such a modeling approach
126 allows—at least in theory—one to rule out indirect genetic effects (e.g., effects that are in fact
127 solely mediated through some other trait) and study only the genetic effects that directly affect the
128 trait of interest.

129 Here, we present a statistical model in which a given trait depends on a genetic
130 polymorphism and in which a number of related traits are included in the model as covariates. In
131 this model, the relationship between the genotype and the trait of interest is our primary focus.
132 Analyzing such a model allows us to identify the direct effect of genetics on the trait of interest.
133 Mathematically, the model is equivalent to the model used by Deng and Pan [15]. We first compare
134 this conditional genetic association (cGAS) approach with the standard model in which a trait of
135 interest depends solely on genotype, without other traits used as covariates (i.e., the univariate
136 genetic association—or uGAS—model). We do so by mathematically deriving expressions that
137 allow us to examine the relative power of the uGAS and cGAS approaches, and we identify the
138 situations in which these models are expected to yield different results.

139 As might be expected—and as demonstrated here—the choice of covariates plays a critical
140 role in conditional analyses. First, we used the assumption that the covariates (i.e., biologically
141 relevant traits) are known. Second, we explored the problem of selecting appropriate covariates,
142 and we tested the approaches by performing a proof-of-principle study using metabolomics data
143 consisting of 151 metabolites (Biocrates assay) obtained from the KORA F4 study (n=1785

144 individuals). Specifically, we selected covariates based on existing knowledge from metabolite
145 biochemical networks (BN-cGAS) and using a data-driven approach based on Gaussian Graphical
146 Modeling (GGM-cGAS). Finally, we compare and discuss the obtained results, and we discuss
147 possible applications for this analysis based on biologically and/or statistically relevant traits.
148
149

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

150 Results

151 The power of performing a conditional analysis of genetic associations

152 We start with the theoretical substantiation and identification of specific scenarios in which
153 adjusting for biologically relevant covariates can modify the power of an association analysis.

154 Let us consider a trait of interest, y , covariate c , and genotype g . We can formulate this
155 problem in terms of a linear regression as follows: $y = \mu + \beta_g * g + \beta_c * c + e$, where β_g and β_c
156 are the effects of the genotype and covariate, respectively, and e is the residual noise. Without a
157 loss of generality, we assume that all random variables in this equation are distributed with a mean
158 of zero and a standard deviation of 1, making (partial) regression coefficients equal to (partial)
159 correlation coefficients. Given these assumptions made, the joint distribution of y , g , and c can be
160 specified using a set of three correlation coefficients, ρ_{yg} (the correlation between the trait and the
161 genotype), ρ_{cg} (the correlation between the covariate and the **genotype**), and ρ_{yc} (the correlation
162 between the trait and the covariate). To test the association between y and g , we use the Wald test,
163 which is defined as the square of the ratio between the effect estimate and its standard error, with
164 the latter estimated under the alternative hypothesis (see [16]). The value of the “univariate” Wald
165 test statistic is calculated as $T_u^2 = \frac{n \hat{\rho}_{yg}^2}{\hat{\sigma}_u^2}$, where n is the sample size and $\hat{\sigma}_u^2 = 1 - \hat{\rho}_{yg}^2$ is the
166 **estimated** residual variance of y . For the conditional test, the Wald test is $T_c^2 = \frac{n \hat{\beta}_g^2}{\hat{\sigma}_c^2}$, where $\hat{\beta}_g$ is
167 the **estimated** partial correlation between the trait y and the genotype g (estimated from the
168 conditional model) and $\hat{\sigma}_c^2$ is the estimated residual variance of y . Note that under the null
169 hypothesis, **when n is large**, both T_u^2 and T_c^2 **are well approximated by** chi-square distribution with
170 one degree of freedom. **For genetic association studies, n is thousands or orders or magnitude**
171 **more.**

172 For the conditional model, $\hat{\beta}_g = \hat{\rho}_{yg} - \hat{\beta}_c \hat{\rho}_{cg}$; thus, we can rewrite $T_c^2 = n(\hat{\rho}_{yg} -$
173 $\hat{\beta}_c \hat{\rho}_{cg})^2 / \hat{\sigma}_c^2$. Consequently, the log-ratio of the conditional and univariate test statistics can be
174 partitioned into two components:

$$175 \log\left(\frac{T_c^2}{T_u^2}\right) = \log\left(\frac{\hat{\sigma}_u^2}{\hat{\sigma}_c^2}\right) + \log\left(\left[1 - \frac{\hat{\beta}_c \hat{\rho}_{cg}}{\hat{\rho}_{yg}}\right]^2\right) \quad (1)$$

176 Because the first term in Eq. (1) is dependent only upon residual variances of the two
177 models, we call this term the “noise” component. The second term depends upon the correlations
178 between traits and between the traits and the genotype; we call this term the “pleiotropic”
179 component. Because the noise component ($\hat{\sigma}_u^2 / \hat{\sigma}_c^2$) is always ≥ 1 , any possible decrease in the ratio
180 between univariate and conditional tests is determined by the sign and the magnitude of the term

181 $\hat{\beta}_c \hat{\rho}_{cg}/\hat{\rho}_{yg}$. If this term is negative, there will always be an increase in the power of the conditional
 182 analysis.

183 We can re-write $\hat{\beta}_c \hat{\rho}_{cg}/\hat{\rho}_{yg}$ as $\hat{\beta}_c \hat{\rho}_{yc}^*$, where $\hat{\rho}_{yc}^* = \hat{\rho}_{cg}/\hat{\rho}_{yg}$ is the component of the
 184 correlation between trait y and covariate c , which is induced by the variation in the genotype g .
 185 This quantity takes a central place in a Mendelian randomization analysis, which uses a genetic
 186 variation to anchor the causality arrow and consequently infers a causal relation between various
 187 traits (see for example [17]). Note that whereas $\hat{\rho}_{yc}^*$ reflects the covariance between the trait and
 188 the covariate induced by the effect of the genotype, $\hat{\beta}_c$ is conditional on the genotype and is related
 189 to the residual sources of covariance between y and c .

190 In general, the genetically induced covariance and the residual covariance are expected to
 191 have a concordant sign (see Discussion for details and relevant references). Thus, we conclude
 192 somewhat surprisingly that when genotype-induced and environmental correlations are similar in
 193 sign (i.e., both are positive or both are negative), the product $\hat{\beta}_c \hat{\rho}_{yc}^*$ is positive and the contribution
 194 of the second term in Eq. (1) to the relative power is negative. Note that the contribution of the
 195 first term in Eq. (1) is always positive; therefore, even if $\hat{\beta}_c \hat{\rho}_{yc}^*$ is positive, the power of a
 196 conditional analysis may still be higher than the power of a univariate analysis. In contrast, an
 197 “unexpected” product (in which the signs are different and hence $\hat{\beta}_c \hat{\rho}_{yc}^*$ is negative) contributes
 198 positively to the relative power of the conditional model. Note that in such a situation, the power
 199 of a conditional analysis will always be higher than the power of a univariate analysis.

200 We can readily extend Eq. (1) to a situation in which k covariates are included in the
 201 conditional model. Denoting the estimated coefficients of correlation between g and covariate i as
 202 $\hat{\rho}_{gi}$ and the estimated partial correlation between y and covariate i as $\hat{\beta}_i$ yields the following
 203 equation:

$$204 \quad \log\left(\frac{T_c^2}{T_u^2}\right) = \log\left(\frac{\hat{\sigma}_u^2}{\hat{\sigma}_c^2}\right) + \log\left(\left[1 - \frac{1}{\hat{\rho}_{yg}} \sum_{i=1}^k \hat{\beta}_i \hat{\rho}_{gi}\right]^2\right) \quad (2)$$

205 When appropriate covariates are selected, performing cGAS using individual-level data
 206 becomes rather trivial and can be achieved using standard statistical and software tools in which
 207 one estimates the effects of a SNP and covariates. However, cGAS becomes somewhat less trivial
 208 if one chooses to use summary-level univariate GWAS data such as data available from previously
 209 published studies. The formalization of cGAS in terms of summary univariate GWAS statistics is
 210 described in **Supplementary Note 1**. Here, we used methods based on analyzing summary-level
 211 data.

212

213 Network-based selection of covariates

1 214 The ability to select appropriate covariates is extremely important, as it can have direct
2 215 implications regarding the outcome of the analysis. If the biological/biochemical relationships
3 216 between traits of interest are known and are summarized in a database(s), this knowledge can be
4 217 used directly, for example by using all direct neighbors as covariates. We refer to this approach as
5 218 a biochemical-network driven cGAS (BN-cGAS). Alternatively, the network can be reconstructed
6 219 in a hypothesis-free, empirical manner from the data, for example using a Gaussian Graphical
7 220 Model (GGM) [18]. We refer to this approach as a GGM-cGAS.

