Supporting Information

Johansson et al. 10.1073/pnas.1217238110

SI Methods

Normalization of Peptide Values. To account for the variation in peptide intensities between plates and samples, the following transformations were performed. First, missing values for a peptide were imputed for each individual based on the minimum value for each peptide among all individuals. Second, peptide intensities for each individual were transformed to normality using rank-based inverse normal transformation, as implemented in the R library GenABEL (1). Third, the intensities for each peptide were transformed using the same inverse normal transformation for each plate separately. Peptides that were detected in fewer than 20 of the 96 samples on an analysis plate were classified as missing for all samples on that plate. Inverse normal transformation results in a normal distribution with mean 0 and SD 1. This transformation assumes that all individuals have the same total amount of peptides in the plasma and that all peptides have the same mean value on all plates. In the final dataset, imputed peptide values have the lowest rank for each individual.

Genetic Data and Analyses. Analysis of the raw data was done in BeadStudio software with the recommended parameters for the Infinium assay and using the genotype cluster files provided by Illumina. Quality control and imputation were performed on the two datasets separately using GenABEL (1) and the following cutoffs: genotyping call rate >95% (SNP) and >90% (individual), minor allele frequency (MAF) >0.01, and deviation from Hardy–Weinberg equilibrium ($P < 3.4 \times 10^{-8}$ or 1.4×10^{-8} Bonferroni corrected, per substudy). Pairwise kinship matrixes were calculated for each cohort, using the genotyped SNPs (180,212 genotyped SNPs that overlap between the microarrays for the two cohorts together) and the ibs function implemented in GenABEL, which computes a matrix of average "identical by state" (IBS) for a group of people. The kinship matrix is used to adjust for pedigree structure and to perform multidimensional scaling (MDS). Genetic outliers were investigated by 3D MDS using a false discovery rate (FDR) cutoff of 0.005. MDS plots for the two cohorts separately and together are included in Fig. S4. Duplicates, monozygotic twins, and genetic outliers were removed from further analyses resulting in $n = 691$ and $n = 345$ individuals for each array, respectively. Additionally 4 individuals were removed because they overlapped between the two studies, resulting in a total of 1,032 individuals in the downstream analysis. The genome location data from the Illumina Human Hap 300v2 were lifted from HG18 to HG19 using the UCSC liftOver tool (2). The data were then imputed with a prephasing approach using IMPUTE (version 2.2.2) (3) using the 1000 Genomes Phase I integrated variant set (National Center for Biotechnology Information build b37, March 2012) (4) as reference panel. The 1000 Genome reference panel was accessed from the IMPUTE Web resource. We filtered the imputed SNPs using MAF >0.01, IMPUTE2's info metric >0.3, and dosage >0.9, in KA06 and KA09, separately. Using these quality settings resulted in a total of 7.83 M and 8.78 M SNPs in KA06 and KA09, respectively. Of the 1,032 individuals with genotype data available, 1,029 were also included in the protein quantification.

Estimation of the Number of Unique Peptides. Starting from the protein sequences in the last release (September 2011) of the International Protein Index database (5), we counted the number of exact matches in all protein amino acid sequences ($n = 91,464$) for every quantified peptide in this study ($n = 984$). As a comparison, we also downloaded the complete list of peptides used in the Human Plasma Proteome Project (6) ($n = 20,433$) and performed the same analysis.

Estimating the Fraction of SNPs in Linkage Disequilibrium with a Nonsynonymous SNP. We downloaded the March 2012 version of the 1000 Genomes haplotypes (7) as available from the IMPUTE2's Website (http://mathgen.stats.ox.ac.uk/impute/impute_v2.html#reference). We reformatted the files into a format suitable for PLINK (8) and calculated the linkage disequilibrium (LD) patterns for each chromosome separately using the "–r2" options in windows of 250 kb. The total number of markers used for this was 30,072,738 genome-wide and 93,583 over the genes with peptide measurements. We then used ANNOVAR (9) to annotate any markers that are in LD ($r^2 > 0.8$) with any other SNP and calculated the number of nonsynonymous SNP (nsSNP) annotations to estimate the fraction of SNPs in the genome that is in LD with a nsSNP. The resulting fractions genome-wide and over the investigated genes were found to be 0.0022% and 0.27%, respectively.

1. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM (2007) GenABEL: An R library for genome-wide association analysis. *Bioinformatics* 23(10):1294–1296.
2. Hinrichs AS, et al. (2006) The UCSC Genome Browser Database: Update 2006. *Nucleic Acids Res* 34(Database issue):D590–D598.
3. Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 5(6): e1000529.
4. The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* 467(7319):1061–1073.
5. Kersey PJ, et al. (2004) The International Protein Index: An integrated database for proteomics experiments. *Proteomics* 4(7):1985–1988.
6. Farrah T, et al. (2011) A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Mol Cell Proteomics* 10(9): M110.006353.
7. Abecasis GR, et al.; 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* 467(7319):1061–1073.
8. Purcell S, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3):559–575.
9. Wang K, Li M, Hakonarson H (2010) ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38(16):e164.

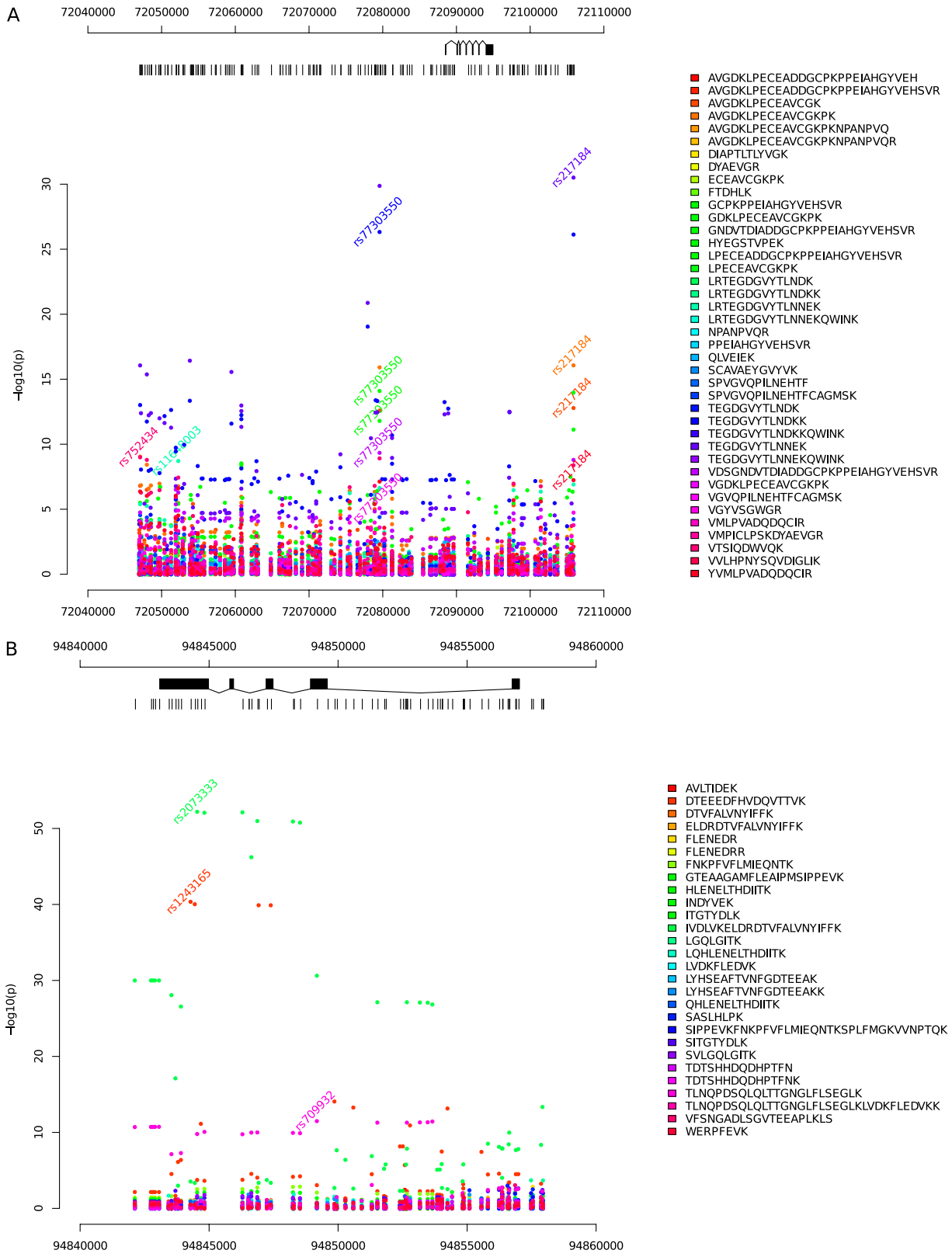

Fig. S1. Manhattan plot for *P* values across the genes for the (*A*) Haptoglobin protein (HPT) and (*B*) Alpha-1-antitrypsin (A1AT) proteins. The dots represent *P* value distributions for association with each of the different peptides listed on the *Right*. Bars indicate the location of SNPs included in the genome-wide association studies (GWAS) and the black boxes, the gene structure. For each peptide, the name of the strongest associated SNP is written out next to the dot.

