Supplementary Data
MetaPhat: Detecting and decomposing multivariate associations from univariate genome-wide association statistics

Jake Lin[1], Rubina Tabassum[1], Samuli Ripatti[1,2,3], Matti Pirinen[1,2,4]

1. Institute for Molecular Medicine Finland FIMM, HiLIFE, University of Helsinki, Helsinki, Finland.
2. Public Health, University of Helsinki, Helsinki, Finland.
3. Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, MA, USA.
4. Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland.

Contact: jake.lin@helsinki.fi and matti.pirinen@helsinki.fi

Outline

Introduction

MetaPhat is an open source application to detect variants with multivariate associations by using summary statistics from univariate genome-wide association studies. The application also performs decomposition by finding statistically optimal subsets and driver traits for the multivariate associations. The results are visualized to gather novel insight and interpretation about multivariate associations. The code is written in Python 2.7+ (syntax consistent with Python 3) and R 3.4+ and is available at https://sourceforge.net/projects/meta-pheno-association-tracer/, and also includes installation instructions and test input data, described further below.

A high-level flow chart is shown in Figure S1. The component boxes outline the user input, quality control (QC) and trait correlation estimation, genome-wide testing, variant decomposition through iterative exclusion of traits and finally plotting. MetaPhat parameters and their descriptions are listed further below in the Program arguments section. Genome-wide significant variants are detected via integration of metaCCA implementation (Cichonska et al. 2016; https://github.com/aalto-ics-kepaco/metaCCA), that does multivariate testing based on canonical correlation analysis (Hotelling 1936). Our manuscript presents results for 21 heritable lipid species containing polyunsaturated fatty acids (Table S1) using their univariate GWAS summary statistics from Tabassum et al. 2018. The details of the decomposition procedure used for finding the optimal subset and driver traits for the identified variants including the *APOE* variant (rs7412), a known lipid-associated variant (Willer et al. 2013), are presented in Data example.

Work flow

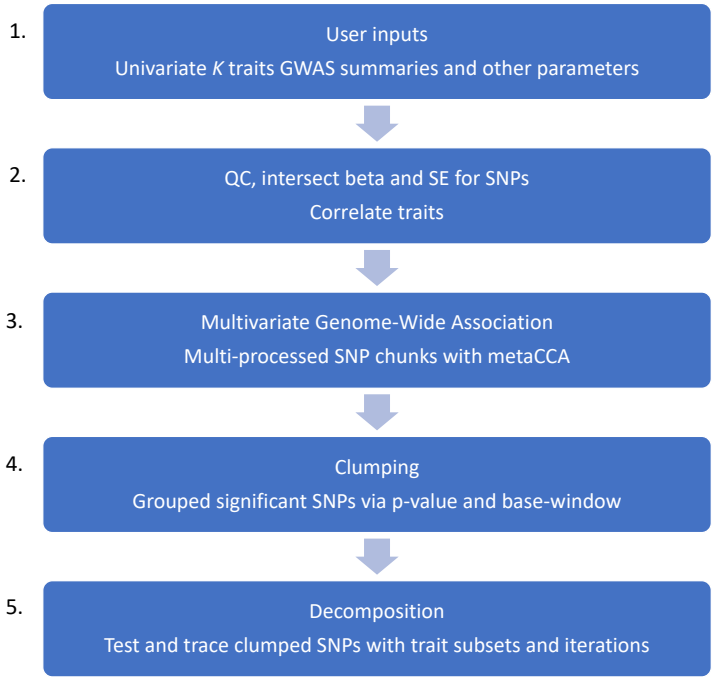

Figure S1: Workflow. 1. MetaPhat requires an input of a summary file defining multiple univariate GWAS summaries. 2. Quality control and trait correlation estimates are done. 3. Genome-wide association tests are performed on the full model with all traits included using metaCCA. The variants are divided into chunks and the association tests are performed in parallel using multiple processors, as defined in the inputs. 4. Significant variants are detected and clumped based on p-value and the given base-pair window size. 5. Decomposition is performed by finding statistically optimal traits and driver traits by tracing the highest and lowest p-values on trait subsets. MetaPhat has been tested on multiple platforms and suitable for cloud computing, more details are listed in the Architectures and Performance section.

Data example

1. Heritable lipid species

We processed univariate GWAS summaries of 21 correlated lipid species with polyunsaturated fatty acids that were reported to exhibit high heritability (Tabassum et al. 2018). These summaries were generated using lipidomics data from 2,045 Finnish subjects with imputed genotypes available at ~8.5 million SNPs. The full lipid names and fatty acid chemical properties are listed in Table S1.

Table S1: Human lipid measures used in MetaPhat analysis. Polyunsaturated lipids species with acyl chains- C20:4 (14 lipids), C20:5 (3 lipids) and 22:6 (4 lipids) are reported to exhibit high heritability (Tabassum et al. 2018). As these lipids also demonstrate considerable correlations among themselves (shown in Figure S2), they are suitable for MetaPhat multivariate analysis and decomposition.