8 221 We compared cGAS and uGAS by performing a genome-wide analysis of genetic effects
9 222 using summary-level data obtained from the KORA F4 study. This study included 151 metabolites
10 223 measured in 1784 individuals using the Biocrates assay and imputed at 1,717,498 SNPs.

11 224 First, we examined the potential of using cGAS when the covariates are selected based on
12 225 a known biochemical network (i.e., BN-cGAS). Thus, our analysis was restricted to a subset of
13 226 105 metabolites for which at least the one-reaction-step immediate biochemical neighbors are
14 227 known [18]. This biochemical network incorporates only lipid metabolites, and the pathway
15 228 reactions cover two groups of pathways: (1) fatty acid biosynthesis reactions, which apply to the
16 229 metabolite classes lyso-PC, diacyl-PC, acyl-alkyl-PC, and sphingomyelins; and (2) β -oxidation
17 230 reactions that reflect fatty acid degradation and apply to acylcarnitines. The β -oxidation model
18 231 consists of a linear chain of C2 degradation steps (C10 to C8 to C6, etc.). The number of covariates
19 232 ranged from 1 to 4, with mean and median values of 2.48 covariates and 2 covariates, respectively.

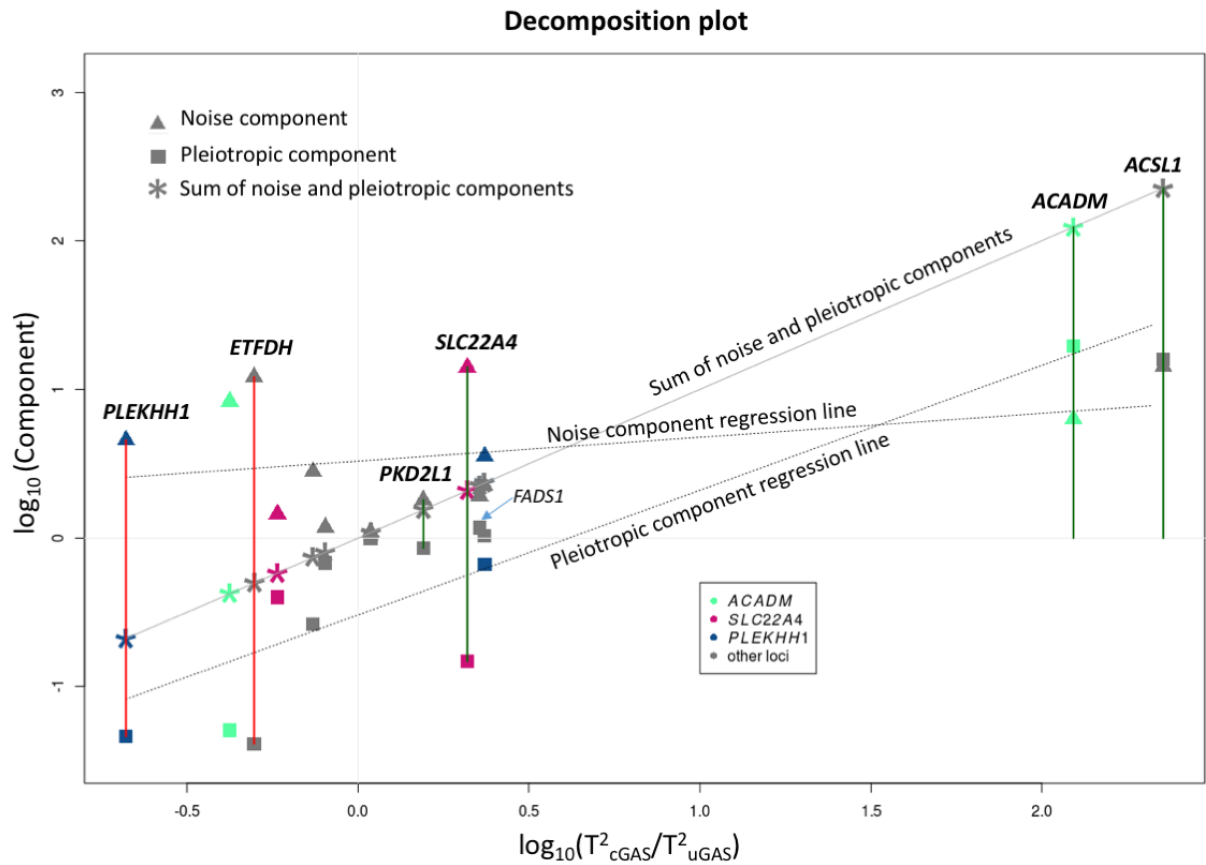
20 233 **Table 1** lists the 11 loci that were significant in either BN-cGAS or uGAS and fell into
21 234 known associated regions (see **Supplementary Note 2**). Of these 11 loci, ten and nine loci could
22 235 be identified by BN-cGAS and uGAS, respectively. Compared to uGAS, BN-cGAS identified one
23 236 fewer locus (*ETFDH*), but identified two more (*ACSL1* for PC ae C42:5 and *PKD2L1* for lyso-
24 237 PC a C16:1). It is interesting to note that for *ACSL1*, the effect of SNP rs4862429 on PC ae C42:5
25 238 was highly significant ($p=7e-11$) with BN-cGAS, but was not significant ($p=0.7$) with uGAS; this
26 239 outcome is to be expected under the model of unexpected pleiotropy.

27 240 Next, to test whether using BN-cGAS increases the average power of the association
28 241 analysis, we compared the BN-cGAS and uGAS chi-square test results for the loci listed in **Table**
29 242 **1**. Within a given locus, we compared the maximum test value. The average ratio of the maximum
30 243 test statistic between BN-cGAS and uGAS was 1.47, indicating that on average, BN-cGAS led to
31 244 higher test statistic values. However, when we used a paired-sample Wilcoxon test to compare the
32 245 best chi-square test results between BN-cGAS and uGAS, the difference between the two methods
33 246 was not significant ($p=0.123$) (see Supplementary Table S1A).

247 For the SNPs listed in **Table 1**, we then used Eq. (2) to partition the log-ratio of the BN-
248 cGAS and uGAS statistics values into “noise” and “pleiotropic” components. As shown in **Figure**
249 **1**, the regression slope of the second (i.e., “pleiotropic”) component is considerably higher than
250 the slope of the noise component; in other words, the ratio is determined primarily by the
251 pleiotropic term in Eq. (2). Moreover, with the exception of the *SLC22A4* locus, the SNP-trait pairs
252 for which BN-cGAS had increased power are the pairs in which the second term in Eq. (2) is either
253 positive or close to zero. In contrast, in the SNP-trait pairs that were not identified using BN-
254 cGAS, the “pleiotropic” term in Eq. (2) had a strong negative contribution.

255 Next, we investigated the variance-covariance structure of the loci with positive and
256 negative pleiotropic terms. We therefore selected a locus in which the pleiotropic component’s
257 contribution to power was positive (rs174547 at *FADS1*) and a locus in which the pleiotropic
258 component’s contribution to power was negative (rs8396 at *ETFDH*). **Figure 2** shows the
259 corresponding correlations between the SNP, the trait, and the covariates involved, together with
260 the partial coefficients for the conditional regression of the trait on the SNP and the covariates.
261 With respect to *FADS1* (**Figure 2A**), the correlations between the SNP and the trait (lyso-
262 PC a C20:4) and between the SNP and the covariate (lyso-PC a C20:3) are in opposite directions,
263 generating negative genetically induced covariance between lyso-PC a C20:4 and lyso-
264 PC a C20:3. In contrast, the residual correlation between the trait and the covariate is positive.
265 Therefore, the value of the partial regression coefficient between the SNP and lyso-PC a C20:4,
266 conditional on lyso-PC a C20:3, is greater than that of the coefficient of regression without
267 covariates.

268 With respect to the second example, *ETFDH* (**Figure 2B**), we found that the conditional
269 regression of C10 on rs8396 and two covariates (C8 and C12, two medium-chain acylcarnitines)
270 led to a smaller SNP partial regression coefficient compared to an unconditional regression; this
271 is because all of the terms in $\sum_{i=1}^k \hat{\beta}_i \hat{\rho}_{gi} / \hat{\rho}_{yg}$ are positive.



273
 274 **Figure 1. Decomposition of the \log - T^2 ratio for cGAS and uGAS into pleiotropic and noise**
 275 **components.** Vertically grouped trios (each composed of a square, triangle, and asterisk)
 276 correspond to one of fourteen associations in Table 1. The position of a trio on the x -axis
 277 corresponds to the log-ratio between conditional and univariate test statistic. On the y -axis,
 278 the asterisk corresponds to the log-ratio of cGAS and uGAS T^2 statistics. The value of the pleiotropic
 279 component is depicted by a square, and the value of the noise component is depicted by a triangle.
 280 Each trio is shown in gray, except the trios representing the *ACADM*, *SLC22A4*, and *PLEKHH1*
 281 loci, for which we have two different associations. The three dotted lines correspond to the
 282 regression lines for the two components and their sum. The four dark-green vertical lines indicate
 283 the associations that were significant in the cGAS analysis but not in the uGAS analysis, and the
 284 two dark-red lines indicates the associations that were significant only in the uGAS analysis.

