Identification of Quantitative Trait Loci Underlying Proteome Variation in Human Lymphoblastoid Cells*\‡

Nikhil Garge‡§, Huaqin Pan§¶, Megan D. Rowland‡, Benjamin J. Cargile‡, Xinxin Zhang‡, Phillip C. Cooley¶, Grier P. Page∥, and Maureen K. Bunger‡§∗

Population-based variability in protein expression patterns, especially in humans, is often observed but poorly understood. Moreover, very little is known about how interindividual genetic variation contributes to protein expression patterns. To begin to address this, we describe elements of technical and biological variations contributing to expression of 544 proteins in a population of 24 individual human lymphoblastoid cell lines that have been extensively genotyped as part of the International HapMap Project. We determined that expression levels of 10% of the proteins were tightly correlated to cell doubling rates. Using the publicly available genotypes for these lymphoblastoid cell lines, we applied a genetic association approach to identify quantitative trait loci associated with protein expression variation. Results identified 24 protein forms corresponding to 15 proteins for which association approach to identify quantitative trait loci associated with protein expression variation. Results identified 24 protein forms corresponding to 15 proteins for which genetic elements were responsible for >50% of the expression variation. The genetic variation associated with protein expression levels were located in \textit{cis} with the gene coding for the transcript of the protein for 19 of these protein forms. Four of the genetic elements identified were coding non-synonymous single nucleotide polymorphisms that resulted in migration pattern changes in the two-dimensional gel. This is the first description of large scale proteomics analysis demonstrating the direct relationship between genome and proteome variations in human cells. \textit{Molecular & Cellular Proteomics} 9: 1383–1399, 2010.

Recent research has shown that gene expression variation in both humans and model organisms behaves as a complex genetic trait (1–5). The genetic basis of gene expression can be treated in a manner similar to that of other complex quantitative phenotypes, such as body weight and blood pressure. The identification of quantitative trait loci associating with individual mRNA expression levels (expression quantitative trait loci (eQTL)\(^1\)) has been reported in yeast and mice using recombinant inbred strains (6–11). It has also been described in humans using lymphocytes, Epstein-Barr virus-transformed lymphoblastoid cell lines (LCLs), and tissue biopsies (10, 12–21). Several recent studies using eQTL analysis have also been used to narrow the list of candidate genes for Type I diabetes and coronary artery disease susceptibility in obese mouse strains (7, 12, 13). These initial studies have established that more than 80% of mRNA expression phenotypes are heritable and that, on average, 30% of the variation is due to the genetic sources. As a result, the mapping of eQTL responsible for gene expression variation in disease suggests that eQTL association studies may be an additional powerful method in uncovering the causal genetic determinants of disease and disease susceptibility (5, 7, 11, 12, 18, 20, 22–26).

Most disease phenotypes have both causal and reactive relationships to several hundred protein changes across multiple tissue types and environmental variables (26–28). Consequently, the genetic components that are involved in complex disease could number in the hundreds. However, disease phenotype variations must relate to variation of individual proteins. One hypothesis is that each protein trait should be less complexly associated with the genome than is a clinical disease phenotype. Therefore, determining associations between genetic polymorphism and individual protein traits affords a method of building the complex relationship between genome and disease from the bottom up. Furthermore, an understanding of this relationship has additional practical implications in disease and drug response protein biomarker research in humans where identification of associating poly-

\(^1\) The abbreviations used are: eQTL, expression quantitative trait locus (loci if plural); LCL, lymphoblastoid cell line; 2D, two-dimensional; BH-FDR, Benjamini and Hochberg false discovery rate; CEPH, Centre d’Etude du Polymorphisme; CEU, Caucasians of European descent living in Utah; GWA, genome-wide association; LD, linkage disequilibrium; NS-SNP, non-synonymous single nucleotide polymorphism; PDT, population doubling time; peQTL, protein expression quantitative trait locus (loci if plural); QTL, quantitative trait locus (loci if plural); SNP, single nucleotide polymorphism; ID, identity; LCP1, L-plastin; CPNE1, Copine-1; HCLS1, hematopoietic cell-specific lyn substrate 1; UTR, untranslated region; EXOC4, exocyst complex component 4.
Genetic Analysis of Proteome Variation

Large scale quantitative proteomics is a challenging methodological process. Shotgun proteomics approaches in which proteins are digested in mixtures and separated by liquid chromatography upstream of mass spectrometric detection is increasingly used for large scale unbiased proteomics. In this approach, relative protein quantitation can be carried out using either differential labeling with isotopic reagents or label-free methods, such as peak height determination or spectral counting (36–41). When using proteomics to uncover the relationship between genomic polymorphisms and proteomic variation, a more appealing approach would be to target intact proteins where differences in isoforms and post-translational modifications can be reliably monitored in contrast to the shotgun approach. Currently, the most reliable method of unbiased quantitative analysis of intact proteins is 2D DIGE. This approach takes advantage of the quantitative properties of fluorescent protein labeling dyes coupled with large format 2D DIGE (42, 43).

In this study, we describe technical, biological, and genetic sources of proteomic variation in a small cohort of human LCLs derived from the Centre d’Etude du Polymorphisme (CEPH) cohort of Caucasians of European descent living in Utah (CEU) using 2D DIGE. The purpose of this work is to identify genetic elements responsible for variation in protein expression in human cells. As mentioned above, 2D DIGE is unique among proteomics methods in allowing the ability to quantitatively monitor differential modifications of intact proteins independently. Therefore, this approach should allow for observation of regulatory polymorphisms influencing protein expression levels in addition to other polymorphisms that could influence the pI and molecular weight of any given protein. To determine how much variation in protein expression levels detected by 2D DIGE analysis can be explained by genetic differences between the cell lines, we performed a protein expression quantitative trait locus (peQTL) analysis of 544 basal protein expression phenotypes using publicly available genotype data and both cis-only and genome-wide association (GWA) approaches. As a result, we identified a total of 24 peQTL associated with individual protein phenotypes. These results affirm that variability in cellular protein expression in human LCLs is influenced by polymorphisms in the genome and provide the first broad scale proteomics analysis to complement previous eQTL efforts using this model system. Furthermore, our results demonstrate that 2D DIGE has both the sensitivity and specificity to be used in understanding sources of biological variation in the proteome.

EXPERIMENTAL PROCEDURES

2D DIGE Materials—Sodium chloride, Trizma (Tris base), bromphenol blue, chloroform, N,N-dimethylformamide, phosphatase inhibitor, sodium carbonate, sodium bicarbonate, ammonium bicarbonate, ammonium monobasic phosphate, L-lysine, and α-cyano-4-hydroxycinnamic acid were obtained from Sigma-Aldrich. Urea, thiourea, SDS, DTT, Tris, and iodoacetamide were obtained from GE Healthcare. Protein labeling dyes (CyDye DIGE fluoros) with an N-hydroxysuccin-
imide ester reactive group for reaction with the ε amino group of lysine of proteins were obtained from GE Healthcare. CHAPS was obtained from USB Corp. Complete Mini protease inhibitor mixture tablets were obtained from Roche Applied Science. Methanol, acetonitrile, and HPLC grade water were obtained from Burdick & Jackson. Acetic acid was obtained from Mallinckrodt Baker. CEPH-CEU LCLs were acquired from the Coriell Institute for Medical Research.

**Cell Cultures**—Twenty-four cell lines from the CEPH-CEU collection were obtained from the Coriell repository. Identification numbers for the 24 cell lines are GM12057, GM07345, GM12145, GM10860, GM11829, GM12056, GM11840, GM11830, GM12004, GM12144, GM10846, GM07357, GM11839, GM12003, GM06994, GM11993, GM10856, GM11992, GM07000, GM13848, GM0854, GM07029, GM10838, and GM10851. Cells were cultured at 37 °C in RPMI 1640 medium supplemented with 15% fetal bovine serum (HyClone), 1 mM GM10856, GM11992, GM07000, GM13848, GM0854, GM07029, GM10838, and GM10851. Cells were washed once with methanol and resuspended in a humidified chamber containing 5% CO₂. Cells were plated at an initial density of 200,000 cells/ml and grown to between 800,000 and 1,100,000 cells/ml at which time they were harvested by washing two times in phosphate-buffered saline. Frozen pellets containing ~10⁷ cells were stored at ~80 °C.

