nature genetics

Genetic basis of proteome variation in yeast

Eric J Foss[1,5], Dragan Radulovic[2,5], Scott A Shaffer[3], Douglas M Ruderfer[4], Antonio Bedalov[1], David R Goodlett[3] & Leonid Kruglyak[4]

Proper regulation of protein levels is essential for health, and abnormal levels of proteins are hallmarks of many diseases. A number of studies have recently shown that messenger RNA levels vary among individuals of a species and that genetic linkage analysis can be used to identify quantitative trait loci that influence these levels. By contrast, little is known about the genetic basis of variation in protein levels in genetically diverse populations, in large part because techniques for large-scale measurements of protein abundance lag far behind those for measuring transcript abundance. Here we describe a label-free, mass spectrometry–based approach to measuring protein levels in total unfractionated cellular proteins, and we apply this approach to elucidate the genetic basis of variation in protein abundance in a cross between two diverse strains of yeast. Loci that influenced protein abundance differed from those that influenced transcript levels, emphasizing the importance of direct analysis of the proteome.

A number of recent studies have shown that transcript levels vary among individuals of a species, and that genetic linkage analysis can be used to identify quantitative trait loci that influence the transcript levels of individual genes and groups of genes[1]. In some cases, genetic differences in transcript levels have been shown to be linked to phenotype or associated with disease[2–5], but in general the functional significance of genetic variation in transcript abundance remains unknown. Much of the work of the cell is performed by proteins, and therefore functionally important changes in transcript levels are expected to be reflected in changes in the levels of corresponding proteins. However, various mechanisms of post-transcriptional regulation can either buffer changes in transcript abundance so that they do not lead to changes in protein abundance or lead to changes in protein abundance in the absence of a corresponding effect on transcripts, as reflected in the weak correlation between transcript and protein levels[6] (however, see also ref. 7). Thus, as a more immediate readout of cellular physiology, direct examination of the proteome is expected to provide biological insights and disease biomarkers that cannot be captured through evaluation of the transcriptome alone. Previous studies of the genetic control of protein levels have been largely qualitative owing to limitations of existing techniques, specifically two-dimensional gels[8,9]. Furthermore, these studies did not include measurements of transcripts to allow the genetics of protein and transcript level variation to be compared in the same population.

Proteome profiling based on mass spectrometry holds great promise for the quantitative measurement of protein abundance[10,11]. In a proteomic experiment, output from a mass spectrometer can be represented as a matrix of peaks (**Fig. 1a**), each of which represents a peptide that is defined by a specific elution time, relative ion intensity value and mass-to-charge ratio. These matrices allow direct quantitative proteomic comparisons because the relative ion intensities of a given peptide in two samples reflect the relative abundance of their corresponding proteins. In theory, one should be able to compare the levels of peptides in complex mixtures simply by analyzing each sample by liquid chromatography-tandem mass spectrometry (LC-MS/MS) and comparing the ion intensities of the peptides of interest from hundreds of different matrices, just as transcript levels are compared across samples by measurements of corresponding spot hybridization intensities on microarrays. Because elution times can differ from experiment to experiment in a nonlinear fashion, however, alignment of the corresponding peptides presents an important challenge to this approach. Alignment of the matrices allows one not only to quantify large numbers of peptides across many datasets but also to increase peptide sequence identifications. Sequencing of peptides is inefficient, and therefore many peptides are sequenced in only a small percentage of the total datasets. However, alignment allows a single sequence identity to be translated across hundreds of datasets. Methods for aligning peptides that are based on labeling two peptide mixtures with different stable isotope tags and then analyzing the combined mixtures in a single LC-MS/MS experiment have been developed[12–14], but these methods, although well suited for pairwise comparisons, have proved inadequate for proteome profiling studies involving large numbers of samples because they do not solve the problem of aligning peptide identities across many matrices. On the other hand, methods for relative quantification that are based on mathematical alignment[15–18], although appealing in principle, have been difficult to put into practice for large-scale studies

[1]Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA. [2]Florida Atlantic University, Boca Raton, Florida 33431, USA. [3]University of Washington, Seattle, Washington 98195, USA. [4]Lewis-Sigler Institute for Integrative Genomics and Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544, USA. [5]These authors contributed equally to this work. Correspondence should be addressed to L.K. (leonid@genomics.princeton.edu) or D.R. (radulovi@fau.edu).