Identifiers	Lipid classes	Lipid species
CE14	Cholesteryl ester	CE(20:4;0)
CE15	Cholesteryl ester	CE(20:5;0)
CE17	Cholesteryl ester	CE(22:6;0)
LPC8	Lysophospatidylcholines	LPC(20:4;0)
LPC9	Lysophospatidylcholines	LPC(22:6;0)
LPE5	Lysophosphatidylethanolamine	LPE(20:4;0)
LPE6	Lysophosphatidylethanolamine	LPE(22:6;0)
PC17	Phosphatidylcholine	PC(16:0;0-20:4;0)
PC18	Phosphatidylcholine	PC(16:0;0-20:5;0)
PC29	Phosphatidylcholine	PC(17:0;0-20:4;0)
PC36	Phosphatidylcholine	PC(18:0;0-20:4;0)
PC37	Phosphatidylcholine	PC(18:0;0-20:5;0)
PC46	Phosphatidylcholine	PC(18:1;0-20:4;0)
PC21	Phosphatidylcholine	PC(16:0;0-22:6;0)
PCO7	Phosphatidylcholine-ether	PC-O(16:0;0-20:4;0)
PCO23	Phosphatidylcholine-ether	PC-O(18:0;0-20:4;0)
PCO29	Phosphatidylcholine-ether	PC-O(18:1;0-20:4;0)
PE7	Phosphatidylethanolamine	PE(18:0;0-20:4;0)
PEO3	Phosphatidylethanolamine-ether	PE-O(16:1;0-20:4;0)
PEO11	Phosphatidylethanolamine-ether	PE-O(18:2;0-20:4;0)
PI9	Phosphatidylinositol	PI(18:0;0-20:4;0)

Correlations among the traits (Figure S2) are estimated via metaCCA (estimateSyy function) using the univariate beta coefficients across the traits as input. MetaPhat performs quality control to filter out variants that are not available for all traits and also automatically aligns the effects to a same allele.

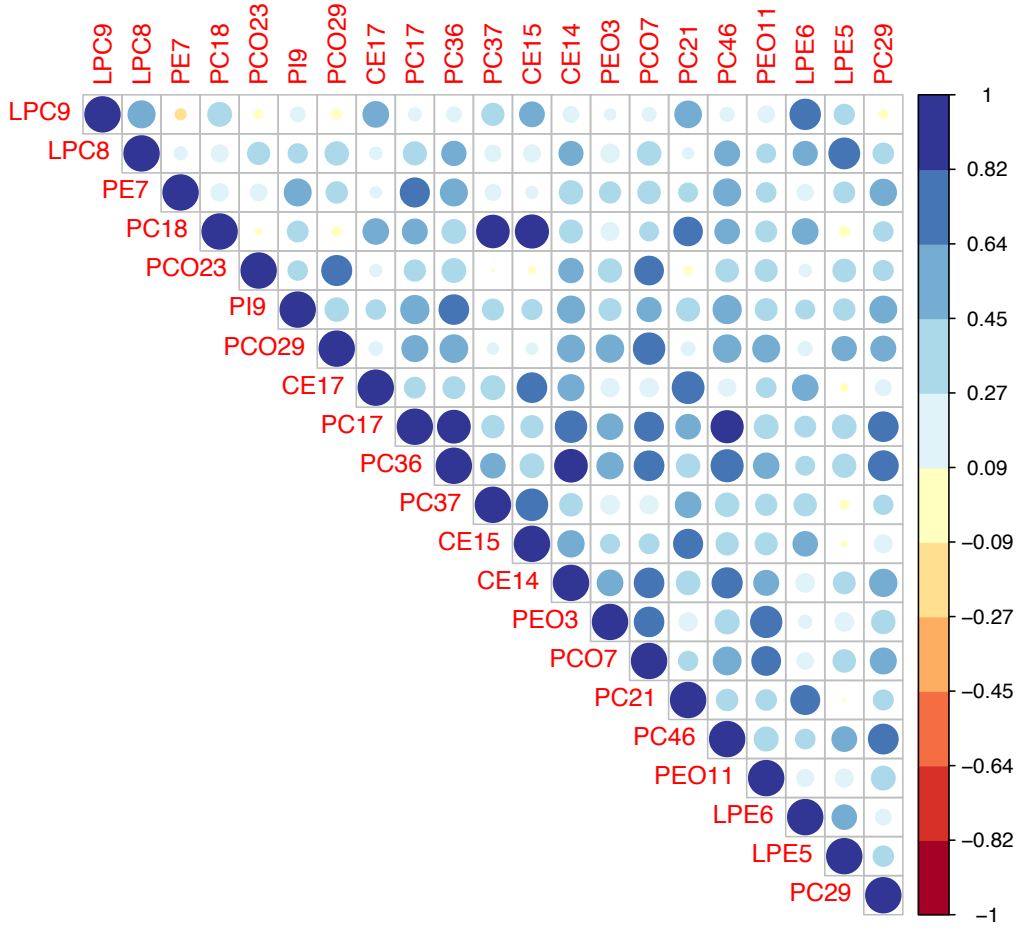

Figure S2: Correlation map of 21 heritable lipid traits computed from univariate beta coefficients using metaCCA.

MetaPhat detected significant associations at 433 SNPs ($p < 5x10^{-8}$) of which 7 independent variants, listed in Table S2, remained after clumping by a window of one million-base pairs. Clumping means that the set of independent variants was generated iteratively as follows: the variant with the smallest p-value $< 5x10^{-8}$ was added among the independent variants and all other variants within 1 Mb of the newly chosen variant were filtered out. This procedure was repeated until no genome-wide significant ($p < 5x10^{-8}$) variant remained.