**Sample Preparation and Protein Labeling**—Upon thawing, cells were lysed with 180 μl of TNE buffer (50 mM Tris-Cl, pH 7.6, 150 mM NaCl, 2 mM EDTA, pH 8.0, 1% Nonidet P-40) with protease and phosphatase inhibitors added. Soluble protein was extracted by centrifugation at 14,000 rpm for 10 min at 4 °C. Protein concentration was determined using the BCA Protein Assay Kit (Pierce). Protein disulfide bonds were reduced in 2 mM DTT. For each replicate, 25 μg of each sample was precipitated using 4:1 (v/v) methanol:chloroform. The samples were washed once with methanol and resuspended in L-buffer (7M urea, 2M thiourea, 4% CHAPS, 30 mM Tris, pH 8.5). An internal standard was created by pooling 8.5 μg of each sample. The remainder of each sample was then labeled with 3 pmol of either Cy3 or Cy5 dye/μg of protein. The internal pooled standard was labeled using 3 pmol/μg Cy2. Reactions were stopped by adding 10 mM lysine in equal volume to the volume of dye. For each gel, a Cy3- and Cy5-labeled protein sample was mixed with an equal volume of Cy2-labeled standard. Samples were then reduced and denatured in R-buffer (7 M urea, 2 M thiourea, 4% CHAPS, 30 mM Tris, pH 8.5) at 10:1 (v/v) with amphotolys (IPG Buffer) (GE Healthcare). The samples were then applied to a 24-cm pH 4–7 Immobiline DryStrip (GE Healthcare).

**2D DIGE and Image Analysis**—First dimension isoelectric focusing was performed using an Ettan IPIphor3 (GE Healthcare) for a total focusing time of 64.5 kV·h at 20 °C and 75 μA/strip maximum. After focusing was complete, each strip was equilibrated with 10 ml of equilibration buffer (6 M urea, 75 mM Tris, pH 8.8, 29.3% glycerol, 2% SDS, 0.002% bromphenol blue) with 10 mg/ml DTT added for 15 min. The strips were then equilibrated with 10 ml of equilibration buffer (6 M urea, 75 mM Tris, pH 8.8, 29.3% glycerol, 2% SDS, 0.002% bromphenol blue) with 10 mg/ml DTT added for 15 min. The strips were then digested using Trypsin Gold (Promega) by overnight incubation. The digests were then desalted and concentrated using C₁₈ P10 ZipTips (Millipore). One-third of each sample was loaded onto the ZipTip by pipetting up and down five times into a fresh 96-well plate. The sample was then washed four times with 10 μl of 0.1% TFA and eluted with 2 μl of 50% acetonitrile, 0.1% trifluoroacetic acid. A total of 1 μl of eluate was pipetted onto a clean MALDI plate and covered with 1 μl of α-cyano-4-hydroxycinnamic acid MALDI matrix. Mass spectra of each spot were acquired using an Applied Biosystems 4700 Proteome Analyzer MALDI-TOF-TOF instrument. Data were acquired in reflectron positive mode for the mass range 900–3500 Da. Shots per spectrum for both MS and MS/MS totaled 2000. Data-dependent MS/MS analysis was performed on the top 10 peaks from each MS spectrum. Peaks excluded from analysis correlated with trypsin autolysis peaks at 844.0, 1047.5, and 2257.5 Da. Peak lists were generated using the 4000 Series Explorer v3.0. The data were searched using Mascot v1.9.0 (Matrix Sciences) against the human NCBI protein RefSeq database build 36.3 consisting of 37,742 sequences. Search parameters included trypsin enzyme specificity with two permitted misscleavages. Variable modifications included carbamidomethyl-Cys and oxidized Met. Mass tolerances for both precursor and fragment ions were 1 Da. Identification criteria included a Mascot score >65 (selected based on a corrected p value <0.01), at least 5 orders of magnitude difference in the expect score between non-isoformic first and second hits, at least two unique MS/MS peptides scoring at least 20 (p < 0.01), and the protein having a molecular weight similar to experimentally derived values. Where the second hit was an isoform of the first and had identical scores, the first is reported. Where the second hit was an isoform with a lesser score, the higher scoring isoform is reported. Where a peptide matched to multiple proteins in the database, the peptide was assigned to the higher scoring protein. All peptide mass matches for an accepted protein are reported in supplemental Table 2 regardless of whether the matching peptide from the MS/MS search achieved the score cutoff.

**Association Analysis**—Protein expression levels from 544 spots derived from the average between two technical replicates were used as protein expression phenotypes for genetic association analysis. Genotype information from 3.8 million SNPs was downloaded from the International HapMap Project web site. Because of our small sample size, we removed SNPs with minor allele frequency <0.1 and genotype missingness >0.1 prior to association. This filtering resulted in a set of 1.7 million SNPs that were used in the final association analysis. For the cis-only analysis, SNPs were limited to the 200-kb region surrounding the transcript-coding allele. For each protein, expression levels were log-transformed and treated as the response variable. Based on the allele composition of the genotypes, each SNP (coded as 0, 1, and 2) constituted the single dependent variable, and a linear regression was performed using the genetic toolkit PLINK (44). Statistical significance of the strength of the relationship between the response variable and the dependent (genotype) variable was measured using the Wald test. Multiple testing corrections were made using the Benjamini and Hochberg false discovery rate (BH-FDR) procedure (45). Results were visualized and annotated using WGAViewer (46). Power calculations were performed using the genetic power calculator (47). We used the following settings: (a) locus-specific heritability ranging from 0.03 to 0.8, (b) sibling correlation = 0.5, (c) linkage disequilibrium (LD) (D') = 1, (d) QTL increaser allele frequency = 0.3, (e) marker allele frequency = 0.3, (f) sample size = 16, sibship size = singletons (both parents genotyped), and (g) type I error rate (α) = 0.0001 and 0.0000001. Higher sibling correlation, a measure of total heritability of the traits, increases the power. Hence, we considered a sibling correlation of 0.5 as a reasonable setting: not too high, not too low, and typical for many traits. Locus-specific
Genetic Analysis of Proteome Variation

Results

Heritability values were considered from 0.03 (very small) to 0.8 (large). Because we considered 1.7 million SNPs, there was a high chance that the markers and the pQTL were near to each other; therefore, we set LD ($D'$) to 1. LD is the non-random association of alleles at two or more loci. $D'$ is a standardized measure of the degree of LD with a maximal value of 1. Supplemental Table 5 summarizes the power for cis and genome-wide analyses. We considered an $\alpha$ of 1.00e−07 for genome-wide analysis because it involved 1.7 million SNPs per trait. The cis-only analysis involved far fewer SNPs per trait; therefore, power calculations considered an $\alpha$ of 0.0001 for cis-only analysis.

**Lyphoblast Proteome**—An overview of the experimental approach is shown in Fig. 1. A 2D DIGE analysis was performed to determine base-line proteome expression levels in LCLs derived from 24 individual cell lines derived from 16 unrelated individuals and their eight children. Gel spots were numerically identified and matched between gels using the internal standard. Gel images were manually examined to determine the quality of the matching. Incorrectly matched spots were eliminated or corrected if possible. Spots were then eliminated from analysis if they appeared to be artifacts from streaking or were affected by streaking of another spot. Spots were also eliminated if not detected in more than one gel. As a result of these quality control measures, a total of 544 protein spots were subsequently analyzed (Fig. 2).

A separate preparative gel of the pool of all 24 samples was run and poststained with Deep Purple. On this gel, 492 of the 544 proteins appeared with high spot resolution and abundance. Gel plugs for these 492 proteins were isolated and subjected to in-gel trypsin digestion followed by MALDI-MS/MS analysis. Identification of proteins was determined by combined peptide mass fingerprinting and MS/MS fragmentation pattern using Mascot. Using a Mascot cutoff score of 65 (false positive rate <1%), proteins within 311 spots were identified. These 311 spots corresponded to 243 different proteins. Predicted and actual molecular weights and pl values for proteins were compared for verification. All protein identifications are presented in supplemental Table 1. Protein and individual peptide IDs are presented in supplemental Table 2.

**Proteome Transcription Correlation Analysis**—The correlation of basal expression levels of the proteins with corresponding transcripts was analyzed using the publicly available raw transcript data as reported by Morley et al. (20) and acquired from the Gene Expression Omnibus accession number GSE2552. Pearson’s correlation coefficient was calculated to compare protein expression values for the 16 individual parents with the gene expression values for the same 16 individuals. Each expression value was derived from the average of two replicates. Of the 311 identified spots, 251 could be directly compared with transcript values from the published microarray analysis. Of these, 41 showed a significant and highly positive correlation with transcript data ($r > 0.5$ and $p < 0.05$), and only two showed a significant negative correlation ($r < −0.5$ and $p < 0.05$). The remaining proteins did not correlate significantly with transcript levels (supplemental Table 3).