Received 31 May; accepted 20 September; published online 21 October 2007; doi:10.1038/ng.2007.22

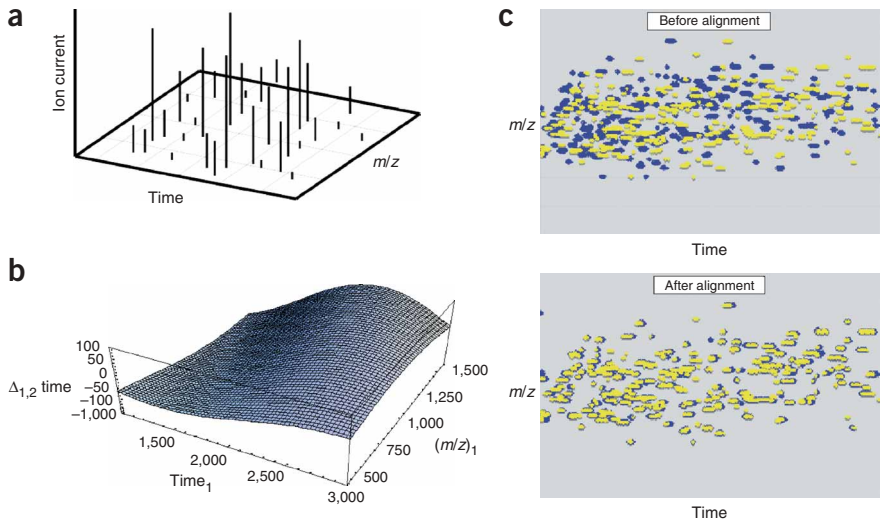

Figure 1 Schematic of method for alignment of MS matrices. (**a**) Mass spectrometric analysis of a complex mixture can be represented as a collection of peaks, each of which corresponds to a peptide that is defined by a specific elution time and mass-to-charge (m/z) ratio. (**b**) Graph shows the amount of time ($\Delta_{1,2}$ time), as a function of the elution times ($time_1$) and mass to charges (m/z_1) of the peptides in sample 1, that must be added to the elution time of peptides in sample 2 to make the two sets of peptides superimposable. (**c**) Two-dimensional representations of the LC-MS data obtained by taking a 'slice' at a single level of relative ion intensity. The blue and yellow spots represent peptides from different LC-MS analyses that are approximately, but not exactly, superimposable. To align LC-MS data, a given amount of time, which depends on both elution time and m/z, must be added to each yellow spot.

because computational requirements rise quadratically with increasing numbers of samples.

Building on our previous algorithm[19], which calculates a continuous function to determine how much time must be added to or subtracted from the elution times for peaks in two complex mixtures to make them superimposable (**Fig. 1b,c**), we have now developed a robust and accurate computational method for aligning MS matrices and correlating peptides in hundreds of LC-MS datasets. We used graph theory to overcome a mathematical challenge that arises when the number of samples to be aligned exceeds approximately 20 (see **Supplementary Methods** and **Supplementary Fig. 1** online). Notably, this algorithm has modest computational requirements: using only a typical laptop computer, we were able to align 408 LC-MS/MS datasets of total unfractionated yeast proteins. This allowed us to map more than 100 loci that regulate protein levels and thereby to gain insight into the genetic basis of proteome diversity in yeast on a global scale. Furthermore, because the genetic basis of transcript abundance had been previously analyzed in the same population, our study enabled us to compare directly the genetic circuitry that underlies proteome and transcriptome diversity in an outbred population. It is important to note that, even if we had quantified these peptides using stable isotopes rather than our label-free quantification, a large-scale mapping study such as this would not have been possible without the alignment algorithm, as discussed above.

RESULTS

To analyze the proteomes in a genetically diverse population, we used a cross between a laboratory strain of yeast, BY4716 (ref. 20), and a vineyard isolate, RM11-1a (ref. 21). These strains have both been sequenced, and they differ at ~0.6% of base pairs[22]. Furthermore, these strains and more than 100 segregants from a cross between them have been densely genotyped and extensively studied with regard to the genetic basis of variation in transcript levels[5,23,24]. We chose a direct 'lyse and go' approach for the proteome analysis that circumvented proteome pre-fractionation. The simplicity of this approach allowed us to analyze more than 400 samples of total cellular proteins by shotgun proteomics. We isolated total proteins from eight independent logarithmic-phase cultures of each parent and from two independent cultures of each of 98 segregants, digested them with trypsin, and analyzed the resulting mixture directly by LC-MS/MS on a linear ion trap (LTQ)–Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometer. Each parental protein preparation was analyzed once (eight replicates per parent) and each segregant preparation was analyzed twice (four replicates per segregant). We identified 6,898 peptides, derived from 1,693 proteins, among which we quantified 1,873 peptides corresponding to 569 proteins. These 569 proteins ranged from very low to very high abundance, as determined by tagging with green fluorescent protein (GFP) and/or by tandem affinity purification (TAP)[25], although there was a clear bias toward high-abundance proteins. Peptide ionization efficiencies vary widely between peptides, and our ability to quantify even low-abundance proteins might be partly attributable to each of those proteins having at least one peptide that ionizes particularly efficiently. Of the proteins we quantified, 225 were involved in protein biosynthesis, 81 in energy metabolism, 41 in cellular structure, 32 in response to stress, 35 in transport, 18 in RNA metabolism and 13 in DNA metabolism; 124 either had other functions or had unknown functions[26].