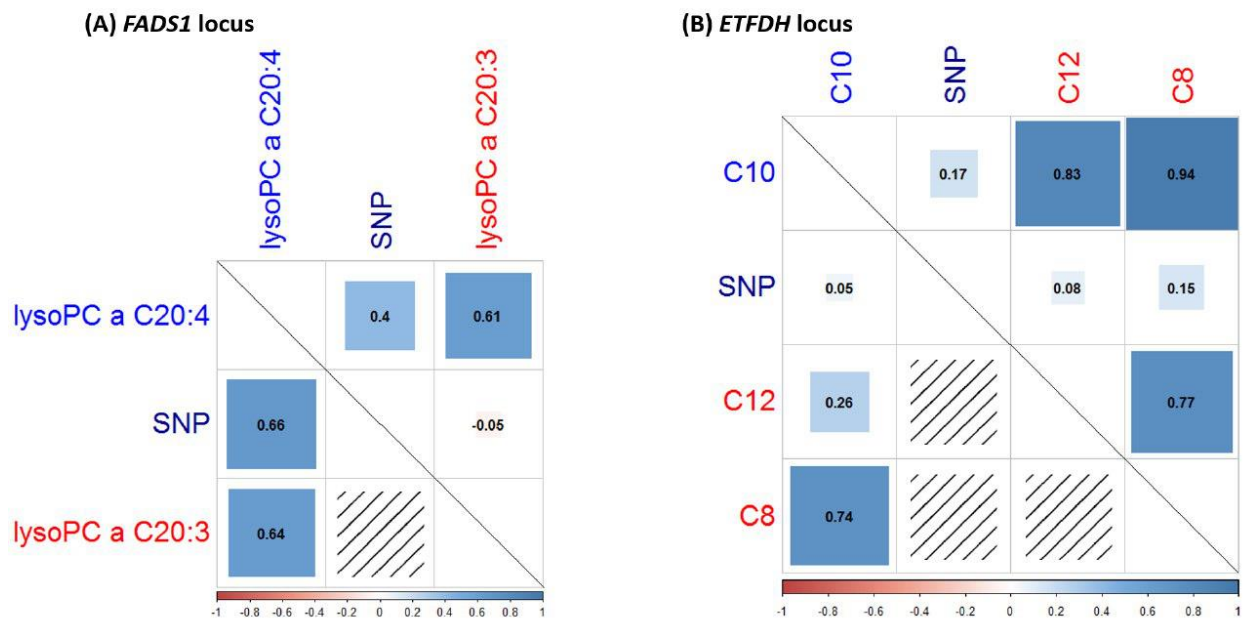


Figure 2. Matrix of correlations (above diagonal line) and the partial regression coefficients of the trait of interest on the SNP genotype and covariate(s) (the first column) for the *FADS1* (A) and *ETFDH* (B) loci. The result of the univariate analysis of regression of the corresponding traits onto SNPs is presented in Supplementary Table S1C. Names of traits used as covariates are in red. The number in a cell indicates the value of correlation (partial regression coefficient). The area of a square is proportional to the absolute value of correlation (partial regression coefficient); the effect magnitude is also reflected by square's color (the scale provided at the bottom of the graph). The *FADS1* locus represents scenario in which the pleiotropic term in Eq. (2) is strongly positive, while for *ETFDH* this term is negative.

Although using a known biochemical network to select covariates has many advantages, it may be somewhat unpractical and perhaps even harmful, as our biochemical knowledge is still relatively incomplete. Therefore, we explored the potential of performing a cGAS in which the covariates are selected using a data-driven approach (GGM-cGAS). The network of metabolites was reconstructed using Gaussian Graphical Models based on partial correlations. For a given metabolite, we selected covariates based on significant partial correlations. Specifically, we used the following threshold as proposed previously [18]: a p -value \leq (0.01/number of calculated partial correlations), which corresponds to a cut-off at $p \leq 8.83 \times 10^{-7}$. The network used in our analysis is shown in **Supplementary Figure S1**.

To compare GGM-cGAS with BN-cGAS, we used the same set of metabolites that we used for BN-cGAS to run our GGM-cGAS analysis; these results are presented in **Supplementary Table S1B**. We found 15 SNP-trait pairs clustered at 10 known loci (see **Supplementary Note 2**) that were detected by either GGM-cGAS or BN-cGAS. More covariates were included in the GGM-cGAS analysis (ranging from 1 to 18, with mean and median values of 7.6 covariates and 7 covariates, respectively) than in the BN-cGAS analysis. Thus, we predicted that GGM-cGAS would have relatively more power than BN-cGAS due to reduced noise (term 1 in Eq. (2)); on the

313 other hand, GGM-cGAS might lose power because of reduced occurrence of unexpected
314 pleiotropy (term 2 in Eq. (2)).

315 For the best SNP-trait pairs detected by GGM-cGAS or BN-cGAS, we computed the
316 components in Eq. (2) and compared these components using a paired-sample Wilcoxon test. We
317 found that the noise component in Eq. (2) was always larger for GGM-cGAS, with a mean
318 difference of 0.29 ($p=6 \times 10^{-5}$). Moreover, the second “pleiotropic” component in Eq. (2) was
319 generally smaller for GGM-cGAS than for BN-cGAS, with a mean difference of -0.47 ($p=0.015$);
320 nevertheless, for three out of 15 GGM-cGAS SNP-trait pairs, the pleiotropic component was
321 positive. The average chi-square value was 25% smaller for GGM-cGAS than for BN-cGAS,
322 indicating an average loss of power (although this loss was not significant; $p=0.5$ based on a paired
323 Wilcoxon test).

324 Next, we investigated further the potential of using cGAS under realistic conditions to a
325 full extent by analyzing all 151 available metabolites using GGM-cGAS and comparing these
326 results with the results of uGAS (Table 2 and Supplementary Figure S2). In total, uGAS detected
327 15 loci at the genome-wide significance level $p \leq 5 \times 10^{-8}/151$ (i.e., $p < 3.3 \times 10^{-10}$). On the other
328 hand, GGM-cGAS identified 19 significant loci using the same threshold. As expected, the
329 standard errors of the genetic effect estimates were smaller for GGM-cGAS than for uGAS (Table
330 2 and Supplementary Figure S3). A total of 14 loci were detected by both uGAS and GGM-
331 cGAS. GGM-cGAS failed to identify one locus that was identified by uGAS (C5:1-DC at
332 rs2943644), but identified five loci that were missed by uGAS. Three of the five loci identified
333 solely by GGM-cGAS affect amino acids, and the remaining two loci affect acylcarnitines. It is
334 important to note that the loci identified by BN-cGAS (when we analyzed 105 metabolites) are a
335 subset of the 19 loci that were identified by GGM-cGAS (when we used all 151 metabolites).

336 Finally, we searched the available literature for the loci listed in Table 2 (see
337 Supplementary Note 2 for details). From the 20 loci that we report here, 15 were found to be
338 significant at the genome-wide level in a recent large (n=7478) meta-analysis of Biocrates
339 metabolomics data reported by Draisma et al. [7]. Some of the metabolites analyzed in our study
340 were not analyzed by Draisma et al. [7]; nevertheless, for 11 out of these 15 loci, we observed a
341 significant association for the same SNP-metabolite pair; for three loci, the strongest association
342 was with a metabolite in the same class, and for one locus the strongest association was with a
343 metabolite from a different lipid class (see Supplementary Table S2). For the other five loci that
344 were not significant in the study by Draisma et al. [7], we determined whether these five loci were
345 significant and replicated in a study by Tsepilov et al. [19]. It should be noted that Tsepilov et al.
346 analyzed the ratios of metabolites and also used the KORA F4 data set in their discovery stage,
347 although they used another cohort (TwinsUK) for replication. Of these five loci, two were also

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

348 significant in the study by Tsepilov et al. [19]; moreover, for both of these loci the metabolite
349 analyzed in our study was included in the ratios analyzed by Tsepilov et al. One of the five loci
350 was associated with the same trait in two other studies [20,21]. Finally, we found no prior
351 published evidence of any association with metabolites for rs2943644 (*LOC646736*) or
352 rs17112944 (*LOC728755*). Taking into account that this association was not found in (much)
353 bigger meta-analysis, we conclude the observed associations with rs17112944 and rs2943644 as
354 likely false positives, and these two loci were excluded from further consideration.

