Variation and genetic control of protein abundance in humans

Linfeng Wu1*, Sophie I. Candille1*, Yoonha Choi1, Dan Xie1, Lihua Jiang1, Jennifer Li-Pook-Than1, Hua Tang1 & Michael Snyder1

Gene expression differs among 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 analysed extensively in human populations1–3, our knowledge is limited regarding the differences in human protein abundance and the genetic basis for this difference. Variation in messenger RNA expression is not a perfect surrogate for protein expression because the latter is influenced by an array of post-transcriptional regulatory mechanisms, and, empirically, the correlation between protein and mRNA levels is generally modest4–6. Here we used isobaric tag-based quantitative mass spectrometry to determine relative protein levels of 5,953 genes in lymphoblastoid cell lines from 95 diverse individuals genotyped in the HapMap Project5,9,10. 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 covary 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 analysis of 2,159,989 peptide spectra (Supplementary Table 1). To

**Figure 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 labelled with one of the TMT-sixplex tags. Labelled 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 of the proteins that passed the threshold are highlighted with larger dots and labelled with gene names. Proteins that differed between CEU and YRI are distributed throughout the genome.

*These authors contributed equally to this work.

Reference

©2013 Macmillan Publishers Limited. All rights reserved
ensure adequate sample size and statistical power, most of the analyses described below focused on the 4,053 proteins that were detected in more than 50% of the 74 unrelated individuals.

To assess reproducibility, we analysed 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 was much less than that of biological replicates, with median values 0.19 versus 0.56 (Supplementary Fig. 1a), suggesting that the 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 probably an underestimate because of precursor ion interference10,11. 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 to be 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 post-translational modification (Supplementary Fig. 2).

To characterize the most and least variable proteins, we performed Gene Ontology (GO) 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 those observed in previous mRNA studies22. However, caution should be taken when comparing variability between proteins, because peptide ratios measured by isotopic tag-based mass spectrometry can be distorted during precursor ion isolation10,11. Because precursor ion interference mostly compresses the peptide ratio towards one, the underlying variation in some 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 a false discovery rate (FDR) of 10%, 12 proteins are differentially expressed between sexes, among which seven have a Bonferroni corrected FDR of 10%, 12 proteins are differentially expressed in protein metabolic process (P = 4 × 10^{-4}), translation (P = 2 × 10^{-3}) and glycolysis (P = 2 × 10^{-11}), respectively. We also found many smaller clusters that consisted of subunits of protein complexes, for example, minichromosome maintenance complex components. Many of these edges connect known interacting proteins. Enrichment analysis showed that the known interacting proteins are significantly enriched in the protein coexpression network (P = 5 × 10^{-5}). 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 1,012 edges connecting 944 proteins, featuring a ‘megcluster’ of proteins that is enriched in translation (P = 2 × 10^{-6}) (Supplementary Table 6). These results demonstrated that protein expression in a cell is highly coordinated and that, for several important biological processes (for example, 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 LCLs2,3. 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, indicating that variation in protein levels is not entirely regulated through RNA expression.

![Protein covariation network generated by sparse partial correlation estimation](image)

**Figure 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 is annotated by node colour. Edge colour is categorized according to correlation value. Known protein–protein interacting pairs are highlighted in larger nodes and labelled with gene names.
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 estimated as the slope of regressing the offspring trait values on the average trait values of their parents. 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 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, 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. We limited the search for protein quantitative trait loci (pQTLs) to those single nucleotide polymorphisms (SNPs) located between ±20 kilobases (kb) of the gene region with minor allele frequency (MAF) >10% in our samples. We performed a cis-pQTL analysis separately in CEU and 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 and Supplementary Fig. 7a). At a 10% FDR threshold, we detected 33, 17 and 77 genes with at least one significant pQTL in CEU, YRI and in all three populations combined, respectively (Table 1 and Supplementary Table 8). Of the 77 genes with a pQTL in the analysis combining all three populations, 34 were also identified in the 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, c). 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 probably 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. 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 and Supplementary Table 8), and we estimate that approximately one-half of pQTLs are probably also eQTLs. However, 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. 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. Our data indicate 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, but no IMPA1 genetic variant has been associated with bipolar disease, nor has an eQTL been identified for this gene in recent RNA-seq studies.

**Figure 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 are plotted in the Manhattan plot. pQTLs with protein level multiple testing corrected P value <0.001 are highlighted with a bigger dot size and a black outline. Multiple loci throughout the genome displayed an excess of small P values. The 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 phases I, II and III genotypes release 28. Each dot represents a tested SNP. Dot colours 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 three populations combined (CEU, YRI and ASN). 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 labelled 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. Error bars indicate standard error of the mean.
studies\textsuperscript{3,5}. We found that SNP rs1058401, located at the 3’ untranslated region (UTR) of the IMPA1 gene, is associated with protein levels. We first explored a fine cis-pQTL mapping of the IMPA1 gene using denser SNP coverage. We selected all the SNPs within ±200kb of the IMPA1 gene from HapMap phases 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, d and 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 (r = 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 whether this SNP is associated with mRNA decay rate using results from a recent report\textsuperscript{15}, and found no support for such a hypothesis. Therefore, this pQTL may have a significant role in regulating gene expression at the translational level.

