Meta-Analysis of Gene Level Tests for Rare Variant Association

Dajiang J. Liu[1*], Gina M. Peloso[2,3,4*], Xiaowei Zhan[1*], Oddgeir L. Holmen[5,6*] Matthew Zawistowski[1], Shuang Feng[1], Majid Nikpay[7], Paul L. Auer[8,9], Anuj Goel[10,11], He Zhang[12,13], Ulrike Peters[8,14], Martin Farrall[10,11], Marju Orho-Melander[11,15], Charles Kooperberg[8,16], Ruth McPherson[7], Hugh Watkins[10,11], Cristen J. Willer[12,13], Kristian Hveem[5,17], Olle Melander[11,15], Sekar Kathiresan[2,3,4,18+] and Gonçalo R. Abecasis[1,+]

[1]Center for Statistical Genetics, Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI 48109;

[2]Broad Institute of Harvard and MIT, Cambridge, MA;

[3]Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA

[4]Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA;

[5]Department of Public Health and General Practice, Norwegian University of Science and Technology, Trondheim 7489, Norway;

[6]St. Olav Hospital, Trondheim University Hospital, Trondheim, Norway

[7]University of Ottawa Heart Institute, Ottawa, Ontario, Canada;

[8]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle WA 98109, USA;

[9]School of Public Health, University of Wisconsin-Milwaukee

[10]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, United Kingdom;

[11]Department of Cardiovascular Medicine, University of Oxford, Oxford, UK;

[12]Division of Cardiology, Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, MI 48109;

[13]Department of Human Genetics, University of Michigan Medical School, Ann Arbor, MI 48109;

[14]Department of Epidemiology, University of Washington School of Public Health, Seattle, WA;

[15]Department of Clinical Sciences, Lund University, Malmö, Sweden;

[16]Department of Biostatistics, University of Washington School of Public Health, Seattle, WA;

[17]Levanger Hospital, Levanger, Norway

[18]Harvard Medical School, Cambridge, MA

Nature Genetics: doi:10.1038/ng.2852

* These authors contributed equally and should be considered joint first authors.

+ These authors jointly directed the study.

Correspondence To:

Dajiang J. Liu (dajiang@umich.edu)

Or

Gonçalo R. Abecasis (goncalo@umich.edu)

Center for Statistical Genetics,

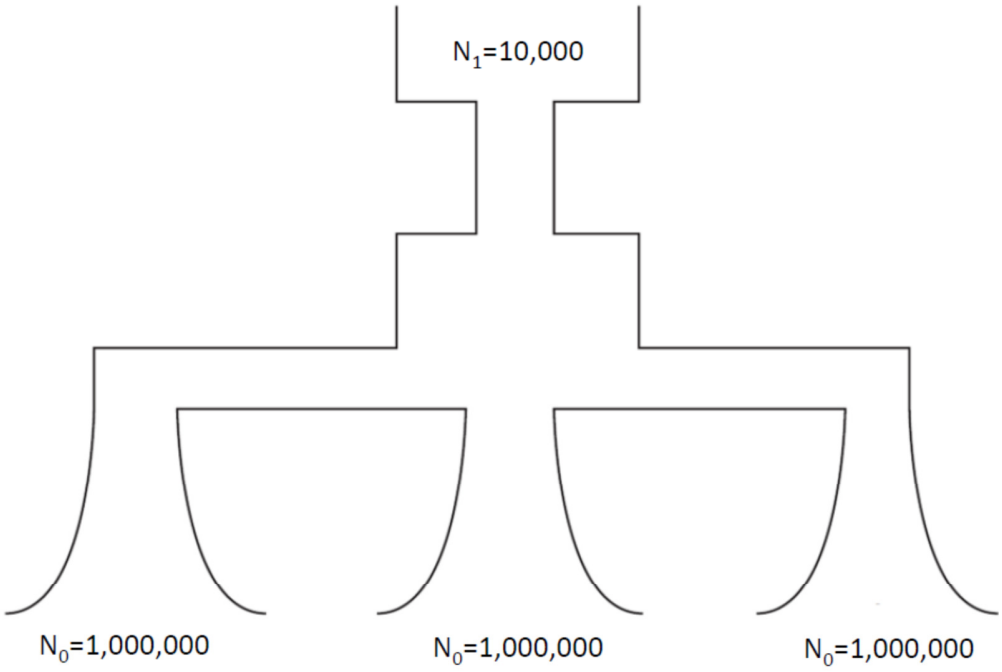
Department of Biostatistics,

1415 Washington Heights,

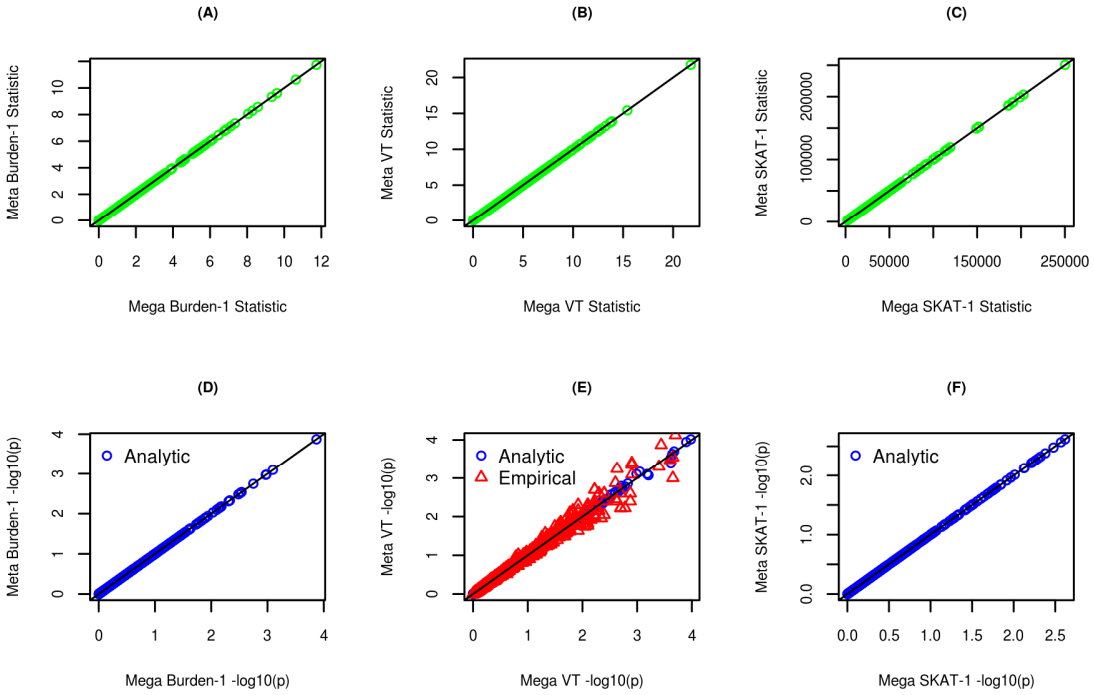
Ann Arbor, MI 48109

Supplementary Figures

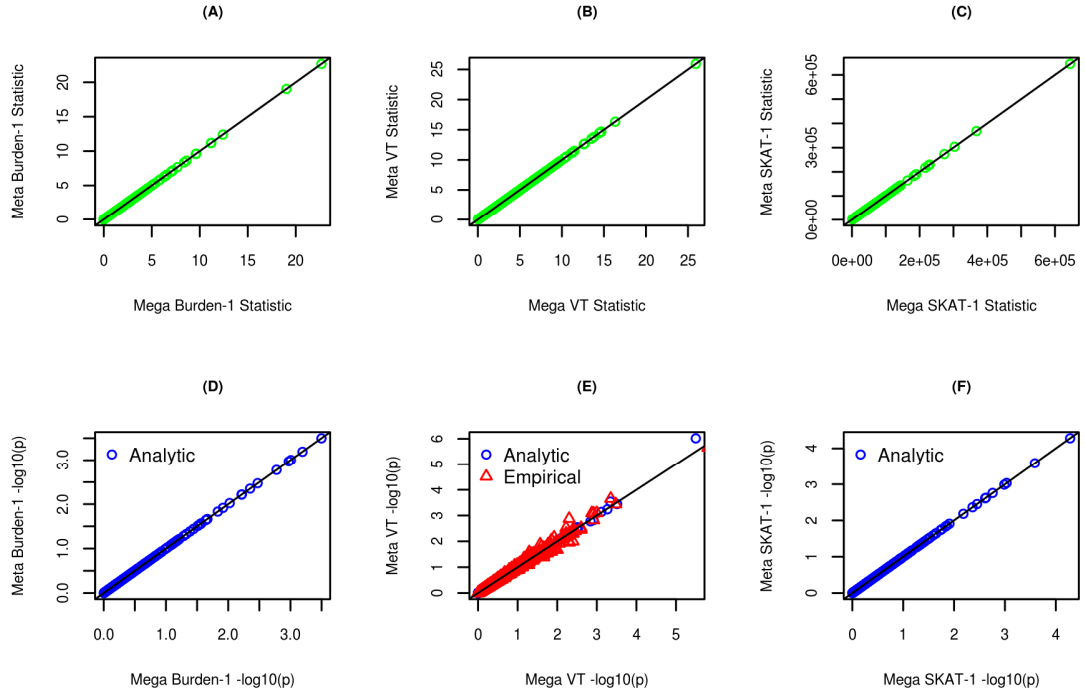
Supplementary Figure 1: Demographic model for simulated European populations. The demographic model includes an ancient population bottleneck, recent exponential growth, differentiation and migration. The model parameters were calibrated to mimic populations sampled in continental Europe.



Supplementary Figure 2: Comparison of test statistics and p-values in meta-analysis and in joint analysis of individual data that allows for heterogeneity in nuisance parameters. We considered three different association tests: a simple burden test grouping variants with MAF < 1% (burden-1) (Panel A and D), variable threshold tests (VT) (Panel B and E) and tests allowing for variants with opposite effects (SKAT) (Panel C and F). Three samples of 1000 European-ancestry individuals were simulated. Traits were simulated assuming that 50% of the variants in the gene region are causal and that each causal variant increases trait means by 0.125 standard deviations. Scatter plots compare our meta-analysis statistics (top row) and p-values (bottom row) with those calculated after pooling individual level data. For the VT test (Panel E), we compare both asymptotic (blue circles) and empirical p-values (red triangles). Empirical p-values were obtained using our Monte-Carlo procedure to generate replicates until 100 simulated statistics exceeded the original observation or 40,000,000 statistics were simulated.

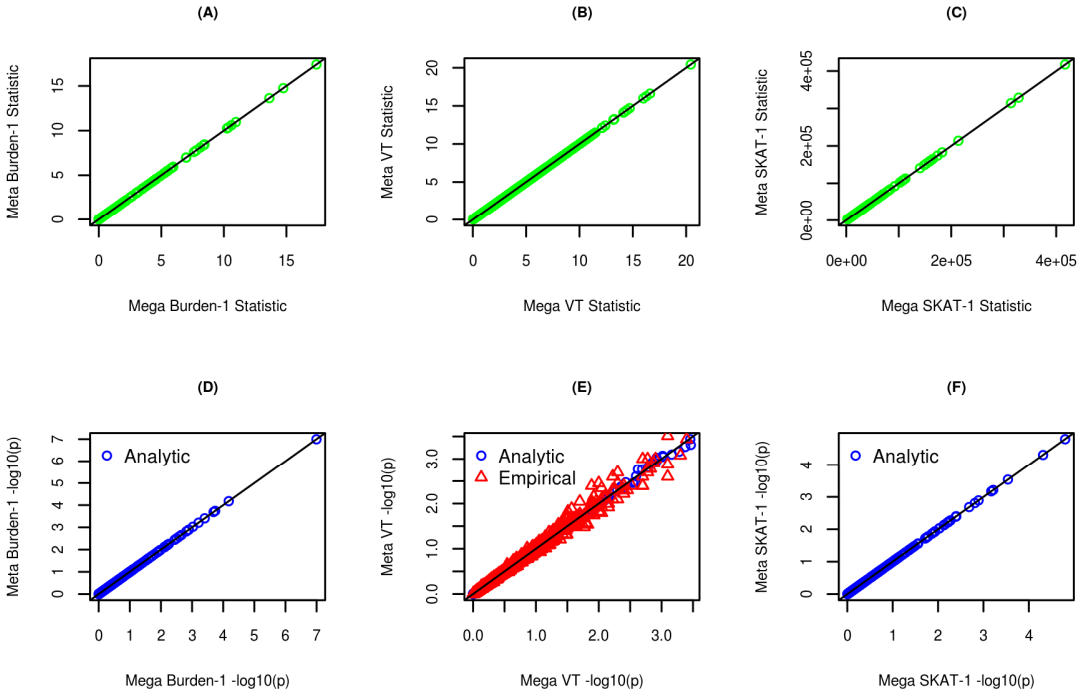


Supplementary Figure 3: Comparison of statistics and p-values for a simple burden test, variable threshold (VT) and sequence kernel association test (SKAT) in analysis of pooled samples (X-axis) and in meta-analysis (Y-axis). Three samples of 1000 individuals were simulated. The figure is analogous to **Supplementary Figure 1** but allows for variants with opposite phenotypic effects to reside within each gene. Traits were generated assuming that 50% of the variants are causal and that, among these, 80% of the variants increase the trait values by 0.25 standard deviation units and the remaining 20% decrease trait values by the same amount.

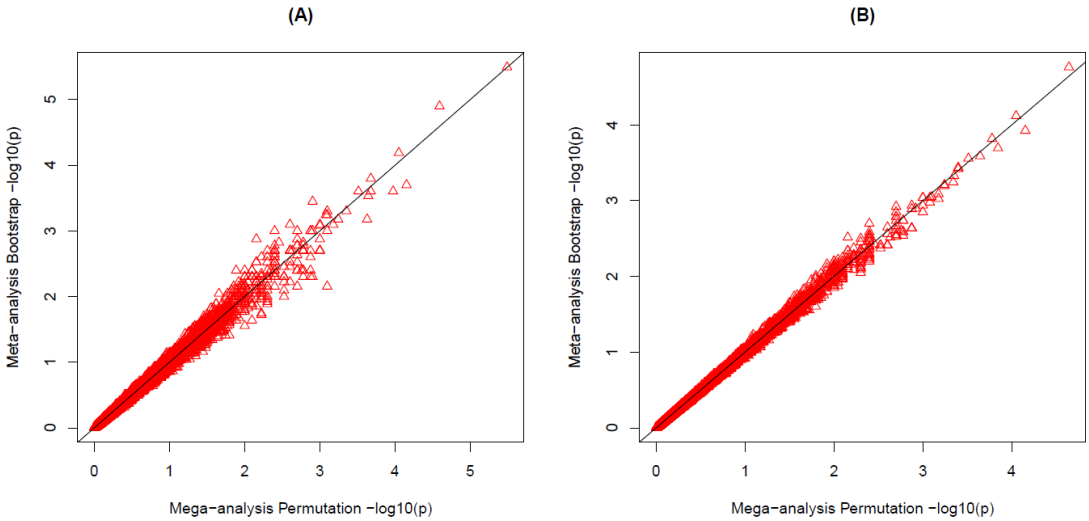


.

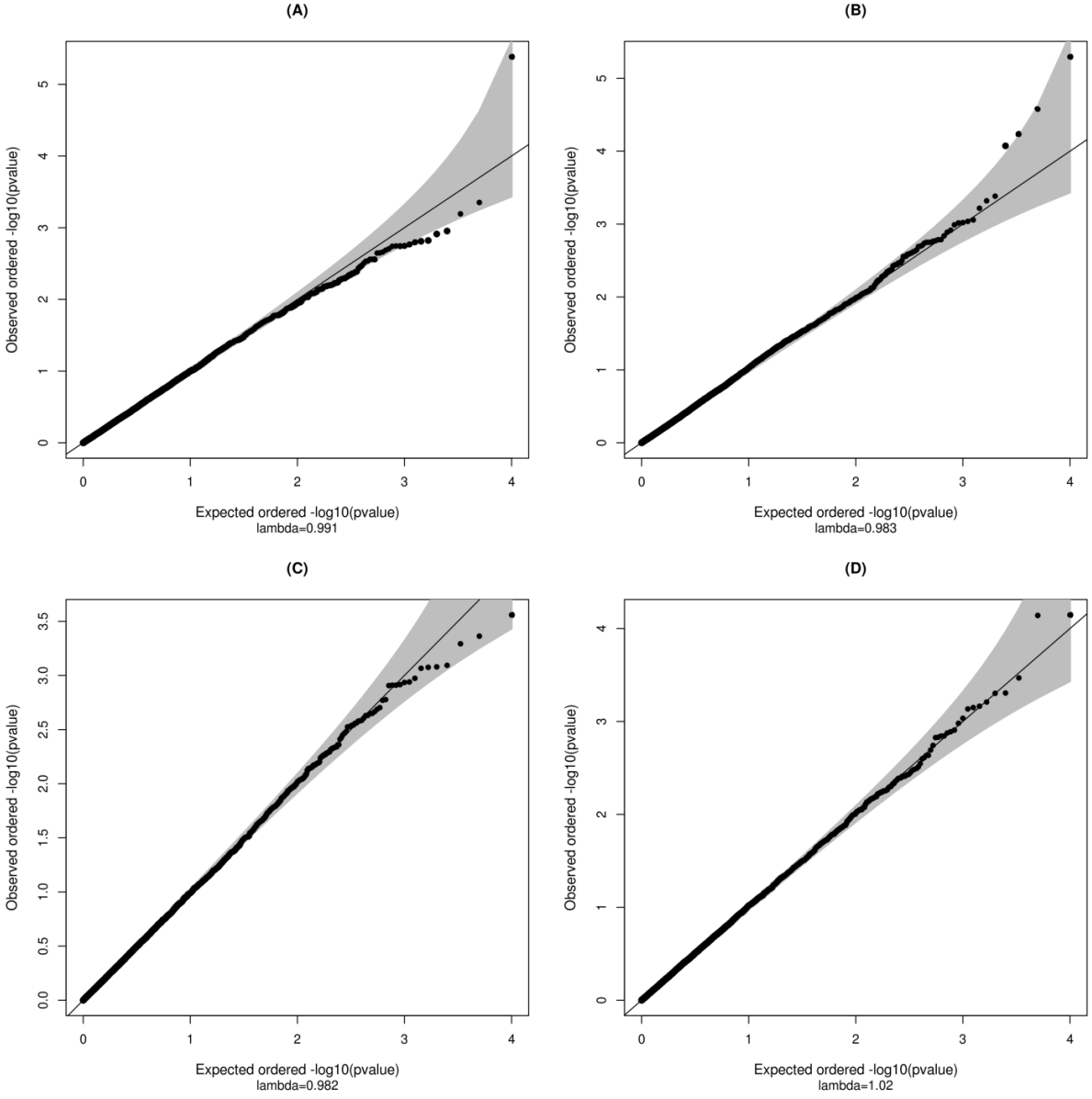
Supplementary Figure 4: Comparison of statistics and p-values for a simple burden test, variable threshold (VT) and sequence kernel association test (SKAT) in analysis of pooled samples (X-axis) and in meta-analysis (Y-axis). This Figure is analogous to **Supplementary Figures** 2 and 3, but assumes a random effect for each causal variant, distributed as Normal(0.25, 0.01) in standard deviation units.