**Sources of Variation**—To address a concern that lower intensity spots would exhibit greater technical variability and therefore reduce the power to determine sources of biological variation among lower abundance proteins, a linear regression was performed to determine the role of spot volume in population variance. Averaged volume of the same gel spot detected across the population of 24 individuals was used as the predictor variable. Population variance for each spot calculated from the standardized and normalized log expression abundance values across the 24 individuals was used as the dependent variable. As shown in Fig. 3A, following normalization and standardization, spot volume accounts for only 1% of the experimental differences ($r^2 = 0.01$).

Another variable that could contribute to protein expression variation is the cell population doubling time (PDT). As reported by Choy et al. (48), many transcript phenotypes in these same lymphoblast cell lines were linked to the growth rate and ATP levels of the individual cell lines. Such an observation confounds the ability to find true genetic effects. To determine the PDT, cell lines were plated at 2^5 cells/ml and were counted daily until cells reached 10^6 cells/ml. PDTs ranged from 0.21 to 0.82 doublings/day across the 24 cell lines. A regression analysis was performed using expression values for each protein as the response variable and the PDT of each cell line as the predictor variable. The expression levels of 53 proteins were significantly related to PDT ($p < 0.05$, $r^2 > 0.5$). The 53 proteins represent ~10% of the total number of proteins analyzed, slightly more than that reported for transcripts by Choy et al. (48). Protein expression was evenly split between positive PDT associations and negative
associations (supplemental Table 4). Two examples of PDT-associated protein expression patterns are shown in Fig. 3. B and C. Among these 53 spots, we were able to identify 26 proteins. Functional annotation of these proteins revealed enrichment for nucleotide-interacting proteins and proteins with ATPase activity (p = 0.003; not shown).

**cis-Acting Association Tests for peQTL**—Traditionally, an eQTL is defined as a mappable locus that affects the expression level of a particular transcript. However, a protein appearing on a 2D gel at a location other than expected due to a splice variation or non-synonymous coding polymorphism that alters the overall charge of the protein or because it is differentially modified by phosphorylation or glycosylation represents important functional consequences of polymorphisms not detected if only considering protein expression levels alone. Therefore, we define a peQTL as a mappable locus that affects the level, molecular weight, charge, or modification of a protein. We first performed an association study of each of the 243 identified proteins and SNPs that resided within 200 kb of the genetic coding region for each protein (cis-acting). For association analysis, only data for the 16 unrelated parents were used. Where multiple spots corresponded to a single protein, each spot was considered independently. Each spot trait (i.e. expression level) was regressed against the cis genotypes using PLINK (44). From this analysis, expression levels of 19 spots corresponding to 11 proteins were significantly associated with polymorphisms in cis with the protein-coding gene region (BH-FDR < 0.01).

To compare the regression models developed from the association of expression values with individual SNPs within the parent population with that of parents and children considered together, we computed the Pearson’s correlation coefficient ($r^2$) for each model with the expectation that there should be little change in $r^2$ when the children are added to the model. This $r^2$ value also represents the locus specific heritability measure for each phenotype (49). A full

![Fig. 2. Two-dimensional electrophoresis map of human lymphoblastoid cell line proteome. Soluble cellular proteins were labeled with either Cy3 or Cy5 through free amine groups. A pool of all lysates was labeled with Cy2 and run along with each sample as an internal standard. Electrophoresis was performed in two dimensions with separation by pI in the first dimension and by molecular weight in the second dimension. The image is a representative gel of the Cy2-labeled pooled internal standard. Spot numbers for spots used in the analysis are indicated on the image, and annotations of each spot are found in supplemental Tables 1 and 2. Red spot numbers indicate the location of spots that associated with peQTL (Tables I and II).](image-url)
list of the 19 proteins or protein forms, the expression of which is associated with *cis*-acting polymorphisms, is found in Table I.

Among the proteins associating with *cis*-acting SNPs was a six-spot cluster in the 2D DIGE gel identified as L-plastin (LCP1) (spots 675, 677, 687, 688, 691, and 693). Several polymorphisms in the region of the LCP1 gene were observed to be significantly associated with expression of each of these six spots (for an example of spot 691, see Fig. 4A). One of these (rs4941543) is a non-synonymous single nucleotide polymorphism (NS-SNP) that causes a Lys to Glu alteration at position 533 in the LCP1 protein. To verify genetic association with this SNP, we determined $r^2$ for the 16 parental cell lines in addition to the eight children. Relative concordance was observed between the results of the parents alone and parents plus children (Table I, last two columns). In Fig. 4B, the LCP1 protein abundance from all 24 individuals in the population for all six spots clearly shows differential allele-specific expression patterns based on the presence of the Lys-coding allele. This substitution is predicted to shift the pI of LCP1 protein from 5.3 to 5.2. Such a shift is evident on the 2D DIGE gels and segregates with Lys-coding versus Glu-coding genotypes selected from the population (Fig. 4C).

Another of the proteins associating with a *cis*-acting QTL was Copine-1 (CPNE1). Expression levels of the CPNE1 protein (spot 780) associated with a *cis*-acting peQTL on chromosome 20 (rs3746410, $p$ value = 4.94e-09). This region continues for about 600 kb upstream through intron 1 of the CPNE1 transcript coding region (Fig. 4A) and coincides with the transcript eQTL identified for the CPNE1 transcript when microarray analysis of these same cell lines was performed by two independent groups (15, 20). To determine the exact overlap between the peQTL and the eQTL, we obtained the raw transcript expression data for CPNE1 available from the GeneVar database (15, 17) and repeated the *cis* association analysis using data from only the 16 parental lines used in our peQTL analysis. Fig. 5 shows the genomic region that is significantly associated with both protein (bottom panel) and transcript expression (top panel) of CPNE1 and indicates that the regions are identical. Correlation of both the protein expression and transcript expression with a representative SNP (rs3746410) is shown in Fig. 5, B and C.

Two spots (495 and 502) identified as hematopoietic cell-specific lyn substrate 1 (HCLS1) proteins are associated with SNPs located in *cis* to the HCLS1 gene. These spots were part of a four-spot cluster that included spots 517 and 490.
| Spot | ID     | Chr | Top SNP | Top SNP p value | cis-only FDR\(^a,b\) | Chr. | Top SNP | Top SNP type | Sig. SNPs in cis region\(^b\) | Sig. SNPs within gene | Range of sig. SNPs | r\(^2\)-parents\(^c\) | r\(^2\)-all\(^d\) |
|------|--------|-----|---------|-----------------|-------------------|-----|---------|---------------|------------------------|-------------------|------------------|-----------------|-----------------|----------------|
| 495  | HCLS1  | 3   | rs1128163 | 2.78e-05        | 2.89e-04          | 3' UTR | 25      | 7             | 134,605                | 0.78              | 0.79             |                  |                 |
| 502  | HCLS1  | 3   | rs1128163 | 4.62e-05        | 4.82e-04          | 3' UTR | 25      | 8             | 134,605                | 0.306             | 0.479            |                  |                 |
| 528  | GARS   | 7   | rs17158396 | 2.04e-04        | 4.88e-03          | Downstream | 5      | 0             | 17,292                 | 0.638             | 0.2              |                  |                 |
| 675  | LCP1   | 13  | rs4941543 | 7.41e-06        | 7.37e-04          | Non-synonymous coding (Glu/Lys) | 4      | 4             | 20,617                 | 0.772             | 0.739            |                  |                 |
| 677  | LCP1   | 13  | rs4941543 | 3.10e-06        | 2.26e-04          | Non-synonymous coding (Glu/Lys) | 9      | 9             | 28,365                 | 0.798             | 0.816            |                  |                 |
| 687  | LCP1   | 13  | rs4941543 | 5.48e-05        | 5.45e-03          | Non-synonymous coding (Glu/Lys) | 4      | 4             | 20,617                 | 0.664             | 0.644            |                  |                 |
| 688  | LCP1   | 13  | rs4941543 | 2.01e-10        | 1.68e-08          | Non-synonymous coding (Glu/Lys) | 12     | 9             | 32,976                 | 0.948             | 0.901            |                  |                 |
| 691  | LCP1   | 13  | rs4941543 | 6.24e-11        | 6.21e-09          | Non-synonymous coding (Glu/Lys) | 14     | 10            | 36,172                 | 0.956             | 0.918            |                  |                 |
| 693  | LCP1   | 13  | rs4941543 | 1.79e-08        | 1.78e-06          | Non-synonymous coding (Glu/Lys) | 11     | 10            | 33,233                 | 0.902             | 0.815            |                  |                 |
| 780  | CPNE1  | 20  | rs3746410 | 4.94e-09        | 1.65e-08          | Intron | 88     | 23            | 230,220                | 0.919             | 0.815            |                  |                 |
| 1001 | CNDP2  | 18  | rs2278161 | 2.09e-05        | 2.94e-04          | Non-synonymous coding (Tyr/His) | 19     | 18            | 103,845                | 0.809             | 0.77             |                  |                 |
| 1014 | ANNX11 | 10  | rs2236556 | 3.02e-06        | 7.63e-05          | Intron | 20     | 13            | 47,939                 | 0.803             | 0.794            |                  |                 |
| 1020 | CNDP2  | 18  | rs2278161 | 9.16e-06        | 1.20e-04          | Non-synonymous coding (Tyr/His) | 21     | 18            | 56,255                 | 0.765             | 0.724            |                  |                 |
| 1505 | GLRX3  | 10  | rs13991  | 4.49e-06        | 9.25e-04          | Non-synonymous coding (Gln/His) | 1      | 1             | 1                      | 0.788             | 0.825            |                  |                 |
| 1509 | GLRX3  | 10  | rs13991  | 1.10e-05        | 2.27e-03          | Non-synonymous coding (Gln/His) | 2      | 1             | 62,185                 | 0.759             | 0.78             |                  |                 |
| 1685 | PPA2   | 4   | rs13787  | 1.45e-04        | 1.23e-03          | Non-synonymous coding (Lys/Asn) | 36     | 36            | 80,349                 | 0.829             | 0.819            |                  |                 |
| 1734 | NAPA   | 19  | rs2974225 | 6.09e-04        | 1.08e-02          | Upstream     | 4      | 0             | 22,406                 | 0.58              | 0.38             |                  |                 |
| 1753 | ANXA5  | 4   | rs3811746 | 4.14e-04        | 1.82e-02          | Intergenic   | 12     | 0             | 155,254                | 0.582             | 0.324            |                  |                 |
| 1971 | PSMB4  | 1   | rs4603   | 2.57e-06        | 5.55e-06          | Non-synonymous coding (Ile/Thr) | 34     | 2             | 112,398                | 0.866             | 0.866            |                  |                 |

