Published in final edited form as:

Nature. 2013 July 4; 499(7456): 79–82. doi:10.1038/nature12223.

Variation and Genetic Control of Protein Abundance in Humans

Linfeng Wu[1,*], **Sophie I Candille**[1,*], **Yoonha Choi**[1], **Dan Xie**[1], **Jennifer Li-Pook-Than**[1], **Hua Tang**[1], and **Michael Snyder**[1]

[1]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305

Abstract

Gene expression differs among both individuals and populations and is thought to be a major determinant of phenotypic variation. Although variation and genetic loci responsible for RNA expression levels have been analyzed extensively in human populations[1–5], our knowledge is limited regarding the differences in human protein abundance and their genetic basis. Variation in mRNA expression is not a perfect surrogate for protein expression because the latter is influenced by a battery of post-transcriptional regulatory mechanisms, and, empirically, the correlation between protein and mRNA levels is generally modest[6,7]. Here we used isobaric tandem mass tag (TMT)-based quantitative mass spectrometry to determine relative protein levels of 5953 genes in lymphoblastoid cell lines (LCLs) from 95 diverse individuals genotyped in the HapMap Project[8,9]. We found that protein levels are heritable molecular phenotypes that exhibit considerable variation between individuals, populations, and sexes. Levels of specific sets of proteins involved in the same biological process co-vary among individuals, indicating that these processes are tightly regulated at the protein level. We identified cis-pQTLs (protein quantitative trait loci), including variants not detected by previous transcriptome studies. This study demonstrates the feasibility of high throughput human proteome quantification which, when integrated with DNA variation and transcriptome information, adds a new dimension to the characterization of gene expression regulation.

We used TMT-based quantitative mass spectrometry to determine protein expression variation of LCL derived from 95 ethnically-diverse individuals genotyped in the HapMap Consortium. The samples consisted of 53 Caucasians of northern and western European ancestry (CEU); 33 Yorubans of African ancestry from Ibadan, Nigeria (YRI); eight Han Chinese from Beijing (CHB) and one Japanese from Tokyo (JPT). CHB and JPT were

Correspondence and requests for materials should be addressed to: M.S. (mpsnyder@stanford.edu) or H.T. (huatang@stanford.edu).

[*]These authors contributed equally to this work

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Author Contributions L.W. performed all the experimental work, mass spectrometry-related data analysis and part of the statistical data analysis. S.I.C. performed most of the statistical data analysis. Y.C. performed the protein covariation analysis. D.X. performed the screening of allele-specific peptides. J.L. contributed text and comments to the manuscript. L.W. and S.I.C. wrote the manuscript and participated in detailed discussion of study design and data analysis at all stages of the study. M.S. and H.T. designed and supervised the project.

Mass spectrometry raw data and searching results have been deposited in http://www.peptideatlas.org/PASS/PASS00230.

Reprints and permissions information is available at www.nature.com/reprints.

grouped together as East Asians (ASN). The ASN individuals were unrelated whereas the CEU and YRI groups included trios, and had 42 and 23 unrelated individuals, respectively. In each experiment, we used unique TMT tags to label trypsin-digested peptides from six cell lines, including a reference cell line (GM12878) and five other cell lines followed by two-dimensional liquid chromatography tandem mass spectrometry (2D LC-MS/MS) analysis (Fig. 1a).

Fifty-one experiments were performed that included biological replicates; each resulted in an average of 54,000 high-confidence peptide identifications and quantifications. Protein expression in a cell line was quantified relative to the reference cell line, using peptides that uniquely mapped to a gene and lacked any known polymorphic protein coding variant among the 95 individuals (Supplementary Methods). A total of 5953 proteins were quantified based on the analysis of 2,159,989 peptide spectra (Supplementary Table 1). To ensure adequate sample size and statistical power, most of the analyses described below focused on the 4053 proteins that were detected in more than 50% of the 74 unrelated individuals.

To assess reproducibility, we analyzed the correlation of protein level measurements between replicate and non-replicate cell lines. We observed that the Spearman’s rank correlation coefficient among non-replicates were much less than that of biological replicates, with median values 0.19 vs. 0.56 (Supplementary Fig. 1a), suggesting that TMT-based quantitative mass spectrometry technique can reproducibly detect variation in protein expression across individuals.