Decomposition

Table S2 below lists the significant SNP associations and their decomposition results by MetaPhat using univariate GWAS summaries for the 21 lipid species (Table S1). Gene information is captured for these SNPs programmatically using an Ensemble (Hunt et al. 2018; GRCH37 and GRCH38 builds supported) resource lookup. Variants within 3,000 bases to a gene start codon site are labeled with the gene name and the exact base pair distance from the start codon (here 12:21414585:ATTTC:A_2949_SLCO1A2). Outside this 3,000 base-window and within 25K base-window, variants are labeled by the gene name and distance in kilo-bases(KB) (here 13:90477047:G:T_13KB_ENSG00000200733).

Table S2: List of 7 independent variants with their optimal and driver traits that are identified by MetaPhat in the multivariate analyses of 21 polyunsaturated lipid species.

Variant Gene GRCH37 (RSID)	Sample missing	All traits Neg. log10(P value)	Optimal traits	Optimal Neg. log10(P value)	Drivers
11:61593005:G:A FADS2 (rs174567)	1.3%	144.62	LPC8_PE7_PI9_CE17_PC17_ PC36_PC37_CE15_CE14_PEO3_ PC21_PC46_PEO11_LPE5	148.70	PC36_CE14_PC17_LPC8_PEO11_ PEO3_LPE5_PC21_PC46_PC29_ CE15_PC37_PC18_PCO7_PCO29_ PCO23_PI9_PE7_CE17_LPC9_LPE6
11:116623213:TA:T BUD13 (rs66505542)	0.1%	7.81	LPC9_PI9_PC36	11.44	PI9
12:21414585:ATTTC :A_2949_SLCO1A2 (rs146327691)	1.2%	7.37	LPC9_PE7_LPE6_LPE5	10.26	LPE5
13:90477047:G:T 13KB_ENSG0000020 0733 (rs188167837)	1.0%	7.53	CE17_PC17_CE14_PC21	8.32	PC17
15:58678720:C:T ALDH1A2 (rs261290)	0.6%	40.60	PE7_PI9_PCO29_PC17_CE15_LPE6	46.87	PE7
19:45412079:C:T APOE (rs7412)	0	12.38	PC18_PCO23_PC36_CE14	17.82	CE14_PCO23
19:54677189:C:T MBOAT7 (rs8736)	23.6%	49.04	PE7_PI9_PC36_PCO7_PC46_LPE6	57.73	PI9

Figure S3 repeats Figure 1A from the main text and shows the result of a variant in the *APOE* gene (rs7412) with the highest (green) and lowest (orange) p-value traces, from the full set of 21 traits through the iterated subsets that exclude one trait at a time until only a single trait remains. The highest trace optimizes for the highest association statistic (smallest p-value) whereas the lowest trace optimizes for the lowest association statistic (largest p-value). By following these traces, we can decompose the set of traits into smaller subsets of traits that contribute the most to the association statistic. To assist the interpretation, we define two concepts: *optimal traits* and *driver traits*.

We define the *optimal traits* (here PC18, PC36, CE14 and PCO23) as those that remain on the highest trace when the association statistic is increasing the last time. For this variant, it can be seen that after dropping LPC8 we have remaining Traits=4 and a negative log10 p-value of 17.82, that is larger than at Traits=5 and also larger than at any value Traits<4. Thus those 4 traits form the set of *optimal traits*. Table S3 below lists the negative log10 p-values from Traits 4 down to univariate statistics. We note that the lowest univariate p-value for this *APOE* variant for all 21 lipids is about 0.0001 and hence this variant was detected at significance level $5x10^{-8}$ only by applying a multivariate model.

We define the *driver traits* (here CE14 and PCO23) as those that have been removed on the lowest trace when the p-value first becomes non-significant (> $5x10^{-8}$). Starting from the full model, (21 Traits), the negative log10 p-value drops to 8.58 after dropping CE14 (20 Traits), and then to 5.74 after dropping PCO23 (19 Traits) which is below the genome-wide significance threshold (7.3 =-log10($5x10^{-8}$)). As it requires the removal of both CE14 and PCO23 to get below this threshold, these two traits thus formed the *driver traits*. For clarity, Table S4 below lists all the models and their p-values that MetaPhat considered when Traits=21, 20 or 19.