\(^{a}\) FDR calculated using Benjamini and Hochberg method (45).

\(^{b}\) cis-only refers to the 200-kb region surrounding location of gene encoding for the identified protein.

\(^{c}\) r\(^2\), square of the Pearson’s product moment correlation coefficient; equivalent measure of heritability.

\(^{d}\) r\(^2\)-all includes data derived from 16 parents and their eight children.
Genetic Analysis of Proteome Variation

A

B

LCP1 Spot 675

LCP1 Spot 677

LCP1 Spot 687

LCP1 Spot 688

LCP1 Spot 691

LCP1 Spot 693

C

687 675 693 677 688 691
Spot 517 was also identified as HCLS1, and spot 490 remains technically unidentified. When expression values derived from spot 490 were used as a phenotype trait in association with cis-acting SNPs in the HCLS1 gene, the same set of SNPs was highly significant (e.g., rs1128163, p value = 9.39e–09), suggesting that the protein represented by spot 490 is likely to be HCLS1. Additional SNPs in the 200-kb region surrounding the HCLS1 gene were significantly associated with expression (Fig. 6B) and ranged from the 5′-upstream region to the 3′-UTR. The correlation of SNP rs1128163 to protein expression level across the full 24 individuals in our population for spots 490, 495, and 502 (Fig. 6, C–E) supported the model. The exception was spot 517 (Fig. 6F). We also analyzed raw transcript data for HCLS1 transcript from analyses with these same 16 parental cell lines from the GeneVar database and did not find SNPs within the HCLS1 locus that were associated with transcript variation (rs1128163, p value = 0.28).

GWA Tests for peQTL—A full GWA analysis was then performed for each of the 544 analyzed spots regardless of whether the corresponding protein was identified. Each individual protein trait (i.e., spot) in the 16 parental cell lines was tested by linear regression against a total of 1.7 million SNPs available from the International HapMap Project. We applied the BH-FDR correction for multiple testing, setting a cutoff of 0.01. Because of the small sample size and potential for false positive results due to random association, we also considered only peQTL that were identified with three or more SNPs within a 100-kb region. To locate the QTL associated with each trait, the −log10(p value) for association of each SNP with each protein trait was plotted according to the position of the SNP in each chromosome. As with the cis-only analysis, to verify the model, a representative SNP identified through association analysis with the 16 parents was plotted against protein expression from all 24 individuals, and a new Pearson’s correlation coefficient was calculated. This resulted in the identification of eight proteins with associating QTL meeting GWA significance (Table II).

Several of the cis-acting QTL for LCP1, CPNE1, and HCLS1 that were identified through the cis-targeted analysis also met statistical significance in the GWA study. The standardized abundance of three spots, 688, 691, and 693, that were identified as the LCP1 protein showed a statistically significant association with rs4941543 (as described above) in addition to a second trans-acting QTL. This second strongly associating QTL region for all three of these spots was an intragenic region of chromosome 3 (p values <10e–08). However, when protein abundance from all 24 individuals in the population was plotted against genotype for the most significant SNPs in the chromosome 3 QTL region, the correlation was less apparent than when only the 16 parents were used (supplemental Fig. 1). The cis-acting QTL identified for spots 780 (CPNE1) and 490 (HCLS1 inferred) were also significant in the GWA analysis with no additional trans-acting QTL identified.

Two putative trans-acting peQTL were identified only within the GWA analysis (Table II). The expression level of vinculin (spot 175) was significantly associated with a trans-acting QTL located within the exocyst complex component 4 (EXOC4) gene and directly upstream of LRGUK, a leucine-rich repeats and guanylate kinase-containing gene on chromosome 7 (rs12539993, p = 3.75e–10). The EXOC4 gene has been previously identified in a GWA analysis of arterial stiffness phenotypes within the Framingham Heart Study (51). However, the correlation between the genotype of a representative SNP (rs12539993) and vinculin protein expression with children included was weaker (parents r² = 0.94 versus all r² = 0.40; Table II).

The expression level of LCP1 protein in spot 675 was associated more strongly with a trans-acting allele on chromosome 14 than it was for the cis-acting allele identified in our cis-only results (compare the top spot 675 results in Tables I and II). At the chromosome 14 location, five SNPs located around or within the NCP2 gene were significantly associated (p = 1.7e–08) with expression levels of spot 675. However, as was the case with vinculin, the correlation between genotype and expression values for spot 675 with children included was weaker than when considering the parents alone (parents r² = 0.88 versus all r² = 0.37; Table II and supplemental Fig. 2).

Spots 1415, 729, and 741 were associated with QTL on chromosomes 10, 11, and 2, respectively, but the proteins in these spots remain unidentified (Table II). The QTL associated with the expression level of spot 1415 lies entirely within the coding region of the MPP7 gene encoding a membrane-associated guanylate kinase, and the QTL associated with the expression level of spot 729 lies entirely within the gene coding for GUCA1A, a guanylate cyclase protein. The region on chromosome 2 that is associated with expression levels of spot 741 is intergenic with no nearby gene features. Correlations between a representative genotype and protein expression levels among these unknowns are shown in supplemental Fig. 3.

**Fig. 4.** LCP1 protein expression association with SNP rs4941543. A, annotation of QTL on chromosome (Chr) 13 using WGAViewer showing ideogram of Chr 13 showing relative position of LCP1 gene (red vertical line) with shaded areas depicting chromosomal band patterns clustering of significant p values for association of spot 691 expression abundance near the 3′-end of the L-plastin gene, LCP1. B, plots of all six LCP1 forms against rs4941543, a non-synonymous polymorphism in the LCP1 gene resulting in either a Glu or Lys at position 533 in the LCP1 protein. The parental lines are shown in blue, and children are shown in red. C, image of gel spots in individuals representative of the each of the three genotypes for rs4941543 showing the spot pattern differences. To the right is a schematic of the LCP1 gene showing location of putative phosphorylation sites (denoted pS and pY for serine and tyrosine residues, respectively) and the E533K substitution that could result in the spot pattern observed in the 2D gel. std, standard.
DISCUSSION

To evaluate the molecular links between the genome and cellular phenotypes, we assessed the proteomic variation in a small cohort of the human HapMap CEU LCLs using 2D DIGE. Our results indicate that technical variation within the 2D DIGE approach was minimal. In this regard, we demonstrated that CPNE1 protein and transcript levels are both associated with SNPs from CPNE1 regulatory region.

