The American Journal of Human Genetics, Volume *101*

Supplemental Data

Prospects of Fine-Mapping Trait-Associated Genomic Regions by Using Summary Statistics from Genome-wide Association Studies

Christian Benner, Aki S. Havulinna, Marjo-Riitta Järvelin, Veikko Salomaa, Samuli Ripatti, and Matti Pirinen

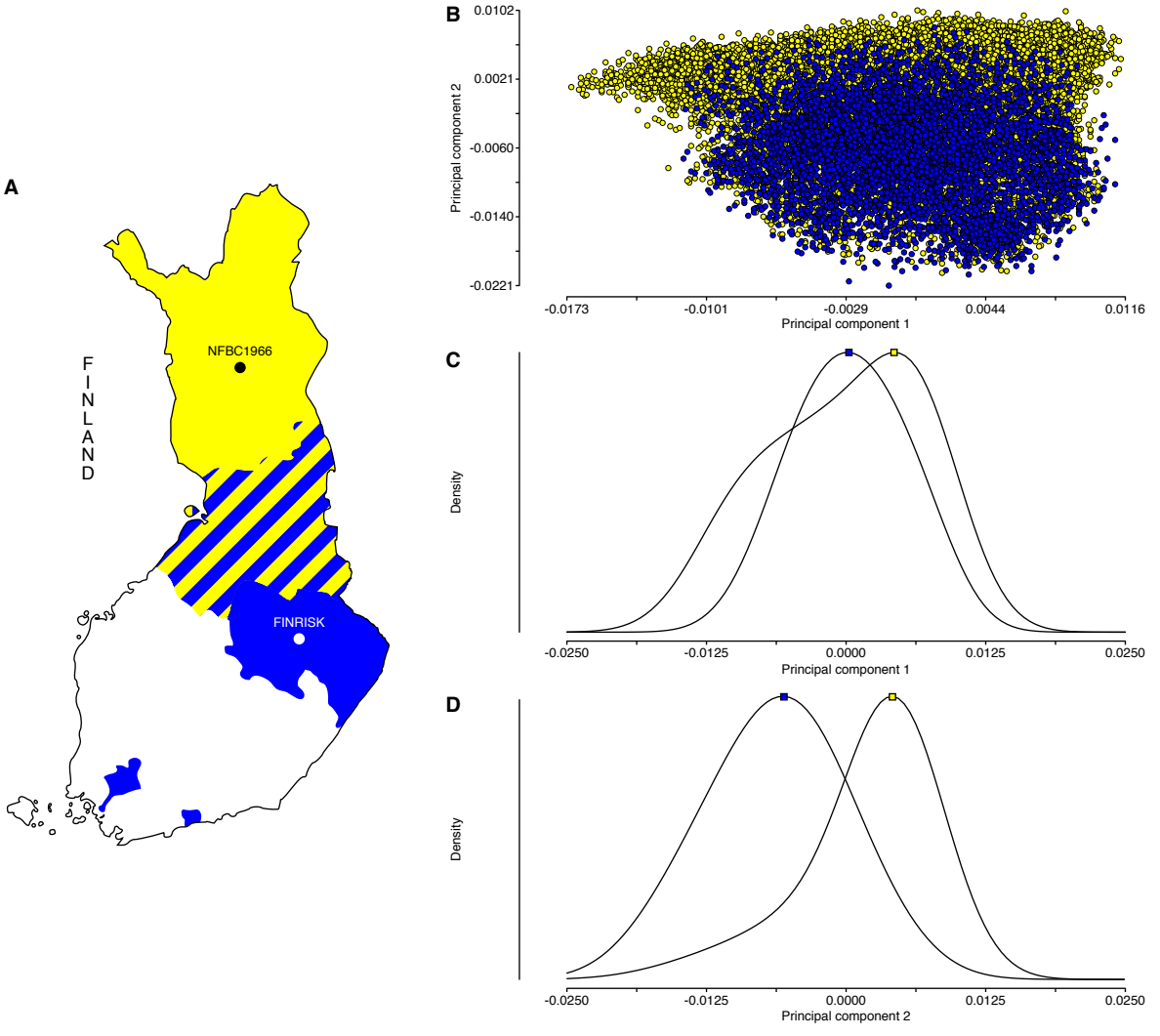

Figure S1 | Genetic structure of individuals in two Finnish cohorts. (**A**) The FINRISK study is a cross-sectional survey of the Finnish working age population collected at five centers (●). The Northern Finland Birth Cohort 1966 (NFBC1966) is a longitudinal study of individuals from the provinces of Oulu and Lapland in the north of Finland (○). Genetically, NFBC1966 is not a perfect match to FINRISK although individuals in both studies are collected within Finland. (**B-D**) First two principal components from an analysis of genotype data on 20,626 individuals included in the FINRISK surveys from 1992 to 2007 and 5,363 individuals from NFBC1966.

