

ARTICLE

<https://doi.org/10.1038/s41467-019-11954-8>

OPEN

# Genetic architecture of human plasma lipidome and its link to cardiovascular disease

Rubina Tabassum et al.<sup>#</sup>

Understanding genetic architecture of plasma lipidome could provide better insights into lipid metabolism and its link to cardiovascular diseases (CVDs). Here, we perform genome-wide association analyses of 141 lipid species ( $n = 2,181$  individuals), followed by phenome-wide scans with 25 CVD related phenotypes ( $n = 511,700$  individuals). We identify 35 lipid-species-associated loci ( $P < 5 \times 10^{-8}$ ), 10 of which associate with CVD risk including five new loci-*COL5A1*, *GLTPD2*, *SPTLC3*, *MBOAT7* and *GALNT16* (false discovery rate  $< 0.05$ ). We identify loci for lipid species that are shown to predict CVD e.g., *SPTLC3* for CER(d18:1/24:1). We show that lipoprotein lipase (LPL) may more efficiently hydrolyze medium length triacylglycerides (TAGs) than others. Polyunsaturated lipids have highest heritability and genetic correlations, suggesting considerable genetic regulation at fatty acids levels. We find low genetic correlations between traditional lipids and lipid species. Our results show that lipidomic profiles capture information beyond traditional lipids and identify genetic variants modifying lipid levels and risk of CVD.

---

Correspondence and requests for materials should be addressed to , S.R. (email: [samuli.ripatti@helsinki.fi](mailto:samuli.ripatti@helsinki.fi)). <sup>#</sup>A full list of authors and their affiliations appears at the end of the paper.

Cardiovascular diseases (CVDs) encompass many pathological conditions of impaired heart function, vascular structure and circulatory system. CVDs are the leading cause of mortality and morbidity worldwide<sup>1</sup>, necessitating the need for better preventive and predictive strategies. Plasma lipids, the well-established heritable risk factors for CVDs<sup>2</sup>, are routinely monitored to assess CVD risk. Standard lipid profiling measures traditional lipids (referred to LDL-C, HDL-C, total triglycerides and total cholesterol), but does not capture the functionally and chemically diverse molecular components—the lipid species<sup>3</sup>. These molecular lipid species may independently and specifically affect different manifestations of CVD, such as ischaemic heart disease and stroke. Lipid species including cholesterol esters (CEs), lysophosphatidylcholines (LPCs), phosphatidylcholines (PCs), phosphatidylethanolamines (PEs), ceramides (CERs), sphingomyelins (SMs) and triacylglycerols (TAGs) potentially improve CVD risk assessment over traditional lipids<sup>4–9</sup>.

Understanding of the genetic architecture and genetic regulation of these lipid species could help guide tool development for CVD risk prediction and treatment. Genetic studies of traditional lipids have identified over 250 genomic loci and improved our understanding of CVD pathophysiology<sup>10,11</sup>. For the majority of the lipid loci, however, their effects on detailed lipidome beyond traditional lipids are unknown. Only a few studies have reported genetic associations for lipid species either through studies on subsets of the lipidome<sup>12,13</sup> or GWASs on metabolome<sup>14–20</sup>.

In light of the limited information about the genetics of lipidomic profiles and their relationship with CVDs, we carried out a GWAS of lipidomic profiles of 2181 individuals using ~9.3 million genetic markers followed by PheWAS including 25 CVD-related phenotypes in up to 511,700 individuals (Fig. 1). We aimed to (1) determine heritability of lipid species and their genetic correlations; (2) identify genetic variants influencing the plasma levels of lipid species; (3) test the relationship between identified lipid-species-associated variants and CVD manifestations and (4) gain mechanistic insights into established lipid variants. We find that lipid species are heritable, suggesting a considerable role of endogenous regulation in lipid metabolism. We report association of new genomic loci with lipid species and CVD risk in humans. In addition to enhancing the current understanding of genetic regulation of circulating lipids, our

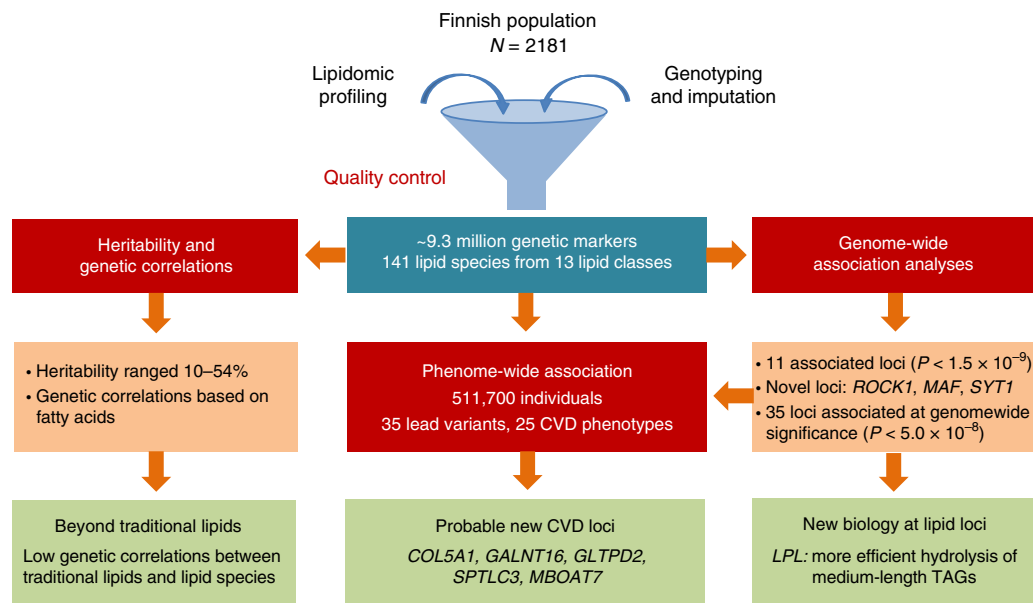
study emphasises the need of lipidomic profiling in identifying additional variants influencing lipid metabolism.

## Results

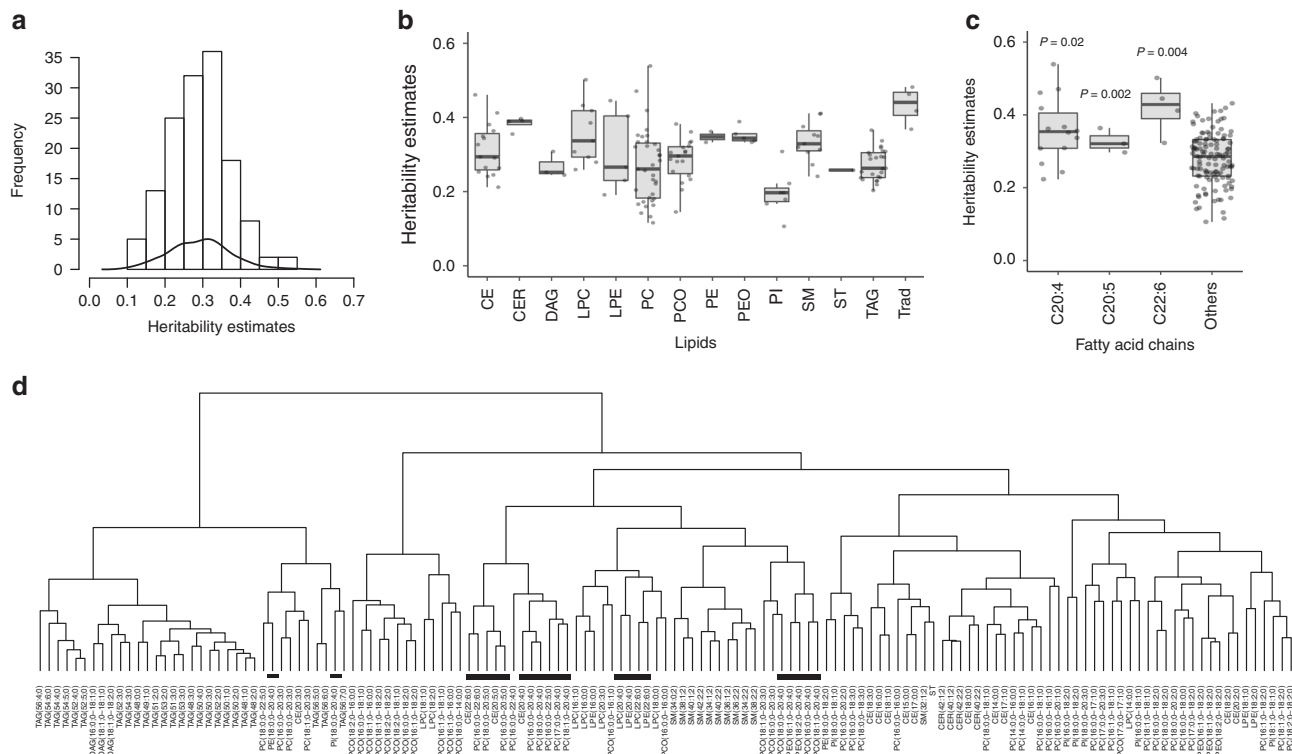
**Heritability of lipid species.** First, we determined SNP-based heritability for each of the lipid species and traditional lipids using genetic relationship matrix for all the study participants. The demographic characteristics of the study participants are provided in Supplementary Table 1. SNP-based heritability estimates ranged from 0.10 to 0.54 (Fig. 2a; Supplementary Table 2), showing considerable variation across lipid classes (Fig. 2b), with similar trends as reported previously<sup>21,22</sup>. CERs showed the greatest estimated heritability (median = 0.39, range = 0.35–0.40), whereas phosphatidylinositols (PIs) showed the least heritability (median = 0.19, range = 0.11–0.31). Sphingolipids had higher heritability than glycerolipids ranging from 0.24 to 0.41 (Fig. 2b), which is similar to a previous study that reported higher heritability for sphingolipids ranging from 0.28 to 0.53 estimated based on pedigrees<sup>21</sup>. Lipids containing polyunsaturated fatty acids, particularly C20:4, C20:5 and C22:6, had significantly higher heritability compared with other lipid species (Fig. 2c). For instance, PC (17:0;0–20:4;0) and LPC (22:6;0) had the highest heritability (>0.50), whereas PC (16:0;0–16:1;0) and PI (16:0;0–18:2;0) had the lowest heritability estimates (<0.12) (Supplementary Table 2).

**Genetic correlations between lipid species.** Longer, polyunsaturated lipids (those with four or more double bonds) had stronger genetic correlations with each other than with other lipid species (Supplementary Fig. 1, Supplementary Data 1). This can be seen in the hierarchical clustering based on genetic correlations that segregate TAG subspecies into two clusters based on carbon content and degree of unsaturation (Fig. 2d). These patterns were not seen in phenotypic correlations that were estimated based on the plasma levels of lipid species (Supplementary Fig. 2).

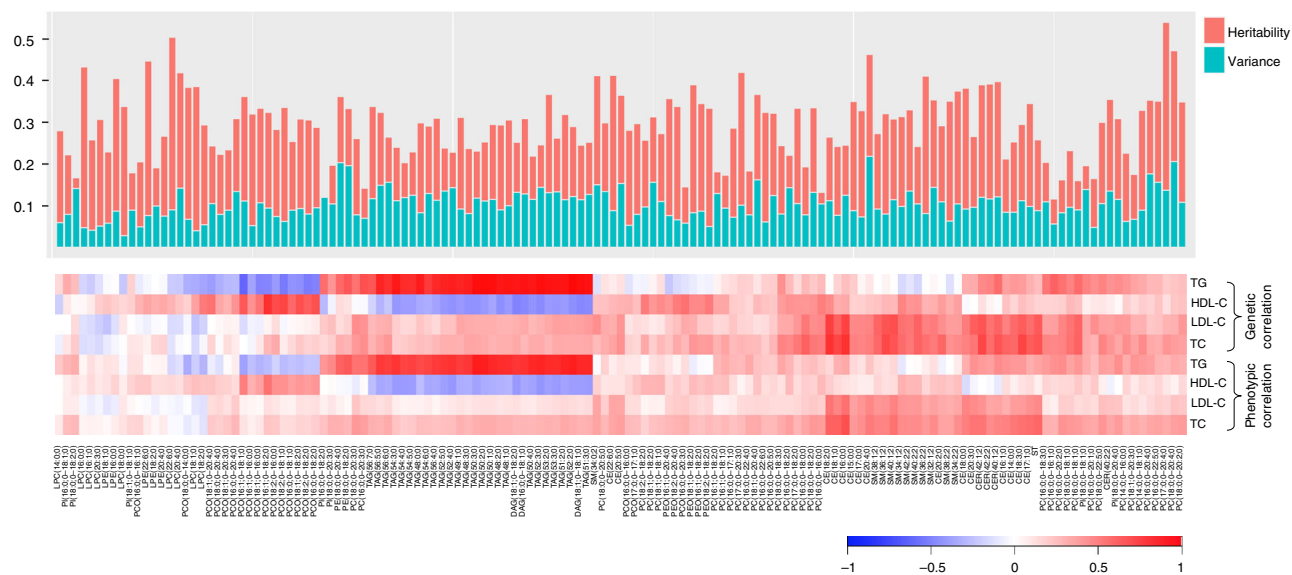
We observed low phenotypic and genetic correlation between traditional lipids and molecular lipid species, except strong positive genetic correlations of triglycerides with TAGs and DAGs (average  $r = 0.88$ ) (Fig. 3). However, triglycerides had low genetic correlation with other lipid species (average (abs)  $r = 0.26$ ). HDL-C and LDL-C levels had low genetic and phenotypic