Fig. 5. CPNE1 protein and transcript levels are both associated with SNPs from CPNE1 regulatory region. A, annotation of QTL on chromosome (Chr) 20 using WGAViewer showing ideogram of Chr 13 showing relative position of LCP1 gene (red vertical line) with shaded areas depicting chromosomal band patterns clustering of significant p values for association of spot 780 expression abundance near the 5′-upstream region of the CPNE1 gene. Association of both CPNE1 transcript (top) and protein expression (bottom) levels with polymorphisms surrounding the CPNE1 gene (circled) is shown. Normalized transcript expression levels were obtained from the GeneVar public database for cell lines corresponding to the 16 parental lines used for the protein expression analysis. The approximate location of the CPNE1 gene (circled) relative to significant p value SNPs (red lines) is shown. B and C, expression levels of both CPNE1 protein (B) and transcript (C) relative to the genotype of rs3746410 in the 5′-region of the CPNE1 gene. The parental cell lines are shown in blue, and children are shown in red. std, standard.
spot volume alone was not correlated to population variation. Consistent with what has been reported for base-line transcript expression in these cells, a proportion (10%) of the proteins exhibited strong correlation of expression with cell line PDT (Fig. 2) (48). Finally, using genetic association methods and the publicly available genotype data, we identified 24 protein forms for which >50% of the expression variation could be explained by genetic differences between cell lines.

We localized associating peQTL and, in some cases, the putative causative SNPs that contributed to variation in allelic protein expression levels.

Using LCLs, others have identified transcript QTL associating with drug response phenotypes (48, 52, 53). Those studies have also shown that a number of technical and biological variables, if not controlled for, could confound any attempts to identify a complex genetic contribution to pheno-
**TABLE II**

Summary of top results for genome-wide association analysis of protein expression traits

Chr., chromosome; Sig., significant; VCL, vinculin.

| Spot ID | Top SNP Chr. | GWA p value | FDR | Closest gene | Top SNP type | Sig. SNPs in gene region | Sig. SNPs within gene | Range of sig. SNPs | r²-parents⁵ | r²-all⁶ |
|---------|--------------|-------------|------|--------------|--------------|-------------------------|-----------------------|---------------------|------------|--------|
| 490     | HCLS1(1) rs1128163 | 3 9.39e-09 | 4.04e-04 | HCLS1 | 3'-UTR | 44 | 4 | 373,465 | 0.897 | 0.921 |
| 688     | LCP1 rs4941543 | 13 2.01e-10 | 2.08e-05 | LCP1 | Non-synonymous coding (Glu/Lys) | 3 | 3 | 2,601 | 0.948 | 0.901 |
| 691     | LCP1 rs4941543 | 13 6.24e-11 | 6.61e-06 | LCP1 | Non-synonymous coding (Glu/Lys) | 3 | 3 | 2,601 | 0.956 | 0.918 |
| 693     | LCP1 rs4941543 | 13 1.79e-08 | 1.86e-03 | LCP1 | Non-synonymous coding (Glu/Lys) | 3 | 3 | 2,601 | 0.902 | 0.815 |
| 780     | CPNE1 rs3746410 | 20 4.94e-09 | 3.02e-04 | CPNE1 | Intrinsic | 31 | 19 | 140,084 | 0.919 | 0.815 |
| 175     | VCL rs12539993 | 7 3.75e-10 | 2.24e-04 | EXOC4 | Intrinsic | 5 | 5 | 112,633 | 0.939 | 0.398 |
| 675     | LCP1 rs917394 | 14 1.70e-08 | 5.90e-03 | NCP2 | Intrinsic | 5 | 4 | 11,596 | 0.881 | 0.376 |
| Unknown ID | 729 rs11211933 | 11 2.97e-08 | 5.89e-03 | GUCY1A2 | Intrinsic | 7 | 7 | 153,602 | 0.895 | 0.746 |
| 741     | rs6707443 | 2 5.99e-08 | 8.32e-03 | Intergenic | Intergenic | N/A⁷ | N/A | 84,945 | 0.814 | 0.521 |
| 1415    | rs7068246 | 10 5.77e-08 | 9.25e-03 | MPP7 | Intergenic | 6 | 6 | 24,894 | 0.888 | 0.842 |

---

⁴ FDR calculated using Benjamini and Hochberg method (45).

⁵ Refers to the 200-kb region surrounding location of closest gene.

⁶ r², square of the Pearson’s product moment correlation coefficient; equivalent measure of heritability.

⁷ r²—all includes data derived from 16 parents and their eight children.

⁸ N/A, not applicable because no gene was found within 400 kb of QTL.
types, and this was most evident for growth characteristics of the individual cell lines. We addressed this issue by measuring the PDT of each of the 24 cell lines used in our study and correlated protein expression phenotypes with the results. Our analysis indicated that the proteins most strongly associated with PDT belonged to families of ATP-binding proteins and other nucleic acid-binding proteins, consistent with variation in the control cell cycles and metabolic rates. Our analysis also indicated that 10% of all proteins analyzed correlated with PDT. This is higher than was reported for transcripts (48), suggesting a stronger relationship of the proteome to metabolism and cell cycle dynamics than what is reflected at the transcript level. However, the total number of proteins strongly correlating to PDT (e.g. $r^2 > 0.5$) did not limit our ability to detect peQTL.

We then performed an analysis of genetic determinants associated with individual protein expression levels. Because of our small sample size, we selected alleles with a minor allele frequency >10%. This resulted in a large number of SNPs being excluded from the analysis and ensured that those SNPs considered would be reasonably represented in our small population. We also applied a very stringent BH-FDR cutoff of 0.01 and included verification of the association using expression and genotype values for the eight children of the 16 parents used in the analysis. These criteria were applied in an effort to avoid false positives from random associations. In light of these stringent acceptance criteria, our results are not meant to be an exhaustive study of genetic elements driving proteome variation but should be taken as a proof of concept that proteomic variation is influenced by genetics, and the details of this influence can be uncovered using our approach.

We performed the genetic association analysis using two approaches: 1) a cis-only approach that used SNPs within a 200-kb region surrounding the genes that coded for each protein to identify cis-acting SNPs that influenced downstream protein expression and 2) a second genome-wide approach that used high density SNPs across the genome to identify SNPs throughout the genome that influenced protein expression. Of the 311 proteins and protein forms that were analyzed, 19 protein forms corresponding to 11 proteins showed significant association with a cis-acting peQTL when only cis-acting SNPs were considered (Table I). Using a genome-wide approach in which 1.7 million SNPs were regressed against expression of all proteins, we identified 10 protein forms corresponding to seven proteins showing association at the genome-wide significance level. For three of these protein forms, the cis-acting QTL identified in the cis-only analysis was the most predominant QTL when the whole genome was considered. Two trans-acting peQTL were identified in the initial GWA screen; however, genotype-phenotype correlation dropped when the eight additional data points from the children were added to the original 16 parental data points (Table II). The relationship of QTL to encoding gene for another three protein forms could not be determined because the proteins within these spots have not been successfully identified.

The selection of this population of cell lines was based on the availability of high density, genome-wide, validated SNP genotypes for all the cell lines through the International HapMap Project (54). Previously reported studies using these same cell lines have shown transcript variation to be genetically heritable, and many transcript eQTL have been identified using similar methods (15, 20). Because of this, we expected some degree of overlap between previously reported eQTL results and ours. Indeed, we detected several SNPs that were significantly associated with CPNE1 protein levels that were identical to those previously identified as associating with CPNE1 mRNA levels (15, 20). The peQTL for CPNE1 expression covered the region mainly 5′ of the CPNE1 gene in both previous studies and in our results. This indicates that cis-acting transcription regulation variation contributes most prominently to population-based variation in both CPNE1 transcript and protein expression levels and further validates our methodological approach.