16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Table 1. Eleven loci identified by BN-cGAS and uGAS on metabolites for which at least one one-reaction-step neighbor was available.

| Locus | SNP | Metabolite | chr:pos | Gene | effA/refA | EAF | uGAS | | cGAS | | N _{cov} | Noise | Pleiotropic |
|------------------------|------------|-----------------|--------------|---------|-----------|------|--------------|------------|--------------|------------|------------------|-------|-------------|
| | | | | | | | beta (se) | P-value GC | Beta (se) | P-value GC | | | |
| <i>uGAS & cGAS</i> | | | | | | | | | | | | | |
| 1 | rs211718 | C8 | 1:75879263 | ACADM | T/C | 0.3 | -0.45(0.034) | 6.35E-39 | -0.10(0.012) | 4.45E-17 | 1 | 0.92 | -1.29 |
| 1 | rs211718 | C12 | 1:75879263 | ACADM | T/C | 0.3 | -0.04(0.036) | 2.21E-01 | 0.20(0.014) | 4.07E-42 | 3 | 0.80 | 1.29 |
| 2 | rs7705189 | PC ae C42:5 | 5:131651257 | SLC22A4 | G/A | 0.47 | 0.15(0.034) | 8.83E-06 | 0.06(0.009) | 9.63E-11 | 3 | 1.16 | -0.83 |
| 2 | rs419291 | C5 | 5:131661254 | SLC22A4 | T/C | 0.38 | 0.26(0.035) | 6.62E-14 | 0.17(0.029) | 1.40E-08 | 1 | 0.16 | -0.40 |
| 3 | rs9368564 | PC aa C42:5 | 6:11168269 | ELOVL2 | G/A | 0.25 | -0.29(0.039) | 4.64E-14 | -0.15(0.024) | 1.06E-10 | 3 | 0.45 | -0.58 |
| 4 | rs12356193 | C0 | 10:61083359 | SLC16A9 | G/A | 0.17 | -0.51(0.046) | 4.93E-28 | -0.42(0.042) | 8.83E-23 | 1 | 0.07 | -0.17 |
| 5 | rs174547 | lyso-PC a C20:4 | 11:61327359 | FADS1 | C/T | 0.7 | 0.61(0.033) | 2.12E-75 | 0.66(0.024) | 2.65E-169 | 1 | 0.29 | 0.07 |
| 6 | rs2066938 | C4 | 12:119644998 | ACADS | G/A | 0.27 | 0.73(0.033) | 1.07E-104 | 0.72(0.031) | 4.26E-116 | 1 | 0.05 | 0.00 |
| 7 | rs10873201 | PC ae C36:5 | 14:67036352 | PLEKHH1 | T/C | 0.45 | -0.26(0.034) | 6.34E-14 | -0.21(0.018) | 5.72E-31 | 2 | 0.55 | -0.18 |
| 7 | rs1077989 | PC ae C32:2 | 14:67045575 | PLEKHH1 | C/A | 0.46 | -0.30(0.034) | 9.22E-19 | -0.06(0.016) | 5.39E-05 | 3 | 0.66 | -1.34 |
| 8 | rs4814176 | PC ae C40:2 | 20:12907398 | SPTLC3 | T/C | 0.36 | 0.24(0.035) | 5.60E-12 | 0.25(0.023) | 1.28E-25 | 4 | 0.35 | 0.02 |
| <i>Only uGAS</i> | | | | | | | | | | | | | |
| 9 | rs8396 | C10 | 4:159850267 | ETFDH | C/T | 0.71 | 0.26(0.037) | 1.32E-12 | 0.05(0.010) | 5.08E-07 | 2 | 1.09 | -1.39 |
| <i>Only cGAS</i> | | | | | | | | | | | | | |
| 10 | rs4862429 | PC ae C42:5 | 4:186006834 | ACSL1 | T/C | 0.31 | 0.02(0.037) | 6.63E-01 | -0.06(0.010) | 7.01E-11 | 3 | 1.15 | 1.20 |
| 11 | rs603424 | Lyso-PC a C16:1 | 10:102065469 | PKD2LI | A/G | 0.8 | 0.23(0.042) | 4.83E-08 | 0.21(0.031) | 1.76E-11 | 1 | 0.26 | -0.07 |

Notes: The best SNP-metabolite pair is shown for each locus. chr:pos refers to the physical position of the SNP; EAF, effect allele frequency; beta (se), the estimated effect and standard error of the SNP; effA/refA, effect allele/reference allele; P-value, the p-value of the additive model; Gene, the most likely (according to DEPICT) associated gene in the region; N_{cov}, the number of covariates used in cGAS; Noise/Pleiotropic, the values of noise and pleiotropic components of the log-ratio of cGAS and uGAS T² statistics.

16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Table 2. Twenty loci identified by GGM-cGAS and uGAS.

| LOCUS | SNP | Metabolite | chr:pos | Gene | effA/refA | EAF | uGAS | | cGAS | | N _{cov} | Noise | Pleiotropic |
|------------------------|------------|-----------------|--------------|-----------|-----------|------|--------------|------------|--------------|------------|------------------|-------|-------------|
| | | | | | | | beta (se) | P-value GC | beta (se) | P-value GC | | | |
| <i>uGAS & cGAS</i> | | | | | | | | | | | | | |
| 1 | rs211718 | C6 (C4:1-DC) | 1:75879263 | ACADM | T/C | 0.30 | -0.48(0.034) | 3.31E-44 | -0.13(0.017) | 1.21E-13 | 7 | 0.61 | -1.16 |
| 1 | rs7552404 | C6 (C4:1-DC) | 1:75908534 | ACADM | G/A | 0.30 | -0.48(0.034) | 2.14E-44 | -0.12(0.017) | 2.34E-13 | 7 | 0.61 | -1.17 |
| 2 | rs483180 | Ser | 1:120069028 | PHGDH | G/C | 0.30 | -0.24(0.037) | 2.26E-11 | -0.24(0.028) | 1.10E-17 | 2 | 0.24 | -0.02 |
| 2 | rs477992 | Ser | 1:120059099 | PHGDH | A/G | 0.70 | 0.24(0.037) | 3.50E-11 | 0.24(0.028) | 2.52E-18 | 2 | 0.24 | 0.00 |
| 3 | rs2286963 | C9 | 2:210768295 | ACADL | G/T | 0.63 | -0.49(0.032) | 4.76E-52 | -0.48(0.027) | 7.41E-73 | 3 | 0.16/ | -0.01 |
| 4 | rs8396 | C10 | 4:159850267 | ETFDH | C/T | 0.71 | 0.26(0.037) | 1.32E-12 | 0.04(0.010) | 1.23E-05 | 8 | 1.11 | -1.53 |
| 4 | rs8396 | C7-DC | 4:159850267 | ETFDH | C/T | 0.71 | -0.09(0.037) | 1.67E-02 | -0.13(0.019) | 2.93E-11 | 8 | 0.56 | 0.33 |
| 5 | rs419291 | C5 | 5:131661254 | SLC22A4 | T/C | 0.38 | 0.26(0.035) | 6.62E-14 | 0.17(0.026) | 2.25E-10 | 3 | 0.25 | -0.40 |
| 5 | rs270613 | C5 | 5:131668482 | SLC22A4 | A/G | 0.61 | -0.26(0.035) | 7.48E-14 | -0.17(0.026) | 8.24E-11 | 3 | 0.25 | -0.38 |
| 6 | rs9393903 | PC aa C42:5 | 6:11150895 | ELOVL2 | A/G | 0.75 | 0.29(0.039) | 9.13E-14 | 0.18(0.020) | 1.32E-19 | 6 | 0.56 | -0.38 |
| 6 | rs9368564 | PC aa C42:5 | 6:11168269 | ELOVL2 | G/A | 0.25 | -0.29(0.039) | 4.64E-14 | -0.19(0.021) | 3.04E-19 | 6 | 0.56 | -0.40 |
| 7 | rs816411 | Ser | 7:56138983 | PHKG1 | C/T | 0.51 | -0.22(0.034) | 1.53E-10 | -0.19(0.026) | 4.83E-13 | 2 | 0.23 | -0.12 |
| 7 | rs1894832 | Ser | 7:56144740 | PHKG1 | C/T | 0.51 | 0.21(0.034) | 2.33E-10 | 0.19(0.026) | 1.55E-13 | 2 | 0.23 | -0.09 |
| 8 | rs12356193 | C0 | 10:61083359 | SLC16A9 | G/A | 0.17 | -0.51(0.046) | 4.93E-28 | -0.27(0.034) | 1.03E-15 | 3 | 0.26 | -0.53 |
| 9 | rs174547 | lyso-PC a C20:4 | 11:61327359 | FADS1 | C/T | 0.70 | 0.61(0.033) | 2.12E-75 | 0.07(0.011) | 2.08E-10 | 9 | 0.98 | -1.90 |
| 9 | rs174556 | PC ae C44:4 | 11:61337211 | FADS1 | T/C | 0.27 | 0.09(0.038) | 1.61E-02 | 0.21(0.014) | 1.17E-48 | 3 | 0.84 | 0.73 |
| 10 | rs2066938 | C4 | 12:119644998 | ACADS | G/A | 0.27 | 0.73(0.033) | 1.07E-104 | 0.71(0.024) | 6.95E-189 | 2 | 0.28 | -0.02 |
| 11 | rs12879147 | PC aa C28:1 | 14:63297349 | SYNE2 | A/G | 0.85 | -0.46(0.050) | 1.83E-19 | -0.12(0.019) | 6.87E-11 | 14 | 0.86 | -1.14 |
| 11 | rs17101394 | SM(OH) C14:1 | 14:63302139 | SYNE2 | A/G | 0.83 | -0.32(0.050) | 1.02E-10 | -0.10(0.011) | 9.23E-18 | 7 | 1.30 | -1.05 |
| 12 | rs1077989 | PC ae C36:5 | 14:67045575 | PLEKHH1 | C/A | 0.46 | -0.26(0.034) | 4.96E-14 | -0.08(0.010) | 2.56E-15 | 10 | 1.05 | -1.00 |
| 12 | rs1077989 | PC ae C32:2 | 14:67045575 | PLEKHH1 | C/A | 0.46 | -0.30(0.034) | 9.22E-19 | -0.05(0.016) | 1.35E-03 | 6 | 0.67 | -1.55 |
| 13 | rs4814176 | SM(OH).C22:1 | 20:12907398 | SPTLC3 | T/C | 0.36 | 0.03(0.035) | 4.53E-01 | -0.07(0.009) | 9.11E-17 | 10 | 1.22 | 0.87 |
| 13 | rs4814176 | SM(OH) C24:1 | 20:12907398 | SPTLC3 | T/C | 0.36 | 0.24(0.035) | 4.29E-12 | 0.09(0.013) | 2.85E-11 | 9 | 0.86 | -0.90 |
| 14 | rs5746636 | Pro | 22:17276301 | PRODH | T/G | 0.24 | -0.31(0.039) | 1.89E-15 | -0.32(0.034) | 5.05E-21 | 2 | 0.11 | 0.03 |
| <i>Only uGAS</i> | | | | | | | | | | | | | |
| 15 | rs2943644 | C5:1-DC | 2:226754586 | LOC646736 | C/T | 0.68 | 0.32(0.042) | 3.99E-14 | 0.09(0.022) | 3.97E-05 | 5 | 0.56 | -1.08 |