**Fig. 1** Study design and work flow. The figure illustrates the study design and key findings of the study



**Fig. 2** Heritability of lipidomic profiles and genetic correlations among the lipid species. **a** Histogram and kernel density curve showing the distribution of heritability estimates across all the lipid species. **b** Boxplot showing the heritability estimates in each lipid class. **c** Boxplot showing comparison of the median heritability estimates of lipid species containing C20:4, C20:5 and C22:6 acyl chains and all others. The *P*-values were calculated using the Wilcoxon rank-sum test. **d** Hierarchical clustering of lipid species based on genetic correlations among lipid species. Lipids containing polyunsaturated fatty acids C20:5, C20:4 and C22:6 are highlighted with black bars. The data presented in the boxplots represent the interquartile range (IQR) defined by the bounds of the box with the median (middle line of the box) and whiskers extending to the largest/smallest values no further than 1.5 times the IQR. CER ceramide, DAG diacylglyceride, LPC lysophosphatidylcholine, LPE lysophosphatidylethanolamine, PC phosphatidylcholine, PCO phosphatidylcholine-ether, PE phosphatidylethanolamine, PEO phosphatidylethanolamine-ether, PI phosphatidylinositol, CE cholesteryl ester, SM sphingomyelin, ST sterol, TAG triacylglycerol, Trad traditional lipids



**Fig. 3** Lipidomic profiles capture information beyond traditional lipids. The genetic and phenotypic correlations between traditional lipids and molecular lipid species are shown in lower panel. The bar plot in the upper panel shows the heritability estimates of each lipid species (red bars) and the variance explained by all the known loci together (green bars). The lipid species are ordered based on the hierarchical clustering showing the correlations between the lipid species and traditional lipids. TC total cholesterol, TG triglycerides

**Table 1 Genomic loci associated with molecular lipid species at genome-wide significance**

| SNP          | Position     | Gene              | Change     | Ref | Alt | AF    | Lipid species      | Effect | SE    | P                     |
|--------------|--------------|-------------------|------------|-----|-----|-------|--------------------|--------|-------|-----------------------|
| rs201385366  | 1:897866     | <i>KLHL17</i>     | Intronic   | C   | T   | 0.019 | LPE(22:6;0)        | -0.87  | 0.16  | $3.6 \times 10^{-8}$  |
| rs187163948  | 1:14399146   | <i>KAZN</i> *     | Intronic   | G   | A   | 0.011 | TAG(53:3;0)        | 0.95   | 0.17  | $3.5 \times 10^{-8}$  |
| rs76866386#  | 2:44075483   | <i>ABCG5/8</i>    | Intronic   | T   | C   | 0.077 | CE(20:2;0)         | -0.39  | 0.06  | $3.9 \times 10^{-10}$ |
| rs58029241   | 2:98701245   | <i>VWA3B</i> *    | Intergenic | T   | A   | 0.062 | TAG(50:1;0)        | 0.37   | 0.07  | $1.9 \times 10^{-8}$  |
| rs13070110   | 3:21393248   | <i>ZNF385D</i> *  | Intergenic | T   | C   | 0.085 | Total CER          | 0.33   | 0.06  | $3.9 \times 10^{-9}$  |
| rs10212439   | 3:142655053  | <i>PAQR9</i>      | Intergenic | T   | C   | 0.602 | PI(18:0;0-18:1;0)  | 0.18   | 0.03  | $3.1 \times 10^{-8}$  |
| rs13151374   | 4:8122221    | <i>ABLIM2</i> *   | Intronic   | G   | A   | 0.153 | TAG(50:1;0)        | 0.25   | 0.04  | $3.7 \times 10^{-8}$  |
| rs186689484  | 4:97033701   | <i>PDHA2</i> *    | Intergenic | G   | A   | 0.051 | TAG(52:4;0)        | -0.40  | 0.07  | $4.2 \times 10^{-8}$  |
| rs543895501  | 6:74120350   | <i>DDX43</i> *    | Intronic   | C   | T   | 0.013 | Total LPC          | 0.87   | 0.16  | $2.9 \times 10^{-8}$  |
| rs4896307    | 6:138297840  | <i>TNFAIP3</i> *  | Intergenic | C   | T   | 0.216 | PCO(16:1;0-16:0;0) | -0.23  | 0.04  | $3.3 \times 10^{-8}$  |
| rs534693155  | 7:101081274  | <i>COL26A1</i> *  | Intronic   | A   | G   | 0.010 | LPC(16:1;0)        | 1.24   | 0.23  | $3.9 \times 10^{-8}$  |
| rs10281741   | 7:157793122  | <i>PTPRN2</i> *   | Intronic   | G   | A   | 0.225 | TAG(54:6;0)        | 0.21   | 0.04  | $2.2 \times 10^{-8}$  |
| rs1478898    | 8:11395079   | <i>BLK</i> *      | Intronic   | G   | A   | 0.440 | PC(16:0;0-16:0;0)  | 0.17   | 0.03  | $2.5 \times 10^{-8}$  |
| rs11570891   | 8:19822810   | <i>LPL</i>        | Intronic   | C   | T   | 0.075 | TAG(52:3;0)        | -0.33  | 0.06  | $2.9 \times 10^{-8}$  |
| rs146717710  | 9:137549865  | <i>COL5A1</i> *   | Intronic   | C   | T   | 0.011 | PC(16:0;0-16:1;0)  | -1.03  | 0.19  | $2.8 \times 10^{-8}$  |
| rs140645847  | 10:118863255 | <i>SHTN1</i>      | Intronic   | G   | T   | 0.101 | LPE(20:4;0)        | -0.32  | 0.06  | $3.3 \times 10^{-8}$  |
| rs28456#     | 11:61589481  | <i>FADS2</i>      | Intronic   | A   | G   | 0.405 | CE(20:4;0)         | -0.59  | 0.03  | $1.1 \times 10^{-77}$ |
| rs964184     | 11:116648917 | <i>APOA5</i>      | Intergenic | G   | C   | 0.855 | TAG(52:3;0)        | -0.258 | 0.045 | $9.5 \times 10^{-9}$  |
| rs10790495   | 11:122198706 | <i>MIR100HG</i> * | Intronic   | A   | G   | 0.590 | TAG(56:4;0)        | -0.20  | 0.04  | $2.1 \times 10^{-8}$  |
| rs117388573# | 12:78980665  | <i>SYT1</i> *     | Intergenic | A   | G   | 0.020 | LPC(14:0;0)        | -0.77  | 0.13  | $9.8 \times 10^{-10}$ |
| rs512948     | 13:52374489  | <i>DHRS12</i> *   | Intronic   | T   | C   | 0.225 | LPE(18:2;0)        | -0.22  | 0.04  | $1.4 \times 10^{-8}$  |
| rs8008070#   | 14:64233720  | <i>SYNE2</i>      | Intronic   | A   | T   | 0.132 | SM(32:1;2)         | 0.48   | 0.05  | $2.9 \times 10^{-26}$ |
| rs3902951    | 14:69789755  | <i>GALNT16</i>    | Intronic   | T   | G   | 0.361 | PEO(18:1;0-18:2;0) | 0.19   | 0.03  | $1.9 \times 10^{-8}$  |
| rs35861938   | 15:45637343  | <i>GATM</i> *     | Intergenic | T   | C   | 0.398 | PCO(18:2;0-18:1;0) | 0.18   | 0.03  | $2.7 \times 10^{-8}$  |
| rs261290#    | 15:58678720  | <i>LIPC</i>       | Intronic   | T   | C   | 0.617 | PE(18:0;0-20:4;0)  | -0.37  | 0.03  | $4.0 \times 10^{-31}$ |
| rs35221977#  | 16:79563576  | <i>MAF</i> *      | Intronic   | G   | C   | 0.054 | LPC(16:0;0)        | -0.46  | 0.08  | $1.3 \times 10^{-9}$  |
| rs79202680#  | 17:4692640   | <i>GLTPD2</i>     | Intronic   | G   | T   | 0.032 | SM(34:0;2)         | -0.85  | 0.09  | $3.4 \times 10^{-22}$ |
| rs143203352  | 17:77293933  | <i>RBFOX3</i> *   | Intronic   | T   | C   | 0.024 | PC(16:0;0-18:1;0)  | 0.60   | 0.11  | $3.2 \times 10^{-8}$  |
| rs151223356# | 18:18627427  | <i>ROCK1</i> *    | Intronic   | A   | C   | 0.013 | LPC(14:0;0)        | 0.97   | 0.15  | $1.9 \times 10^{-10}$ |
| rs7246617#   | 19:8272163   | <i>CERS4</i>      | Intergenic | G   | A   | 0.402 | SM(38:2;2)         | 0.25   | 0.03  | $2.5 \times 10^{-15}$ |
| rs2455069    | 19:51728641  | <i>CD33</i> *     | Missense   | A   | G   | 0.383 | TAG(52:5;0)        | -0.19  | 0.03  | $9.3 \times 10^{-9}$  |
| rs8736#      | 19:54677189  | <i>MBOAT7</i>     | UTR        | C   | T   | 0.388 | PI(18:0;0-20:4;0)  | -0.38  | 0.03  | $9.8 \times 10^{-28}$ |
| rs4374298    | 19:55738746  | <i>TMEM86B</i> *  | Synonymous | G   | A   | 0.166 | PEO(16:1;0-20:4;0) | -0.25  | 0.04  | $2.3 \times 10^{-8}$  |
| rs364585#    | 20:12962718  | <i>SPTLC3</i>     | Intergenic | A   | G   | 0.670 | Total CER          | -0.20  | 0.03  | $9.1 \times 10^{-10}$ |
| rs186680008  | 22:39754367  | <i>SYNGR1</i> *   | Intronic   | A   | C   | 0.015 | CE(20:3;0)         | -0.81  | 0.15  | $2.6 \times 10^{-8}$  |

Ref reference allele, Alt alternate allele, AF alternate allele frequency, SE standard error, UTR untranslated region

The strongest association between SNP and lipid species in the genome-wide significant loci ( $P < 5.0 \times 10^{-8}$ ) are presented. The P-values were calculated from the meta-analyses using the inverse variance weighted method for fixed effects. The study-wide significant associations are marked by hash symbol. The SNPs are annotated to the nearest gene if identified in this study (marked by asterisk symbol) or to previously known gene if in linkage disequilibrium with the known loci for any lipid measure. The effect sizes presented are change in standard deviation of the lipid species per alternate allele. Chromosomal positions are based on hg19 reference sequence

correlations with most of the lipid species (Fig. 3; Supplementary Data 1). Consistently, all of the known lipid variants explained 2–21% of variances in plasma levels of various lipid species, with the least variance accounting for LPCs (Fig. 3). To rule out the possibility that lipid-lowering medications resulted in the observed low genetic correlations between traditional lipids and lipid species, we also calculated the genetic correlations after excluding the individuals using lipid lowering medications ( $N = 172$ ). This re-analysis provided the similar results as the primary analysis (Supplementary Fig. 3). It is to be noted that this sample size might not provide sufficient power for heritability estimations in unrelated samples. Our study also included the family samples which provides higher statistical power in heritability estimation than unrelated samples.

