Supplementary Information Guide:

1. Supplementary Methods

2. Supplementary Figures 1-11

   2.1. Supplementary Fig. 1: Protein variation between individuals and populations.

   2.2. Supplementary Fig. 2: Protein phosphorylation level in different protein variation quantiles.

   2.3. Supplementary Fig. 3: Go ontology enrichment analysis of protein groups with different variation levels.

   2.4. Supplementary Fig. 4: Protein level differences between sexes.

   2.5. Supplementary Fig. 5: Correspondence between protein-protein covariation and RNA-RNA covariation.

   2.6. Supplementary Fig. 6: Distribution of protein heritability in CEU and YRI.

   2.7. Supplementary Fig. 7: Loci associated with protein expression levels in CEU and YRI.

   2.8. Supplementary Fig. 8: Comparison of pQTLs and eQTLs.

   2.9. Supplementary Fig. 9: OAS1 peptide expression levels against rs10774671 genotypes.

   2.10. Supplementary Fig. 10: Validation of IMPA1 protein level in the YRI population.

   2.11. Supplementary Fig. 11: Correlation between protein and RNA for IMPA1.

3. Supplementary Tables 1-8

   3.1. Supplementary Table 1: Log2 protein ratio for 5953 proteins in 95 cell lines (MS Excel spreadsheet, 6.8 MB).

3.2. Supplementary Table 2: Coefficients of variation for 4053 proteins (MS Excel spreadsheet, 1.0 MB).

3.3. Supplementary Table 3: Proteins different between males and females at FDR 10% (MS Excel spreadsheet, 27 KB).

3.4. Supplementary Table 4: Proteins different between CEU and YRI at FDR 10% (MS Excel spreadsheet, 80 KB).

3.5. Supplementary Table 5: Sparse protein covariation network (MS Excel spreadsheet, 56 KB).

3.6. Supplementary Table 6: Dense protein covariation network (MS Excel spreadsheet, 163 KB).

3.7. Supplementary Table 7: Protein level heritability (MS Excel spreadsheet, 315 KB).

3.8. Supplementary Table 8: pQTLs in CEU, YRI and 3 populations combined at FDR 10% (MS Excel spreadsheet, 103 KB).

1. Supplementary Methods

Cell culture, sample preparation and labeling with TMT

Lymphoblastoid cell lines (LCLs) from 95 HapMap individuals were obtained from the Coriell Institute for Medical Research. The samples include 53 Caucasians of northern and western European ancestry (CEU), consisting of 10 trios, one duo, and 21 unrelated individuals; 33 Yorubans from Ibadan, Nigerian (YRI), consisting of 9 trios, one duo, and four unrelated individuals; 9 eastern Asians (ASN), consisting of eight unrelated Han Chinese from Beijing (CHB) and one Japanese from Tokyo (JPT). Anti-IMPA1 antibody was from Sigma-Aldrich (Product No: SAB2103588). The LCLs were grown to a density of 0.6-1.4 x 10^6/mL in RPMI medium supplemented with 15% fetal bovine serum, L-glutamine and 1 x antibiotic-antimycotic solution at 37°C, 5% CO_2. Biological replicates were grown separately. Cell pellets were lysed in 10x volume of lysis buffer containing 4% SDS and 100 mM dithiotreitol (DTT) in 100 mM Tris-HCl pH 8.0. Lysates were incubated at 95°C for 5 min. Insoluble remains were removed by centrifugation at 16,000xg for 15 min at room temperature.

Protein lysates were subjected to detergent cleanup, cysteine alkylation, trypsin digestion and TMT labeling as described previously with modifications[19]. In brief, 30 µl protein lysates were mixed with 200 µl urea buffer (8M Urea, 0.1 M Tris-HCl pH 8.0) in YM-30 microcon filter unit (Millipore), and then concentrated at 12,000xg for 15 min. This step was repeated twice to remove SDS. The protein lysates were incubated in 100 µl urea buffer supplied with 0.05 M iodoacetamide in darkness for 1 hour. Protein concentration was measured using a bicinchoninic acid (BCA) assay (Thermo Scientific). Then protein lysates were cleaned up by urea buffer twice followed by 20 mM triethylammonium bicarbonate buffer (TEAB) at least three times to get rid of all the remaining urea and Tris-HCl. The protein lysates were mixed in 20 mM TEAB with trypsin at an enzyme:protein ratio 1:50, and incubated at 37°C overnight. The filter was transferred to a new tube and spun at 14,000xg for 10 min. Digested peptides were labeled with sixplex TMT reagents (Thermo Scientific) as recommended by the manufacturer. Finally, six TMT-labeled samples (TMT126-TMT131) were equally mixed to generate the final digest mixture. TMT126 was used in each experiment to label the peptides from the reference cell line GM12878.

Two dimensional liquid chromatography tandem mass spectrometry

All digest mixtures were analyzed on an LTQ Orbitrap Velos (Thermo Scientific) equipped with an online 2D nanoACQUITY UPLC System (Waters) as previously described with the following modifications[19]. Peptides were separated by using a dual reversed-phase (RP) approach. In the first dimension, peptides were separated on an Xbridge 300 μm x 5 cm C18 5.0 μm column (Waters) using 11 discontinuous step gradient at 2 μl/min. Acetonitrile concentration for each step was adjusted to ensure nearly equivalent peptide load and MS intensity for each second-dimension run. In the second dimension, peptides were separated on an in-house packed 75 μm ID/15 μm tip ID x 20cm C18-AQ 3.0 μm resin column by applying a 8-30% acetonitrile gradient in 0.1% formic acid over 120 min at 300 nl/min. A total of 51 digest mixtures were analyzed in this study.

During data acquisition by LTQ Orbitrap, the full MS scan was performed in the orbitrap in the range of 400-1800 m/z at a resolution of 60000, followed by the selection of the 10 most intense ions for HCD-MS2 fragmentation using a precursor isolation width window of 1.5 m/z. The normalized collision energy for HCD was set to 38% at 0.1 ms activation time. The signal for MS2 requires a minimum of 5000 counts. Ions with singly charged state or unassigned charge states were rejected for MS2. Ions within 10 ppm m/z window around ions selected for MS2 were excluded from further selection for fragmentation for 60 s.