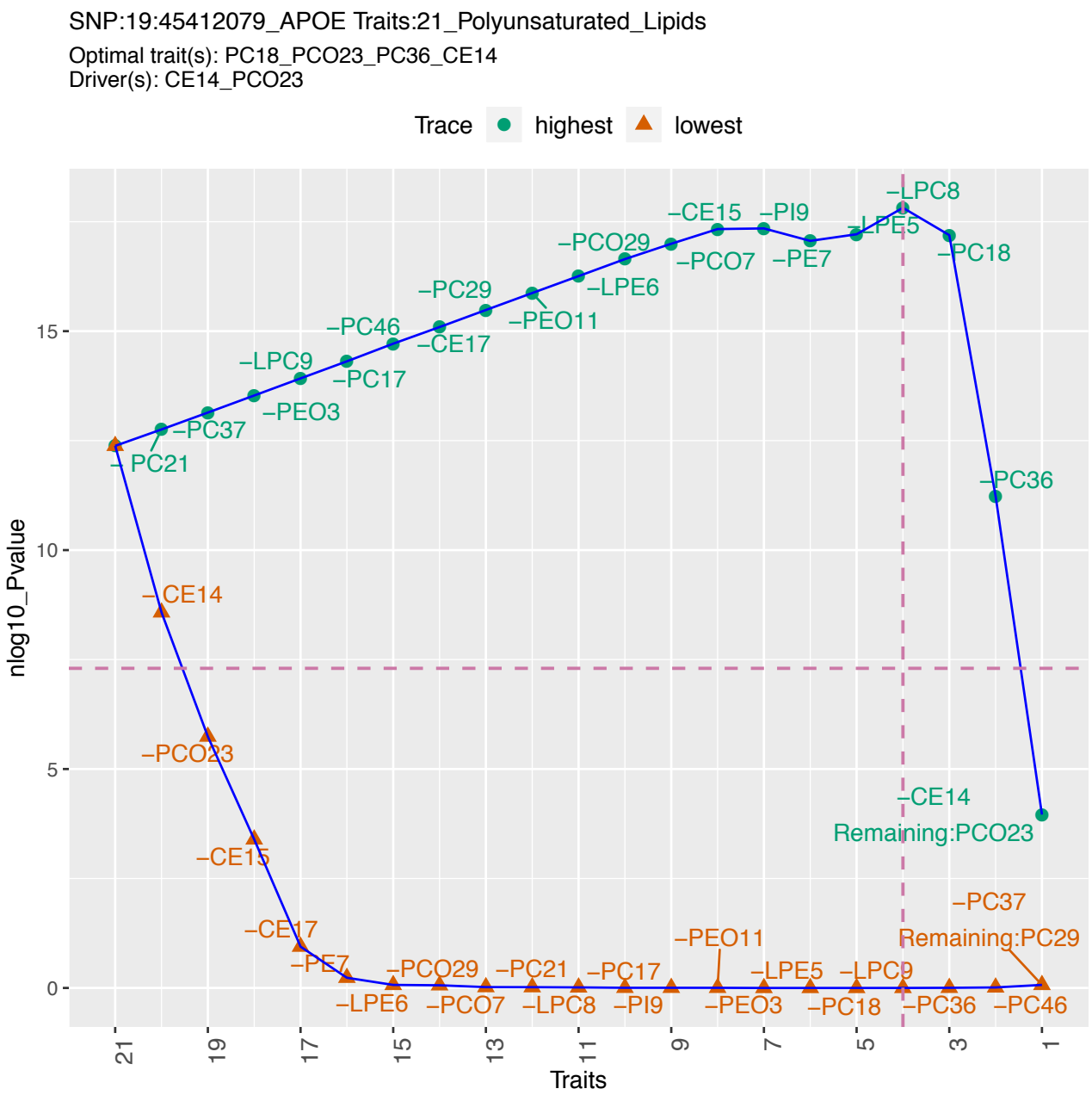

Figure S3: Trace plot of *APOE* variant rs7412 showing CE14 and PCO23 as the driver traits and LPC8, PC18, PC36, CE14 and PCO23 as the optimal traits.

Table S3: Defining optimal traits. Using the highest trace for *APOE* variant in Figure S3, MetaPhat detects PC18, PC36, CE14 and PCO23 as the optimal traits. The highest trace peaks after dropping LPC8 when 4 traits remain and begins a continuous descent, crossing the standard GWAS p-value $5x10^{-8}$ after PC36 is dropped and 2 traits PCO23 and CE14 remain. These two traits are also the driver traits defined by the lower trace (see Table S3). Table shows the association statistic (-log10 p-value) for all combinations that MetaPhat considers after there are 5 traits left. The traits on the highest trace are shown in bold at each remaining iteration.

Trait dropped	**Neg. log10(p value)**	**Traits left**
None	12.75	Full model, 21 polyunsaturated lipids
LPE5	17.21	5
LPC8	**17.82**	4 (PC18_PCO23_PC36_CE14_LPE5)
PC18	16.60	4
PC36	13.24	4
PCO23	10.59	4
CE14	3.23	4

PC18	**17.19**	3
PC36	13.60	3
PCO23	10.38	3
CE14	3.70	3
PC36	**11.23**	2
PCO23	11.04	2
CE14	3.81	2
PCO23	**3.96** (p-value 0.0001 in univariate GWAS)	1
CE14	3.36 (p-value 0.0004 in univariate GWAS)	1

Table S4: Defining driver traits. Using the lowest trace for *APOE* variant in Figure S3, MetaPhat detects CE14 and PCO23 as the driver traits. Dropping PCO23 provides the lowest -log10 p-value out of all sets with 20 Traits followed by dropping CE14 to get to 19 traits. After these 2 traits were dropped, the p-value was no longer genome-wide significant and hence these two traits form the *driver traits*.