Genetic differences in protein abundance

We next looked for differences in protein abundance between the two parents. Of 1,010 quantified peptides from 376 proteins that were well measured in both parents, 196 peptides from 137 proteins had significant differences in intensity ($P < 0.005$). If these proteins constitute a representative sample of the proteome, more than one-third of the proteome differs between the two parents. To confirm that the differences in peptide intensity between the parental strains reflect the differences in the abundance of the corresponding proteins, we compared the mass spectrometric measurements of the peptides with the protein levels measured by western blots (**Fig. 2**). We introduced a triple hemagglutinin (HA) tag at the C termini of nine proteins whose peptides were found to be different between the strains by integrating a tag into the corresponding sites in the genome of each strain[27]. These proteins were chosen randomly from the set of 137 proteins that were determined to be different between the two parents with a probability $P < 0.005$. Measurement of protein levels using western blots unequivocally confirmed the differences in protein abundance for eight of nine measured proteins and correlated well with the mass spectrometric measurements. For one protein, 6-phosphogluconate dehydrogenase (Gnd1), the difference between laboratory and vineyard strain measured by western blotting was significantly less than expected from mass spectrometric measurements. The introduction of the tag, which changes the amino acid composition of the

protein and alters the 3′ untranslated region, could account for this single discrepancy, as might noise in the mass spectrometric measurements or a combination of the two. These results show that our technology can identify and accurately quantify differences in protein abundance (also see **Supplementary Fig. 2** online and **Supplementary Methods**).

Protein levels varied continuously among the segregants, showing that protein levels, like transcript levels, behave as quantitative traits. To study the genetic basis of variation in protein levels, we focused on those proteins that were quantified in at least 40 segregants and that, within each segregant, were quantified in at least three of the four replicate analyses. We also selected only the best peptide for a given protein, yielding 221 unique peptides with high-quality data. These 221 peptides correspond to 278 proteins. (There are more proteins than peptides because 57 peptides arise from ribosomal genes that are present in two copies.) We estimated the genetic contribution to the observed variability in protein abundance for a subset of 156 of these proteins for which high-quality data from the parent strains was also available. This contribution, as measured by heritability, averaged 62%, implying that variation in abundance among the segregants is primarily due to genetic differences rather than to stochastic variability among biological replicates or to measurement noise.

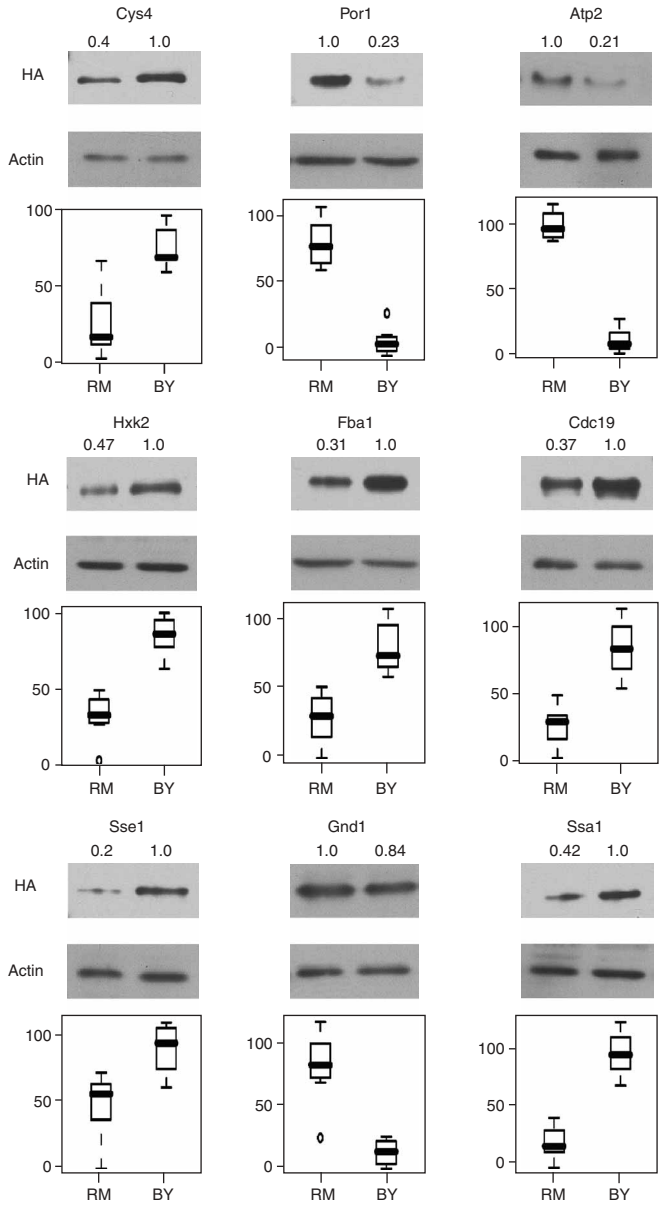

Figure 2 Protein quantifications by western blotting and by MS measurements are comparable. Cys4, Por1, Atp2, Hxk2, Fba1, Cdc19, Sse1, Gnd1 and Ssa1 were tagged with three HA epitopes[27] and analyzed by western blotting with anti-HA and anti-actin antibodies. Loading was normalized according to Ponceau staining (data not shown) and actin. The bottom panels show the median (black line) surrounded by one s.d. (rectangle) and the range (dashed line) for ion current quantification of eight mass spectrometric measurements of the untagged versions of the proteins.