Mass Spectrometry data processing and filtering

The acquired raw data from each digest mixture were searched against a human International Protein Index (IPI) database, version 3.75[20], concatenated with a decoy database with all the protein sequences in reverse order, using the SEQUEST algorithm[21] (Proteome Discoverer software, version 1.2, Thermo Scientific). Searches were performed using a 10 ppm mass tolerance for precursor ions and 0.02 Da for fragment ions, allowing up to two trypsin missed cleavages. Sixplex TMT tags on lysine residues and peptide N termini (+ 229.163 Da) and oxidation of methionine residues (+ 15.995 Da) were set as a variable modification; carbamidomethylation of cysteine residues (+57.021 Da) were set as static modifications. Peptides with minimum seven amino acid lengths and rank 1st were considered for protein identification. We first grouped the redundant proteins and filtered low confidence identifications based on the IPI database: proteins identified with the same sequence were grouped together to eliminate redundancy in the protein list. Proteins matched to decoy sequences were considered as false discovery. Peptides were further filtered based on SEQUEST parameter XCorr vs. charge state to achieve final protein identification FDR less than 1% for each mixed sample. For quantification, peptides with all the quantification channels present were used. Intensity of reporter ions in each channel was integrated by most confident centroid method with 10 ppm window tolerance. Peptide ratios (based on reporter ion intensity/ reference cell line reporter ion intensity) were median normalized across all the quantified proteins in each sample to eliminate sample mixing bias. Outlier peptides which were above 100 fold change were removed. The filtered high confidence peptides in each mixed sample were then exported from the Proteome Discoverer software for the next step analysis. Combining all the 51 2D LC-MS/MS experiments, 2,726,242 high confidence peptide spectra were analyzed, corresponding to 71,800 unique peptide sequences.

Calculation of the protein levels

The correspondence between peptide sequences, proteins, genes (Ensembl gene IDs) and genomic coordinates was established based on the protein and gene cross-reference tables of IPI database version 3.87 and the transcript sequences of Ensembl database release 62. Among the 71,800 unique peptide sequences, 67,075 were mapped to unique Ensembl gene ID and selected for further analyses.

We next sought to identify the peptide sequences overlapping with known protein coding variants predicted to alter the protein sequence and polymorphic in our samples. Such peptides may lead to false positive pQTL discoveries by our method (in a matter analogous in RNA expression studies to the issue of SNPs overlapping with microarray gene expression probes). The genotypes of 94 out of 95 of our cell lines were obtained from HapMap release 28[9,22](combining phase I II and III, there are up to ~4 million SNPs genotyped). Furthermore, 62 of our 95 cell lines had whole genome resequencing information available from the 1000 Genomes Project phase 1v3 (56 cell lines) and pilot 2 trios (6 cell lines)[23]. Variant consequence annotations were obtained from the Ensembl database release 62 and the 1000 Genomes Project. Peptides overlapping with a nonsynonymous SNP or mapping 3’ of exon-coding-change SNPs (e.g. a stop codon gained or lost, or an essential splice site mutation in the first two or last two base pairs of an intron) which were polymorphic in our samples were removed from the quantification, leaving 60740 unique peptide sequences, mapping to 5953 genes, and quantified in 2,159,989 MS2 spectra. Based on this final peptide list, for each sample and each gene, we re-quantified proteins based on the median ratio of all peptides mapping to the same gene. The log2-transformed protein ratio was considered as the relative protein level. Note that by this approach, we obtained a single protein level quantification per gene and excluded many of the HLA proteins.

Our study included 2 to 5 biological replicates for each cell line (biological replicates mean the cell line was independently cultured, prepared and analysed in different LC-MS/MS runs). The measure of protein level for each individual used throughout the manuscript was the average log2 protein ratio of biological replicates.

Protein variation

For most of the analyses below, we focused on the 4053 proteins measured in at least half of the unrelated individuals. We have a total of 74 unrelated individuals in our study: 42 CEU (20 parents of trios, 1 parent of a duo, and 21 other unrelated), 23 YRI (18 parents of trios, 1 parent of a duo, and 4 other unrelated), and 9 ASN. Protein level data from the 19 children in trios were used only for heritability estimations. We noted that these 4053 proteins, consistently detected by mass spectrometry, are likely the more abundant proteins in the LCLs.

Protein variation was quantified by calculating for each protein the coefficient of variation (CV) of the protein levels on the ratio scale within each population and in all populations combined after adjusting the protein levels for population averages. GO ontology categories enrichment analyses were performed in PANTHER[25] using the Bonferroni correction for multiple testing and the list of 4053 proteins as the background set.

Since differential posttranslational modifications could affect protein expression measured by our method, we examined the potential contribution of posttranslational modifications to protein variation in our dataset. We focused on peptide phosphorylation and used two different approaches to estimate the potential phosphorylation level of each protein. In the first approach, we calculated the STY amino acid (serine, threonine, and tyrosine) percentage among the sum of all used peptide sequences for each protein. In the second approach, we computed the fraction of potential phosphopeptides used for quantifying each protein, treating all peptides containing a phosphorylable site in the Phospho.EML database as a potential phosphopeptide. We observed that neither the STY percentage nor the phosphopeptide percentage shows enrichment in the sets of highly variable proteins (Supplementary Fig. 2). Although indirect, this observation indicates that the contribution of phosphorylation to our measurements of protein expression variation is likely to be small.

Differences between sexes and populations

To identify proteins differentially expressed between males and females, we used a linear model and regressed protein levels on sex, adjusting for population differences by using the population label as a covariate. To identify proteins differentially expressed between CEU and YRI LCLs, we regressed protein levels on population label. The FDR for both analyses was calculated using the QVALUE Bioconductor package[24].