Trait dropped	**Neg. log10(p value)**	**Traits**
None	12.37	Full model, 21 polyunsaturated lipids
CE14	**8.58**	**20**
PCO23	8.82	20
PC36	11.01	20
LPE5	11.88	20
LPC8	12.21	20
PE7	12.45	20
PI9	12.53	20
PC18	12.62	20
LPE6	12.63	20
PCO7	12.65	20
PEO11	12.65	20
PCO29	12.66	20
CE15	12.66	20
PC29	12.70	20
PC46	12.71	20
CE17	12.71	20
PC17	12.74	20
LPC9	12.75	20
PEO3	12.75	20
PC37	12.76	20
PC21	12.76	20
PCO23	**5.74**	19
CE15	6.12	19
PC18	6.87	19
LPE5	7.97	19
CE17	8.05	19
PE7	8.08	19
PC17	8.31	19
PC36	8.43	19
PI9	8.52	19
LPC8	8.60	19
LPE6	8.71	19
PC46	8.73	19
PC21	8.76	19
PEO11	8.81	19
PCO29	8.84	19
PC29	8.87	19

PC37	8.89	19
LPC9	8.90	19
PEO3	8.90	19
PCO7	8.90	19

This example showed that (1) we needed a multivariate test to detect association of *APOE* variant with lipid species and (2) only 2-4 of the 21 traits contribute to most of the multivariate association statistic. With MetaPhat we could seamlessly carry out both of these tasks from existing univariate GWAS summary statistics.

In addition to *APOE* variant, we found 6 other significant variants after clumping (Table S2). For 5 out of the 7 reported SNPs, the driver set has only one trait. Shown in the left panel of Figure S4, *BUD13* is clearly driven by trait PI9 while *FADS2* (Figure S4 right panel) variant decomposition resulted in 18 driver traits, including 8 traits that were also forming the optimal traits (LPC8, PE7, PEO3, PC17, PC36, PC37, CE15, CE14, PI9, PC21, PC46, PEO11, LPE5). Notably, FADS2 is an essential gene for fatty acid metabolism (Lattka et al. 2010). Trait importance map of each of these 7 SNPs are shown in Figure S5 and the SNP similarity clusters are shown in Figure S6.
The detailed trace plots, cluster maps and tabular outputs are available at:
https://sourceforge.net/projects/meta-pheno-association-tracer/files/test_outputs/21_polyunsaturated_lipids.tar.gz

MetaPhat output formats and intermediate results are further described at:
https://sourceforge.net/p/meta-pheno-association-tracer/wiki/output

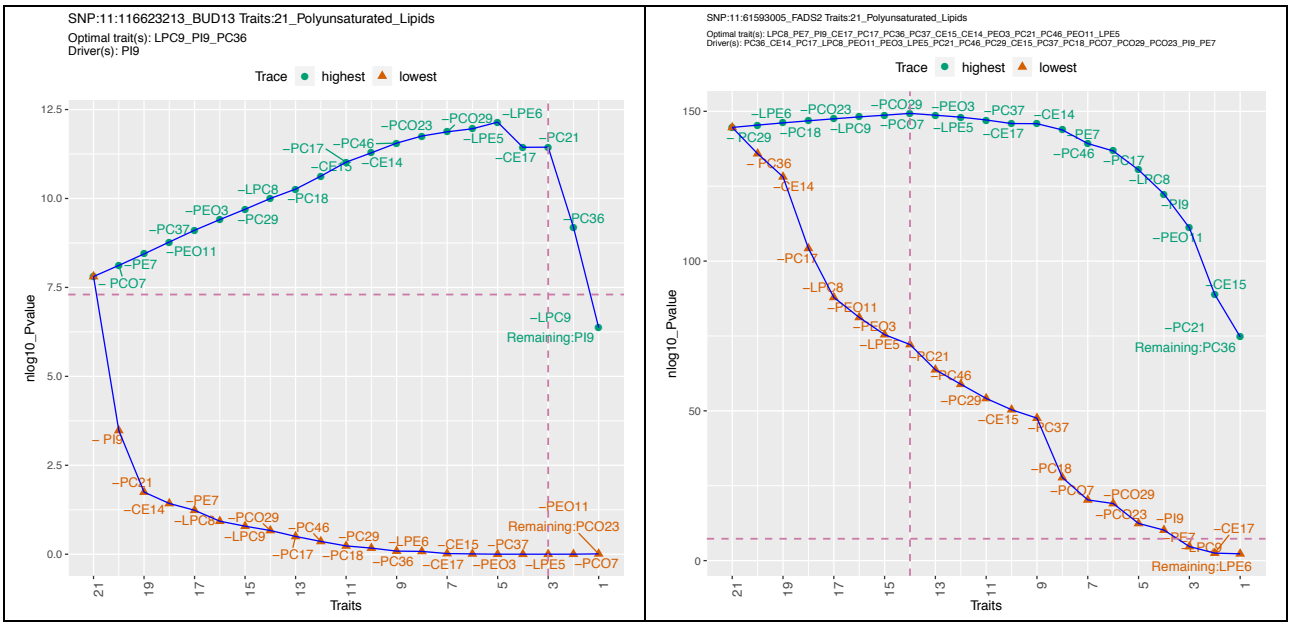

Figure S4: Decomposition of *BUD13* and *FADS2* variants. Like most of the decomposed variants (5 out of 7), *BUD13* variant association (left panel) is also driven by single trait-PI9. *FADS2* association (right panel) has the most complex signal, as there are 14 optimal traits and 18 driver traits. Notably, polyunsaturated lipids LPC8, PE7, PEO3, PC17, PC36, PC37, CE15, CE14, PI9, PC21, PC46, PEO11, LPE5 were found in both sets.

