nature biotechnology

© 2007 Nature Publishing Group http://www.nature.com/naturebiotechnology

A high-quality catalog of the *Drosophila melanogaster* proteome

Erich Brunner[1,10], Christian H Ahrens[1,2,10], Sonali Mohanty[1,3,10], Hansruedi Baetschmann[2], Sandra Loevenich[1,3], Frank Potthast[2], Eric W Deutsch[4], Christian Panse[2], Ulrik de Lichtenberg[5,6], Oliver Rinner[3], Hookeun Lee[3], Patrick G A Pedrioli[3], Johan Malmstrom[3], Katja Koehler[3], Sabine Schrimpf[1,7], Jeroen Krijgsveld[8], Floyd Kregenow[4], Albert J R Heck[8], Ernst Hafen[3], Ralph Schlapbach[2] & Ruedi Aebersold[3,9]

Understanding how proteins and their complex interaction networks convert the genomic information into a dynamic living organism is a fundamental challenge in biological sciences. As an important step towards understanding the systems biology of a complex eukaryote, we cataloged 63% of the predicted *Drosophila melanogaster* proteome by detecting 9,124 proteins from 498,000 redundant and 72,281 distinct peptide identifications. This unprecedented high proteome coverage for a complex eukaryote was achieved by combining sample diversity, multidimensional biochemical fractionation and analysis-driven experimentation feedback loops, whereby data collection is guided by statistical analysis of prior data. We show that high-quality proteomics data provide crucial information to amend genome annotation and to confirm many predicted gene models. We also present experimentally identified proteotypic peptides matching ~50% of *D. melanogaster* gene models. This library of proteotypic peptides should enable fast, targeted and quantitative proteomic studies to elucidate the systems biology of this model organism.

Understanding how, when and where genomic information is translated into proteins and how these molecules interact remains the primary challenge in systems biology. Despite the important conceptual and technical contributions of genome sequencing projects, the generation of accurate and complete models of physiological processes from genomic information alone seems unrealistic. In contrast to transcriptomics approaches that can measure the expression of all elements of a system, current proteomics technologies are incapable of analyzing complete proteomes, even of species with relatively simple genomes. Moreover, owing to a lack of experimental data and the limited accuracy of current gene prediction methods, uncertainty surrounds the validity of a substantial percentage of the protein sequences and their processed forms predicted from the genome alone[1,2]. These considerations underscore the importance of generating definitive proteomic catalogs for the functional analysis of biological systems.

We describe an extensive high-quality catalog of *D. melanogaster* proteins, based on 498,000 redundant and 72,281 distinct, fully tryptic peptide identifications (known and low cumulative false discovery rate of 1.37%) corresponding to 9,124 distinct proteins, or ~63% of the predicted fruitfly gene models. This level of proteome coverage, which to our knowledge has not been achieved for any other complex eukaryote, was achieved by maximizing sample diversity through analysis of different cell types and developmental states, applying multiple fractionation techniques to reduce sample complexity and a novel iterative feedback strategy that integrates the current results through detailed statistical analyses into the design of the next set of proteomics experiments.

We demonstrate the usefulness of this resource by confirming and correcting predicted fruitfly gene models and identifying a novel short gene. Furthermore, our data set identified a set of proteotypic peptides (PTPs) that matches to 6,980 proteins. PTPs are experimentally observed peptides that contain sufficient information to unambiguously identify a specific protein[3]. This unique PTP library forms the basis for developing a high-throughput, low-redundancy proteome screening approach based on targeted mass spectrometry. The accumulated data set will be made publicly available through PeptideAtlas[4].

