Cell Reports
Resource

Quantitative and Qualitative Proteome Characteristics Extracted from In-Depth Integrated Genomics and Proteomics Analysis

Teck Yew Low,[1,2,6] Sebastiaan van Heesch,[3,6] Henk van den Toorn,[1,2,6] Piero Giansanti,[1,2] Alba Cristobal,[1,2] Pim Toonen,[3] Sebastian Schafer,[4] Norbert Hübner,[4,5] Bas van Breukelen,[1,2] Shabaz Mohammed,[1,2,7] Edwin Cuppen,[3,*] Albert J.R. Heck,[1,2,*] and Victor Guryev[3,8]

[1]Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Padualaan 8, 3584 CH Utrecht, the Netherlands
[2]Netherlands Proteomics Center, Padualaan 8, 3584 CH Utrecht, the Netherlands
[3]Hubrecht Institute-KNAW & University Medical Center Utrecht, Uppsalalaan 8, 3584 CT Utrecht, the Netherlands
[4]Max-Delbruck-Center for Molecular Medicine (MDC), Robert-Rossle-Strasse 10, 13125 Berlin, Germany
[5]DZHK (German Centre for Cardiovascular Research), Partner Site Berlin, Berlin, Germany
[6]These authors contributed equally to this work
[7]Present address: Departments of Chemistry and Biochemistry, University of Oxford, Physical & Theoretical Chemistry Laboratory, South Parks Road, OX1 3QZ Oxford, UK
[8]Present address: European Research Institute for the Biology of Ageing, University of Groningen, University Medical Center Groningen, Antonius Deusinglaan, 1, Building 3226, 9713 AV Groningen, the Netherlands
*Correspondence: e.cuppen@hubrecht.eu (E.C.), a.j.r.heck@uu.nl (A.J.R.H.)
http://dx.doi.org/10.1016/j.celrep.2013.10.041

SUMMARY

Quantitative and qualitative protein characteristics are regulated at genomic, transcriptomic, and post-transcriptional levels. Here, we integrated in-depth transcriptome and proteome analyses of liver tissues from two rat strains to unravel the interactions within and between these layers. We obtained peptide evidence for 26,463 rat liver proteins. We validated 1,195 gene predictions, 83 splice events, 126 proteins with nonsynonymous variants, and 20 isoforms with nonsynonymous RNA editing. Quantitative RNA sequencing and proteomics data correlate highly between strains but poorly among each other, indicating extensive nongenetic regulation. Our multi-level analysis identified a genomic variant in the promoter of the most differentially expressed gene *Cyp17a1*, a previously reported top hit in genome-wide association studies for human hypertension, as a potential contributor to the hypertension phenotype in SHR rats. These results demonstrate the power of and need for integrative analysis for understanding genetic control of molecular dynamics and phenotypic diversity in a system-wide manner.

INTRODUCTION

Mass spectrometry (MS)-based proteomics and next-generation sequencing (NGS) are rapidly maturing techniques, each enabling comprehensive measurements of gene products at a system level (Altelaar et al., 2013; Cox and Mann, 2011; Soon et al., 2013). Although MS and NGS are highly complementary, they are still rarely applied integrated in large-scale studies (Ning et al., 2012). State-of-the-art MS approaches can currently identify over 10,000 proteins in a single experiment (Munoz et al., 2011; Nagaraj et al., 2011), which brings the analysis of complete proteomes within reach (Ahrens et al., 2010; Cox and Mann, 2011). However, as long as noncustomary protein databases that are derived from (typically incomplete) reference genome assemblies and annotations remain the sole source used for MS spectra matching, true completeness will not be reached. For example, protein isoforms arising from genetic polymorphisms, posttranscriptional events such as RNA-editing and posttranslational modifications are largely missed (Jensen, 2004; Uhlen and Ponten, 2005).

Recent advances in NGS techniques, including whole genome sequencing (WGS) and total RNA sequencing (RNA-seq) allow for the generation of near-complete inventories of genetic variation in a system and its transcribed repertoire (Ozsolak and Milos, 2011). However, from such analyses, the effects on the proteins cannot be predicted with high confidence. For example, the consequence of a single nucleotide variant (SNV) on the coding capacity of a transcript can be predicted accurately, but not the potential effect on the stability of the corresponding protein. Systematic comparison of RNA-seq data with genomic data reveals another layer of complexity. It has now been convincingly demonstrated that certain transcripts are modified by posttranscriptional editing, primarily by targeted A to I deamination (Farajollahi and Maas, 2010). Most likely, all these types of variation will not only affect the composition and function of

a protein, but also influence expression levels. However, additional layers of translation control may dampen or completely abolish such effects (Kleinman and Majewski, 2012; Lin et al., 2012; Pickrell et al., 2012).

An integrative analysis of different data modalities, ideally derived from samples of a single source, is required for correctly deciphering the effects of genomic and transcriptomic variation on molecular processes and cellular functioning. An example of such data integration is the use of proteomic data derived from MS in combination with complete genome data to improve gene annotation (Jaffe et al., 2004; Renuse et al., 2011). This approach has so far been sparsely performed and mainly in organisms with smaller genomes (Merrihew et al., 2008; Venter et al., 2011). On the other hand, integrative investigations of messenger RNA levels and the proteins they encode reveal only modest correlations, implying an unresolved level of complexity in regulation of expression (Nesvizhskii et al., 2006; Ning et al., 2012; Schwanhäusser et al., 2011; de Sousa Abreu et al., 2009; Vogel and Marcotte, 2012).

For this study, we selected two rat inbred strains BN-*Lx*/Cub (BN-*Lx*) and SHR/OlaIpcv (SHR) (Printz et al., 2003), representing widely studied, renewable, and genetically homogeneous resources. Both strains have previously been extensively characterized at the genomic (Atanur et al., 2010; Gibbs et al., 2004) and phenotypic level (Hubner et al., 2005; Johnson et al., 2009; Pravenec and Kurtz, 2010; Simonis et al., 2012). The BN-*Lx* strain is derived from, and thus very closely related to, the Brown Norway (BN) strain. The latter strain was used for creating the rat reference genome assembly (Gibbs et al., 2004) and is commonly used as the protein reference data set in rat proteomics studies. The spontaneously hypertensive rat (SHR) is more diverged from BN and is a widely used disease model for hypertension studies. Whereas several blood pressure quantitative trait loci (QTLs) have been mapped to the SHR genome, no functional variants driving elevated blood pressure have been validated to date. Here, we combine in-depth genomic, transcriptomic, and proteomic analyses from inbred rats of two different genetic backgrounds using the same sets of rat liver tissues (Figure 1A). The liver is a large and relatively homogeneous tissue source that is well known to be involved in both hypertension and metabolic syndrome - the phenotypes associated with the SHR strain. We determine quantitative and qualitative molecular dynamics at different functional levels and achieve a level of proteome completeness by adding variation information derived from WGS and RNA-seq data. These data allow us to apply a genome-wide genetic-genomics approach (Jansen and Nap, 2001) to start understanding multilevel systems regulation and to identify candidate genes that are potentially involved in the hypertension phenotype of the SHR rat.

RESULTS AND DISCUSSION

Extension of the Rat Protein Database

In proteomics, tandem mass spectra are typically annotated by searching against in silico-generated spectra based on a publicly available protein database. For rat, such a database is derived from the reference genome assembly of the BN rat (Gibbs et al., 2004). To create a sample-specific database for MS peptide searching, we extended the existing RefSeq-based peptide database by incorporating strain-specific peptides and predicted peptides. We first obtained all strain-specific genetic variation of the BN-*Lx* and SHR strains including single nucleotide variants (SNVs) and in-frame indels.

