Published in final edited form as:
Nat Genet. 2009 January ; 41(1): 35–46. doi:10.1038/ng.271.

Genome-wide association analysis of metabolic traits in a birth cohort from a founder population

Chiara Sabatti[1,2,28], **Susan K Service**[3,28], **Anna-Liisa Hartikainen**[4], **Anneli Pouta**[5], **Samuli Ripatti**[6], **Jae Brodsky**[2], **Chris G Jones**[3,7], **Noah A Zaitlen**[7], **Teppo Varilo**[8,9], **Marika Kaakinen**[10], **Ulla Sovio**[11], **Aimo Ruokonen**[12], **Jaana Laitinen**[13], **Eveliina Jakkula**[6], **Lachlan Coin**[11], **Clive Hoggart**[11], **Andrew Collins**[14], **Hannu Turunen**[6], **Stacey Gabriel**[15], **Paul Elliot**[11], **Mark I McCarthy**[16,17,18], **Mark J Daly**[15,19,20,21,22,23], **Marjo-Riitta Järvelin**[5,11,24], **Nelson B Freimer**[3,25,26], and **Leena Peltonen**[6,15,27]

[1]Department of Human Genetics, University of California, Los Angeles, Los Angeles, California 90095, USA [2]Department of Statistics, University of California, Los Angeles, Los Angeles, California 90095, USA [3]Center for Neurobehavioral Genetics, University of California, Los Angeles, Los Angeles, California 90095, USA [4]Department of Obstetrics and Gynaecology, University of Oulu, Fin-90220, Oulu, Finland [5]National Public Health Institute, Department of Child and Adolescent Health, Box 310, FIN-90101 Oulu, Finland [6]Institute for Molecular Medicine (FIMM), P.O. Box 20, FIN-00014 University of Helsinki, Finland [7]Department of Computer Science, University of California, Los Angeles, Los Angeles, California 90095, USA [8]Department of Medical Genetics, University of Helsinki, Helsinki, Finland [9]Department of Molecular Medicine, National Public Health Institute, Biomedicum Helsinki, 00290 Helsinki, Finland [10]Institute of Health Sciences, P.O. Box 5000, 90014 University of Oulu, Fin-90220, Oulu, Finland [11]Department of Epidemiology and Public Health, Imperial College London, Norfolk Place, St. Mary's Campus, London W2 1PG, UK [12]Department of Clinical Chemistry, University of Oulu, Fin-90220, Oulu, Finland [13]Finnish Institute of Occupational Health, Aapistie 1, Fin-90220, Oulu, Finland [14]Human Genetics Research Division, University of Southampton, Southampton General Hospital, Duthie Building (808), Tremona Road, Southampton SO16 6YA, UK [15]Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA [16]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Old Road, Headington, Oxford OX3 7LJ, UK [17]Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Headington, Oxford OX3 7BN, UK [18]Oxford NIHR Biomedical Research Centre, Churchill Hospital, Old Road, Headington, Oxford OX3 7LJ, UK [19]Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts 02114,USA [20]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA [21]Department of Psychiatry, Harvard Medical School, Boston, Massachusetts 02115, USA [22]Department of Medicine, Harvard Medical School, Boston, Massachusetts 02115, USA [23]Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts 02114, USA [24]Institute of

Correspondence should be addressed to N.F. (nfreimer@mednet.ucla.edu) or L.P. (leena@sanger.ac.uk)..
[28]These authors contributed equally to this work.

AUTHOR CONTRIBUTIONS

The study was designed by L.P., N.B.F., C.S., M.I.M., M.J.D., P.E., M.-R.J. and S.K.S. Information on traits was collected and maintained by M.-R.J., A.-L.H., A.P., A.R. and J.L. Genotyping was supervised by S.G. Database support was provided by H.T., U.S. and M.K. Statistical analysis was performed by C.S., S.K.S., A.C., C.G.J., J.B., N.A.Z., and M.J.D. S.R., E.J. and T.V. contributed to the analysis of the Finnish genetic signature. L.C., C.H. and P.E. participated in discussion of results. The manuscript was written by C.S., S.K.S., N.B.F. and L.P. All authors reviewed the manuscript.

Note: Supplementary information is available on the Nature Genetics website.

Published online at http://www.nature.com/naturegenetics/

Europe PMC Funders Author Manuscripts

Health Sciences and Biocenter Oulu, University of Oulu, Box 5000, Fin-90014, Oulu Finland [25]The Jane and Terry Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, Los Angeles, California 90095, USA [26]Department of Psychiatry, University of California, Los Angeles, Los Angeles, California 90095, USA [27]Wellcome Trust Sanger Institute, Cambridge, CB10 1HH, UK

Abstract

Genome-wide association studies (GWAS) of longitudinal birth cohorts enable joint investigation of environmental and genetic influences on complex traits. We report GWAS results for nine quantitative metabolic traits (triglycerides, high-density lipoprotein, low-density lipoprotein, glucose, insulin, C-reactive protein, body mass index, and systolic and diastolic blood pressure) in the Northern Finland Birth Cohort 1966 (NFBC1966), drawn from the most genetically isolated Finnish regions. We replicate most previously reported associations for these traits and identify nine new associations, several of which highlight genes with metabolic functions: high-density lipoprotein with *NR1H3* (*LXRA*), low-density lipoprotein with *AR* and *FADS1-FADS2*, glucose with *MTNR1B*, and insulin with *PANK1*. Two of these new associations emerged after adjustment of results for body mass index. Gene-environment interaction analyses suggested additional associations, which will require validation in larger samples. The currently identified loci, together with quantified environmental exposures, explain little of the trait variation in NFBC1966. The association observed between low-density lipoprotein and an infrequent variant in *AR* suggests the potential of such a cohort for identifying associations with both common, low-impact and rarer, high-impact quantitative trait loci.

Genetic variation and environmental exposures together determine the distribution of common diseases and disease-related traits within a population. Genome-wide association studies of cohort samples offer an opportunity to dissect the complex etiology and epidemiology of such traits. We report here a GWAS of metabolic traits in the Northern Finland Birth Cohort 1966 (NFBC1966)1,2, a sample that enrolled almost all individuals born in 1966 in the two northernmost Finnish provinces.

Several features advantageous for GWAS characterize NFBC1966: (i) participants derive from a genetic isolate that is relatively homogeneous in genetic background and environmental exposures, and that has more extensive linkage disequilibrium (LD) than most other populations3,4; (ii) cohort ascertainment reflects the composition of the population from which individuals are drawn, facilitating evaluation of effect sizes of identified genetic variants and comparison with environmental effects5; (iii) such ascertainment avoids selection and survival biases characteristic of case-control studies5; (iv) participants were born in the same year, eliminating age as a potential confounder and minimizing the effects of interindividual variability due to secular changes in environmental exposures; (v) longitudinal data collected over several decades 2,6-9 facilitate investigation of relationships between genetic variation, early life events and progression of disease risks.

The NFBC1966 study design enables GWAS of diverse traits. We focus here on nine quantitative traits that are heritable10-13 risk factors for cardiovascular disease (CVD) or type 2 diabetes (T2D): body mass index (BMI), fasting serum concentrations of lipids (triglycerides (TG), high-density lipoproteins (HDL) and low-density lipoproteins (LDL)), indicators of glucose homeostasis (glucose (GLU), and insulin (INS)) and inflammation (CRP), and systolic (SBP) and diastolic (DBP) blood pressure. Extreme values of these traits, in combination, identify a ‘metabolic syndrome’, hypothesized to increase risks for both CVD and T2D14,15.

Although environmental exposures such as diet influence these traits, the magnitude of their effects is controversial, largely because of uncertainty regarding critical time points for measuring exposures. For example, few datasets contain the longitudinal assessments needed to test the hypothesis that early life growth patterns may, through gene-environment interactions, affect values, in adulthood, of the traits comprising metabolic syndrome. NFBC1966 not only provides an ideal dataset for investigating interactions between genetic and environmental variables contributing to metabolic syndrome and its trait components, it offers the opportunity to prospectively investigate such interactions as participants attain ages characterized by high prevalence of CVD and T2D.

Prior GWAS results from case-control samples for these traits16-22 can be compared with findings from NFBC1966, in relation to differences in study design and sample composition. Specifically, NFBC1966 affords the opportunity for GWA analysis of the components of metabolic syndrome in a single cohort of young adults. We replicate most previously reported findings for these traits and report several newly identified associations, most of which are located in or near genes with postulated roles in metabolism. Our results suggest the value of cohorts such as NFBC1966 for elucidating genetic and environmental influences on quantitative traits.

RESULTS

Genotypic and phenotypic data

After applying exclusion criteria and quality control procedures (see Methods), we obtained Illumina Infinium genotypes on 329,091 SNPs in 4,763 individuals (for further information on the final sample, see **Supplementary Note** and Supplementary Table 1 online).

On the basis of literature results and available measurements in NFBC1966 and their patterns of missing values, we focused on six covariates that influence quantitative metabolic traits: alcohol use, smoking, BMI, sex, oral contraceptive use and pregnancy status. All were assessed contemporaneously with and highly associated to the nine traits (Table 1)2,23. Given current interest in the relationship between fetal growth and risk for CVD and T2D24, and as previous investigations in NFBC1966 and other Finnish cohorts have identified such associations2,25,26, we also considered effects of gestational age and birth BMI on the nine traits (see Methods, Table 1 and Supplementary Table 2 online).

Population structure

Recent studies using genome-wide SNPs have identified fine-scale population substructure that could affect association analyses27. We investigated the presence of such substructure in NFBC1966 by examining patterns of genetic similarity between different individuals (see Methods).

This analysis failed to identify the clearly separated clusters expected with strong subpopulation structure, but the genetic similarity between individuals correlates well with geographical and linguistic subdivisions of Northern Finland (Fig. 1, **Supplementary Note** and Supplementary Fig. 1 online). The information plotted in Figure 1 provides a measure of population of origin, which is available for the entire sample and which we use to correct our association analyses.

Genome-wide association analyses of metabolic traits

We analyzed the association between each trait and genomic locations using multiple approaches, with the aim of identifying genetic signals distinct from the effects of different covariates, correcting for possible population stratification, maximizing genomic coverage

and investigating gene-environment interactions. Our primary genome-wide analysis tested the additive effect of each SNP on the traits, adjusted for sex, use of oral contraceptives and pregnancy (see Methods).