16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

| Only cGAS | | | | | | | | | | | | | |
|-----------|------------|-------------|--------------|-----------|-----|------|--------------|----------|--------------|----------|---|------|-------|
| 16 | rs1374804 | Gly | 3:127391188 | ALDH1L1 | A/G | 0.64 | 0.20(0.036) | 1.46E-08 | 0.21(0.029) | 3.65E-13 | 3 | 0.17 | 0.05 |
| 17 | rs4862429 | PC ae C42:5 | 4:186006834 | ACSL1 | T/C | 0.31 | 0.02(0.037) | 6.63E-01 | -0.06(0.008) | 1.15E-12 | 8 | 1.34 | 1.09 |
| 18 | rs603424 | C16:1 | 10:102065469 | PKD2L1 | A/G | 0.80 | 0.16(0.042) | 9.00E-05 | 0.14(0.018) | 9.32E-14 | 9 | 0.71 | -0.15 |
| 19 | rs2657879 | Gln | 12:55151605 | GLS2 | G/A | 0.21 | -0.24(0.042) | 2.65E-08 | -0.27(0.030) | 5.88E-18 | 5 | 0.29 | 0.10 |
| 20 | rs17112944 | C6:1 | 14:27179297 | LOC728755 | A/G | 0.90 | -0.28(0.059) | 1.98E-06 | -0.21(0.031) | 3.74E-11 | 9 | 0.54 | -0.26 |

Notes: The best SNP-metabolite pair is shown for each locus. chr:pos refers to the physical position of the SNP; EAF, effect allele frequency; beta (se), the estimated effect and standard error of the SNP; effA/refA, effect allele/reference allele; *P*-value, the *p*-value of the additive model; Gene, the most likely (according to DEPICT) associated gene in the region; N_{cov}, the number of covariates used in cGAS; Noise/Pleiotropic, the values of noise and pleiotropic components of the log-ratio of cGAS and uGAS T² statistics.

1 367 **Discussion**

2
3
4 368 We report a new “trait-centric” approach for analyzing genetic determinants of multivariate
5 369 “omics” traits by performing a network-based conditional genetic association analysis (cGAS). In
6
7 370 the context of metabolomics, for each trait we selected a set of other metabolites to be used as
8
9 371 covariates in our genetic association analysis. The selection of covariates can be either mechanistic
10
11 372 (e.g., based on known biological relationships between traits of interest) or data-driven (e.g., based
12
13 373 on partial correlations). Importantly, this approach can use either individual-level or summary-
14
15 374 level data. We first mathematically compared the power of conditional and standard single-trait
16
17 375 genetic association analyses (univariate genetic association, uGAS), and we identified scenarios
18
19 376 in which these analyses are expected to produce different results; next, we applied cGAS to 151
20
21 377 metabolomics traits (Biocrates panel) in a large (n=1784 individuals) population-based KORA
22
23 378 cohort.

24 379 We found that the log-ratio between the cGAS and uGAS test statistic can be decomposed
25
26 380 in a “noise” component (which depends on residual variance of the trait and is always positive)
27
28 381 and a “pleiotropic” component. The pleiotropic component is negative in cases in which
29
30 382 genetically induced covariance (between the trait of interest and the trait used as the covariate) and
31
32 383 the residual covariance have the same sign (i.e., act in the same direction). The pleiotropic
33
34 384 component is positive in cases in which the genetically induced covariance and residual covariance
35
36 385 act in opposite directions.

37 386 Should one expect that genetically induced and residual covariance act in the same or
38
39 387 opposite directions? In essence, this is a question about the architecture of pleiotropy: is a
40
41 388 pleiotropic genetic variant expected to induce the same covariance as would be induced by non-
42
43 389 genetic mechanisms? It has been reasoned that in randomly bred populations, the genetic
44
45 390 correlations are expected to arise primarily from pleiotropic gene action [22]. In such populations,
46
47 391 a study and comparison of genetic and environmental correlations—while unable to provide
48
49 392 single-variant resolution—may provide a general notion of what may be expected for
50
51 393 consistency/anti-consistency between genetic and residual covariance. Based on published
52
53 394 literature, Cheverud [23] and Roff [24] concluded that genetic and environmental correlations
54
55 395 normally have both the same sign and the same magnitude. This pattern is particularly clear for
56
57 396 morphologic traits, as opposed to life-history traits (see [25] for review and additional references).
58
59 397 These observations are consistent with recent studies of genetic correlations between complex
60
61 398 human polygenic traits (see [26]).
62
63
64
65

399 Consequently, for complex traits, one may expect that the sign of the pleiotropic
400 component of the log-ratio between the cGAS and uGAS tests (individual summands in the second
401 term of the equation (2)) is generally negative. It should be noted, though, that a negative sign for
402 the pleiotropic component does not necessarily indicate higher power of the uGAS, as the noise
403 component (the first term in equation (2)) may still dominate the log-ratio between the cGAS and
404 uGAS tests. This will happen, for example, when $\hat{\rho}_{cg}$ (the effect of the genotype on the covariate)
405 is small while $\hat{\beta}_{yc}$ (partial residual regression between the trait and covariate) is relatively large,
406 thereby reducing $\hat{\sigma}_c^2$.

407 Nevertheless, in the case of metabolomic traits, genetic and environmental sources do not
408 necessarily generate consistent covariance. Moreover, for a given locus that affects the activity of
409 an enzyme involved in a biochemical reaction, the unexpected inconsistency between genetically
410 induced covariance and residual covariance may not be so unexpected after all. Indeed, consider
411 an allele associated with an increased activity of an enzyme that converts substrate A into product
412 B. One would expect that the levels of A and B are positively correlated; one would also expect
413 that the allele is positively correlated with the level of product B and negatively correlated with
414 the level of substrate A. This is precisely the scenario that yields a positive value for the second
415 term in Eq. (1), thus providing an additional increase in power above and beyond the power
416 provided by the first term in Eq. (1) (noise reduction).

417 Our empirical investigation of real data on the genetic association between the genome and
418 metabolites confirmed the existence of both scenarios. An extreme example of concordance
419 between genetic covariance and residual covariance is provided by the effects of rs8396 on C10,
420 with C8 and C12 used as covariates (see Figure 2B). The *ETFDH* gene, which was prioritized by
421 DEPICT software (see Materials and Methods) as the best candidate in this region (with a false-
422 discover rate <5%), encodes the enzyme electron transfer flavoprotein (ETF) dehydrogenase,
423 which plays a role in mitochondrial fatty acid oxidation. During this process, the acyl group is
424 transferred from a long chain acylcarnitine to a long-chain acetyl-CoA, which is then catabolized.
425 ETF dehydrogenase participates in the catabolic process by transferring electrons from acyl-CoA
426 dehydrogenase to the oxidative phosphorylation pathway. Thus, the *ETFDH* gene should affect all
427 forms of long-chain acylcarnitines in the same way, and we can expect that the pleiotropic effect
428 of this gene on the acylcarnitines in our example (C8, C10, C12, etc.) will be unidirectional. The
429 presence of unidirectional genetic effects and the positive correlation between these acylcarnitines
430 makes the second term in Eq. (2) negative, which determines that—in this situation—univariate
431 GAS has more power than cGAS.