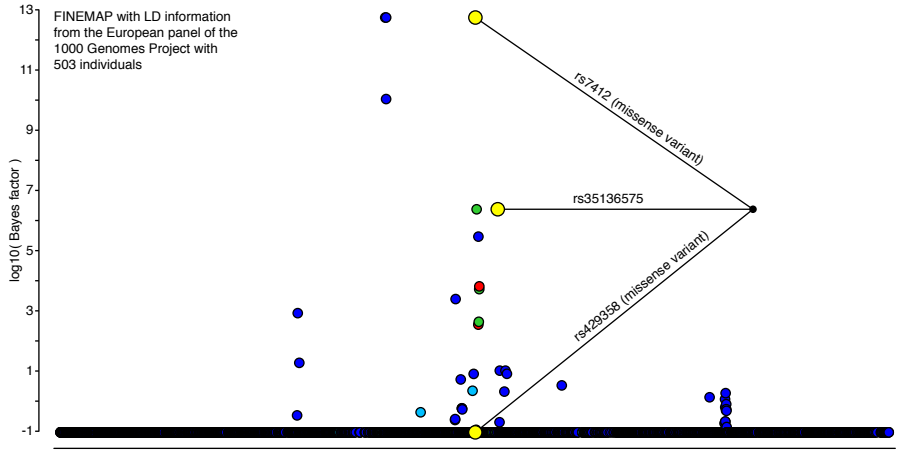

Figure S2 | Fine-mapping 19q13/*APOE* region associated with LDL-C. Bayes factors (log10) are shown from a FINEMAP analysis of 3,078 variants with minor allele frequency above 1% covering 1 Mb of the genome. GWAS summary statistics were computed using 15,626 individuals from the FINRISK study. LD information was obtained from the reference genotypes of 503 European individuals in the 1000 Genomes Project. Variants identified by a standard conditional analysis are highlighted in yellow. All other variants are colored with respect to their LD (absolute value of Pearson correlation) with the lead variant rs7412.