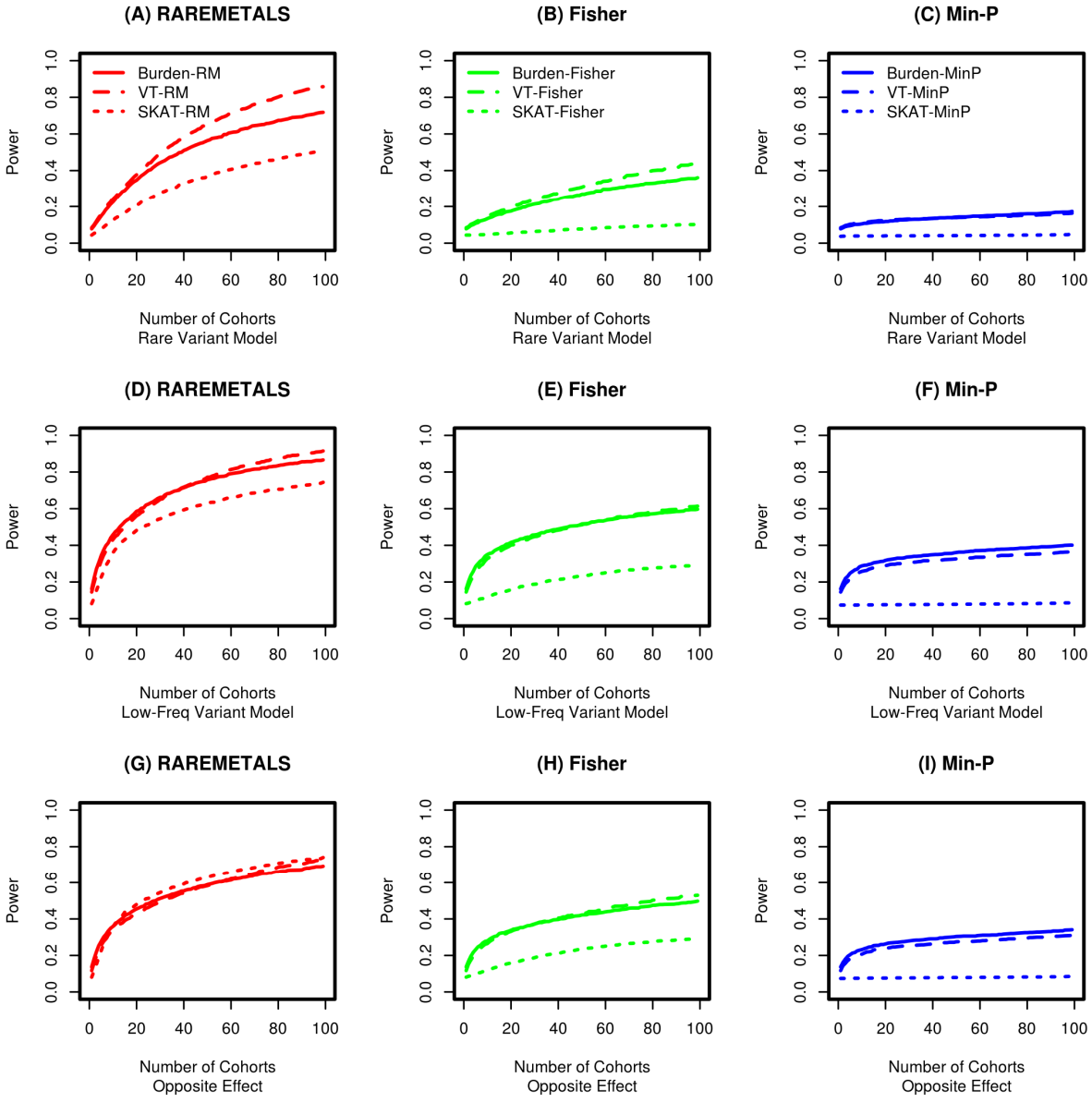
Supplementary Figure 5: Comparison of our Monte-Carlo estimates of p-values in meta-analysis with permutation-based p-values estimated after pooling individual level data. Traits were generated assuming that 50% of the variants are causal and that each causal variant increases the trait values by 0.25 standard deviation units. Empirical p-values in Panel A were obtained using the adaptive Monte Carlo procedure in meta-analysis and using permutation in mega-analysis, stopping after 100 simulated statistics were greater than the original statistic or the number of simulations exceeded 40,000,000. In panel B, we stopped only after 400 simulated statistics exceeded the original observation or 160,000,000 simulations were carried out (demonstrating that accuracy our p-value estimates increases as more replicates are generated, as expected).



Supplementary Figure 6: Evaluation of our method in settings where there are many very rare alleles. These quantile-quantile plots show that p-values calculated using our approach are well calibrated even when only singleton alleles are analyzed. We simulated quantitative phenotypes from a standard normal distribution. We then analyzed the association between these simulated phenotypes and five singleton variants in a sample of 500 individuals. The empirical distribution of $-\log_{10}$ transformed p-values was obtained using 10,000 replicates and plotted against their theoretical expectations. Results are displayed for (A) single site association test statistics, (B) burden tests, (C) the variable threshold test and (D) the SKAT test.

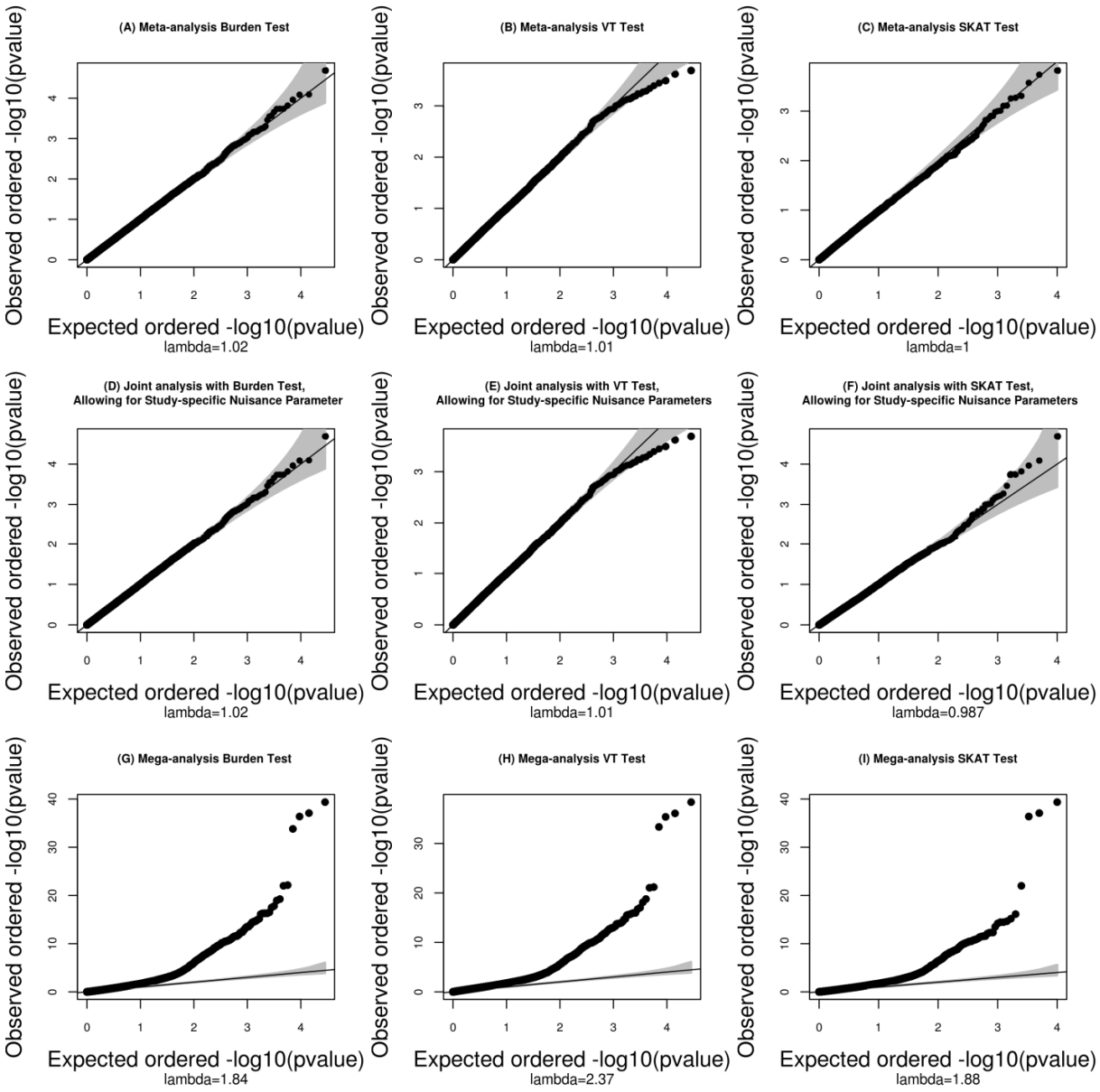


Supplementary Figure 7: Power comparison for our approach, Fisher's method and the minimal p-value approach. Three phenotype models were simulated: half of low frequency variants with MAF < 0.5% are causal, each increasing expected trait values by 1/4 standard deviation (Panels A, B, C); half of all variants are causal, irrespective of frequency, and increase trait values by 1/4 standard deviation (Panels D, E, F); 50% of the variants are casual, irrespective of frequency, and 80% of these increase expected trait values by 1/4 standard deviation, while the remaining 20% decrease trait values by the same amount (Panels G, H, I). Between 2 and 100 studies of 1000 individuals each were simulated. Meta-analysis for variants with MAF<5% was performed using our approach or using Fisher's method and the minimal p-value approach to combine burden test, SKAT and variable threshold (VT) test statistics. Power was evaluated at threshold $\alpha=2.5\times10^{-6}$ using 10,000 replicates. Note that differences between our approach and these alternatives become more marked when more studies are meta-analyzed.

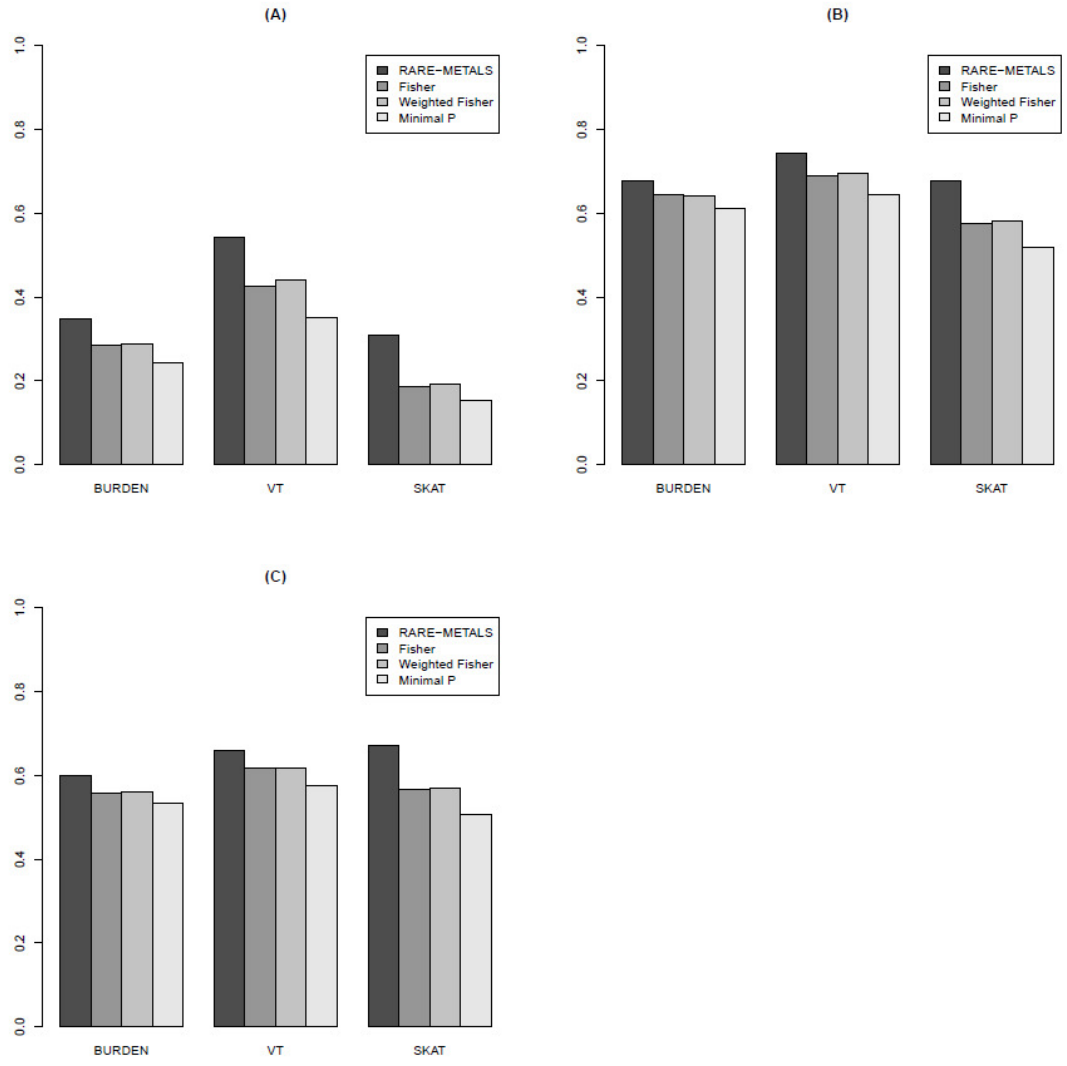


Nature Genetics: doi:10.1038/ng.2852

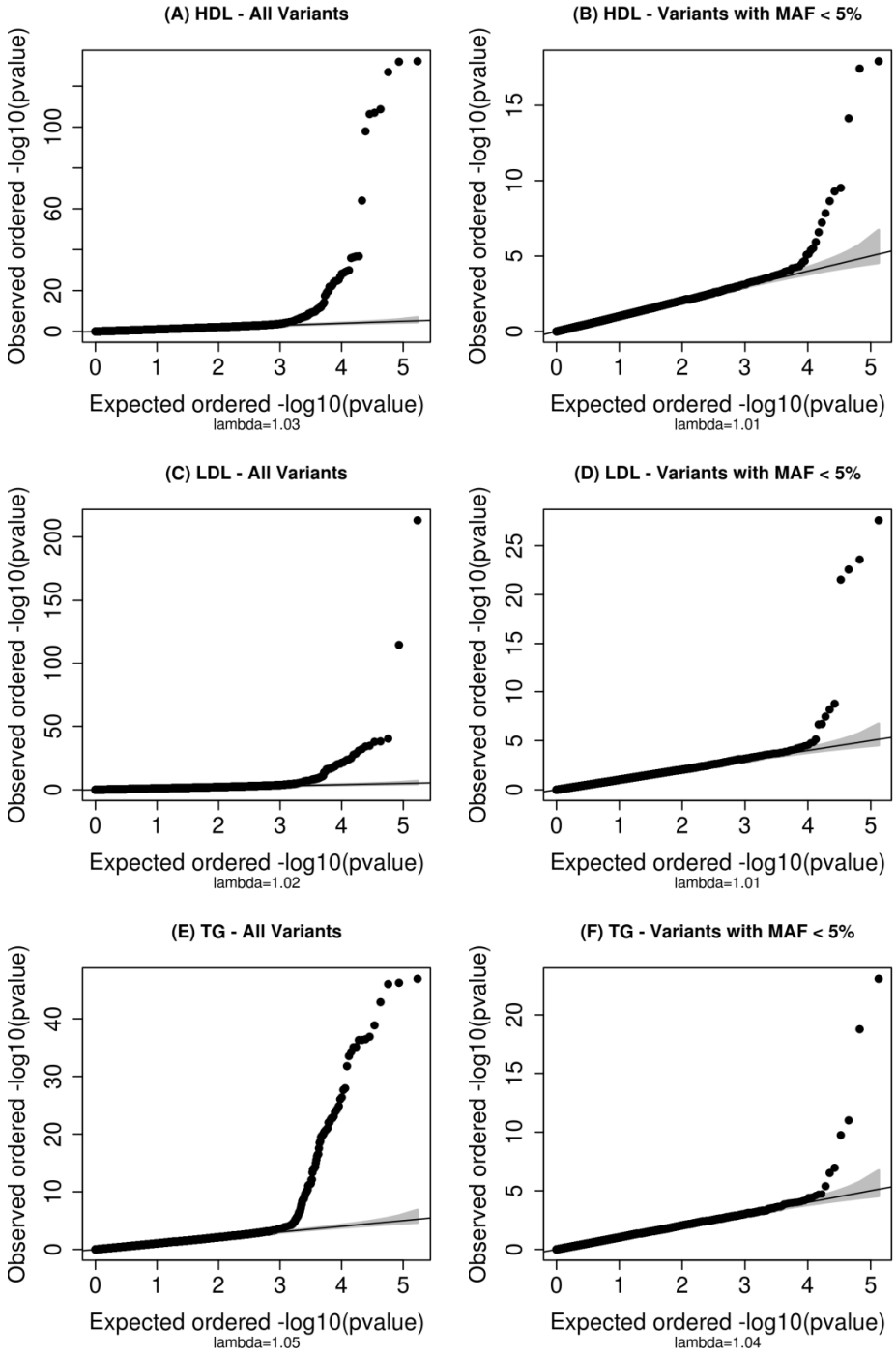
Supplementary Figure 8: Comparison of meta-analysis and analysis of pooled individual data in the presence of between study heterogeneity. We evaluated three analysis plans using genotype data from the MDC and HUNT studies and simulated phenotype under the null hypothesis. For the MDC study, we simulated phenotypes from the distribution N(0,1) and for HUNT study, we simulated phenotypes from a distribution with a shifted mean value, N(0.2,1). Gene-level association analysis was performed for variants with MAF < 5% in each gene using burden (Panels A, D, G), variable threshold (Panels B, E, H) and SKAT tests (Panels C, F, I). We evaluate our meta-analysis approach (panels A-C), and analyses of pooled data that allow for study specific nuisance parameters (panels D-F), both of which performed well. We also considered a naïve analysis strategy, where analysis proceeded directly after pooling of individual level data without using study specific nuisance parameters (G-I), this resulted in markedly inflated test statistics.