We observed considerable inter-individual protein variation: a median of 5.7% of the proteome changed more than 1.5 fold between pairs of individuals (Supplementary Fig. 1b). This figure is likely an underestimate because of precursor ion interference[10,11]. Although the CEU, YRI, and ASN HapMap cell lines were established in separate batches and differ in age, the coefficients of variation (CV) estimated in the different populations are highly correlated (Spearman’s rank correlation coefficients 0.68–0.82, Supplementary Fig. 1c and Supplementary Table 2), indicating that the level of inter-individual protein variation is similar across populations; therefore the observed pattern of protein variation is unlikely dominated by these exogenous factors. Furthermore, by estimation of potential peptide phosphorylation, we found little evidence that the measurements of protein variation were influenced by posttranslational modification (Supplementary Fig. 2).

To characterize the most and least variable proteins, we performed GO Ontology category analysis and found that the most variable proteins were enriched in immune response, whereas the least variable proteins were enriched in housekeeping processes (Supplementary Fig. 3). These findings are similar to that observed in previous mRNA studies[12]. However, caution should be taken when comparing variability between proteins, because peptide ratios measured by isobaric tag-based mass spectrometry can be distorted during precursor ion isolation[10,11]. Since precursor ion interference mostly compresses the peptide ratio towards one, the underlying variation in some of protein expressions may be substantially underestimated. Nonetheless, our results demonstrate a considerable variation in protein levels, particularly in immune response proteins.

As a proof of principle demonstrating that the protein measurements reflect biological variation, we sought to detect protein variation associated with biological attributes such as sex and ethnicity. To avoid the correlation between parents and offspring, we only used unrelated individuals for the analyses below, with the exception of the heritability calculations, which were based on the trios.

To identify proteins differentially expressed between males (n=36) and females (n=38), we regressed protein levels on sex, adjusting for average population differences (Supplementary Table 3). The distribution of *P* values for proteins exhibiting sex differences shows a modest enrichment at small *P* values (Supplementary Fig. 4a). At an FDR of 10%, 12 proteins are differentially expressed between sexes, among which seven have Bonferroni corrected *P* value <0.05 and all seven map to the X or the Y chromosome (Supplementary Fig. 4b). These results indicate our study captures *bona fide* variation in protein expression.

Similarly, we examined population differences in protein expression. We focused on the CEU and YRI unrelated individuals (42 CEU vs. 23 YRI), as the ASN sample size was smaller. At an FDR of 10%, 247 proteins are differentially expressed between CEU and YRI (Supplementary Table 4). The distribution of *P* values for population differences shows a much greater enrichment of small *P* values than for sex differences, and they are distributed throughout the genome (Fig. 1b, 1c). This finding further corroborates that our study can detect meaningful biological differences in protein expression.

Proteins that are part of the same complex or in the same biological process might be expected to vary synchronously, suggestive of a coordinated regulation of biological components and pathways. To determine if this is the case and to identify proteins that exhibit covariation, we constructed protein covariation networks using sparse partial correlation estimation[13]. In a sparse network, which connects proteins showing the strongest evidence of direct correlation (Supplementary Methods), 223 edges connect 278 proteins; these include five major clusters, each with at least 9 proteins (i.e. nodes) (Fig. 2, Supplementary Table 5). We performed GO ontology category analysis for the five clusters; three were enriched in protein metabolic process ($P = 4 \times 10^{-4}$), translation ($P = 2 \times 10^{-9}$), and glycolysis ($P = 2 \times 10^{-11}$), respectively. We also found many smaller clusters that consisted of subunits of protein complexes, e.g. minichromosome maintenance complex components. Many of these edges connect known interacting proteins. Enrichment analysis showed the known interacting proteins are significantly enriched in the protein covariaton network ($P = 5 \times 10^{-6}$). Relaxing the stringency of direct correlation while maintaining high statistical confidence, assessed by permutation and sub-sampling analyses (Supplementary Methods), yielded a denser network with 1012 edges connecting 944 proteins, featuring a "megacluster" of proteins that is enriched in translation ($P = 2 \times 10^{-6}$) (Supplementary Table 6). These results demonstrated that protein expression in a cell is highly coordinated and that, for several important biological processes (e.g. translation and glycolysis), tight control of protein levels is maintained.