432 An empirical example of discordance between genetically induced covariance and residual
433 covariance is provided by the effects of the SNP rs174547 on lyso-PC a C20:4, with lyso-PC a

434 C20:3 used as a covariate. This SNP exhibits opposite correlations with lyso-PC a C20:4 and lyso-
435 PC a C20:3, resulting in negative genetically induced covariance between these traits. At the same
436 time, the residual correlation between these traits is positive, resulting in steep increase in the
437 power of conditional analysis. In this region, the *FADS1/2/3* gene cluster is an attractive candidate,
438 providing the detected model with biological relevance. The *FADS1* gene encodes the enzyme
439 fatty acid desaturase 1, whereas the two traits differ by only one double bond. Thus, this example
440 mimics perfectly the biochemical scenario in which we would expect a conditional analysis to
441 have increased power.

442 The trait-centric methods considered here provide an attractive framework to identify and
443 study direct genetic effects on a trait of interest. Conditional analysis is an attractive option in cases
444 in which we wish to clearly interpret the results in terms of the effect of the genotype on a particular
445 trait. Such specific interpretation may be important when comparing genetic association results
446 obtained for our trait of interest with results obtained for other traits (e.g., using the methods
447 described in [27–29]). It should be noted, though, that a trait-centric approach is not intended to
448 maximize the power of identifying genes that affect metabolomics as a whole. Such a gene-centric
449 view would favor analysis using joint—and not conditional—modeling of sets of traits. Such an
450 approach can maintain power across a wide range of scenarios, including the scenario of
451 concordance between genetically induced and residual covariance [13]. In this gene-centric
452 framework, other formulations of conditional analysis have also been proposed [30] in order to
453 specifically increase power of gene identification by selecting covariates that—using our
454 terminology—affect the “noise reduction” component of the model while avoiding the problems
455 associated with the pleiotropic component.

456 The proper selection of sets of biologically related traits is extremely important for the
457 conditional genetic association analysis method described here, as well as for multivariate methods
458 that model the joint effects of genotype on an ensemble of traits. Here, we considered two
459 alternative approaches—knowledge based and data-driven—to finding the networks of related
460 traits, with a subnetwork centered around a trait of interest used as the analyzed set. In principle,
461 in the context of analyzed metabolomics data, the knowledge-based network approach has slightly
462 higher power in the context of trait-centric genetic association analysis. However, we believe that
463 our analysis revealed that both approaches are suboptimal. The knowledge-based network
464 reconstruction has many advantages, but it may be somewhat unpractical, as our biochemical
465 knowledge is still relatively incomplete. Secondly, by reconstructing the network while relying
466 only on current knowledge, we may be missing new knowledge that may be revealed by the data.
467 Finally, by including neighbors that are based only on biochemical information, we may miss
468 covariance induced by technical confounders; adjusting for this may increase the power of analysis

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

469 [30]. Learning the network from the same data that were used for genetic analysis has the
470 disadvantages of potentially ignoring existing knowledge and being sensitive to sample size.
471 Finally, we note that the total observed correlation between metabolites is determined by the
472 balance between genetic and environmental sources of covariance; it is possible to imagine a
473 situation in which total correlation is smaller than one or more of its components, and our analysis
474 provides examples of such a situation. We may speculate that—ideally—one should use a method
475 that allows one to combine prior knowledge and new information obtained from the data, thereby
476 allowing the simultaneous learning of the structure of dependencies between different metabolites
477 and between the metabolites and the genome. Such learning from the data while allowing for the
478 incorporation of previous knowledge (e.g., biochemical relations between traits) might be achieved
479 (for example, by applying a machine-learning approach that allows for differential shrinkage). It
480 is also important to note that the proper application of such an approach would require the
481 availability of vast samples of data, thereby allowing for separate training, validation, verification,
482 testing, and replication of detected dependencies and associations.

483 **Materials and Methods**

484 **KORA study**

485 The KORA study (Cooperative Health Research in the region of Augsburg) is a series of
486 population-based studies in the region of Augsburg in Southern Germany [31]. KORA F4 is a
487 follow-up survey (conducted from 2006 through 2008) of the baseline KORA S4 survey, which
488 was conducted from 1999 through 2001. All study protocols were approved by the ethics
489 committee of the Bavarian Medical Chamber, and all participants provided written informed
490 consent.

491 The concentration of 163 metabolites were measured in 3061 serum samples obtained from
492 KORA F4 participants using flow injection electrospray ionization tandem mass spectrometry and
493 the AbsoluteIDQ p150 Kit (Biocrates Life Sciences AG, Innsbruck, Austria) [32]. After applying
494 quality control screening, a total of 151 metabolite measurements were used in our analysis.
495 Details regarding the methods and quality control of the metabolite measurements, as well as
496 details regarding the metabolite nomenclature, have been published previously [32]. The
497 nomenclature for the metabolites in this study is provided in **Supplementary Table S3**.

498 Genotyping was performed using the Affymetrix 6.0 SNP array (534,174 SNP markers
499 after quality control), with further imputation using HapMap2 (release 22) as a reference panel,
500 resulting in a total of 1,717,498 SNPs (for details, see ref. [33]). Both the metabolite concentrations
501 and genotype were available for 1785 participants in the KORA F4 study.

502

503 **Statistical analysis**

504 Partial correlation coefficients and their p -values were calculated using the “ppcor” package [34]
505 in R. Graphical representations were generated using the “ggm” [35] package in R. Consistent
506 with previous studies [18], we considered a partial regression coefficient to be significant at p
507 $< 0.01/(151*150/2)$ (i.e., $p < 8.83 \times 10^{-7}$).

508 For the GWAS analysis, we used OmicABEL software [36]. Prior to GWAS, all traits were
509 first adjusted for the participant’s sex, age, and batch effect; subsequently, the residual traits were
510 transformed using an inverse-normal transformation [37]. The genotypes from the KORA F4
511 cohort were used. Only SNPs that had a call rate ≥ 0.95 , $R^2 \geq 0.3$, Hardy–Weinberg equilibrium
512 (HWE) $p \geq 10^{-6}$, and MAF ≥ 0.1 (1,717,498 SNPs in total) were included in the analysis. The
513 genomic control method was used to correct for any possible inflation of the test statistics. The
514 genomic control [38] lambda value for all traits was between 1.00 and 1.03.

515 In a specific analysis (i.e., cGAS or uGAS), we defined independent loci as groups of
516 genome-wide significant associations that were separated by at least 500 kb or were located on

517 different chromosomes. The strongest association (i.e., the association with the lowest p -value)
518 was selected to represent this locus. The cGAS and uGAS results were considered to reflect
519 different loci if the strongest associations were in loci that were separated by at least 500 kb. The
520 threshold for the genome-wide significance for 151 traits was set to $p=5 \times 10^{-8}/151$ (i.e.,
521 $p=3.31 \times 10^{-10}$).

522 When partitioning the log(cGAS/uGAS) test statistics into the noise and pleiotropic
523 components (see Eq. (2) and **Figure 1**), we used all known loci that were significant in either the
524 cGAS or uGAS analysis (see Table 1). If a locus included two SNPs associated with different
525 traits, we included both associations during partitioning. If a locus included two SNPs associated
526 with the same trait, to be conservative we included only the SNP with the lower uGAS p -value
527 during partitioning. After partitioning, we determined whether the value of the pleiotropic and
528 noise components were statistically different using the paired-samples Wilcoxon test. For
529 comparing the chi-square test results for the two methods, for each locus we selected **the largest**
530 **chi-square value for selected SNP among all analyzed traits. If locus had two SNPs we selected**
531 **for each method only the largest chi-square value.**

532 The code for BN-cGAS and GGM-cGAS analyses, and the code for producing the
533 summary tables and graphs, was implemented in R and is available as a workflow from
534 CodeOcean, a cloud-based computational reproducibility platform.

536 ***In silico* functional annotation**

537 We conducted functional annotation for our findings. To prioritize genes in associated regions,
538 gene set enrichment, and tissue/cell-type enrichment analyses, we used DEPICT (Data-driven
539 Expression-Prioritized Integration for Complex Traits) software [39] (release 140721) with the
540 following settings: flag_loci = 1; flag_genes = 1; flag_genesets = 1; flag_tissues = 1; param_ncores
541 = 2; and further manual annotation (h37 assembly). All 27 SNPs (clustered in 20 loci) identified
542 by cGAS or uGAS (see **Table 2**) were included in the analysis. If more than one gene was
543 annotated for a SNP by DEPICT, we selected the gene with the lowest nominal DEPICT P -value.
544 In most cases, the results of manual annotation matched the annotation results using DEPICT
545 annotation (see **Supplementary Note 2**). In addition, we looked up each SNP using the
546 Phenoscanner [40] database to check whether it was previously reported to be associated with
547 metabolic traits at $p < 5 \times 10^{-8}$ and proxy $r^2 < 0.7$.