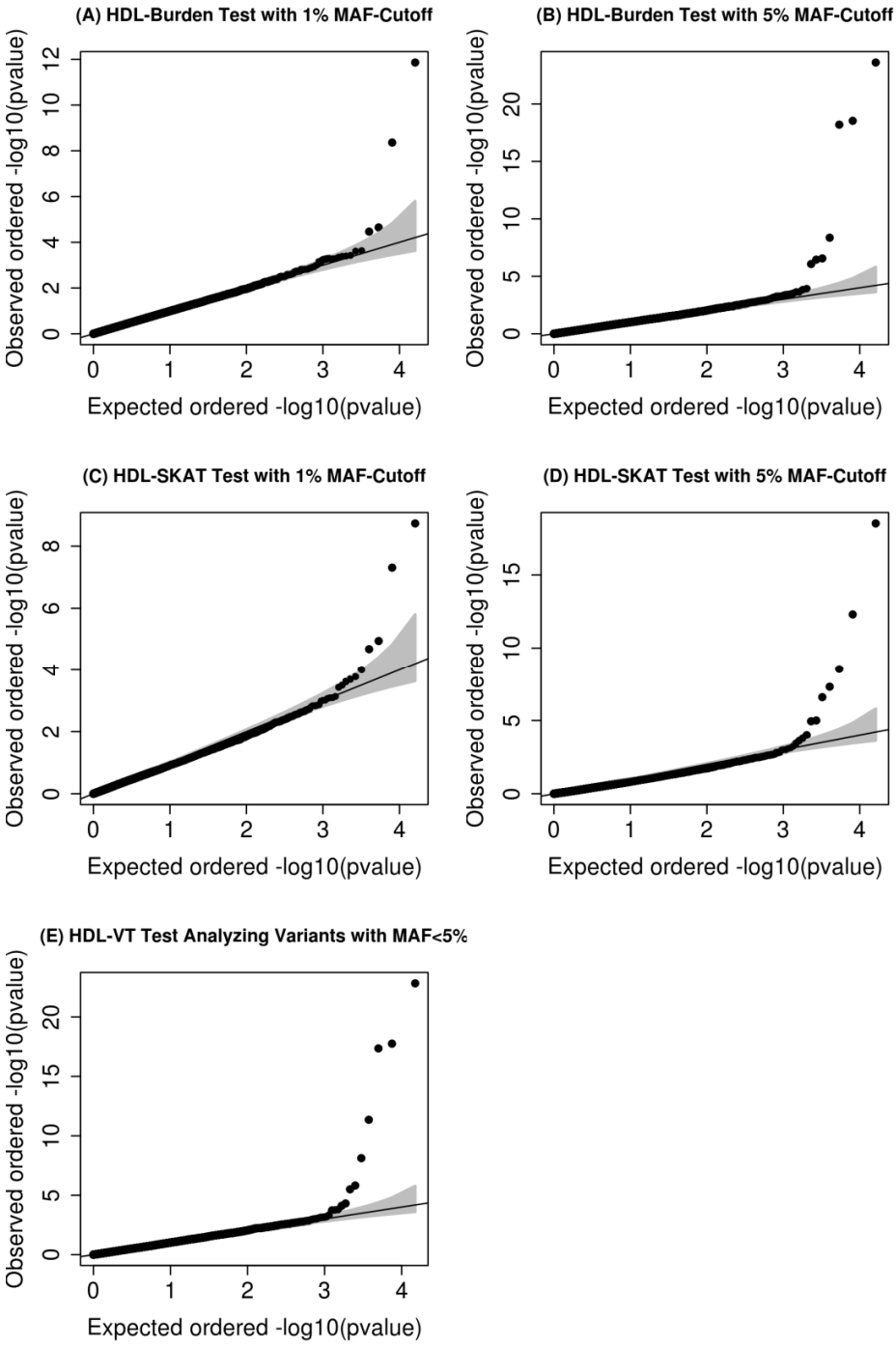
Supplementary Figure 9: Power comparison for our approach, Fisher's method, weighted Fisher's method and the minimal p-value approach. Power was compared using genotype data from the MDC and HUNT studies. In each replicate, one gene from the dataset was analyzed. Three phenotype models were simulated: (A) half of low frequency variants with MAF < 0.5% are causal, each increasing expected trait values by 1/4 standard deviation; (B) half of all variants are causal, irrespective of frequency, and increase trait values by 1/4 standard deviation; (C) 50% of the variants are casual, irrespective of frequency, and 80% of these increase expected trait values by 1/4 standard deviation, while the remaining 20% decrease trait values by the same amount and Meta-analysis was performed using either our approach, Fisher's method, a modified Fisher's method taking into account unequal sample sizes and the minimal p-value approach to combine burden test, SKAT and variable threshold (VT) test statistics for variants with MAF<5%. Power was evaluated based upon 16,153 genes in the dataset using a significance threshold of $\alpha=3.1\times10^{-6}$.



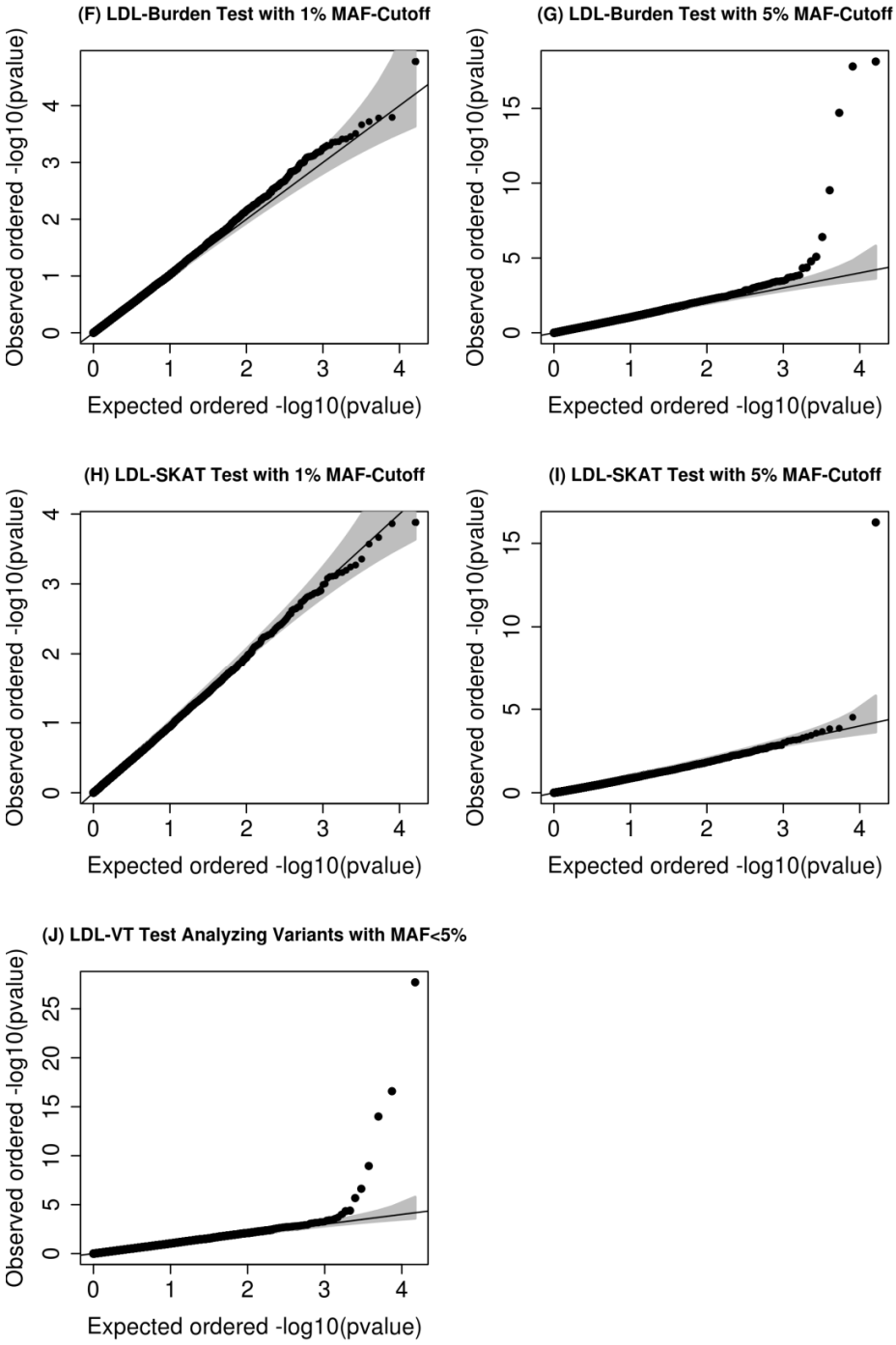
Supplementary Figure 10: Quantile-Quantile plot of p-values for single variant meta-analysis. Log-transformed observed and expected p-values are displayed for high density lipoprotein cholesterol (panels A-B), low density lipoprotein cholesterol (C-D) and triglyceride levels (E-F), either using all variants (left column) or variants with MAF < 5% (right column).



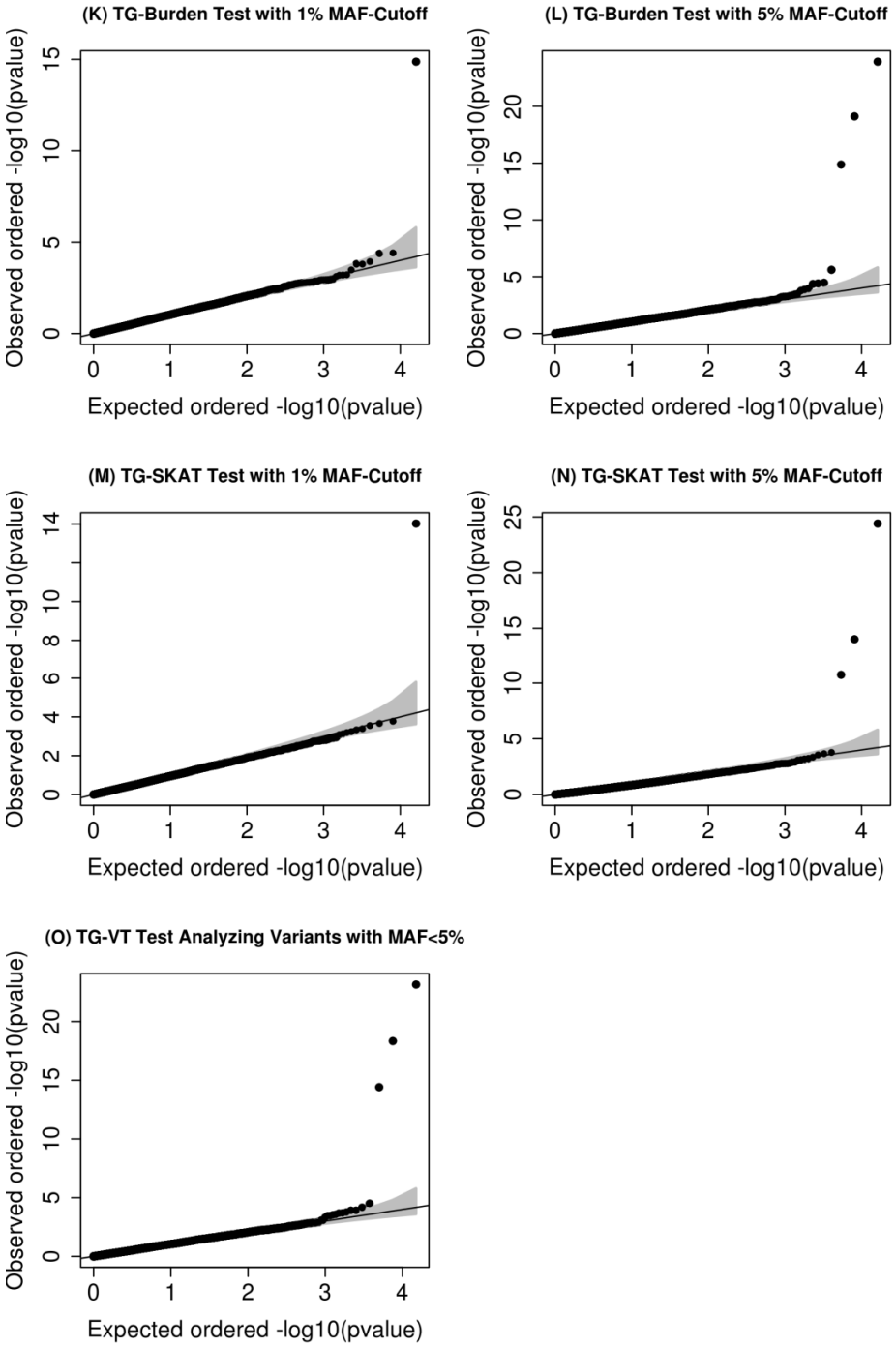
Supplementary Figure 11 (Part 1 of 3): Quantile-quantile plot of p-values for gene-level meta-analysis. Log-transformed observed and expected p-values are displayed for high density lipoprotein cholesterol (panels A-E), low density lipoprotein cholesterol (F-J) and triglyceride levels (K-O).



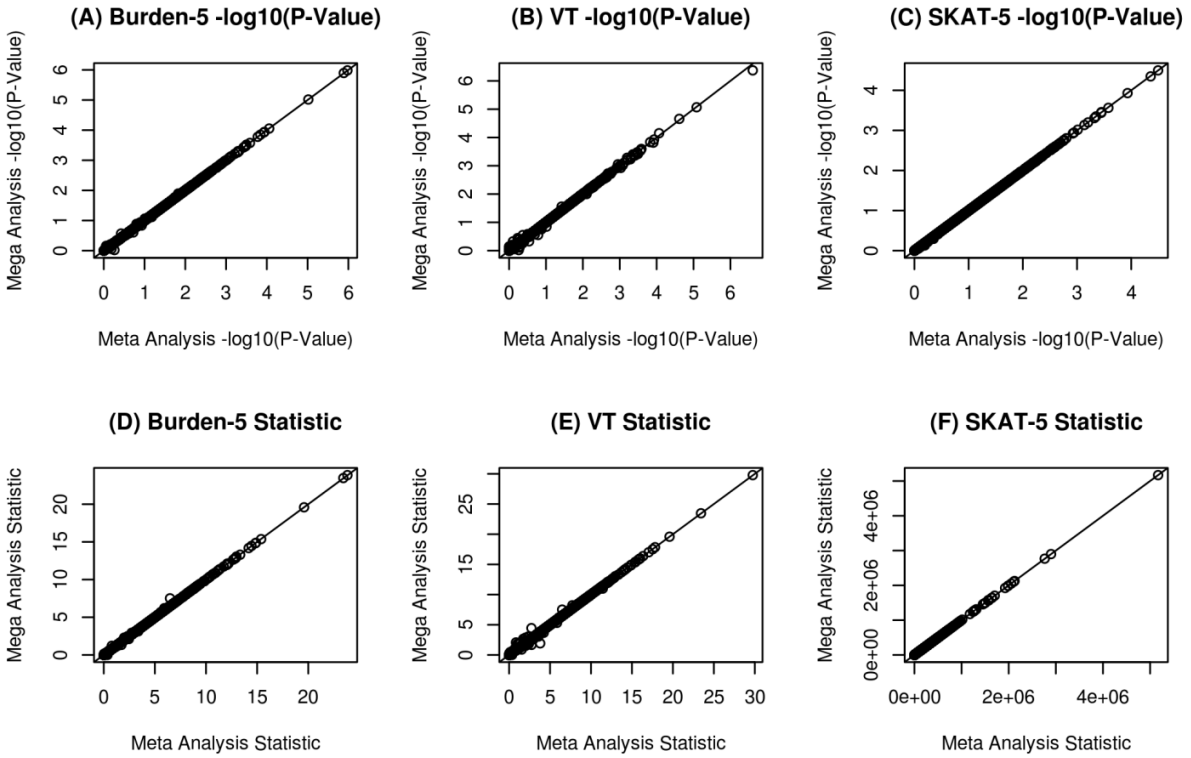
Supplementary Figure 11 (Part 2 of 3):



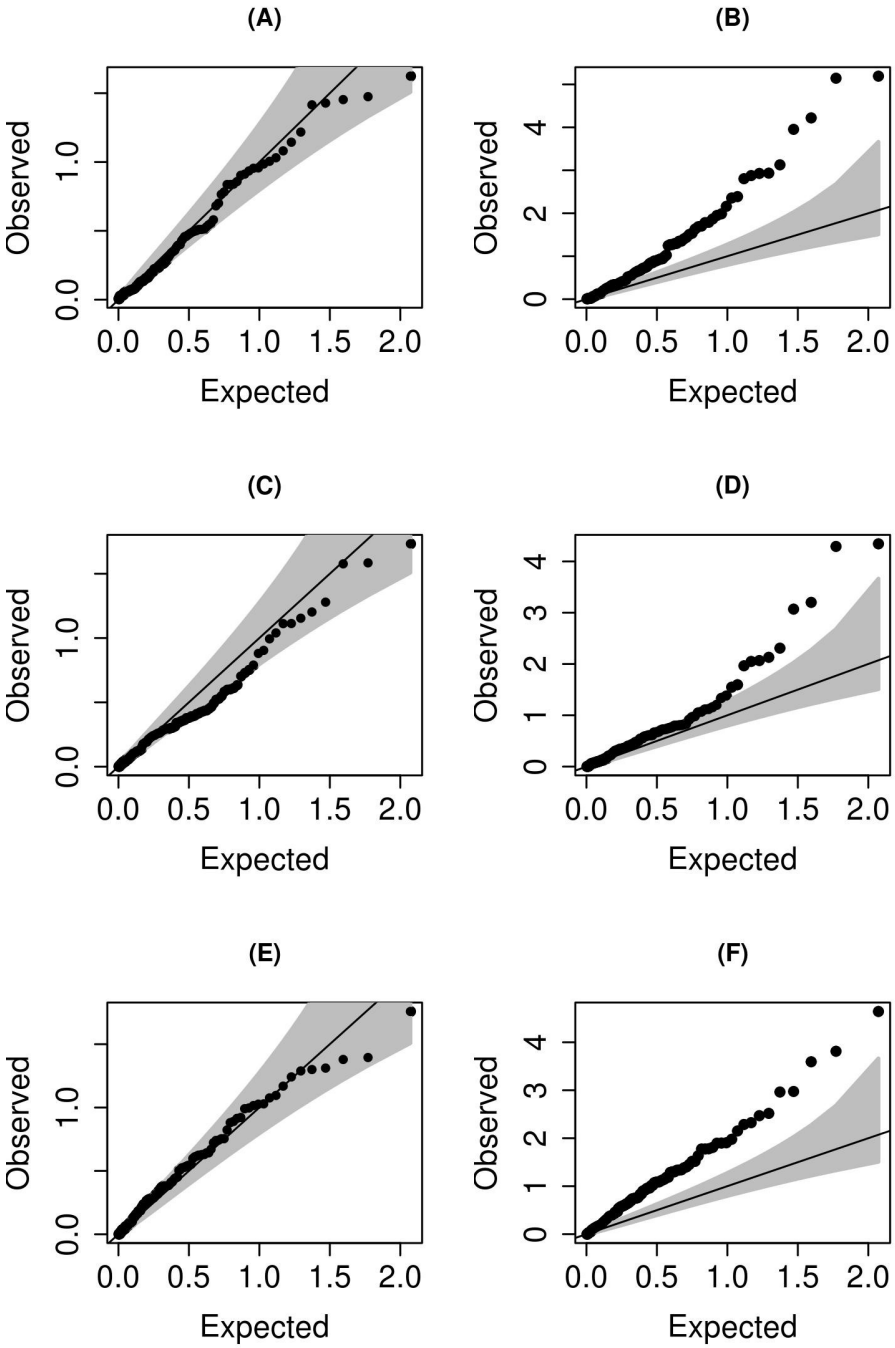
Supplementary Figure 11 (Part 3 of 3):