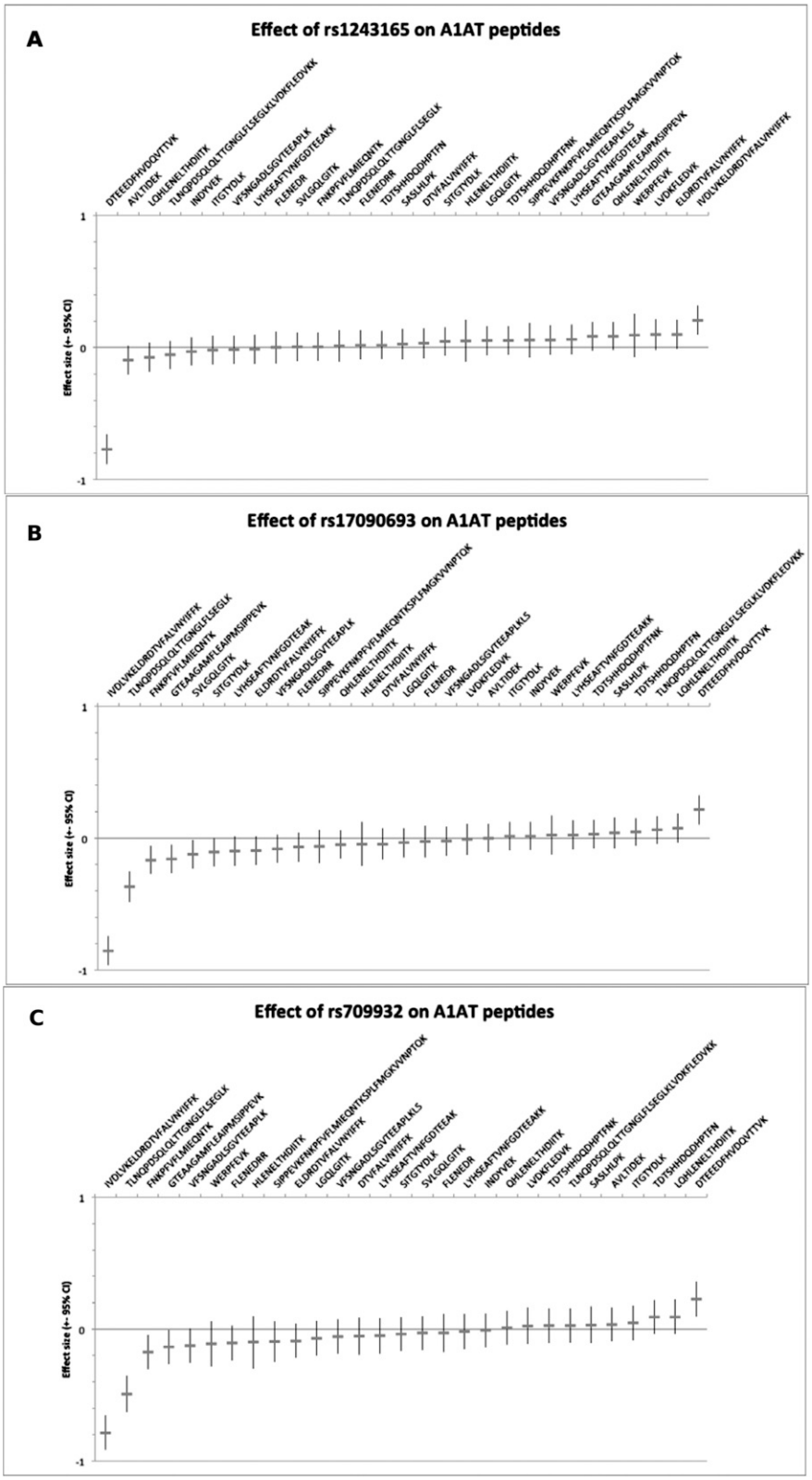

Fig. S2. Effect of SNPs near the serpin peptidase inhibitor, clade A, member 1 (*SERPINA1*) gene on the abundance level of peptides from the A1AT protein. (*A*) rs1243165, (*B*) rs17090693, and (*C*) rs709932. Each peptide is represented by the effect (beta) and the 95% confidence interval (CI) of the effect. All effects are reported for the minor allele in each marker. Observations where the CI does not include zero represent nominally significant observations ($P < 0.05$).