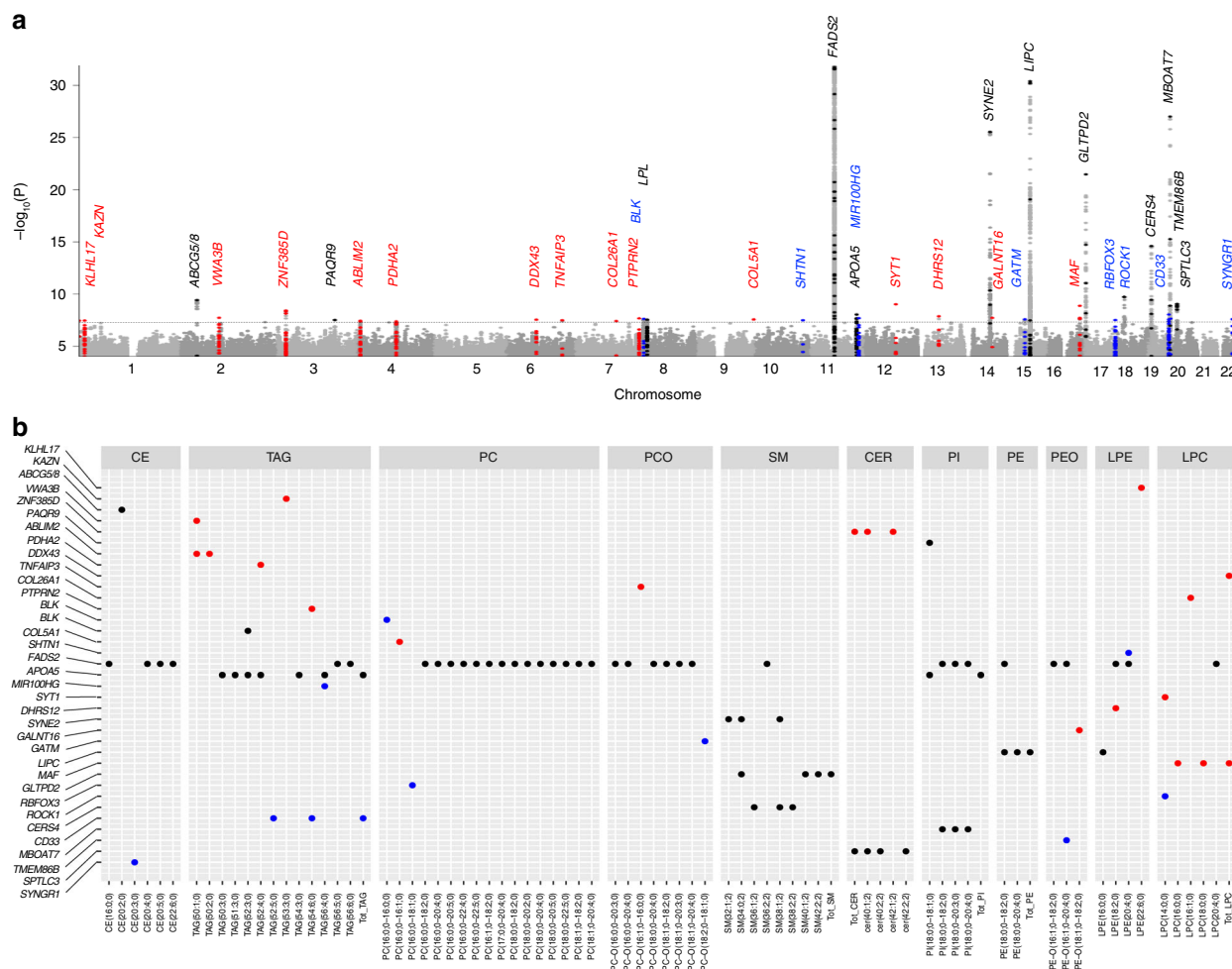
**Lipid species associated variants.** Next, we performed the genome-wide association analyses for 141 lipid species with ~9.3 million genetic markers. We identified 2817 associations between 518 variants located within 11 genomic loci (1MB blocks) and 42 lipid species from 10 lipid classes at study-wide significance ( $P < 1.5 \times 10^{-9}$  accounting for 34 principal components that explain 90% of the variance in lipidome) (Table 1; Supplementary Data 2, 3). These included three new loci (*ROCK1*, *MAF* and *SYT1*) that are not previously reported for any lipid measure or related metabolite (Fig. 4). Among the new loci, the strongest association was at an intronic variant rs151223356 near *ROCK1* with short acyl-chain LPC(14:0;0) ( $P = 1.9 \times 10^{-10}$ ). *ROCK1* encodes for a serine/threonine kinase that plays key role in glucose metabolism<sup>23</sup>. In line with our

observation of higher heritability for lipids with C20:4, C20:5 or C22:6 acyl chains, we detected associations for 15 out of 21 lipids with these acyl chains.

We also replicated the previous associations of *FADS2*, *SYNE2*, *LIPC*, *CERS4* and *MBOAT7* with the same lipid species<sup>13–20</sup>. The previously reported associations at the known loci identified in previous metabolomics GWASs are provided in Supplementary Data 4. This information was obtained from the databases-SNiPA (<http://snipa.org>) using block annotation and PhenoScanner v2 (<http://www.phenoscaner.medschl.cam.ac.uk/>), and were manually curated to include associations from literature search. In addition, we also identified new locus–lipid species associations at previously reported lipid loci including new associations of variants at *ABCG5/8* with CE (20:2;0) ( $P = 3.9 \times 10^{-10}$ ), *MBOAT7* with PI (18:0;0-20:3;0) ( $P = 3.0 \times 10^{-12}$ ) and *GLTPD2* with SM (34:0;2) ( $P = 3.4 \times 10^{-22}$ ) (Supplementary Data 2, 3).

Further, we systematically evaluated the associations of variants previously identified in metabolomics GWAS (126 variants from 46 loci available in our data set out of 132 reported) with 141 lipid species. Of these known variants, 76 variants from 12 loci showed association with 98 different lipid species with  $P < 3.2 \times 10^{-5}$  (correcting for 46 loci and 34 PCs for lipid species) (Supplementary Data 5). Of the 134 previously reported variant–lipid species pair associations that could be examined in our data set, 94 of such associations were replicated with the same direction of effect with  $P < 3.7 \times 10^{-4}$  (accounting for 134 comparisons) in our study (Supplementary Data 6).

In addition, 24 further loci were associated with at least one lipid species at regularly used genome-wide significance level ( $1.5 \times 10^{-9} > P < 5.0 \times 10^{-8}$ ). Among these additional loci, 13 loci



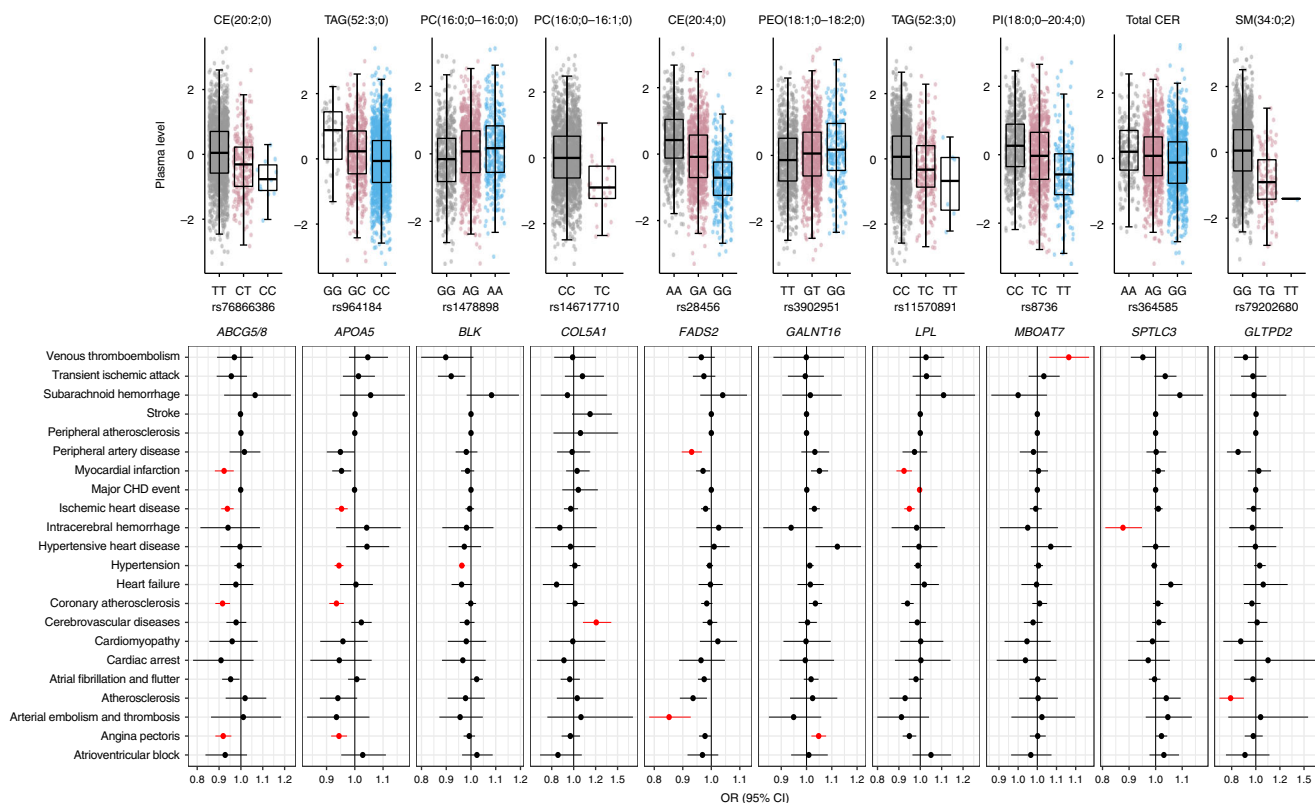
**Fig. 4** Genetic architecture of the lipidome. **a** Manhattan plot showing associations for all 141 lipid species. Only the associations with  $P < 1.0 \times 10^{-4}$  in the meta-analysis and consistent in directions in all three batches are plotted. The y-axis is capped at  $-\log_{10} P$ -value = 30 for better representation of the data. The dotted line represents the threshold for genome-wide significant associations at  $P < 5.0 \times 10^{-8}$ . **b** Genome-wide significant associations between the identified lipid species-associated loci and lipid species showing effect of the loci on the lipidome. The plotted  $P$ -values were calculated from the meta-analyses using the inverse variance weighted method for fixed effects. New hits with  $P < 5.0 \times 10^{-8}$  are shown as red dots, new independent hits in previously reported loci are presented as blue dots and hits in previously known loci are presented as black dots

were located in genomic regions not previously reported for any lipid measure or related metabolite, and 8 loci were located near known loci for lipids but were independent of any previously reported variant (Table 1; Supplementary Data 3). The regional association plots for all 35 loci with  $P < 5.0 \times 10^{-8}$  are presented in Supplementary Data 7, and the genotype–phenotype relationships for the lead variants in these 35 loci are provided in Supplementary Fig. 4.

**Relationship between identified variants and risk of CVD.** As many of the lipid species have previously been shown to predict CVD risk, we determined if the variants associated with lipid species affect individuals’ susceptibility to CVD-related phenotypes in FinnGen and UK Biobank cohorts. We identified 25 CVD-related phenotypes from the clinical outcomes derived from health registry data in the FinnGen and UK Biobanks (Supplementary Table 3). The follow-up PheWAS analyses included lead variants from all of the 35 independent loci that showed associations with  $P < 5.0 \times 10^{-8}$  (Table 1). Overall, 10 of the 35 lipid–species variants (*APOA5*, *ABCG5/8*, *BLK*, *LPL*, *FADS2*, *COL5A1*, *GALNT16*, *GLTPD2*, *MBOAT7* and *SPTLC3*) were associated with at least one of the CVD outcomes (FDR < 5%)

(Fig. 5; Supplementary Data 8). These included novel associations of variants at *COL5A1* with cerebrovascular disease ( $P = 4.6 \times 10^{-4}$ ), *GALNT16* with angina ( $P = 9.3 \times 10^{-4}$ ), *MBOAT7* with venous thromboembolism ( $P = 1.3 \times 10^{-3}$ ), *GLTPD2* with atherosclerosis ( $P = 5.3 \times 10^{-4}$ ) and *SPTLC3* with intracerebral haemorrhage ( $P = 1.0 \times 10^{-3}$ ) (Fig. 5). *FADS1-2-3* is a well-known lipid modifying locus; however, like many other known lipid loci, its effects on CVD risk has been unclear. We found an association of *FADS2* rs28456-G with peripheral artery disease ( $P = 2.2 \times 10^{-4}$ ) and arterial embolism and thrombosis ( $P = 2.5 \times 10^{-4}$ ). *BLK* (rs1478898-A) was also found to be associated with decreased risk of obesity (OR = 0.97,  $P = 5.6 \times 10^{-8}$ ) and type 2 diabetes (OR = 0.96,  $P = 4.5 \times 10^{-5}$ ).

Several studies have suggested a role for sphingolipids, including CERs and SMs, in the pathogenesis of CVDs. CER (d18:1/24:0) and CER (d18:1/24:1) have been reported to be associated with the increased risk of CVD events<sup>9</sup>. We found that the CER (d18:1/24:1) decreasing variant *SPTLC3* rs364585-G was associated with decreased risk of intracerebral haemorrhage, while CER (d18:1/24:0) increasing variant *ZNF385D* rs13070110-C was nominally associated with increased risk of intracerebral haemorrhage. Furthermore, consistent with the observation



**Fig. 5** Relationship between lipid species-associated variants and risk of CVDs. The upper panel shows the association of the identified variants with the strongest associated lipid species. Boxplots show the interquartile range (IQR) defined by the bounds of the box with the median (middle line of the box) of plasma levels of the respective lipid species for each genotype of the variants; whiskers extend to the largest/smallest values no further than 1.5 times the IQR. The lower panel depicts the relationship between the identified variants with CVD phenotypes. The effect sizes (odds ratio) with 95% confidence interval are plotted with respect to the alternate alleles. The associations with CVD phenotypes highlighted in red colour are significant at FDR < 0.05

that elevated plasma SMs levels are atherogenic<sup>24</sup>, we identified association of *GLTPD2* rs79202680-T (associated with reduced levels of SMs) with reduced risk of atherosclerosis.