The quantile-quantile plot (Fig. 2) of the tails of the distribution of observed *P* values (on the $\log_{10}$ scale) from the GWA tests departed from the expected quantiles only at the extreme, suggesting that population structure does not substantially affect the results. Indeed, the estimated genomic control parameter28 for each trait only modestly deviated from the null value of 1.0 (TG: 1.0203, HDL: 1.0419, LDL: 1.077, CRP: 1.009, GLU: 1.040, INS: 1.019, BMI: 1.031, SBP: 1.056, DBP: 1.020). Furthermore, association tests adjusted for population structure produced markedly similar *P* values (Table 2). An analysis of the entire *P*-value distribution allowed us to reject the global null of no association for most of the nine traits (see Methods, **Supplementary Note** and Supplementary Fig. 2 online).

Identification of specific loci associated with each of nine traits requires adoption of a criterion to adjust for multiple comparisons, with a resulting significance threshold. Controlling false discovery rate (FDR) across the nine traits maximizes the power to detect relevant loci with our relatively small sample by allowing a small percentage of false findings. Our rationale, based on an FDR procedure, for adopting a threshold of 5×10^{-7} is detailed in the Methods (**Supplementary Note** and Supplementary Fig. 3 online).

We identified in our main analysis (Table 2 and Fig. 3) 21 associations with a *P* value $<5 \times 10^{-7}$ to one or more traits (four for TG, five for HDL, six for LDL, three for CRP and three for GLU), in 20 genome regions; one region, *APOB*, shows association to both TG and LDL. In 14 of these regions, previous studies have identified association at $P<5 \times 10^{-7}$, whereas seven represent new observations. Genome-wide analysis adjusting for BMI identified two additional regions, one for LDL and one for INS (Table 2, in italics), for a total of 23 association findings.

Imputation analysis (see Methods) identified additional associated SNPs in most regions highlighted by directly genotyped SNPs (Supplementary Fig. 4 online) and, in several cases, provided significant evidence for replication of known loci for which genotyped SNPs had shown less significant results. This analysis, however, did not uncover additional new regions (ignoring imputation signals found in the absence of even nominal association to genotyped SNPs, see **Supplementary Note**).

The study design offered the opportunity to test whether these association results depended on either epidemiologic variables known to affect these traits or the geographic origin of NFBC1966 participants. All locations in Table 2 continued to show association when we used multivariate regression, which modeled the trait of interest as a linear function of all relevant epidemiological and geographical variables (Supplementary Table 3 online).

For most associations in Table 2, multiple SNPs provided association evidence. We identified genomic ‘boundaries’ for each finding, using an LD map29 to evaluate whether associations with multiple SNPs within a region likely represent distinct findings. Figure 4 presents, for the nine newly identified loci, information describing the extent of the signal and its relation to local LD (in LD-map distances surrounding the SNP with the smallest *P* value) as well as its position relative to genes (for analogous information for previously reported associations, see Supplementary Fig. 5 online).

For most of the newly identified loci, the effect size of the individual associations is similar in magnitude to those observed in other GWAS of these traits (**Tables** 2 and 3 and Supplementary Fig. 6 online). An exception is the association observed between LDL and a

variant in *AR* on chromosome X, which has a much larger effect size than other associations observed in NFBC1966 or in previous GWAS of lipid traits17,21.

Replication of previously reported associations

Previous GWAS found 37 loci (represented by 42 SNPs) associated with six of the traits (Table 3). In most cases the causal variant remains unidentified, and several of these SNPs are located in genes known to have substantial allelic heterogeneity30. We therefore evaluated the evidence for replication in two ways, assessing the association signal for the specific SNP in question as well as for the region it implicates.

To investigate SNP replication (Table 3), we test for association with the same SNP (genotyped or imputed) or to a proxy SNP ($r^2 \geq 0.8$ with the originally reported SNP) (see Methods). SNP imputations as well as identification of the best proxy SNP rely on correlations estimated in the HapMap CEU sample, which may in some instances be imprecise. Results of this analysis, including comparisons of effect sizes, are in the first columns of Table 3 and summarized below by trait (**Supplementary Note**). In all instances, when the same SNP or a perfect proxy was analyzed in NFBC1966, the direction of effect was the same as the reported direction, and in most cases the effect sizes were of a similar magnitude.

We defined gene regions on the basis of the LD pattern around the previously reported SNP (a window of 4 LD-map units, see Methods). We tested for association in these windows using the strongest association signal among the SNPs genotyped in NFBC1966 within the window, evaluating its significance through permutations (see Methods; last two columns of Table 3). In several cases where SNP replication was equivocal, we obtained strong evidence of association with the gene region (for example, TG with *APOB* and *APOA1/C3/A4/A5*, HDL with *LIPC*, and CRP with *CRP*; Table 3).

Summary of association results for each trait

For each trait we summarize our findings and present newly identified candidate genes (Table 2 and Fig. 4). Additionally, to evaluate the portion of explained variance for each trait, we incorporate into a multivariate regression model environmental variables hypothesized to contribute to total trait variability (see Methods) together with genetic associations (previously known and newly discovered, **Tables** 2 and 3): the results of this analysis are in Figure 5 and Supplementary Table 4 online. The effect of the multiple associated SNPs in these models is additive across loci and associations in Table 2 remain significant in models that include all associated SNPs. As expected, associated SNPs in different genetic regions are not correlated with each other.

Triglycerides

We replicated SNP associations with *GCKR* and *LPL* at $P < 5 \times 10^{-7}$, and with *ANGPTL3-DOCK7-ATG4C* and *BCL7B-TBL2-MLXIPL* at $5 \times 10^{-7} < P < 0.001$. We did not observe any SNP association with *GALNT2, LIPC* and *NCAN-CILP2-PBX4* ($P > 0.05$). The remainder of SNP associations for TG showed equivocal replication evidence in NFBC1966 ($0.001 < P < 0.05$). A newly identified TG association on chromosome 15 depends solely on evidence from rs2624265 (Table 2 and Fig. 4), and its significance decreases considerably when covariates are taken into account (Table 2). This SNP shows modest association to TG ($P = 0.0027$), with similar effect sizes, in the ENGAGE consortium meta-analysis of >10,000 European samples31. Together, the associated loci explain 4.2% of trait variability.

High-density lipoprotein

We replicated SNP associations with *CETP* and *LCAT* at $P < 5 \times 10^{-7}$, and with *GALNT2, LPL* and *ABCA1* at $5 \times 10^{-7} < P < 0.001$. We did not observe any SNP association with *APOA1/C3/A4/A5, ZNF259* or *BUD13* ($P > 0.05$). The remainder of SNP associations for HDL showed equivocal replication evidence in NFBC1966 ($0.001 < P < 0.05$). We identified association in two newly identified gene regions. An associated region on chromosome 11 includes *NR1H3*, also termed *LXRA*, a transcriptional regulator of cholesterol metabolism32. A second associated region is on chromosome 17. In this region only one additional genotyped SNP approached the 5×10^{-7} threshold, and imputation analysis identified several SNPs showing weaker association at this location (Fig. 4). Both locations showed associations with HDL to the same SNPs as in the ENGAGE consortium meta-analysis31 ($P = 6.96 \times 10^{-8}$ and $P = 0.003$, respectively). Together, the associated loci explain 6% of trait variability.

Low-density lipoprotein

We replicated SNP associations with *CELSR2-PSRC1-SORT1, APOB* and *LDLR* at $P < 5 \times 10^{-7}$, and with *HMGCR* at $5 \times 10^{-7} < P < 0.001$. We did not observe any SNP association with *B3GALT4* and *NCAN-CILP2-PBX4* ($P > 0.05$). For two SNPs, we observed convincing association; however, our proxy SNP is not in strong LD with the reported SNP (located in *PCSK9* and the *APO* cluster).

We identified three new LDL associations. SNP rs4844614 on chromosome 1 lies in an intron of *CR1L*, which encodes a complement receptor protein not known to have a metabolic function. On chromosome X, association was observed for rs5031002 located in intron 6 of *AR*, which encodes a ligand-dependent transcription factor with several functions, including control of circulating androgen levels; alterations in such levels are associated with sex-specific dyslipidemias33. This low-frequency (minor allele frequency (MAF) = 0.017) variant is associated with markedly increased LDL, primarily among 38 males possessing it. Only one female subject is homozygous for the minor allele, and female heterozygotes show modest elevations in LDL (Supplementary Fig. 7 online). Association using females only identifies a similar effect size (0.3547 mmol/l for males and 0.325 mmol/l for females), albeit with a less significant *P* value (males-only $P = 4 \times 10^{-7}$, females-only $P = 0.0039$). Evidence for association with *AR* comes entirely from rs5031002, which shows low levels of LD with surrounding SNPs. Imputed HapMap SNPs in this region also showed relatively little association with LDL (Supplementary Fig. 6).

A locus on chromosome 11, including *FADS1-FADS2*, is associated at $P < 5 \times 10^{-7}$ when we adjust for BMI. These genes encode desaturases that have previously demonstrated strong association with various fatty acids present in serum phospholipids34. Association in this region with LDL, of a comparable magnitude, was also found by the ENGAGE consortium meta-analysis31. The entire collection of associated loci explains 6% of total variability.

C-reactive protein

We replicated SNP associations with *LEPR* and *LEF1* at $P < 5 \times 10^{-7}$ and with *IL6R* at $5 \times 10^{-7} < P < 0.001$. We did not observe any SNP association with *GCKR* ($P > 0.05$). For one SNP in *CRP*, we observed convincing association, but our proxy SNP is not in strong LD with the reported SNP. The remainder of SNP associations for CRP showed equivocal replication evidence in NFBC1966 ($0.001 < P < 0.05$). These loci explained 4% of trait variability.

Glucose

We replicated with $P < 5 \times 10^{-7}$ the previously reported SNP association of glucose with *G6PC2-ABCB1* and observed associations with new loci on chromosomes 7 and 11. The latter association ($P = 5.9 \times 10^{-8}$) with variants in *MTNR1B* has also been identified in a larger sample35. *MTNR1B* is transcribed in human islets and rodent insulinoma cell lines, and the translated receptor is thought to mediate the inhibitory effect of melatonin on insulin secretion. The two associated SNPs on chromosome 7 travel on a single haplotype, as do the two associated SNPs on chromosome 11. These loci explained 1.6% of trait variability.