RESULTS

An improved strategy to catalog proteomes

Although current transcriptomics technologies readily measure the expression level of all genes and splice variants in parallel, the generation of a comprehensive proteome catalog is much more

[1]Center for Model Organism Proteomes, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland. [2]Functional Genomics Center, ETH and University of Zurich, Winterthurerstrasse 190, Y55L70, 8057 Zurich, Switzerland. [3]Institute for Molecular Systems Biology, Swiss Federal Institute of Technology, ETH Honggerberg, Wolfgang-Pauli-Str. 16, HPT, CH-8093 Zurich, Switzerland. [4]Institute for Systems Biology, 1441 N. 34th Street, Seattle, Washington 98103, USA. [5]Center for Biological Sequence Analysis, BioCentrum-DTU, Technical, University of Denmark, Kemitorvet, Building 208, DK-2800 Lyngby, Denmark. [6]LEO Pharma, Industriparken 55, DK2750 Ballerup, Denmark. [7]Institute for Molecular Biology, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland. [8]Department of Biomolecular Mass Spectrometry, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Sorbonnelaan 16, 3584 CA Utrecht, the Netherlands. [9]Faculty of Science, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland. [10]These authors contributed equally to this work. Correspondence should be addressed to R.A. (aebersold@imsb.biol.ethz.ch), E.B. (erich.brunner@molbio.uzh.ch) or C.A. (christian.ahrens@fgcz.ethz.ch).

Received 17 December 2006; accepted 20 March 2007; published online 22 April 2007; doi:10.1038/nbt1300

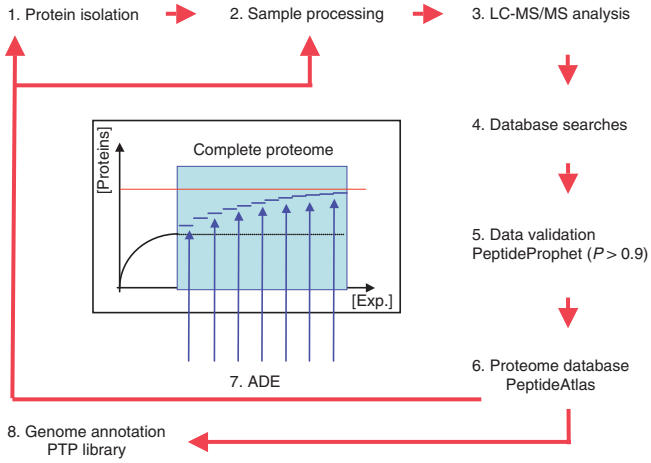

Figure 1 Description of the directed shotgun proteomics workflow. 1. Proteins were extracted from different tissues, cell lines and various developmental stages, enabling us to detect proteins of low abundance and with restricted expression during fruitfly development. 2. The inherent complexity of protein extracts was reduced using a combination of approaches, including subcellular fractionation and further biochemical separation of proteins and peptides by, for example, strong cation exchange chromatography (SCX) or free-flow electrophoresis in conjunction with the selective isolation of cysteine-containing peptides by the isotope coded affinity tag (ICAT)[24] technology (see **Supplementary Results** for details). 3. The samples are analyzed by reversed-phase chromatography-MS/MS in an automated fashion under standardized conditions. 4. The uninterpreted collision-induced dissociation (CID) spectra are searched against a standard *D. melanogaster* protein database using the TurboSequest algorithm. 5,6. Statistical filtering using PeptideProphet ($P > 0.9$) is used for data validation. The validated peptides are then integrated into the PeptideAtlas[4]. For data storage and initial data mining, we used an open source integrated database system called SBEAMS (systems biology experiment analysis management system)[4]. 7. ADE feedback strategy (see text): A detailed statistical analysis of the current data set allows one to identify areas of the proteome that are under-represented by current experimentation, and that can subsequently be targeted with an adapted experimentation protocol (steps 1,2). Repetitive cycling through this workflow (red arrows) will eventually lead to the annotation of a 'complete' proteome catalog. 8. Each cycle delivers new data that are used for genome annotation and for the extension of PTP libraries. ADE, analysis-driven experimentation.