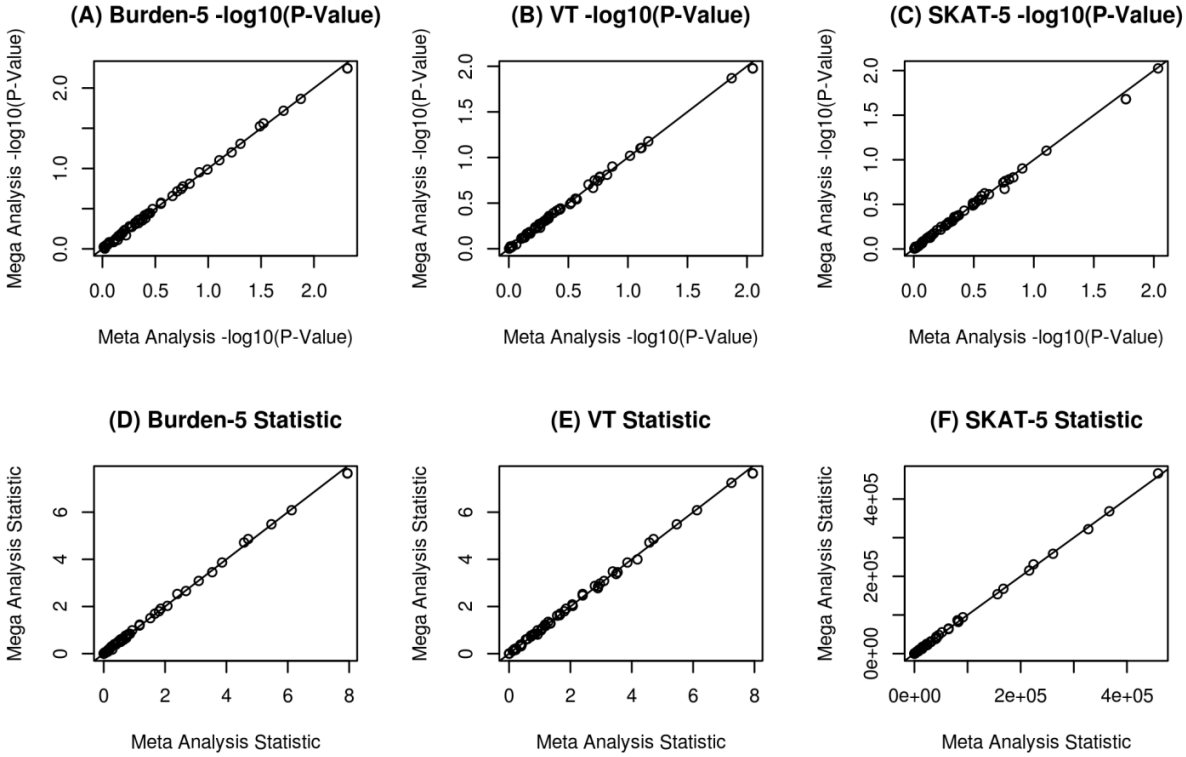
Supplementary Figure 12: Comparison of meta-analysis and analysis of pooled individual level data in the Malmö Diet and Cancer Study (MDC) and Ottawa Heart Study. LDL-cholesterol values in both cohorts were standardized using an inverse-normal transformation. Summary statistics were generated for both cohorts and meta-analyses were performed combining summary statistics from the two studies. As a comparison, analysis of pooled individual data was also performed. Scatter plots of -$\log_{10}$(p-values) from meta-analysis and mega-analysis of burden, VT and SKAT test are shown in panels (A-C), while the scatter plots for the test statistics are shown in panels (D-F). Note that applying the inverse-normal transformation to residuals in each study separately guards against potential artifacts due to heterogeneity between studies.



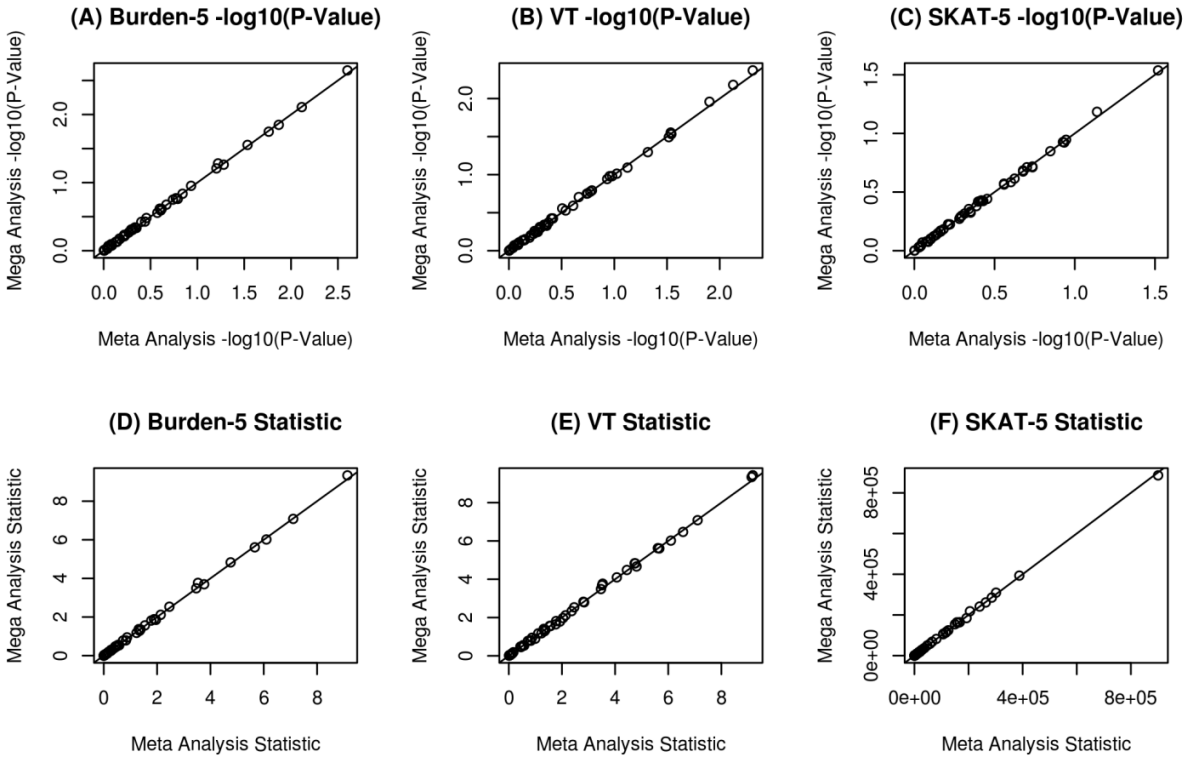
Supplementary Figure 13: Quantile-quantile (QQ) plot for log-transformed p-values comparing the distribution of p-values in a conditional analysis (on the left) and an unconditional analysis (on the right). One hundred samples were simulated. In each simulation, a single common variant (with MAF>10%) was marked as causal, increasing expected trait values by 0.25 standard deviation units. A series of rare variant association analysis were then carried out, with (left) or without (right) conditioning on the effect of this common variant, and panels A-B/C-D/E-F displaying QQ plots of $-\log_{10}$(p-values) form burden/VT/SKAT tests. The result clearly shows that, without conditioning, rare variant association test statistics are inflated.



Supplementary Figure 14: *APOE* region: comparison of conditional analysis in our meta-analysis framework and conditional analysis using pooled individual level data, in the MDC and Ottawa studies. LDL-cholesterol values in each cohort were standardized using an inverse-normal transformation. Summary statistics were generated for MDC and Ottawa cohorts. Conditional association analysis was performed in the pooled sample by controlling for the most significant single nucleotide polymorphism (rs7412) in *APOE*. Conditional evidence for association at sixty-six genes within 1Mb of rs7412 was evaluated, either using our meta-analysis approach (X axis) or by analyzing pooled individual level data directly (Y axis). The genes examined were *TOMM40, APOE, OPA3, ERCC1, MARK4, FOSB, PVRL2, CKM, CLPTM1, RTN2, ZNF155, ZNF230, ZFP112, ZNF225, ZNF223, ZNF221, ZNF222, DMPK, ZNF45, CEACAM19, CLASRP, BCL3, EML2, SIX5, GEMIN7, PPP1R13L, FBXO46, PVR, CBLC, LOC100379224, ZNF227, ZNF235, ZNF285, CEACAM20, ZNF296, NKPD1, TRAPPC6A, BLOC1S3, KLC3, PPM1N, IRF2BP1, MYPOP, ERCC2, ZNF226, CD3EAP, GIPR, ZNF180, DMWD, BCAM, SYMPK, ZNF229, RSPH6A, ZNF234, VASP, APOC4, APOC1, FOXA3, EXOC3L2, RELB, ZNF224, APOC4-APOC2, APOC2, CEACAM16, ZNF284, QPCTL, ZNF233*. Scatter plots of -$\log_{10}$(p-values) from meta-analysis and mega-analysis of burden, VT and SKAT tests are shown in panels (A-C), while the scatter plots for the test statistics are shown in panels (D-F).



Supplementary Figure 15: *LDLR* region: comparison of conditional analysis in our meta-analysis framework and conditional analysis using pooled individual level data, in the MDC and Ottawa studies. LDL-cholesterol values in each cohort were standardized using an inverse-normal transformation. Summary statistics were generated for MDC and Ottawa cohorts. Conditional association analysis was performed in the pooled sample conditional on three common single nucleotide polymorphism (rs6511720, rs2228671 and rs72658855) in gene *LDLR* that attain significant evidence for association. A total of 59 genes within 1Mb of these top 3 SNPs were analyzed, either using our meta-analysis approach (X axis) or by analyzing pooled individual level data directly (Y axis). Test statistics were evaluated at *DNM2, S1PR5, LOC55908, ZNF844, CCDC151, TYK2, ZNF440, ZNF491, CNN1, SLC44A2, SMARCA4, KEAP1, ICAM3, KANK2, ICAM1, RAVER1, EPOR, ICAM4, ZNF439, ILF3, DNMT1, PRKCSH, TMED1, KRI1, QTRT1, C19orf38, C19orf52, SPC24, TMEM205, C19orf39, ECSIT, ZNF441, ZNF69, ZNF700, ZNF433, AP1M2, TSPAN16, ACP5, RGL3, LDLR, YIPF2, CDKN2D, S1PR2, ZGLP1, ZNF653, LPPR2, DOCK6, PDE4A, ZNF627, CCDC159, ATG4D, RAB3D, CARM1, ICAM5, MRPL4, ZNF823, FDX1L, ZNF763, ZNF878*. Scatter plots of -$\log_{10}$(p-values) from meta-analysis and mega-analysis using burden, VT and SKAT tests are shown in panels (A-C), while the scatter plots for the test statistics are shown in panels (D-F).



Supplementary Tables

Supplementary Table 1: Evaluation of type I error rates for meta-analysis methods. Type I error rates were evaluated for three rare variant tests (burden-1: a simple burden test group variants with <1% frequency, VT: a variable threshold association test, SKAT-1: a sequence kernel association test focused on variants with frequency <1% and allowing for variants with opposite effects to reside in the same gene). Significance levels α=0.001, 0.0001, and 2.5×10^{-6} were considered. Data were generated for meta-analysis of 3, 6 and 9 samples of 1000 individuals. The type I error estimates are based upon 5×10^{7} null simulations.

Number of Studies	Burden-1	VT	SKAT-1
$\alpha=1\times10^{-3}$			
3	9.9×10^{-4}	1.0×10^{-3}	9.4×10^{-4}
6	1.0×10^{-3}	1.0×10^{-3}	9.8×10^{-4}
9	1.0×10^{-3}	1.0×10^{-3}	1.0×10^{-3}
$\alpha=1\times10^{-4}$			
3	9.9×10^{-5}	1.1×10^{-4}	9.2×10^{-5}
6	1.1×10^{-4}	1.2×10^{-4}	9.9×10^{-5}
9	1.0×10^{-4}	1.1×10^{-4}	9.9×10^{-5}
$\alpha=2.5\times10^{-6}$			
3	2.2×10^{-6}	2.6×10^{-6}	2.5×10^{-6}
6	2.4×10^{-6}	2.6×10^{-6}	2.6×10^{-6}
9	2.2×10^{-6}	1.6×10^{-6}	2.2×10^{-6}

Supplementary Table 2: Summary trait and variant information. In each study, medians and interquartile ranges are tabulated for age, sex and lipid traits, together with the number of genotyped non-synonymous and loss-of-function variants. Participating studies were the Malmo Diet and Cancer (MDC) study, the Ottawa Heart study, the Women's Health Initiative Sequencing Project (WHISP), Procardis and HUNT. Genotyped samples in Procardis and HUNT are separated into heart disease cases and controls.

Study	Age	HDL	LDL	TG	Total Number of Individuals	Proportion of Males	Number of Variants		
	Median (Interquartile Range) Lipids Level (mg/dL)						All	Nonsynonymous + Loss of Function	Nonsynonymous + Loss of Function, MAF <1%
Malmo Diet and Cancer	58 (10.4)	51.4 (18.9)	158.3 (50.2)	102.7 (65.5)	4924	40.8%	130,621	111,127	90,317
Ottawa Heart Study	72 (13.1)	50.6 (20.9)	139.4 (48.3)	121.2 (100.0)	2938	60.0%	116,173	97,628	77,866
WHISP European Americans	68 (7.0)	56.0 (22.0)	140.2 (48.1)	139.0 (87.0)	2031	0.0%	110,678	91,998	70,421
PROCARDIS (Cases)	58 (10.0)	46.0 (17.0)	142.0 (54.3)	159.0 (120.0)	2070	48.9%	97,887	79,551	58,864
PROCARDIS (Controls)	66 (7.0)	53.9 (20.5)	129.9 (42.2)	123.1 (88.5)	1299	62.1%	105,255	86,639	66,114
HUNT (Cases)	66(18.0)	46.3(15.4)	162.2(57.9)	177 (123.9)	2659	65.5%	90,340	72,866	52,133
HUNT (Controls)	65(19.0)	50.2(19.3)	154.4(54.1)	141.6 (86.7)	2778	66.1%	91,902	74,366	53,682

Supplementary Table 3 Variants sites shared between studies, by frequency. The number of shared variant nucleotide sites are displayed respectively for each pair of studies and for variant sites with MAF > 1% and with MAF≤ 1%. Tabulated studies include the Malmo Diet and Cancer (MDC) study, the Ottawa Heart Study (Ottawa), European American Samples from the WHISP study, and case and control samples from Procardis and HUNT.

	MDC	Ottawa	WHISP European Americans	PROCARDIS (Cases)	PROCARDIS (Controls)	HUNT (Cases)	HUNT (Controls)
Variants with MAF > 1%							
MDC	36,520	33,546	34,414	34,925	34,903	34,263	34,251
Ottawa		34,981	34,472	34,311	34,265	33,178	33,183
WHISP European Americans			37,256	35,090	35,020	33,942	33,929
PROCARDIS (Cases)				36,484	35,609	34,370	34,390
PROCARDIS (Controls)					36,283	34,317	34,327
HUNT (Cases)						36,140	35,516
HUNT (Controls)							36,098
Variants with MAF ≤ 1%							
MDC	94,101	57,932	53,613	48,682	53,051	44,583	45,773
Ottawa		81,192	55,396	47,528	52,283	39,414	40,493
WHISP European Americans			73,422	43,892	47,858	36,627	37,614
PROCARDIS (Cases)				61,403	45,837	35,867	36,670
PROCARDIS (Controls)					68,972	38,204	39,061
HUNT (Cases)						54,200	45,054
HUNT (Controls)							55,804

Supplementary Table 4: Results for single variant meta-analysis. Loci that are statistically significant after Bonferroni correction (with $p < 3\times10^{-7}$) are shown. In each locus, p-value, annotation, reference and alternative allele, alternative allele frequency as well as genetic effect estimate and standard deviation are displayed for the variant with the most significant p-value.

Gene	Gene Position[a]	P-value	rs#	Annotation	Ref/Alt	Frequency for Alt Allele	Estimated Effect Size for Alt Allele (in standard deviation units)	Standard Error for Estimated Effect (in standard deviation units)
				HDL				
LPL	chr8:19.8Mb	1.17×10^{-18}	rs268	Nonsynonymous	A/G	.025	-0.296	.001129
ANGPTL4	chr19:8.4Mb	3.61×10^{-18}	rs116843064	Nonsynonymous	G/A	.026	0.281	.001045
LIPG	chr18:47.1Mb	7.26×10^{-15}	rs77960347	Nonsynonymous	A/G	.013	0.348	.002006
CD300LG	chr17:41.9Mb	3.00×10^{-10}	rs72836561	Nonsynonymous	C/T	.033	-0.182	.000830
LIPC	chr15:58.9Mb	5.10×10^{-10}	rs113298164	Nonsynonymous	C/T	.0037	0.536	.007425
APOB	chr2:21.2Mb	2.24×10^{-9}	rs533617	Nonsynonymous	T/C	.040	0.159	.000709
HNF4A	chr20:43.0Mb	2.64×10^{-7}	rs1800961	Nonsynonymous	C/T	.041	-0.134	.000680
				LDL				
PCSK9	chr1:55.5Mb	2.60×10^{-28}	rs11591147	Nonsynonymous	G/T	.013	-0.511	.002146
BCAM	chr19:45.3Mb	2.64×10^{-24}	rs28399653	Nonsynonymous	G/A	.035	-0.290	.000810
CBLC	chr19:45.3Mb	2.99×10^{-22}	rs3208856	Nonsynonymous	C/T	.037	-0.270	.000774
PVR	chr19:45.2Mb	1.71×10^{-9}	rs1058402	Nonsynonymous	G/A	.049	-0.146	.000586
APOB	chr2:21.2Mb	6.70×10^{-9}	rs5742904	Nonsynonymous	C/T	.00063	1.211	.043597
				TG				
ANGPTL4	chr19:8.4Mb	8.55×10^{-24}	rs116843064	Nonsynonymous	G/A	.027	-0.325	.001043
LPL	chr8:19.8Mb	1.70×10^{-19}	rs268	Nonsynonymous	A/G	.025	0.302	.001120
APOB	chr2:21.2Mb	1.81×10^{-10}	rs533617	Nonsynonymous	T/C	.040	-0.170	.000708

a. Gene position is defined based upon hg19, GRCh37 Genome Reference Consortium Human Reference 37

Supplementary Table 5 (Part 1 of 3): Comparison of meta-analysis and analysis of individual studies for gene-level tests. Results for six rare variant tests are shown (burden-5: a simple burden test group variants with <5% or <1% frequency, VT: a variable threshold association test, SKAT-5: a sequence kernel association test focused on variants with <5% or <1% frequency and allowing for variants with opposite effects to reside in the same gene). For tests that assume that model the average effect of variants in a gene, a + or – sign indicates whether these variants raised (+) or lowered (-) trait levels on average. Overall, the results show that, for these genes, meta-analysis results in a substantially stronger signal than analysis of any single sample and that the direction of effect for these top signals is generally consistent across studies. Study abbreviations are as in previous tables.