Most genomic SNVs are located in the noncoding sequences, and only the less frequent nonsynonymous variants in coding regions give rise to altered amino acid sequences (Hurst et al., 2009; Su et al., 2011; Valentine et al., 2006). We collected 10,493 nonsynonymous variants from recently generated high-coverage WGS data of the BN-*Lx* and SHR genomes (Atanur et al., 2010; Simonis et al., 2012), which are predicted to affect 6,187 protein isoforms derived from 4,566 genes. Furthermore, to be able to detect in silico gene predictions using the proteomics data as evidence (Volkening et al., 2012), we added 44,993 GENSCAN gene predictions to our rat database (Burge and Karlin, 1997).

Next, we performed RNA-seq (Table S1) on RNA extracted from liver tissue of both rat strains (two males and two females per strain). To this end, paired-end sequencing data were generated to construct de novo transcriptome assemblies for each strain, In total, we found expression evidence for 18,116 known genes (12,052 with fragments per kilobase of exon per million fragments mapped [FPKM] >1), of which 2,612 (1,820 with FPKM >1) overlap the nonsynonymous variants previously detected by genome resequencing. Also, we identified 2,545 transcript splicing events affecting 1,015 genes. Although the majority of the identified splice events (1,687) were detected in both strains, 220 and 638 events were specific to BN-*Lx* and SHR rats, respectively (Table S2). Independent RT-PCR-based Sanger sequencing confirmed 74.1% (43 out of the 58 successful PCR assays) of a randomly sampled subset as true transcript isoforms (Table S3A). In addition, the same transcriptome assembly data provided expression evidence for 2,903 GENSCAN predictions (Table S4).

The de novo assembled transcriptome data also allow for characterization of transcriptomes at nucleotide resolution. Because both BN-*Lx* and SHR strains are fully inbred, observed changes at the transcript level are unlikely to be allele-specific variation and can thus be attributed to technical artifacts (introduced during sequencing or mapping) or to RNA editing (Farajollahi and Maas, 2010). We find a total of 799 canonical (A to I or C to T) RNA-editing variants (Table S5) of which 176 and 354 are specifically observed in BN-*Lx* and SHR, respectively. As expected, a large proportion of edits resides in the noncoding UTR parts of transcripts or do not change the coding capacity of a transcript. Yet, they might be affecting RNA secondary structure, stability, or miRNA binding. Only 196 edits were nonsynonymous and therefore included in our protein database as potentially detectable by MS. Of a subset of 169 candidate editing events tested by independent RT-PCR-based amplicon resequencing, most (104) showed reads corresponding to expected edited transcripts, and another 12 likely represent germline variants that missed detection during genome resequencing (Table S3B).

All peptide variants and isoforms derived from genome and transcriptome variation and all newly predicted peptides based on GENSCAN and de novo transcriptome assembly data were appended to the Ensembl rat database (3.4.63) to create our

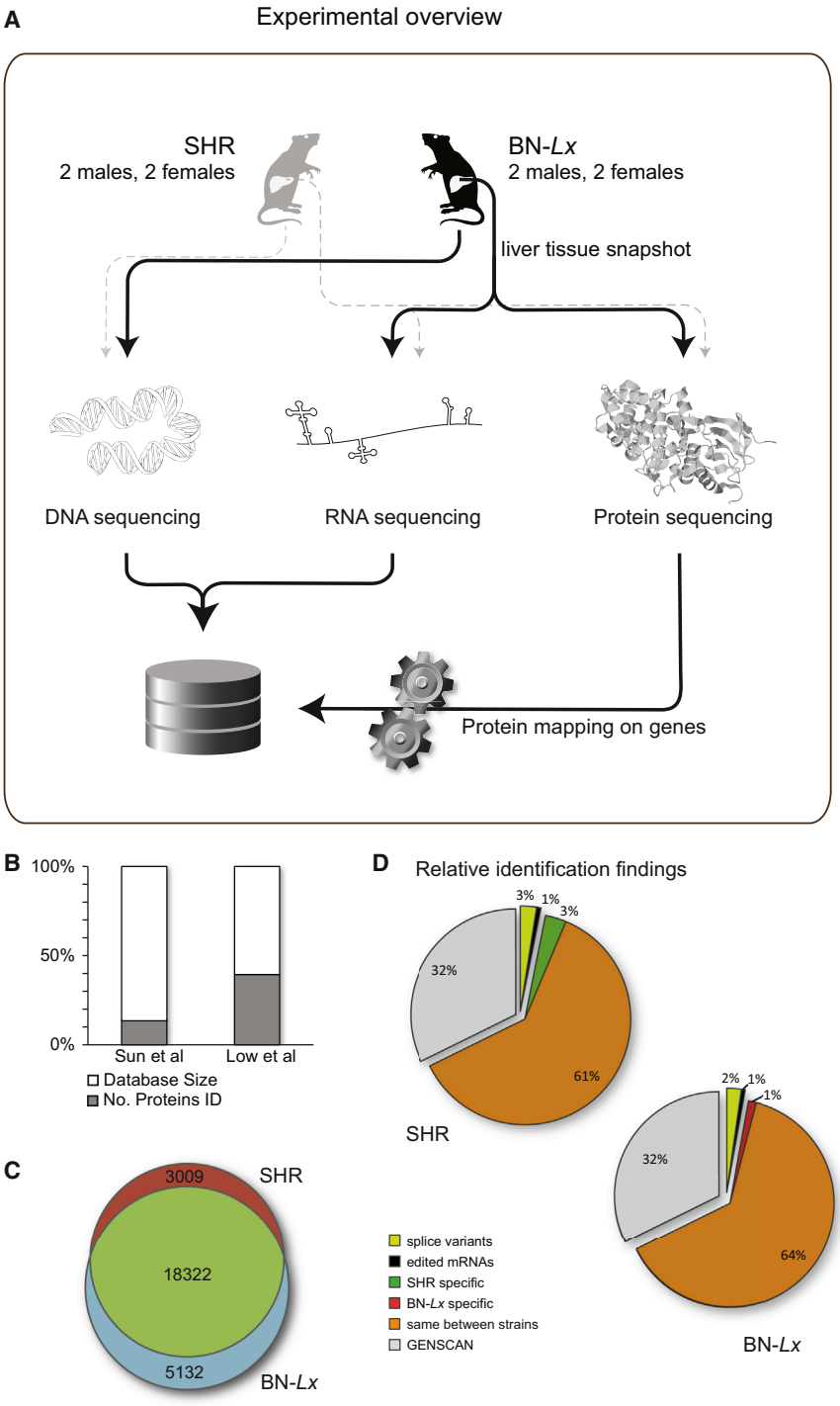


Figure 1. Integrated Proteomics, Genomics, and Transcriptomics to Improve Sample-Specific Protein Identification

(A) Schematic representation of the integrated genome and proteome analysis of BN-*Lx* and SHR rat liver using NGS and deep-proteome profiling.

(B) Bar plot showing the percentage of the current reference database that is covered by the experimentally derived proteomes, with respect to recent other proteomics efforts (Sun et al., 2010). For BN-*Lx* and SHR, 39.7% of the Ensembl database is represented (13,088 out of 32,971 entries; release 3.4.63). The human liver proteome generated by the Chinese Human Liver Proteome Profiling Consortium cover 13.5% of the IPI human database (version 3.07; 7,050 out of 50,225 entries).

(C) Diagram displaying identified proteins specific to BN-*Lx* (blue), SHR-specific proteins (red), and proteins shared between both strains (green).

(D) Relative contribution (%) to the BN-*Lx* and SHR rat proteomes (containing unique peptides) of each additional layer of genomics- and transcriptomics-derived protein variants.

See also Tables S1, S2, S3, S4, S5, S6, S8, S9, S10, S11, and S12.

customized RAT_COMBINED database, which was used for all subsequent proteomic analyses.