Another goal of this work was to determine the suitability of quantitative proteomics methods for QTL discovery. Many quantitative proteomics approaches have been described that could be applicable in a broad scale peQTL effort. Some of the more popular quantitative proteomics approaches are based on digested proteins separated in one or two dimensions where quantitation is facilitated either by stable isotope labeling of the proteins/peptides, the use of retention time measurements coupled with peak area (label-free), or spectral counting. The only peQTL approach described prior to this work that was unbiased toward any given protein was a single dimensional shotgun approach using label-free quantitation methods (34). This effort yielded the identification of 635 proteins and 154 QTL linkages in yeast (34, 55). A more targeted approach to peQTL analysis could also prove fruitful because such an approach could provide information about specific proteins as opposed to proteins detectable by proteomics methods. In this regard, a major peQTL effort in humans wherein 42 serum proteins were targeted for analysis mainly used established ELISAs (33). Likewise, an MS-based single reaction monitoring approach could yield similar results. No single quantitative proteomics approach yields the ability to analyze all proteins simultaneously, and this is an overall limitation of all proteomics approaches when compared with transcriptomics in which microarrays allow a more controlled and comprehensive result.

Whereas the 2D DIGE approach is somewhat limited in proteome coverage compared with a 2D shotgun approach, the capacity to analyze expression of intact proteins has obvious advantages in an effort to define genetic components that lead to proteome variation. Multiple spots corresponding to a single protein often occur in 2D gels. These are a result of

Molecular & Cellular Proteomics 9.7 1395
NS-SNPs or post-translational modifications, such as phosphorylation and glycosylation, which can affect protein pl and/or molecular weight. Therefore, quantitative intact proteomics afforded by the 2D DIGE method allows insight not only into overall protein expression variation but also variation in post-translational modifications such as phosphorylation and glycosylation that cause spot shifts on 2D gels. Whereas the depth of proteome coverage when using a 2D shotgun approach may be superior to 2D gel coverage, there are limitations in the capacity of shotgun approaches to detect post-translational modifications as separate protein entities. In this regard, shotgun methods cannot account for the full complexity of the proteome. Therefore, 2D DIGE and shotgun methods should be considered complementary methods in peQTL efforts.

We identified several representative examples in which NS-SNPs resulted in pl shifts evident in the 2D gel. For example, the positions of multiple spots associated with LCP1 protein within the 2D gel appear to be a result of a QTL located on chromosome 13. The QTl on chromosome 13 is in cis with the LCP1 gene, and the causative SNP is likely rs9491543, which codes for either a Glu (major allele) or a Lys (minor allele) amino acid at position 533 in LCP1 protein. Depending on which is present, the amino acid substitution changes the pl of the protein, which would be expected to cause a shift in protein migration on the 2D gel. Indeed, such a shift in protein migration is evident in images of the LCP1 spots grouped by pI of the protein, which would be expected to cause a shift in electrophoretic mobility. Among these, however, only 41 spots were shown to correlate specifically to phosphorylated forms. Two other examples where likely NS-SNPs resulted in pl shifts include the CDNP2 (spots 1001 and 1020) and GLRX1 (spots 1505 and 1509) proteins (supplemental Fig. 4).

Another critical advantage of utilizing 2D DIGE is the capacity to quantitatively monitor protein modifications as distinct phenotypes. Several of the peQTL that were not associated with NS-SNPs included the four-spot train identified as HCLS1. Among these, expression levels of spots 490, 495, and 502 showed significant association with SNPs in the 200-kb region of the HCLS1 gene on chromosome 3. Association of these polymorphisms with spot 490, which is inferred to be HCLS1, also met genome-wide significance in our GWA analysis. Neither the fourth spot in this HCLS1 cluster, spot 517, nor the HCLS1 transcript significantly associated with polymorphisms in the HCLS1 region. Interestingly, a closer examination of GWA results from spot 502 revealed a region of suggestive trans-acting association located on chromosome 1 ($p = 5.17e^{-07}$, FDR = 0.08) (supplemental Fig. 5). This peQTL had a lower $p$ value with spot 502 than the best cis-acting peQTL ($p = 4.62e^{-05}$, FDR = 0.33). The chromosome 1 peQTL is within the TNNI3K gene, which codes for a tyrosine kinase, although the HCLS1 protein is not known to be a direct target of TNNI3K-mediated phosphorylation. Collectively, these results imply that regulation of expression of these multiple forms of HCLS1 is complex, potentially due to differential regulation of post-translational modifications. This observation could lead to a model in which some of the HCLS1 forms are driven mostly by genetic factors within the HCLS1 gene, whereas other proteins modify the overall regulation. The fact that the HCLS1 mRNA was not associated with cis-acting polymorphisms suggests that the protein and mRNA regulation is uncoupled (15). Moreover, these results represent a good example of the increased depth of information gained by utilizing protein expression traits, which may be among the first indicators of the genetics underlying more complex molecular regulation. With increased sample sizes, the additional genetic or environmental factors underlying these more complex regulatory processes may become more apparent.

We also observed some inconsistencies with other reported peQTL discovery efforts, including the proportion of cis versus trans QTL identified as well as the proportion of associating SNPs falling in coding versus non-coding regions. Previously reported efforts to map genetic determinants of protein expression on a large scale have been performed for 665 mouse brain proteins from 200 animals using 2D gels (35), 42 human serum proteins from 1200 subjects using mostly ELISAs (33), and 221 peptides from 98 yeast segregants using single dimension LC-MS/MS (34). In the study of the mouse brain proteome, Klose et al. (35) evaluated shifts in brain protein spot patterns on 2D gels resulting from outcrossing two divergent mouse species and identified 688 spots that showed a shift in electrophoretic mobility. Among these, however, only 41 spots were shown to correlate with genetic heterogeneity in backcross data. Most of these did not map to the location of the protein-coding gene, suggesting that trans-acting alleles accounted for the majority of the electrophoretic mobility variation. In the study of yeast proteomic variation, Foss et al. (34) monitored 221 peptides in replicates of 98 segregants of a cross between two laboratory strains of yeast using single dimension, high mass accuracy LC-MS/MS. These results were compared with transcriptome data derived from the same samples. When compared with each other, loci were detected for 56% of transcripts but only 38% of peptides, and there was little overlap between them. Linkage for peptide abundance was more likely to be trans-acting than what was observed for transcripts. Conversely, in the study by Melzer et al. (33) that focused on serum levels of 42 proteins, eight cis-acting loci were identified, and only one trans-acting locus was identified. Similar to Melzer et al. (33), in our results, we detected mostly cis-acting genetic elements that could be verified as associating with protein expression variation.

Another inconsistency between our peQTL discovery efforts versus previous eQTL or peQTL discovery efforts is the proportion of coding versus non-coding SNPs discovered. Previously reported eQTL studies revealed that the majority of genome variation associated with transcript expression lies...
within the regulatory regions of genes rather than coding NS-SNPs (14, 15, 19). This result is consistent with the observation that coding NS-SNPs account for very few disease associations and suggests that regulation of expression is an important factor in disease development (56). Similarly, according to Melzer et al. (33), of the nine identified SNPs associating with individual serum protein levels, none were found in the coding sequence of the relevant nearest gene (33). Our peQTL discovery efforts, however, resulted in an even split between the association of coding NS-SNPs and non-coding SNPs with protein expression variation. The 2D DIGE approach is highly sensitive to NS-SNPs that cause electrophoretic shifts due to protein pl changes. Thus, the choice of using 2D DIGE, combined with the higher density in SNP genotypes in our study, provided a greater power to detect associating NS-SNPs than did other proteomics methods. To determine how well we were able to detect NS-SNPs in our data, we determined how many identified spots potentially contain NS-SNPs and compared this with the number we actually identified through QT analysis. Using the dbSNP database (NCBI, build 130) as a resource and based on whether the NS-SNP affected a charged or polar amino acid, we predicted that there are 81 proteins in our data that could produce a shift in protein migration and therefore be detectable by 2D DIGE. However, of these NS-SNPs, only 19 were present and variable in our population at a minor allele frequency >10%. We identified four proteins that were associated with NS-SNPs through peQTL analysis. Therefore, the false negative rate was estimated at (19 – 4)/19 = 78%. Considering all possible NS-SNPs, even ones that produce only neutrally charged amino acid shifts, we expected 43 among our 311 identified proteins leading to a false negative rate of 93%. Therefore, although our analysis showed some bias toward NS-SNPs, we were not capable of comprehensively identifying all NS-SNPs in the population.

Collectively, the differences in our results compared with those previously reported may be a reflection of the smaller sample size in our study. In this regard, the smaller sample size (e.g. 16 here versus 60–270 in previous studies) effectively reduced the power to detect more complex regulation due to trans-acting factors and increased bias toward identification of NS-SNPs where the allelic contribution to phenotype is expected to be high.