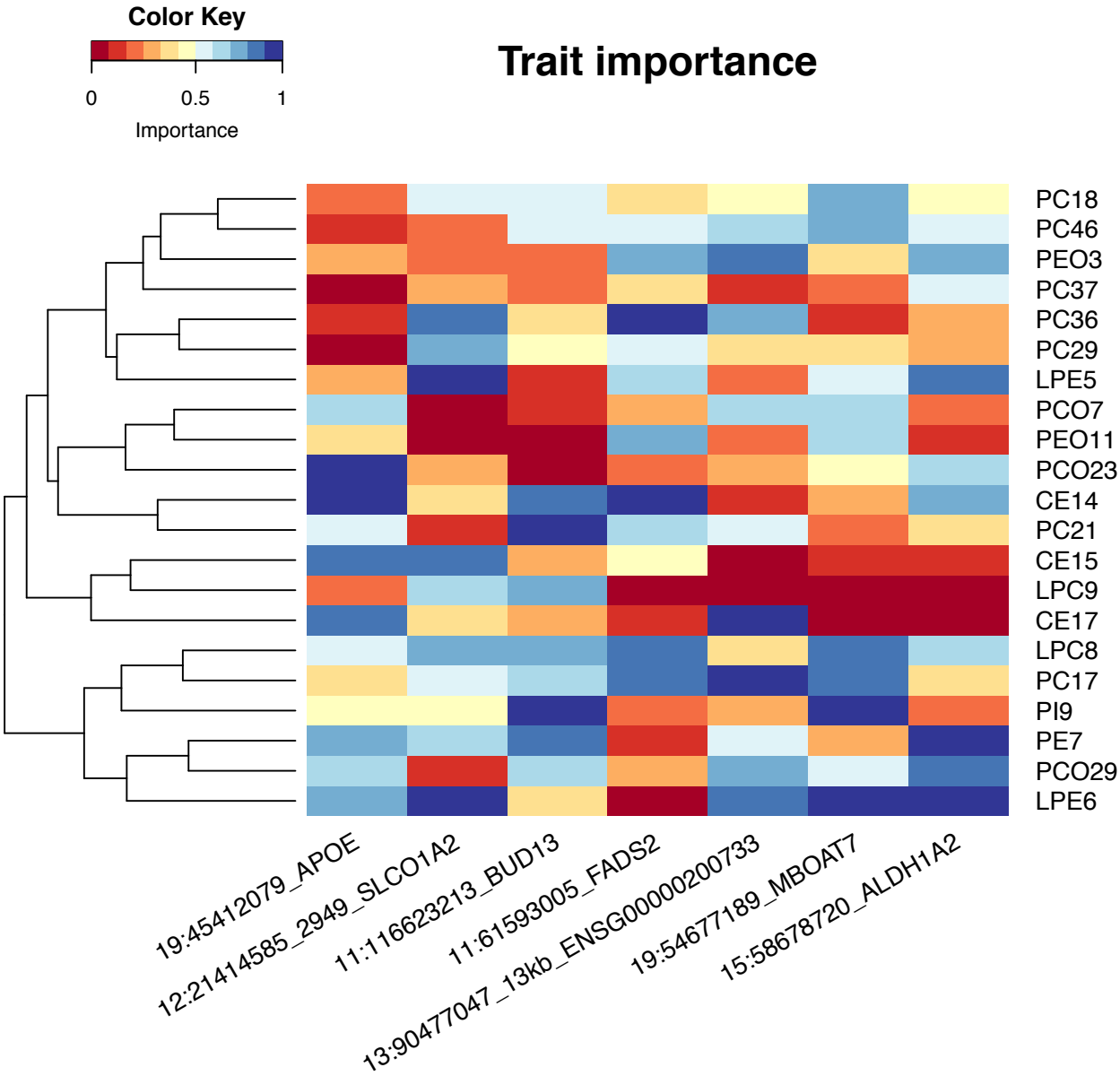

Figure S5: Trait importance map, color coded from more important (blue) to less important (red), for the 7 independent variants based on the rank of the traits on the lowest trace.