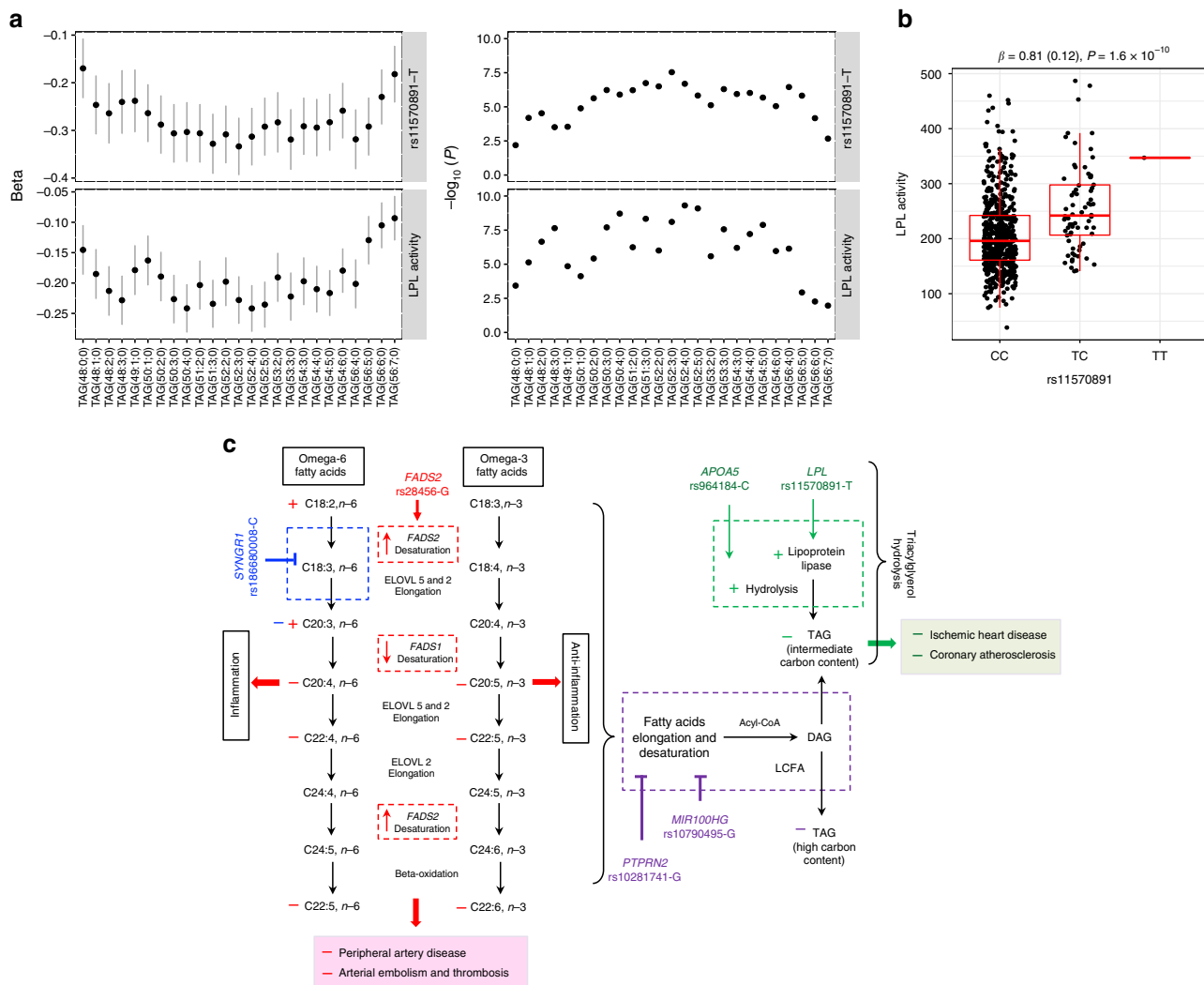
**Mechanistic insights into lipid variants.** Next, we determined if the detailed lipidomic profiles could provide new mechanistic insights into the role of known lipid variants in lipid biology. We present two examples of well-established lipid variants here. First is the fatty acid desaturase (*FADS*) gene cluster that has been consistently reported to be associated with omega-3 and omega-6 fatty acids levels with inverse effects on different PUFAs. Its mechanism, however, has not been fully deciphered. Here, we found that the *FADS2* rs28456-G was associated with increased levels of lipids with a C20:3 acyl chain and decreased levels of lipids with C20:4, C20:5 and C22:6 acyl chains (Supplementary Fig. 5). The rs28456-G is also an eQTL that increases *FADS2* expression while reduces the expression of *FADS1* [GTEx v7]. These data together explain the inverse relationship of *FADS2* variants with lipids containing different polyunsaturated fatty acids (PUFAs) (Fig. 6).

Another example is lipoprotein lipase (*LPL*). *LPL* codes for lipoprotein lipase that is the master lipolytic factor of TAGs in TAG-enriched chylomicrons and VLDL particles. We found that *LPL* rs11570891-T was associated with reduced levels of medium length TAGs (C50–C56), with strongest associations with TAG (52:3;0). This suggested that *LPL* enzyme might have different efficiency in hydrolysis of TAGs of different length. We explored this possibility by evaluating (1) the effect of *LPL* rs11570891-T on *LPL* enzymatic activity and (2) the relationship between *LPL*

activity and plasma levels of TAGs of different length, using post-heparin *LPL* measured in the EUFAM cohort. We found that *LPL* rs11570891-T (an eQTL increasing *LPL* expression) was associated with increased *LPL* activity, which in turn was associated with TAG species with stronger effect on medium length TAGs than other TAGs (Fig. 6). Consistent with a previous report by Rhee et al.<sup>16</sup>, variant rs964184-C at *APOA5*, which codes for the activator that stimulates *LPL*-mediated lipolysis of TAG-rich lipoproteins and their remnants, also showed association with medium length TAGs (Fig. 6). These results provide first clues to the probable variable role of *LPL* and *APOA5* in the hydrolysis of different TAG species.

Similarly, the association patterns of some of the newly mapped loci suggested their underlying functions. For example, *SYNGR1* rs186680008-C showed strongest associations with decreased levels of lipid species with C20:3 acyl chain from different lipid classes, including CEs, PCs and PCOs (Supplementary Fig. 5), suggesting its role in PUFA metabolism (Fig. 6). *PTPRN2* rs10281741-G and *MIR100HG* rs10790495-G showed associations with reduced levels of long polyunsaturated TAG species, suggesting their role in negative regulation of either elongation and desaturation of fatty acids or incorporation of long-chain unsaturated fatty acids during TAG biosynthesis.

**Lipidomics provide higher statistical power.** As intermediate phenotypes are known to provide more statistical power, we assessed whether the lipid species could help to detect genetic associations with greater power than traditional lipids using

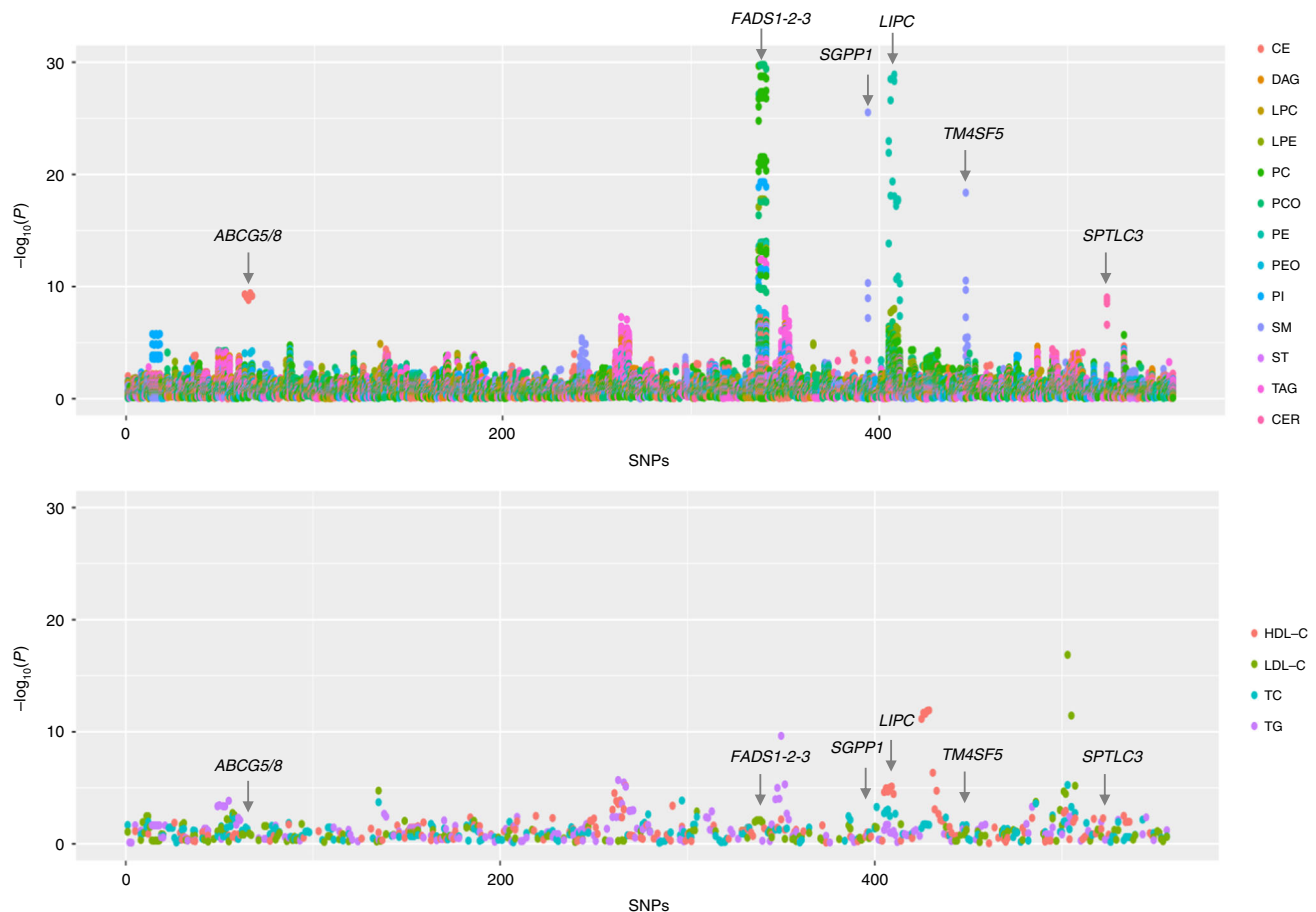


**Fig. 6** Patterns in associations and proposed mechanisms for the effect of identified variants on lipid metabolism and clinical outcomes. **a** Associations of *LPL* rs11570891-T and *LPL* activity with TAGs. Change (beta and standard errors) in plasma levels of TAGs per increase in standard deviation of *LPL* activity with their corresponding *P*-values, as calculated using linear regression model, are plotted in lower panel. The upper panel shows change (beta and standard errors) in plasma levels of TAGs per T allele with their corresponding *P*-values, as obtained from meta-analyses of genome-wide association analysis. **b** Association of *LPL* variant rs11570891 with *LPL* activity. The effect size (beta in standardised units and standard error in parenthesis) and *P*-value were calculated using linear mixed model. Boxplot depicts the interquartile range (IQR) defined by the bounds of the box, median (middle line) and whiskers extending to the largest/smallest values no further than 1.5 times the IQR. **c** Based on the patterns of the association of lipid species-associated loci with different lipid species, we propose that: (1) *LPL* rs11570891-T and *APOA5* rs964184-C might result in more efficient hydrolysis of medium length TAGs which might result in reduced CVD risk, (2) *FADS2* rs28456-G may have observed effect on PUFA metabolism through its inverse effect on *FADS2* and *FADS1* expressions, (3) *SYNGR1* rs18680008-C might have a role in the negative regulation of either desaturation of linoleic acid (C18:2,*n*-6) or elongation of gamma linolenic acid (C18:3,*n*-6). (4) *PTPRN2* rs10281741-G and *MIR100HG* rs10790495-G, which have very similar patterns of association with reduced level of long polyunsaturated TAGs, might have a role in negative regulation of either elongation and desaturation of fatty acids or incorporation of long chain unsaturated fatty acids in glycerol backbone during TAG biosynthesis. The positive (+) and negative (–) signs indicate increase or decrease, respectively, in level of lipid species or risk of disease as observed in our study, with different colours for different genetic variant

variants previously identified for traditional lipids (number of variants = 557; Supplementary Data 9). We found that molecular lipid species have much stronger associations than traditional lipids with the same sample size, except for well-known *APOE* and *CETP* (Fig. 7; Supplementary Data 10). The associations were several orders of magnitudes stronger for the variants in or near genes involved in lipid metabolism, such as *FADS1-2-3*, *LIPC*, *ABCG5/8*, *SGPPI* and *SPTLC3*. This shows that the lipidomics provides higher chances to identify lipid-modulating variants, particularly the ones with direct role in lipid metabolism, with much smaller sample size than traditional lipids.

**Discussion**

We present findings from a large-scale study that integrate lipidome, genome and phenome revealing detailed description of genetic regulation of lipidome and its associations with CVD risk. In addition to enhancing the current understanding of genetic determinants of circulating lipids, our study highlights the potential of lipidomics in gene mapping for lipids and CVDs over traditional lipids. The study generates a publicly available knowledgebase of genetic associations of molecular lipid species and their relationships with thousands of clinical outcomes.