Proteomics Analysis

We generated proteomics data with the same liver tissues used for RNA-seq. Each lysate was proteolyzed with five orthogonal proteases, and the resulting 36 SCX fractions per digest were analyzed with LC-MS/MS, cumulating in 180 runs per strain, yielding ~12 million tandem MS spectra. By using multiple proteases, not only the identification and sequence coverage of each protein increase, but also the chance of capturing evidence for predicted peptides/proteins and consequences of RNA editing (Mohammed et al., 2008; Peng et al., 2012; Swaney et al., 2010). To ensure comprehensive coverage, two different but complementary algorithms for spectra-to-peptide assignment were applied. First, Mascot search engine was used for database searching. Next, remaining unassigned spectra were processed with PEAKS Studio 6.0 which incorporates a proprietary de novo sequencing algorithm. The large amount of data allowed us to apply a false discovery rate (FDR) filter of 0% (q = 0) and still identify over 2 million peptide-spectral matches (PSMs), corresponding to ~175,000 nonredundant peptides (Tables S6A and S6B). By performing a merged BN-*Lx* and SHR data set search against our custom RAT_COMBINED database, we obtained peptide evidence for 26,463 database entries. Of these, 18,322 are shared between BN-*Lx* and SHR (Figure 1C; Table S7), whereas 3,009 and 5,132 appear strain specific for SHR and BN-*Lx*, respectively. For comparison, we counted the number of identifications matching entries in the Ensembl database (3.4.63), disregarding

the variants. Out of the 32,971 original database entries, 13,088 were matched, representing 39.7% of database entries. In contrast, the most extensive liver proteome so far (the human liver proteome generated by the Chinese Human Liver Proteome Profiling Consortium) covers only 13.5% of the human IPI database (version 3.07; 50,225 entries), illustrating the depth of our data (Figure 1B; Table S6D). Over 86.5% of all proteins could be supported by evidence of gene expression in the RNA-seq data. As expected, identified peptides are evenly distributed over the rat chromosomes, concordant with the distribution of genes and transcripts (Figure S1A). The median coverage of all proteins is 15.6% with roughly equal contributions from each protease data set (Figure S1B).

Identification of Predicted Proteins and Protein Isoforms

Approximately 5,700 unique peptides (Table S8A) provide experimental evidence for 1,195 in silico predicted GENSCAN proteins (Table S7 and S8B). For 1,187 (99%) of those, RNA-seq data support the observed expression. Fifty of them show best reciprocal hits with known mouse proteins, and another 32 with known human proteins (Table S8C). Furthermore, we detect N-terminally acetylated peptides for 69 of these 1,187 proteins, with A, M, S, and T as their N-terminal residues (Table S8D) (Dormeyer et al., 2007; Starheim et al., 2012). N-terminal peptides validate these putative genes by confirming their translational start sites. A different class of proteins with largely uncertain existence is the short expressed proteins (SEPs) encoded by short open reading frames (Slavoff et al., 2013). Of all peptides in our data set, 0.25% could be assigned to 86 known SEPs and 37 identified SEPs (Figure S2; Tables S7 and S8B).

The proteomics data also provide support for 83 transcript splicing events (0% FDR) that were previously not annotated (Figure 1D; Tables S6C, S7, S10A, and S10D). From all predicted proteins and splice isoform identifications, 309 and 15 respectively were unique for BN-*Lx*, and 193 and 13 were specific to SHR (Table S6C).

Detection of Nonsynonymous Protein Variants

Next, we explored to what extent the addition of strain-specific variants affected protein detectability and stability. Of the uniquely assigned spectra, 3.5% did discriminate between allele-specific protein isoforms. We detected 126 nonsynonymous variants in our proteomic data, 38 for BN-*Lx*, and 88 for SHR (Tables S10A and S10B). By applying a 0% FDR cutoff, we reassuringly did not find any BN-*Lx* variants in the SHR samples, and vice versa (Table S6C). The fact that only a portion of nonsynonymous variants was confirmed by peptide-based evidence can be explained by our experimental design in which only genes expressed in the liver could be detected. Clearly, the inclusion of allele-specific variants has a measurable impact on protein discovery and results in more balanced peptide count per strain. The latter is most notable for the SHR rat because its genome is more diverged from the reference strain (BN).

We used SIFT and Polyphen2 to predict if nonsynonymous SNVs could affect protein stability (Tables S9 and S11). Potentially damaging mutations were clearly overrepresented in differentially expressed proteins with nondifferential transcript levels ($p < 0.002$) (Table S11). This illustrates that nonconservative and structural missense variants may have limited influence on the abundance of a transcript yet can show a pronounced effect on protein stability.

Peptide-Based Evidence for RNA Editing

To identify functional RNA-editing events, we mapped our peptide spectra to the set of potential RNA-editing events. In total, 20 out of the 196 nonsynonymous editing events could be confirmed by unique peptide-based evidence (Tables S6C, S9, S10A, and S10C). Because unique peptide evidence needs to overlap with the predicted editing site, many of the remaining 176 edits are likely missed because of incomplete coverage or redundancy in peptide data. Whereas limitations in the MS technology obviously result in an underrepresentation of identified RNA edits, MS still provides the best means to confirm the presence of such posttranscriptional modifications in the expressed proteins. On the other hand, we cannot rule out a possibility that the relatively low percentage of confirmed events is a true representation of the actual level of posttranscriptional modifications that make it to mature proteins. This may be due to negative selection against modified mRNA molecules. The high level of RNA sequencing coverage and the strict calling settings used to define editing events make it unlikely that an overestimation of editing events is introduced during the RNA sequencing procedure and analysis.

It is worth noting that our comparison of de novo assembled and the annotated transcriptomes may not only reveal genetic differences, transcript isoforms, and common edited sites. Sequence and annotation imperfections within the current assembly and gene build can also be detected because the proteogenomics approach used in this study accounts for differences between observed and annotated transcriptome that originate from both biological and technical sources. Also, we emphasize that the de novo transcriptome assembly approach should be supplemented by regular transcriptome profiling if one aims to discover transcript variants that correspond to low-abundance transcripts and low-frequency events. To this end, we performed direct alignment of RNA-seq data to the rat transcriptome (known proteins and GENSCAN predictions) and predicted additional modifications of annotated transcripts (Table S5).

Relation between Transcriptome and Proteome Levels

Next, we studied quantitative aspects by investigating the abundance of mRNA and protein levels. Although being derived from two different strains of rats, we observed a very high correlation of liver mRNA between BN-*Lx* and SHR ($r = 0.98$). Similarly, the correlation coefficient for protein expression between BN-*Lx* and SHR is also remarkably high ($r = 0.94$) (Figure 2), providing confidence in our quantification strategy based on spectral counts.

Next, we sought to define a correlation between mRNA and protein expression levels in our data. Making a direct correlation between mRNA and protein levels is hampered by the fact that in peptide-based proteomics many proteins contain similar peptide sequences. It is therefore hard to assign any of the shared peptides unambiguously to a protein, the so-called