Insulin

Adjusting for BMI, we identified an INS association on chromosome 10 at rs11185790, which lies in an intron of *PANK1*. In imputation analyses, when INS was not adjusted for BMI, this region showed association at similar levels. *PANK1* encodes panthothenate kinase, a critical enzyme in the synthesis of coenzyme A that is induced by bezafibrate, a hypolipidemic agent36. Additionally, mouse chemical knockout studies of panthothenate kinase resulted in a hypoglycemic phenotype, providing functional evidence supporting the role of this gene in glucose metabolism37; *PANK1* alone explains 0.56% of total variability.

Body mass index

No individual loci attained genome-wide significance in NFBC1966, including two established BMI loci, *FTO* (NFBC1966 was part of the sample used in the original *FTO* finding) and *MC4R*, which we replicate at significance levels of $P = 2.4 \times 10^{-4}$ and $P = 1.78 \times 10^{-3}$, respectively. *FTO* and *MC4R* together explain 0.55% of the trait variability.

Systolic and diastolic blood pressure

No individual locus achieved genome-wide significance in NFBC1966.

Genome-wide analysis of interactions

The genome-wide association analyses described above identify loci accounting for a limited fraction of total trait variance in NFBC1966. Additional loci may be identified by analyzing interactions between genotypes and environment. Interaction analyses address whether the effect of a SNP on a trait depends on the level of an environmental covariate. We evaluated interaction with factors hypothesized to interact specifically with metabolic traits: measures of early growth (birth BMI and growth in the first six months of life), BMI at age 31, oral-contraceptive use and sex.

We consider genome-wide interaction analysis in NFBC1966 to be exploratory for two reasons: (i) given the level of multiple testing in these additional analyses, our sample is underpowered to achieve confirmatory significance levels for effect sizes similar to those reported above; (ii) replication of NFBC1966 interaction results may be difficult, as not all GWAS of these traits have data on the same covariates. Nevertheless, we present findings to generate hypotheses for future studies, focusing on tests resulting in $P < 5 \times 10^{-7}$ (**Supplementary Note** and Supplementary Table 5a-d online). In brief, all environmental factors except growth in the first six months of life interact with at least one SNP to influence the mean level of at least one trait.

DISCUSSION

This GWAS identifies variants influencing risk for metabolic traits in a population cohort of young adults. We replicate most prior findings and identify nine previously unreported associations. Five of these associations—HDL with *NR1H3* (*LXRA*), LDL with *AR* and

FADS1-FADS2, glucose with *MTNR1B* and insulin with *PANK1*—implicate genes with known or postulated roles in metabolism and, therefore, support heightened investigation of pathways through which they are hypothesized to function.

We highlight 23 findings passing a significance threshold (5×10^{-7}) accounting for the multiple comparisons involved in GWAS of several traits. Choosing a threshold involves judging the degree of dependence among the tests. This dependence varies across studies, for both phenotypes and genotypes. We investigated component traits for metabolic syndrome, which are relatively dependent. More importantly, LD structure, which varies substantially between populations, determines the degree of dependence of SNPs in a GWAS; a 'one size fits all' threshold is overconservative in some situations and under-conservative in others. The extensive LD in Northern Finland4 supports the threshold we selected.

More important than the significance threshold in any one study is the accumulation of evidence that particular loci contribute to trait variation. Replication most precisely refers to a specific SNP (or another SNP in strong LD with that SNP) with the same direction of effect38. The 12 instances in which we replicate previously reported SNP associations (at $P < 5 \times 10^{-7}$, Table 3) are the most statistically significant findings in our study, with generally the largest effect sizes (which are roughly similar to those observed in the GWAS initially reporting these associations).

At four additional genes we observed associations ($<5 \times 10^{-7}$ threshold), but not for SNP(s) for which association was previously reported (Table 3). SNPs in different studies may each be strongly associated with a trait and in strong LD with an unknown causal variant but not with one another, resulting in nonreplication at the SNP level. Additionally, association with different SNPs across studies may reflect multiple causal variants within the same gene.

Our study differs from previous investigations in power and study design, each possibly accounting for our nonreplication of previously reported associations, and our identification of nine previously unknown loci. Previously reported associations with these traits derive largely from meta-analyses of samples much larger (up to ~20,000 individuals17) than that of NFBC1966; however, published data do not indicate obvious differences that could reflect sample size (for example, level of statistical support or effect size) between associations that we replicate and those that we do not replicate (**Supplementary Note**).

Cohort studies such as NFBC1966, unlike case-control studies, investigate individuals drawn from the full distribution of disease-associated quantitative traits. Differences between these study designs in the strength of association, in terms of significance level, at particular loci may provide clues regarding their function; stronger association in cohort studies may characterize loci with broader phenotypic effects, whereas stronger association in case-control studies may characterize loci that more directly influence the diseases that define cases.

Birth cohorts, unlike other study designs, factor out age-specific effects, both age-varying associations and age-gene interactions39. New associations identified in NFBC1966 may indicate loci that particularly influence metabolic traits in early adulthood, when individuals are mostly free of late-onset diseases. Additionally, variability in metabolic traits in younger adults may be more genetically determined than in older adults, simply because of less accumulation of environmental exposures. At the same time, the impact of other loci, potentially reflecting genetically mediated responses to such accumulations, may not yet be discernable in NFBC1966, for example previous associations observed for CRP in older individuals19.

The metabolic syndrome concept assumes that correlations between the traits assessed here identify an important source of risk for CVD and T2D. We confirmed overlapping associations between TG and HDL at *LPL, GALNT2* and *APOA1/C3/A4/A5*, and between TG and LDL at *NCAN-CILP2-PBX4* and *APOB*, and we observed HDL association at *APOB*, which to our knowledge has not been previously reported. No additional gene regions showed significant association with multiple traits, suggesting that common variants with sizable effects do not substantially account for trait correlations that characterize metabolic syndrome. More powerful methods might identify pleiotropic effects among these traits. Principal Component of Heritability (PCH), a technique to identify the linear combination of phenotypic values that maximizes heritability with a genotyped SNP40, seems powerful for uncovering pleiotropy, but in its current form it is computationally impractical for GWAS.

The longitudinal study design facilitates additional types of analyses that may illuminate the relationships between multiple influences on trait variation. For example, NFBC1966 has, from birth through age 31, assessed participants' BMI. Elevated adult BMI, a core component of metabolic syndrome, is a well-established risk factor for CVD and T2D, and abnormal early-life BMI is also hypothesized to predispose to these disorders2,24,26.

In GWAS one can either adjust traits for such a covariate or condition on them in interaction analyses. By adjusting traits for adult BMI (in addition to sex, pregnancy status and oral-contraceptive use), we discovered two new associations (for LDL and INS) not identified in the primary analyses. In an interaction analysis, where individuals were stratified as having normal or elevated BMI, three loci not previously highlighted showed associations at $P < 5 \times 10^{-7}$ (Supplementary Table 5c and Supplementary Fig. 8a online). A similar analysis identified a locus for which the effect on CRP depended on the level of BMI at birth (**Supplementary Note** and Supplementary Fig. 8b). We consider these interaction analyses exploratory, as power is limited by several factors: the small effects characterizing the relationship between covariates and the traits, the relatively small sample of NFBC1966 given such effect sizes, and the requirement for a smaller alpha level given the multiple testing involved in interaction analyses of several covariates.

The currently identified loci, singly and cumulatively, explain little of the trait variability in NFBC1966 (at most ~6% based on multivariate regression). The estimated heritability for each trait is in the range of 50% or greater10,11,13,41,42. Our analyses indicate that environmental variables account for less than ~30% of trait variability. Even assuming that these analyses do not include all relevant exposures, the discrepancy between estimated heritability and the percentage of trait variance now accounted for by genetic associations suggests that loci contributing the largest proportion of trait variance remain undiscovered. Such loci could include additional common variants (with even smaller effects), require interaction with environmental variables (also presumably with small effects) or include multiple infrequent variants. Such rare variants could have substantial effects in small numbers of individuals and, in aggregate, represent a substantial proportion of trait variance in the population43.

The relative importance of common and rare variants in the genetic architecture of complex traits is a topic of intense current discussion, but so far with little data. Our results suggest that cohorts from population isolates, such as NFBC1966, may be a valuable source of such data. Current GWAS are designed to investigate association with common variants. Most such variants predate the separation of modern European populations, and, therefore, it is predictable that most of the associations detected in NFBC1966 with common SNPs reflect the actions of loci with roughly comparable effect sizes across such populations.

Rare variants, by contrast, likely arose more recently, and their frequency may vary substantially between populations, particularly considering isolates such as Northern Finland in which repeated bottlenecks resulted in extensive genetic drift at many loci. This phenomenon likely explains the association (not found in other studies) of LDL with an infrequent SNP (MAF < 0.02) in *AR*. In a sample of almost 5,000 individuals, such an infrequent variant must have a large effect to be statistically detectable, as was the case in NFBC1966 (males with the minor allele have a mean LDL elevation of 28 mg/dl).

Current GWAS designs are poorly powered to detect association with infrequent variants, and our finding for LDL and *AR* probably reflects luck as well as distinct genetic features of Northern Finland. Genome-wide resequencing efforts now underway, such as the 1000 Genomes project, will likely permit investigation of associations with infrequent variants for a wide range of complex traits. As with the LDL-*AR* association, these findings may prove difficult to replicate, but may immediately suggest biological pathways for further investigation.

METHODS

Study sample and clinical characteristics

Study subjects were members of the Northern Finnish Birth Cohort of 1966 (NFBC1966) (see **Supplementary Note** for more information on NFBC1966). Informed consent from all study subjects was obtained using protocols approved by the Ethical Committee of the Northern Ostrobothnia Hospital District.

Trait measurements

All traits were measured at the 31-year examination. Blood samples were drawn after overnight fasting in the morning (between 0800 and 1100 h). The concentrations of INS were analyzed by radioimmuno-assay (Pharmacia Diagnostics), and GLU by a glucose dehydrogenase method (Granutest 250, Diagnostica Merck). Serum CRP concentrations were determined by immunoenzymometric assay (Medix Biochemica). Serum GLU, total cholesterol (TC), HDL and TG were determined by enzymatic methods using a Hitachi 911 Clinical Chemistry Analyzer (Boehringer Mannheim). Additional detail on all assays can be found in the **Supplementary Note**.