**Fig. 7** Association of known variants for traditional lipids with lipid species and traditional lipids. The  $P$ -values for the associations of the lead SNPs (557 SNPs available in our data set) identified through different genome-wide or exome-wide studies of traditional lipids (HDL-C, LDL-C, TG and TC) with lipid species (upper panel) and traditional lipids (lower panel) are plotted. The  $y$ -axis in the upper panel is capped at  $-\log_{10} P$ -value = 30 for better representation of the data. The SNPs on the  $x$ -axis are serially arranged based on their chromosomal positions and as listed in the Supplementary Data 8. The points on the plots are colour coded by the lipid classes in the upper panel and traditional lipid in the lower panel. CER ceramide, DAG diacylglyceride, LPC lysophosphatidylcholine, LPE lysophosphatidylethanolamine, PC phosphatidylcholine, PCO phosphatidylcholine-ether, PE phosphatidylethanolamine, PEO phosphatidylethanolamine-ether, PI phosphatidylinositol, CE cholesteryl ester, SM sphingomyelin, ST sterol, TAG triacylglycerol, TC total cholesterol, TG triglycerides

Despite the expected influence of dietary intake on the circulatory lipids, plasma levels of lipid species are found to be heritable, suggesting considerable role of endogenous regulation in lipid metabolism. Importantly, genetic mechanisms do not seem to regulate all lipid species in a lipid class in the same way, as also observed in recent mice lipidomics studies<sup>25,26</sup>. Longer and more unsaturated lipid species from different lipid classes clearly display stronger genetic correlations. These observations are consistent with a previous study based on family pedigrees<sup>21</sup>. Our finding is important in the light of the proposed role of lipids containing PUFAs in CVDs, diabetes and other disorders<sup>27–29</sup>. Identification of genetic factors regulating these particular lipids is important for understanding the subtleties of lipid metabolism and devising preventive strategies including dietary interventions. Our study provides multiple leads in this direction by identifying 11 genomic loci (*KLHL17*, *APOA5*, *CD33*, *SHTN1*, *FADS2*, *LIPC*, *MBOAT7*, *MIR100HG*, *PTPRN2*, *PDHA2* and *TMEM86B*) associated with long, polyunsaturated lipids at genome-wide significance. Of these, *FADS2*, *APOA5*, *LPL* and *MBOAT7* variants were also associated with risk of CVDs (Fig. 5).

Further, we mapped genetic variants for lipid species from several lipid classes, including CERs, CEs, TAGs, SMs and PCs, that are shown to predict CVD risk<sup>4–9</sup>. Our PheWAS analyses

also suggested relationship between many of the mapped genetic variants and CVD outcomes. This knowledge can directly fuel studies on CVD prediction or drug target discovery. For instance, CERs and CEs have also been reported to associate with increased risk of CVD events<sup>5–9</sup>. Our study revealed three loci associated with CEs, including *FADS2* and two novel loci-*ABCG5/8* and *SYNGR1*, and two loci for CERs (*SPTLC3* and *ZNF385D*). CER species, particularly CER (d18:1/24:0) and CER (d18:1/24:1) are recently reported to be associated with the increased risk of CVD<sup>9</sup>. We identified two variants near *SPTLC3* and *ZNF385D* that modulate the plasma levels of CER (d18:1/24:1) and CER (d18:1/24:0), respectively, and risk for intracerebral haemorrhage. This information could also guide future studies to establish the causal relationship between lipid species and CVD.

The detailed lipidomic profile also provided clues towards understanding the mechanisms of effects of well-established lipid loci like *FADS2* and *LPL* on lipid metabolism and CVD risks. We show how the inverse effects of *FADS2* rs28456-G on the expression of two desaturases (*FADS2* and *FADS1*) could explain its opposite effects on lipids with different PUFAs. The delta-6 desaturation by *FADS2* generates gamma-linolenic acid and stearidonic acid that by elongation yield dihomo-gamma-linolenic acid and eicosatetraenoic acid (Fig. 6)<sup>30</sup>. Further,



delta-5 desaturation of dihomo-gamma-linolenic acid by *FADS1* generates arachidonic acid and eicosapentaenoic acid. Thus, as depicted in Fig. 6, the inverse effects of *FADS2* rs28456-G on *FADS2* and *FADS1* expressions explain its opposite effects on different PUFAs. The association of *FADS2* rs28456-G with the reduced levels of lipids containing arachidonic acid may also explain its association with reduced risk of atherosclerotic CVD outcomes—peripheral artery disease (PAD) and arterial embolism and thrombosis.

LPL and APOA5 are the key players in TAG hydrolysis. Our integrated approach suggested that their activity could be different for different TAG species with higher efficiency for medium length TAGs (C50–C56). We show that an *LPL* variant increases the LPL activity resulting in decreased levels of medium length TAGs. The association of the LPL variant with reduced susceptibility to CVD and type 2 diabetes could be mediated through the decrease in medium length TAGs (Fig. 5). This is consistent with a previous report that showed a similar pattern of association of levels of TAG species with type 2 diabetes<sup>31</sup>.

Similarly, the patterns of associations of newly mapped loci also suggested their involvement in the regulation of lipid metabolism. For example, rs10281741-G near *PTPRN2* and rs10790495-G near *MIR100HG* showed distinct association patterns with TAGs, with strongest association with long polyunsaturated TAGs. *PTPRN2* codes for protein tyrosine phosphatase receptor N2 with a possible role in pancreatic insulin secretion and development of diabetes mellitus<sup>32</sup>, while *MIR100HG* rs10790495 is an eQTL for the heat-shock protein HSPA8 that has a role in cell proliferation<sup>33</sup>. However, it is not known if *PTPRN2* and *MIR100HG* or *HSPA8* have any role in lipid metabolism.

Finally, we show that lipidomic profiles capture information beyond traditional lipids and provide an opportunity to identify additional genetic variants influencing lipid metabolism and disease risk. Previously, Petersen et al. showed that lipoprotein subfractions correlate with traditional lipids and strengthen genetic associations at known lipid loci and that these loci explain more of the variance of lipoprotein subfractions than of serum lipids<sup>34</sup>. Similarly, our study demonstrates that molecular lipid species have stronger statistical power compared with traditional lipids at known lipid loci using the same sample size. However, in contrast to Petersen et al., we found that many of the lipid species, including LPCs and PCs that have previously been associated with incident coronary heart disease risk<sup>4–6</sup>, have low phenotypic and genotypic correlations with traditional lipids. We also show that the known lipid variants for traditional lipids explain less of the variance of lipid species than traditional lipids. Altogether, as expected these results suggest that lipidomic profiles could provide novel information that could not be captured by traditional lipids and lipoprotein measurements.

Our study had some potential limitations. Though our study represents one of the largest genetic screen of lipidomic variation, larger cohorts are needed to achieve its full understanding. Blood samples for the EUFAM cohort were drawn after an overnight fast whereas the FINRISK cohort samples had varied fasting duration. This, however, does not seem to have substantial effect on the results and their interpretation as shown in Supplementary Data 11 and Supplementary Fig. 6. Moreover, a recent study by Rämö et al. also demonstrated similar lipidomic profiles for dyslipidemias from the EUFAM and FINRISK cohorts<sup>35</sup>. The UK Biobank cohort is reported to have a “healthy volunteer” effect<sup>36</sup>, which may affect the PheWAS results, however, given the large sample size, this is unlikely to have a substantial effect on genetic association analyses. Furthermore, lipidomic profiles were measured in whole plasma, which does not provide information at the level of individual lipoprotein subclasses and limits our ability to gain detailed mechanistic insights. We also excluded poorly

detected lipid species to ensure high data quality that narrowed the spectrum of lipidomic profiles. Further advances in lipidomics platforms might help to capture more comprehensive and complete lipidomic profiles, including the position of fatty acyl chains in the glycerol backbone of TAGs and glycerophospholipids and detection of sphingosine-1-P species and several other species, that would allow to overcome these limitations.

In conclusion, our study demonstrates that lipidomics enables deeper insights into the genetic regulation of lipid metabolism than clinically used lipid measures, which in turn might help guide future biomarker and drug target discovery and disease prevention.

## Methods

**Subjects and clinical measurements.** The study included participants from the following cohorts: EUFAM, FINRISK, FinnGen and UK Biobank. The EUFAM (The European Multicenter Study on Familial Dyslipidemias in Patients with Premature Coronary Heart Disease) study cohort is comprised of the Finnish familial combined hyperlipidemia families<sup>37</sup>. The families in EUFAM study were identified via probands admitted to Finnish university hospitals with a diagnosis of premature coronary heart disease. The probands had premature coronary heart disease and high levels of the total cholesterol, triglycerides, or both ( $\geq 90$ th Finnish age-specific and sex-specific population percentile), or low HDL-C levels ( $\leq 10$ th percentile). Invitation was extended to all the family members and spouses of the probands if at least one first-degree relative of the proband had high levels of the total cholesterol, triglycerides, or both. Venous blood samples were obtained from all participants after overnight fasting. Triglycerides and total cholesterol were measured by enzymatic methods using an automated Cobas Mira analyser (Hoffman-La Roche, Basel, Switzerland)<sup>37,38</sup>. HDL-C was quantified by phosphotungstic acid/magnesium chloride precipitation procedures, and LDL-C was calculated using the Friedewald formula<sup>39</sup>.

The Finnish National FINRISK study is a population-based survey conducted every 5 years since 1972, and thus far samples have been collected in 1992, 1997, 2002, 2007 and 2012<sup>40</sup>. Collections from the 1992, 1997, 2002, 2007 and 2012 surveys are stored in the National Institute for Health and Welfare (THL) Biobank. Lipidomic profiling was performed for 1142 participants that were randomly selected from the FINRISK 2012 survey (Supplementary Table 1). The participants were advised to fast for at least 4 h before the examination and to avoid heavy meals earlier during the day. Venous blood samples were obtained from all the participants and sera were separated. HDL-C, triglycerides and total cholesterol were measured with enzymatic methods (Abbott laboratories, Abbott Park, IL, USA) with Abbott Architect c8000 clinical chemistry analyser<sup>40</sup>.

The FinnGen data release 2 is composed of 102,739 Finnish participants. The phenotypes were derived from ICD codes in Finnish national hospital registries and cause-of-death registry as a part of FinnGen project. The quality of the CVD diagnoses in these registers has been validated in previous studies<sup>41–45</sup>. The UK Biobank data is comprised of >500,000 participants based in UK and aged 40–69 years, annotated for over 2000 phenotypes<sup>46</sup>. The PheWAS analyses in this study included 408,961 samples from white British participants.

**Ethics statement.** The study was conducted in accordance with the principles of the Helsinki declaration. Written informed consent was obtained from all the study participants. The study protocols were approved by the ethics committees of the participating centres (The Hospital District of Helsinki and Uusimaa Coordinating Ethics committees, approval No. 184/13/03/00/12). For the Finnish Institute of Health and Welfare (THL) driven FinnGen preparatory project (here called FinnGen), all patients and control subjects had provided informed consent for biobank research, based on the Finnish Biobank Act. Alternatively, older cohorts were based on study specific consents and later transferred to the THL Biobank after approval by Valvira, the National Supervisory Authority for Welfare and Health. Recruitment protocols followed the biobank protocols approved by Valvira. The Ethical Review Board of the Hospital District of Helsinki and Uusimaa approved the FinnGen study protocol Nr HUS/990/2017. The FinnGen preparatory project is approved by THL, approval numbers THL/2031/6.02.00/2017, amendments THL/341/6.02.00/2018, THL/2222/6.02.00/2018 and THL/283/6.02.00/2019. All DNA samples and data in this study were pseudonymized.

**Lipidomic profiling.** Mass spectrometry-based lipid analysis of 2181 participants was performed in three batches—353 and 686 EUFAM participants in two batches and 1142 FINRISK participants in third batch at Lipotype GmbH (Dresden, Germany). Samples were analysed by direct infusion in a QExactive mass spectrometer (Thermo Scientific) equipped with a TriVersa NanoMate ion source (Advion Biosciences)<sup>47</sup>. The data were analysed using in-house developed lipid identification software based on LipidXplorer<sup>48,49</sup>. Post processing and normalisation of data were performed using an in-house developed data management system. Only lipids with signal-to-noise ratio >5 and amounts at least fivefold higher than in the corresponding blank samples were considered for further

analyses. Reproducibility of the assay was assessed by the inclusion of reference plasma samples (eight reference samples for EUFAM and three reference samples for FINRISK) per 96-well plate. Median coefficient of variation was <10% across all batches. The data were corrected for batch and drift effects. Lipid species detected in <80% of the samples in any of the batches and samples ( $N = 64$ ) with low lipid contents were excluded. Among the lipid species which passed quality control, a total of 141 lipid species from 13 lipid classes (Supplementary Table 2) were detected consistently in all three batches and were included in all analysis. The total amounts of lipid classes were calculated by summing up the absolute concentrations of all lipid species belonging to each lipid class. The measured concentrations of the lipid species and calculated class total were transformed to normal distribution by rank-based inverse normal transformation.

It is to be noted that Lipotype platform used in the study detected many additional lipid species ( $N = 83$ ) that were not captured previously by other platforms. The list of the lipid species detected by different platforms and overlaps across the platforms are provided in the Supplementary Data 12 and Supplementary Fig. 7.