We also investigated the correspondence between protein-protein covariation and RNA-RNA covariation obtained by RNA sequencing (RNA-Seq) in CEU and YRI LCLs[2,3]. We observed that covarying proteins tend to correspond to covarying RNAs with median correlation 0.42 for CEU and 0.21 for YRI (Supplementary Fig. 5). However, protein and RNA do not correlate perfectly suggesting that variation in protein levels is not entirely regulated through RNA expression.

To assess the extent and nature of the genetic factors that affect protein levels, we estimated the “narrow-sense” heritability of protein levels, which represents the additive genetic component of protein levels and is calculated based on the midparent-offspring regressions in trios. Median heritability of protein levels was 0.06 and 0.17 in CEU and YRI, respectively; 38% of the CEU proteins and 47% of the YRI proteins had a heritability higher than 0.2, respectively (Supplementary Fig. 6, Supplementary Table 7). Overall, proteins in YRI cell lines show greater heritability than in CEU cell lines. Previous analyses on RNA level heritability have shown a similar trend[1], which may be attributable to the newer age of the YRI cell lines relative to the CEU cell lines.

We also tested the association of cis genetic variation with protein levels using HapMap phase III genotypes[9]. We limited the search for protein quantitative trait loci (pQTLs) to those SNPs located between +/− 20 kb of the gene region with minor allele frequency (MAF) > 10% in our samples. We performed a cis-pQTL analysis separately in CEU, YRI, and in CEU, YRI, and ASN combined, in an effort to reveal pQTLs common to all populations. Multiple loci throughout the genome displayed an excess of small *P* values (Fig. 3a, Supplementary Fig. 7a). At a 10% FDR threshold, we detected 33, 13 and 77 genes with at least one significant pQTL in CEU, YRI and in all three populations combined, respectively (Table 1, Supplementary Table 8). Of the 77 genes with a pQTL in the analysis combining all three populations, 34 were also identified in CEU and/or YRI population. Indeed the CEU pQTLs are highly enriched for significant *P* values and tend to have consistent regression coefficients or effect sizes in YRI (Supplementary Fig. 7b, 7c). These results suggest that there is a considerable overlap in the genetic architecture of protein expression across populations. The lower number of significant pQTLs detected in YRI is likely a consequence of the smaller sample size.

To what extent do the genetic determinants that affect RNA levels coincide with those that regulate protein levels? To address this question the genetic regions that affect protein expression (pQTLs) were compared with those that affect RNA expression (eQTLs) previously identified in HapMap individuals using RNA-Seq methods[2,3]. For each pQTL SNP, we obtained the *P* value for its association with RNA expression in CEU and YRI. Overall, we observed enrichment for small *P* values (Supplementary Fig. 8, Supplementary Table 8), and we estimate that approximately one half of pQTLs are likely also eQTLs. On the other hand, many pQTLs do not correspond to eQTLs, even at a relaxed statistical stringency. We note that the numbers of pQTLs detected in this study are relatively small due to the limited sample size. Therefore, the proportions of genetic variants contributing to both protein and mRNA variation and specific to protein variation should be considered as approximations. Nonetheless, our results indicate that despite an overlap between eQTLs and pQTLs, many pQTLs are distinct from eQTLs.

Manual inspection of the individual pQTLs revealed interesting variants in several cases. OAS1 (2′-5′-oligoadenylate synthase 1) is an essential protein involved in the innate immune response to viral infection. Mutations in *OAS1* have been associated with susceptibility to viral infection[14]. We identified a pQTL for OAS1. The variant showing the strongest correlation with OAS1 protein level is located at a splice site (rs10774671), where the G allele is associated with higher protein level than the A allele. OAS1 protein levels were calculated based on the quantification of 14 unique peptides, all of which are located before the splice site variant. Nine of them are shared by all known OAS1 isoforms in the literature. All of the used peptides have the same expression orientation at rs10774671, indicating that this SNP is associated with total protein level variation (Supplementary Fig. 9). The G allele at rs10774671 has previously been associated with higher enzyme activity but the underlying mechanism is unknown[15]. Our data suggest that this variant may influence the overall OAS1 protein expression, in addition to giving rise to different isoforms.