Gene	Meta Analysis	MDC	Ottawa	WHISP	PROCARDIS (Cases)	PROCARDIS (Controls)	HUNT (Cases)	HUNT (Controls)
				Burden-5				
				HDL				
LPL	2×10^{-24}/-	5×10^{-11}/-	4×10^{-5}/-	0.007/-	5×10^{-4}/-	0.004/-	0.002/-	0.001/-
ANGPTL4	3×10^{-19}/+	2×10^{-5}/+	2×10^{-6}/+	0.04/+	0.1/+	0.006/+	0.03/+	4×10^{-6}/+
LIPG	6×10^{-19}/+	1×10^{-8}/+	0.03/+	0.2/+	0.003/+	0.04/+	5×10^{-7}/+	0.001/+
HNF4A	3×10^{-7}/-	0.009/-	5×10^{-4}/-	0.003/-	0.7/+	0.002/-	0.8/-	0.08/-
LIPC	4×10^{-7}/+	8×10^{-4}/+	0.4/+	0.4/+	0.5/+	0.3/+	0.007/+	9×10^{-4}/+
CD300LG	8×10^{-7}/-	0.04/-	0.002/-	0.1/-	0.8/+	0.09/-	0.7/-	1×10^{-4}/-
				LDL				
PCSK9	7×10^{-19}/-	5×10^{-5}/-	1×10^{-9}/-	5×10^{-7}/-	0.02/-	0.001/-	0.02/-	0.06/-
BCAM	2×10^{-18}/-	6×10^{-6}/-	0.4/-	0.02/-	0.01/-	0.01/-	0.03/-	3×10^{-6}/-
CBLC	2×10^{-15}/-	3×10^{-7}/-	0.1/-	0.3/-	0.2/-	5×10^{-5}/-	0.002/-	2×10^{-4}/-
PVR	3×10^{-10}/-	2×10^{-5}/-	0.02/-	0.05/-	0.01/-	0.2/-	0.1/+	0.3/-
				TG				
ANGPTL4	1×10^{-24}/-	3×10^{-6}/-	9×10^{-6}/-	2×10^{-6}/-	0.01/-	0.005/-	0.002/-	2×10^{-6}/-
LPL	8×10^{-20}/+	1×10^{-9}/+	0.001/+	5×10^{-4}/+	0.001/+	0.2/+	0.04/+	5×10^{-5}/+

Supplementary Table 5 (Part 2 of 3).

Gene	Meta Analysis	MDC	Ottawa	WHISP	PROCARDIS (Cases)	PROCARDIS (Controls)	HUNT (Cases)	HUNT (Controls)
				SKAT-5				
				HDL				
ANGPTL4	3×10^{-19}/+	5×10^{-5}/+	2×10^{-6}/+	0.03/+	0.02/+	0.03/+	0.04/+	7×10^{-6}/+
LPL	5×10^{-13}/-	8×10^{-6}/-	0.003/-	0.08/-	0.07/-	0.1/-	0.1/-	0.05/-
LIPG	3×10^{-9}/+	3×10^{-4}/+	0.04/+	0.4/+	0.03/+	0.2/+	0.002/+	0.07/+
HNF4A	3×10^{-7}/-	0.009/-	4×10^{-4}/-	0.003/-	0.8/+	0.003/-	0.8/-	0.08/-
				LDL				
PCSK9	6×10^{-17}/-	8×10^{-4}/-	4×10^{-7}/-	3×10^{-9}/-	0.003/-	3×10^{-5}/-	0.2/-	0.1/-
				TG				
ANGPTL4	4×10^{-25}/-	4×10^{-6}/-	8×10^{-6}/-	8×10^{-6}/-	0.006/-	0.02/-	0.005/-	1×10^{-5}/-
LPL	2×10^{-11}/+	5×10^{-7}/+	0.02/+	0.02/+	0.1/+	0.3/+	0.2/+	0.02/+

Supplementary Table 5 (Part 3 of 3).

				VT				
				HDL				
LPL	1×10^{-23}/-	1×10^{-10}/-	2×10^{-4}/-	0.02/-	0.002/-	0.02/-	0.006/-	0.003/-
ANGPTL4	2×10^{-18}/+	1×10^{-4}/+	2×10^{-5}/+	0.1/+	0.4/-	0.02/+	0.07/+	2×10^{-5}/+
LIPG	4×10^{-18}/+	6×10^{-8}/+	0.04/-	0.4/-	0.01/+	0.1/+	1×10^{-6}/+	0.003/+
LIPC	4×10^{-12}/+	0.003/-	0.2/-	0.5/-	0.8/-	0.7/-	1×10^{-4}/-	4×10^{-7}/-
				LDL				
PCSK9	2×10^{-28}/-	2×10^{-5}/-	3×10^{-9}/-	2×10^{-6}/-	0.001/-	9×10^{-6}/-	0.08/-	0.008/-
BCAM	3×10^{-17}/-	9×10^{-5}/-	0.7/-	0.007/-	0.01/-	0.003/-	0.07/-	4×10^{-6}/-
CBLC	1×10^{-14}/-	7×10^{-7}/-	0.4/-	0.06/+	0.06/-	3×10^{-4}/-	0.01/-	0.001/-
PVR	1×10^{-9}/-	8×10^{-5}/-	0.08/-	0.2/-	0.06/-	0.3/-	0.3/+	0.7/+
LDLR	2×10^{-7}/+	0.1/-	0.001/-	0.2/-	5×10^{-4}/-	0.1/-	0.03/-	0.008/-
				TG				
ANGPTL4	7×10^{-24}/-	1×10^{-5}/-	5×10^{-5}/-	6×10^{-6}/-	0.04/+	0.01/-	0.005/-	7×10^{-6}/-
LPL	5×10^{-19}/+	4×10^{-9}/+	0.006/+	0.001/+	0.005/+	0.6/+	0.06/-	2×10^{-4}/+
				Burden-1				
				HDL				
LIPC	1×10^{-12}/+	0.004/+	0.05/+	0.2/+	0.7/+	0.8/+	4×10^{-5}/+	1×10^{-7}/+
				SKAT-1				
				HDL				
LIPC	2×10^{-9}/+	0.07/+	0.1/+	0.08/+	0.7/+	0.8/+	0.001/+	3×10^{-5}/+

Supplementary Table 6: Comparison of gene-level test results with single variant association tests. For each locus identified using gene-level association tests, we show the rs number, ref/alt allele, alt allele frequency and p-value for the variant site that displays the most significant p-value. The loci where one or more gene-based association signal exceeds the top single variant association signal are labeled with an asterisk.

Gene	Burden-1	Burden-5	SKAT-1	SKAT-5	VT	MAF Cutoff	Top Single Variant Association(MAF<5%)			
							rs Number	Ref/Alt	p-value	AF
HDL										
LIPC*	**1.4×10^{-12}**	**3.5×10^{-7}**	**1.8×10^{-9}**	1.4×10^{-2}	**4.5×10^{-12}**	3.7×10^{-3}	rs113298164	C/T	5.1×10^{-10}	3.68×10^{-3}
LPL*	0.97	**2.5×10^{-24}**	0.35	**5.0×10^{-13}**	**1.5×10^{-23}**	0.025	rs268	A/G	1.2×10^{-18}	0.025
ANGPTL4*	0.022	**2.9×10^{-19}**	0.022	**3.0×10^{-19}**	**1.8×10^{-18}**	0.026	rs116843064	G/A	3.6×10^{-18}	0.027
LIPG*	2.2×10^{-5}	**6.4×10^{-19}**	2.1×10^{-5}	**2.9×10^{-9}**	**4.4×10^{-18}**	0.013	rs77960347	A/G	7.3×10^{-15}	0.014
HNF4A	0.74	**2.8×10^{-7}**	0.68	**2.5×10^{-7}**	**1.5×10^{-6}**	0.041	rs1800961	C/T	2.6×10^{-7}	0.041
CD300LG	0.49	**8.5×10^{-7}**	0.52	1.0×10^{-5}	**3.1×10^{-6}**	0.033	rs72836561	C/T	3.0×10^{-10}	0.033
LDL										
PCSK9*	1.8×10^{-2}	**7.4×10^{-19}**	0.081	**5.5×10^{-17}**	**2.0×10^{-28}**	0.013	rs11591147	G/T	2.6×10^{-28}	0.013
BCAM	0.17	**1.6×10^{-18}**	0.15	3.0×10^{-5}	**2.6×10^{-17}**	0.036	rs28399653	G/A	2.6×10^{-24}	0.035
CBLC	0.94	**2.0×10^{-15}**	0.44	1.5×10^{-4}	**1.0×10^{-14}**	0.044	rs3208856	C/T	3.0×10^{-22}	0.037
PVR	0.061	**3.0×10^{-10}**	0.048	0.063	**1.1×10^{-9}**	0.049	rs1058402	G/A	1.7×10^{-9}	0.049
LDLR*	1.8×10^{-3}	4.7×10^{-5}	0.038	0.25	**2.4×10^{-7}**	5.2×10^{-4}	rs139791325	G/A	7.68×10^{-4}	5.2×10^{-4}
TG										
ANGPTL4*	0.026	**1.2×10^{-24}**	0.037	**3.9×10^{-25}**	**7.1×10^{-24}**	0.026	rs116843064	G/A	8.6×10^{-24}	0.027
LPL*	0.68	**7.7×10^{-20}**	0.26	**1.8×10^{-11}**	**4.6×10^{-19}**	0.025	rs268	A/G	1.7×10^{-19}	0.025

Nature Genetics: doi:10.1038/ng.2852

Supplementary Table 7: Results of meta-analysis of gene-level association using our approach and Fisher's method for combining p-values. Meta-analysis were carried out for a simple burden test, for a SKAT test using 1% and 5% allele frequency cutoffs, and for the variable threshold (VT) test that analyzes variants with MAF<5%. Significant p-values (using a threshold of 3.1×10^{-6}) are displayed in **bold**.

	Burden-1		Burden-5		SKAT-1		SKAT-5		VT	
Gene	Fisher	Our Approach	Fisher	Our Approach	Fisher	Our Approach	Fisher	Our Approach	Fisher	Our Approach
					HDL					
LIPC	**8.0×10^{-10}**	**1.4×10^{-12}**	.05	**3.5×10^{-7}**	**2.1×10^{-7}**	**1.8×10^{-9}**	.16	.014	**1.2×10^{-8}**	**4.5×10^{-12}**
LPL	.41	.97	**7.9×10^{-21}**	**2.5×10^{-24}**	.49	.35	**3.6×10^{-8}**	**5.0×10^{-13}**	.22	**1.5×10^{-23}**
ANGPTL4	.14	.022	**2.7×10^{-16}**	**2.9×10^{-19}**	.39	.022	**2.3×10^{-14}**	**3.0×10^{-19}**	.11	**1.8×10^{-18}**
LIPG	5.5×10^{-4}	2.2×10^{-5}	**2.2×10^{-16}**	**6.4×10^{-19}**	3.5×10^{-3}	2.1×10^{-5}	**2.2×10^{-6}**	**2.9×10^{-9}**	1.9×10^{-3}	**4.4×10^{-18}**
HNF4A	.73	.75	**1.1×10^{-6}**	**2.8×10^{-7}**	.72	.68	**1.7×10^{-6}**	**2.5×10^{-7}**	.72	**1.5×10^{-6}**
CD300LG	.99	.49	2.1×10^{-5}	**8.5×10^{-7}**	.99	.52	4.0×10^{-3}	1.0×10^{-5}	1.0	**3.1×10^{-6}**
					LDL					
PCSK9	5.8×10^{-3}	.018	**2.3×10^{-16}**	**7.4×10^{-19}**	5.6×10^{-3}	.081	**5.4×10^{-7}**	**5.5×10^{-17}**	5.7×10^{-4}	**2.0×10^{-28}**
BCAM	.46	.17	**5.6×10^{-16}**	**1.6×10^{-18}**	.67	.15	3.8×10^{-3}	3.0×10^{-5}	.62	**2.6×10^{-17}**
CBLC	.056	.94	**2.0×10^{-13}**	**2.0×10^{-15}**	.23	.44	.05	1.5×10^{-4}	.19	**1.0×10^{-14}**
PVR	.045	.061	1.5×10^{-3}	**3.0×10^{-10}**	.054	.048	.63	.063	.24	**1.1×10^{-9}**
LDLR	2.9×10^{-5}	1.8×10^{-3}	2.3×10^{-3}	4.7×10^{-5}	7.5×10^{-4}	.038	.92	.25	3.5×10^{-5}	**2.4×10^{-7}**
					TG					
ANGPTL4	.28	.026	**1.6×10^{-21}**	**1.2×10^{-24}**	.24	.037	**7.7×10^{-21}**	**3.9×10^{-25}**	.18	**7.1×10^{-24}**
LPL	.60	.68	**2.7×10^{-17}**	**7.7×10^{-20}**	.13	.26	**8.5×10^{-9}**	**1.8×10^{-11}**	.12	**4.6×10^{-19}**

Nature Genetics: doi:10.1038/ng.2852

Supplementary Table 8: Results of meta-analysis of gene-level association using our method and a weighted version of Fisher's method. P-values are displayed for meta-analysis of a simple burden test, the SKAT test using 1% and 5% allele frequency cutoffs, and a variable threshold (VT) test that analyzes variants with MAF<5%. Significant p-values (using a threshold of 3.1×10^{-6}) are displayed in **bold**.

	Burden-1		Burden-5		SKAT-1		SKAT-5		VT	
Gene	Weighted Fisher	Our Approach	Weighted Fisher	Our Approach	Weighted Fisher	Our Approach	Weighted Fisher	Our Approach	Weighted Fisher	Our Approach
					HDL					
LIPC	$\mathbf{1.5\times10^{-9}}$	$\mathbf{1.4\times10^{-12}}$	.057	$\mathbf{3.5\times10^{-7}}$	1.2×10^{-5}	$\mathbf{1.8\times10^{-9}}$	.052	.014	$\mathbf{2.7\times10^{-7}}$	$\mathbf{4.5\times10^{-12}}$
LPL	.50	.97	$\mathbf{2.3\times10^{-21}}$	$\mathbf{2.5\times10^{-24}}$	.58	.35	$\mathbf{2.6\times10^{-9}}$	$\mathbf{5.0\times10^{-13}}$	$\mathbf{8.2\times10^{-19}}$	$\mathbf{1.5\times10^{-23}}$
ANGPTL4	.12	.022	$\mathbf{8.4\times10^{-16}}$	$\mathbf{2.9\times10^{-19}}$	.42	.022	$\mathbf{7.8\times10^{-14}}$	$\mathbf{3.0\times10^{-19}}$	$\mathbf{2.5\times10^{-13}}$	$\mathbf{1.8\times10^{-18}}$
LIPG	1.3×10^{-4}	2.2×10^{-5}	$\mathbf{5.7\times10^{-17}}$	$\mathbf{6.4\times10^{-19}}$	1.2×10^{-3}	2.1×10^{-5}	1.2×10^{-5}	$\mathbf{2.9\times10^{-9}}$	$\mathbf{3.6\times10^{-14}}$	$\mathbf{4.4\times10^{-18}}$
HNF4A	.74	.75	1.6×10^{-5}	$\mathbf{2.8\times10^{-7}}$	.75	.68	1.7×10^{-5}	$\mathbf{2.5\times10^{-7}}$	1.9×10^{-4}	$\mathbf{1.5\times10^{-6}}$
CD300LG	.97	.49	3.7×10^{-5}	$\mathbf{8.5\times10^{-7}}$	.95	.52	1.7×10^{-3}	1.0×10^{-5}	5.1×10^{-4}	$\mathbf{3.1\times10^{-6}}$
					LDL					
PCSK9	.0054	.018	$\mathbf{7.2\times10^{-15}}$	$\mathbf{7.4\times10^{-19}}$	2.6×10^{-3}	.081	2.2×10^{-5}	$\mathbf{5.5\times10^{-17}}$	$\mathbf{1.0\times10^{-18}}$	$\mathbf{2.0\times10^{-28}}$
BCAM	.47	.17	$\mathbf{4.5\times10^{-15}}$	$\mathbf{1.6\times10^{-18}}$	.66	.15	4.3×10^{-3}	3.0×10^{-5}	$\mathbf{2.0\times10^{-11}}$	$\mathbf{2.6\times10^{-17}}$
CBLC	.11	.94	$\mathbf{2.5\times10^{-14}}$	$\mathbf{2.0\times10^{-15}}$	.26	.44	.027	1.5×10^{-4}	$\mathbf{4.1\times10^{-12}}$	$\mathbf{1.0\times10^{-14}}$
PVR	.13	.061	1.5×10^{-4}	$\mathbf{3.0\times10^{-10}}$	.11	.048	.5	.063	1.5×10^{-3}	$\mathbf{1.1\times10^{-9}}$
LDLR	2.3×10^{-4}	1.8×10^{-3}	2.3×10^{-3}	4.7×10^{-5}	1.4×10^{-3}	.038	.89	.25	1.5×10^{-5}	$\mathbf{2.4\times10^{-7}}$
					TG					
ANGPTL4	.32	.026	$\mathbf{3.3\times10^{-19}}$	$\mathbf{1.2\times10^{-24}}$	.28	.037	$\mathbf{3.4\times10^{-19}}$	$\mathbf{3.9\times10^{-25}}$	$\mathbf{1.4\times10^{-16}}$	$\mathbf{7.1\times10^{-24}}$
LPL	.68	.68	$\mathbf{7.5\times10^{-18}}$	$\mathbf{7.7\times10^{-20}}$	.18	.26	$\mathbf{1.1\times10^{-10}}$	$\mathbf{1.8\times10^{-11}}$	$\mathbf{8.4\times10^{-16}}$	$\mathbf{4.6\times10^{-19}}$

Nature Genetics: doi:10.1038/ng.2852

Supplementary Table 9: Results of meta-analysis of gene-level association using our approach and the minimum p-value approach. P-values are displayed for meta-analysis of a simple burden test, a SKAT test using 1% and 5% allele frequency cutoffs, and a variable threshold (VT) test that analyzes variants with MAF<5%. Significant p-values (using a threshold of $p<3.1\times10^{-6}$) are displayed in **bold**.