**Genotyping and imputation.** Genotyping for both EUFAM and FINRISK cohorts was performed using the HumanCoreExome BeadChip (Illumina Inc., San Diego, CA, USA). The genotype calls were generated together with other available data sets using zCall at the Institute for Molecular Medicine Finland (FIMM). Genotype data underwent stringent quality control (QC) before imputation that included exclusion of samples with low call rate (<95%), sex discrepancies, excess heterozygosity and non-European ancestry. Variants with low call rate (<95%) and deviation from Hardy–Weinberg Equilibrium ( $HWE P < 1 \times 10^{-6}$ ) were excluded. Imputation was performed using IMPUTE2<sup>50</sup>, which used two population-specific reference panels of 2690 high-coverage whole-genome and 5093 high-coverage whole-exome sequence data. Variants with imputation info score <0.70 were filtered out. After QC on population profiles and imputed variants, all subsequent analyses included 2045 individuals and ~9.3 million variants with  $MAF > 0.005$  that were available in both cohorts.

FinnGen samples were genotyped with Illumina and Affymetrix arrays (Thermo Fisher Scientific, Santa Clara, CA, USA). Genotype calls were made with GenCall and zCall algorithms for Illumina and AxiomGT1 algorithm for Affymetrix chip genotyping data. Genotyping data produced with previous chip platforms were lifted over to build version 38 (GRCh38/hg38) following the protocol described here: [dx.doi.org/10.17504/protocols.io.nqtddwn](https://doi.org/10.17504/protocols.io.nqtddwn). Samples with sex discrepancies, high genotype missingness (>5%), excess heterozygosity (+4SD) and non-Finnish ancestry were removed. Variants with high missingness (>2%), deviation from HWE ( $P < 1e-6$ ) and low minor allele count ( $MAC < 3$ ) were removed. Pre-phasing of genotyped data was performed with Eagle 2.3.5 (<https://data.broadinstitute.org/alkesgroup/Eagle/>) with the default parameters, except the number of conditioning haplotypes was set to 20,000. Imputation was carried out by using the population-specific SISu v3 imputation reference panel with Beagle 4.1 (version 08Jun17.d8b, [https://faculty.washington.edu/browning/beagle/b4\\_1.html](https://faculty.washington.edu/browning/beagle/b4_1.html)) as described in the following protocol: [[dx.doi.org/10.17504/protocols.io.nmndc5e](https://doi.org/10.17504/protocols.io.nmndc5e)]. SISu v3 imputation reference panel was developed using the high-coverage (25–30x) whole-genome sequencing data generated at the Broad Institute of MIT and Harvard and at the McDonnell Genome Institute at Washington University; and jointly processed at the Broad Institute. Variant calling was produced with GATK HaplotypeCaller algorithm by following GATK best-practices for variant calling. Genotype-, sample- and variant-wise QC was applied in an iterative manner by using the Hail framework v0.1 [<https://github.com/hail-is/hail>]. The resulting high-quality WGS data for 3775 individuals were phased with Eagle 2.3.5 as described above. Post-imputation quality control involved excluding variants with INFO score < 0.7.

Genotyping for the majority of the UK Biobank participants was done using the Affymetrix UK Biobank Axiom Array, while a subset of participants was genotyped using the Affymetrix UK BiLEVE Axiom Array. Details about the quality control and imputation of UK Biobank cohort are described by Bycroft et al.<sup>51</sup>.

**Heritability estimates and genetic correlations.** For heritability and genetic correlation estimation, rank-based inverse-transformed measures of lipid species, computed separately for the EUFAM and FINRISK cohorts, were combined to increase statistical power. The residuals of inverse-transformed measures after regressing for age, sex, first ten principal components (PCs) of genetic population structure, lipid medication, hormone replacement therapy, thyroid condition and type 2 diabetes were used as phenotypes. SNP-based heritability estimates were calculated using the variance component analysis using a genetic relationship matrix (GRM) as implemented in biMM<sup>52</sup>. Only the good quality variants with missingness <10% and  $MAF > 0.005$  were used to generate the GRM. The GRM was generated using GCTA by setting the off-diagonal elements that are <0.05 to 0 as proposed by Zaitlen et al.<sup>53</sup>. This allows to estimate SNP-based heritability in family data without removing closely related individuals. The heritability estimates of lipid species in different groups were compared using Wilcoxon rank-sum test.

The genetic correlation between each pair of lipid species and between each lipid species and traditional lipids was determined using the generated GRM with bivariate linear mixed model as implemented in biMM. The correlations based on the plasma levels (termed as phenotypic correlations) between all the pairs of the

lipid species and traditional lipids were calculated using Pearson's correlation coefficient. The heatmaps and hierarchical clustering based on genetic and phenotypic correlations were generated using heatmap.2 in R. As lipid-lowering medications could affect the plasma levels of lipid species, all analyses were adjusted for the usage of lipid-lowering medications, and separate analyses were also performed after excluding individuals using lipid-lowering medications ( $N = 172$ ).

**Lipidomics GWAS.** We performed univariate association tests for 141 individual lipid species, 12 total lipid classes and 4 traditional lipid measures (HDL-C, LDL-C, total cholesterol and triglycerides), in all batches to control for possible batch effects and combined the summary statistics by meta-analysis. The association analyses for the EUFAM cohort were performed using linear mixed models, including the above-mentioned covariates as fixed effects and kinship matrix as random effect as implemented in MMM<sup>54</sup>. The kinship matrices for the GWAS analyses were computed separately for each chromosome to include the variants from the other chromosomes using directly genotyped variants with  $MAF > 0.01$  and missingness <2%. The FINRISK cohort was analysed with linear regression model adjusting for age, sex, first ten PCs, lipid medication and diabetes using SNPTEST v2.5<sup>55</sup>. Meta-analyses were performed using the inverse variance weighted method for fixed effects adjusted for genomic inflation factor in METAL<sup>56</sup>. In addition, analyses adjusting for the traditional lipids (in addition to above-mentioned covariates) were also performed for the identified variants to determine the independent effect on lipid species.

Test statistics were adjusted for  $\lambda$  values if  $> 1.0$  before meta-analyses. Genomic inflation factor ( $\lambda$ ) ranged from 0.98 to 1.19 across the batches whereas the final  $\lambda$  values for meta-analysis ranged from 0.998 to 1.045 (Supplementary Data 13). The  $P$ -values obtained from the meta-analysis were considered to determine the SNP–lipid species associations. To account for multiple tests, the study-wide  $P$ -value threshold was set at  $< 1.5 \times 10^{-9}$  after correcting for 34 principal components (PCs) that explain over 90% of the variance in lipidomic profiles. Only the associations consistent in effect direction in all three batches were considered significant. Variants were designated as new if not located within 1 Mb of any previously reported variants for lipids (any of the traditional lipids and molecular lipid species); and as independent signal in known locus if located within 1 Mb but  $r^2 < 0.20$  with the previous lead variants and confirmed by conditional analysis. Variants with the strongest association in the identified lipid species loci was identified as the lead variants, and were annotated to the nearest gene for the new loci.

**PheWAS.** We identified 25 CVD-related outcomes from the derived phenotypes in the FinnGen and UK Biobanks (Supplementary Table 3). Associations between the 35 lead variants from the identified loci and 25 selected CVD phenotypes in FinnGen cohort were obtained from the ongoing analyses as a part of the FinnGen project. The associations were tested using saddle point approximation method adjusting for age, sex and first 10 PCs as implemented in SPAtest R package<sup>57</sup>. Associations between selected binary phenotypes and 35 lead variants in UK Biobank were obtained from Zhou et al. that were tested using logistic mixed model in SAIGE with a saddle point approximation and adjusting for first four principal components, age and sex (<https://www.leelabsg.org/resources>)<sup>58</sup>. Data for four phenotypes were not available from Zhou et al. and hence were obtained from <http://www.nealelab.is/uk-biobank/>. Associations of quantitative traits were tested using linear regression models with the same covariates as mentioned above, both for Finnish and UK Biobank cohorts. Meta-analyses of both cohorts were performed using the inverse variance weighted method for fixed effects model in METAL. The  $P$ -values obtained from the meta-analyses of the two cohorts are reported for PheWAS associations. All the PheWAS associations with false discovery rate (FDR) <5% evaluated using the Benjamini–Hochberg method and consistent direction of effects were considered significant.

**Variance explained.** To determine the variance explained by the known loci for traditional lipids, we included all the lead variants with  $MAF > 0.005$  in 250 genomic loci that have previously been associated with one or more of the four traditional lipids. Of the 636 reported variants, 557 variants with  $MAF > 0.005$  (including six proxies) were available in our QC passed imputed genotype data (Supplementary Data 10). A genetic relationship matrix (GRM) based on these 557 variants was generated using GCTA that was used to determine the variance in plasma levels of all lipid species explained by the known variants using variance component analysis in biMM.

**LPL activity.** The post-heparin lipoprotein lipase (LPL) after 15 min of heparin load was measured for 630 individuals in the EUFAM cohort using the ELISA method developed by Antikainen et al.<sup>59</sup>. The measured values were transformed using rank-based inverse normal transformation. Associations between the LPL activity and plasma levels of TAGs were determined using linear regression model adjusted for age, sex, lipid medication, hormone replacement therapy, thyroid condition and type 2 diabetes. Association between the LPL variant rs11570891 and LPL activity was tested using linear mixed model adjusted for age, sex, first ten PCs of genetic population structure, lipid medication, hormone replacement therapy,

thyroid condition and type 2 diabetes as fixed effect and kinship matrix as random effect as implemented in MMM.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

### Data availability

The full lipidomics GWAS summary level data are available on the web-based database [<https://mql.fimm.fi>]. Similarly, the PheWAS summary data can be obtained through [<https://www.leelabsg.org/resources>] and [<http://www.nealelab.is/uk-biobank/>]. The data presented in the figures and other summary level data are contained within the Supplementary Files and Supplementary Data. Other data are available through the Institute for Molecular Medicine Finland Data Access Committee on reasonable request after appropriate ethical approval.