A second example, IMPA1 (Inositol monophosphatase 1), is a putative target for lithium in the treatment of bipolar disorder[16], but no *IMPA1* genetic variant has been associated with bipolar disease[17], nor has an eQTL been identified for this gene in recent RNA-Seq studies[2,3]. We found that SNP rs1058401, located at the 3′ UTR of the *IMPA1* gene, is associated with protein levels. We first explored a fine cis-pQTL mapping of *IMPA1* gene using denser SNP coverage. We selected all the SNPs within +/− 200 kb of the *IMPA1* gene from HapMap phase I, II and III with a MAF > 5%. Several SNPs on or near the 3′UTR show significant pQTL effect in CEU and in the three populations combined (Fig. 3b). We validated this pQTL by immunoblot analyses in both CEU and YRI (Fig. 3c, 3d, Supplementary Fig. 10). The results are consistent with the data obtained using mass spectrometry, confirming that rs1058401 is indeed associated with IMPA1 protein levels.

We also evaluated the correlation between IMPA1 protein and mRNA levels, and observed a poor correlation between protein and mRNA in the combined sample ($r = 0.04$, $P = 0.76$, Supplementary Fig. 11) or in CEU alone ($r = −0.19$, $P = 0.27$). However, protein and mRNA levels do show moderate correlation in YRI ($r= 0.50$, $P = 0.02$). The rs1058401 SNP showed no evidence of association with RNA levels measured in CEU ($P = 0.56$), moderate evidence of association with RNA levels in YRI ($P = 0.008$), and much stronger evidence of association with protein levels ($P = 3 \times 10^{-7}$, in the combined populations analysis). We checked if this SNP is associated with mRNA decay rate using results from a recent report[18], and found no support for such a hypothesis. Therefore this pQTL may have a significant role in regulating gene expression at the translational level.

In summary, we describe the first systematic interrogation of the genetic effects on the human proteome using isobaric tag-based quantitative mass spectrometry. Our results demonstrate the power of quantitative mass spectrometry data for analysis of protein co-regulation and uncovering genetic effects influencing protein abundance. With a larger number of cell lines and improvement of mass spectrometry technology, the number of pQTLs is likely to increase substantially. Some, but not all pQTLs overlap with those identified in eQTL studies. These results indicate that distinct and diverse genetic mechanisms control gene expression at many different levels, suggesting that important and complementary knowledge can be acquired by systematically characterizing the human proteome.

Methods summary

Lymphoblastoid cell lines (LCLs) from 95 HapMap individuals were obtained from the Coriell Institute for Medical Research. All trypsin-digest mixtures were analyzed on an LTQ Orbitrap Velos (Thermo Scientific) equipped with an online 2D nanoACQUITY UPLC System (Waters) as previously described with modifications[19]. The acquired mass spectrometry raw data 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 SEQUEST algorithm[21] (Proteome Discoverer software, version 1.2, Thermo Scientific). The correspondence between proteins, genes (Ensembl gene IDs) and genomic loci was established based on the protein and gene cross-reference tables of IPI database version 3.87 and transcript sequences of Ensembl database release 62. Screening of peptides overlapping with protein coding changes was based on genotypes and annotations releases by the HapMap and 1000 Genomes Project[9,22,23]. To estimate the false discovery rate for sex, population and pQTL analyses, the QVALUE Bioconductor package was used[24]. For full methods, see Supplementary Information.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank S. Montgomery for providing gene-level RNA sequencing measurements of CEU LCLs. This work was supported by NIH grants to Drs. Snyder and Tang.

References

1. Stranger BE, et al. Population genomics of human gene expression. Nature Genet. 2007; 39:1217–1224. [PubMed: 17873874]
2. Montgomery SB, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010; 464:773–777. [PubMed: 20220756]
3. Pickrell JK, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010; 464:768–772. [PubMed: 20220758]
4. Stranger BE, et al. Patterns of cis regulatory variation in diverse human populations. PLoS Genet. 2012; 8:e1002639. [PubMed: 22532805]
5. Kasowski M, et al. Variation in transcription factor binding among humans. Science. 2010; 328:232–235. [PubMed: 20299548]
6. Schwanhausser B, et al. Global quantification of mammalian gene expression control. Nature. 2011; 473:337–342. [PubMed: 21593866]
7. de Sousa Abreu R, Penalva LO, Marcotte EM, Vogel C. Global signatures of protein and mRNA expression levels. Mol Biosyst. 2009; 5:1512–1526. [PubMed: 20023718]
8. Ong SE, Mann M. Mass spectrometry-based proteomics turns quantitative. Nature Chem Biol. 2005; 1:252–262. [PubMed: 16408053]
9. Altshuler DM, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010; 467:52–58. [PubMed: 20811451]
10. Ow SY, et al. iTRAQ underestimation in simple and complex mixtures: “the good, the bad and the ugly”. J Proteome Res. 2009; 8:5347–5355. [PubMed: 19754192]
11. Karp NA, et al. Addressing accuracy and precision issues in iTRAQ quantitation. Mol Cell Proteomics. 2010; 9:1885–1897. [PubMed: 20382981]