Linkage analysis of protein abundance

We next looked for linkage between protein levels in the segregants and 2,951 genetic markers for which these strains have been genotyped[23]. Linkages were identified with the widely used R-QTL software package[28]. At a false-discovery rate (FDR) of 0.042, corresponding to a lod score ≥ 5.1, we detected 24 linkages, with a single expected false positive. For further analyses, we chose to look at a larger number of more liberal linkages, in order to be able to look for hot spots and carry out other analyses. The abundance of 85 proteins mapped to at least one locus with a LOD score of 3 (at this threshold, 19 linkages are expected by chance on the basis of permutation tests). We identified 109 loci when we counted multiple loci per protein. An example of a mapping result is shown in **Figure 3a**. These mapping results provide further validation of protein quantification because highly 'noisy' data would not permit detection of linkage[29].

Among the 221 peptides used for mapping, 37 differed between the parents at $P < 0.005$. Twenty-three of these 37 peptides (62%) showed linkage to at least one locus. Because this cross has high power to map parental differences that are due to single loci, the lack of linkage for 38% of peptides with differences in abundance between the parents indicates considerable complexity in the underlying genetics, as observed for transcript abundance[23,24]. An additional 39 peptides were not called different between the parents at $P < 0.005$ but showed linkage in the cross. This result is probably a combination of false-positive linkages (19 are expected), false negatives in the test for differences between the parent strains, and transgressive segregation. Transgressive segregation is a common phenomenon in which each parent has alleles that exert opposing effects on a trait; these opposing effects can leave the parents appearing to be identical for the trait, but large differences can arise in the progeny when new combinations of the opposing alleles unmask their effects. Most heritable transcript levels showed transgressive segregation in this cross[23]. An example of transgressive segregation is shown in **Figure 3b**.

To explore the genetic regulation of protein levels further, we determined whether the loci that affect protein abundance mapped near the genes that encodes the corresponding proteins (local linkage) or elsewhere in the genome (distant linkage). Local linkage is consistent with a genetic change in a regulatory region of the structural gene or in the coding sequence that affects the stability or regulation of the protein. To test for local linkage, we measured the distances between the genetic marker with the highest linkage score and the gene that encodes the affected protein. Only 7% (6/85) of the peptides with linkage linked to a marker within 20 kb of the encoding gene, and the majority of the linkages were to a locus on a different chromosome (**Supplementary Table 1** online). Polymorphisms that result in amino acid differences are expected to manifest themselves as locally linked proteins because they will be absent in those segregants

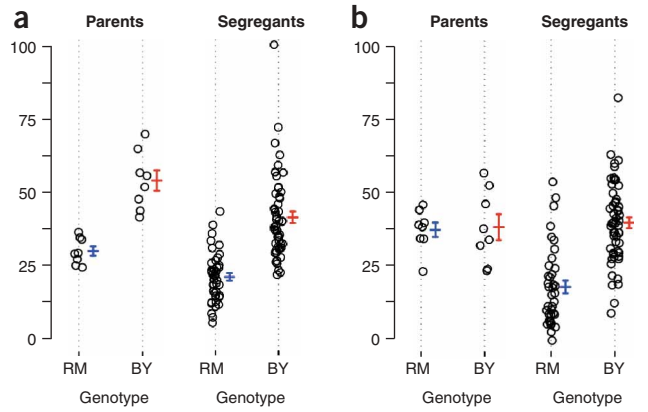

Figure 3 Peptide levels in parents and segregants. (**a**) An example of linkage between the levels of a peptide and the inheritance of a genetic locus. The levels of peptide AGGECITLDQLAVR, derived from Rpl18A, are shown for both parents and then for two groups of segregants that were divided on the basis of inheritance of a specific genetic marker. Those progeny that inherited the marker from RM11-1a had lower abundance of the peptide than those that inherited the marker from BY4716, consistent with the differences between the parents. (**b**) An example of transgressive segregation. Levels of peptide NILAESNSSLDNIVK, derived from Mmf1, are shown for both parents and then for two groups of segregants that were divided on the basis of inheritance of a specific genetic marker. This linkage was identified despite the lack of difference between the parents. The *y* axes in both panels show relative peptide abundance in arbitrary units, scaled from 0 (minimum) to 100 (maximum).

that inherit the polymorphism (that is, this apparent local linkage is just an artifact of the technique and not a biologically meaningful result). Furthermore, because the amino acid composition of the two forms of the protein will be different, it is not possible to use ion intensities for relative quantification of the two forms. As expected, we saw polymorphic peptides that showed local linkage. However, because apparent local linkage due to such polymorphisms does not reflect genetic control of differences in the abundance of the corresponding protein, such results were not taken into account in calculating the numbers cited above.