Protein covariation

For the protein covariation network analysis, we selected the 2279 proteins which were quantified in all 74 unrelated individuals. To adjust for population stratification, a linear regression of protein level on population label was performed and the residuals were normalized by transforming the quantiles of the residual values to their respective quantiles of a N(0,1) distribution. We constructed the protein coexpression networks based on the Gaussian graphical model and the statistical approaches implemented in the R package sparse partial correlation estimation (SPACE)[13]; we also analyzed the data using the method of Meinshausen and Bühlmann[26]. Since both methods gave similar results, we present the results from SPACE only. Both algorithms required choosing a tuning parameter, which controls the stringency with which two proteins are considered as co-varying. To approximate the FDR of the network analysis, we augmented the protein data with permutations of 944 protein expressions. We treated an edge connecting two permuted proteins or an edge connecting a permuted and an observed protein as known false edge; thus the network analysis included 2,591,752 a priori known false edges. In applying SPACE, no known false edges were detected until ~ 1,000 edges were in the inferred, thus we used this criterion to construct the dense network (Supplementary Table 6). We further assessed the stability of both sparse and dense network using a stability concept similar to the method of Meinshausen and Bühlmann[27]. Specifically, we repeatedly sampled half of the individuals (n=37), applied SPACE to these smaller datasets, and compared the edges detected in the full datasets with those detected in sub-samples. The idea of stability is that high-confident edges are likely to be stably detected using a subset of the sample. Because the size of the sub-samples was small, the statistical power to detect a true edge was substantially reduced. To alleviate the problem, we chose the tuning parameters to obtain smaller networks with an average of 520 edges to be compared with the full-data sparse network (Figure 2, Supplementary

Table 5), and bigger networks with an average of 2105 edges to be compared with the full-data dense network. Over 100 sub-samples, we found that an average of 134 of the 223 edges (60%) detected in the sparse network were selected in each sub-sample network. As a negative control, we considered the 448 edges that were selected in a sub-sample but not in the full dataset as low-confident edges; over the 100 sub-samples, an average of 44 edges (~10%) re-occurred in the sub-sample networks. Similarly, we found an average of 482 of the 1012 edges (48%) in the full-data dense network to be represented in each sub-sample network, compared to an average of 143 of the 1813 low-confident edges (8%) that were selected in a sub-sample network. For the comparison with RNA expression, we used data generated in the two RNA-Seq experiments: the CEU gene-level expression[2] and the YRI normalized gene-level expression[3], quantile-normalized to fit a N(0,1) distribution. These included expression of 2147 genes in CEU and 2156 genes in YRI in common with the protein dataset.

For known protein interaction dataset in the literature, we considered protein-protein interactions from three public databases, i.e. BioGRID, HPRD and IntAct[28-30]. To estimate the degree of known protein interaction enrichment in the sparse protein covariation network, we selected 278 proteins randomly from 2279 proteins and then selected 223 edges randomly from the possible 38,503 edges (278 x 277/2 =38,503). In this way, the generated random network was matched with the inferred sparse network. Then we counted the number of known protein interactions in the simulated network, and replicated 20,000 times. The true sparse protein covariation network had 29 known protein-protein interactions, while the maximum number of interacting pairs found in a simulated network was 10. Therefore the protein covariation network is highly enriched for known interacting protein pairs ($P < 5 \times 10^{-6}$).

Heritability estimations

The heritability of protein levels was calculated as the slope of the regression line of the children protein level on the mid-parent protein level (log2 ratios) based on the 10 CEU trios and 9 YRI trios. We focused on the proteins that are detected in all the trios, 2395 and 2534 proteins in CEU and YRI, respectively, with an overlap of 2292 proteins. The slope of the regression line is an estimate of the narrow sense heritability, the proportion of total phenotypic variation due to the additive effects of genes. The heritability estimates are a statistical estimation, and are subject to sampling errors. However, aggregated across all proteins, the "average" protein level heritability is significantly greater than 0 in both CEU (P = 2 x 10^{-9}) and YRI (P = 3 x 10^{-62}), two-tailed t-test. In CEU, the average heritability is 0.06 (95% CI = 0.04-0.09). In YRI, the average heritability is 0.17 (95% CI = 0.15-0.19).

Cis-pQTL mapping

We initially searched for cis-pQTLs +/- 200 kb of the gene region and found that the majority of cis-pQTLs lie within 20 kb of the gene region, similarly to what has been found for cis-eQTLs. Therefore, for the analysis presented in the main text, we tested genetic variants within 20 kb of the gene region for cis-pQTL effect, thereby limiting the multiple testing burdens and increasing our power to detect pQTLs.

For each of the 4053 proteins with at least 50% data, we tested the association between protein levels and the genotypes of SNPs located in the corresponding gene region +/- 20 kb and with MAF > 10%. SNPs genotypes were obtained from HapMap III release 3[9], and were available for 72 out of the 74 unrelated individuals. Throughout the manuscript, we report genomic coordinates relative to the NCBI36 human genome assembly. Where needed, genomic coordinates were converted between the NCBI36 (or hg18) and the GRCh37 (or hg19) human genome assemblies using the UCSC browser liftOver utility[31].

The tests for genetic association were performed in R (http://www.r-project.org/) or PLINK[32], using a linear model where protein level (log2 ratio) was regressed on SNP genotype assuming an additive genetic model. In Supplementary Tables S8, we report the regression coefficient beta for the genetic effect, interpretable as the mean increase in protein level (log2 protein ratio) per copy of the SNP minor allele. In the analysis of the three populations combined, population structure was adjusted for by introducing two covariates in the linear model, coding for the CEU and YRI population label. For association testing with X chromosome SNPs, sex was used as an additional covariate in the model and the SNP genotype was coded as 0, 1, or 2 minor alleles in females and 0 or 1 minor allele in males. In total 116556, 121405, and 130505 tests were performed in CEU, YRI, and three populations combined analyses, respectively. At the protein level, we corrected the nominal *P* values of each SNP for multiple testing using a permutation procedure that accounts for the number of SNPs tested for that protein and local linkage disequilibrium. The corrected *P* values for multiple testing at the protein level were calculated using the max(T) permutation procedure. Briefly, for each protein, protein levels were permuted between cell lines, and for each permutation the minimum *P* value over all SNPs was recorded. A corrected *P* value was calculated for each SNP as the count of *P* values identified in the permutations that were smaller than the original *P* value for that SNP divided by the number of permutations. In the CEU and YRI analyses, corrected *P* values were obtained using the PLINK software implementation of the max(T) procedure and 10,000 permutations of the phenotype. For all three populations combined, we implemented the permutation procedure in R, permuting phenotypes within each population adaptively up to 10,000 times. The pQTL FDR was calculated based on the distribution of the minimum max(T) corrected *P* values (one per protein) using the QVALUE Bioconductor package[24].

For the 77 proteins with detected cis-pQTLs in the combined population at FDR 10%, the median heritability was 0.12 in the CEU trios, 0.34 in the YRI trios, and 0.31 for a combined analysis of the 19 CEU and YRI trios correcting for population averages. Therefore as expected proteins with detected cis-pQTLs tend to have greater heritability than on average.