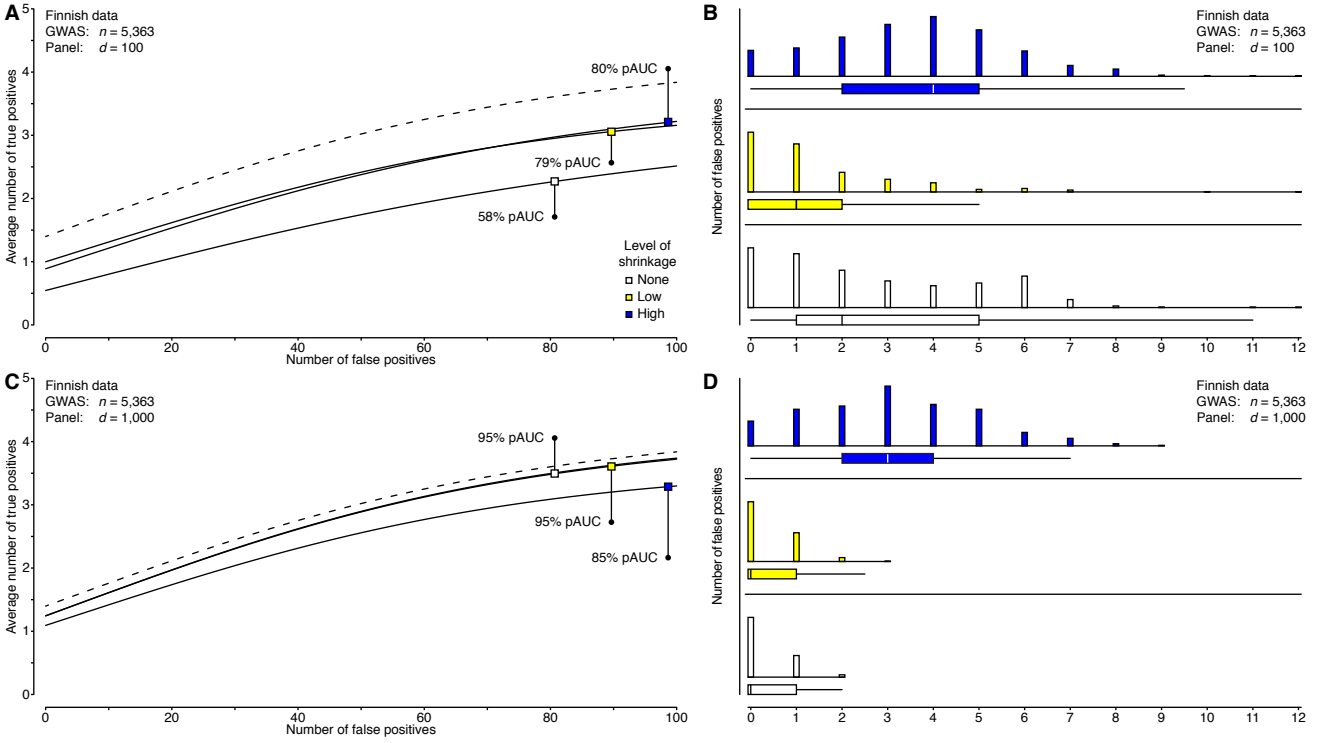

Figure S3 | Shrinkage estimation of reference panel correlations in simulations with Finnish data. Genotype data on 5,363 individuals from NFBC1966 were used to generate phenotype data affected by five causal SNPs. Results with different LD information are shown by plotting the number of selected causal SNPs (true positives) against the number of selected non-causal SNPs (false positives) using the list of SNPs ranked by their posterior probability of being causal (left panels) and by highlighting the number of false positives (non-causal SNPs with posterior probability of being causal larger than 0.5) over the simulated data sets (right panels). Correlation estimates from reference panels (—) using constant shrinkage factors of 1.00 (□), 0.80 (■) or 0.25 (■) are compared to the original genotype data (···) with respect to the achieved partial Area Under the Curve (pAUC) measure. Partial AUCs and curves are averaged over the simulated data sets. (**A-B**) Accuracy with NFBC1966 summary statistics and LD information either from the original genotype data or from a subset of FINRISK reference genotype data comprising 100 individuals. (**C-D**) Accuracy with NFBC1966 summary statistics and LD information either from the original genotype data or from a subset of FINRISK reference genotype data comprising 1,000 individuals.

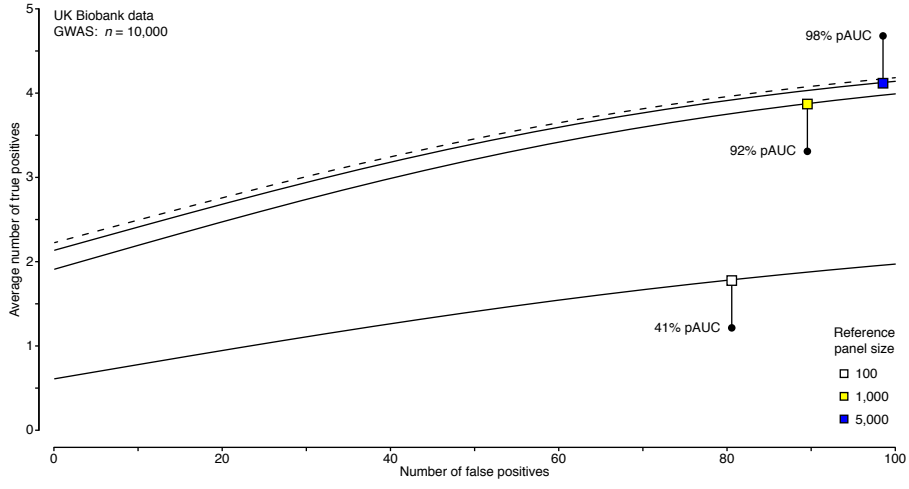

Figure S4 | Fine-mapping accuracy in UK Biobank simulations. Genotype data on 82,199 individuals covering 9q34/*ABO* region were used to generate phenotype data affected by five causal SNPs. Results with different LD information are shown by plotting the number of selected causal SNPs (true positives) against the number of selected non-causal SNPs (false positives) using the list of SNPs ranked by their posterior probability of being causal. Reference genotype panels (—) are compared to the original genotype data (---) with respect to the achieved partial Area Under the Curve (pAUC) measure. Results are shown for summary statistics from GWAS on 10,000 individuals with LD information either from the original genotype data or from a subset of UK Biobank individuals not included in the GWAS. Partial AUCs and curves are averaged over the simulated data sets.