Trans hot spots of linkage

Loci that affect protein levels were not evenly distributed throughout the genome; instead, we found four hot spots that were significantly enriched for the number of linkages (**Fig. 4**). The 'hot spot' bins with five or more compound linkages are 'hot' at $P < 0.001$ for observing any such bins, as computed on the basis of a Poisson distribution with a Bonferroni correction for 611 bins. These hot spots on chromosomes 2, 3, 4 and 13 affected the abundance of 8, 35, 6 and 25 proteins, respectively. The locus that affected the largest number of proteins mapped to *LEU2*, a gene involved in leucine biosynthesis that had been deleted in one of the two parents, and 9 of the 35 proteins whose abundance linked to this locus were involved in amino acid biosynthesis. The amino acid biosynthesis regulon can be activated by the lack of a single amino acid[30]. All strains in this experiment were grown in synthetic medium that, although supplemented with amino acids, is limiting for amino acids. Lower levels of leucine in strains in which *LEU2* is deleted are expected to activate the amino acid biosynthesis regulon. A growth advantage phenotype has been mapped to *LEU2* when these strains were grown on standard synthetic medium, whereas doubling the amount of leucine in the medium abolished this growth advantage phenotype (E. Smith, unpublished data). In addition to the enrichment for amino acid biosynthesis genes among those regulated by the *LEU2* hot spot, we saw enrichment (11/25) of such genes linked to the hot spot on chromosome 13. We saw no other evidence of enrichment of specific functions for the genes regulated by a single hot spot.

Comparison of variation in proteome and transcriptome

We found both similarities and intriguing differences between genetic regulation of proteins and transcripts. The average correlation between transcript levels and protein levels was 0.186, which is comparable to published results[6] (**Supplementary Fig. 3** online). The extent of differences in abundance between the parental strains was comparable for the transcriptome and the proteome (around one-third in both cases), but only 59 of 137 proteins (43%) that differed in abundance between the parents corresponded to transcripts that were different between the parents[21]. This overlap is significantly greater than what would be expected by chance ($P < 0.001$), showing that differentially expressed proteins are more likely to correspond to differentially expressed transcripts. The remaining 78 cases are probably a combination of false negatives in the transcript data and real instances where proteins differ and transcripts do not. As observed with transcripts, differences in protein abundance were heritable and showed evidence of complex segregation.

Linkage analysis that was restricted to the set of genes and segregants for which we had measurements of both transcripts and proteins detected loci for 156 of 278 transcripts (56%) compared to 85 of 221 peptides (38%). Most loci influenced either peptide abundance or transcript abundance but not both, and most of the peptide linkages remained significant even when the relative abundances of the corresponding transcripts were included as covariates in the linkage analysis (**Supplementary Note** online). The set of all transcripts showed a higher frequency of local linkage than was observed for peptides, but the subset of transcripts with measured peptide abundance did not (9 of 156 transcripts versus 6 of 85 peptides). This subset is located in regions of the genome with lower polymorphism rates (**Supplementary Note**), which reduces the frequency of local linkages[31]. Whether local regulatory variation is less important for protein abundance than for transcript abundance overall remains to be determined.

A common feature in the genetics of both transcripts and proteins is the existence of linkage hot spots: a small number of loci, each of which affects the abundance of a disproportionately large number of transcripts or proteins. Although the locations of three of the four protein hot spots also represent hot spots for transcriptional regulation (**Fig. 4**), there are several notable differences between the two sets of hot spots. First, one hot spot (on chromosome 4) is seen only at the protein level, implying that some polymorphisms exert their effects without altering transcript abundance and can only be discovered through direct measurement of proteins. (Although the mapping confidence interval is too large to attribute this hot spot to a particular gene, we note that there are two genes in the region, *RPL27B* and *MRPL28*, that are involved in translation.) Similarly, many hot spots observed at the transcript level show no effect on the levels of proteins, presumably reflecting buffering of protein abundance by post-transcriptional regulatory mechanisms. For example, there are ten transcripts whose levels link to *HAP1*, a transcriptional activator on chromosome 12 that contains a functional polymorphism with effects on gene expression[21], but there is no corresponding hot spot for protein levels.