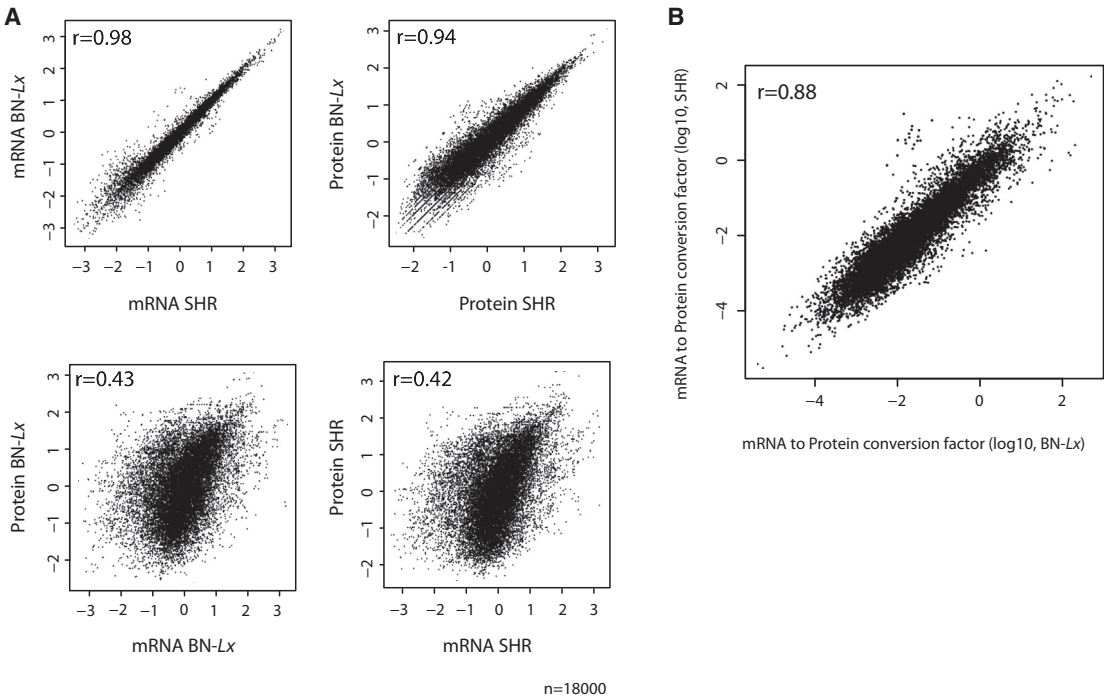


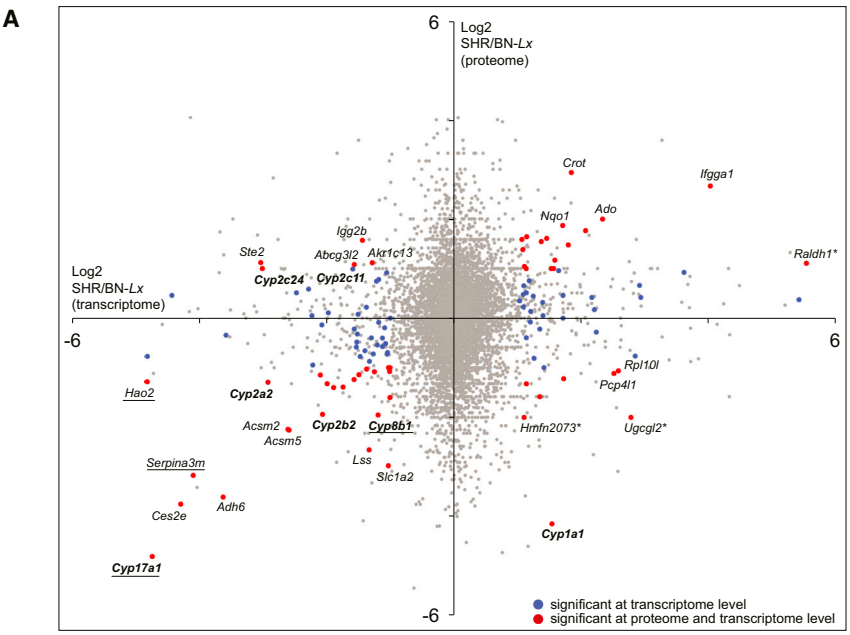
Figure 2. Global Correlation Plots Displaying the Complexity of mRNA and Protein Abundance

(A) The top two panels display the high correlations between BN-*Lx* and SHR mRNA (left, r = 0.98) and protein levels (right, r = 0.94), estimated using log10 normalized spectral counts (Log10SAF) and normalized RNA seq counts (Log10FPKM). The bottom two panels show the poor correlations between mRNA and protein abundance for BN-*Lx* (r = 0.43) and SHR (r = 0.42), respectively.

(B) Scatterplot depicting the correlation between experimentally determined gene-specific mRNA to protein abundance conversion factors as calculated for both BN-*Lx* and SHR (r = 0.88).

See also Table S15.

protein-inference problem (Grobei et al., 2009; Nesvizhskii and Aebersold, 2005). Consequently, it is hard to integrate the quantitative measurements, which are necessarily restricted to peptides, to a protein measurement. Still, numerous studies conclude that the global correlation between mRNA and protein is certainly not linear and often an r of 0.4–0.5 is reported (Ning et al., 2012; de Sousa Abreu et al., 2009; Vogel and Marcotte, 2012). Such findings are corroborated by results that show that indeed only part of the variation in the protein levels can be explained by mRNA levels (Schwanhäusser et al., 2011). Here, we use a spectra-count method for quantification of protein levels. We use data derived from five different proteolytic enzymes, which is sufficient to exclude a proteolytic digest-specific bias (Peng et al., 2012). Although we did identify unique peptides per protein (Table S12), we chose to take the total number of PSMs for every peptide matching a protein as a measurement of its abundance to increase the quantitative resolution per protein. Subsequently, we determined the proteome-transcriptome correlation for BN-*Lx* and SHR to be r = 0.43 and 0.42, respectively (Figure 2A). This correlation is thus weak, albeit in line with the previous studies in other systems. Based on these quantitative comparisons, we also found that 3′ UTR expression levels correlate increasingly better with protein levels (r = 0.54) than do 5′ UTR levels (r = 0.43) or reads derived from the coding sequence (r = 0.47) (Figures S3C–S3E). We speculate that the abundance of 3′ UTR reads depends on transcript integrity and reflects both transcript count and stability. Transcript levels could also be reproducibly converted to predicted protein levels using a gene-specific conversion factor, which showed high correlation between the two strains (r = 0.88) (Figure 2B; Table S15). The high correlation between strains for this gene-specific factor illustrates the conservation of quantitative mRNA levels in relation to protein levels, independent of intermediate (less understood) levels of expression regulation. Although the conversion factor cannot be analyzed in-depth within the scope of this article, we postulate that translation efficiency, RNA, and protein degradation (and thus stability) are likely to play an important role. The top 100 proteins with the lowest and highest conversion factors were subjected to gene ontology (GO) overrepresentation analysis. We observed a trend in cellular localization toward cytoskeleton (highest 100) or the membrane (lowest 100), although the observations were not significant (Figure S3B; Table S15). We can only speculate that the conversion factor appears to be protein specific and conserved between strains. This factor combines the aforementioned levels of gene expression regulation in one value. One particular group of proteins appears to behave differently, representing the family of α_{2u}-globulins (known as rat major urinary proteins; Figures S3A and S3B). Unfortunately, none of these proteins could be identified by unique peptides due to high protein sequence homology



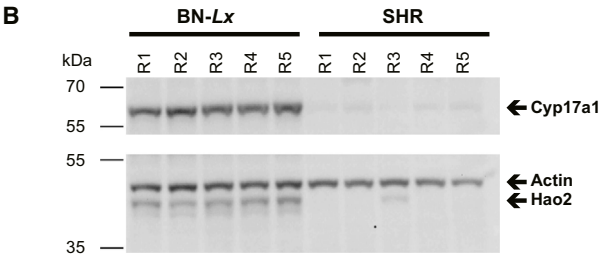


Figure 3. Gene-Centric Strain-to-Strain Comparison of Significantly Differentially Expressed Genes

(A) Genes in BN-*Lx* and SHR with significantly deviating mRNA levels (blue dots; n = 59) or mRNA and protein levels (red dots; n = 54) are highlighted. Gene names marked by an asterisk are based on GENSCAN blast predictions derived from the closest predicted homology to human and mouse genes. Genes belonging to the CYP450 superfamily of catalytic enzymes are in bold and genes associated with hypertension in human or rat literature (*Hao2*, *Serpina3m*, *Cyp8b1*, *and Cyp17a1*) are underscored.

(B) Western blot performed with liver tissues from five animals each for BN-*Lx* and SHR. Both Cyp17a1 and Hao2 are downregulated in all the biological replicates in the SHR strain compared to BN-*Lx*, consistent with the proteomics data. Actin was used as a loading control.

See also Tables S7 and S13.

within this class of genes. Therefore, although this class stands out in the quantitative comparisons, absolute differences between BN-*Lx* and SHR cannot be determined with high confidence at this point.