© 2007 Nature Publishing Group http://www.nature.com/naturebiotechnology

demanding. Both the broad dynamic range of protein expression, spanning six to ten orders of magnitude[5], and the enormous complexity of protein extracts derived from biological samples represent serious obstacles. Therefore, the development of a suitable strategy to generate a comprehensive proteome catalog remains of high importance[6]. We started to catalog the *D. melanogaster* proteome using a tandem mass spectrometry (MS/MS)-based shotgun approach (**Fig. 1** and **Supplementary Table 1** online). For an initial set of 12 experiments, we prepared protein extracts from exponentially growing *D. melanogaster* S2 cells following an elaborate multidimensional fractionation scheme (**Supplementary Fig. 1**, **Supplementary Table 2** and **Supplementary Results** online) and identified a total of 5,795 proteins. To assess whether additional experiments could substantially increase the number of identified proteins, we plotted the total number of distinct protein identifications as a function of the number of overall high-quality peptide identifications for these 12 experiments (red dots, **Fig. 2**) and performed several Monte Carlo simulations. These simulations indicated (i) that a random shotgun approach is highly redundant and (ii) that more experimental trials could still identify additional proteins, in return, requiring an ever-increasing amount of effort (all simulated curves are saturation type curves) (**Fig. 2** and **Supplementary Results**). Similar saturation curves were observed for 4,455 proteins identified in the Kc167 *D. melanogaster* cell line and 3,955 proteins identified from *yw* third-instar larvae. Despite an increase in overall protein identifications, the three data sets showed a very substantial overlap, indicating that increasing sample diversification alone will not be sufficient to overcome the inherent redundancy of a shotgun approach. Therefore, a different strategy is needed that is capable of adding new protein identifications at a stage when a very substantial part of the proteome has already been identified.

Our approach to address this need, which we call analysis-driven experimentation (ADE), is an iterative feedback strategy that cycles between experimentation, thorough bioinformatics and statistical analysis, and redirected experimentation. This enables us to optimize experimental conditions and specifically target parts of a proteome that are under-represented in prior data sets. To accomplish this, several physicochemical and functional protein parameters (including

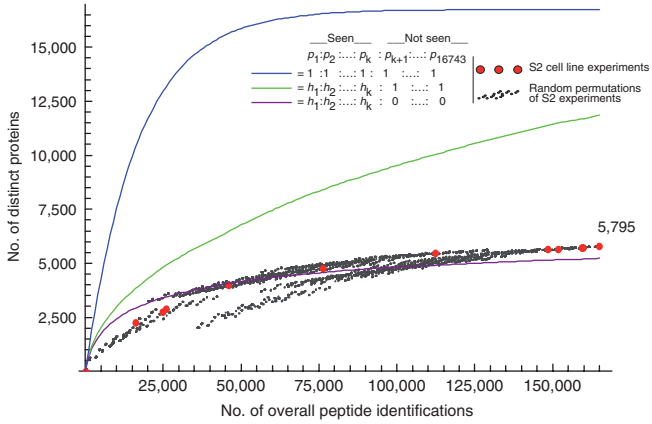

Figure 2 Count of distinct protein identifications as a function of overall high-quality peptide identifications and results of several Monte Carlo simulations. The red dots show the increase of distinct protein identifications for one specific order of 12 S2 cell line experiments, whereas the gray dots show this increase for 100 random permutations of the experiment order (**Supplementary Results**). Results of Monte Carlo simulations based on the assumption that the number of identifications of each protein in n identification trials are multinomially distributed with parameters n, and $p_1,.,p_{16743}$ with p_i = probability of identification of protein i. (i) Blue curve: the probability of identifying a protein is the same for all proteins ($p_{1:\ldots:}\ p_{16743} = 1$). For the purple and green curves the probabilities are based on the overall empirical frequencies of protein identifications in the 12 experiments (h_i = no. of identifications of protein i in all S2 experiments, $h_1,.,h_k > 0$ for $k = 5{,}795$ proteins seen). The probabilities for proteins not identified by the 12 experiments were calculated based on either assuming an overall number of identifications (ii) of 1 for each of these proteins (green curve, $h_{k+1}{:}\ldots{:}\ h_{16743} = 1$) or (iii) of 0 (purple curve, $h_{k+1}{:}\ldots{:}\ h_{16743} = 0$) (**Supplementary Results**). In the case of the green curve, all proteins can be identified, but this needs many more identification trials than for the blue curve. The simulations (purple curve) increase the evidence that at least a subset of those proteins, which was not identified in the 12 experiments, has $P > 0$ of being identified.

