Title: 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[1,3,4]

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.

Summary: MetaPhat detects genetic variants with multivariate associations by using summary statistics from univariate genome-wide association studies, and performs phenotype decomposition by finding statistically optimal subsets of the traits behind each multivariate association. An intuitive trace plot of traits and a similarity measure of variants are provided to interpret multivariate associations.

Availability and Implementation: MetaPhat is implemented in Python and released under the MIT license at https://sourceforge.net/projects/meta-pheno-association-tracer/

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

Introduction

Genome-wide association studies (GWAS) of common diseases and complex traits in large population cohorts have linked many genetic variants to individual phenotypes. The statistical power in these discovery efforts can be boosted considerably by multivariate tests (Inouye et al., 2012, O'Reily et al., 2012) which have become more practical through recent implementations that require only univariate summary statistics, such as TATES (van der Sluis et al., 2013) and metaCCA (Cichonska et al., 2016). A remaining challenge is to interpret which traits are driving the multivariate association and which traits are just passengers contributing little to the association statistic. To address this question, we introduce MetaPhat (Meta-Phenotype Association Tracer), a novel software package to efficiently and systematically

1. Identify variants via multivariate GWAS from univariate summary statistics using metaCCA,
2. Trace the traits of highest and lowest importance to identify subsets of driver and optimal traits at each associated variant,
3. Cluster the variants based on the similarity of their traces.

Materials and Methods

MetaPhat is an open source application written in Python with built-in support for multi-processing, quality control, clumping and intuitive visualizations using R. As detailed in the supplement, the software is easy to use and ideal for cloud computing environments. We demonstrate MetaPhat using univariate GWAS summaries for 21 heritable and correlated polyunsaturated lipid species from 2,045 Finnish samples (Tabassum et al., 2018).

Results

We processed univariate GWAS summaries of 21 correlated lipid species at about 8.5 million SNPs. MetaPhat detected 433 significant SNPs ($p < 5x10^{-8}$) of which 7 independent variants remained after clumping by a window of one million-base pairs. Figure 1 shows the result of a variant in the *APOE* gene (rs7412) that is known to associate with low-density lipoprotein

cholesterol (Willer et al. 2013) but shows no significant univariate association to any of the 21 polyunsaturated lipid traits (smallest p-value is 1.1×10^{-4}). To accomplish an interpretable decomposition of the multivariate association, Figure 1A shows 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 absolute association statistic (smallest p-value) whereas the lowest trace optimizes for the lowest absolute association statistic (largest p-value). 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. 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. Our interpretation is that the driver traits are making the multivariate association significant whereas the optimal traits form a relatively small subset that are jointly leading to a high association statistic. Figure 1B shows the trait importance map, which is simply the ranking on the lowest trace. Figure 1C depicts SNP clustering based on the rank correlation of traits on the lowest trace. Full details and decomposed results for other variants are provided in the supplementary data.