Alternative cis-pQTL analyses subsetting on the proteins with high signal to noise ratios

We reasoned that our power to detect pQTLs might be greater for proteins with higher variation among individuals and high reproducibility among replicates. To explore this possibility, for each of the 4053 proteins, we calculated a signal to noise ratio (SNR) by comparing the variation among individuals (the signal) to the variation among replicate measurements (the noise). Specifically, the signal was calculated as the coefficient of variation of protein levels on the ratio scale among the 74 unrelated individuals after correcting for population differences and the noise was calculated as the median coefficient of variation of the protein level on the ratio scale over all biological replicate measurements. The SNR ranges from 0.66 to 8.97, with a median of 1.79. We found that the cis-pQTLs we identified are enriched for proteins that have higher SNR, with 79%, 69%, and 68% of the proteins with cis-pQTLs in the CEU, YRI, and combined population (FDR10%) having an SNR greater than the median of 1.79, respectively. This indicates that the cis-pQTLs we identified tend to be associated with proteins more reproducibly measured and/or more variable among individuals. To assess whether only considering proteins with high SNR would improve our power to detect pQTLs, we performed the cis-pQTL analysis subsetting on the 2027 proteins with SNR greater than the median of 1.79. At FDR 10%, the counts of cis-pQTLs identified are almost identical to that of the analysis based on all 4053 proteins (35, 13, and 73 pQTLs found in CEU, YRI, and three populations combined respectively). Therefore we presented the analysis based on all 4053 proteins in the main text.

Comparison of pQTLs and eQTLs

We evaluated whether the same genetic variants may be associated with both protein and RNA levels by comparing our proteomics results to that of two RNA-Seq studies of CEU and YRI LCLs[2,3]. First, starting with the pQTLs we identified, we obtained the *P* value for the association of the same SNPs with RNA levels. In CEU, we used the *P* values found in published result tables for the matched SNP and gene. In YRI, we recalculated the *P* value for the SNP/ RNA association using a linear model and gene-level normalized RNA-Seq data. In cases where the SNP was tested for association with multiple transcripts (in CEU) or in cases where multiple SNPs had equally most significant *P* values as pQTLs, we selected the minimum *P* value of all possible associations with RNA levels. Results from this analysis are presented in Supplementary Fig. 8.

Reciprocally, to determine the *P* values as pQTLs of SNPs identified as eQTLs, we used the CEU and YRI RNA-seq result files of significant eQTLs (FDR 10% in YRI and permutation corrected $P \leq 0.05$ in CEU), matching with our protein data on both SNP and corresponding gene. To make the YRI and CEU RNA studies more comparable, we only considered the most significant SNP within 200kb of the gene region in both RNA studies. We observed that eQTLs are enriched for significant associations with protein levels (not shown).

Finally, the *IMPA1* RNA levels were obtained from the same RNA-Seq studies that included 37 CEU LCLs and 22 YRI LCLs in common with our proteomics study[2,3]. The correlation between RNA and protein levels was plotted for the combined CEU and YRI LCLs by standardizing RNA and protein levels to have a mean of 0 and standard deviation of 1 within each population.

Supplementary References

25 Thomas, P. D. *et al.* PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res* **31**, 334-341 (2003).

26 Meinshausen, N. & Buhlmann, P. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34**, 1436-1462 (2006).

27 Meinshausen, N. & Bühlmann, P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**, 417-473 (2010).

28 Stark, C. *et al.* The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res* **39**, D698-704 (2011).

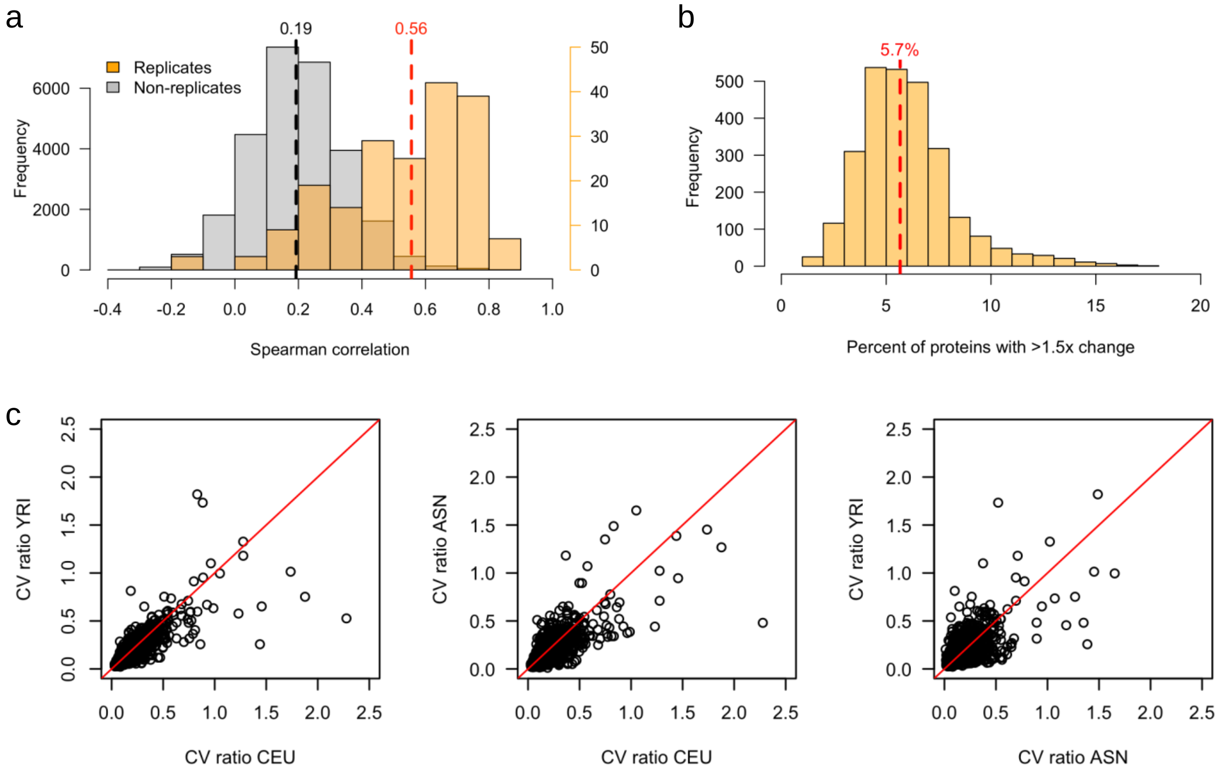
29 Keshava Prasad, T. S. *et al.* Human Protein Reference Database--2009 update. *Nucleic Acids Res* **37**, D767-772 (2009).