Genetic Control of Quantitative Proteome Characteristics

To determine the effects of genetic variation on quantitative transcriptome and proteome characteristics, we compared the difference of mRNA and protein expression between the two rat strains. First, we filtered our quantitative data with more stringent criteria retaining only genes quantifiable at both the protein and the RNA level (reliable expression level estimates by Cuffdiff and nonzero SAF counts). This allowed us to compare 6,743 genes (Figure 3A; Table S13), 113 of which were differentially expressed at the RNA level (at least 2-fold change in expression; and $q < 0.01$) and 205 at the protein level (at least 2-fold change and $q = 0$). The majority of the differentially expressed transcripts (59/113) do not show comparable changes at proteomics level. This group of proteins likely acquires stable expression through regulation of at the level of translation or through proteostasis. A small proportion of the genes (13/113) shows discordant behavior with opposite expression patterns for transcripts and proteins. Both groups do not show any overrepresentation in GO terms or pathways. The limited number of genes with significantly altered expression indicates the high global genome and proteome similarity between the two inbred rat strains. However, it also illustrates that interindividual differences may be in the details, such as represented by changes in posttranslational protein modifications and protein networks (Altelaar et al., 2013; Bensimon et al., 2012). Finally, 41 out of the 113 differential genes show strain-specific expression changes that are consistent between transcriptome and proteome (Figure 3A; Table S13). The products of these 41 genes relate to catalytic activity (28 genes, GO-term enrichment p value 1.4e-5) and metabolic pathways (13 genes, p = 2.6e-4).

A Germline Promoter Variant Deregulates Cyp17a1 Expression in Spontaneously Hypertensive Rats

This set of 41 genes likely underlies some of the phenotypic differences known to exist between BN-*Lx* and SHR rats, like spontaneous hypertension (Okamoto and Aoki, 1963) and metabolic syndrome (Aitman et al., 1997, 1999). We therefore investigated which genes were previously reported to be associated with hypertension in human or rat. First, three out of the 41 genes that are differential at both the mRNA and protein level were found to be associated with hypertension in the rat. Those three genes, *Hao2* (Lee et al., 2003), *Serpina3m*, and *Cyp8b1* (Kinoshita et al., 2011), came out as top hits when studying SHR (-related) strains or a panel of congenic rat strains to define candidates for hypertension. All three genes also overlap known blood pressure QTLs in the rat (Dwinell et al., 2009), and two of them (*Serpina3m* and *Hao2*) show a very strong connection to the SHR genotype based on eQTL data derived from the BXH/HXB

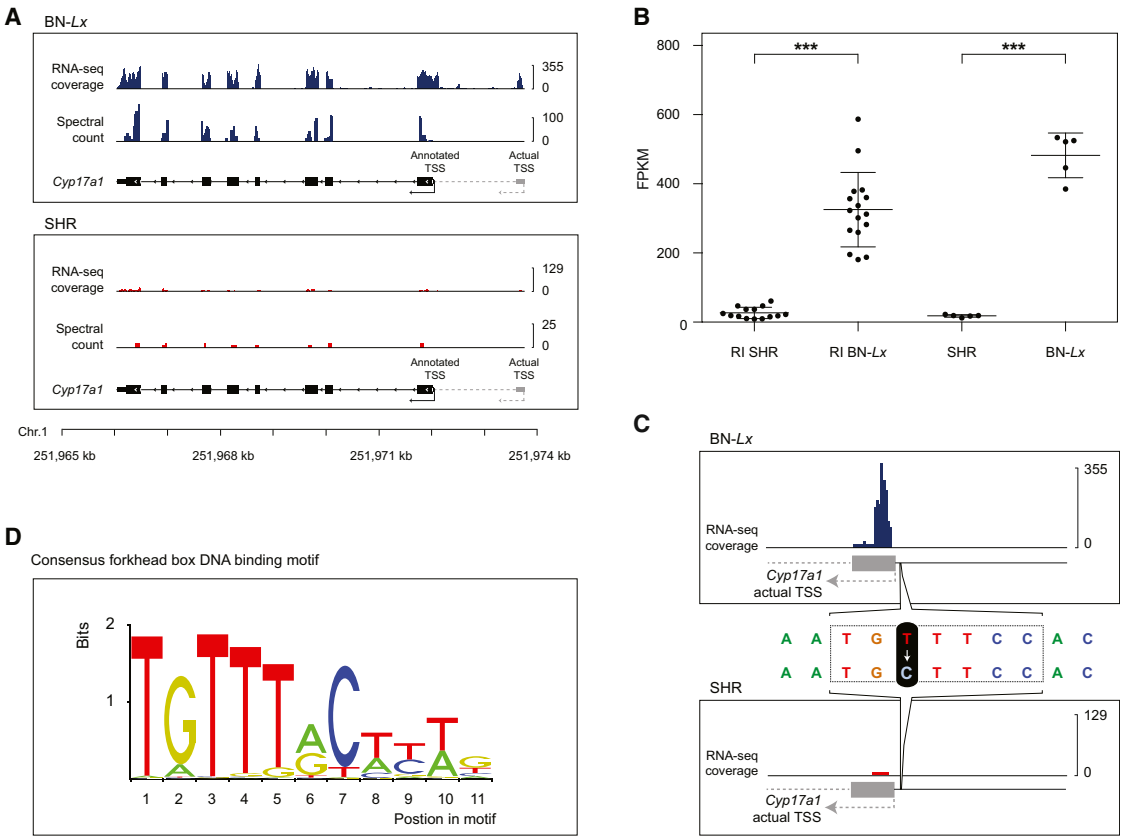


Figure 4. A Germline Promoter Variant Deregulates Cyp17a1 Expression in Spontaneously Hypertensive Rats

(A) Experimental evidence covering this part of the genome from RNA sequencing and the proteomics data (spectral counts) are plotted along the gene body of *Cyp17a1* for BN-*Lx* (blue) and SHR (red). The transcript is positioned on the reverse strand. Both the annotated transcription start site (TSS, black arrow) and the actual TSS (gray arrow) are shown.

(B) Expression QTL analysis of Cyp17a1 expression in the HXB/BXH recombinant inbred panel. Gene expression is plotted based on RNA-seq for the ancestral strains (n = 5) and the RI strains split by ancestral haplotype at the Cyp17a1 locus (n = 16 for BN-*Lx* and n = 14 for SHR).

(C) Zoomed-in view of the actual TSS, with the position of the germline T/C SNV shown. The dashed box (gray) shows the core part of the forkhead box DNA binding motif.

(D) Consensus forkhead box DNA binding motif, obtained from the JASPAR database FOXA1 motif (Sandelin et al., 2004).

See also Table S14.

recombinant inbred panel (founded by the BN-*Lx* and SHR strains) (Figure S4A–S4C). This implies that the gene expression regulation of Serpina3m and Hao2 is regulated in *cis* and thus strongly related to the genotype of the SHR strain. A fourth gene, *Cyp17a1*, was identified as a top hit in relation to blood pressure and hypertension in human genome-wide association studies on European, Japanese, and Chinese individuals (Li et al., 2013; Liu et al., 2011; Newton-Cheh et al., 2009; Takeuchi et al., 2010) (Table S14A). *Cyp17a1* also overlaps a blood pressure QTL in the rat and shows the most extreme downregulation in SHR compared to BN-*Lx* in our analysis (Figure 3A). The differential expression of Hao2 and Cyp17a1 was verified independently by western blot, using liver samples of five animals from each strain (Figure 3B). Like Cyp8b1, Cyp17a1 is a member of the cytochrome P450 (CYP450) superfamily (Danielson, 2002) of catalytic enzymes that mediate monooxygenase reactions and regulate drug metabolism. Interestingly, mutations in human *CYP17A1* are known to lead to congenital adrenal hyperplasia due to 17 alpha-hydroxylase deficiency, which results in hypogonadism, pseudohermaphroditism, and severe hypertension (Biglieri, 1997; Biglieri et al., 1966; Geller et al., 1997; Goldsmith et al., 1967). To determine the genetic basis of the *Cyp17a1* expression differences between BN-*Lx* and SHR, we sought for genetic variants in the annotated exons and flanking regulatory sequences, but none were present. Exploration of eQTL data, however, revealed a very strong *cis*-effect (Figure 4B; Table S14B), indicating that the measured expression difference is due to genetic variants in the gene itself or in neighboring regulatory elements. Upon closer inspection of the RNA-seq data, we found that the transcriptional start site (TSS) of the *Cyp17a1* gene was incorrectly annotated and resides approximately 2 kb upstream of the currently annotated most 5′ exon (Figure 4A). The true location of the promoter could be confirmed by H3K4me3 ChIP data that show specific enrichment of this active promoter mark surrounding the nucleosome-free region of the unannotated TSS (Figure S4D). Interestingly, this promoter