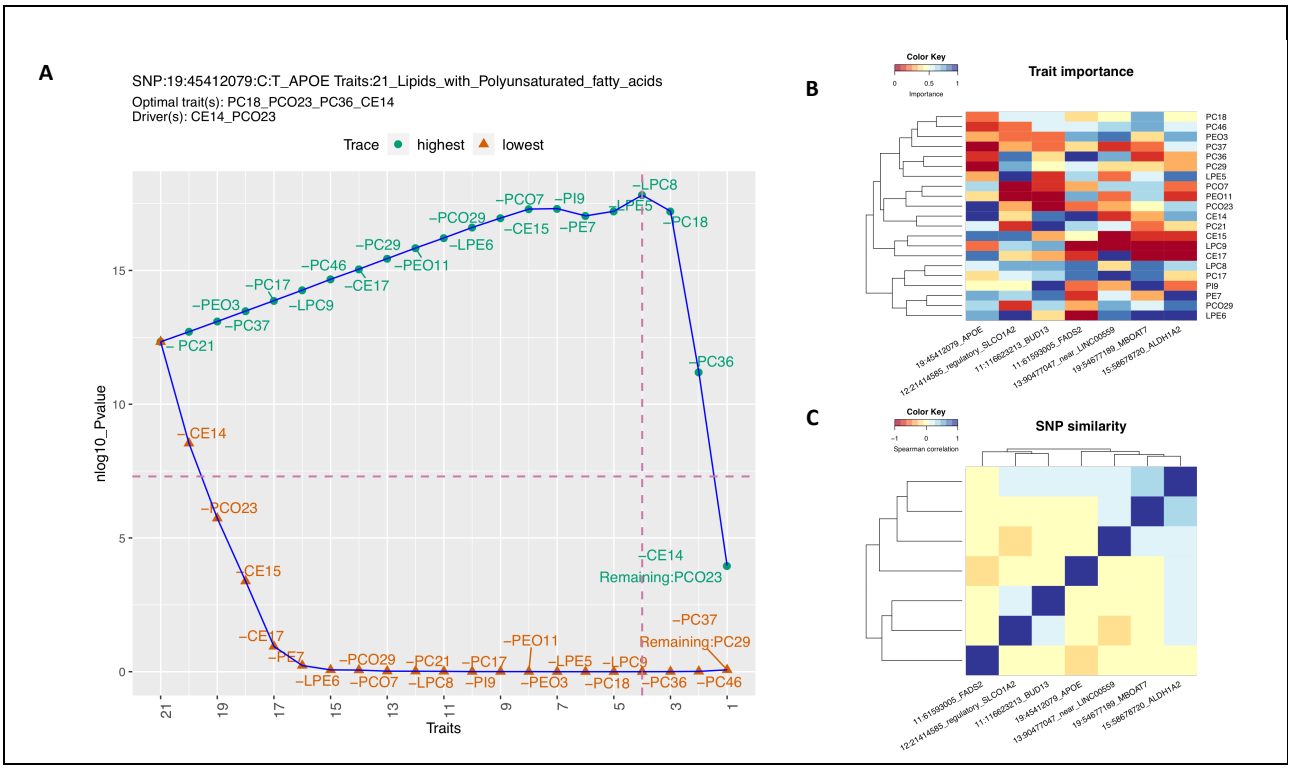

Figure 1 MetaPhat multivariate results using GWAS summaries of 21 heritable polyunsaturated lipids. A. Trace plot of *APOE* variant identifies CE14 and PCO23 as the driver traits and PC18, PC36, CE14 and PCO23 as the optimal subset. B. Trait importance map of each SNP is the ranking on the lowest trace. C. SNP similarity based on rank correlation on the lowest trace.

To summarize the results, this *APOE* SNP association would not have been identified as significant ($p < 5x10^{-8}$) from univariate GWAS of the 21 traits as the smallest univariate p-value was $1.1x10^{-4}$. On the other hand, multivariate genome scans of over 2 million possible subsets of 21 traits would not have been feasible neither computationally nor statistically. With MetaPhat we identified this SNP using a single multivariate genome-wide scan on all traits and then efficiently decomposed and traced this association to drivers and an optimal subset that contain only 2-4 central traits.

Conclusion

MetaPhat systematically detects, decomposes and visualizes statistically significant multivariate genome-phenome associations from univariate GWAS summary statistics, and brings novel interpretability to the powerful multivariate GWAS methods.

Funding

This work was supported by the Academy of Finland (grants no. 288509, 312076, 319181).

Conflicts of Interest: none declared.

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

Inouye M et al. (2012) Novel Loci for Metabolic Networks and Multi-Tissue Expression Studies Reveal Genes for Atherosclerosis. PLoS Genet 8(8): e1002907. https://doi.org/10.1371/journal.pgen.1002907

O'Reilly PF et al. (2012) MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. PLoS One 7: e34861.

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

van der Sluis S et al. (2013) TATES: Efficient Multivariate Genotype-Phenotype Analysis for Genome-Wide Association Studies. PLOS Genetics 9(1): e1003235. https://doi.org/10.1371/journal.pgen.1003235

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