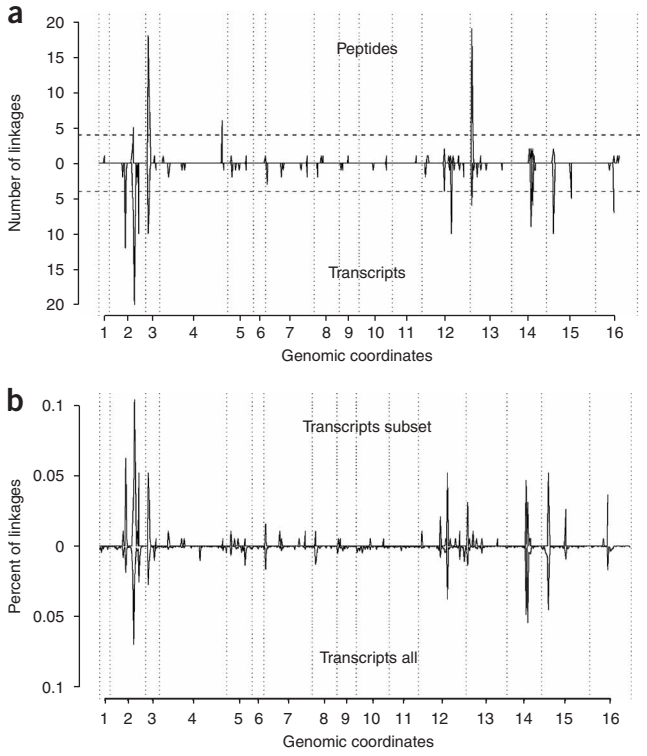

Figure 4 Linkage hot spots plotted against genome location. The genome was divided into 20-kb bins and arranged in chromosomal order. Vertical dotted lines separate the chromosomes and horizontal dotted lines indicate the cutoff for what is considered a hot spot. (**a**) Hot spots of protein and transcript regulation. The number of transcripts and peptides regulated by different areas of the genome is shown with peptides on the top and transcripts on the bottom. Only transcripts for which we have measurements for the corresponding protein are included. (**b**) The fraction of total linkages identified by that subset of transcripts for which we have corresponding protein measurements is shown at the top and the fraction for total transcripts is shown at the bottom.

Second, even when the location of the hot spots is shared (on chromosomes 2, 3 and 13), both the number and the identity of the proteins that are affected by these hot spots can be markedly different than those for the transcripts. For example, although 8 proteins are affected by a locus on chromosome 2 that also affects 42 transcripts, only 1 of the 8 proteins is represented among the 42 transcripts. The overlap between transcripts and proteins is more substantial but nonetheless incomplete for the hot spots on chromosomes 3 and 13, with 5 (of 35 linked proteins and 18 linked transcripts) and 4 (of 25 linked proteins and 7 linked transcripts) genes affected by the locus at both the protein and transcript levels, respectively. The fact that the same locus can affect a different set of genes at the protein and transcript levels implies either that the polymorphism responsible for the difference in transcript levels is not the same as that responsible for the difference in protein levels, which is unlikely because of the precise overlap of the hot spots, or that the same genetic change manifests itself differently at the transcriptome and proteome levels. It is also possible that the two detection methods (microarrays versus mass spectrometry) have different biases for the changes they can most easily detect, although this is less likely in the high-quality subset of the data used for linkage analysis. (For further comparisons of proteome and transcriptome data, see **Supplementary Note**.)

Finally, the protein linkages are concentrated in fewer hot spots than the transcript linkages, indicating that fewer polymorphisms have large effects on protein abundance than on transcript abundance. This suggests that any given polymorphism is more likely to affect the transcriptome than the proteome.

DISCUSSION

It is notable that analysis of a subset of only 278 transcripts identified virtually all of the hot spots that are observed when all transcripts are analyzed (**Fig. 4b**). This finding can be explained with the observation that relevant genetic differences most often change not one but rather several transcripts, hot spots being the most extreme example. Our study of proteome variation in which a single locus often controls the levels of many proteins (hot spots) indicates that the proteome and the transcriptome behave similarly in this regard. Furthermore, this implies that despite testing linkage for only 221 peptides, we might have identified most or all of the polymorphisms that have important effects on protein abundance segregating in this cross. This finding might have implications for the application of global proteome profiling in other biological systems, including the identification of disease biomarkers. Because it is likely that disease processes, like genetic changes at hot spots, alter several proteins, even measurements that include a relatively limited fraction of the proteome are likely to provide global insight into proteome perturbation and to identify relevant disease biomarkers.

We have described a label-free, mass spectrometry–based approach to the measurement of protein levels in total unfractionated cellular proteins. We have validated the quantitative accuracy of this approach through several independent lines of evidence. First, we confirmed the relative abundance of several proteins by western blotting. Second, we found reproducible differences in peptide abundance between the parent strains, demonstrated high heritability of abundance, and mapped with high confidence specific loci that affect abundance; none of this would be possible without accurate quantification. Some of the mapped loci are further validated by their precise coincidence with loci that affect transcript abundance, which could not have occurred by chance ($P < 0.0001$ for coincidence among hot spot locations). Although we are currently examining only a small fraction of the proteome, our approach opens the door to myriad proteome profiling experiments, as was the case in transcriptome profiling with the introduction of microarrays that examined only a fraction of the transcriptome. Our results provide a first look at the genetic circuitry that underlies both proteome and transcriptome diversity in an outbred population.

METHODS

Protein isolation. Yeast were grown in synthetic complete medium to mid log phase, washed, and lysed with 10% trichloroacetic acid. The pellet was washed twice with 90% trichloroacetic acid and proteins were denatured by incubating for 30 min at 56° in 8 M urea and 10 mM DTT. Cysteines are alkylated by incubating for 30 min with 15 mM iodoacetamide. Volume was then increased eightfold with 50 mM NH_4HCO_3 to dilute urea and DTT and then incubated overnight at 37° with trypsin and 0.5 mM $CaCl_2$. The reaction was stopped by addition of 3 μl of glacial acetic acid and peptides were desalted and purified on a disposable C_{18} column.

Protein tagging and western blotting. Protein tagging with HA was performed as described[27]. Western blotting was performed using standard laboratory procedures.