	Burden-1		Burden-5		SKAT-1		SKAT-5		VT	
Gene	Minimal-P	Our Approach	Minimal-P	Our Approach	Minimal-P	Our Approach	Minimal-P	Our Approach	Minimal-P	Our Approach
					HDL					
LIPC	8.7×10^{-7}	**1.4×10^{-12}**	.073	**3.5×10^{-7}**	4.1×10^{-6}	**1.8×10^{-9}**	.091	.014	4.0×10^{-6}	**4.5×10^{-12}**
LPL	.72	.97	**3.2×10^{-10}**	**2.5×10^{-24}**	.87	.35	3.7×10^{-5}	**5.0×10^{-13}**	**1.1×10^{-9}**	**1.5×10^{-23}**
ANGPTL4	.45	.022	**1.5×10^{-6}**	**2.9×10^{-19}**	.51	.022	3.7×10^{-5}	**3.0×10^{-19}**	6.6×10^{-6}	**1.8×10^{-18}**
LIPG	.012	2.2×10^{-5}	**9.2×10^{-8}**	**6.4×10^{-19}**	.06	2.1×10^{-5}	2.2×10^{-3}	**2.9×10^{-9}**	**5.5×10^{-7}**	**4.4×10^{-18}**
HNF4A	.43	.75	3.8×10^{-3}	**2.8×10^{-7}**	.44	.68	3.2×10^{-3}	**2.5×10^{-7}**	.013	**1.5×10^{-6}**
CD300LG	.54	.49	2.8×10^{-3}	**8.5×10^{-7}**	.62	.52	.028	1.0×10^{-5}	.01	**3.1×10^{-6}**
					LDL					
PCSK9	2.2×10^{-5}	.018	1.6×10^{-5}	**7.4×10^{-19}**	.011	.081	3.4×10^{-3}	**5.5×10^{-17}**	**8.4×10^{-8}**	**2.0×10^{-28}**
BCAM	.60	.17	3.7×10^{-6}	**1.6×10^{-18}**	.64	.015	.18	3.0×10^{-5}	4.5×10^{-5}	**2.6×10^{-17}**
CBLC	.11	.94	**7.8×10^{-7}**	**2.0×10^{-15}**	.39	.44	.22	1.5×10^{-4}	**2.2×10^{-6}**	**1.0×10^{-14}**
PVR	.23	.061	8.2×10^{-4}	**3.0×10^{-10}**	.28	.048	.77	.063	4.5×10^{-3}	**1.1×10^{-9}**
LDLR	3.2×10^{-3}	1.8×10^{-3}	.019	4.7×10^{-5}	.02	.038	.92	.25	8.2×10^{-4}	**2.4×10^{-7}**
					TG					
ANGPTL4	.38	.026	1.0×10^{-5}	**1.2×10^{-24}**	.44	.037	**1.4×10^{-6}**	**3.9×10^{-25}**	5.0×10^{-5}	**7.1×10^{-24}**
LPL	.15	.68	**8.8×10^{-9}**	**7.7×10^{-20}**	.20	.26	**8.7×10^{-7}**	**1.8×10^{-11}**	**2.6×10^{-8}**	**4.6×10^{-19}**

Nature Genetics: doi:10.1038/ng.2852

Supplementary Table 10: Results of conditional association analysis for LDL and variants in *LDLR*. We performed conditional association analysis for variants in *LDLR*, conditioning on 3 common variants (rs6511720, rs2228671 and rs72658855) that are strongly associated in single variant analyses (i.e. with p-value $<3\times10^{-7}$). The rs number, reference and alternate alleles, minor allele frequencies, p-values before and after conditioning, estimates of effect size per copy of the alternative alleles, and annotation information are displayed for non-synonymous and loss-of-function variants. Gene level association test results are at the bottom of the table.

Single Variant Association Analysis								
RS	**Ref**	**Alt**	**MAF**	**Original P-value**	**Conditional P-value**	**Original Effect Estimate[a]**	**Conditional Effect Estimate[a]**	**Annotation**
rs6511720	G	T	.11	2×10^{-38}	-	-0.22	-	Intron
rs2228671	C	T	.11	4×10^{-22}	-	-0.16	-	Synonymous
rs2738459	A	C	.49	4×10^{-8}	-	-0.06	-	Intron
rs11669576	G	A	.046	8×10^{-4}	0.201	0.08	0.03	Nonsynonymous
rs139624145	G	A	.0001	8×10^{-4}	0.001	1.68	1.61	Nonsynonymous
rs139791325	G	A	.0005	8×10^{-4}	0.002	0.77	0.7	Nonsynonymous
rs199774121	C	A	3×10^{-5}	0.004	0.002	2.88	3.14	Stop_Gain
rs144172724	G	A	3×10^{-5}	0.024	0.037	2.26	2.11	Nonsynonymous
rs141673997	G	A	3×10^{-5}	0.048	0.056	1.98	1.92	Nonsynonymous
rs150673992	C	T	6×10^{-5}	0.056	0.031	1.35	1.54	Nonsynonymous
rs28942084	C	T	6×10^{-5}	0.151	0.158	1.02	1.01	Nonsynonymous
rs139043155	T	A	.0001	0.21	0.241	0.63	0.59	Nonsynonymous
rs139361635	G	A	3×10^{-5}	0.266	0.2	1.11	1.29	Nonsynonymous
rs143992984	G	A	8×10^{-5}	0.358	0.233	0.53	0.7	Nonsynonymous
rs137853963	G	A	.0018	0.391	0.212	-0.11	-0.16	Nonsynonymous
rs13306505	C	T	6×10^{-5}	0.511	0.47	0.47	0.51	Nonsynonymous
rs148698650	G	A	.0001	0.539	0.585	0.27	0.25	Nonsynonymous
rs200727689	G	A	3×10^{-5}	0.603	0.667	0.52	0.43	Nonsynonymous
rs5928	G	A	3×10^{-5}	0.892	0.851	-0.14	-0.19	Nonsynonymous
rs146200173	C	G	.0002	0.997	0.99	0	0	Nonsynonymous

Gene-level Test						
Gene	**P-value before Conditioning**	**P-value after Conditioning**	**MAF cutoff before conditioning**	**MAF cutoff after conditioning**	**Estimate of Genetic Effect Before Conditioning[a]**	**Estimate of Genetic Effect After Conditioning[a]**
LDLR	2.4×10^{-7}	4.6×10^{-7}	5.2×10^{-4}	5.2×10^{-4}	0.75	0.73

a. In standard deviation units.

Nature Genetics: doi:10.1038/ng.2852

Supplementary Table 11: Results of conditional association analysis for trait LDL and genes *BCAM, CBLC* and *PVR* near the *APOE* locus. We performed conditional association analysis conditioning on the top variant (rs7412) in the locus. The p-values before and after conditional analysis for burden test and SKAT tests with 5% MAF cutoff and a variable threshold (VT) test that analyzes variant with MAF<5% are shown. The rs number, reference, alternative alleles, p-values before and after conditioning on rs7412 were also displayed for each gene.

	Burden-5		SKAT-5		VT						
							Top SNP				
Gene	Original P-value	Conditional P-value	Original P-value	Conditional P-value	Original P-value	Conditional P-value	RS#	Ref	Alt	Original P-value	Conditional P-value
BCAM	1.57×10^{-18}	.89	3.01×10^{-5}	.42	2.61×10^{-17}	.80	rs28399653	G	A	2.64×10^{-24}	.67
CBLC	1.98×10^{-15}	.02	1.47×10^{-4}	.41	9.99×10^{-15}	.09	rs3208856	C	T	2.99×10^{-22}	.76
PVR	2.97×10^{-10}	.14	6.30×10^{-2}	.62	1.13×10^{-9}	.39	rs1058402	G	A	1.71×10^{-9}	.27

Nature Genetics: doi:10.1038/ng.2852

Supplementary Table 12: Additional conditional association analyses in the *APOE* and *LDLR* loci. For the top variant (rs7412) in *APOE*, we repeated association analysis conditioning on rare non-synonymous variants with MAF< 5% in *BCAM*, *CBLC*, and *PVR*. For the top variants (rs6511720, rs2228671 and rs2738459) in *LDLR*, we repeated association analysis conditioning on rare non-synonymous variants in *LDLR* with MAF<5%. The p-values before and after conditional analysis are shown. The rs number, p-values before and after conditional analyses and the variant genotypes that are controlled for are displayed for each gene.

Gene	Variant	Conditional P-Value	Unconditional P-Value	Variants Conditioned On
APOE	rs7412	3.5×10^{-181}	8.2×10^{-214}	***BCAM***: rs28399653, rs28399654, rs200421757, rs143379896, rs200458600, rs199854072, rs139610351, rs138302587, rs199600463, rs149302547, rs200947707, rs199922856, rs145626518, rs117737673, rs150798131, rs200634102, rs139746192, rs144124876, rs28399626, rs148391498, rs28399630, rs9967601, rs141133602 ***CBLC***: rs3208856, rs35106910, rs149074838, rs114569424, rs116023028, rs115775900, rs35457630, rs137908794 ***PVR***: rs1058402, rs35365841, rs139267469, rs139528439, rs149458939, rs142546426
LDLR	rs6511720	4.9×10^{-39}	2.0×10^{-38}	***LDLR***: rs139791325, rs11669576, rs139624145, rs144172724, rs141673997, rs150673992, rs28942084, rs139043155, rs139361635, rs143992984, rs137853963, rs13306505, rs148698650, rs200727689, rs5928, rs146200173
LDLR	rs2228671	9.6×10^{-23}	4.1×10^{-22}	
LDLR	rs2738459	5.3×10^{-8}	3.8×10^{-8}	

Nature Genetics: doi:10.1038/ng.2852

Supplementary Note

We describe a framework for meta-analysis of rare variant association tests. The approach starts with meta-analysis of single variant association test statistics and then uses these to construct test statistics for genes or other functional units. We describe the implementation of several rare variant association tests and strategies for conditional analysis, which can provide a useful means of disentangling nearby signals. Finally, we propose a Monte Carlo simulation based strategy to evaluate significance levels empirically. The document also includes a brief summary of the simulations carried out in preparing our manuscript.

Notation

We consider constructing joint analysis statistics of rare variant association tests using multiple studies. For simplicity, we describe our strategy for analysis of a single gene, but the approach naturally extends to multiple genes. Let J be number of variant nucleotide sites genotyped (using arrays or sequencing) in at least one of the studies. For study k, let N_k denote the number of samples phenotyped and genotyped, and let the vector $\mathbf{y_k} = \left(Y_{1,k}, \cdots, Y_{N_k,k}\right)$ denote the quantitative trait (or quantitative trait residuals) each with variance σ_k^2. In all analyses reported here, we applied an inverse normal transformation to trait residuals prior to analysis. In our preliminary analyses, this transformation reduced the impact of non-normally distributed phenotypes and led to better-behaved quantile-quantile plots.

Within each study, we encode genotype information in a matrix:

$$\boldsymbol{X_k} = \begin{pmatrix} X_{1,1,k} & \cdots & X_{1,j,k} & \cdots & X_{1,J,k} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{i,1,k} & \cdots & X_{i,j,k} & \cdots & X_{i,J,k} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{N_k,1,k} & \cdots & X_{N_k,j,k} & \cdots & X_{N_k,J,k} \end{pmatrix}$$

Each entry in matrix $X_{i,j,k}$ represents the genotype individual i at site j, coded as the number of alternative alleles carried by the individual. We encode missing genotypes in the dataset as the average number of alternative alleles in individuals who are genotyped for that marker; alternatively, more advanced imputation algorithms (as implemented in MaCH[1], IMPUTE2[2] or BEAGLE[3]) could be used.

Nature Genetics: doi:10.1038/ng.2852

The multi-site genotype for individual i is denoted by the row vector $\mathbf{x}_{i,\bullet,k}$, and the genotypes for all N_k individuals at site j are given by column vector $\mathbf{x}_{\bullet,j,k}$. For ease of presentation, we define the mean genotype matrix $\bar{\boldsymbol{X}}_{\boldsymbol{k}}$, where the (i, j)-th element is $\left(\sum_i X_{i,j,k}\right)/N_k$ and the centered genotype matrix is $\boldsymbol{X}_{\boldsymbol{k}} - \bar{\boldsymbol{X}}_{\boldsymbol{k}}$.

Summary Statistics

For each study, we first calculate a vector of score statistics $\mathbf{u_k} = \left(\boldsymbol{X}_{\boldsymbol{k}} - \bar{\boldsymbol{X}}_{\boldsymbol{k}}\right)^T \mathbf{y_k}$ and a corresponding variance-covariance matrix $\boldsymbol{V}_{\boldsymbol{k}} = \hat{\sigma}_k^2 N_k \operatorname{cov}\left(\boldsymbol{X}_{\boldsymbol{k}}\right) = \hat{\sigma}_k^2 \left(\boldsymbol{X}_{\boldsymbol{k}} - \bar{\boldsymbol{X}}_{\boldsymbol{k}}\right)^T \left(\boldsymbol{X}_{\boldsymbol{k}} - \bar{\boldsymbol{X}}_{\boldsymbol{k}}\right)$. Then, to enable meta-analysis, we share the following summary statistics between studies:

a) Score statistics $\mathbf{u_k}$, which can be meta-analyzed across studies and then combined into gene-level statistics.

b) The covariance matrix for single variant score statistics $\boldsymbol{V}_{\boldsymbol{k}}$. This variance-covariance matrix will later allow us to calculate the distribution of gene-level statistics that result from combining several single variant score statistics. In principle, sharing the full matrix would allow the most flexibility when grouping variants into genes during meta-analysis and when executing conditional analyses. In practice, we make two simplifications. First, because the matrix is symmetric, we share only its upper triangle. Second, because most gene level tests group nearby variants, we share only covariance information for markers <1 Mb apart.

c) Estimated alternative allele frequencies for each marker $p_{j,k} = \sum_i X_{i,j,k} / 2N_k$, which can be used to decide which variants to analyze based on frequency.

d) Mean and variance for the quantitative trait residuals, for debugging purposes and for quality control in multi-sample analyses. As usual, these are $\hat{\mu}_k = \sum_i Y_{i,k} / N_k$ and $\hat{\sigma}_k^2 = \sum_i \left(Y_{i,k} - \hat{\mu}_k\right)^2 / N_k$.

e) Genotype call rate and p-values for testing Hardy-Weinberg equilibrium at each variant site, for quality control and to aid in variant filtering.

Meta-analysis of Single Variant Association Test Statistics

We first combine single variant association test statistics across studies using the Cochran-Mantel-Haenszel method. Specifically, we calculate a score statistic at each site as:

$$T_{j,\bullet} = U_{j,\bullet} \Big/ \sqrt{V_{j,j,\bullet}}$$

where $U_{j,\bullet} = \sum_k U_{j,k}$ and $V_{j,j,\bullet} = \sum_k V_{j,j,k}$. Cochran-Mantel-Haenszel statistics deal gracefully with very rare variants because $U_{j,k}$. and $V_{j,j,k}$ remain defined (as zero) even when a variant is monomorphic or missing in a study. For ease of presentation, we denote the vector of single variant association tests after meta-analysis as $\mathbf{u} = \sum_k \mathbf{u_k}$. Under the null, this vector is distributed as multivariate normal $\mathbf{u} \sim \mathrm{MVN}\left(\mathbf{0}, \sum_k \boldsymbol{V_k}\right)$.

Meta-Analysis of Gene-level Rare-Variant Association Tests

We consider two major types of rare variant association methods: (i) burden tests that assume all variants in a gene influence the trait in the same direction, such as the GRANVIL test by Morris and Zeggini[4] and (ii) methods that allow variants with opposite effects to reside in the same gene, such as the variance component score test implemented in SKAT by Wu et al[5]. Below, we show that both types of method can be derived in a regression model, which allows adjusting for covariates. Furthermore, we illustrate how the corresponding gene level statistics can be derived from single site meta-analysis statistics and how the information stored in the variance-covariance matrix is used when evaluating statistical significance.

Burden Tests That Assume Variants Have Similar Effect Sizes

For a simple burden test in study *k*, the impact of multiple rare variants in a region can be modeled using a shared regression coefficient β_{BURDEN} in a regression model that takes the form:

$$Y_{i,k} = \beta_{0,k} + \beta_{BURDEN} C_{BURDEN}\left(\mathbf{x_{i,\bullet,k}}\right) + \varepsilon_{i,k} \text{, where } \varepsilon_{i,k} \sim \mathrm{N}\left(0, \sigma_k^2\right)$$

Nature Genetics: doi:10.1038/ng.2852

$C_{BURDEN}\left(\mathbf{x_{i,\bullet,k}}\right)$ is a function that takes genotypes for a single individual as input and returns the rare variant burden for the gene being examined. Popular definitions for $C_{BURDEN}\left(\mathbf{x_{i,\bullet,k}}\right)$ include a simple sum statistic and a weighted sum statistic $C_{BURDEN}\left(\mathbf{x_{i,\bullet,k}}\right)=\sum_j \omega_j X_{i,j,k}$, where ω_j is the weight assigned to variant *j* according to its allele frequency or its computationally predicted functional impact[6,7]. Note that in this regression model, we allow the intercept $\beta_{0,k}$ and residual error σ_k^2 to vary between studies, but assume that β_{BURDEN} is shared across studies. For convenience of notation, we define a vector of nuisance parameters $\boldsymbol{\theta}_\mathbf{k}=\left(\beta_{0,k},\sigma_k^2\right)$ and $\boldsymbol{\theta}=\left(\boldsymbol{\theta}_\mathbf{1},\cdots,\boldsymbol{\theta}_\mathbf{K}\right)$, which are used in the likelihood function below.

As usual, the likelihood factors into a product of per study likelihoods:

$$L\left(\beta_{BURDEN},\boldsymbol{\theta}\middle|\mathbf{y},\boldsymbol{X}\right)=\prod_k L\left(\beta_{BURDEN},\boldsymbol{\theta}\middle|\mathbf{y_k},\boldsymbol{X_k}\right)$$

In joint analysis with individual level data, the score statistic is thus a sum for per study score statistics:

$$\begin{aligned}U_{BURDEN}&=\frac{\partial \log L\left(\beta_{BURDEN},\boldsymbol{\theta}\middle|\mathbf{y},\boldsymbol{X}\right)}{\partial\beta_{BURDEN}}\\&=\sum_k\frac{\partial \log L\left(\beta_{BURDEN},\boldsymbol{\theta}_\mathbf{k}\middle|\mathbf{y_k},\boldsymbol{X_k}\right)}{\partial\beta_{BURDEN}}\\&=\sum_k U_{BURDEN,k}\end{aligned}$$

Its variance can be derived using the Fisher information matrix, and following the derivations in Lin and Tang[7] (who studied a general framework for performing rare variant association tests), it can be shown that the variance for the score statistic in joint analysis of all individuals equals the sum of variances in each individual study $V_{BURDEN}=\sum_k V_{BURDEN,k}$. Therefore, when nuisance parameters are allowed to vary between studies, a score test for joint analysis of individual level

Nature Genetics: doi:10.1038/ng.2852

data (and allowing for study specific nuisance parameters) is equivalent to combining per study score statistics via the Cochran-Mantel-Haenszel method.

The arguments above show that the joint analysis statistic for gene-level tests can be constructed when a per-study $U_{BURDEN,k}$ statistic is shared. But, because of the simple relationship between burden and single variant score statistics in each study (specifically, $U_{\beta_{BURDEN},k} = \boldsymbol{\omega}^{\mathbf{T}}\mathbf{u_k}$), the joint analysis statistic for gene-level association tests can also be calculated when only single marker statistics are shared. Specifically, the burden test score statistics becomes:

$$U_{BURDEN} = \sum_k U_{BURDEN,k} = \sum_k \boldsymbol{\omega}^{\mathbf{T}}\mathbf{u_k} = \boldsymbol{\omega}^{\mathbf{T}}\mathbf{u}$$

Under the null, this statistic is approximately normally distributed with mean 0 and variance $V_{BURDEN} = \boldsymbol{\omega}^{\mathbf{T}}\left(\sum_k \boldsymbol{V_k}\right)\boldsymbol{\omega}$, enabling significance tests. Note that the regression coefficient β_{BURDEN} can be interpreted as a weighted average of single variant effects[8].