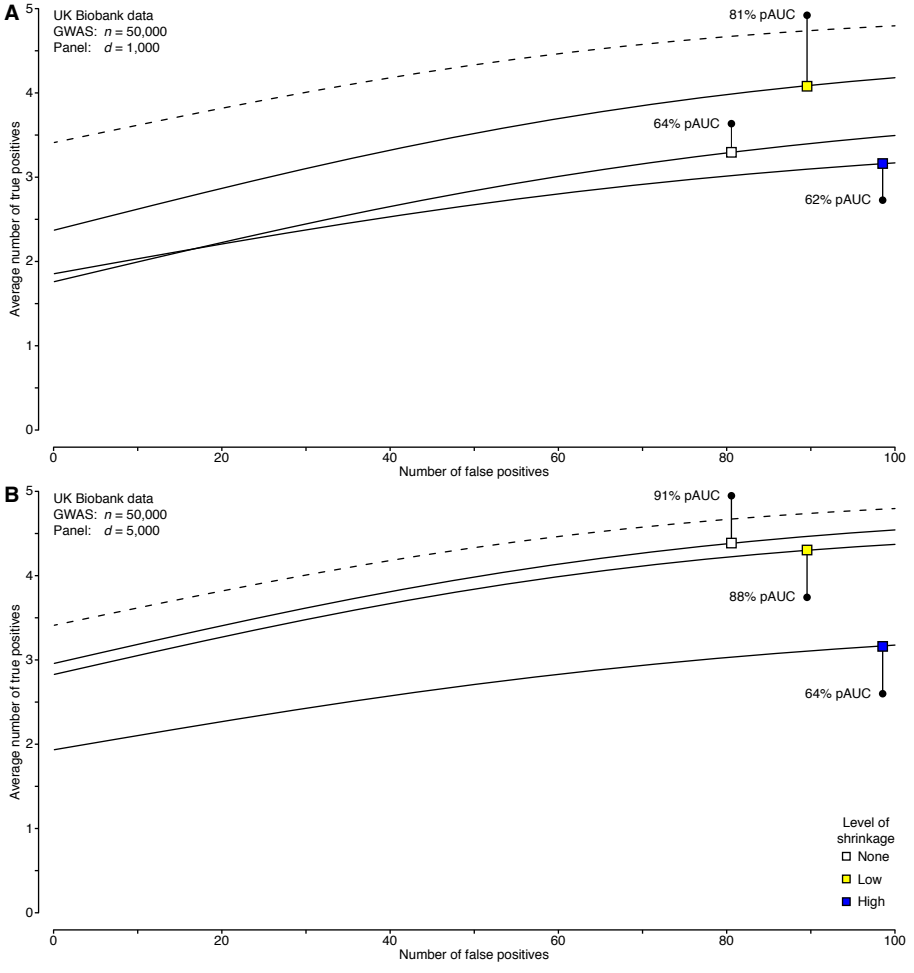

Figure S5 | Shrinkage estimation of reference panel correlations in UK Biobank simulations. Genotype data on 82,199 individuals covering 9q34/*ABO* region were used to generate phenotype data affected by five causal SNPs. GWAS summary statistics were computed using 50,000 individuals. Results with different LD information are shown by plotting the number of selected causal SNPs (true positives) against the number of selected non-causal SNPs (false positives) using the list of SNPs ranked by their posterior probability of being causal. Correlation estimates from reference genotype panels (—) using constant shrinkage factors of 1.00 (□), 0.80 (■) or 0.25 (■) are compared to the original genotype data (---) with respect to the achieved partial Area Under the Curve (pAUC) measure. Partial AUCs and curves are averaged over the simulated data sets. (**A**) Accuracy with LD information either from the original GWAS data or from a subset of UK Biobank genotype data not included in the GWAS comprising 1,000 individuals. (**B**) Accuracy with LD information either from the original GWAS data or from a subset of UK Biobank genotype data not included in the GWAS comprising 5,000 individuals.

Table S1. Average size of credible sets from FINEMAP analysis of 9q34/*ABO* region with 5 causal SNPs among 1,000 variants using two sources for LD information

GWAS sample size	Credible set	LD information from reference panels with size *d*			LD information from the original GWAS data
		d = 100	*d* = 1000	*d* = 5000	
n = 5363	90%	324	430	422	416
	95%	383	498	489	481
	99%	456	559	553	546
n = 50000	90%	14	174	273	237
	95%	14	248	465	467
	99%	16	340	605	628

Table S2. Average causal SNP coverage of credible sets from FINEMAP analysis of 9q34/*ABO* region with 5 causal SNPs among 1,000 variants using two sources for LD information

GWAS sample size	Credible set	LD information from reference panels with size *d*			LD information from the original GWAS data
		d = 100	*d* = 1000	*d* = 5000	
n = 5363	90%	68%	93%	93%	93%
	95%	73%	95%	95%	95%
	99%	79%	97%	97%	97%
n = 50000	90%	15%	70%	94%	96%
	95%	16%	74%	96%	98%
	99%	17%	80%	98%	99%