Liquid chromatography and mass spectrometry. Peptide digests were analyzed by electrospray ionization in the positive ion mode on a hybrid linear ion

trap–Fourier transform–ion cyclotron resonance mass spectrometer (LTQ-FT; Thermo Electron Corp.). Nanoflow HPLC was performed using a Michrom Bioresources Paradigm MS4B MDLC coupled to a Michrom Paradigm Endurance autosampler. Peptides were trapped and captured using a 300 μm i.d. × 50 mm long precolumn (Michrom) packed with 200 Å (5-mm Magic C18 particles). Peptides were separated on a 100 μm i.d. × 200 mm long analytical column (Michrom) packed with 100 Å (5 μm C18AQ particles). Electrospray ionization was achieved using a 50-μm-i.d. tapered stainless steel capillary (Michrom) located immediately after the analytical column. The voltage was applied through a stainless steel micro-tee union between the precolumn and the analytical column. With an injection volume of 10 μl, peptides were loaded on the precolumn at ~30 μl min^{-1} in H_2O/CH_3CN (95/05) with 0.1% (v/v) formic acid. Peptides were eluted using an acetonitrile gradient flowing at ~500 nl min^{-1} using mobile phase consisting of A, H_2O; B, CH_3CN; and C, 1.0% (v/v) formic acid. The gradient program was 0–5 min, A (85%), B (5%), C (10%); 60 min, A (55%), B (35%), C (10%); 65–74 min, A (10%), B (80%), C (10%); 75 min, A (85%), B (5%), C (10%); 90 min (stop). Ion source conditions were optimized using a tuning solution composed of caffeine (Sigma), MetArgPheAla (MRFA; Bachem) and Ultramark 1621 (Lancaster Synthesis). Injection waveforms for the LTQ-FT linear ion trap and ICR cell were kept on for all acquisitions. For MS, ICR resolution was set to 100,000 (*m/z* 400) and ICR ion populations were held at 1×10^6 through the use of automatic gain control (AGC). For MS/MS in the linear ion trap, the ion population was set to 1×10^4, the precursor isolation width was set to 2 Da, and the collision energy set to 35%. Data was acquired using an MS 'survey' scan in the ICR followed by MS/MS data-dependent selection of the three most abundant precursors from the survey scan in the linear ion trap. Singly charged ions were excluded from data-dependent analysis. Data redundancy was minimized by excluding previously selected precursor ions (–0.1 Da/+1.1 Da) for 120 s after their selection for MS/MS. Data were acquired using Xcalibur, version 1.4 (Thermo).

Peptide sequencing. Peptide mixtures derived from protein digests were pumped through a chromatography column packed with reverse-phase beads and electrosprayed into an online MS/MS instrument. The instrument was set to record the fragmentation pattern of an individual peptide subject to collision-induced dissociation three times for each scan of unfragmented total precursor ions. Database search and comparison of the fragmentation patterns were then used for peptide sequencing using SEQUEST[32]. PeptideProphet software was used to score the SEQUEST hits[33]. In this study we used only scores that were ≥0.99.

Protein quantification. Details of the algorithms and software are available in the **Supplementary Methods**. Original Fortran code is available upon request from D.R.

Peptide abundance comparison between parent strains. Data for the parents consisted of 1,287 aligned peaks, which corresponded to 1,010 unique peptides and 376 unique proteins. Each of the two parents (BY and RM) had 0–8 replicate abundance values for a given peptide. We performed two tailed *t*-tests on all peaks for which there were at least four replicate values for each parent.

Analysis of peptides in the segregants. Data for the segregants consisted of 3,451 aligned peaks, which corresponded to 1,744 unique peptides and 547 unique proteins. We selected the best peak for each unique protein from among the peaks that had at least 3 of 4 replicate abundance measurements for at least 40 segregants. We needed at least 3 out of 4 replicates to compensate for the high signal-to-noise ratio inherent to MS analysis, and we needed at least 40 segregants to achieve significant mapping results[21]. As proteins are usually represented by more than one peptide, we chose the best peak (peptide) on the basis of sequence coverage. After this filtering, a set of 221 unique peptides remained, which corresponded to 278 unique proteins, as 57 peptides exactly matched two proteins each. (These peptides with two matches all came from ribosomal proteins that are present in two nearly identical copies.) Heritabilities were computed for the set of 156 peptides with both segregant and parental data as described[21].

Linkage analysis. We performed linkage analysis on the 221 peptides using the software package R/qtl. Linkage was performed using the scanone function with default settings. Significance was determined by empirical permutation tests, with ten permutations of the entire dataset. For local linkages, we treated duplicate protein matches by testing whether the locus was located near either gene; none of the six observed local linkages involved duplicates.