30 Kerrien, S. *et al.* The IntAct molecular interaction database in 2012. *Nucleic Acids Res* **40**, D841-846 (2012).

31 Fujita, P. A. *et al.* The UCSC Genome Browser database: update 2011. *Nucleic Acids Res* **39**, D876-882 (2011).

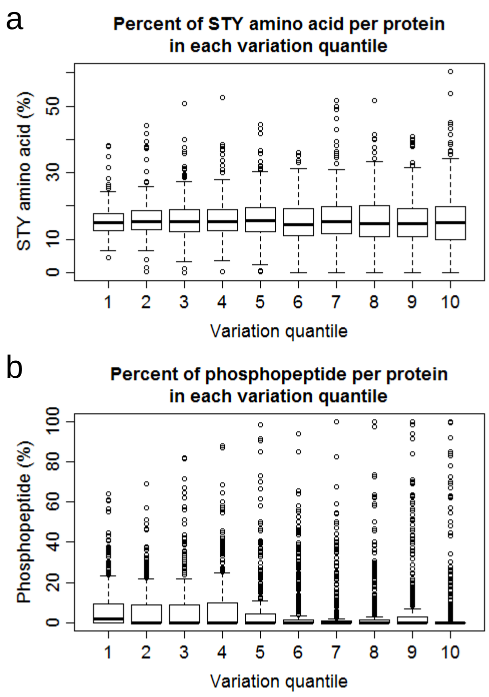
32 Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575 (2007).

2. Supplementary Figures



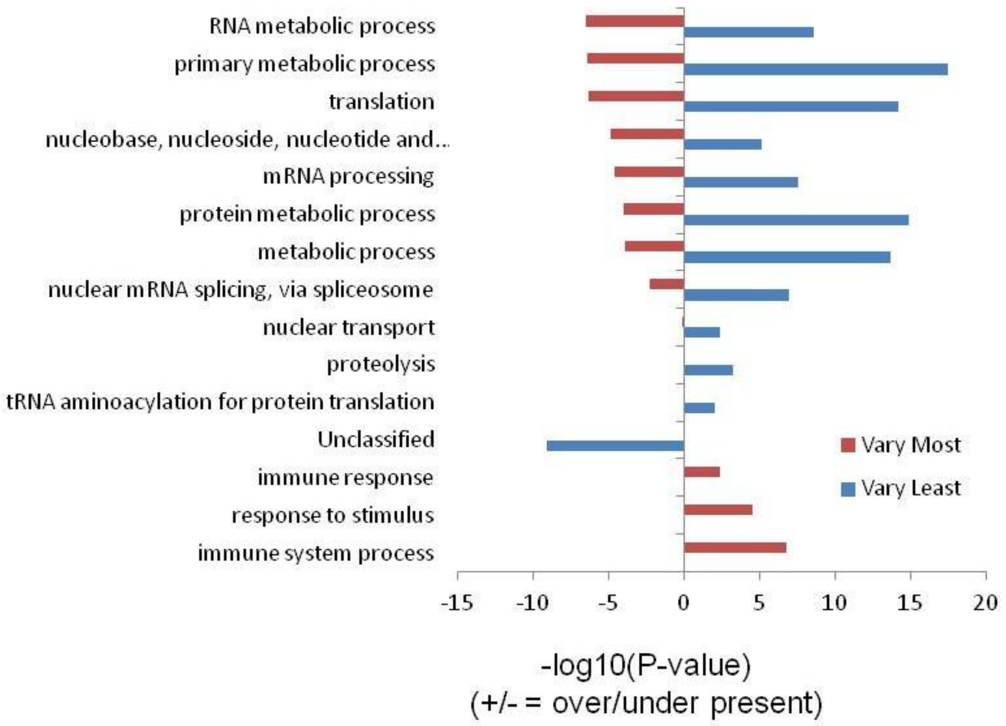
Supplementary Fig. 1 | Protein variation between individuals and populations.

a) Comparison of Spearman's rank correlation coefficient distributions between pairs of biological replicates and non-replicates. The dashed lines are at median values. Biological replicates (in orange) have higher correlation than non-replicates (in grey). **b**) Percentage of proteins that change between pairs of individuals. The percent of proteins that change more than 1.5 folds between pairs of individuals was calculated based on 4053 proteins in 74 unrelated individuals. A median of 5.7% of the proteome changed more than 1.5 folds between a pair of individuals. **c**) Proteome variation among populations. The coefficient of variation (CV) of each protein ratio was calculated in each population (CEU, YRI, and ASN). Variable proteins in one population tend to also be variable in the other two populations.



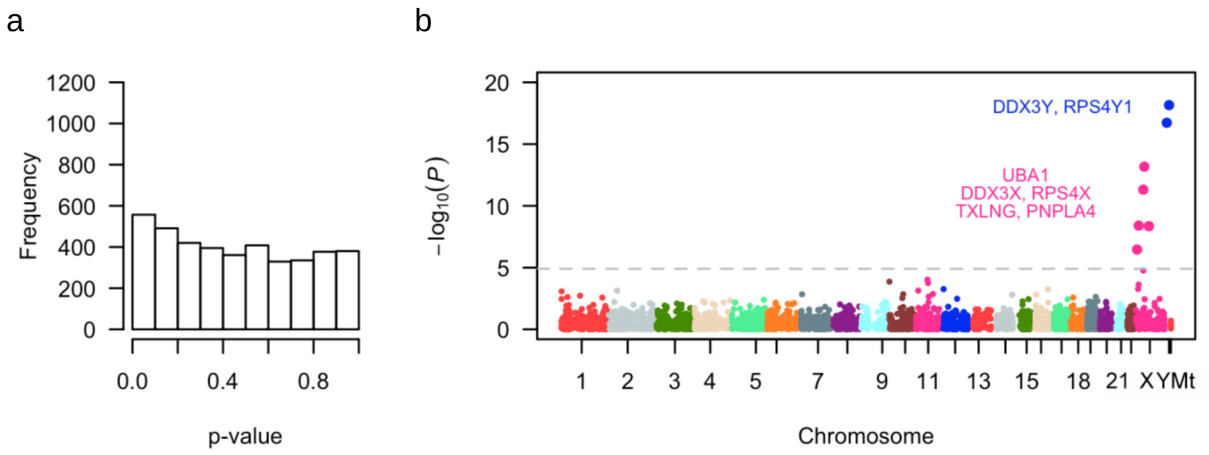
Supplementary Fig. 2 | Protein phosphorylation level in different protein variation quantiles.

a) Percent of STY amino acids per variation quantile. **b**) Percent of phosphopeptide per variation quantile. X axes, protein variation levels increase from left quantile to right quantile.