Height and body weight were measured using a standardized height measure and scale. BMI was calculated as kg m^{-2}. Weight was not measured for pregnant women. SBP and DBP were measured by trained nurses with a mercury sphygmomanometer using a standardized procedure and ongoing quality control2. Measurements were made on the right arm, after the subject had rested in a sitting position for 15 min. The average of duplicate measures was used as the trait measure. Blood pressure of individuals who were on blood pressure medication (1.9% of study subjects) was adjusted by adding 15 mm Hg to SBP and 10 mm Hg to the DBP44.

For association analyses, TG, BMI, INS and GLU were all natural log transformed. One-half of the detection limit (0.002 mg/l) of CRP was added to 0 values, and then all values were natural log transformed for association analysis. Results use SI units for all traits.

Individuals were excluded from analysis of specific traits on the basis of criteria that were established separately for each trait (**Supplementary Note**). Individuals were excluded from analysis of lipid traits (TG, HDL, LDL) if they had not fasted before blood collection, or if they were diabetic. Individuals were excluded from analysis of GLU and INS if the blood sample was nonfasting, if they were diabetic, on diabetic medication, pregnant, or if their glucose/insulin measurement (after correction for sex, oral-contraceptive use, and pregnancy

status) was in excess of three s.d. from the mean. Individuals were excluded from analysis of BMI if their weight was not directly measured. No exclusion criteria were applied to CRP or to SBP and DBP.

Alcohol consumption was measured in the absolute amount of alcohol (grams per day), as reported in a questionnaire administered at age 31. Smoking habits were measured in the same questionnaire by using the answer to question: “Have you ever smoked in your life?” Use of oral contraceptive and pregnancy was determined on the basis of self-reported data at age 31.

Birth BMI was regressed on gestational age, mother’s parity, smoking status, height and weight before pregnancy and by the infant’s sex. Residuals from this regression were ordered and placed into five quantiles, defining the ordered factor we use in the analysis. Early growth was examined by taking the residuals from a regression of BMI at six months on birth BMI, as done previously26. Gestational age of the subject was dichotomized into >37 weeks (full term) and ≤37 weeks (preterm).

The NFBC1966 database records the municipality of origin of both parents of each subject. Most parents of NFBC1966 individuals were born in municipalities that can be assigned to one of six linguistic and geographic regions that can be distinguished in Northern Finland: East Lapland, Central Lapland, West Lapland, North Oulu, South Oulu and Kainuu. A subject in NFBC1966 is assigned to one of these groups when both parents belong to the same group (see **Supplementary Note**).

Sample preparation, genotyping and quality control

Genomic DNA was extracted from whole blood using standard methods. All DNA samples for the Illumina Infinium 370cnvDuo array were prepared for genotyping by the Broad Institute Biological Sample Repository (BSP). Details on genotyping procedures can be found in the **Supplementary Note**.

All individuals in the study were genotyped with call rates >95%. Individuals with discrepancy between their reported sex and the sex determined from the X chromosome were excluded from analysis. We used the identity-by-descent (IBD) analysis option of PLINK45 to determine possible relatedness among our sample subjects and to identify sample duplications and sample contamination (the latter identified as individuals who seemed to be related to nearly everyone in the sample). If the sample duplication issue could not be resolved by external means, both samples were excluded. All apparently contaminated samples were excluded. We identified individuals related at the level of half-sibs or closer with the IBD analysis and excluded one subject from each pair (the subject with less complete genotyping). Subsequent to this overall exclusion, individuals may have been excluded from analysis of specific traits, as detailed above.

Markers were excluded from analysis if the call rate in the final sample was <95%, if the *P* value from a test of Hardy-Weinberg Equilibrium (HWE) was <0.0001, or if the MAF was <1%. A complete description of the individuals available for analysis can be found in the **Supplementary Note**.

Genotype analysis

Using individuals and markers that passed inclusion criteria, we assessed population structure using classical multidimensional scaling on the matrix of identity-by-state sharing of all pairs of individuals, as estimated by the program PLINK45. The first two dimensions of the result were plotted, color coding individuals known to belong to different linguistic/geographic groups. Unlike geographical/linguistic groups that can be assigned on the basis

of parental birth records only for a subset of the sample, multidimensional scaling (MDS) components are available for the entire sample. We adjusted for ancestry using the first two components of MDS, as they allow us to distinguish the different linguistic/geographic groups, and are significantly associated with a large fraction of our traits.

We constructed a genome-wide LD map in LD units (LDU) from the SNP data to evaluate the extent of LD on each chromosome. Map distances are an LD analog of the centimorgan scale of linkage maps, with one LD unit corresponding to the average distance over which LD declines to 'background' levels. The LD maps were used to gauge the independence of signals in nearby genomic regions. Multiple SNPs located within two LDUs of each other were (conservatively) considered likely to represent the same signal. LD maps were constructed from weighted pairwise association data following Maniatis *et al.29* as implemented in the LDMAP program. The details of LDU map construction precisely followed the methods described by Service *et al.4*. Fine-scale genetic map information for the CEU population was downloaded from the HapMap website.

Statistical analysis

Following a standard genetic approach, we selected a main model that maximizes the power to detect any genetic association, even when the pathway connecting gene and trait may be complex and involve other traits. That is, we selected a model in which the traits of interest are minimally adjusted: we correct for sex, pregnancy status and use of oral contraceptive. Adjusting for sex is standard practice. Oral contraceptive (OC) use has been previously associated with metabolic traits in this population23 and, together with pregnancy status (PG), is important in our comparatively young sample. The cumulative effect of these variables is captured by a factor (SexOCPG in Table 1) that has one level for males and nine levels for females corresponding to all combinations of reported use of oral contraceptives (yes, no, unknown) and pregnancy status (yes, no, unknown). As noted from Table 1, sex, oral-contraceptive use, pregnancy status and BMI are, by far, the covariates with strongest effects on the metabolic traits.

In our main analysis (analysis A), the residuals from the regression of each trait on SexOCPG were used as the outcome in a main effects model that tested the association between SNP markers and trait. SNP genotypes were coded as 0, 1 or 2 copies of the minor allele. An additive effect of genotype was assumed where the effect on the trait of the heterozygote was estimated to be midway between the levels of the two homozygotes. Association testing proceeded via regression analysis, and was done in PLINK45. On the basis of the strong effect of BMI on each of the traits documented in Table 1, we also carried out a genome-wide scan adjusting traits for BMI at age 31, in addition to SexOCPG (analysis B). To rule out the possibility of spurious associations due to population stratification, we corrected the *P* values of analysis A using the genomic-control parameter λ*28*, estimated as the median of the χ2 statistics and divided by 0.456. Furthermore, we conducted a genome-wide scan adjusting for the first two components of MDS (analysis C).

To maximize coverage, it is common to test for association at nongenotyped SNPs whose genotype can be imputed on the basis of haplotypes of nearby markers. We used a custom modification of weighted haplotype association (WHAP)46 that can be directly applied to quantitative traits and implements a novel quality control procedure (analysis D). Details can be found in the **Supplementary Note**. Haplotype phasing was done with Beagle47.

We established a threshold of significance of 5×10^{-7}, conservatively applying the Benjamini Hochberg48 procedure to control FDR at the 0.05 level across the *P* value from analysis A of the nine traits (see **Supplementary Note**). *P* values from the other analysis are highly correlated with those obtained in our main model and treating them as independent

would artificially reduce power. To validate this theoretically established threshold, we ran 1,000 genome scans under the complete null hypothesis, using permutations. Across the nine traits, we obtained an average of 1.6 SNPs that lead to smaller *P* values than 5×10^{-7}; this can be taken as the expected number of false positives among the results we report.

Known loci associated with the nine traits were identified through Internet and Pubmed search, beginning with a catalog of published genome-wide association studies compiled by the National Human Genome Research Institute (see URLs section below). Literature searches were conducted for the names of the nine traits together with key words 'genome-wide association' and 'genetic association'. Loci were considered 'known' if they had a *P* value $<5 \times 10^{-7}$. We were able to identify 42 previously reported associations.

Quantile-quantile plots were produced plotting the ordered observed *P* values (on the $-\log_{10}$ scale) for each trait against the expected value of the corresponding ordered statistics for 329,091 independent observations from the uniform distribution. The 0.025 and 0.975 quantiles of the beta approximation for the distribution of these order statistics under the hypothesis of independence was used to obtain point-wise confidence bounds. We used a set of 1,000 permutations for each of the traits to obtain empirical point-wise confidence intervals for the ordered *P* values.

To assess the amount of genetic signal remaining in our data, besides the signals corresponding either to our new genome-wide hits or to known locations, we excluded all SNPs with $P < 0.001$ that are within 4 LD units (2 LDU on either side) of either a new hit or a known location and compared the distribution of the *P* values for remaining SNPs with that expected under the complete null. We tested the global null hypothesis of no genetic effect using the Higher Criticism49 statistic (**Supplementary Note**).

To identify which of the covariates in Table 1 best contributes to explaining each of the nine traits, we used a step-wise model selection procedure as implemented in the R function step (). Results are listed in Supplementary Table 3. For each of the associations reported in Table 2, we investigate if the SNP remains significant in a multivariate regression that includes all the covariates identified as above. *P* values from the corresponding F test are listed in Table 2. To assess the overall impact of all genetic association and the relative importance of each locus, we carried out a multivariate regression that includes the covariates above and the SNPs carrying the strongest signal from each of the regions in Table 2 and each of the previously reported regions for which we obtained an empirical *P* value >0.05 in Table 3.

To investigate SNP replication, we examined the *P* value in our study of the originally reported SNP (either directly genotyped or imputed). When this was neither genotyped nor reliably imputable, we looked at the best proxy SNP, defined as the SNP in HapMap that is in highest LD with the original SNP (r^2 of at least 0.8) and that is either genotyped or imputable in our sample.

To investigate gene-region replication, we identified all SNPs in our sample that reside within 2 LD units to either side (4 LDU total) of the known location. We identified the smallest association *P* value among these SNPs (NFBC min *P* in Table 3). We evaluated the probability that the smallest *P* value among all the SNPs in each of these regions is smaller then NFBC min *P* under the null hypothesis using 5,000 permutations. The proportion of permutations that lead to a smaller min *P* defines the empirical *P* in Table 3.