548

549 **Additional Files**

1 550 Supplementary Note 1 – cGAS using summary level data
2

3 551 Supplementary Note 2 – Literature search for loci identified by cGAS and uGAS
4

5 552 Supplementary Tables
6

7 553 Supplementary Table S1 – BN- cGAS and GGM- cGAS for 105 metabolites
8

9 554 Supplementary Table S2 – GGM-cGAS and uGAS for 151 metabolites
10

11 555 Supplementary Table S3 - List of metabolites measured using the AbsoluteIDQ p150 Kit
12

13 556 Supplementary Figures
14

15 557 Supplementary Figure S1 – Partial correlations network
16

17 558 Supplementary Figure S2 – Manhattan plots for cGAS and uGAS for 151 metabolites
18

19 559 Supplementary Figure S3 – Comparison of effect estimates and their standard errors for
20

21 560 SNPs from Table 2
22

23 561

24 562 **Abbreviations**
25

26 563 GWAS – genome-wide association study
27

28 564 cGAS – conditional GWAS
29

30 565 uGAS – univariate GWAS (trait-by-trait)
31

32 566 BN-cGAS – cGAS based on biochemical networks
33

34 567 GGM-cGAS – Gaussian Graphical Modeling cGAS based on partial correlations network
35

36 568

37 569 **Acknowledgments**
38

39 570 We thank Athina Spilopoulou and Felix Agakov for helpful discussions. We also thank Alexander
40

41 571 Zlobin and Alexander Grishenko for help preparing the tables and figures in the manuscript, and
42

43 572 we thank Sophie Molnos for help with data management. **We thank the reviewers - Simina Boca**
44

45 573 **and Tim Ebbels—for their very helpful input.**
46

47 574

48 575 **Funding**
49

50 576 The KORA study was initiated and financed by the Helmholtz Center Munich – German
51

52 577 Research Center for Environmental Health, which is funded by the German Federal Ministry of
53

54 578 Education and Research (BMBF) and by the State of Bavaria. Furthermore, the KORA study
55

56 579 was supported by the Munich Center of Health Sciences (MC-Health), Ludwig Maximilian
57

58 580 University of Munich, as part of the LMUinnovativ project.
59

60 581 This work was supported by the European Union FP7 framework project Pain-Omics (grant
61
62
63
64
65

582 number 602736).

1 583 SS was supported by the Russian Ministry of Science and Education under the 5-100 Excellence
2 Programme. YA and YT were supported by the Federal Agency of Scientific Organisations via
3 584 the Institute of Cytology and Genetics (project number 0324-2018-0017).
4
5

6 586

7
8
9 **587 Authors Contributions**

10
11 588 YT, CG, and YA designed and supervised the study; PC, CP, JA, KG, and RW-S collected the
12 data; CG and KS contributed data for the analysis; YT, OZ, and SS analyzed the data; YT, YA,
13 589 CG, OZ, JK, and KS discussed and interpreted the results; YT, OZ, CG, and YA wrote the
14 manuscript. All authors contributed to and approve the final version of the manuscript.
15 590
16 591

17 592

18
19
20 **593 Availability of Data and Materials**

21
22
23 594 The code produced in relation to this work is distributed under the MIT license and is available
24 from [https://codeocean.com/2018/04/02/a-network-based-conditional-genetic-association-](https://codeocean.com/2018/04/02/a-network-based-conditional-genetic-association-analysis-of-the-human-metabolome/)
25 595 [analysis-of-the-human-metabolome/](https://codeocean.com/2018/04/02/a-network-based-conditional-genetic-association-analysis-of-the-human-metabolome/). All summary statistics and association data that are
26 596 necessary to reproduce our results are licensed under CC0 and are accessible online at
27 [https://codeocean.com/2018/04/02/a-network-based-conditional-genetic-association-analysis-of-](https://codeocean.com/2018/04/02/a-network-based-conditional-genetic-association-analysis-of-the-human-metabolome/)
28 597 [the-human-metabolome/](https://codeocean.com/2018/04/02/a-network-based-conditional-genetic-association-analysis-of-the-human-metabolome/) . The informed consent given by the KORA study participants does not
29 598 cover the posting of participant-level phenotype or genotype data in public databases. However,
30 600 the KORA data are available upon request from KORA-gen ([https://www.helmholtz-](https://www.helmholtz-muenchen.de/en/kora/index.html)
31 601 [muenchen.de/en/kora/index.html](https://www.helmholtz-muenchen.de/en/kora/index.html)). Requests can be submitted online and are subject to approval
32 602 by the KORA board.
33
34 603
35
36 604

37
38
39
40
41
42
43 **605 Competing Interests**

44
45
46 606 Y. Aulchenko is the founder and co-owner of PolyOmica, a private research organization that
47 specializes in computational and statistical (gen)omics.
48 607
49
50 608

609 References

- 1 610 1. Visscher PM, Brown MA, McCarthy MI, Yang J. Five Years of GWAS Discovery. *Am J Hum*
2 611 *Genet* [Internet]. 2012;90:7–24. Available from:
3 <http://linkinghub.elsevier.com/retrieve/pii/S0002929711005337>
4 612
- 5 613 2. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to
6 614 uncover genotype–phenotype interactions. *Nat Rev Genet* [Internet]. 2015;16:85–97. Available
7 615 from: <http://www.nature.com/doifinder/10.1038/nrg3868>
8 616
- 9 617 3. van der Sijde MR, Ng A, Fu J. Systems genetics: From GWAS to disease pathways. *Biochim*
10 617 *Biophys Acta - Mol Basis Dis* [Internet]. 2014;1842:1903–9. Available from:
11 618 <http://linkinghub.elsevier.com/retrieve/pii/S0925443914001124>
12 619
- 13 620 4. Hicks AA, Pramstaller PP, Johansson A, Vitart V, Rudan I, Ugocsai P, et al. Genetic
14 620 determinants of circulating sphingolipid concentrations in European populations. *PLoS Genet*
15 621 [Internet]. 2009 [cited 2013 Dec 19];5:e1000672. Available from:
16 622 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2745562&tool=pmcentrez&renderty>
17 623 [pe=abstract](#)
18 624
- 19 625 5. Suhre K, Shin S-Y, Petersen A-K, Mohny RP, Meredith D, Wägele B, et al. Human metabolic
20 625 individuality in biomedical and pharmaceutical research. *Nature* [Internet]. 2011 [cited 2013 Dec
21 626 19];477:54–60. Available from:
22 627 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3832838&tool=pmcentrez&renderty>
23 628 [pe=abstract](#)
24 629
- 25 630 6. Inouye M, Ripatti S, Kettunen J, Lyytikäinen L-P, Oksala N, Laurila P-P, et al. Novel Loci for
26 630 metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis. Visscher
27 631 PM, editor. *PLoS Genet* [Internet]. 2012;8:e1002907. Available from:
28 632 <http://dx.plos.org/10.1371/journal.pgen.1002907>
29 633
- 30 634 7. Draisma HHM, Pool R, Kobl M, Jansen R, Petersen A-K, Vaarhorst AAM, et al. Genome-wide
31 634 association study identifies novel genetic variants contributing to variation in blood metabolite
32 635 levels. *Nat Commun* [Internet]. England; 2015;6:7208. Available from:
33 636 <http://www.ncbi.nlm.nih.gov/pubmed/26068415>
34 637
- 35 638 8. Kettunen J, Demirkan A, Würtz P, Draisma HHMM, Haller T, Rawal R, et al. Genome-wide
36 638 study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat*
37 639 *Commun* [Internet]. 2016;7:11122. Available from:
38 640 <http://www.ncbi.nlm.nih.gov/pubmed/27005778> <http://www.nature.com/doifinder/10.1038/ncomms11122>
39 641
- 40 642 9. Cichonska A, Rousu J, Marttinen P, Kangas AJ, Soininen P, Lehtimäki T, et al. metaCCA:
41 643 summary statistics-based multivariate meta-analysis of genome-wide association studies using
42 644 canonical correlation analysis. *Bioinformatics* [Internet]. 2016;32:1981–9. Available from:
43 645 <http://www.ncbi.nlm.nih.gov/pubmed/27153689>
44 646
- 45 646 10. Stephens M. A unified framework for association analysis with multiple related phenotypes.
46 647 *PLoS One* [Internet]. 2013;8:e65245. Available from:
47 648 <http://www.ncbi.nlm.nih.gov/pubmed/23861737>
48 649
- 49 649 11. O’Reilly PF, Hoggart CJ, Pomyen Y, Calboli FCF, Elliott P, Jarvelin M-R, et al. MultiPhen:
50 650 joint model of multiple phenotypes can increase discovery in GWAS. *PLoS One* [Internet]. 2012
51 651 [cited 2014 Sep 20];7:e34861. Available from:
52 652 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3342314&tool=pmcentrez&renderty>
53 653 [pe=abstract](#)
54 654
- 55 654 12. Galesloot TE, van Steen K, Kiemeny LALM, Janss LL, Vermeulen SH. A comparison of
56 655 multivariate genome-wide association methods. *PLoS One* [Internet]. 2014 [cited 2014 Sep
57 656 20];9:e95923. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24763738>
58 657
- 59 657 13. Shen X, Klarić L, Sharapov S, Mangino M, Ning Z, Wu D, et al. Multivariate discovery and
60 658 replication of five novel loci associated with Immunoglobulin G N-glycosylation. *Nat Commun*
61 659 [Internet]. 2017;8:447. Available from: <http://www.nature.com/articles/s41467-017-00453-3>
62
63
64
65