© 2007 Nature Publishing Group http://www.nature.com/naturebiotechnology

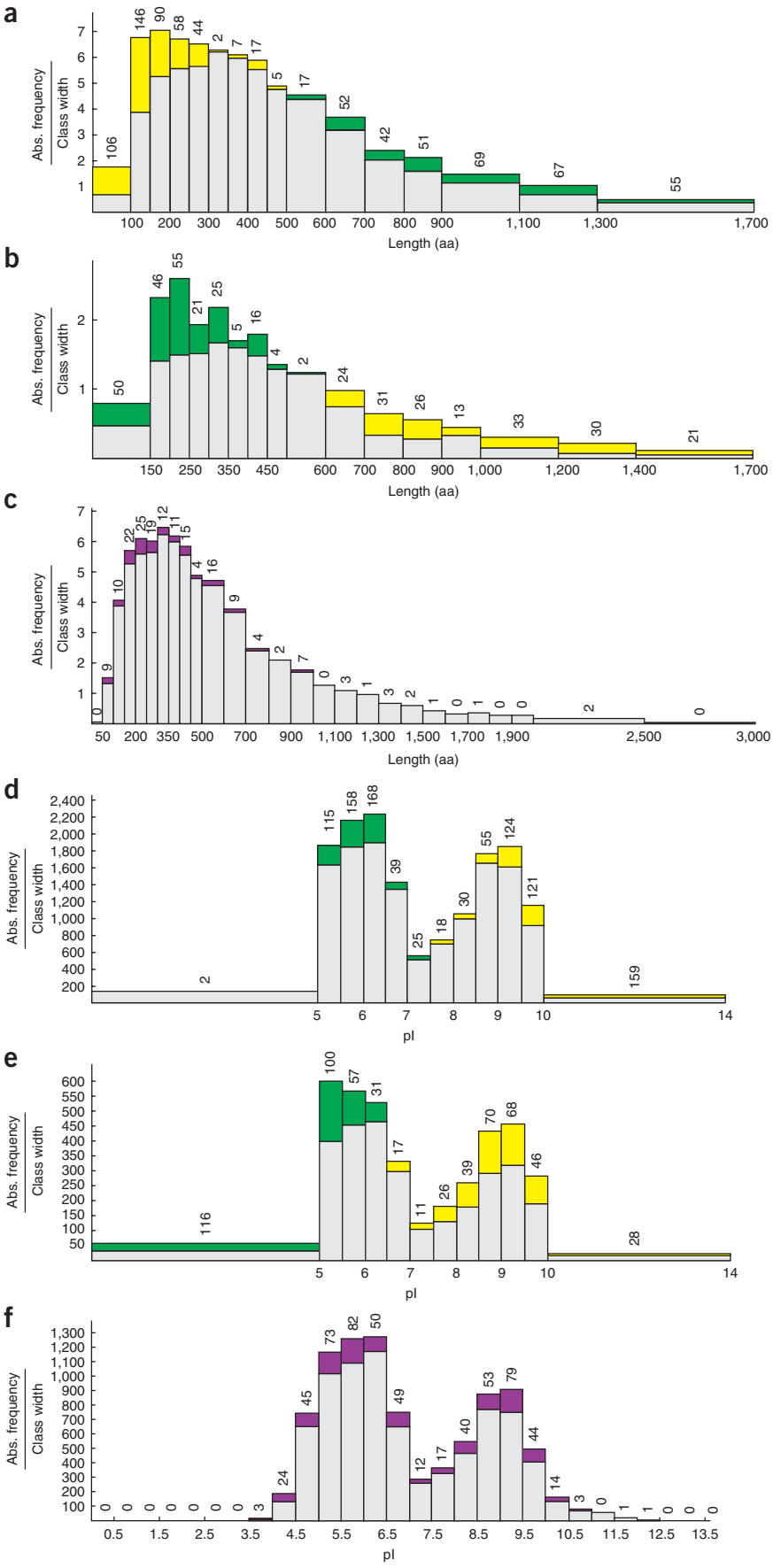

Figure 3 Benefits of the ADE strategy visualized by combined protein parameter histograms and difference histograms. (**a**) Comparison of length distribution of proteins identified in five Kc167 cell line experiments versus all proteins (**Supplementary Fig. 2**; under-represented areas are shown in yellow, over-represented areas in green). The numbers indicate how many proteins are under- or over-represented, respectively. (**b**) 1,194 proteins identified by gel-filtration experiments versus 4,455 proteins identified in five previous experiments. (**c**) Protein length class association of newly identified proteins (purple box) by the gel filtration experiment. The values indicate absolute numbers of added identifications. (**d**) Comparison of isoelectric point (pI) distribution of 7,998 proteins identified from various experiments versus all proteins. (**e**) 1,960 larval proteins identified by free-flow electrophoresis pH 4–7 versus all proteins. (**f**) Isoelectric point class association of newly identified larval proteins (purple box) by the protein free-flow electrophoresis experiment compared to the proteins identified by the undirected larval experiments (3,955). The values indicate absolute numbers of added identifications.