URLs. North Finland Birth Cohort Studies, http://kelo.oulu.fi/NFBC; ENGAGE consortium, http://www.euengage.org; NHGRI catalog of GWAS, http://www.genome.gov/26525384.

We considered gene-environment interactions with six variables: sex, use of oral contraceptives, an indicator variable for overweight (BMI > 25) individuals, subject's gestational age (dichotomized as pre-term or term), birth BMI and early growth (see above). These epidemiological covariates, population structure covariates, and early life covariates were shown to affect mean trait levels in NFBC1966 (Table 1 and **Supplementary Tables 2** and **3**). Interaction of genotype with sex was analyzed in a sample that excluded women on oral contraceptives and pregnant women. Interaction of genotype with oral contraceptive use was analyzed in a sample of women that excluded pregnant women and women whose oral-contraceptive use was unknown. Interaction of genotype with the overweight indicator, gestational age, birth BMI and early growth was analyzed in the full sample, where traits had been adjusted for sex, oral-contraceptive use and pregnancy status. When analyzing interaction with binary variables (sex, oral-contraceptive use, overweight indicator, preterm or full-term gestational age), we evaluated interaction by comparing the effect size of the loci in the two groups, as implemented in the PLINK 'gxe' procedure45. Continuous covariates (subject's birth BMI and early growth) were evaluated in a linear regression model that included a term for the additive effect of genotype, a term for the continuous covariate, and a term for the interaction of genotype and covariate.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We acknowledge the support of US National Heart, Lung, and Blood Institute grant HL087679 through the STAMPEED program, grants MH083268, GM053275-14 and U54 RR020278 from the US National Institutes of Health, grant DMS-0239427 from the National Science Foundation, the Medical Research Council of the UK, EURO-BLCS, QLG1-CT-2000-01643 and the European Community's Seventh Framework Programme (FP7/2007-2013), ENGAGE project and grant agreement HEALTH-F4-2007-201413. The authors would like to thank the Center of Excellence in Common Disease Genetics of the Academy of Finland and Nordic Center of Excellence in Disease Genetics, the Sydantautisaatio (Finnish Foundation of Heart Diseases), the Broad Genotyping Center, D. Mirel, H. Hobbs, J. DeYoung, P. Rantakallio, M. Koiranen and M. Isohanni for advice and assistance.

References

1. Rantakallio P. Groups at risk in low birth weight infants and perinatal mortality. Acta Paediatr. Scand. Suppl. 1969; 193:43.
2. Jarvelin MR, et al. Early life factors and blood pressure at age 31 years in the 1966 northern Finland birth cohort. Hypertension. 2004; 44:838–846. [PubMed: 15520301]
3. Varilo T, Peltonen L. Isolates and their potential use in complex gene mapping efforts. Curr. Opin. Genet. Dev. 2004; 14:316–323. [PubMed: 15172676]
4. Service S, et al. Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. Nat. Genet. 2006; 38:556–560. [PubMed: 16582909]
5. Manolio TA, Bailey-Wilson JE, Collins FS. Genes, environment and the value of prospective cohort studies. Nat. Rev. Genet. 2006; 7:812–820. [PubMed: 16983377]
6. Laitinen J, Pietilainen K, Wadsworth M, Sovio U, Jarvelin MR. Predictors of abdominal obesity among 31-y-old men and women born in Northern Finland in 1966. Eur. J. Clin. Nutr. 2004; 58:180–190. [PubMed: 14679384]
7. Pouta A, et al. Manifestations of metabolic syndrome after hypertensive pregnancy. Hypertension. 2004; 43:825–831. [PubMed: 14981067]
8. Taponen S, et al. Metabolic cardiovascular disease risk factors in women with self-reported symptoms of oligomenorrhea and/or hirsutism: Northern Finland Birth Cohort 1966 Study. J. Clin. Endocrinol. Metab. 2004; 89:2114–2118. [PubMed: 15126528]

9. Saari KM, et al. A 4-fold risk of metabolic syndrome in patients with schizophrenia: the Northern Finland 1966 Birth Cohort study. J. Clin. Psychiatry. 2005; 66:559–563. [PubMed: 15889940]
10. Heller DA, de Faire U, Pedersen NL, Dahlen G, McClearn GE. Genetic and environmental influences on serum lipid levels in twins. N. Engl. J. Med. 1993; 328:1150–1156. [PubMed: 8455681]
11. Souren NY, et al. Anthropometry, carbohydrate and lipid metabolism in the East Flanders Prospective Twin Survey: heritabilities. Diabetologia. 2007; 50:2107–2116. [PubMed: 17694296]
12. Korkeila M, Kaprio J, Rissanen A, Koskenvuo M. Effects of gender and age on the heritability of body mass index. Int. J. Obes. 1991; 15:647–654. [PubMed: 1752725]
13. Wessel J, et al. C-reactive protein, an 'intermediate phenotype' for inflammation: human twin studies reveal heritability, association with blood pressure and the metabolic syndrome, and the influence of common polymorphism at catecholaminergic/beta-adrenergic pathway loci. J. Hypertens. 2007; 25:329–343. [PubMed: 17211240]
14. Alberti KG, Zimmet P, Shaw J. Metabolic syndrome-a new world-wide definition. A Consensus Statement from the International Diabetes Federation. Diabet. Med. 2006; 23:469–480. [PubMed: 16681555]
15. Haffner SM. Relationship of metabolic risk factors and development of cardiovascular disease and diabetes. Obesity (Silver Spring). 2006; 14(Suppl. 3):121S–127S. [PubMed: 16931493]
16. Frayling TM, et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science. 2007; 316:889–894. [PubMed: 17434869]
17. Kathiresan S, et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat. Genet. 2008a; 40:189–197. [PubMed: 18193044]
18. Loos RJ, et al. Common variants near MC4R are associated with fat mass, weight and risk of obesity. Nat. Genet. 2008; 40:768–775. [PubMed: 18454148]
19. Ridker PM, et al. Loci related to metabolic-syndrome pathways including LEPR,HNF1A, IL6R, and GCKR associate with plasma C-reactive protein: the Women's Genome Health Study. Am. J. Hum. Genet. 2008; 82:1185–1192. [PubMed: 18439548]
20. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447:661–678. [PubMed: 17554300]
21. Willer CJ, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat. Genet. 2008; 40:161–169. [PubMed: 18193043]
22. Kathiresan S, Musunuru K, Orho-Melander M. Defining the spectrum of alleles that contribute to blood lipid concentrations in humans. Curr. Opin. Lipidol. 2008b; 19:122–127. [PubMed: 18388691]
23. Morin-Papunen L, et al. Comparison of metabolic and inflammatory outcomes in women who used oral contraceptives and the levonorgestrel-releasing intrauterine device in a general population. Am. J. Obstet. Gynecol. 2008; 199:529.e1–529.e10. [PubMed: 18533124]
24. Barker DJ, Osmond C, Forsen TJ, Kajantie E, Eriksson JG. Trajectories of growth among children who have coronary events as adults. N. Engl. J. Med. 2005; 353:1802–1809. [PubMed: 16251536]
25. Tzoulaki I, et al. Size at birth, weight gain over the life course, and low-grade inflammation in young adulthood: northern Finland 1966 Birth Cohort study. Eur. Heart J. 2008; 29:1049–1056. [PubMed: 18403494]
26. Kajantie E, Barker DJ, Osmond C, Forsen T, Eriksson JG. Growth before 2 years of age and serum lipids 60 years later: the Helsinki Birth Cohort study. Int. J. Epidemiol. 2008; 37:280–289. [PubMed: 18267964]
27. Helgason A, Yngvadottir B, Hrafnkelsson B, Gulcher J, Stefansson K. An Icelandic example of the impact of population structure on association studies. Nat. Genet. 2005; 37:90–95. [PubMed: 15608637]
28. Bacanu SA, Devlin B, Roeder K. Association studies for quantitative traits in structured populations. Genet. Epidemiol. 2002; 22:78–93. [PubMed: 11754475]
29. Maniatis N, et al. The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. Proc. Natl. Acad. Sci. USA. 2002; 99:2228–2233. [PubMed: 11842208]

Europe PMC Funders Author Manuscripts

30. Aulchenko YS, et al. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nat. Genet. Dec 7.2008 advance online publication, doi:10.1038/ng.269.
31. Cohen JC, et al. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science. 2004; 305:869–872. [PubMed: 15297675]
32. Zelcer N, Tontonoz P. Liver X receptors as integrators of metabolic and inflammatory signaling. J. Clin. Invest. 2006; 116:607–614. [PubMed: 16511593]
33. Wu FC, von Eckardstein A. Androgens and coronary artery disease. Endocr. Rev. 2003; 24:183–217. [PubMed: 12700179]
34. Schaeffer L, et al. Common genetic variants of the FADS1 FADS2 gene cluster and their reconstructed haplotypes are associated with the fatty acid composition in phospholipids. Hum. Mol. Genet. 2006; 15:1745–1756. [PubMed: 16670158]
35. Prokopenko I, et al. Variants in *MTNR1B* influence fasting glucose levels. Nat. Genet. Dec 7.2008 advance online publication, doi:10.1038/ng.290.
36. Ramaswamy G, Karim MA, Murti KG, Jackowski S. PPARalpha controls the intracellular coenzyme A concentration via regulation of PANK1alpha gene expression. J. Lipid Res. 2004; 45:17–31. [PubMed: 14523052]
37. Zhang YM, et al. Chemical knockout of pantothenate kinase reveals the metabolic and genetic program responsible for hepatic coenzyme A homeostasis. Chem. Biol. 2007; 14:291–302. [PubMed: 17379144]
38. NCI-NHGRI Working Group on Replication in Association Studies. et al. Replicating genotype-phenotype associations. Nature. 2007; 447:655–660. [PubMed: 17554299]
39. Lasky-Su J, et al. On the replication of genetic associations: timing can be everything! Am. J. Hum. Genet. 2008; 82:849–858. [PubMed: 18387595]
40. Klei L, Luca D, Devlin B, Roeder K. Pleiotropy and principal components of heritability combine to increase power for association analysis. Genet. Epidemiol. 2008; 32:9–19. [PubMed: 17922480]
41. Katoh S, et al. Genetic and environmental effects on fasting and postchallenge plasma glucose and serum insulin values in Finnish twins. J. Clin. Endocrinol. Metab. 2005; 90:2642–2647. [PubMed: 15728197]
42. Hottenga JJ, Whitfield JB, de Geus EJ, Boomsma DI, Martin NG. Heritability and stability of resting blood pressure in Australian twins. Twin Res. Hum. Genet. 2006; 9:205–209. [PubMed: 16611489]
43. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nat. Genet. 2008; 40:695–701. [PubMed: 18509313]
44. Tobin MD, Sheehan NA, Scurrah KJ, Burton PR. Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. Stat. Med. 2005; 24:2911–2935. [PubMed: 16152135]
45. Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 2007; 81:559–575. [PubMed: 17701901]
46. Zaitlen N, Kang HM, Eskin E, Halperin E. Leveraging the HapMap correlation structure in association studies. Am. J. Hum. Genet. 2007; 80:683–691. [PubMed: 17357074]
47. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 2007; 81:1084–1097. [PubMed: 17924348]
48. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. A. 1995; 57:289–300.
49. Donoho D, Jin J. Higher criticism for detecting sparse heterogeneous mixtures. Ann. Statist. 2004; 32:962–994.
50. Bouatia-Naji N, et al. A polymorphism within the G6PC2 gene is associated with fasting plasma glucose levels. Science. 2008; 320:1085–1088. [PubMed: 18451265]