12. Li J, Liu Y, Kim T, Min R, Zhang Z. Gene expression variability within and between human populations and implications toward disease susceptibility. PLoS Comput Biol. 2010; 6

13. Peng J, Wang P, Zhou N, Zhu J. Partial Correlation Estimation by Joint Sparse Regression Models. J Am Stat Assoc. 2009; 104:735–746. [PubMed: 19881892]

14. Lim JK, et al. Genetic variation in OAS1 is a risk factor for initial infection with West Nile virus in man. PLoS Pathog. 2009; 5:e1000321. [PubMed: 19247438]

15. Bonnevie-Nielsen V, et al. Variation in antiviral 2′,5′-oligoadenylate synthetase (2′5′AS) enzyme activity is controlled by a single-nucleotide polymorphism at a splice-acceptor site in the OAS1 gene. Am J Hum Genet. 2005; 76:623–633. [PubMed: 15732009]

16. Agam G, et al. Knockout mice in understanding the mechanism of action of lithium. Biochem Soc Trans. 2009; 37:1121–1125. [PubMed: 19754464]

17. Sjoholt G, et al. Examination of IMPA1 and IMPA2 genes in manic-depressive patients: association between IMPA2 promoter polymorphisms and bipolar disorder. Mol Psychiatry. 2004; 9:621–629. [PubMed: 14699425]

18. Pai AA, et al. The contribution of RNA decay quantitative trait Loci to inter-individual variation in steady-state gene expression levels. PLoS Genet. 2012; 8:e1003000. [PubMed: 23071454]

19. Chen R, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell. 2012; 148:1293–1307. [PubMed: 22424236]

20. Kersey PJ, et al. The International Protein Index: an integrated database for proteomics experiments. Proteomics. 2004; 4:1985–1988. [PubMed: 15221759]

21. Jimmy K, Eng ALM, John R, Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom. 1994; 5:976–989. [PubMed: 24226387]

22. Frazer KA, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007; 449:851–861. [PubMed: 17943122]

23. Abecasis GR, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491:56–65. [PubMed: 23128226]

24. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA. 2003; 100:9440–9445. [PubMed: 12883005]

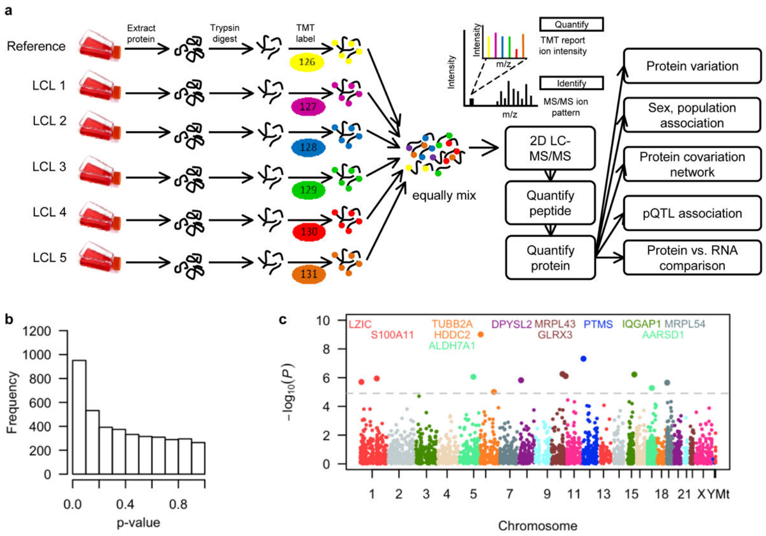

Fig. 1. Overview of workflow and protein association with ethnicity

a) Flow chart of experimental scheme. In each experiment, peptide digests from a reference cell line (GM12878) and five other cell lines were each labeled with one of the TMT-sixplex tags. Labeled peptides were equally mixed and subjected to identification and quantification by mass spectrometry, and then used for protein quantification. A total of 51 experiments were performed.

b) The *P* value distribution for the difference in protein levels between CEU and YRI shows enrichment at small *P* values.

c) *P* value of protein level differences between CEU and YRI 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 labeled with gene names. Proteins that differed between CEU and YRI are distributed throughout the genome.