In summary, because of the generation of high density human SNP maps and genome-wide sequencing efforts in the past decade, many genetic determinants of complex phenotypes and diseases have been identified (50, 54, 57–59). Although there has been much attention paid to the role of genetics in transcript expression variation, very few large scale efforts have attempted the same for protein expression variation. The most successful approach for peQTL discovery is one that allows for both a high degree of technical reproducibility and the opportunity to capture proteomic variation due to alternative splicing and post-translational modifications. Our results demonstrate that 2D DIGE has both the sensitivity and specificity to be used for understanding the genetic underpinnings of proteome variation.

Acknowledgments—We thank James L. Stephenson and David Friedman for critical review of the manuscript and Susanna Cantor for editing.

* This work was supported by RTI International Grants 0270900.059 and 0270800.040 and NIH Grant R21CA135503 to MKB.

This article contains supplemental Figs. 1–5 and Tables 1–5.

These authors contributed equally to this work.

** To whom correspondence should be addressed: Center for Biomarkers and Systems Biology, Research Triangle Inst., 3040 Cornwallis Rd., Research Triangle Park, NC 27709-2194. Tel.: 919-990-8364; E-mail: mbunger@rti.org.

REFERENCES

1. Monks, S. A., Leonardson, A., Zhu, H., Cundiff, P., Pietrusiak, P., Edwards, S., Phillips, J. W., Sachs, A., and Schadt, E. E. (2004) Genetic inheritance of gene expression in human cell lines. Am. J. Hum. Genet. 75, 1094–1105

2. Bergen, A. W., Baccarelli, A., McDaniel, T. K., Kuhn, K., Pfeiffer, R., Kakol, J., Bender, P., Jacobs, K., Pocker, B., Chanock, S. J., and Yeager, M. (2007) cis sequence effects on gene expression. BMC Genomics 8, 296

3. Brem, R. B., and Kruglyak, L. (2005) The landscape of genetic complexity across 6,700 gene expression traits in yeast. Proc. Natl. Acad. Sci. U.S.A. 102, 1572–1577

4. Brem, R. B., Yvert, G., Clinton, R., and Kruglyak, L. (2002) Genetic dissection of transcriptional regulation in budding yeast. Science 296, 752–755

5. Gilad, Y., Ritkin, S. A., and Pritchard, J. K. (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet. 24, 408–415

6. Ronald, J., Brem, R. B., Whittle, J., and Kruglyak, L. (2005) Local regulatory variation in Saccharomyces cerevisiae. PLoS Genet. 1, e25

7. Cervino, A. C., Li, G., Edwards, S., Zhu, J., Laurie, C., Tokiwa, G., Lum, P. Y., Wang, S., Castellani, L. W., Lucis, A. J., Carlson, S., Sachs, A. B., and Schadt, E. E. (2005) Integrating QT and high-density SNP analyses in mice to identify Insig2 as a susceptibility gene for plasma cholesterol levels. Genomics 86, 505–517

8. Davis, R. C., Schadt, E. E., Cervino, A. C., Petferly, M., and Lucis, A. J. (2005) Ultrafine mapping of SNPs from mouse strains C57BL/6J, DBA/2J, and C57BLKS/J for loci contributing to diabetes and atherosclerosis susceptibility. Diabetes 54, 1191–1199

9. Ghazalpour, A., Doss, S., Sheth, S. S., Ingram-Drake, L. A., Schadt, E. E., Lucis, A. J., and Schadt, T. A. (2005) Genomic analysis of metabolic pathway gene expression in mice. Genome Biol. 6, R59

10. Schadt, E. E., Monks, S. A., Drake, T. A., Lucis, A. J., Che, N., Colinavoy, V., Ruff, T. G., Milligan, S. B., Lamb, J. R., Cavet, G., Linsley, P. S., Mao, M., Stoughton, R. B., and Friend, S. H. (2003) Genomics of gene expression surveyed in maize, mouse and man. Nature 422, 297–302

11. Doss, S., Schadt, E. E., Lucis, A. J., and Schadt, T. A. (2005) Cis-acting expression quantitative trait loci in mice. Genome Res. 15, 681–691

12. Emilion, V., Thorleifsson, G., Zhang, B., Leonardson, A. S., Zink, F., Zhu, J., Carlson, S., Heigason, A., Walters, G. B., Gunnarsdottir, S., Mous, M., Steinthorsdottir, V., Eiriksdottir, G. H., Bjornsdottir, G., Reynisdottir, I., Gudbjartsson, D., Helgadottir, A., Jonasdottir, A., Jonasson, T., Styrkarsdottir, U., Gretarsdottir, S., Magnusson, K. P., Stefansson, H., Fossdal, R., Kristjansson, K., Gislason, H. G., Stefansson, T., Leifsson, B. G., Thorsteinsdottir, U., Lamb, J. R., Gulerch, J. R., Retinam, M. L., Kong, A., Schadt, E. E., and Stefansson, K. (2008) Genomics of gene expression and its effect on disease. Nature 452, 423–428

13. Schadt, E. E., Molony, C., Chudin, E., Hao, K., Yang, X., Lun, P. Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., Zhu, J., Millstein, J., Sieberts, S., Lamb, J., GuhaThakurta, D., Derry, J., Storey, J. D., Avila-Campillo, I., Kruger, M. J., Johnson, J. M., Rohli, C. A., van Nas, A., Mehrabian, M., Drake, T. A., Lucis, A. J., Smith, R. C., Guengerich, F. P.,
Strom, S. C., Schuetz, E., Rushmore, T. H., and Ulrich, R. (2008) Mapping the genetic architecture of gene expression in human liver. PLoS Biol. 6, e107

14. Dermitzakis, E. T., and Stranger, B. E. (2006) Genetic variation in human gene expression. Mamm. Genome 17, 503–508

15. Stranger, B. E., Forrest, M. S., Clark, A. G., Minichiello, M. J., Deutsch, S., Lyle, R., Hunt, S., Kahi, B., Antonarakis, S. E., Tavaré, S., Deloukas, P., and Dermitzakis, E. T. (2005) Genome-wide associations of gene expression variation in humans. PLoS Genet. 1, e78

16. Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., Redon, R., Bird, C. P., de Grassi, A., Lee, C., Tyler-Smith, C., Carter, N., Scherer, S. W., Tavaré, S., Deloukas, P., Hurles, M. E., and Dermitzakis, E. T. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Nat. Genet. 39, 1369–1375

17. Klose, J., Nock, C., Hermann, M., Stuhler, K., Marcus, K., Blüggel, M., Kiseo, E., Schlalkyey, L. C., Rastan, S., Brown, S. D., Büssow, K., Himmelbauer, H., and Lehrach, H. (2002) Genetic analysis of the mouse brain proteome. Nat. Genet. 30, 385–393

18. Mann, M. (2006) Functional and quantitative proteomics using SILAC. Nat. Rev. Mol. Cell Biol. 7, 952–958

19. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using iso-topped affinity tags. Nat. Biotechnol. 17, 994–999

20. Benjamini, Y., and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B 57, 289–300

21. Washburn, M. P., Wolters, C., and Yates, J. R. (2001) Large-scale analysis of protein mixtures using chromatographic alignment of high-resolution muLC-MS data. Anal. Chem. 73, 894–905

22. Carvalho, P. C., Jaworski, D., and Yates, J. R. (2003) Identifying differences in protein expression levels by spectral counting and feature selection. Genet. Mol. Res. 7, 342–356

23. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., and Sham, P. C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575

24. Benjamini, Y., and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B 57, 289–300

25. Cheung, V. G., Stratton, M. R., Nyholt, D. R., O'Donovan, M. C., and Sklar, P. (2007) Common genetic variants account for differences in gene expression among ethnic groups. Nat. Genet. 39, 226–231

26. Spielman, R. S., and Cheung, V. G. (2004) Genetic analysis of genome-wide association studies. Nat. Rev. Genet. 5, 450–461

27. Cheung, V. G., Stratton, M. R., Nyholt, D. R., O'Donovan, M. C., and Sklar, P. (2007) Common genetic variants account for differences in gene expression among ethnic groups. Nat. Genet. 39, 226–231

28. Spielman, R. S., Radulovic, D., Shaffer, S. A., Ruderfer, D. M., Bedalov, A., Kingsley, J., Harmsen, E., Labuda, D., Morgan, K., Vohl, M. C., Bandinelli, S., Frayling, T. M., Singleton, A., and Ferrucci, L. (2008) A genome-wide association study identifies protein quantitative trait loci (pQTLs). Plos Genet. 4, e1000729