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 analysed on an LTQ Orbitrap Velos (Thermo Scientific) equipped with an online 2D nanoACQUITY UPLC System (Waters) as previously described, with modifications\textsuperscript{16}. The acquired mass spectrometry raw data were searched against the human International Protein Index (IPI) database, version 3.75\textsuperscript{20}, concatenated with a decoy database with all the protein sequences in reverse order, using SEQUEST\textsuperscript{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 annotation releases by the HapMap and 1000 Genomes Project\textsuperscript{17,18}. To estimate the false discovery rate for sex, population and pQTL analyses, the QVALUE Bioconductor package was used\textsuperscript{19}. For full methods, see Supplementary Information.

**Table 1** Number of cis-pQTLs at different FDR

| Group | Number of LCLs | Number of proteins | Number of tests | Number of genes with a pQTL |
|-------|----------------|--------------------|----------------|-----------------------------|
| CEU   | 41             | 3,984              | 116,556        | 33                         |
| YRI   | 22             | 4,017              | 121,405        | 13                         |
| Three populations* | 72 | 4,021 | 130,505 | 77 |

* CEU, YRI and ASN populations combined.

1. Stranger, B. E. et al. Population genomics of human gene expression. Nature Genet. 39, 1217–1224 (2007).
2. Montgomery, S. B. et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 464, 773–777 (2010).
3. Pickrell, J. K. et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464, 768–772 (2010).
4. Stranger, B. E. et al. Patterns of cis regulatory variation in diverse human populations. PLoS Genet. 8, e1002639 (2012).
5. Kasowski, M. et al. Variation in transcription factor binding among humans. Science 328, 232–235 (2010).
6. Schwanhäusser, B. et al. Global quantification of mammalian gene expression control. Nature 473, 337–342 (2011).
7. de Sousa Abreu, R., Penalva, L. O., Marcotte, E. M. & Vogel, C. Global signatures of protein and mRNA expression levels. Mol. Biosyst. 5, 1512–1526 (2009).
8. Ong, S. E. & Mann, M. Mass spectrometry-based proteomics turns quantitative. Nature Chem. Biol. 1, 252–262 (2005).
9. Arftsburger, D. M. et al. Integrating common and rare genetic variation in diverse human populations. Nature 477, 52–58 (2011).
10. Ow, S. Y. et al. iTRAQ underestimation in simple and complex mixtures: “the good, the bad and the ugly”. J. Proteome Res. 8, 5347–5355 (2009).
11. Karp, N. A. et al. Addressing accuracy and precision issues in iTRAQ quantitation. Mol. Cell. Proteomics 9, 1885–1897 (2010).
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. 6, e1000910 (2010).
13. Peng, J., Wang, P., Zhou, N. & Zhu, J. Partial correlation estimation by joint sparse regression models. J. Am. Stat. Assoc. 104, 735–746 (2009).
14. Lim, J. K. et al. Genetic variation in OAS1 is a risk factor for initial infection with West Nile virus in man. PLoS Pathog. 5, e1000321 (2009).
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. 76, 623–633 (2005).
16. Agam, G. et al. Knockout mice in understanding the mechanism of action of lithium. Biochem. Soc. Trans. 37, 1121–1125 (2009).
17. Sjöholt, G. et al. Examination of IMPA1 and IMPA2 genes in manic-depressive patients: association between IMPA2 promoter polymorphisms and bipolar disorder. Mol. Psychiatry 9, 621–629 (2004).
18. Pai, A. A. et al. The contribution of RNA decay quantitative trait Loci to interindividual variation in steady-state gene expression levels. PLoS Genet. 8, e1003000 (2012).
19. Ren, R. et al. Personal omics profiling reveals dynamic molecular and cellular phenotypes. Cell 148, 1293–1307 (2012).
20. Kersky, P. J. et al. The International Protein Index: an integrated database for proteomics experiments. Proteomics 4, 1895–1908 (2004).
21. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. 5, 976–989 (1994).
22. Frazer, K. A. et al. A second generation human haplotype map of over 3.1 million SNPs. Nature 449, 851–861 (2007).
23. Abecasis, G. R. et al. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65 (2012).
24. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. Proc. Natl Acad. Sci. USA 100, 9440–9445 (2003).

Received 16 August 2012; accepted 26 April 2013.

Published online 15 May 2013.