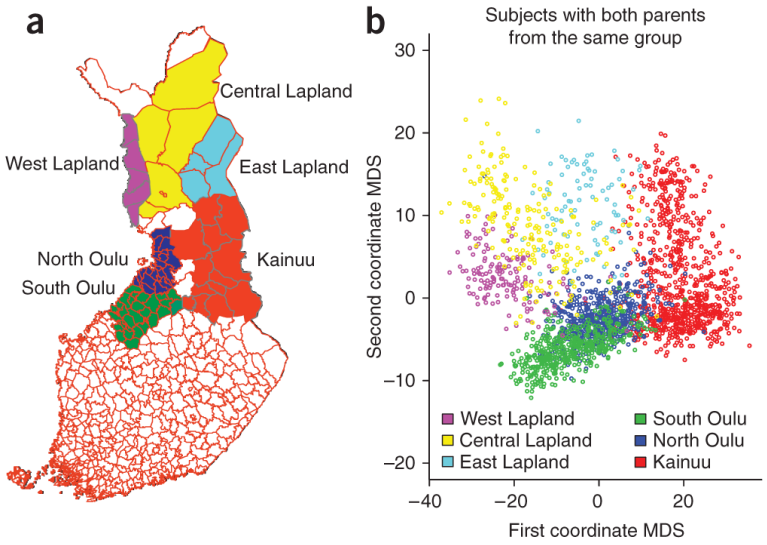

Figure 1.

Linguistic/geographic groups of Northern Finland and their genetic signature. (**a**) Map of Finland with county boundaries. The subjects in NFBC1966 were all born in the two northern provinces. Counties in Northern Finland are color coded to correspond to the six linguistic/geographical groups that can be identified. (**b**) Scatterplot of the two first components identified by MDS on the matrix of genetic similarity between individuals. Only subjects with both parents born in the same population group are plotted, and they are color coded according to the group of origin.

Europe PMC Funders Author Manuscripts

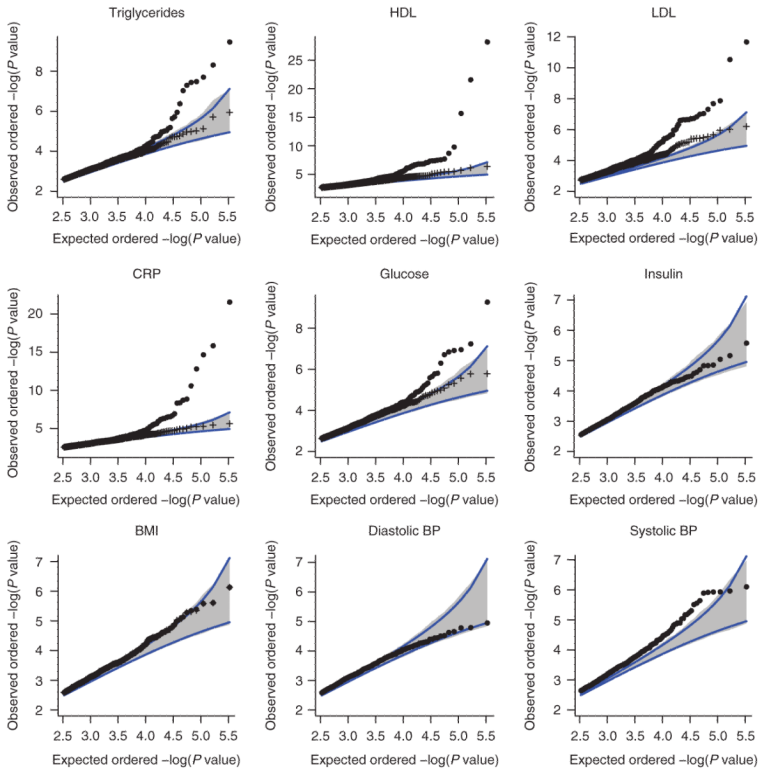

Figure 2.

Quantile-quantile plots of the tails of the *P*-value distribution for the nine traits. Scatter plot of the observed ordered $-\log_{10}$ *P* values versus the $-\log_{10}$ expected ordered *P* values under the complete null for $N = 329{,}091$ tests (solid circles). Only the 1,000 lowest *P* values are considered. Blue lines and gray shaded areas indicate the 0.025 and 0.975 point-wise quantiles of the ordered *P* value under the complete null distribution. The percentiles depicted with blue lines were calculated using a beta approximation for the distribution of ordered statistics of uniform variates and assuming independence across tests. The gray shaded areas were obtained with 1,000 permutations. Crosses indicate the observed ordered $-\log_{10}$ *P* values excluding tests that correspond to a known region or to a newly identified hit. These are also plotted against the $-\log_{10}$ expected ordered *P* values for the same number of tests under the complete null distribution.

Europe PMC Funders Author Manuscripts

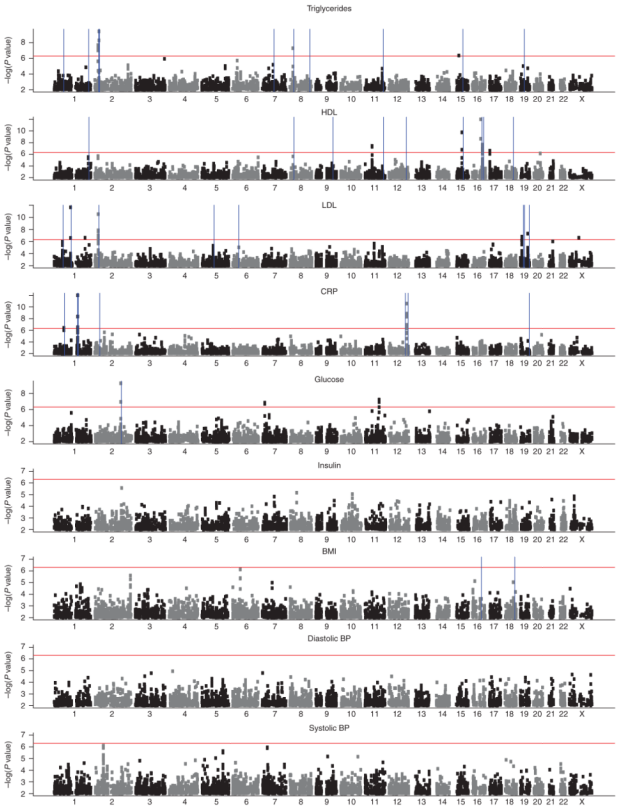

Figure 3.

Association *P* values for genotyped SNPs for the nine traits. Genomic position is on the *x* axis (chromosome number is indicated at the bottom of the plot). The $-\log_{10}$ of the association *P* value is on the *y* axis. Only *P* values $<10^{-2}$ are displayed. To increase readability, the $-\log_{10}$(*P* values) are truncated at 12. The horizontal red line corresponds to a *P* value of 5×10^{-7}. Blue vertical lines indicate position of loci recently identified in GWAS (Table 3).

Europe PMC Funders Author Manuscripts

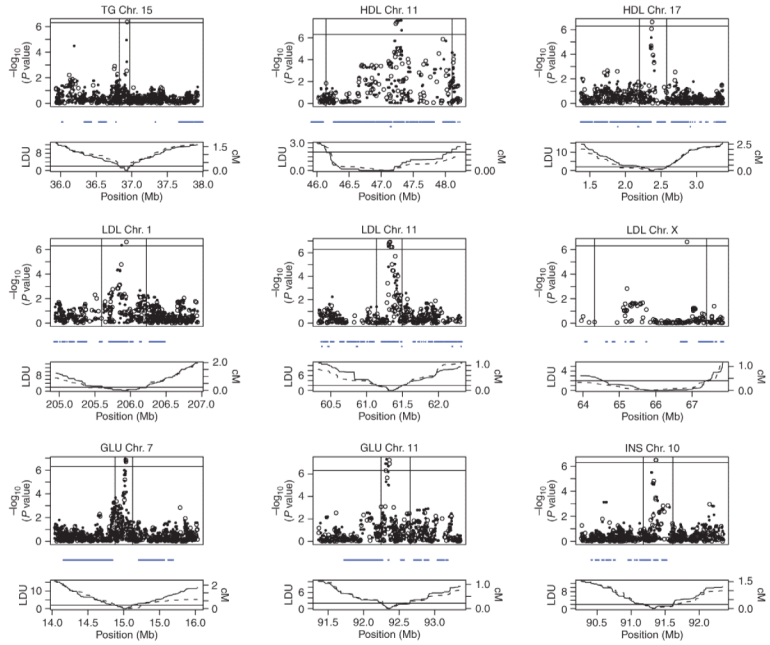

Figure 4.