29. Foss, E. J., Radulovic, D., Shaffer, S. A., Ruderfer, D. M., Bedalov, A., Goodlett, D. R., and Kruglyak, L. (2007) Genetic basis of proteome variation in yeast. Nat. Genet. 39, 1369–1375

30. Man, M. (2006) Functional and quantitative proteomics using SILAC. Nat. Rev. Mol. Cell Biol. 7, 952–958

31. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat. Biotechnol. 17, 994–999

32. Benjamini, Y., and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B 57, 289–300

33. Melzer, D., Perry, J. R., Hernandez, D., Corsi, A. M., Stevens, K., Rafferty, I., Laurenti, F., Murray, A., Gibbs, J. R., Paolissio, G., Rafiq, S., Simon-Sanchez, J., Lango, H., Schozl, S., Weidon, M. N., Arepalli, S., Rice, N., Washecka, N., Hurst, A., Britton, A., Henley, W., van de Leemput, J., Li, R., Newman, A. B., Tranah, G., Harris, T., Panicker, V., Dayan, C., Bennett, A., McCarthy, M. I., Rukkonen, A., Jarvelin, M. R., Guralnik, J., Bandodkar, S., Freyling, T. M., Singleton, A., and Ferrucci, L. (2008) A genome-wide association study identifies protein quantitative trait loci (pQTLs). Plos Genet. 4, e1000729

34. Foss, E. J., Radulovic, D., Shaffer, S. A., Ruderfer, D. M., Bedalov, A., Goodlett, D. R., and Kruglyak, L. (2007) Genetic basis of proteome variation in yeast. Nat. Genet. 39, 1369–1375

35. Klose, J., Nock, C., Herrmann, M., Stuhler, K., Marcus, K., Büggel, M., Kiseo, E., Schlalkyey, L. C., Rastan, S., Brown, S. D., Büssow, K., Himmelbauer, H., and Lehrach, H. (2002) Genetic analysis of the mouse brain proteome. Nat. Genet. 30, 385–393

36. Mann, M. (2006) Functional and quantitative proteomics using SILAC. Nat. Rev. Mol. Cell Biol. 7, 952–958
Study 100K Project: genome-wide associations for blood pressure and arterial stiffness. *BMC Med. Genet.* 8, 53

52. Huang, R. S., Duan, S., Shukla, S. J., Kistner, E. O., Clark, T. A., Chen, T. X., Schwetzler, A. C., Blume, J. E., and Dolan, M. E. (2007) Identification of genetic variants contributing to cisplatin-induced cytotoxicity by use of a genomewide approach. *Am. J. Hum. Genet.* 81, 427–437

53. Shukla, S. J., Duan, S., Badner, J. A., Wu, X., and Dolan, M. E. (2008) Susceptibility loci involved in cisplatin-induced cytotoxicity and apoptosis. *Pharmacogenet. Genomics* 18, 253–262

54. International HapMap Consortium (2003) The International HapMap Project. *Nature* 426, 789–796

55. Khan, Z., Bloom, J. S., Garcia, B. A., Singh, M., and Kruglyak, L. (2009) Protein quantification across hundreds of experimental conditions. *Proc. Natl. Acad. Sci. U.S.A.* 106, 15544–15548

56. Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W. H., Samani, N. J., Todd, J. A., Donnelly, P., Barrett, J. C., Davison, D., Easton, D., Evans, D. M., Leung, H. T., Marchini, J. L., Morris, A. P., Spencer, C. C., Tobin, M. D., Altwood, A. P., Boorman, J. P., Cant, B., Everson, U., Hussey, J. M., Jolley, D. J., Knight, A. S., Meek, E., Nuttall, S., Prowse, C. V., Stevens, H. E., Taylor, N. C., Walters, G. R., Walker, N. M., Watkins, N. A., Winzer, T., Jones, R. W., Mcardle, W. L., Ring, S. M., Strachan, D. P., Pembrey, M., Breen, G., St Clair, D., Caesar, S., Gordon-Smith, K., Jones, L., Fraser, C., Green, E. K., Grozdeva, R., Hamshere, M. L., Holmans, P. A., Jones, I. R., Kirov, G., Moskivina, V., Nikolov, I., O’Donovan, M. C., Owen, M. J., Collier, D. A., Elkin, A., Farmer, A., Williamson, R., Mccaffrey, P., Young, A. H., Ferrier, I. N., Ball, S. G., Balmforth, A. J., Barrett, J. R., Bishop, T. D., Iles, M. M., Maqbool, A., Yuldasheva, N., Hall, A. S., Braund, P. S., Dixon, R. J., Mangino, M., Stevens, S., Thompson, J. R., Bredin, F., Tremelling, M., Parkes, M., Drummond, H., Lees, C. W., Nimmo, E. R., Satsangi, J., Fisher, S. A., Forbes, A., Lewis, C. M., Omnie, C. M., Prescott, N. J., Sanderson, J., Matthew, C. G., Barbour, J., Moliuddin, M. K., Todhunter, C. E., Mansfield, J. G., Ahmad, T., Cummings, F. R., Jewell, D. P., Webster, J., Brown, M. J., Lathrop, M. G., Collin, J., Dominiczak, A., Marcano, C. A., Burke, B., Dobson, R., Gungadoo, J., Lee, K. L., Munroe, P. B., Newhouse, S. J., Onipinla, A., Wallace, C., Xue, M., Caulfield, M., Farrall, M., Barton, A., Bruce, I. N., Donovan, H., Eyer, S., Gilbert, P. D., Hilder, S. L., Hinks, A. M., John, S. L., Potter, C., Silman, A. J., Symmons, D. P., Thomson, W., Worthington, J., Dunger, D. B., Widmer, B., Frayling, T. M., Freathy, R. M., Lango, H., Perry, J. R., Shields, B. M., Weedon, M. N., Hattersley, A. T., Hitman, G. A., Walker, M., Elliott, K. S., Groves, C. J., Lindgren, C. M., Rayner, N. W., Timpson, N. J., Zeggini, E., Newton, M., Slrugo, G., Lyons, E., Vannenberg, F., Hill, A. V., Bradbury, L. A., Farrar, C., Pointon, J. J., Wordsworth, P., Brown, M. A., Franklin, J. A., Heward, J. M., Simmonds, M. J., Gough, S. C., Seal, S., Stratton, M. R., Rahman, N., Ban, M., Goris, A., Sawyer, S. J., Compston, A., Conway, D., Jallow, M., Pembrey, M., Sirugo, G., Rockett, K. A., Bumpstead, S. J., Chaney, A., Downes, K., Ghori, M. J., Gwilliam, R., Hunt, S. E., Inouye, M., Keniry, A., King, E., Mcginnis, R., Potter, S., Ravanhazaj, R., Whitaker, P., Widdden, C., Withers, D., Cardin, N. J., Davison, D., Ferreira, T., Pereira-Gale, J., Hallgrimsson, I. B., Howie, B. N., Su, Z., Teo, Y. Y., Vukcevic, D., Bentley, D., Brown, M. A., Compston, A., Farrall, M., Hall, A. S., Hattersley, A. T., Hill, A. V., Parkes, M., Pembrey, M., Stratton, M. R., Mitchell, S. L., Newby, P. R., Brand, O. J., Carr-Smith, J., Pearce, S. H., Mcginnis, R., Keniry, A., Deloukas, P., Reville, J. D., Zhou, X., Sims, A. M., Dowling, A., Taylor, J., Doan, T., Davis, J. C., Savage, L., Ward, M. M., Leach, T. L., Weissman, M. H., and Brown, M. (2007) Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nat. Genet.* 39, 1329–1337

57. Cargill, M., Altschuler, D., Ireland, J., Sklar, P., Airdle, K., Patil, N., Wallace, C., Lane, C. R., Lim, E. P., Kalyanaraman, N., Nemes, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G. Q., and Lander, E. S. (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* 22, 231–238

58. Collins, F. S., Lander, E. S., Rogers, J., and Waterston, R. H. (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431, 931–945

59. Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z., Rogers, J., Bentley, D. R., Kwok, P. Y., Mardis, E. R., Yeh, R. T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fultan, L., Hillier, L., Waterston, R. H., McPherson, J. D., Gilman, B., Schaaffner, S., Van Etten, W. J., Reich, D., Higgins, J., Daly, M. J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M. C., Linton, L., Lander, E. S., and Altschuler, D. (2001) A map of human genome sequence variation containing 1.4 million single nucleotide polymorphisms. *Nature* 409, 928–933