does harbor a germline variant in SHR that disrupts the core part of an evolutionary conserved forkhead-box DNA binding domain (Figures 4C and 4D) (Sandelin et al., 2004), specifically deregulating transcription in SHR (Figure 4A). Because this expression trait is regulated in *cis* and this SNV is the only germline variant in the vicinity of the gene, our integrated genomics, transcriptomics, and proteomics approach has most likely identified the source of expression variation. The overlap with the RGD blood pressure QTL (http://rgd.mcw.edu/), top GWAS loci in humans, and known link to hypertension as a result of renal hyperplasia in patients carrying *CYP17A1* mutations are good indications that this promoter mutation in the SHR *Cyp17a1* gene contributes to the observed hypertensive phenotype of SHR rats.

Conclusions

Technological advances in both the proteomics and the sequencing community now provide the ability to discriminate genetic and posttranscriptional polymorphisms at the proteome level. These advances also allow improved quantitation of gene expression, which is generally restricted by the imprecise proxy of transcriptome data alone. We here show that the synergistic use of genomic, transcriptomic, and proteomic technologies significantly improves the information load that can be gained from proteomics as well as genomics efforts. By matching deep MS-based proteomics to a personalized database built from a sample-specific genome and transcriptome, we identify thousands of peptides that would otherwise escape identification. We believe that future efforts on both platforms benefit largely from our proof-of-concept approach, which brings integrated proteogenomics to a higher level. To highlight the strength of this approach, we present a link of a genomic variant in the *Cyp17a1* gene promoter and associate it with the hypertension phenotype of the extensively studied SHR rats.

EXPERIMENTAL PROCEDURES

An extended version of the experimental procedures can be found in the Supplemental Information.

Custom Rat Protein Database Construction

We modified and appended the Ensembl (Birney et al., 2004; Curwen et al., 2004; Hubbard et al., 2002) rat protein FASTA (build 3.4.63), which was derived from the reference (BN) genome assembly, with DNA resequencing and RNA-seq data of the BN-*Lx* and SHR strains. Single nucleotide variants and indels were obtained from previous genome sequencing efforts (Atanur et al., 2010; Simonis et al., 2012). RNA was isolated from liver tissues of 6-week-old inbred BN-*Lx*/Cub and SHR/Olalpcv rats. SOLiD RNA-seq libraries were prepared with ribosomal RNA depleted RNA and sequenced on the SOLiD V4 system. Next, we used CLC assembly cell version 4 (CLC Bio) to de novo assemble each rat liver transcriptome. Merged BN-*Lx* and SHR transcriptomes were mapped against the reference genome assembly using BLAT software (Kent, 2002). Splicing and RNA-editing events were detected using alignments between the assembled transcriptome and genome and compared to their corresponding proteins. For all nonsynonymous genomic and transcriptome variants, individual entries were added to the extended protein search database. Also, we included 44,993 GENSCAN gene predictions of which 17,100 (FPKM >0.1) or 4,998 (FPKM >1.0) show evidence of expression.

Quantification of Transcriptome Data and Identification of eQTLs

To quantify expression differences, RNA-seq data for each sample were aligned to reference genome using TopHat2. Expressed and differentially expressed genes were defined by Cuffdiff using all four transcriptomes per strain (Trapnell et al., 2012). Determination of eQTLs in the HXB/BXH recombinant inbred panel consisting of 30 rat strains was performed exactly as previously described (Heinig et al., 2010).

Strong Cation Exchange Chromatography

After sonication and centrifugation, liver tissue lysates (300 μg each) were proteolyzed using trypsin, LysC, GluC, AspN, and chymotrypsin. After desalting, peptides were fractionated using a strong cation exchange (SCX) column (Zorbax BioSCX-Series II; 0.8 mm inner diameter × 50 mm length, 3.5 μm), and 36 fractions were collected per digest.

Mass Spectrometry Analysis

The first 26 fractions were analyzed with an Agilent 1290 Infinity (Agilent Technologies) LC, operating in reverse-phase (C18) mode, coupled to a TripleTOF 5600 (AB Sciex). MS spectra (350–1,250 m/z) were acquired in high-resolution mode (R > 30,000), whereas MS2 in high-sensitivity mode (R > 15 000). The next ten fractions were analyzed with a Proxeon EASY-nLC 1000 (Thermo Scientific) operating in reverse phase (C18) and connected to an LTQ-Orbitrap Velos (Thermo Fisher Scientific). For MS analysis, MS spectra (350–1,500 m/z) were acquired at a resolution of 30,000 and for MS2, R = 7,500.

Protein Database Searching

Peak lists (MGFs) were submitted to the Mascot (version 2.3) via Proteome Discoverer version 1.3 (Thermo Fisher Scientific) and searched against RAT_COMBINED with the respective proteases chosen. Peptide tolerance was 50 ppm, and MS/MS tolerance was 0.1 Da (TOF), 0.02 Da (Orbitrap), and 0.5 Da (ion trap). All PSMs were validated with Percolator (Käll et al., 2007) based on q = 0 (0% FDR). Only PSMs ranked first by the search engine with at least six amino acids were kept. Unmatched spectra were exported for analysis with PEAKS Studio (version 6.0). Peak lists were filtered with a quality value of 0.65, followed by a tag database search. The maximum allowed variable PTM per peptide was set to 3. De novo interpreted PSMs were submitted to PEAKS DB database matching, allowing semienzymatic specificity and a maximum cleavage per peptide of 2. The FDR was estimated using a concatenated decoy database and according to a threshold of 0.0%.

Quantitative Comparison of Proteome and Transcriptome Data

To combine quantitative data from all methods, we developed a relational database schema (Figure S8) for data storage. The database schema was converted to Java (Java SE 7, Oracle) entities, using Java Persistence API (JPA version 2) implemented in EclipseLink version 2.3.2 (http://www.eclipse.org/eclipselink), with the tools provided in Netbeans IDE 7.3 (http://www.netbeans.org). The database used was MySQL version 5.5 (Oracle).

ACCESSION NUMBERS

The ProteomeXchange (Vizcaíno et al., 2013) accession number for the MS data reported in this paper is PXD000131. The Sequence Read Archive accession numbers for the DNA data are ERP001355 (BN-*Lx* genome), ERP001371 (SHR genome), and ERP000510 (BN reference genome). RNA sequencing data were stored in ArrayExpress under the accession number E-MTAB-1666.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, five figures, and fifteen tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2013.10.041.

AUTHOR CONTRIBUTIONS

T.Y.L designed, performed, and analyzed the proteomics experiments; S.v.H. designed, performed, and analyzed the RNA-seq experiments. V.G. performed bioinformatics analysis on genomics, transcriptomics, and proteomics data and is responsible for generating the protein database. H.v.d.T. and V.G. performed qualitative, quantitative, and bioinformatics analysis on both

transcriptomics and proteomics data. P.G. and A.C. performed MS and data analysis. B.v.B. and S.M. provided consultation and support for bioinformatics and MS. S.v.H, P.T., and V.G. performed and analyzed RNA-seq validation experiments. S.S. and N.H. provided eQTL data and interpretation. T.Y.L., S.v.H., H.v.d.T., S.M., V.G., E.C., and A.J.R.H. contributed to conceptual design and scientific discussions. T.Y.L., S.v.H., H.v.d.T., B.v.B., S.M., A.J.R.H., E.C., and V.G. wrote the manuscript. E.C. and A.J.R.H. supervised the study.