Association signal in the nine newly identified loci. Association results are shown in relation to RefSeq genes and LD and genetic maps in the region of our findings. For each region the top panel depicts the 2-Mb region around each of our findings. The horizontal line corresponds to $P < 5 \times 10^{-7}$. The vertical lines demarcate the 2 LD unit area around each new finding. Open symbols represent association results from our genotyped SNPs; smaller closed symbols represent association results of nongenotyped, imputed HapMap SNPs using WHAP. The middle panel depicts the location of known RefSeq genes. Genes that overlap are presented on two lines. The bottom panel displays the LD (solid line) and genetic (dashed line) maps. The midpoint of our evidence was set to be zero on both maps. The horizontal line is at 2 LD units. Note that for the HDL locus on chromosome 11 and for the LDL locus on chromosome X LD was so extensive that the 2 LD unit boundaries exceeded a 2-Mb region; therefore, the physical extent of the area plotted for these two regions was increased in order for the 2 LD unit boundaries to be shown.

Europe PMC Funders Author Manuscripts

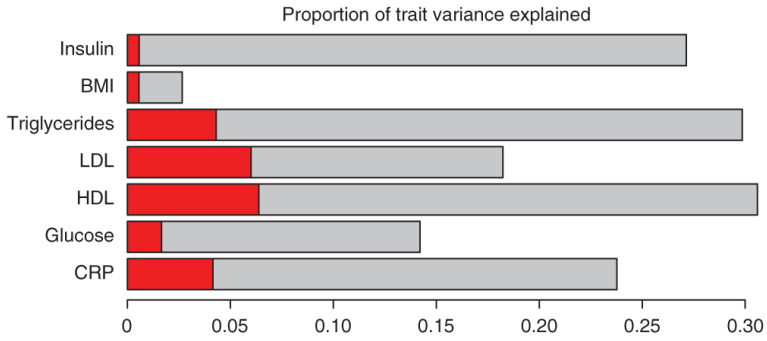

Figure 5.

Graphical representation of the proportion of explained variance for each of the five traits for which genetic loci have been identified. The total explained variance with a model that includes all the genetic loci (either described in **Table** 2 or 3) and the variables listed in Supplementary Table 3 is indicated in gray; the proportion of variance explained by the genetic loci is indicated in red.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Table 1

***P* values for epidemiological variables considered to influence the mean level of nine traits**

Trait	SexOCPG	Smoking	Alcohol	BMI	Birth BMI	Gestational age	C1	C2
TG	4.56E-106	0.00147	1.86E-07	2.88E-125	2.97E-06	0.613	0.297	0.048
HDL	2.19E-150	5.11E-05	6.24E-17	1.89E-78	0.0353	0.893	0.262	0.0915
LDL	1.15E-45	0.312735	0.0145	1.79E-39	0.354	0.177	0.536	1.85E-06
CRP	6.27E-98	0.377	1.28E-05	6.92E-123	0.0135	0.391	0.402	0.722
GLU	1.68E-53	0.083	0.216	7.99E-39	0.069	0.889	0.00707	0.878
INS	7.33E-11	0.521	0.089	2.67E-226	2.79E-05	0.778	0.891	0.0707
BMI	4.39E-16	0.027	0.508	n.a.	0.00332	0.207	0.912	0.585
SBP	3.80E-112	8.12E-08	0.000132	5.60E-103	5.13E-06	0.0225	0.0038	0.00670
DBP	3.09E-43	4.85E-06	5.08E-05	5.26E-98	0.00238	0.506	0.728	0.636

SexOCPG is a factor that accounts for sex, oral-contraceptive use and pregnancy status. Birth BMI was adjusted as described in the text and placed into five quantiles. Gestational age was dichotomized into ≤37 weeks and >37 weeks. C1 and C2 are the first two principal components from an MDS analysis of genome-wide identity-by-state data and are used as a proxy for geographical origin of the subject. Each *P* value was obtained by comparing the residual sum of squares between a model that includes all explanatory variables and a model that omits the variable under consideration.

Europe PMC Funders Author Manuscripts

Nat Genet. Author manuscript; available in PMC 2009 July 01.

Table 2

Loci reaching our significance threshold in the analysis of NFBC traits

Chr.	SNP	Position	Gene	*N*	BETA (s.e.m.)	*P*	GC *P*	MDS *P*	Epi *P*	A1	MAF	Known or new
SNPs associated with mmol/l TG												
2	rs673548	21091049	*APOB*	4,527	−0.081 (0.016)	2.01E-08	2.78E-08	4.69E-08	3.37E-07	A	0.271	K
2	rs1260326	27584444	*GCKR*	4,501	0.093 (0.015)	3.56E-10	5.32E-10	4.70E-10	5.18E-11	A	0.354	K
8	rs10096633	19875201	*LPL*	4,531	−0.123 (0.025)	5.16E-08	7.00E-08	4.78E-08	9.54E-09	A	0.098	K
15	rs2624265	36935941		4,526	0.069 (0.015)	4.31E-07	5.62E-07	7.06E-07	1.00E-04	G	0.419	N
SNPs associated with mmol/l HDL												
11	rs2167079	47226831	*NR1H3*	4,449	0.04 (0.0074)	5.13E-08	9.48E-08	2.23E-08	3.52E-07	A	0.417	N
11	rs7120118	47242866	*NR1H3*	4,525	0.04 (0.0073)	3.57E-08	6.70E-08	1.53E-08	2.30E-07	G	0.417	N
15	rs1532085	56470658	*LIPC*	4,529	0.047 (0.0074)	1.77E-10	4.08E-10	3.05E-10	2.74E-11	A	0.443	K
16	rs3764261	55550825	*CETP*	4,527	0.092 (0.0081)	6.97E-29	8.44E-28	7.81E-29	8.29E-29	A	0.275	K
16	rs255049	66570972	*LCAT*	4,326	0.05 (0.0009)	3.06E-08	5.78E-08	5.98E-08	2.06E-07	G	0.220	K
17	rs9891572	2375258		4,525	0.052 (0.0099)	2.33E-07	4.06E-07	2.14E-07	1.84E-05	A	0.163	N
SNPs associated with mmol/l LDL												
1	rs646776	109620053	*CELSR2-PSRC1-SORT1*	4,507	−0.155 (0.022)	2.19E-12	1.32E-11	6.59E-13	7.24E-11	G	0.213	K
1	rs4844614	205941798	*CR1L*	4,472	0.1 (0.019)	2.38E-07	6.42E-07	2.11E-07	1.09E-06	A	0.321	N
2	rs693	21085700	*APOB*	4,512	0.123 (0.018)	2.99E-11	1.50E-10	2.32E-11	3.20E-12	A	0.412	K
11	*rs174537*	61309256	*FADS1-FADS2*	4,270	−0.094 (0.018)	2.10E-07	6.29E-07	3.85E-08	8.97E-08	A	0.434	N
11	*rs102275*	61314379	*FADS1-FADS2*	4,270	−0.095 (0.018)	1.52E-07	4.66E-07	2.77E-08	6.73E-08	G	0.436	N
11	*rs174546*	61326406	*FADS1-FADS2*	4,268	−0.096 (0.018)	1.30E-07	4.04E-07	2.25E-08	5.33E-08	A	0.436	N
11	*rs174556*	61337211	*FADS1-FADS2*	4,273	−0.092 (0.018)	3.49E-07	1.01E-06	6.38E-08	1.34E-07	A	0.412	N
11	*rs1535*	61354548	*FADS1-FADS2*	4,273	−0.092 (0.018)	3.65E-07	1.05E-06	6.76E-08	1.60E-07	G	0.435	N
19	rs11668477	11056030	*LDLR*	4,494	−0.127 (0.024)	1.51E-07	4.21E-07	2.55E-07	7.17E-09	G	0.174	K
19	rs157580	50087106	*APO cluster*	4,514	−0.109 (0.02)	4.96E-08	1.49E-07	5.90E-08	1.14E-08	G	0.292	K
X	rs5031002	66859350	*AR*	4,518	0.296 (0.057)	2.37E-07	6.38E-07	4.77E-07	3.42E-07	A	0.017	N
SNPs associated with mg/l CRP												
1	rs12753193	65942267	*LEPR*	4,739	−0.126 (0.075)	3.76E-07	4.29E-07	3.57E-07	5.14E-06	G	0.448	K
1	rs2794520	157945440	*CRP*	4,724	−0.466 (0.077)	2.92E-22	4.65E-22	3.21E-22	1.30E-19	A	0.361	K
12	rs2650000	119873345	*LEF1*	4,734	−0.398 (0.075)	2.71E-11	3.39E-11	1.64E-11	1.08E-08	A	0.451	K

Europe PMC Funders Author Manuscripts

Chr.	SNP	Position	Gene	*N*	BETA (s.e.m.)	*P*	GC *P*	MDS *P*	Epi *P*	A1	MAF	Known or new
SNPs associated with mmol/l GLU												
2	rs560887	169471394	*G6PC2-ABCB11*	4,308	−0.056 (0.009)	5.69E-10	1.21E-09	3.41E-10	4.25E-10	A	0.306	K
7	rs10244051	15030358		4,314	0.047 (0.0087)	1.45E-07	2.52E-07	3.83E-07	1.51E-06	C	0.458	N
7	rs2191348	15030780		4,311	0.046 (0.0087)	1.99E-07	3.42E-07	5.06E-07	1.90E-06	A	0.458	N
11	rs1447352	92362409	*MTNR1B*	4,312	−0.047 (0.009)	5.90E-08	1.06E-07	8.60E-08	8.35E-08	G	0.422	N
11	rs7121092	92363999	*MTNR1B*	4,249	−0.046 (0.009)	1.23E-07	2.15E-07	1.72E-07	1.52E-07	A	0.419	N
SNPs associated with pmol/l INS												
10	*rs11185790*	91362513	*PANK1*	4,196	−0.312 (0.066)	3.31E-07	3.62E-07	6.9E-07	5.38E-06	A	0.206	N
10	*rs1075374*	91366955	*PANK1*	4,200	−0.31 (0.066)	3.32E-07	3.63E-07	6.9E-07	5.72E-06	A	0.206	N

Position in base pairs is from NCBI build 36. BETA is the effect on the untransformed trait for each copy of the minor allele (A1). The name of the nearest gene, if any, is indicated. *N* is the number of individuals included in the association analysis of the SNP and the trait. *P* is the *P* value from the primary test of association, GC *P* is the genomic control-corrected *P* value and MDS *P* is the *P* value from an association analysis that includes the first two principal components from an MDS analysis of population stratification. Epi *P* is the *P* value from an association analysis that includes relevant epidemiological covariates from Table 1 in the model. K indicates a previously known association; N indicates a new association discovered in our analyses. For previously known genes, only the most significant SNP in our study is presented. For new findings, all significant SNPs are presented. The r^2 between SNPs in the same gene was >0.90. SNPs in italics were identified after correcting traits for BMI at age 31.