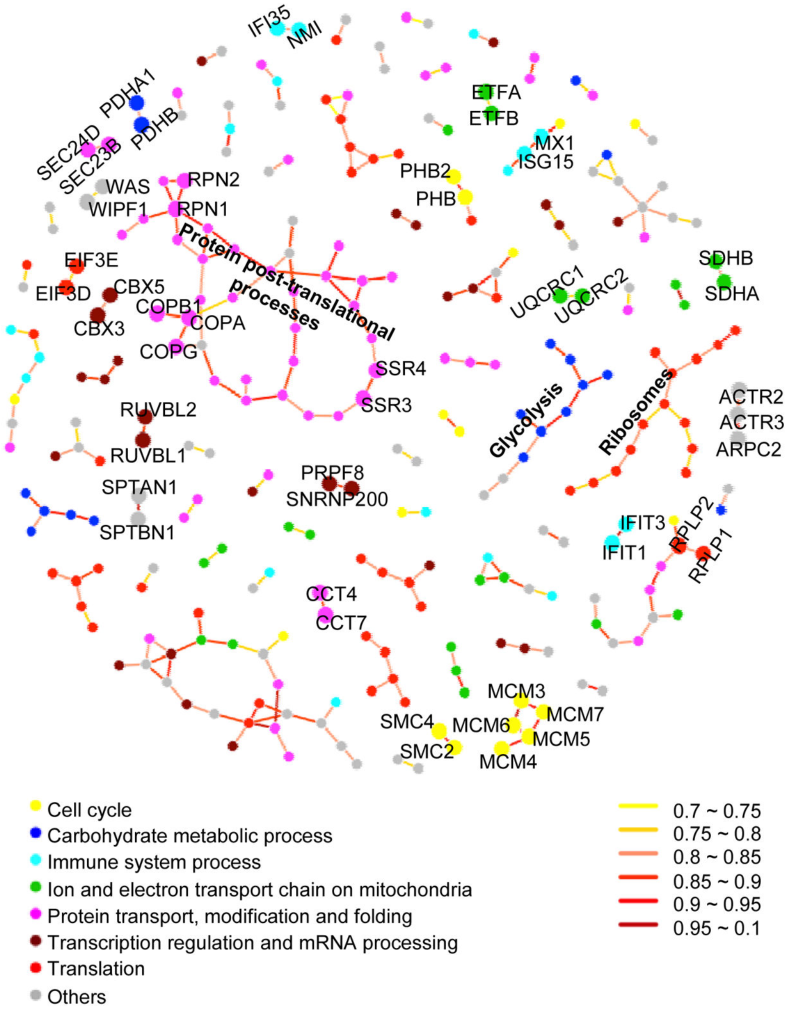

Fig. 2. Protein covariation network generated by sparse partial correlation estimation

Nodes represent proteins. Edges represent connection by covariation. This sparse network displays the 223 strongest connections among 278 proteins. Protein function was annotated by node color. Edge color was categorized according to correlation value. Known protein-protein interacting pairs were highlighted in larger nodes and labeled with gene names.