length, isoelectric point, favored codon frequency (FCF) as a measure for protein abundance[7], transmembrane domains and signal peptides) were computed for all distinct proteins. Next, the distribution of these parameters among experimentally identified proteins was compared to that for all proteins to identify biases of our shotgun approach (**Supplementary Fig. 2** and **Supplementary Results** online). Statistical analysis revealed a significant under-representation of short proteins and basic proteins (**Fig. 3**) in the data set ($P < 10^{-10}$). Next, experiments were designed to specifically target the determined classes of both under- and over-represented proteins (**Fig. 3a–f** and **Supplementary Results**). Using this targeted approach, we could enrich for under-represented short proteins by using gel filtration (**Fig. 3b**,**c**) and identify additional acidic proteins (which are over-represented) by using protein free-flow electrophoresis experiments in the pH range 4–7 (**Fig. 3e–f**). For the significantly under-represented low-abundance proteins (**Fig. 4a**) and unstable proteins (**Fig. 4b**) ($P < 10^{-10}$), which are less amenable to specific targeting by current experimentation, new experimental strategies have to be designed.

The improvement of proteome coverage through focusing on parts that are under- or over-represented, especially at a stage where several thousand proteins have already been identified, establishes the validity of the ADE concept. ADE will thus be an essential component in our further analyses of the *D. melanogaster* proteome (see Discussion),