Variable Threshold Tests with an Adaptive Frequency Threshold

In variable threshold tests, rare variant burden statistics are calculated for each potential definition of "rare variant" and significance is evaluated for the maximum of these statistics. Typically, to calculate these statistics, all unique variant frequencies observed in a gene are listed and each of these frequencies is used as a potential frequency threshold. Frequency thresholds can be defined in terms of the pooled minor allele frequency or, sometimes, the pooled minor allele count (the two can differ depending on whether samples where a variant is missing are assumed to be wild type or unknown).

Given a specific variant frequency threshold *F* we define a corresponding burden score statistic as:

$$U_{BURDEN(F)} = \mathbf{v_F^T u}.$$

Here, $\mathbf{v}_{\mathbf{F}}$ is a vector of indicators where the j^{th} element equals 1 if the pooled minor allele frequency at variant site j is less than F and zero otherwise. For convenience of presentation and without loss of generality, we also define a matrix of indicators $\boldsymbol{\Phi} = \left(\mathbf{v}_{\mathbf{F}_1}, \mathbf{v}_{\mathbf{F}_2}, \cdots, \mathbf{v}_{\mathbf{F}_J}\right)$. The covariance between burden score statistics $U_{BURDEN(\phi)}$ and $U_{BURDEN(\phi^*)}$ calculated for thresholds F and F^*, is equal to $\Omega_{BURDEN(F),BURDEN(F^*)} = \mathbf{v}_{\mathbf{F}}^{\mathbf{T}}\left(\sum_k \boldsymbol{V}_k\right)\mathbf{v}_{\mathbf{F}^*}$. After burden statistics are calculated for each potential frequency threshold, they are standardized, dividing each statistic by its corresponding variance, and the maximum statistic is identified:

$$T_{VT} = \max_F \left\{T_{BURDEN(F)}\right\}, \text{ where } T_{BURDEN(F)} = U_{BURDEN(F)} \Big/ \sqrt{\Omega_{BURDEN(F),BURDEN(F)}}.$$

Significance for this statistic can be evaluated using the cumulative distribution function for multivariate normal distribution. Specifically, given the definition of the covariance between burden statistics calculated using different allele frequency thresholds, we have:

$$\left(T_{BURDEN(F_1)}, \cdots, T_{BURDEN(F_M)}\right) \sim \text{MVN}\left(\mathbf{0}, \boldsymbol{\Phi}^{\boldsymbol{T}}\left(\sum_k \boldsymbol{V}_k\right)\boldsymbol{\Phi}\right).$$

Significance tests can be calculated using standard methods for calculating multivariate normal integrals[9]:

$$\Pr\left(T_{VT} < t \mid \left(T_{BURDEN(F_1)}, \cdots, T_{BURDEN(F_J)}\right) \sim \text{MVN}\left(\mathbf{0}, \boldsymbol{\Phi}^{\boldsymbol{T}}\left(\sum_k \boldsymbol{V}_k\right)\boldsymbol{\Phi}\right)\right) =$$
$$\Pr\left(T_{BURDEN(F_1)} < t, \cdots, T_{BURDEN(F_J)} < t \mid \left(T_{BURDEN(F_1)}, \cdots, T_{BURDEN(F_J)}\right) \sim \text{MVN}\left(\mathbf{0}, \boldsymbol{\Phi}^{\boldsymbol{T}}\left(\sum_k \boldsymbol{V}_k\right)\boldsymbol{\Phi}\right)\right)$$

In practice, the covariance matrix for burden score statistics calculated using different frequency thresholds can be singular or nearly so, even when the variants are not in linkage disequilibrium. This occurs because $U_{BURDEN(\phi)}$ and $U_{BURDEN(\phi^*)}$ will typically be strongly correlated whenever ϕ and ϕ^* are similar and small. Evaluating the corresponding integrals can be numerically challenging. Thus, we recommend verifying analytical p-values using simulations (described later in this document).

Gene-level Tests that Assume a Distribution of Variant Effect Sizes (e.g. SKAT tests)

The simple burden test and variable threshold test described above can be underpowered when variants with opposite effects on the phenotype reside in the same gene and are grouped together, because the shared regression coefficient can average close to zero in that situation. One option in this setting would be to model the effects of each rare variant individually – but that strategy consumes many degrees of freedom and thus loses efficiency. Instead, we assume an underlying distribution of rare variance effect sizes with mean zero and test whether the variance of this distribution, τ^2, is greater than zero.

Specifically, we consider the model:

$$Y_{i,k} = \beta_{0,k} + \sum_j \beta_j X_{i,j,k} + \varepsilon_{i,k}, \text{ where } \varepsilon_{i,k} \sim \mathrm{N}\left(0, \sigma_k^2\right),$$

and make inferences about rare variant effect sizes $\boldsymbol{\beta} = \left(\beta_1, \beta_2, \cdots, \beta_J\right)$ by assuming these follow a common distribution with mean zero and variance τ^2. Under the null $\boldsymbol{\beta} = \mathbf{0}$, which is equivalent to $\tau^2 = 0$. Following Wu et al[5], we consider the likelihood:

$$\begin{aligned} L\left(\tau, \vec{\theta} \middle| \vec{Y}, \mathbf{X}\right) &= \int L\left(\boldsymbol{\beta}, \boldsymbol{\theta} \middle| \mathbf{y}, X\right) p\left(\boldsymbol{\beta} \middle| \tau\right) d\boldsymbol{\beta} \\ &= \int \prod_k L\left(\boldsymbol{\beta}, \boldsymbol{\theta}_\mathbf{k} \middle| \mathbf{y}_\mathbf{k}, \boldsymbol{X}_k\right) p\left(\boldsymbol{\beta} \middle| \tau\right) d\boldsymbol{\beta} \\ &= \int \exp\left(\sum_k \log\left(L\left(\boldsymbol{\beta}, \boldsymbol{\theta}_\mathbf{k} \middle| \mathbf{y}_\mathbf{k}, \boldsymbol{X}_k\right)\right)\right) p\left(\boldsymbol{\beta} \middle| \tau\right) d\boldsymbol{\beta} \end{aligned}$$

In order to derive the variance component score statistics for this likelihood, we apply a Laplace transformation to the marginal likelihood function, as suggested by Lin[10]. If repeating this calculation, please note that the integrand satisfies:

$$\exp\left(\log L\left(\boldsymbol{\beta},\hat{\boldsymbol{\theta}}_{\mathbf{k}}\middle|\mathbf{y}_{\mathbf{k}},\boldsymbol{X}_k\right)\right)=\exp\left(\log L\left(\mathbf{0},\hat{\boldsymbol{\theta}}_{\mathbf{k}}\middle|\mathbf{y}_{\mathbf{k}},\boldsymbol{X}_k\right)\right)\times$$
$$\left(1+tr\left(\vec{\beta}^T\left(\left(\frac{\partial\log L\left(\boldsymbol{\beta},\hat{\boldsymbol{\theta}}_{\mathbf{k}}\middle|\mathbf{y}_{\mathbf{k}},\boldsymbol{X}_k\right)}{\partial\boldsymbol{\beta}}\right)^T\frac{\partial\log L\left(\boldsymbol{\beta},\hat{\boldsymbol{\theta}}_{\mathbf{k}}\middle|\mathbf{y}_{\mathbf{k}},\boldsymbol{X}_k\right)}{\partial\boldsymbol{\beta}}+\frac{\partial^2\log L\left(\boldsymbol{\beta},\hat{\boldsymbol{\theta}}_{\mathbf{k}}\middle|\mathbf{y}_{\mathbf{k}},\boldsymbol{X}_k\right)}{\partial\boldsymbol{\beta}^2}\right)\vec{\beta}\right)+o\left(|\boldsymbol{\beta}|^2\right)\right)$$

Then, following the argument in Lin[10], it can be shown that the variance component score statistics in the joint analysis with individual level data for testing $\tau=0$ is:

$$\begin{aligned}Q&=\left(\sum\nolimits_k\frac{\partial\log L\left(\boldsymbol{\beta},\hat{\boldsymbol{\theta}}_{\mathbf{k}}\mathbf{y}_{\mathbf{k}},\boldsymbol{X}_k\right)}{\partial\boldsymbol{\beta}}\right)^T\mathbf{K}\left(\sum\nolimits_k\frac{\partial\log L\left(\boldsymbol{\beta},\hat{\boldsymbol{\theta}}_{\mathbf{k}}\mathbf{y}_{\mathbf{k}},\boldsymbol{X}_k\right)}{\partial\boldsymbol{\beta}}\right)\\&=\left(\sum\nolimits_k\mathbf{u}_{\mathbf{k}}\right)^T\left(\sum\nolimits_k\mathbf{u}_{\mathbf{k}}\right)\end{aligned}$$

Therefore, gene level statistics that allow for a distribution of rare variant effect sizes (rather than assuming a shared regression coefficient) can also be constructed after meta-analysis of single variant statistics. In this test, Q is a quadratic function of single site meta-analysis statistics, in contrast to burden statistics defined in previous sections, which were linear functions of single site statistics. In practice, weights can also be assigned in the variance component score statistics, and the test statistic takes the form of $Q=\left(\sum_k\mathbf{u}_{\mathbf{k}}\right)^T\boldsymbol{K}\left(\sum_k\mathbf{u}_{\mathbf{k}}\right)$. The matrix $\boldsymbol{K}$ is the kernel that compares multi-site genotypes. A default choice is $\boldsymbol{K}=diag\left(\omega_1,\omega_2,\cdots,\omega_J\right)$, with ω_j being the weight assigned to variant site j[5]. The statistic Q follows a mixture chi-square distribution[11], with mixture proportions given by the eigenvalues for the matrix $\left(\sum_k\boldsymbol{V}_k\right)^{1/2}\boldsymbol{K}\left(\sum_k\boldsymbol{V}_k\right)^{1/2}$.

Monte Carlo Method for Empirical Assessment of Significance

The previous sections describe how a series of gene-level test statistics can be calculated and, for each one, proposes a strategy for evaluating significance. In practice, evaluating the required numerical integrals can be challenging, because variance-covariance matrices describing the relationship between single variant statistics or burden scores evaluated at different thresholds can be singular or nearly so. In this section, we describe a simple strategy for re-sampling plausible sets of single marker test statistics. Gene level statistics can then be evaluated for each of these simulated vectors of single

marker test statistics and used to assess significance empirically, avoiding some of the problems inherent with numerical integration.

Recall that test statistics are distributed as:

$$\sum_k \mathbf{u_k} = \sum_k \mathbf{y_k^T}\left(\boldsymbol{X_k} - \bar{\boldsymbol{X}}_k\right) \sim \text{MVN}\left(\mathbf{0}, \sum_k \boldsymbol{V_k}\right)$$

To evaluate significance empirically, we sample random vectors from the distribution $\text{MVN}\left(\mathbf{0}, \sum_k \boldsymbol{V_k}\right)$ and calculate gene level rare variant test statistics for each of these sampled random vectors, allowing us to obtain an empirical distribution for any gene-level statistic[12]. As usual, p-values can then be evaluated by comparing the test statistics for the original data with those in this empirical distribution. For computational efficiency, we use an adaptive algorithm where a larger number of vectors are sampled when assessing small p-values and fewer vectors are sampled when assessing larger p-values[13]. Specifically, we continue sampling new vectors until the number of sampled statistics that are greater than the statistic in the original data exceeds a particular threshold (100, unless noted otherwise) or the total number of sampled vectors exceeds a predefined limit (40,000,000; unless noted otherwise).

Conditional Analyses

It is well known that, due to linkage disequilibrium, one or more common causal variants can result in shadow association signals at other nearby common variants. As illustrated in our analysis of the *APOE* locus in the text, common variant association signals can also result in shadow rare variant association signals at nearby genes. In unpublished data, we have observed many other examples where rare variant signals are shadows of other nearby common or rare variants. Conditional analysis provides a useful procedure for disentangling neighboring association signals in this setting, by checking whether weaker signals remain significant after conditioning on nearby stronger signals.

For common variants, Yang et al[14] have shown that linkage disequilibrium relationships between variants, estimated from external reference panels, can be used to enable conditional analysis in meta-analysis settings. For rare variants and gene-level tests, accurately describing relationships between variants is crucial and we recommend against the use of external

reference panels. Instead, we recommend using linkage disequilibrium relationships estimated in the samples being analyzed and summarized in the variance-covariance matrix of single variant score statistics.

To describe our strategy for conditional analyses, we first decompose the genotype matrix into two components: a matrix of genotypes $\boldsymbol{X_k}$ for variants to be tested for independent association and a matrix of genotypes $\boldsymbol{Z_k}$ for variants that should be included as covariates in the null regression model (and thus controlled for). In order to facilitate presentations, we denote $\boldsymbol{W_k} = (\boldsymbol{X_k}, \boldsymbol{Z_k})$.

Conditional Analysis for Gene Level Tests That Use A Shared Regression Coefficient for Rare Variants

When individual level data is available, conditional analysis considers a model similar to:

$$Y_{i,k} = \beta_{0,k} + \beta_{BURDEN} C_{BURDEN}\left(\vec{X}_{i,\bullet,k}\right) + \vec{\alpha}_k^T \vec{Z}_{i,k} + \varepsilon_{i,k}$$

This analysis could be readily carried out by repeating analysis and re-calculating score statistics for each study, but this is not required. Instead, the score statistics that result from the conditional analysis described above can be readily estimated using summary information.

Let $\tilde{T}_{\beta_{BURDEN},k}, \tilde{U}_{\beta_{BURDEN},k}$ and $\tilde{V}_{\beta_{BURDEN},\beta_{BURDEN},k}$ denote test statistics, score statistics and their variances from conditional analysis; analogous to statistics previously defined for unconditional analysis. As usual, $\tilde{T}_{\beta_{BURDEN},k} = \tilde{U}_{\beta_{BURDEN},k} \Big/ \sqrt{\tilde{V}_{\beta_{BURDEN},\beta_{BURDEN},k}}$. To derive the component statistics, we use the approach of Lin and Tang[7], to show that:

$$\tilde{U}_{\beta_{BURDEN},k} = \left(\left(\mathbf{y_k} - \bar{Y}_k\right) - \left(\boldsymbol{Z_k} - \bar{\boldsymbol{Z}}_{\boldsymbol{k}}\right)\hat{\boldsymbol{\alpha}}_{\mathbf{k}}\right)^T \boldsymbol{X_k}$$

$$\text{with } \hat{\boldsymbol{\alpha}}_{\mathbf{k}} = \left(\left(\boldsymbol{Z_k} - \bar{\boldsymbol{Z}}_{\boldsymbol{k}}\right)^{\mathsf{T}}\left(\boldsymbol{Z_k} - \bar{\boldsymbol{Z}}_{\boldsymbol{k}}\right)\right)^{-1}\left(\boldsymbol{Z_k} - \bar{\boldsymbol{Z}}_{\boldsymbol{k}}\right)^{\mathsf{T}} \mathbf{y_k},$$

and that:

$$\tilde{V}_{\beta_{BURDEN},\beta_{BURDEN},k} = \hat{\phi}^2 \boldsymbol{\omega}^{\mathbf{T}} \left(\left(\boldsymbol{X}_k - \bar{\boldsymbol{X}}_k \right)^T \left(\boldsymbol{X}_k - \bar{\boldsymbol{X}}_k \right) \right) \boldsymbol{\omega}$$
$$- \hat{\phi}^2 \boldsymbol{\omega}^{\mathbf{T}} \left(\left(\boldsymbol{X}_k - \bar{\boldsymbol{X}}_k \right)^T \left(\boldsymbol{Z}_k - \bar{\boldsymbol{Z}}_k \right) \right) \left(\left(\boldsymbol{Z}_k - \bar{\boldsymbol{Z}}_k \right)^T \left(\boldsymbol{Z}_k - \bar{\boldsymbol{Z}}_k \right) \right)^{-1} \left(\left(\boldsymbol{Z}_k - \bar{\boldsymbol{Z}}_k \right)^T \left(\boldsymbol{X}_k - \bar{\boldsymbol{X}}_k \right) \right) \boldsymbol{\omega} ,$$

$$\text{with } \hat{\phi}^2 = 1/N_k \times \left(\left(\mathbf{y}_\mathbf{k} - \bar{Y}_k \right) - \left(\boldsymbol{Z}_k - \bar{\boldsymbol{Z}}_k \right) \hat{\boldsymbol{\alpha}}_\mathbf{k} \right)^T \left(\left(\mathbf{y}_\mathbf{k} - \bar{Y}_k \right) - \left(\boldsymbol{Z}_k - \bar{\boldsymbol{Z}}_k \right) \hat{\boldsymbol{\alpha}}_\mathbf{k} \right)$$

Now, we can verify that $\tilde{U}_{\beta_{BURDEN},k}$ and $\tilde{V}_{\beta_{BURDEN},\beta_{BURDEN},k}$ can be calculated using shared summary level statistics, because all key terms in the above equations can be extracted from the list of single variant score statistics and from the variance-covariance matrix of single marker association test statistics (which we have shared), since

$$\left(\boldsymbol{W}_k - \bar{\boldsymbol{W}}_k \right)^{\mathbf{T}} \left(\boldsymbol{W}_k - \bar{\boldsymbol{W}}_k \right) = \begin{pmatrix} \left(\boldsymbol{X}_k - \bar{\boldsymbol{X}}_k \right)^T \left(\boldsymbol{X}_k - \bar{\boldsymbol{X}}_k \right) & \left(\boldsymbol{X}_k - \bar{\boldsymbol{X}}_k \right)^T \left(\boldsymbol{Z}_k - \bar{\boldsymbol{Z}}_k \right) \\ \left(\boldsymbol{Z}_k - \bar{\boldsymbol{Z}}_k \right)^T \left(\boldsymbol{X}_k - \bar{\boldsymbol{X}}_k \right) & \left(\boldsymbol{Z}_k - \bar{\boldsymbol{Z}}_k \right)^T \left(\boldsymbol{Z}_k - \bar{\boldsymbol{Z}}_k \right) \end{pmatrix} .$$

Finally, meta-analysis burden score statistics can be calculated as:

$$\tilde{T}_{\beta_{BURDEN},\bullet} = \sum_k \tilde{U}_{\beta_{BURDEN},k} \Big/ \sqrt{\sum_k \tilde{V}_{\beta_{BURDEN},\beta_{BURDEN},k}} .$$

Thus, conditional meta-analysis statistics can be calculated from shared single variant statistics and their variance-covariance matrix, as desired.