Comparisons with transcripts. We performed linkage analysis on a previous set of expression data[23] for the 94 segregants that overlapped with the segregants used in this study. Correlation between peptide abundance and transcript abundance was computed for the set of 164 peptides that exactly matched only one protein each. For linkage comparisons, we used transcripts that corresponded to each of 278 unique proteins: that is, we included the transcripts for both proteins in the cases of peptides that exactly matched two proteins.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank J. Akey, R. Brem, A. de la Cruz, D. Roberts, J. Ronald and E. Smith for helpful discussions; J. Kim, S. Ryu and G. Taylor for technical assistance; and S. Ryu and E. Smith for sharing unpublished results. This work was supported by the Howard Hughes Medical Institute, by National Center for Research Resources grant 1S10RR17262-01 for purchase of the LTQ-FT (to D.R.G.), by National Institute of Allergy and Infectious Disease grant 1U54 AI57141-01 for Mass Spectrometry Core for the WWAMI Regional Center of Excellence for Biodefense and Emerging Infectious Diseases (to D.R.G.), by National Institute of Environmental Health Science (NIEHS) grant P30ES07033 for the University of Washington NIEHS-sponsored Center for Ecogenetics and Environmental Health (to D.R.G.), by National Cancer Institute grant CA015704 (to A.B.), by National Institute of Mental Health grant R37 MH059520 and a James S. McDonnell Foundation Centennial Fellowship (to L.K.) and by Center grant P50GM071508 from the National Institute of General Medical Science to the Lewis-Sigler Institute.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Rockman, M.V. & Kruglyak, L. Genetics of global gene expression. *Nat. Rev. Genet.* **7**, 862–872 (2006).
2. Mehrabian, M. *et al.* Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. *Nat. Genet.* **37**, 1224–1233 (2005).
3. Schadt, E.E. *et al.* An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* **37**, 710–717 (2005).
4. Schadt, E.E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302 (2003).
5. Yvert, G. *et al.* Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nat. Genet.* **35**, 57–64 (2003).
6. Gygi, S.P., Rochon, Y., Franza, B.R. & Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **19**, 1720–1730 (1999).
7. Lu, P., Vogel, C., Wang, R., Yao, X. & Marcotte, E.M. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* **25**, 117–124 (2007).
8. Damerval, C., Maurice, A., Josse, J.M. & de Vienne, D. Quantitative trait loci underlying gene product variation: a novel perspective for analyzing regulation of genome expression. *Genetics* **137**, 289–301 (1994).
9. Klose, J. *et al.* Genetic analysis of the mouse brain proteome. *Nat. Genet.* **30**, 385–393 (2002).
10. Domon, B. & Aebersold, R. Mass spectrometry and protein analysis. *Science* **312**, 212–217 (2006).
11. Ong, S.E. & Mann, M. Mass spectrometry-based proteomics turns quantitative. *Nat. Chem. Biol.* **1**, 252–262 (2005).
12. DeSouza, L. *et al.* Search for cancer markers from endometrial tissues using differentially labeled tags iTRAQ and cICAT with multidimensional liquid chromatography and tandem mass spectrometry. *J. Proteome Res.* **4**, 377–386 (2005).
13. Gygi, S.P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17**, 994–999 (1999).
14. Ong, S.E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1**, 376–386 (2002).
15. Bellew, M. *et al.* A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics* **22**, 1902–1909 (2006).

16. Fischer, B. *et al.* Semi-supervised LC/MS alignment for differential proteomics. *Bioinformatics* **22**, e132–e140 (2006).
17. Wang, P. *et al.* A statistical method for chromatographic alignment of LC-MS data. *Biostatistics* **8**, 357–367 (2007).
18. Wang, W. *et al.* Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal. Chem.* **75**, 4818–4826 (2003).
19. Radulovic, D. *et al.* Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **3**, 984–997 (2004).
20. Brachmann, C.B. *et al.* Designer deletion strains derived from *Saccharomyces cerevisiae* S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast* **14**, 115–132 (1998).
21. Brem, R.B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752–755 (2002).
22. Ruderfer, D.M., Pratt, S.C., Seidel, H.S. & Kruglyak, L. Population genomic analysis of outcrossing and recombination in yeast. *Nat. Genet.* **38**, 1077–1081 (2006).
23. Brem, R.B. & Kruglyak, L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl. Acad. Sci. USA* **102**, 1572–1577 (2005).
24. Brem, R.B., Storey, J.D., Whittle, J. & Kruglyak, L. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* **436**, 701–703 (2005).
25. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003).
26. The Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Res.* **11**, 1425–1433 (2001).
27. Longtine, M.S. *et al.* Additional modules for versatile and economical PCR-based gene deletion and modification in *Saccharomyces cerevisiae*. *Yeast* **14**, 953–961 (1998).
28. Broman, K.W., Wu, H., Sen, S. & Churchill, G.A. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**, 889–890 (2003).
29. Lander, E.S. & Botstein, D. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199 (1989).
30. Hinnebusch, A.G. & Natarajan, K. Gcn4p, a master regulator of gene expression, is controlled at multiple levels by diverse signals of starvation and stress. *Eukaryot. Cell* **1**, 22–32 (2002).
31. Ronald, J., Brem, R.B., Whittle, J. & Kruglyak, L. Local regulatory variation in *Saccharomyces cerevisiae*. *PLoS Genet.* **1**, e25 (2005).
32. 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).
33. Keller, A., Nesvizhskii, A.I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383–5392 (2002).