ACKNOWLEDGMENTS

This work was supported by the Netherlands Proteomics Centre, which is part of the Netherlands Genomics Initiative and a TOP grant from NWO-CW (N° 700.58.303) to E.C. This work was financially supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. HEALTH-F4-2010-241504 (EURATRANS) to E.C. and N.H. and the PRIME-XS project grant agreement number 262067 to A.J.R.H. We would like to thank Dr. Vincentius A. Halim for technical assistance. We would also like to thank the PRIDE Team for assistance.

Received: July 8, 2013
Revised: September 28, 2013
Accepted: October 24, 2013
Published: November 27, 2013

REFERENCES

Ahrens, C.H., Brunner, E., Qeli, E., Basler, K., and Aebersold, R. (2010). Generating and navigating proteome maps using mass spectrometry. Nat. Rev. Mol. Cell Biol. *11*, 789–801.

Aitman, T.J., Gotoda, T., Evans, A.L., Imrie, H., Heath, K.E., Trembling, P.M., Truman, H., Wallace, C.A., Rahman, A., Doré, C., et al. (1997). Quantitative trait loci for cellular defects in glucose and fatty acid metabolism in hypertensive rats. Nat. Genet. *16*, 197–201.

Aitman, T.J., Glazier, A.M., Wallace, C.A., Cooper, L.D., Norsworthy, P.J., Wahid, F.N., Al-Majali, K.M., Trembling, P.M., Mann, C.J., Shoulders, C.C., et al. (1999). Identification of Cd36 (Fat) as an insulin-resistance gene causing defective fatty acid and glucose metabolism in hypertensive rats. Nat. Genet. *21*, 76–83.

Altelaar, A.F.M., Munoz, J., and Heck, A.J.R. (2013). Next-generation proteomics: towards an integrative view of proteome dynamics. Nat. Rev. Genet. *14*, 35–48.

Atanur, S.S., Birol, I., Guryev, V., Hirst, M., Hummel, O., Morrissey, C., Behmoaras, J., Fernandez-Suarez, X.M., Johnson, M.D., McLaren, W.M., et al. (2010). The genome sequence of the spontaneously hypertensive rat: Analysis and functional significance. Genome Res. *20*, 791–803.

Bensimon, A., Heck, A.J.R., and Aebersold, R. (2012). Mass spectrometry-based proteomics and network biology. Annu. Rev. Biochem. *81*, 379–405.

Biglieri, E.G. (1997). 17 alpha-Hydroxylase deficiency: 1963-1966. J. Clin. Endocrinol. Metab. *82*, 48–50.

Biglieri, E.G., Herron, M.A., and Brust, N. (1966). 17-hydroxylation deficiency in man. J. Clin. Invest. *45*, 1946–1954.

Birney, E., Andrews, T.D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T., et al. (2004). An overview of Ensembl. Genome Res. *14*, 925–928.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. *268*, 78–94.

Cox, J., and Mann, M. (2011). Quantitative, high-resolution proteomics for data-driven systems biology. Annu. Rev. Biochem. *80*, 273–299.

Curwen, V., Eyras, E., Andrews, T.D., Clarke, L., Mongin, E., Searle, S.M.J., and Clamp, M. (2004). The Ensembl automatic gene annotation system. Genome Res. *14*, 942–950.

Danielson, P.B. (2002). The cytochrome P450 superfamily: biochemistry, evolution and drug metabolism in humans. Curr. Drug Metab. *3*, 561–597.

de Sousa Abreu, R., Penalva, L.O., Marcotte, E.M., and Vogel, C. (2009). Global signatures of protein and mRNA expression levels. Mol. Biosyst. *5*, 1512–1526.

Dormeyer, W., Mohammed, S., Breukelen, Bv., Krijgsveld, J., and Heck, A.J.R. (2007). Targeted analysis of protein termini. J. Proteome Res. *6*, 4634–4645.

Dwinell, M.R., Worthey, E.A., Shimoyama, M., Bakir-Gungor, B., DePons, J., Laulederkind, S., Lowry, T., Nigram, R., Petri, V., Smith, J., et al.; RGD Team (2009). The Rat Genome Database 2009: variation, ontologies and pathways. Nucleic Acids Res. *37* (Database issue), D744–D749.

Farajollahi, S., and Maas, S. (2010). Molecular diversity through RNA editing: a balancing act. Trends Genet. *26*, 221–230.

Geller, D.H., Auchus, R.J., Mendonça, B.B., and Miller, W.L. (1997). The genetic and functional basis of isolated 17,20-lyase deficiency. Nat. Genet. *17*, 201–205.

Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E., et al.; Rat Genome Sequencing Project Consortium (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature *428*, 493–521.

Goldsmith, O., Solomon, D.H., and Horton, R. (1967). Hypogonadism and mineralocorticoid excess. The 17-hydroxylase deficiency syndrome. N. Engl. J. Med. *277*, 673–677.

Grobei, M.A., Qeli, E., Brunner, E., Rehrauer, H., Zhang, R., Roschitzki, B., Basler, K., Ahrens, C.H., and Grossniklaus, U. (2009). Deterministic protein inference for shotgun proteomics data provides new insights into Arabidopsis pollen development and function. Genome Res. *19*, 1786–1800.

Heinig, M., Petretto, E., Wallace, C., Bottolo, L., Rotival, M., Lu, H., Li, Y., Sarwar, R., Langley, S.R., Bauerfeind, A., et al.; Cardiogenics Consortium (2010). A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. Nature *467*, 460–464.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., et al. (2002). The Ensembl genome database project. Nucleic Acids Res. *30*, 38–41.

Hubner, N., Wallace, C.A., Zimdahl, H., Petretto, E., Schulz, H., Maciver, F., Mueller, M., Hummel, O., Monti, J., Zidek, V., et al. (2005). Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nat. Genet. *37*, 243–253.

Hurst, J.M., McMillan, L.E.M., Porter, C.T., Allen, J., Fakorede, A., and Martin, A.C.R. (2009). The SAAPdb web resource: a large-scale structural analysis of mutant proteins. Hum. Mutat. *30*, 616–624.

Jaffe, J.D., Berg, H.C., and Church, G.M. (2004). Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics *4*, 59–77.

Jansen, R.C., and Nap, J.P. (2001). Genetical genomics: the added value from segregation. Trends Genet. *17*, 388–391.

Jensen, O.N. (2004). Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. Curr. Opin. Chem. Biol. *8*, 33–41.

Johnson, M.D., He, L., Herman, D., Wakimoto, H., Wallace, C.A., Zidek, V., Mlejnek, P., Musilova, A., Simakova, M., Vorlicek, J., et al. (2009). Dissection of chromosome 18 blood pressure and salt-sensitivity quantitative trait loci in the spontaneously hypertensive rat. Hypertension *54*, 639–645.

Käll, L., Canterbury, J.D., Weston, J., Noble, W.S., and MacCoss, M.J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat. Methods *4*, 923–925.

Kent, W.J. (2002). BLAT—the BLAST-like alignment tool. Genome Res. *12*, 656–664.

Kinoshita, K., Ashenagar, M.S., Tabuchi, M., and Higashino, H. (2011). Whole rat DNA array survey for candidate genes related to hypertension in kidneys from three spontaneously hypertensive rat substrains at two stages of age and with hypotensive induction caused by hydralazine hydrochloride. Exp Ther Med *2*, 201–212.

Kleinman, C.L., and Majewski, J. (2012). Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". Science *335*, 1302, author reply 1302.

Lee, S.J., Liu, J., Qi, N., Guarnera, R.A., Lee, S.Y., and Cicila, G.T. (2003). Use of a panel of congenic strains to evaluate differentially expressed genes as candidate genes for blood pressure quantitative trait loci. Hypertens. Res. *26*, 75–87.