Conditional Analysis for Tests that Assume a Distribution of Rare Variant Effect Sizes (e.g. SKAT)

Similar arguments can be used to derive formulae for conditional analysis of rare variant association tests in settings where direction of effect and effect sizes are allowed to vary between markers. In that setting, we follow the approach of Wu et al[5]. The variance component score test takes the form:

$$\tilde{Q}_k = \left(\left(\mathbf{y}_\mathbf{k} - \bar{Y}_k \right) - \left(\boldsymbol{Z}_k - \bar{\boldsymbol{Z}}_k \right) \hat{\boldsymbol{\alpha}}_\mathbf{k} \right)^T \boldsymbol{X}_k \times \mathbf{K} \times \boldsymbol{X}_k^T \left(\left(\mathbf{y}_\mathbf{k} - \bar{Y} \right) - \left(\boldsymbol{Z}_k - \bar{\boldsymbol{Z}}_k \right) \hat{\boldsymbol{\alpha}}_\mathbf{k} \right)$$

Following the derivation for unconditional analysis, the meta-analysis test statistic is given by :

Nature Genetics: doi:10.1038/ng.2852

$$\tilde{Q} = \left(\sum_k \left(\left(\mathbf{y_k} - \bar{Y}_k\right) - \left(\boldsymbol{Z}_k - \bar{\boldsymbol{Z}}_k\right)\hat{\boldsymbol{\alpha}}_\mathbf{k}\right)^T \boldsymbol{X}_k\right) \times \boldsymbol{K} \times \left(\sum_k \boldsymbol{X}_k^T \left(\left(\mathbf{y_k} - \bar{Y}_k\right) - \left(\boldsymbol{Z}_k - \bar{\boldsymbol{Z}}_k\right)\hat{\boldsymbol{\alpha}}_\mathbf{k}\right)\right)$$

Then, noting that the single variant score statistics $\sum_k \left(\left(\mathbf{y_k} - \bar{Y}_k\right) - \left(\boldsymbol{Z}_k - \bar{\boldsymbol{Z}}_k\right)\hat{\boldsymbol{\alpha}}_\mathbf{k}\right)^T \boldsymbol{X}_k$ follow a multivariate normal distribution with mean zero and variance-covariance matrix $\sum_k \tilde{\boldsymbol{V}}_k$, where

$$\tilde{\boldsymbol{V}}_k = \hat{\phi}^2 \left(\left(\boldsymbol{X}_k - \bar{\boldsymbol{X}}_k\right)^T \left(\boldsymbol{X}_k - \bar{\boldsymbol{X}}_k\right)\right)$$
$$-\hat{\phi}^2 \left(\left(\boldsymbol{X}_k - \bar{\boldsymbol{X}}_k\right)^T \left(\boldsymbol{Z}_k - \bar{\boldsymbol{Z}}_k\right)\right)\left(\left(\boldsymbol{Z}_k - \bar{\boldsymbol{Z}}_k\right)^T \left(\boldsymbol{Z}_k - \bar{\boldsymbol{Z}}_k\right)\right)^{-1} \left(\left(\boldsymbol{Z}_k - \bar{\boldsymbol{Z}}_k\right)^T \left(\boldsymbol{X}_k - \bar{\boldsymbol{X}}_k\right)\right)$$

It is straightforward to show that $\tilde{Q}$ follows a mixture chi-square distribution with mixture proportions being the eigenvalues of $\left(\sum_k \tilde{\boldsymbol{V}}_k\right)^{1/2} \boldsymbol{K} \left(\sum_k \tilde{\boldsymbol{V}}_k\right)^{1/2}$. Therefore, score statistics and the variance-covariance matrix for single marker statistics, are sufficient to enable derivation of the test statistics and p-values for conditional meta-analysis.

Extension to Analyze Samples with Known or Cryptic Relateness

Using linear mixed models, our methods and tools can also be used to perform meta-analysis of studies that include related samples. The linear mixed model for analyzing single variant associations takes the form

$$Y_{i,k} = \beta_{0,k} + \beta_{i,j,k} X_{i,j,k} + g_{i,k} + e_{i,k},$$

where $g_{i,k}$ and $e_{i,k}$ are the polygenic and non-shared residual variance component, respectively. For two individuals with a shared genetic background, the unobserved polygenic variance component can be correlated and modeled as $\mathbf{g}_{\bullet,\mathbf{k}} \sim \mathrm{MVN}\left(\mathbf{0}, \sigma_g^2 \boldsymbol{K}\right)$, where the kinship matrix $\boldsymbol{K}$ can be inferred from available pedigree structure or deduced from available genotype data. The non-shared residual variance component is assumed to be independent between individuals and satisfies $\mathbf{e}_{\bullet,\boldsymbol{k}} \sim \mathrm{MVN}\left(\vec{0}, \sigma_e^2 \boldsymbol{I}\right)$.

The covariance matrix for the quantitative trait is given by

$$\boldsymbol{\Sigma}_\boldsymbol{k} = \sigma_g^2 \boldsymbol{K} + \sigma_e^2 \boldsymbol{I}.$$

Then, we defined the single variant score statistics to be shared as $\mathbf{u_k} = \left(\boldsymbol{X}_k - \bar{\boldsymbol{X}}_k\right)^T \boldsymbol{\Sigma}_k^{-1} \mathbf{y_k}$ and their variance-covariance matrix as $\boldsymbol{V}_k = \left(\boldsymbol{X}_k - \bar{\boldsymbol{X}}_k\right)^T \boldsymbol{\Sigma}_k^{-1} \left(\boldsymbol{X}_k - \bar{\boldsymbol{X}}_k\right)$. These statistics can be readily incorporated in all the previous formulae for

Nature Genetics: doi:10.1038/ng.2852

burden tests, variable threshold tests, sequence kernel association tests, and conditional analysis. Furthermore, just like with unrelated individuals, meta-analysis of studies that include related individuals, remains equivalent to an analysis that pools individual level data and allows for study specific nuisance parameters, provided no related or duplicated individuals are present across studies.

Analysis of Binary Traits

For binary trait analyses, it is common to use logistic regression:

$$\log\left(\frac{\Pr\left(Y_{i,k}=1\right)}{\Pr\left(Y_{i,k}=0\right)}\right)=\beta_{0,k}+\beta_{j,k}X_{i,j,k}+\sum_l \gamma_{l,k}C_{i,l,k}$$

In this setting, single variant statistics for marker *j* are defined as:

$U_j=\sum_i X_{i,j,k}\left(Y_{i,k}-\hat{Y}_{i,k}\right)$, where $\hat{Y}_{i,k}=\frac{\exp\left(\hat{\beta}_{0,k}+\sum_l \hat{\gamma}_{l,k}C_{i,l,k}\right)}{1+\exp\left(\hat{\beta}_{0,k}+\sum_l \hat{\gamma}_{l,k}C_{i,l,k}\right)}$, and the parameters $\hat{\beta}_{0,k}$, $\hat{\gamma}_{l,k}, l=1,\cdots,L$ are

estimated using the null model

$$\log\left(\frac{\Pr\left(Y_{i,k}=1\right)}{\Pr\left(Y_{i,k}=0\right)}\right)=\beta_{0,k}+\sum_l \gamma_{l,k}C_{i,l,k}$$

The variance-covariance matrix between single variant statistics $\mathbf{u_k}=\left(U_{1,k},\cdots,U_{J,k}\right)$ has the form

$\boldsymbol{V_k}=\boldsymbol{X_k^T P_k X_k}-\left(\boldsymbol{X_k^T P_k \tilde{C}_k}\right)\left(\boldsymbol{\tilde{C}_k^T P_k \tilde{C}_k}\right)^{-1}\left(\boldsymbol{\tilde{C}_k^T P_k X_k}\right)$, where

$\boldsymbol{P_k}=diag\left(\hat{Y}_{1,k}\left(1-\hat{Y}_{1,k}\right),\cdots,\hat{Y}_{N_k,k}\left(1-\hat{Y}_{N_k,k}\right)\right)$, and $\boldsymbol{\tilde{C}_k}$ is the matrix of covariates augmented by a column of 1's.

Burden, variable threshold and SKAT tests can then be implemented the same way as in the analysis of quantitative traits. An important caution is that when the total number of rare variants carriers is small or the number of cases and controls in each study is very unbalanced, asymptotic distributions for the burden, variable threshold and sequence kernel association tests may not be accurate. Because it assumes a continuous distribution for test statistics, we do not recommend

application of our current Monte-Carlo framework to binary trait analyses. This is concordant with observations by Lin and Tang [7] and Lee et al [11].

Simulation of Population Genetic Data

We simulated haplotypes using a coalescent model and the program *ms*[15]. We chose a demographic model consistent with European demographic history[16], including an ancestral bottleneck followed by more recent population differentiation and exponential growth (**Supplementary Figure 1**). Model parameters were based upon estimates from large scale sequencing studies[17], tuned such that measures of genetic diversity between simulated sub-populations match estimates from European samples[18]. The simulated haplotypes had an average pairwise sequence difference of $\pi = 0.001$ and an average $F_{ST} = 0.004$. Furthermore, when 5000 haplotypes were sampled, a typical simulated 5000 base pair region included ~100 variant sites, of which 80% had MAF<1% and ~49% were singletons. For any two pools of 5000 haplotypes sampled from different subpopulations, ~3% of variants with MAF<1% and 35% of the variants with MAF>1% are shared, consistent with expectations from population genetics and observations from real data[17,19].

Our model assumes an ancestral population with effective population size of $N_1 = 10{,}000$ where an instantaneous bottleneck event 3,000 generations in the past reduced population size to $N_{bottleneck}$ = 75. Then, our simulations assume that this population simultaneously split into present day populations 500 generations before the present. Following the divergence from the ancestral population, the present-day populations underwent recent exponential growth, each growing to a present day effective population size of $N_0 = 1 \times 10^6$ over 400 generations. We assume equal, symmetric migration rates between the sub-populations with a per-haplotype, per-generation migration rate of 5×10^{-4}. We also assume a per-basepair, per-generation mutation rate of 2.5×10^{-8} and a recombination rate equivalent to 1cM/Mb.

Type I Error Rate and Power

Using simulated genetic data, we estimated power and type I error for each of the methods described here. We considered three representative rare variant tests association tests: a simple burden test (with MAF thresholds of 1% and 5%), a variable threshold association test, and the SKAT test. First, we generated 50,000,000 null replicates to evaluate type I

error rates in meta-analyses of 3, 6 and 9 samples of 1000 individuals at significance level α=0.001, 0.0001, or 2.5×10^{-6}. As shown in **Supplementary Table 1**, the type I error rates are well controlled.

Next, the power for different rare variant tests in meta-analysis was evaluated for a series of genetic models. Three phenotype models were simulated: 1.) half of low frequency variants with MAF < 0.5% are causal, each increasing expected trait values by 1/4 standard deviation; 2.) half of all variants are causal, irrespective of frequency, and increase trait values by 1/4 standard deviation; 3.) 50% of the variants are casual, irrespective of frequency, and 80% of these increase expected trait values by 1/4 standard deviation, while the remaining 20% decrease trait values by the same amount. Between 2 and 100 studies of 1000 individuals each were simulated. Meta-analysis for variants with MAF<5% was performed using our approach or using Fisher's method and the minimal p-value approach to combine burden test, SKAT and variable threshold (VT) test statistics. Power was evaluated at threshold $\alpha=2.5\times10^{-6}$ using 10,000 replicates.

Figure 1 and **Supplementary Figure 7** summarize the results of our power simulations, considering meta-analysis of up to 100 samples of 1000 individuals each. Several patterns are clear from the figures. First, for the effect sizes simulated here, very large sample sizes may be required to ensure adequate power. In some settings, power only reaches ~60% in analyses that include ~100,000 individuals, even using the most powerful available test. Second, we did not find a universally most powerful method, emphasizing the utility of approaches such as ours that can be extended to implement a diverse set of test statistics. Typically, we found that when the proportion of non-causal variants is high or causal variants can have opposite effects, the SKAT was more powerful. When causal variants have effects in the same direction, simpler burden tests were more powerful. Third, in all the simulation scenarios considered, our method greatly outperforms these alternative methods for meta-analysis, especially when information is combined across a large number of samples.

Evaluation of Conditional Analysis Strategy

As described in our analysis of genes neighboring *APOE*, common variant association signals can produce inflated rare variant test statistics at nearby genes due to linkage disequilibrium. To evaluate our strategy for conditional analysis of rare variant association tests, we selected one common variant with pooled MAF>10% as causal and increasing mean trait values by 0.25 standard deviation per copy. We then evaluated the type I error rate of gene-level rare variant association

test statistics (**Supplementary Figure 13**). The results show that, without conditioning, p-values deviate substantially from null expectations. The results also show that, after conditioning, p-values for rare variant association tests behave as expected under the null.

Meta-Analysis of Lipid Traits

Summary statistics were calculated for each participating study and shared to enable a central meta-analysis. In single variant and gene-base rare variant association analysis, age, age^2, sex and cohort specific covariates, such as principal components of ancestry were included in the analysis. Trait residuals were standardized using an inverse normal transformation.

STUDY DESCRIPTIONS

To illustrate applications of our method, we perform meta-analysis of lipids traits using 18,699 individuals, a sample size that is expected to identify several signals (see Willer et al[20] and Kathiresan et al[21] for example).

Malmö Diet and Cancer Study – Cardiovascular Cohort (MDC-CC)

The Malmö Diet and Cancer Study[22] is a community-based prospective epidemiologic cohort of 28,449 persons recruited for a baseline examination between 1991 and 1996. From this cohort, 6,103 persons were randomly selected to participate in the cardiovascular cohort, which sought to investigate risk factors for cardiovascular disease. All participants underwent a medical history assessment and a physical examination.

Women's Health Initiative

The WHI[23] encompasses four randomized clinical trials as well as a prospective cohort study of 161,808 post-menopausal women aged 50–79, recruited (1993–1998) and followed up at 40 centers across the US. Samples examined here were genotyped as part of the NHLBI Exome Sequencing Project.

Ottawa Heart Study

Cases and controls were recruited from either the lipid clinic at the University of Ottawa Heart Institute or the cardiac catheterization laboratory[24]. All cases were required to have at least one of: a stenosis in a major epicardial vessel of at least 50%; have had a percutaneous intervention (PCI); have had coronary artery bypass surgery (CABG); or have had a myocardial infarction (MI). Cases with diabetes mellitus were excluded. Age of onset of CAD was required to be ≤55 years old for men and ≤65 years old for women. Controls were either healthy asymptomatic elderly individuals or were recruited through the catheterization laboratory with no stenosis ≥50% in any major epicardial vessel and were required to be ≥65 years old for men and ≥70 years old for women.

PROCARDIS

The PROCARDIS[25] "genetically-enriched" case collection is composed of sibships (proband and at least one affected sibling) with coronary disease. Ascertainment criteria for PROCARDIS probands were myocardial infarction (MI) or symptomatic acute coronary syndrome before the age of 66 years. For each of the coronary disease cases included in the "genetically-enriched" case-control study, it was planned to recruit one control of the same sex, ethnicity and within 5 years of age of cases, with no personal or sibling history of coronary disease before age 66 years. In the UK, controls were identified by mailing a self-administered questionnaire to spouses or siblings of spouses or male friends of any individuals who had previously returned a completed questionnaire to the PROCARDIS study. Eligible respondents were asked to attend their general practice to have their blood pressure, height and weight recorded, and to provide a blood sample. In Sweden, Italy and Germany, controls were selected from population registers and invited to attend a special clinic to have their blood pressure, height and weight recorded, to provide a blood sample and to complete a self-administered questionnaire.

HUNT – The Nord-Trøndelag Health Study

The HUNT study has been described in detail previously[26]. The HUNT study is a population based health study with personal and family medical histories on 106,436 people from Nord-Trøndelag County, Norway, collected during three phases from 1984 to 2008. A subsample of 5,869 individuals were successfully genotyped on the iSelect Exomechip V1.0 (Illumina, San Diego, CA), 2,928 cases with retrospectively hospital diagnosed myocardial infarction and 2,941 healthy

Nature Genetics: doi:10.1038/ng.2852

controls matched on sex, birth year and municipality. Genotype calling was done using GenTrain version 2.0 in GenomeStudio V2011.1 (Illumina, San Diego, CA) in combination with zCall version 2.2 [27].

References:

1. Yan, Y.N. *et al.* Synthesis and optical behaviors of 2-(9-phenanthrenyl)-, 2-(9-anthryl)-, and 2-(1-pyrenyl)-1-alkylimidazole homologues. *Spectrochim Acta A Mol Biomol Spectrosc* **74**, 233-42 (2009).
2. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
3. Browning, B.L. & Browning, S.R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* **84**, 210-23 (2009).
4. Morris, A.P. & Zeggini, E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* (2009).
5. Wu, M.C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**, 82-93 (2011).
6. Madsen, B.E. & Browning, S.R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* **5**, e1000384 (2009).
7. Lin, D.Y. & Tang, Z.Z. A general framework for detecting disease associations with rare variants in sequencing studies. *Am J Hum Genet* **89**, 354-67 (2011).
8. Liu, D.J. & Leal, S.M. Estimating genetic effects and quantifying missing heritability explained by identified rare-variant associations. *Am J Hum Genet* **91**, 585-96 (2012).
9. Genz, A. Numerical Computation of Multivariate Normal Probabilities. *Journal of Computational and Graphical Statistics* **1**, 141-149 (1992).
10. Lin, X. Variance component testing in generalised linear models with random effects. *Biometrika* **84**, 309-326 (1997).
11. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* **91**, 224-37 (2012).
12. Zou, F., Fine, J.P., Hu, J. & Lin, D.Y. An efficient resampling method for assessing genome-wide statistical significance in mapping quantitative trait Loci. *Genetics* **168**, 2307-16 (2004).
13. Besag, J. & Clifford, P. Sequential Monte Carlo p-values. *Biometrika* **78**, 301-304 (1991).
14. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* **44**, 369-75, S1-3 (2012).
15. Hudson, R.R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337-8 (2002).
16. Kryukov, G.V., Shpunt, A., Stamatoyannopoulos, J.A. & Sunyaev, S.R. Power of deep, all-exon resequencing for discovery of human trait genes. *Proc Natl Acad Sci U S A* **106**, 3871-6 (2009).
17. Coventry, A. *et al.* Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun* **1**, 131 (2010).
18. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98-101 (2008).
19. Nelson, M.R. *et al.* An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337**, 100-4 (2012).
20. Willer, C.J. *et al.* Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet* **40**, 161-9 (2008).
21. Kathiresan, S. *et al.* Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat Genet* **40**, 189-97 (2008).
22. Kathiresan, S. *et al.* Polymorphisms associated with cholesterol and risk of cardiovascular events. *N Engl J Med* **358**, 1240-9 (2008).
23. Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. *Control Clin Trials* **19**, 61-109 (1998).
24. McPherson, R. *et al.* A common allele on chromosome 9 associated with coronary heart disease. *Science* **316**, 1488-91 (2007).
25. Clarke, R. *et al.* Genetic variants associated with Lp(a) lipoprotein level and coronary disease. *N Engl J Med* **361**, 2518-28 (2009).
26. Krokstad, S. *et al.* Cohort Profile: The HUNT Study, Norway. *Int J Epidemiol* (2012).

27. Goldstein, J.I. *et al.* zCall: a rare variant caller for array-based genotyping: genetics and population analysis. *Bioinformatics* **28**, 2543-5 (2012).