Received: 18 April 2019 Accepted: 13 August 2019

Published online: 24 September 2019

### References

- Global Burden of Disease 2016 Causes of Death Collaborators. Global, regional, and national age-specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet* **390**, 1151–1210 (2017).
- Ference, B. A. et al. Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement from the European atherosclerosis society consensus panel. *Eur. Heart J.* **38**, 2459–2472 (2017).
- Quehenberger, O. & Dennis, E. A. The human plasma lipidome. *N. Engl. J. Med.* **365**, 1812–1823 (2011).
- Alshehry, Z. H. et al. Plasma lipidomic profiles improve on traditional risk factors for the prediction of cardiovascular events in type 2 diabetes mellitus. *Circulation* **134**, 1637–1650 (2016).
- Stegemann, C. et al. Lipidomics profiling and risk of cardiovascular disease in the prospective population-based bruneck study. *Circulation* **129**, 1821–1831 (2014).
- Laaksonen, R. et al. Plasma ceramides predict cardiovascular death in patients with stable coronary artery disease and acute coronary syndromes beyond ldl-cholesterol. *Eur. Heart J.* **37**, 1967–1976 (2016).
- Havulinna, A. S. et al. Circulating ceramides predict cardiovascular outcomes in the population-based FINRISK 2002 cohort. *Arterioscler. Thromb. Vasc. Biol.* **36**, 2424–2430 (2016).
- Razquin, C. et al. Plasma lipidome patterns associated with cardiovascular risk in the PREDIMED trial: a case-cohort study. *Int. J. Cardiol.* **253**, 126–132 (2018).
- Wang, D. D. et al. Plasma ceramides, mediterranean diet, and incident cardiovascular disease in the PREDIMED Trial (Prevención con Dieta Mediterránea). *Circulation* **135**, 2028–2040 (2017).
- Surakka, I. et al. The impact of low-frequency and rare variants on lipid levels. *Nat. Genet.* **47**, 589–597 (2015).
- Liu, D. J. et al. Exome-wide association study of plasma lipids in >300,000 individuals. *Nat. Genet.* **49**, 1758–1766 (2017).
- Demirkan, A. et al. Genome-wide association study identifies novel loci associated with circulating phospho- and sphingolipid concentrations. *PLoS Genet.* **8**, e1002490 (2012).
- Hicks, A. A. et al. Genetic determinants of circulating sphingolipid concentrations in European populations. *PLoS Genet.* **5**, e1000672 (2009).
- Gieger, C. et al. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet.* **4**, e1000282 (2008).
- Kettunen, J. et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat. Commun.* **7**, 11122 (2016).
- Rhee, E. P. et al. A genome-wide association study of the human metabolome in a community-based cohort. *Cell Metab.* **18**, 130–143 (2013).
- Draisma, H. H. M. et al. Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. *Nat. Commun.* **6**, 7208 (2015).
- Illig, T. et al. A genome-wide perspective of genetic variation in human metabolism. *Nat. Genet.* **42**, 137–141 (2010).
- Shin, S. Y. et al. An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46**, 543–550 (2014).
- Long, T. et al. Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nat. Genet.* **49**, 568–578 (2017).
- Bellis, C. et al. Human plasma lipidome is pleiotropically associated with cardiovascular risk factors and death. *Circ. Cardiovasc. Genet.* **7**, 854–863 (2014).
- Frahnow, T. et al. Heritability and responses to high fat diet of plasma lipidomics in a twin study. *Sci. Rep.* **7**, 3750 (2017).
- Chun, K. H. et al. In vivo activation of ROCK1 by insulin is impaired in skeletal muscle of humans with type 2 diabetes. *Am. J. Physiol. Endocrinol. Metab.* **300**, E536–E542 (2011).
- Jiang, X. C. et al. Plasma sphingomyelin level as a risk factor for coronary artery disease. *Arterioscler. Thromb. Vasc. Biol.* **20**, 2614–2618 (2000).
- Jha, P. et al. Systems analyses reveal physiological roles and genetic regulators of liver lipid species. *Cell Syst.* **6**, 722–733 (2018). e6.
- Jha, P. et al. Genetic Regulation of plasma lipid species and their association with metabolic phenotypes. *Cell Syst.* **6**, 709–721 (2018). e6.
- Ander, B. P., Dupasquier, C. M., Prociuk, M. A. & Pierce, G. N. Polyunsaturated fatty acids and their effects on cardiovascular disease. *Exp. Clin. Cardiol.* **8**, 164–172 (2003).
- Forouhi, N. G. et al. Association of plasma phospholipid n-3 and n-6 polyunsaturated fatty acids with type 2 diabetes: The EPIC-InterAct Case-Cohort Study. *PLoS Med.* **13**, e1002094 (2016).
- Dyall, S. C. Long-chain omega-3 fatty acids and the brain: a review of the independent and shared effects of EPA, DPA and DHA. *Front Aging Neurosci.* **7**, 52 (2015).
- Saini, R. K. & Keum, Y. S. Omega-3 and omega-6 polyunsaturated fatty acids: dietary sources, metabolism, and significance—a review. *Life Sci.* **203**, 255–267 (2018).
- Rhee, E. P. et al. Lipid profiling identifies a triacylglycerol signature of insulin resistance and improves diabetes prediction in humans. *J. Clin. Invest.* **4**, 1402–1411 (2011).
- Doi, A. et al. IA-2beta, but not IA-2, is induced by ghrelin and inhibits glucose-stimulated insulin secretion. *Proc. Natl Acad. Sci. USA* **103**, 885–890 (2006).
- Rohde, M. et al. Members of the heat-shock protein 70 family promote cancer cell growth by distinct mechanisms. *Genes Dev.* **19**, 570–582 (2005).
- Petersen, A. K. et al. Genetic associations with lipoprotein subfractions provide information on their biological nature. *Hum. Mol. Genet.* **21**, 1433–1443 (2012).
- Ramo, T. J. et al. Coronary artery disease risk and lipidomic profiles are similar in hyperlipidemias with family history and population-ascertained hyperlipidemias. *J. Am. Heart Assoc.* **8**, e012415 (2019).
- Fry, A. et al. Comparison of sociodemographic and health-related characteristics of UK biobank participants with those of the general population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).
- Porkka, K. V. et al. Phenotype expression in familial combined hyperlipidemia. *Atherosclerosis* **133**, 245–253 (1997).
- Ripatti, P. et al. The contribution of gwas loci in familial dyslipidemias. *PLoS Genet.* **12**, e1006078 (2016).
- Friedewald, W. T., Levy, R. I. & Fredrickson, D. S. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin. Chem.* **18**, 499–502 (1972).
- Borodulin, K. et al. Forty-year trends in cardiovascular risk factors in Finland. *Eur. J. Public Health* **25**, 539–546 (2015).
- Pajunen, P. et al. The validity of the Finnish hospital discharge register and causes of death register data on coronary heart disease. *Eur. J. Cardiovasc. Prev. Rehabil.* **12**, 132–137 (2005).
- Tolonen, H. et al. The validation of the Finnish hospital discharge register and causes of death register data on stroke diagnoses. *Eur. J. Cardiovasc. Prev. Rehabil.* **14**, 380–385 (2007).
- Mähönen, M. et al. The validity of heart failure diagnoses obtained from administrative registers. *Eur. J. Prev. Cardiol.* **20**, 254–259 (2013).
- Mähönen, M. et al. The validity of hospital discharge register data on coronary heart disease in Finland. *Eur. J. Epidemiol.* **13**, 403–415 (1997).
- Sund, R. Quality of the Finnish hospital discharge register: a systematic review. *Scand. J. Public Health* **40**, 505–515 (2012).
- Sudlow, C. et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
- Surma, M. A. et al. An automated shotgun lipidomics platform for high throughput, comprehensive, and quantitative analysis of blood plasma intact lipids. *Eur. J. Lipid Sci. Technol.* **117**, 1540–1549 (2015).
- Herzog, R. et al. A novel informatics concept for high-throughput shotgun lipidomics based on the molecular fragmentation query language. *Genome Biol.* **12**, R8 (2011).
- Herzog, R. et al. Lipidexplorer: a software for consensual cross-platform lipidomics. *PLoS One* **7**, e29851 (2012).
- Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
- Bycroft, C. et al. The UK biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

52. Pirinen, M. et al. biMM: efficient estimation of genetic variances and covariances for cohorts with high-dimensional phenotype measurements. *Bioinformatics* **33**, 2405–2407 (2017).
53. Zaitlen, N. et al. Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet.* **9**, e1003520 (2013).
54. Pirinen, M., Donnelly, P. & Spencer, C. C. A. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *Ann. Appl. Stat.* **7**, 369–390 (2012).
55. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
56. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
57. Dey, R., Schmidt, E. M., Abecasis, G. R. & Lee, S. A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. *Am. J. Hum. Genet.* **101**, 37–49 (2017).
58. Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).
59. Antikainen, M., Suurinkeroinen, L., Jauhiainen, M., Ehnholm, C. & Taskinen, M. R. Development and evaluation of an ELISA method for the determination of lipoprotein lipase mass concentration-comparison with a commercial, one-step enzyme immunoassay. *Eur. J. Clin. Chem. Clin. Biochem* **34**, 547–553 (1996).

## Acknowledgements

We would like to thank Sari Kivikko, Hwei-Yi Shen and Ulla Tuomainen for management assistance. We thank all study participants of the study for their participation. The FINRISK and FinnGen data used for the research were obtained from THL Biobank. We thank the THL DNA laboratory for its skillful work to produce the DNA samples used in the genotyping work, which was used in this study. Part of the genotyping was performed by the Institute for Molecular Medicine Finland FIMM Technology Centre, University of Helsinki. This research has been conducted using the UK Biobank Resource with application number 22627. This work was supported by National Institutes of Health [grant HL113315 to S.R., M.R.T., N.F. and A.P.]; Finnish Foundation for Cardiovascular Research [to S.R., V.S., M.R.T., M.J. and A.P.]; Academy of Finland Center of Excellence in Complex Disease Genetics [grant 312062 to S.R.]; Academy of Finland [285380 to S.R., 288509 to M.P.]; Jane and Aatos Erkko Foundation [to M.J.]; Sigrid Jusélius Foundation [to S.R. and M.R.T.]; Horizon 2020 Research and Innovation Programme [grant 692145 to S.R.]; EU-project RESOLVE (EU 7th Framework Programme) [grant 305707 to M.R.T.]; HiLIFE Fellowship [to S.R.]; Helsinki University Central Hospital Research Funds [to M.R.T.]; Magnus Ehrnrooth Foundation [to M.J.]; Leducq Foundation [to M.R.T.]; Ida Montin Foundation [to P.R.]; MD-PhD Programme of the Faculty of Medicine, University of Helsinki [to J.T.R.]; Doctoral Programme in Population Health, University of Helsinki [to J.T.R. and P.R.]; Finnish Medical Foundation [to J.T.R.]; Emil Aaltonen Foundation [to J.T.R. and P.R.]; Biomedicum Helsinki Foundation [to J.T.R.]; Paulo Foundation [to J.T.R.]; Idman Foundation [to J.T.R.]; Veritas Foundation [to J.T.R.]; FIMM-EMBL PhD Fellowship grant [to S.H.]. The FinnGen project is funded by two grants from Business Finland (HUS 4685/31/2016 and UH 4386/31/2016) and nine industry partners (AbbVie, AstraZeneca, Biogen, Celgene, Genentech, GSK, Merck,

Pfizer and Sanofi). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the paper.

## Author contributions

R.T., S.R., M.R.T., K.S., A.P. and N.B.F. conceived and designed the study and wrote the paper; R.T. and J.T.R. performed the statistical analyses; J.T.K., M.K., J.K., T.T.J.K., A.S.H., H.L. and V.S. were involved in FinnGen phenotype definitions and PheWAS analyses; S.H., J.N. and P.P. performed the genotype imputation; M.J.G., C.K. and M.A.S. performed lipidomic profiling and processed the raw data; S.K.S. and M.P. provided critical inputs in statistical analyses; M.J. provided critical inputs in the interpretation of the data; P.R., S.S., N.M., N.O.S., M.J.D., V.S., N.B.F., A.P., M.R.T., K.S. and S.R. acquired the data. All authors read, revised and approved the paper.

## Additional information

**Supplementary Information** accompanies this paper at <https://doi.org/10.1038/s41467-019-11954-8>.

**Competing interests:** V.S. has participated in a conference trip sponsored by Novo Nordisk and received an honorarium from the same source for participating in an advisory board meeting. M.J.G. is an employee of Lipotype GmbH, C.K. is a shareholder and employee of Lipotype GmbH, K.S. is a shareholder and CEO of Lipotype GmbH. M.A.S. is a shareholder of Lipotype GmbH and an employee of Łukasiewicz Research Network-PORT Polish Center for Technology Development. The remaining authors have no relevant competing interests.

**Reprints and permission** information is available online at <http://npg.nature.com/reprintsandpermissions/>

**Peer review information** *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019

Rubina Tabassum<sup>1</sup>, Joel T. Rämö<sup>1</sup>, Pietari Ripatti<sup>1</sup>, Jukka T. Koskela<sup>1</sup>, Mitja Kurki<sup>1,2,3</sup>, Juha Karjalainen<sup>1,4,5</sup>, Priit Palta<sup>1,6</sup>, Shabbeer Hassan<sup>1</sup>, Javier Nunez-Fontarnau<sup>1</sup>, Tuomo T.J. Kiiskinen<sup>1,7</sup>, Sanni Söderlund<sup>8</sup>, Niina Matikainen<sup>8,9</sup>, Mathias J. Gerl<sup>10</sup>, Michal A. Surma<sup>10,11</sup>, Christian Klose<sup>10</sup>, Nathan O. Stitzel<sup>12,13,14</sup>, Hannele Laivuori<sup>1,15,16</sup>, Aki S. Havulinna<sup>1,7</sup>, Susan K. Service<sup>17</sup>, Veikko Salomaa<sup>7</sup>, Matti Pirinen<sup>1,18,19</sup>, FinnGen Project, Matti Jauhiainen<sup>7,20</sup>, Mark J. Daly<sup>1,4</sup>, Nelson B. Freimer<sup>17</sup>, Aarno Palotie<sup>1,4,21</sup>, Marja-Riitta Taskinen<sup>8</sup>, Kai Simons<sup>10,22</sup> & Samuli Ripatti<sup>1,4,23</sup>

<sup>1</sup>Institute for Molecular Medicine Finland, HiLIFE, University of Helsinki, Helsinki, Finland. <sup>2</sup>Program in Medical and Population Genetics and Genetic Analysis Platform, Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>3</sup>Psychiatric & Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. <sup>4</sup>Broad Institute of the Massachusetts Institute of Technology and Harvard University, Cambridge, MA, USA. <sup>5</sup>Analytic and Translational Genetics Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. <sup>6</sup>Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia. <sup>7</sup>National Institute for Health and Welfare, Helsinki, Finland. <sup>8</sup>Research Programs Unit, Diabetes & Obesity, University of Helsinki and Department of Internal Medicine, Helsinki University Hospital, Helsinki, Finland. <sup>9</sup>Endocrinology, Abdominal Center, Helsinki University Hospital, Helsinki, Finland. <sup>10</sup>Lipotype GmbH, Dresden, Germany. <sup>11</sup>Łukasiewicz Research Network-PORT Polish Center for Technology Development, Stalowicka 147 Str., 54-066 Wrocław, Poland. <sup>12</sup>Cardiovascular Division, Department of Medicine, Washington University School of Medicine, Saint Louis, MO, USA. <sup>13</sup>Department of

Genetics, Washington University School of Medicine, Saint Louis, MO, USA. <sup>14</sup>McDonnell Genome Institute, Washington University School of Medicine, Saint Louis, MO, USA. <sup>15</sup>Department of Obstetrics and Gynecology, Tampere University Hospital and Tampere University, Faculty of Medicine and Health Technology, Tampere, Finland. <sup>16</sup>Medical and Clinical Genetics, University of Helsinki and Helsinki University Hospital, Helsinki, Finland. <sup>17</sup>Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, CA, USA. <sup>18</sup>Department of Public Health, University of Helsinki, Helsinki, Finland. <sup>19</sup>Helsinki Institute for Information Technology HIIT and Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland. <sup>20</sup>Minerva Foundation Institute for Medical Research, Biomedicum, Helsinki, Finland. <sup>21</sup>Psychiatric & Neurodevelopmental Genetics Unit, Department of Psychiatry, Analytic and Translational Genetics Unit, Department of Medicine, and the Department of Neurology, Massachusetts General Hospital, Boston, MA, USA. <sup>22</sup>Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany. <sup>23</sup>Department of Public Health, Clinicum, Faculty of Medicine, University of Helsinki, Helsinki, Finland. A full list of consortium members appears at the end of the paper.