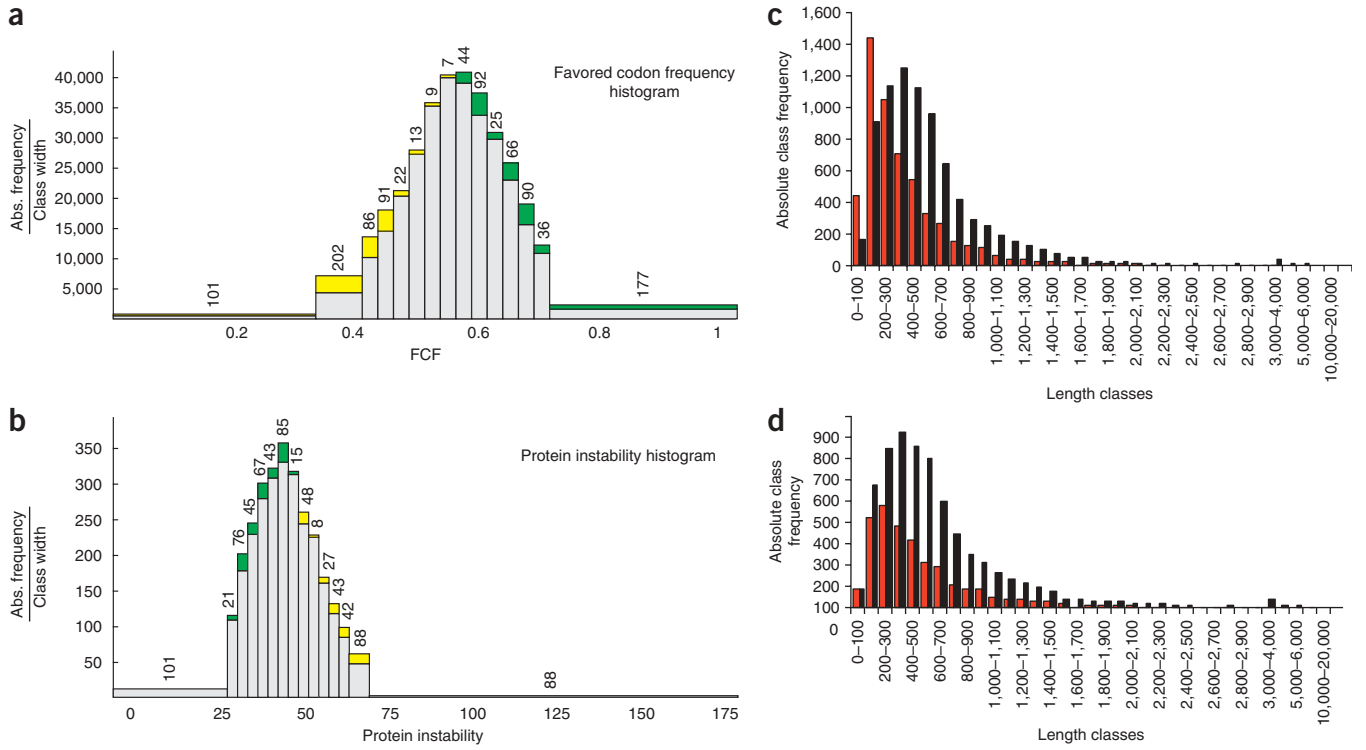

Figure 4 Additional biases among the experimentally identified proteins. (**a**,**b**) Distributions of protein parameters of experimentally identified proteins (9,124) versus all proteins (16,743) for FCF (**a**) and protein instability (**b**) based on dipeptide composition. Low FCF values indicate low-abundance proteins. Proteins with instability scores (>40) are typically unstable (ExPASy, ProtParam tool). Under-represented areas are shown in yellow, over-represented areas in green. The numbers indicate how many proteins are under- or over-represented, respectively. (**c**,**d**) Gene Ontology (GO) comparison of gene models according to the average length of their encoded proteins. Gene models having at least one GO assignment (GO+) in the database are represented by black bars; gene models lacking a GO assignment (GO⁻) are represented by red bars. Comparison of the GO distributions of all gene models in the database (13,792) (**c**) versus the distribution of GO classes of the experimentally observed 8,672 gene models (**d**) reveal a statistically highly significant under-representation of GO⁻ gene models encoding for proteins in the length classes below 500 aa (**Supplementary Table 3f** and **Supplementary Fig. 2**).

© 2007 Nature Publishing Group http://www.nature.com/naturebiotechnology

and we expect it to be a very efficient strategy for minimizing the high levels of redundancy in similar shotgun proteomics projects.