Supplementary Fig. 3 | Go ontology enrichment analysis of protein groups with different variation levels.

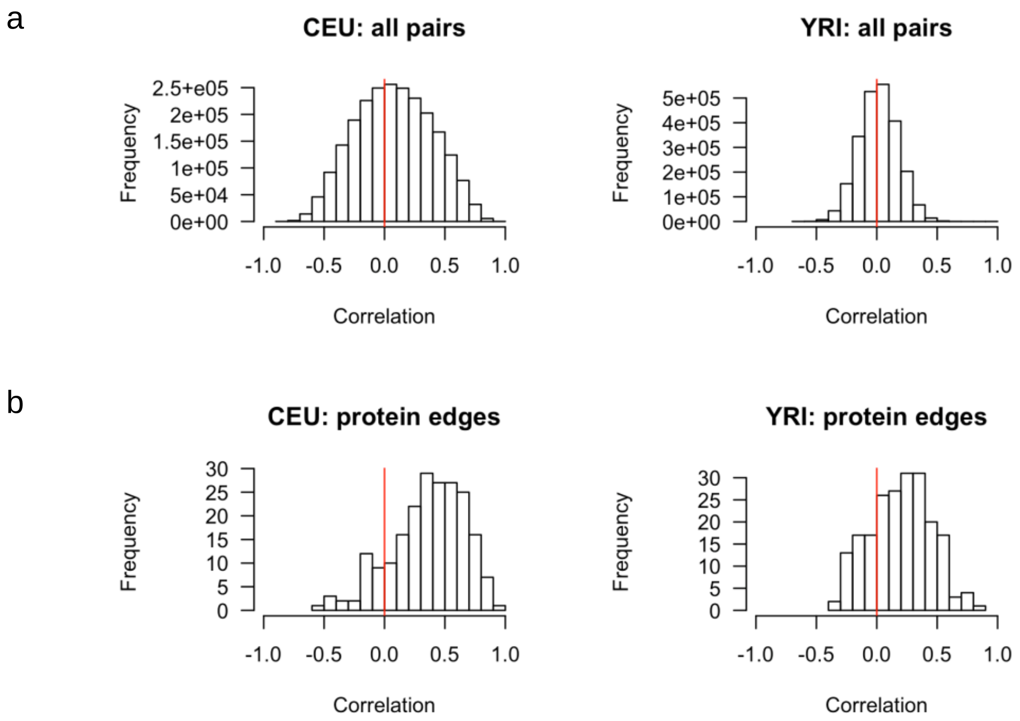
Proteins varying the most and the least (top 20%) based on the 4053 proteins in 74 unrelated individuals were subjected to GO enrichment analysis on biological processes. Plotted are the functional categories with enrichment *P* value < 0.01. Bonferroni correction was used for multiple testing. These two groups of proteins show very different GO enrichment patterns. Proteins that vary the most are enriched in immune response processes; proteins that vary the least are enriched in metabolic processes.



Supplementary Fig. 4 | Protein level differences between sexes.

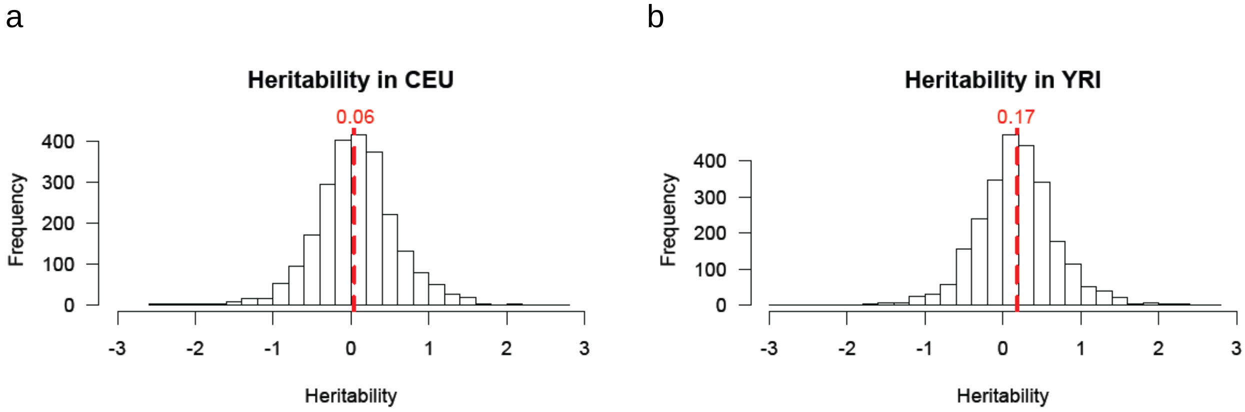
a) The *P* value distribution for the test of the difference in protein levels between males and females shows modest enrichment at small *P* values.

b) *P* values for the test of protein level differences between males and females plotted as a function of the genomic coordinate for each protein. The dashed line is at significance threshold Bonferroni *P* = 0.05. All the proteins that passed the threshold are highlighted with larger dots and labelled with gene names. All of the seven proteins that significantly differed between males and females mapped to either the X or Y chromosome.