Li, X., Ling, Y., Lu, D., Lu, Z., Liu, Y., Chen, H., and Gao, X. (2013). Common polymorphism rs11191548 near the CYP17A1 gene is associated with hypertension and systolic blood pressure in the Han Chinese population. Am. J. Hypertens. *26*, 465–472.

Lin, W., Piskol, R., Tan, M.H., and Li, J.B. (2012). Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". Science *335*, 1302, author reply 1302.

Liu, C., Li, H., Qi, Q., Lu, L., Gan, W., Loos, R.J., and Lin, X. (2011). Common variants in or near FGF5, CYP17A1 and MTHFR genes are associated with blood pressure and hypertension in Chinese Hans. J. Hypertens. *29*, 70–75.

Merrihew, G.E., Davis, C., Ewing, B., Williams, G., Käll, L., Frewen, B.E., Noble, W.S., Green, P., Thomas, J.H., and MacCoss, M.J. (2008). Use of shotgun proteomics for the identification, confirmation, and correction of C. elegans gene annotations. Genome Res. *18*, 1660–1669.

Mohammed, S., Lorenzen, K., Kerkhoven, R., van Breukelen, B., Vannini, A., Cramer, P., and Heck, A.J.R. (2008). Multiplexed proteomics mapping of yeast RNA polymerase II and III allows near-complete sequence coverage and reveals several novel phosphorylation sites. Anal. Chem. *80*, 3584–3592.

Munoz, J., Low, T.Y., Kok, Y.J., Chin, A., Frese, C.K., Ding, V., Choo, A., and Heck, A.J.R. (2011). The quantitative proteomes of human-induced pluripotent stem cells and embryonic stem cells. Mol. Syst. Biol. *7*, 550.

Nagaraj, N., Wisniewski, J.R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Pääbo, S., and Mann, M. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. Mol. Syst. Biol. *7*, 548.

Nesvizhskii, A.I., and Aebersold, R. (2005). Interpretation of shotgun proteomic data: the protein inference problem. Mol. Cell. Proteomics *4*, 1419–1440.

Nesvizhskii, A.I., Roos, F.F., Grossmann, J., Vogelzang, M., Eddes, J.S., Gruissem, W., Baginsky, S., and Aebersold, R. (2006). Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. Mol. Cell. Proteomics *5*, 652–670.

Newton-Cheh, C., Johnson, T., Gateva, V., Tobin, M.D., Bochud, M., Coin, L., Najjar, S.S., Zhao, J.H., Heath, S.C., Eyheramendy, S., et al.; Wellcome Trust Case Control Consortium (2009). Genome-wide association study identifies eight loci associated with blood pressure. Nat. Genet. *41*, 666–676.

Ning, K., Fermin, D., and Nesvizhskii, A.I. (2012). Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with RNA-Seq gene expression data. J. Proteome Res. *11*, 2261–2271.

Okamoto, K., and Aoki, K. (1963). Development of a strain of spontaneously hypertensive rats. Jpn. Circ. J. *27*, 282–293.

Ozsolak, F., and Milos, P.M. (2011). RNA sequencing: advances, challenges and opportunities. Nat. Rev. Genet. *12*, 87–98.

Peng, M., Taouatas, N., Cappadona, S., van Breukelen, B., Mohammed, S., Scholten, A., and Heck, A.J.R. (2012). Protease bias in absolute protein quantitation. Nat. Methods *9*, 524–525.

Pickrell, J.K., Gilad, Y., and Pritchard, J.K. (2012). Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". Science *335*, 1302, author reply 1302.

Pravenec, M., and Kurtz, T.W. (2010). Recent advances in genetics of the spontaneously hypertensive rat. Curr. Hypertens. Rep. *12*, 5–9.

Printz, M.P., Jirout, M., Jaworski, R., Alemayehu, A., and Kren, V. (2003). Genetic Models in Applied Physiology. HXB/BXH rat recombinant inbred strain platform: a newly enhanced tool for cardiovascular, behavioral, and developmental genetics and genomics. J. Appl. Physiol. *94*, 2510–2522.

Renuse, S., Chaerkady, R., and Pandey, A. (2011). Proteogenomics. Proteomics *11*, 620–630.

Sandelin, A., Alkema, W., Engström, P., Wasserman, W.W., and Lenhard, B. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. *32* (Database issue), D91–D94.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342.

Simonis, M., Atanur, S.S., Linsen, S., Guryev, V., Ruzius, F.-P., Game, L., Lansu, N., de Bruijn, E., van Heesch, S., Jones, S.J.M., et al. (2012). Genetic basis of transcriptome differences between the founder strains of the rat HXB/BXH recombinant inbred panel. Genome Biol. *13*, r31.

Slavoff, S.A., Mitchell, A.J., Schwaid, A.G., Cabili, M.N., Ma, J., Levin, J.Z., Karger, A.D., Budnik, B.A., Rinn, J.L., and Saghatelian, A. (2013). Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat. Chem. Biol. *9*, 59–64.

Soon, W.W., Hariharan, M., and Snyder, M.P. (2013). High-throughput sequencing for biology and medicine. Mol. Syst. Biol. *9*, 640.

Starheim, K.K., Gevaert, K., and Arnesen, T. (2012). Protein N-terminal acetyltransferases: when the start matters. Trends Biochem. Sci. *37*, 152–161.

Su, Z.-D., Sun, L., Yu, D.-X., Li, R.-X., Li, H.-X., Yu, Z.-J., Sheng, Q.-H., Lin, X., Zeng, R., and Wu, J.-R. (2011). Quantitative detection of single amino acid polymorphisms by targeted proteomics. J. Mol. Cell Biol. *3*, 309–315.

Sun, A., Jiang, Y., Wang, X., Liu, Q., Zhong, F., He, Q., Guan, W., Li, H., Sun, Y., Shi, L., et al. (2010). Liverbase: a comprehensive view of human liver biology. J. Proteome Res. *9*, 50–58.

Swaney, D.L., Wenger, C.D., and Coon, J.J. (2010). Value of using multiple proteases for large-scale mass spectrometry-based proteomics. J. Proteome Res. *9*, 1323–1329.

Takeuchi, F., Isono, M., Katsuya, T., Yamamoto, K., Yokota, M., Sugiyama, T., Nabika, T., Fujioka, A., Ohnaka, K., Asano, H., et al. (2010). Blood pressure and hypertension are associated with 7 loci in the Japanese population. Circulation *121*, 2302–2309.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. *7*, 562–578.

Uhlen, M., and Ponten, F. (2005). Antibody-based proteomics for human tissue profiling. Mol. Cell. Proteomics *4*, 384–393.

Valentine, S.J., Sevugarajan, S., Kurulugama, R.T., Koeniger, S.L., Merenbloom, S.I., Bohrer, B.C., and Clemmer, D.E. (2006). Split-field drift tube/mass spectrometry and isotopic labeling techniques for determination of single amino acid polymorphisms. J. Proteome Res. *5*, 1879–1887.

Venter, E., Smith, R.D., and Payne, S.H. (2011). Proteogenomic analysis of bacteria and archaea: a 46 organism case study. PLoS ONE *6*, e27587.

Vizcaíno, J.A., Côté, R.G., Csordas, A., Dianes, J.A., Fabregat, A., Foster, J.M., Griss, J., Alpi, E., Birim, M., Contell, J., et al. (2013). The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. Nucleic Acids Res. *41* (Database issue), D1063–D1069.

Vogel, C., and Marcotte, E.M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat. Rev. Genet. *13*, 227–232.

Volkening, J.D., Bailey, D.J., Rose, C.M., Grimsrud, P.A., Howes-Podoll, M., Venkateshwaran, M., Westphall, M.S., Ané, J.-M., Coon, J.J., and Sussman, M.R. (2012). A proteogenomic survey of the Medicago truncatula genome. Mol. Cell. Proteomics *11*, 933–944.