Bioinformatic analysis of the proteome catalog

Using sample diversity and diverse biochemical fractionation schemes in combination with the iterative ADE feedback strategy, we present what is, to our knowledge, the largest high-quality shotgun proteomics data set to date. Forty-three experiments from four developmental stages (embryo, larva, pupa, adult) and two cell lines (Kc167, S2) of *D. melanogaster* provided evidence for the expression of 9,124 proteins (**Supplementary Table 2** online), which are encoded by 8,672 distinct gene models (62.9% of all gene models). From ~10 million MS/MS sequencing attempts acquired in close to 1,700 liquid chromatography (LC)-MS/MS runs, almost 500,000 redundant and 72,281 distinct peptide sequences could be assigned with a PeptideProphet probability value of at least 0.9 (ref. 8) (**Supplementary Results**). On average, eight peptides were identified per protein. This data set is publicly available through PeptideAtlas[4] (http://www.mop.unizh.ch/peptideatlas) and FlyBase (http://www.flybase.org/). The PeptideAtlas information for the identified proteins includes the genome sequence coordinates of the corresponding peptides, along with experimental information.

To assess whether our approach identified low-abundance proteins, we compared FCF parameter distributions of the experimentally observed proteins versus those of all proteins[7]. Despite a clear bias toward identifying proteins with higher FCF values, the data set comprises a substantial portion of low-abundance proteins (**Fig. 4a**): 351 of the 1,000 proteins with the lowest FCF values, compared to 725 of the 1,000 proteins with the highest FCF values, have been identified. In addition, the protein instability histogram of experimentally seen versus all proteins shows that we predominantly identify stable proteins (**Fig. 4b**). Furthermore, of the 9,124 experimentally identified proteins, 1,833 (20.1%) have one or more transmembrane domains (predicted using TMHMM2.0)[9], and 1,020 (11.2%) are secreted proteins (predicted using SignalP3.0)[10]. These values are close to the 22.5% transmembrane and 12.4% secreted proteins predicted for the entire set of 16,743 distinct protein sequences. Thus, when combined with adequate multidimensional fractionation strategies, a shotgun approach can identify a very substantial percentage of low-abundance, membrane and secreted proteins, which represents a major advance compared to two-dimensional gel-based approaches.

A first-pass analysis of functional protein domains using the Pfam classification (**Supplementary Table 3a**,**b** online) revealed that ~67% of all distinguishable fly proteins and 75% of the experimentally identified proteins contain a Pfam domain, indicating that we preferentially identified proteins with described functional domains. Under-represented Pfam domain families primarily include domains found in transcription factors and plasma membrane proteins. This finding correlates with the above-reported under-representation of