660 14. Schaid DJ, Tong X, Larrabee B, Kennedy RB, Poland GA, Sinnwell JP. Statistical Methods
661 for Testing Genetic Pleiotropy. *Genetics* [Internet]. 2016;204:483–97. Available from:
662 <http://www.genetics.org/cgi/doi/10.1534/genetics.116.189308>

663 15. Deng Y, Pan W. Conditional analysis of multiple quantitative traits based on marginal GWAS
664 summary statistics. *Genet Epidemiol* [Internet]. 2017;41:427–36. Available from:
665 <http://www.ncbi.nlm.nih.gov/pubmed/28464407>

666 16. Cox DR, Hinkley D V. Theoretical statistics. 1974. London, Chapman Hall. 1:511.

667 17. Smith GD, Ebrahim S. “Mendelian randomization”: can genetic epidemiology contribute to
668 understanding environmental determinants of disease? *Int J Epidemiol* [Internet]. 2003;32:1–22.
669 Available from: <http://www.ncbi.nlm.nih.gov/pubmed/12689998>

670 18. Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ. Gaussian graphical modeling reconstructs
671 pathway reactions from high-throughput metabolomics data. *BMC Syst Biol* [Internet]. BioMed
672 Central Ltd; 2011 [cited 2013 May 23];5:21. Available from:
673 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3224437&tool=pmcentrez&renderty>
674 [pe=abstract](#)

675 19. Tsepilov YA, Shin S-Y, Soranzo N, Spector TD, Prehn C, Adamski J, et al. Nonadditive
676 Effects of Genes in Human Metabolomics. *Genetics* [Internet]. 2015;200:707–18. Available from:
677 <http://www.genetics.org/cgi/doi/10.1534/genetics.115.175760>

678 20. Xie W, Wood AR, Lyssenko V, Weedon MN, Knowles JW, Alkayyali S, et al. Genetic variants
679 associated with glycine metabolism and their role in insulin sensitivity and type 2 diabetes.
680 *Diabetes* [Internet]. 2013;62:2141–50. Available from:
681 <http://www.ncbi.nlm.nih.gov/pubmed/23378610>

682 21. Shin S-Y, Fauman EB, Petersen A-K, Krumsiek J, Santos R, Huang J, et al. An atlas of genetic
683 influences on human blood metabolites. *Nat Genet* [Internet]. 2014 [cited 2014 May 12];46:543–
684 50. Available from: <http://www.nature.com/doi/10.1038/ng.2982>

685 22. Falconer DS, Mackay TFC. *Introduction to Quantitative Genetics* (4th Edition) [Internet]. 4th
686 ed. Pearson; 1996. Available from:
687 <http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0582243025>

688 23. Cheverud JM. A COMPARISON OF GENETIC AND PHENOTYPIC CORRELATIONS.
689 *Evolution* [Internet]. 1988;42:958–68. Available from:
690 <http://www.ncbi.nlm.nih.gov/pubmed/28581166>

691 24. Roff DA. The estimation of genetic correlations from phenotypic correlations: a test of
692 Cheverud’s conjecture. *Heredity* (Edinb) [Internet]. 1995;74:481–90. Available from:
693 <http://www.nature.com/articles/hdy199568>

694 25. Lynch M, Walsh B, others. *Genetics and analysis of quantitative traits*. Sinauer Sunderland,
695 MA; 1998.

696 26. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh P-R, et al. An atlas of
697 genetic correlations across human diseases and traits. *Nat Genet*. Nature Publishing Group;
698 2015;47:1236–41.

699 27. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data
700 from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* [Internet].
701 2016;48:481–7. Available from:
702 <http://www.nature.com/doi/10.1038/ng.3538>
703 <http://www.ncbi.nlm.nih.gov/pubmed/27019110>

704 28. Pickrell JK, Berisa T, Liu JZ, Séguérel L, Tung JY, Hinds DA. Detection and interpretation of
705 shared genetic influences on 42 human traits. *Nat Genet*. 2016;019885.

706 29. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al.
707 Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary
708 Statistics. Williams SM, editor. *PLoS Genet* [Internet]. 2014;10:e1004383. Available from:
709 <http://dx.plos.org/10.1371/journal.pgen.1004383>

710 30. Aschard H, Guillemot V, Vilhjalmsón B, Patel CJ, Skurnik D, Ye CJ, et al. Covariate selection
711 for association screening in multiphenotype genetic studies. *Nat Genet* [Internet]. 2017;49:1789–

- 712 95. Available from: <http://www.nature.com/doi/10.1038/ng.3975>
- 713 31. Wichmann H-E, Gieger C, Illig T. KORA-gen--resource for population genetics, controls and
714 a broad spectrum of disease phenotypes. *Gesundheitswesen* [Internet]. 2005 [cited 2013 Jun 6];67
715 Suppl 1:S26-30. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/16032514>
- 716 32. Illig T, Gieger C, Zhai G, Römisch-Margl W, Wang-Sattler R, Prehn C, et al. A genome-wide
717 perspective of genetic variation in human metabolism. *Nat Genet* [Internet]. Nature Publishing
718 Group; 2010 [cited 2013 May 23];42:137–41. Available from:
719 <http://www.ncbi.nlm.nih.gov/pubmed/20037589>
- 720 33. Kolz M, Johnson T, Sanna S, Teumer A, Vitart V, Perola M, et al. Meta-analysis of 28,141
721 individuals identifies common variants within five new loci that influence uric acid concentrations.
722 *PLoS Genet* [Internet]. 2009 [cited 2013 May 30];5:e1000504. Available from:
723 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2683940&tool=pmcentrez&rendertype=abstract>
- 724
725 34. Kim S. ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients.
726 *Commun Stat Appl Methods* [Internet]. 2015;22:665–74. Available from:
727 <http://www.csam.or.kr/journal/view.html?doi=10.5351/CSAM.2015.22.6.665>
- 728 35. Marchetti GM. Independencies Induced from a Graphical Markov Model after Marginalization
729 and Conditioning: The R Package ggm. *J Stat Softw* [Internet]. 2006;15. Available from:
730 <http://www.jstatsoft.org/v15/i06/>
- 731 36. Fabregat-Traver D, Sharapov SZ, Hayward C, Rudan I, Campbell H, Aulchenko Y, et al. High-
732 Performance Mixed Models Based Genome-Wide Association Analysis with omicABEL
733 software. *F1000Research* [Internet]. 2014;3:200. Available from:
734 <http://f1000research.com/articles/3-200/v1>
- 735 37. Beasley TM, Erickson S, Allison DB. Rank-based inverse normal transformations are
736 increasingly used, but are they merited? *Behav Genet* [Internet]. 2009 [cited 2013 Nov 7];39:580–
737 95. Available from:
738 <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2921808&tool=pmcentrez&rendertype=abstract>
- 739
740 38. Devlin B, Roeder K. Genomic control for association studies. *Biometrics* [Internet]. 1999
741 [cited 2013 Jun 5];55:997–1004. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/11315092>
- 742 39. Pers TH, Karjalainen JM, Chan Y, Westra H-J, Wood AR, Yang J, et al. Biological
743 interpretation of genome-wide association studies using predicted gene functions. *Nat Commun*
744 [Internet]. 2015;6:5890. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25597830>
- 745 40. Staley JR, Blackshaw J, Kamat MA, Ellis S, Surendran P, Sun BB, et al. PhenoScanner: a
746 database of human genotype–phenotype associations. *Bioinformatics* [Internet]. 2016;32:3207–9.
747 Available from:
748 <http://bioinformatics.oxfordjournals.org/lookup/doi/10.1093/bioinformatics/btw373>
- 749



Click here to access/download
Supplementary Material
Supplementary Figures.docx





Click here to access/download
Supplementary Material
Supplementary Note 1.docx





Click here to access/download
Supplementary Material
Supplementary Note 2.docx





Click here to access/download
Supplementary Material
Supplementary Table 1.xlsx





Click here to access/download
Supplementary Material
Supplementary Table 2.xlsx





Click here to access/download
Supplementary Material
Supplementary Table 3.docx