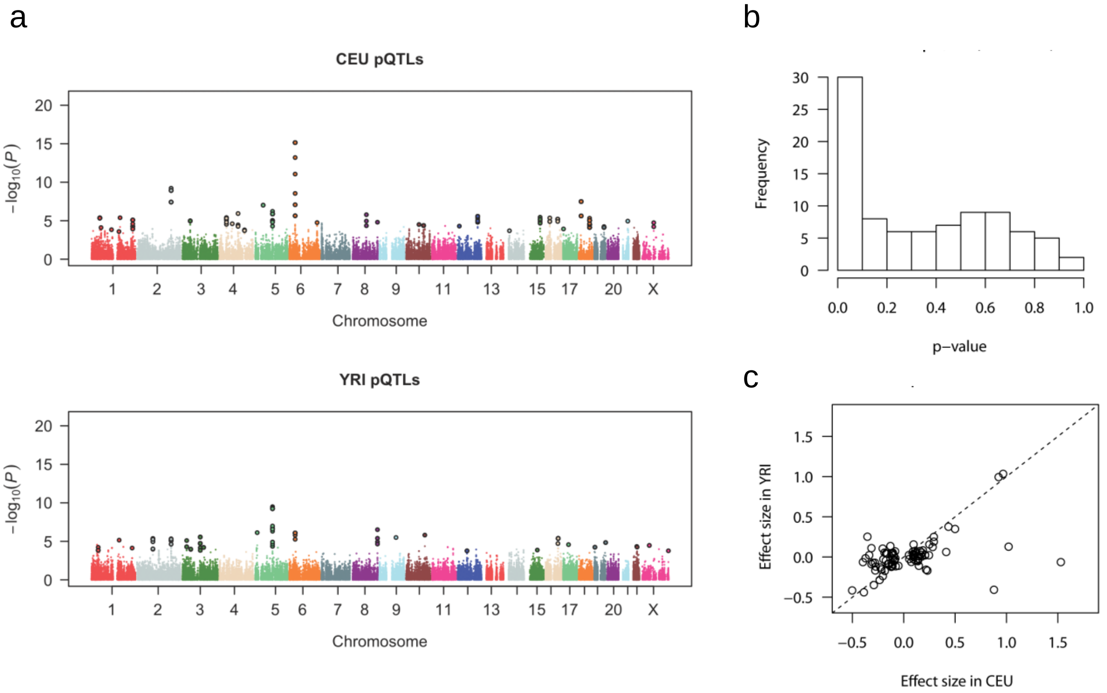
Supplementary Fig. 5 | Correspondence between protein-protein covariation and RNA-RNA covariation.

a) The correlations of all RNA pairs using RNA-Seq data in CEU and YRI is symmetrically distributed around zero. **b**) The distribution of the correlations of RNA pairs corresponding to the pairs in the protein covariation network is skewed toward larger positive values (with median Spearman's rank correlation coefficient 0.42 for CEU and 0.21 for YRI). Therefore, covarying proteins tend to have covarying RNAs on average.



Supplementary Fig. 6 | Distribution of protein heritability in CEU and YRI.

Distributions of the heritability of the levels of 2395 proteins in CEU (**a**) and 2534 proteins in YRI (**b**). Estimates of the narrow-sense heritability were calculated as the slope of the regression line of the child log2 protein ratio plotted as a function of the parents average log2 protein ratio, based on data with no missing measurements from 10 trios in CEU and 9 trios in YRI. Heritability estimates were centered at a median of 0.06 in CEU and 0.17 in YRI respectively. The mean heritability was 0.07 with a 95% confidence interval of [0.04, 0.09] in CEU, and 0.17 with a 95% confidence interval of [0.15, 0.19] in YRI. The mean protein level heritability was significantly different from 0 in both CEU and YRI (P = 2 x 10^{-09} and P = 3 x 10^{-62}, in CEU and YRI respectively).

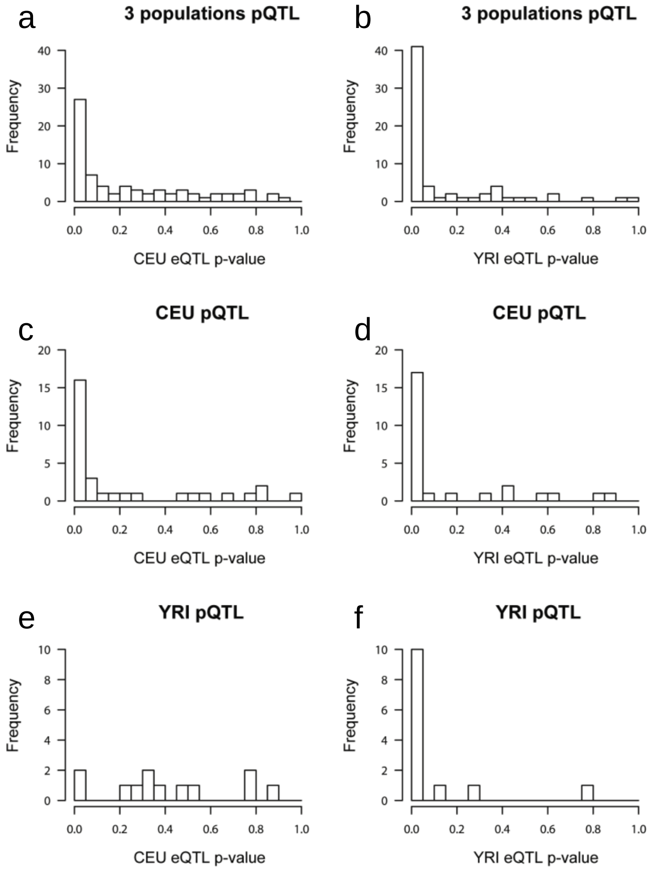


Supplementary Fig. 7 | Loci associated with protein expression levels in CEU and YRI.

a) Identification of cis-pQTLs in CEU (n=41) and YRI (n=22). The *P* value and genomic coordinate for each protein/cis-SNP association test were plotted in Manhattan plots. pQTLs with max(T) corrected *P* value < 0.001 were highlighted with a bigger dot size and a black outline. Multiple loci throughout the genome displayed an excess of small *P* values.