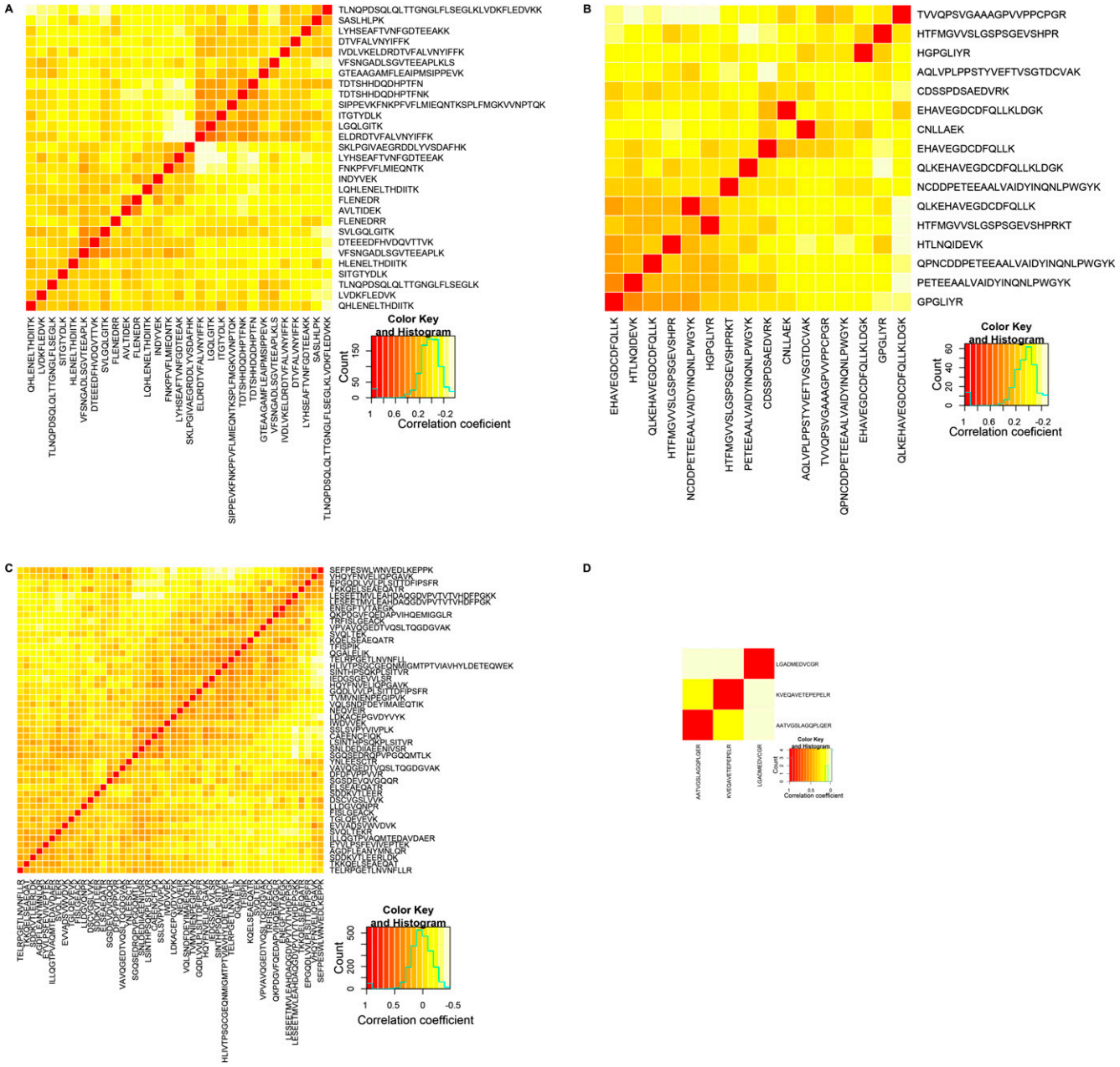

Fig. S3. (*A*) Heatmap of pairwise correlation coefficients between A1AT peptides. Each point in the heatmap represents the correlation coefficient between two peptides ranging from $R = -0.3$ (white) to $R = 1$ (red). The color histogram (*Lower Right*) shows the distribution of correlation coefficients. (*B*) Heatmap of pairwise correlation coefficients between alpha-2-HS-glycoprotein (*FETUA*) peptides. Each point in the heatmap represents the correlation coefficient between two peptides ranging from $R = -0.3$ (white) to $R = 1$ (red). The color histogram (*Lower Right*) shows the distribution of correlation coefficients. (*C*) Heatmap of pairwise correlation coefficients between complement C3 (CO3) peptides. Each point in the heatmap represents the