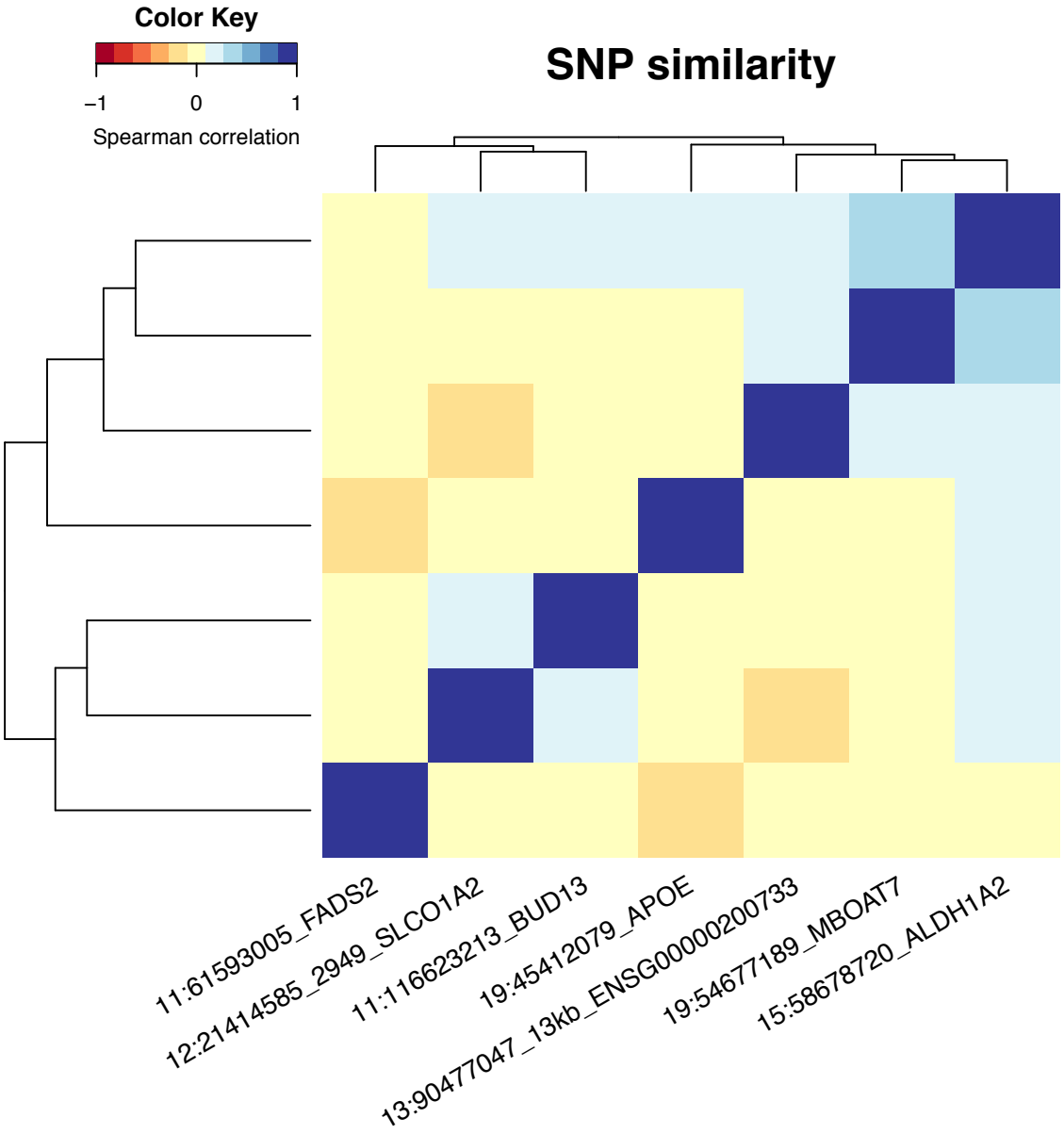

Figure S6: SNP similarities are based on clustering of trait importance in Figure S5. It can be seen that variants annotated to *ALDH1A2* (15:58678720) and *MBOAT7* (19:54677189) are positively correlated and hence could be linked through shared biological mechanisms.

Computational performance

The present analysis of 21 polyunsaturated lipids with 8,576,290 SNPs took 157 minutes on a Google cloud Linux instance with 16 cores. The program parameters and performances are listed in Table S5. For comparison and running on the same cloud instance, a job using 4 essential lipids (HDL, LDL, TC and TG) and 2,161,631 variants from Global Lipids Genetics Consortium (GLGC) (Willer et al. 2013) took 45 minutes. Willer and colleagues reported 155 significant variants, including *APOE* rs7412 reported here. We detected 2,364 variants at a p-value cutoff of 10^{-10} and these were clumped into 56 independent variants. The initial MetaPhat GLGC results, with decomposed trace plots, are available here:
https://sourceforge.net/projects/meta-pheno-association-tracer/files/test_outputs/glgc_results.tar.gz/download

Table S5: Example data sets of GWAS summaries and their processing times with MetaPhat.

Data	Phenotypes	Subjects	Intersected SNPs	Time (Minutes)	Significant loci	Clumped Significant loci
Heritable polyunsaturated lipids[1]	21	2,045	8,576,290	157	433	7
Global Lipid Consortium[2]	4 (HDL, LDL, TG, TC)	93,126	2,161,631	45	2,364	56

1. --parallel 15, --chunks 10000, --waittime 40, –clump 1000000, --cutoff 7, -log10(0.0000001), maf:range 0.01:0.5
2. --parallel 15, --chunks 10000, --waittime 40, –clump 1000000, --cutoff 10, -log10(0.00000001), maf:range 0.01:0.5

Program arguments

Table S6: The input switches of MetaPhat are shown with the --help option. Table below outlines the switches and their default values.