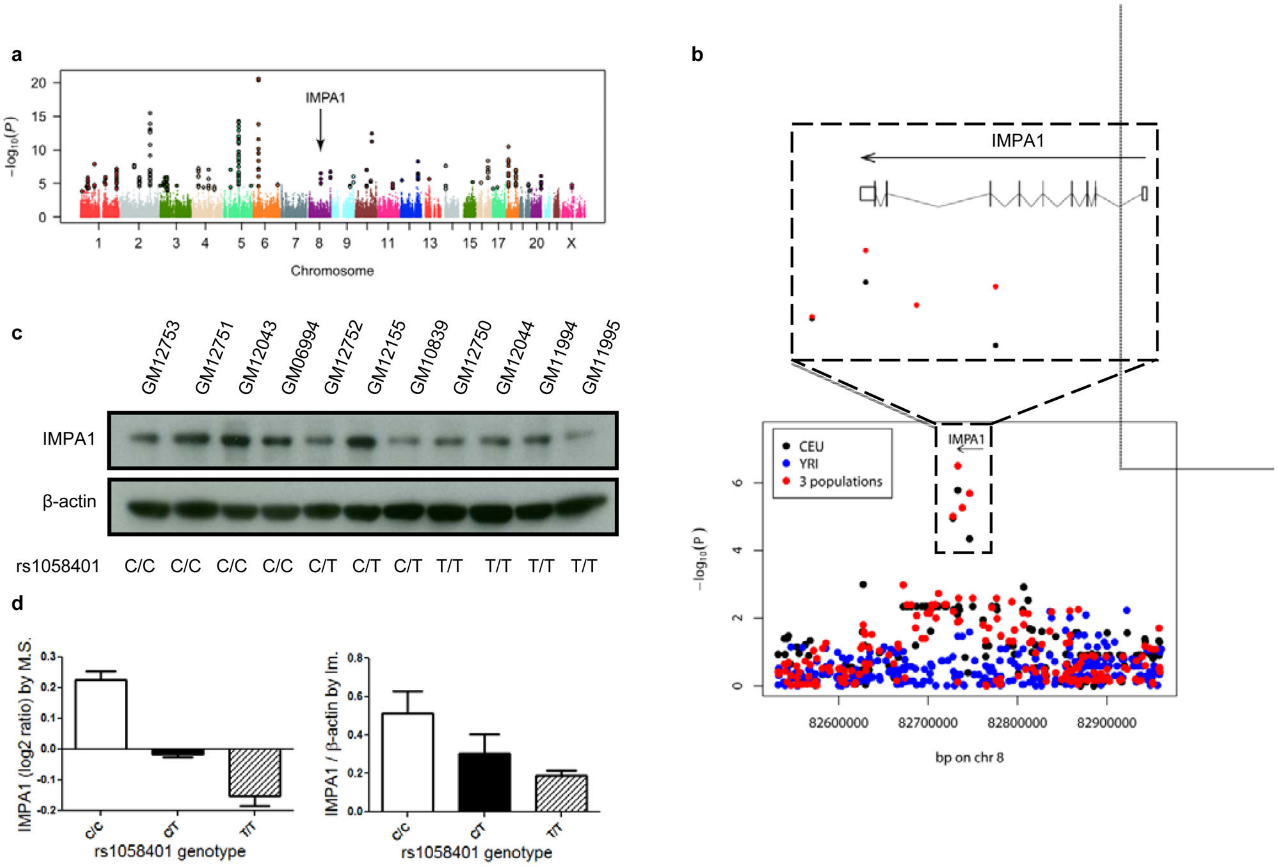

Fig. 3. Loci associated with protein expression levels

a) Identification of cis-pQTLs in all three populations combined (n=72). The *P* value and genomic coordinates for each protein/cis-SNP association test were plotted in the Manhattan plot. 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. Arrow indicates the location of the *IMPA1* gene which contains a significant cis-pQTL.

b) Overview of IMPA1 protein level and SNP genotype association in CEU, YRI, and all populations combined. The bottom plot is the fine mapping of cis-pQTL for IMPA1 based on HapMap I, II and III genotypes release 28. Each dot represents a tested SNP. Dot colors represent testing groups. The arrow is indicative of the chromosome location and transcription direction of the *IMPA1* gene. There are several highly significant associations near the *IMPA1* region in CEU and all populations combined. The exact locations of these associations in the *IMPA1* gene region are illustrated in the top plot. The most significant SNP is rs1058401, located in *IMPA1* 3′UTR.

c) Validation of IMPA1 protein expression level. IMPA1 protein expression level was validated by immunoblotting in 11 CEU individuals, with their genotype at rs1058401 labeled at the bottom.

d) The bar plots show the mean of IMPA1 protein level of these 11 individuals in each rs1058401 genotype, based on data measured by quantitative mass spectrometry and by densitometry of immunoblot blots. Error bar, standard error of the mean. M.S., mass spectrometry. Im., immunoblotting.

Table 1

Number of cis-pQTLs at different FDR

group	No. of LCLs	No. of proteins	No. of tests	No. of genes with a pQTL		
				10% FDR	20% FDR	30% FDR
CEU	41	3,984	116,556	33	54	122
YRI	22	4,017	121,405	13	34	50
3 pop.	72	4,021	130,505	77	134	239