correlation coefficient between two peptides ranging from R = −0.5 (white) to $R = 1$ (red). The color histogram (*Lower Right*) shows the distribution of correlation coefficients. (*D*) Heatmap of pairwise correlation coefficients between apolipoprotein E (*APOE*) peptides. Each point in the heatmap represents the correlation coefficient between two peptides ranging from $R = 0$ (white) to $R = 1$ (red). The color histogram (*Lower Right*) shows the distribution of correlation coefficients.

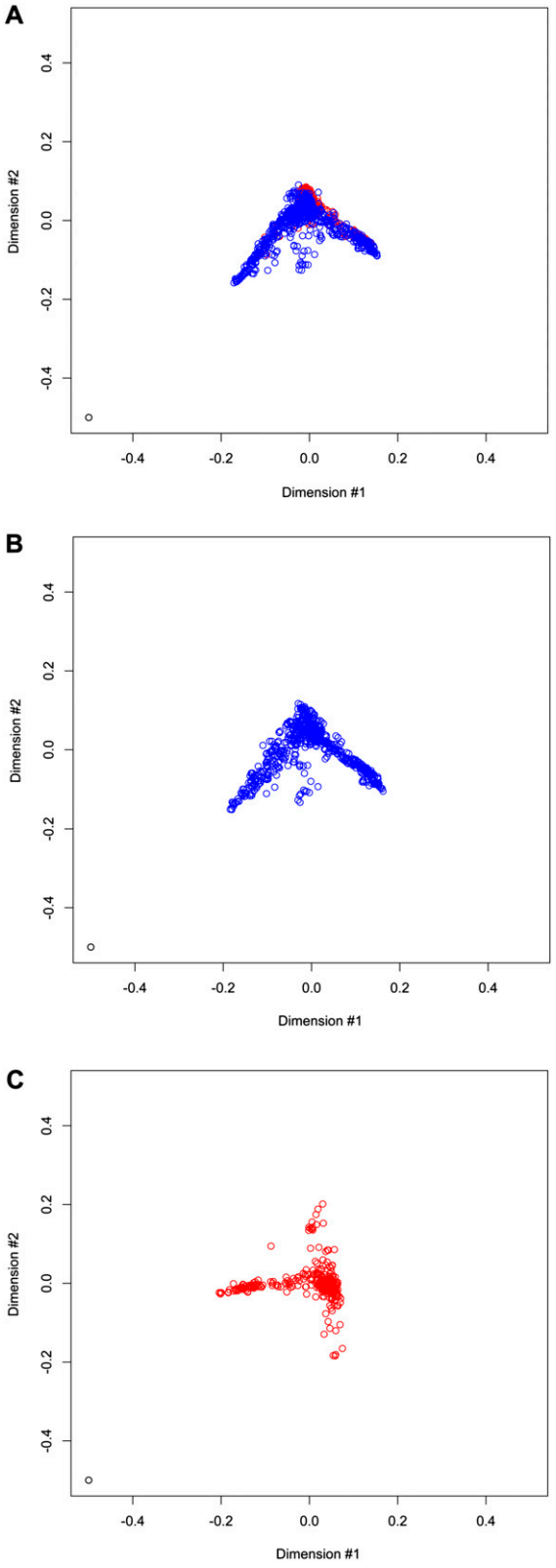

Fig. S4. (*A*) Multidimensional scaling (MDS) plot of the discovery (blue) and replication (red) cohort together. (*B*) MDS plot of the discovery cohort. (*C*) MDS plot of the replication cohort.

Other Supporting Information Files

Table S1 (DOC)
Table S2 (DOC)
Table S3 (DOC)