Switches	Description and default parameter value
--help	Shows all arguments and lists relevant default values
--phenotypes	Required, parameter is a file that defines GWAS summary files for testing (key:full_path) One summary file per line, in this format pheno1:/path/summary1.gz pheno2:/path/summmary2.gz … See example: https://sourceforge.net/projects/meta-pheno-association-tracer/files/test_inputs/global_lipid_summaries2
--outlabel	Required, defines a label for output files and plotting, must be alphanumeric and without spaces.
--nsamples	Required, the number of subjects the GWAS summaries were based on.
--nsamples_dev	Optional, defines the allowed deviating percentage of missing samples for variants to be included in analysis. It only applies if N is defined in GWAS summaries. Defaults to .25 (25%, for example, if nsamples=1000, variants with N<750 or N>1250 will be excluded)
--chunksize	Required, sets the number of SNPs processed by metaCCA on each batch. Defaults to 100000
--parallel	Required, defines the number of threads, or batch jobs, executed at the same time. The entry depends on your processor and whether it is shared. Defaults to 5.
--waittime	Required, defines the number of seconds each batch submission approximately runs. Waittime depends on the chunksize and parallel threads. In our tests we have found that this should be around 120-180 seconds for chunks of 100000 and 6-8 threads on 16Gb 16Core nonshared environment. On shared server, we recommend using higher waittime. Defaults to 180.
--r1	Optional, defines the output column from metaCCA by which to sort the variants during clumping. metaCCA outputs both r1 and pvalue. As pvalue can be set to INF for multiple variants within the same test trait set. We recommend using r1, the leading canonical correlation value. Defaults to 1.
--outdir	Required, defines the complete path to store outputs.
--gwascutoff	Required, defines the pvalue cutoff, -1(math.log(pv, 10)). Defaults to 7.3, ~5e-8 which is the standard GWAS threshold
--grch	Optional, defines the human ENSEMBL (Hunt et al. 2018) build version, possible values are 37 and 38. Defaults to 37.

--exclude	Optional, defines a file whose variants are excluded even if the variant is significant. One variant per line. See example: https://sourceforge.net/projects/meta-pheno-association-tracer/files/test_inputs/exclude.dat
--interested	Optional, defines a file whose variants are included in the output results even if variant's pvalue does not pass the cutoff. One variant per line. See example: https://sourceforge.net/projects/meta-pheno-association-tracer/files/test_inputs/interested.list
--maf_range	Optional, defines the variant frequency range min:max to include. Example values .01:.5, mean that variants with maf values >.5 and <.01 will be excluded. Only applies if all input GWAS summaries include this field, and the field column header needs to be effect_allele_frequency or maf. For possible guidelines, see UK Biobank (item 2): http://www.nealelab.is/blog/2017/9/11/details-and-considerations-of-the-uk-biobank-gwas
--clump	Required, defines the base-pair window to clump variants based on their pvalues/r1 canonical correlation values. Defaults to 500000.
--neglogval	Optional, defines the maximal negative log pvalue for INF values. Defaults to 400.
--Rscript	Required, defines the full path to Rscript executable. Defaults to /usr/bin/Rscript

GWAS summary format and headers

Univariate GWAS summary files need to follow existing EBI format standards:
https://www.ebi.ac.uk/gwas/docs/methods/summary-statistics
Additional information is provided here:
https://sourceforge.net/p/meta-pheno-association-tracer/wiki/GWAS_summary_input/
In addition, the summary files need to be in gzip format for efficiency reasons.

Prerequisites

MetaPhat requires Python 2.7 and R 3.4+. The detailed installation instructions and package dependencies are listed here:
https://sourceforge.net/p/meta-pheno-association-tracer/wiki/Install_dependencies/

Architecture and Performance

MetaPhat has been tested on shared Linux servers, MacBook laptops and Google cloud instances. The results reported have been processed from Google cloud instance with machine type n1-highmen-16. The specs are listed below:

```
Architecture:          x86_64
CPU op-mode(s):        32-bit, 64-bit
Byte Order:            Little Endian
CPU(s):                16
On-line CPU(s) list:   0-15
Thread(s) per core:    2
Core(s) per socket:    8
Socket(s):             1
NUMA node(s):          1
Vendor ID:             GenuineIntel
```

```
CPU family:            6
Model:                 85
Model name:            Intel(R) Xeon(R) CPU @ 2.00GHz
Stepping:              3
CPU MHz:               2000.182
```

More details: https://cloud.google.com/compute/docs/machine-types
For smaller data sets, the standard memory instances are suitable. We recommend machines with higher number of CPUs for optimal multiprocessing performances.

References

Cichonska A et al. (2016) metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. *Bioinformatics*. 2016;32(13):1981–1989. doi:10.1093/bioinformatics/btw052

Hunt S, McLaren W et al. (2018) Ensembl variation resources. Database Volume 2018, doi:10.1093/database/bay119

Hotelling H. (1936) Relations between two sets of variates. Biometrika, 28, 321–377.

Tabassum R et al. (2018) Genetics of human plasma lipidome: Understanding lipid metabolism and its link to diseases beyond traditional lipids
bioRxiv 457960; doi: https://doi.org/10.1101/457960

Teslovich TM, Musunuru K, Smith AV, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010;466(7307):707–713. doi:10.1038/nature09270.

Willer CJ et al. (2013) Discovery and refinement of loci associated with lipid levels. Nat. Genet. doi:10.1038/ng.2797