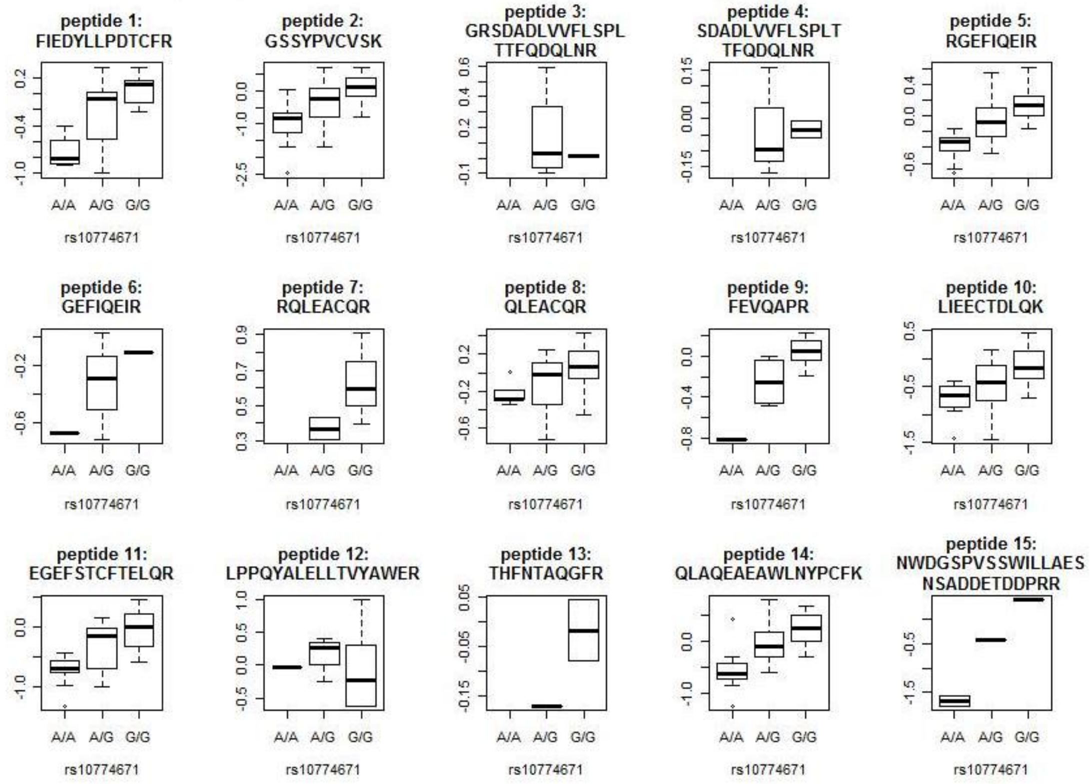
b) Estimation of the reproducibility of CEU pQTLs in YRI. The *P* value distribution for tests in YRI of CEU pQTLs identified at FDR 30% shows that the significant tail is highly enriched.

c) The regression coefficients or effect sizes of CEU pQTLs (FDR 30%) in the CEU and YRI populations are mostly consistent. These results indicate that the genetic loci that affect protein expression are often shared across populations.



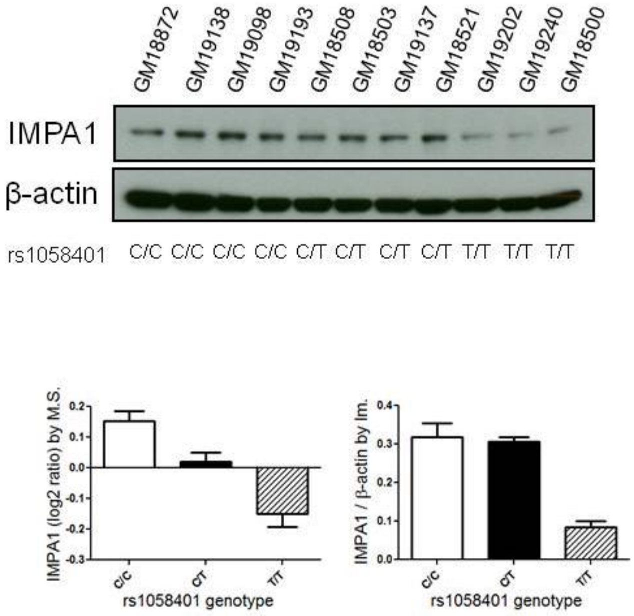
Supplementary Fig. 8 | Comparison of pQTLs and eQTLs.

a,b) Histogram of *P* values for the association of CEU RNA expression (**a**) or YRI RNA expression (**b**) with the SNPs identified as pQTLs in the combined populations (FDR 10%). Enrichment of small *P* values for association of RNA expression with the pQTL SNPs was observed by using either CEU or YRI RNA expression, but not all of the pQTLs correspond to an eQTL. **c,d**) Histogram of *P* values for the association of CEU RNA expression (**c**) or YRI RNA expression (**d**) with the SNPs identified as pQTLs in CEU. **e,f**) Histogram of *P* values for the association of CEU RNA expression (**e**) or YRI RNA expression (**f**) with the SNPs identified as pQTLs in YRI. Due to the limitation of sample size, fewer pQTLs were observed in YRI compared to CEU and the combined population analysis.



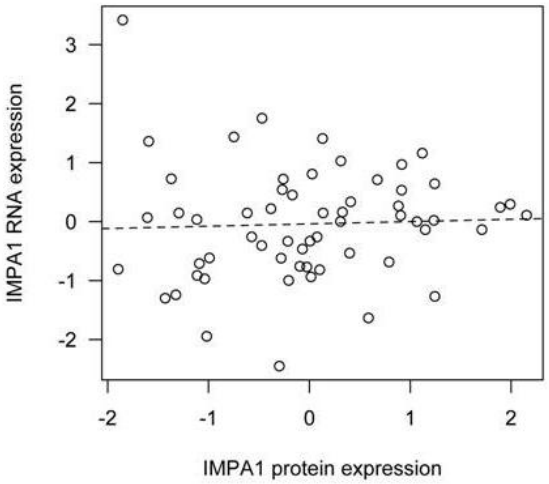
Supplementary Fig. 9 | OAS1 peptide expression levels against rs10774671 genotypes.

Expression levels of 15 unique peptides mapping to the OAS1 gene in 95 LCLs were plotted against rs10774671 genotypes. Of them, peptides 1 to 14 are not allele-specific in the 95 LCLs, which were used to quantify OAS1 protein levels. Peptides 1 to 9 are located in exon 1 and 2, and are shared by all known OAS1 isoforms in the literature. Peptide 15 is located at the junction between exon 5 and 6. SNP rs10774671 is located right before exon 6 and causes alternative splicing of OAS1. At peptide level, almost all peptides show a consistent association of the G allele with increased protein levels.



Supplementary Fig. 10 | Validation of IMPA1 protein level in the YRI population.

IMPA1 protein expression level was validated by immunoblotting in 11 YRI individuals, with their genotype at rs1058401 (the most significant pQTL) labeled at the bottom. The bar plots show the mean of IMPA1 protein level of these 11 individuals in each rs1058401 genotype category, based on data measured by quantitative mass spectrometry (left plot) and by densitometry of immunoblotting figures (right plot). Error bar, standard error of the mean. M.S., mass spectrometry. Im., immunoblotting.



Supplementary Fig. 11 | Correlation between protein and RNA for IMPA1.

Plotted is the correlation between IMPA1 RNA and protein standardized expression levels, combining results from CEU (n=37) and YRI (n=22). There is no significant correlation between protein and RNA levels (Pearson correlation coefficient r = 0.04, *P* = 0.76).