## FinnGen Project

Anu Jalanko<sup>1</sup>, Jaakko Kaprio<sup>1</sup>, Kati Donner<sup>1</sup>, Mari Kaunisto<sup>1</sup>, Nina Mars<sup>1</sup>, Alexander Dada<sup>1</sup>, Anastasia Shcherban<sup>1</sup>, Andrea Ganna<sup>1</sup>, Arto Lehisto<sup>1</sup>, Elina Kilpeläinen<sup>1</sup>, Georg Brein<sup>1</sup>, Ghazal Awaisa<sup>1</sup>, Jarmo Harju<sup>1</sup>, Kalle Pärn<sup>1</sup>, Pietro Della Briotta Parolo<sup>1</sup>, Risto Kajanne<sup>1</sup>, Susanna Lemmelä<sup>1</sup>, Timo P. Sipilä<sup>1</sup>, Tuomas Sipilä<sup>1</sup>, Ulrike Lyhs<sup>1</sup>, Vincent Llorens<sup>1</sup>, Teemu Niiranen<sup>7</sup>, Kati Kristiansson<sup>7</sup>, Lotta Männikkö<sup>7</sup>, Manuel González Jiménez<sup>7</sup>, Markus Perola<sup>7</sup>, Regis Wong<sup>7</sup>, Terhi Kilpi<sup>7</sup>, Tero Hiekkalinna<sup>7</sup>, Elina Järvensivu<sup>7</sup>, Essi Kaiharju<sup>7</sup>, Hannele Mattsson<sup>7</sup>, Markku Laukkanen<sup>7</sup>, Päivi Laiho<sup>7</sup>, Sini Lähteenmäki<sup>7</sup>, Tuuli Sistonen<sup>7</sup>, Sirpa Soini<sup>7</sup>, Adam Ziemann<sup>24</sup>, Anne Lehtonen<sup>24</sup>, Apinya Lertratanakul<sup>24</sup>, Bob Georgantas<sup>24</sup>, Bridget Riley-Gillis<sup>24</sup>, Danjuma Quarless<sup>24</sup>, Fedik Rahimov<sup>24</sup>, Graham Heap<sup>24</sup>, Howard Jacob<sup>24</sup>, Jeffrey Waring<sup>24</sup>, Justin Wade Davis<sup>24</sup>, Nizar Smaoui<sup>24</sup>, Relja Popovic<sup>24</sup>, Sahar Esmaeeli<sup>24</sup>, Jeff Waring<sup>24</sup>, Athena Matakidou<sup>25</sup>, Ben Challis<sup>25</sup>, David Close<sup>25</sup>, Slavé Petrovski<sup>25</sup>, Antti Karlsson<sup>26</sup>, Johanna Schleutker<sup>26</sup>, Kari Pulkki<sup>26</sup>, Petri Virolainen<sup>26</sup>, Lila Kallio<sup>27</sup>, Arto Mannermaa<sup>28</sup>, Sami Heikkinen<sup>28</sup>, Veli-Matti Kosma<sup>29</sup>, Chia-Yen Chen<sup>30</sup>, Heiko Runz<sup>30</sup>, Jimmy Liu<sup>30</sup>, Paola Bronson<sup>30</sup>, Sally John<sup>30</sup>, Sanni Lahdenperä<sup>30</sup>, Susan Eaton<sup>30</sup>, Wei Zhou<sup>4</sup>, Minna Hendolin<sup>31</sup>, Outi Tuovila<sup>31</sup>, Raimo Pakkanen<sup>31</sup>, Joseph Maranville<sup>32</sup>, Keith Usiskin<sup>32</sup>, Marla Hochfeld<sup>32</sup>, Robert Plenge<sup>32</sup>, Robert Yang<sup>32</sup>, Shameek Biswas<sup>32</sup>, Steven Greenberg<sup>32</sup>, Eija Laakkonen<sup>33</sup>, Juha Kononen<sup>33</sup>, Juha Paloneva<sup>33</sup>, Urho Kujala<sup>33</sup>, Teijo Kuopio<sup>34</sup>, Jari Laukkanen<sup>35</sup>, Eeva Kangasniemi<sup>36</sup>, Kimmo Savinainen<sup>36</sup>, Reijo Laaksonen<sup>36</sup>, Mikko Arvas<sup>37</sup>, Jarmo Ritari<sup>37</sup>, Jukka Partanen<sup>37</sup>, Kati Hyvärinen<sup>37</sup>, Tiina Wahlfors<sup>37</sup>, Andrew Peterson<sup>38</sup>, Danny Oh<sup>38</sup>, Diana Chang<sup>38</sup>, Edmond Teng<sup>38</sup>, Erich Strauss<sup>38</sup>, Geoff Kerchner<sup>38</sup>, Hao Chen<sup>38</sup>, Hubert Chen<sup>38</sup>, Jennifer Schutzman<sup>38</sup>, John Michon<sup>38</sup>, Julie Hunkapiller<sup>38</sup>, Mark McCarthy<sup>38</sup>, Natalie Bowers<sup>38</sup>, Tim Lu<sup>38</sup>, Tushar Bhangale<sup>38</sup>, David Pulford<sup>39</sup>, Dawn Waterworth<sup>39</sup>, Diptee Kulkarni<sup>39</sup>, Fanli Xu<sup>39</sup>, Jo Betts<sup>39</sup>, Jorge Esparza Gordillo<sup>39</sup>, Joshua Hoffman<sup>39</sup>, Kirsi Auro<sup>39</sup>, Linda McCarthy<sup>39</sup>, Soumitra Ghosh<sup>39</sup>, Meg Ehm<sup>39</sup>, Kimmo Pitkänen<sup>40</sup>, Tomi Mäkelä<sup>41</sup>, Anu Loukola<sup>42</sup>, Heikki Joensuu<sup>42</sup>, Juha Sinisalo<sup>42</sup>, Kari Eklund<sup>42</sup>, Lauri Aaltonen<sup>42</sup>, Martti Färkkilä<sup>42</sup>, Olli Carpen<sup>42</sup>, Paula Kauppi<sup>42</sup>, Pentti Tienari<sup>42</sup>, Terhi Ollila<sup>42</sup>, Tiinamaija Tuomi<sup>42</sup>, Tuomo Meretoja<sup>42</sup>, Anne Pitkäranta<sup>42</sup>, Joni Turunen<sup>42</sup>, Katariina Hannula-Jouppi<sup>42</sup>, Sampsa Pikkariainen<sup>42</sup>, Sanna Seitsonen<sup>42</sup>, Miika Koskinen<sup>42</sup>, Antti Palomäki<sup>43</sup>, Juha Rinne<sup>43</sup>, Kaj Metsärinne<sup>43</sup>, Klaus Elenius<sup>43</sup>, Laura Pirilä<sup>43</sup>, Leena Koulou<sup>43</sup>, Markku Voutilainen<sup>43</sup>, Markus Juonala<sup>43</sup>, Sirkku Peltonen<sup>43</sup>, Vesa Aaltonen<sup>43</sup>, Andrey Loboda<sup>44</sup>, Anna Podgornaia<sup>44</sup>, Aparna Chhibber<sup>44</sup>, Audrey Chu<sup>44</sup>, Caroline Fox<sup>44</sup>, Dorothee Diogo<sup>44</sup>, Emily Holzinger<sup>44</sup>, John Eicher<sup>44</sup>, Padhraig Gormley<sup>44</sup>, Vinay Mehta<sup>44</sup>, Xulong Wang<sup>44</sup>, Johannes Kettunen<sup>45</sup>, Katri Pylkäs<sup>45</sup>, Marita Kalaoja<sup>45</sup>, Minna Karjalainen<sup>45</sup>, Reetta Hinttala<sup>45</sup>, Riitta Kaarteenaho<sup>45</sup>, Seppo Vainio<sup>45</sup>, Tuomo Mantere<sup>45</sup>, Seppo Vainio<sup>46</sup>, Anne Remes<sup>47</sup>, Johanna Huhtakangas<sup>47</sup>, Juhani Junttila<sup>47</sup>, Kaisa Tasanen<sup>47</sup>, Laura Huilaja<sup>47</sup>, Marja Luodonpää<sup>47</sup>, Nina Hautala<sup>47</sup>, Peeter Karihtala<sup>47</sup>, Saila Kauppila<sup>47</sup>, Terttu Harju<sup>47</sup>, Timo Blomster<sup>47</sup>, Hilikka Soininen<sup>48</sup>, Ilkka Harvima<sup>48</sup>, Jussi Pihlajamäki<sup>48</sup>, Kai Kaarniranta<sup>48</sup>, Margit Pelkonen<sup>48</sup>, Markku Laakso<sup>48</sup>, Mikko Hiltunen<sup>48</sup>, Mikko Kiviniemi<sup>48</sup>, Oili Kaipainen-Seppänen<sup>48</sup>, Päivi Auvinen<sup>48</sup>, Reetta Kälväinen<sup>48</sup>, Valtteri Julkunen<sup>48</sup>, Anders Malarstig<sup>49</sup>, Åsa Hedman<sup>49</sup>, Catherine Marshall<sup>49</sup>, Christopher Whelan<sup>49</sup>,

Heli Lehtonen<sup>49</sup>, Jaakko Parkkinen<sup>49</sup>, Kari Linden<sup>49</sup>, Kirsi Kalpala<sup>49</sup>, Melissa Miller<sup>49</sup>, Nan Bing<sup>49</sup>, Stefan McDonough<sup>49</sup>, Xing Chen<sup>49</sup>, Xinli Hu<sup>49</sup>, Ying Wu<sup>49</sup>, Annika Auranen<sup>50</sup>, Airi Jussila<sup>50</sup>, Hannele Uusitalo-Järvinen<sup>50</sup>, Hannu Kankaanranta<sup>50</sup>, Hannu Uusitalo<sup>50</sup>, Jukka Peltola<sup>50</sup>, Mika Kähönen<sup>50</sup>, Pia Isomäki<sup>50</sup>, Tarja Laitinen<sup>50</sup>, Teea Salmi<sup>50</sup>, Anthony Muslin<sup>51</sup>, Clarence Wang<sup>51</sup>, Clement Chatelain<sup>51</sup>, Ethan Xu<sup>51</sup>, Franck Auge<sup>51</sup>, Kathy Call<sup>51</sup>, Kathy Klinger<sup>51</sup>, Marika Crohns<sup>51</sup>, Matthias Gossel<sup>51</sup>, Kimmo Palin<sup>52</sup>, Manuel Rivas<sup>53</sup>, Harri Siirtola<sup>54</sup> & Javier Gracia Tabuenca<sup>54</sup>

<sup>24</sup>Abbvie, Chicago, IL, USA. <sup>25</sup>Astra Zeneca, Cambridge, UK. <sup>26</sup>Auria Biobank, University of Turku and Hospital District of Southwest Finland, Turku, Finland. <sup>27</sup>Auria Biobank, Turku, Finland. <sup>28</sup>Biobank of Eastern Finland, University of Eastern Finland and Northern Savo Hospital District, Kuopio, Finland. <sup>29</sup>Biobank of Eastern Finland, Kuopio, Finland. <sup>30</sup>Biogen, Cambridge, MA, USA. <sup>31</sup>Business Finland, Helsinki, Finland. <sup>32</sup>Celgene, Summit, NJ, USA. <sup>33</sup>Central Finland Biobank, University of Jyväskylä and Central Finland Health Care District, Jyväskylä, Finland. <sup>34</sup>Central Finland Biobank, Jyväskylä, Finland. <sup>35</sup>Central Finland Health Care District, Jyväskylä, Finland. <sup>36</sup>Finnish Clinical Biobank Tampere, University of Tampere and Pirkanmaa Hospital District, Tampere, Finland. <sup>37</sup>Finnish Red Cross Blood Service and Finnish Hematology Registry and Clinical Biobank, Helsinki, Finland. <sup>38</sup>Genentech, San Francisco, CA, USA. <sup>39</sup>GlaxoSmithKline, Brentford, UK. <sup>40</sup>Helsinki Biobank, Helsinki, Finland. <sup>41</sup>HiLIFE, University of Helsinki, Helsinki, Finland. <sup>42</sup>Hospital District of Helsinki and Uusimaa, Helsinki, Finland. <sup>43</sup>Hospital District of Southwest Finland, Turku, Finland. <sup>44</sup>Merck, Kenilworth, NJ, USA. <sup>45</sup>Northern Finland Biobank Borealis, University of Oulu and Northern Ostrobothnia Hospital District, Oulu, Finland. <sup>46</sup>Northern Finland Biobank Borealis, Oulu, Finland. <sup>47</sup>Northern Ostrobothnia Hospital District, Oulu, Finland. <sup>48</sup>Northern Savo Hospital District, Kuopio, Finland. <sup>49</sup>Pfizer, New York, NY, USA. <sup>50</sup>Pirkanmaa Hospital District, Tampere, Finland. <sup>51</sup>Sanofi, Paris, France. <sup>52</sup>University of Helsinki, Helsinki, Finland. <sup>53</sup>University of Stanford, Stanford, CA, USA. <sup>54</sup>University of Tampere, Tampere, Finland