Europe PMC Funders Author Manuscripts

Table 3

Previously reported associations and evidence for association in NFBC1966

Chr.	Gene	Previous SNP	Allele/ Freq.	Previous effect size (*n*)	NFBC SNP[a]	NFBC *P*	NFBC effect size (s.e.m., *n*)	NFBC min *P*	Empirical *P*
SNPs associated with mmol/l TG									
1	*ANGPTL3-DOCK7-ATG4C*	rs1748195[b]	G/0.30	−0.08 (9,559)	rs1167998 (1) G	1.60E-04	−0.05 (0.016, 4,520)	1.60E-04	0.0066
1	*ANGPTL3-DOCK7-ATG4C*	rs12130333[c]	T/0.22	−0.13 (18,554)	G	3.60E-03	−0.05 (0.021, 4,531)	1.60E-04	0.0066
1	*GALNT2*	rs4846914[c]	G/0.40	0.09 (18,554)	I	2.80E-01	0.001 (0.017, 4,282)	5.66E-03	0.1446
2	*APOB*	rs693[c]	A/0.48	0.09 (18,554)	G	3.40E-03	0.04 (0.015, 4,525)	2.10E-08	0
2	*GCKR*	rs780094[c]	T/0.34	0.15 (18,554)	G	5.03E-09	0.09 (0.015, 4,527)	3.56E-10	0
7	*BCL7B-TBL2-MLXIPL*	rs17145738[c]	T/0.13	−0.16 (18,554)	I	2.70E-05	−0.44 (0.102, 4,531)	6.55E-06	0
8	*LPL*	rs328[c]	G/0.09	−0.22 (18,554)	I	4.50E-08	−0.13 (0.025, 4,531)	5.16E-08	0
8	*TRIB1*	rs17321515[c]	G/0.49	−0.09 (18,554)	rs6982636 (1) G	1.43E-02	−0.05 (0.015, 4,529)	1.43E-02	0.3066
11	*APOA1/C3/A4/A5, ZNF259, BUD13*	rs12286037[b]	T/0.06	0.29 (9,738)	rs12292921 (1) G	1.35E-03	0.1 (0.031, 4,517)	2.07E-05	0.0018
15	*LIPC*	rs4775041[b]	C/0.33	0.04 (8,462)	I	5.85E-01	0.032 (0.028, 4,531)	4.36E-04	0.0132
19	*NCAN-CILP2-PBX4*	rs16996148[c]	T/0.10	−0.12 (18,554)	I	8.40E-01	0.002 (0.029, 4,531)	9.71E-06	0.0004
SNPs associated with mmol/l HDL									
1	*GALNT2*	rs4846914[c]	G/0.40	−0.028 (18,554)	I	5.60E-04	−0.029 (0.0085, 4,284)	2.86E-06	0
8	*LPL*	rs328[c]	G/0.09	0.068 (18,554)	I	7.18E-06	0.06 (0.013, 4,532)	2.25E-06	0
9	*ABCA1*	rs4149268[b]	T/0.36	−0.021 (11,327)	rs2740491 (1) I	3.12E-04	−0.03 (0.008, 4,532)	3.38E-05	0.0012
9	*ABCA1*	rs3890182[c]	A/0.13	−0.04 (18,554)	rs3847303 (1) G	3.22E-03	−0.03 (0.012, 4,531)	3.38E-05	0.0012
11	*APOA1/C3/A4/A5, ZNF259, BUD13*	rs28927680[c,d]	G/0.07	−0.052 (18,554)				1.20E-04	0.0066
12	*MVK-MMAB*	rs2338104[b]	G/0.45	0.012 (11,399)	I	4.40E-02	0.02 (0.007, 4,526)	5.90E-04	0.0182
15	*LIPC*	rs4775041[b]	C/0.33	0.036 (11,426)	I	1.70E-02	0.02 (0.014, 4,532)	1.77E-10	0
15	*LIPC*	rs1800588[c]	T/0.21	0.056 (18,554)	I	4.11E-03	0.04 (0.017, 4,527)	1.77E-10	0
16	*CETP*	rs3764261[b]	A/0.31	0.09 (8,072)	G	6.97E-29	0.09 (0.008, 4,527)	6.97E-29	0
16	*CETP*	rs1800775[c]	C/0.51	−0.072 (18,554)	I	1.32E-09	−0.067 (0.011, 4,531)	6.97E-29	0

Europe PMC Funders Author Manuscripts

Chr.	Gene	Previous SNP	Allele/ Freq.	Previous effect size (*n*)	NFBC SNP[a]	NFBC *P*	NFBC effect size (s.e.m., *n*)	NFBC min *P*	Empirical *P*
16	*LCAT*	rs255052[b]	A/0.17	0.019 (4,534)	G	2.36E-07	0.05 (0.01, 4,416)	3.06E-08	0
18	*LIPG-ACAA2*	rs2156552[c]	A/0.18	−0.028 (18,554)	I	1.15E-02	−0.035 (0.014, 4,455)	1.80E-04	0.0112
SNPs associated with mmol/l LDL									
1	*PCSK9*	rs11206510[c]	T/0.01	−0.47 (18,554)	rs12117661 (0.57) I	1.95E-12	−0.156 (0.022, 4,486)	3.07E-05	0.0016
1	*CELSR2-PSRC1- SORT1*	rs646776[c]	C/0.24	−0.16 (18,554)	G	2.19E-12	−0.16 (0.022, 4,507)	2.19E-12	0
2	*APOB*	rs693[c]	A/0.48	0.12 (18,554)	G	2.99E-11	0.12 (0.018, 4,517)	2.99E-11	0
2	*APOB*	rs562338[b]	A/0.18	−0.126 (10,849)	I	1.97E-07	−0.148 (0.028, 4,517)	2.99E-11	0
5	*HMGCR*	rs12654264[c]	T/0.39	0.1 (18,554)	I	2.63E-05	0.094 (0.022, 4,511)	4.95E-06	0.0002
6	*B3GALT4*	rs2254287[b]	G/0.38	0.049 (7,440)	I	6.98E-01	0.01 (0.018, 4,517)	1.04E-02	0.257
19	*LDLR*	rs6511720[c]	T/0.10	−0.26 (18,554)	I	1.48E-09	−0.16 (0.026, 4,515)	1.51E-07	0
19	*NCAN-CILP2-PBX4*	rs16996148[c]	T/0.10	−0.1 (18,554)	I	8.20E-01	−0.01 (0.037, 4,518)	1.48E-03	0.158
19	*APO cluster*	rs4420638[c]	G/0.20	0.19 (18,554)	rs2075650 (0.37) G	1.05E-05	0.11 (0.026, 4,513)	4.96E-08	0
SNPs associated with log (mg/l) CRP									
1	*LEPR*	rs12753193[e]	G/0.38	−0.154 (6,345)	G	3.76E-07	−0.13 (0.031, 4,739)	3.76E-07	0
1	*IL6R*	rs4129267[e]	A/0.398	−0.101 (6,345)	G	4.35E-04	−0.12 (0.034, 4,742)	4.35E-04	0.02
1	*CRP*	rs1800947[e]	G/0.065	−0.283 (6,345)	rs11588887 (0.59) I	4.92E-08	0.325 (0.059, 4,744)	2.92E-22	0
2	*GCKR*	rs780094[e]	A/0.40	0.14 (6,345)	G	1.50E-01	0.05 (0.032, 4,740)	7.30E-02	0.4986
12	Unknown	rs10778213[e]	A/0.71	0.115 (6,345)	G	2.28E-02	0.07 (0.032, 4,744)	1.58E-02	0.2432
12	*LEF1*	rs1169300[e]	A/0.295	−0.146 (6,345)	G	2.01E-09	−0.2 (0.033, 4,739)	2.71E-11	0
19	*APOE*	rs2075650[e]	G/0.14	−0.209 (6,345)	G	4.40E-03	−0.12 (0.042, 4,739)	6.64E-05	0.0004
SNPs associated with mmol/l GLU									
2	*G6PC2-ABCB11*	rs560887[f]	A/0.30	−0.06 (9,353)	G	5.69E-10	−0.06 (0.009, 4,308)	5.69E-10	0
SNPs associated with kg/m^2 BMI									
16	*FTO*	rs9939609[g]	A/0.39	0.36 (38,759)	rs3751812 (1) G	2.41E-04	0.36 (0.09, 4,490)	2.41E-04	0.0072
18	*MC4R*	rs17782313[h]	C/0.24	0.22 (16,876)	I	1.79E-03	0.33 (0.118, 4,500)	1.92E-04	0.01

We show results for evaluation of SNP replication (columns NFBC SNP, NFBC P and NFBC effect) as well as evaluation of locus replication (last two columns). NFBC *P* is the *P* value from the association test of the given SNP (or its proxy) in NFBC. Locus replication was evaluated by identifying the minimum *P* value (NFBC min *P*) in a 4-LD-unit window around the previously reported SNP

Europe PMC Funders Author Manuscripts

and comparing that *P* value to the distribution of *P* values obtained in 5,000 permutations under the null hypothesis of no association (Empirical *P*). Gene/outcome combinations are listed twice when the r^2 between SNPs reported to be associated to the same region in different previous studies was <0.80.

[a] If the previously reported SNP was genotyped or imputed in NFBC1966, 'G' or 'I', respectively, appears in the column NFBC SNP; otherwise, the best proxy SNP (r^2) and an indicator of whether the proxy was genotyped (G) or imputed (I) is given.

[b] From ref. 21.

[c] From ref. 17.

[d] The tested SNP rs28927680 was not in HapMap. The best proxy SNP reported by ref. 22, rs3133506, was monomorphic in HapMap. We are unable to identify a proxy SNP in NFBC for either rs28927680 or rs3133506; SNP replication of this association is not possible.

[e] From ref. 19.

[f] From ref. 50.

[g] From ref. 16.

[h] From ref. 18.