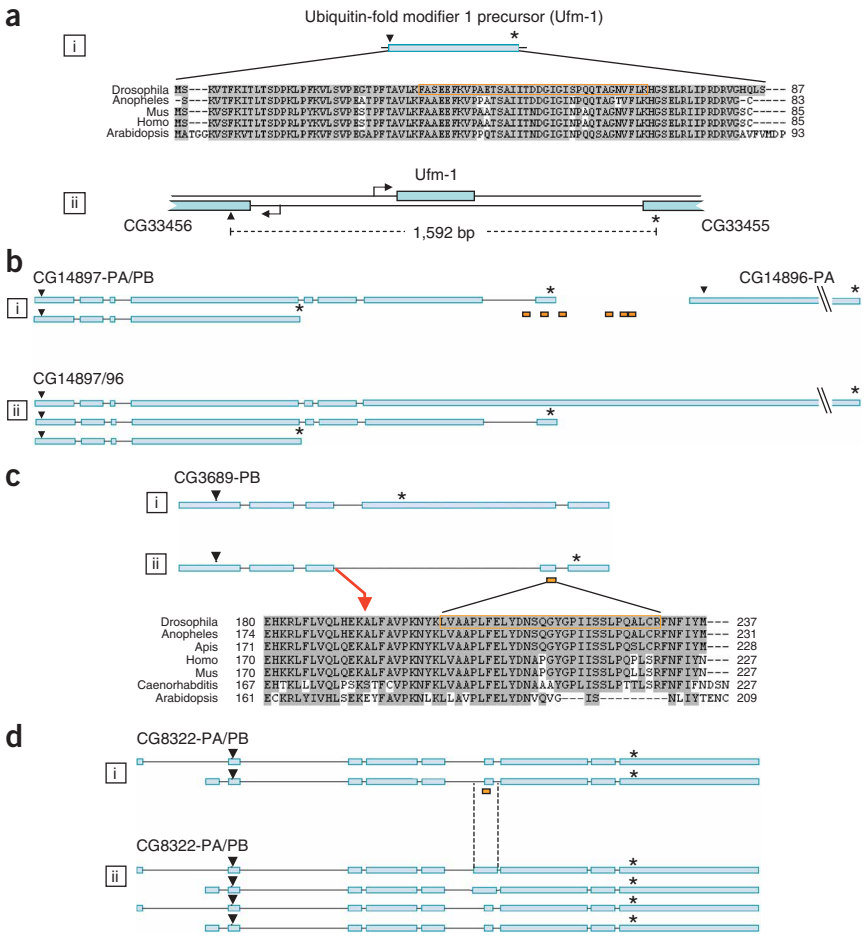

Figure 5 Identification of new genes/exons using proteomics data. (**a**) (i) Peptide-based identification (orange-boxed sequence) of a novel ORF of 87 aa that is highly conserved from plants to human. (ii) The *D. melanogaster* homolog of the human *UFM1* maps within a stretch of 1,592 base pairs between the genes CG33455 (translational stop) and CG33456 (translation initiation). The exact positions of the promoters (arrows) are unknown. The map is not drawn to scale. (**b**) (i) Overprediction of splice sites in a very large exon (>2,600 aa in length) leads to a frameshift and consequently to the premature stop of the predicted protein sequence (CG14897). The remaining part of this exon contained a favorable translation initiation site, and an ORF of 1,298 aa was assigned to a distinct gene model (CG14896). Both ORFs have been identified by various peptides. (ii) Our cross-comparative database searches identified six distinct peptides (orange boxes) that flank the last exon of the first gene model (CG14897). Closer inspection of the genomic locus revealed that this large exon not only includes the three novel peptides but also interconnects the two predicted ORFs fusing them to a single gene model. The longest protein sequence derived from this fused gene has a length of 3,441 aa and is not only anticipated by the Genescan algorithm but has also been computationally predicted[25]. (**c**) (i) Schematic drawing of the predicted gene CG3689 (a homolog of the nudix hydrolase family in mouse) of which only one splice variant is annotated, which is supported by expressed sequence tags. However, the currently annotated *D. melanogaster* protein lacks the highly conserved C terminus (ii). This stretch of 45 aa is present in homologous proteins of other insects, nematodes, plants and mammals. The conservation of the current proteins from the fruitfly compared to other species terminates at the end of coding exon 3. Identification of a peptide that maps to the 3′ region of CG3689 uncovered this highly conserved domain in the fruitfly. A predicted splice site within the last coding exon of CG3689 results in a frameshift of the ORF and consequently perfectly incorporates the missing conserved C-terminal domain into the coding sequence (red arrow). (**d**) (i) **Figure 4d** documents a peptide that partially overlaps with the fifth coding exon present in both protein isoforms encoded by the gene CG8322. A blastp search using this peptide identified a complete match in protein LD21334p that was identified in a high-throughput EST sequencing project[26]. LD21334p itself appeared to be almost identical to CG8322, differing only in the fifth exon. This exon is extended at both its 5′ and 3′ ends and includes in the case of LD21334p the entire peptide sequence. We therefore conclude that CG8322 encodes two different proteins rather than one. ▲, translational start codons; *, translational stop codons. The original gene models are taken from FlyBase. Identified peptides are represented as orange boxes. Identified peptide sequences are boxed. Identical as well as conserved amino acid residues are highlighted in gray.

© 2007 Nature Publishing Group http://www.nature.com/naturebiotechnology

basic proteins and of proteins with transmembrane domains. Conversely, we see an over-representation of domains found in enzymes and cytoplasmic proteins.

Gene Ontology (GO) analysis enables one to group gene sets into coherent functional classes, allowing interpretation of results in a biological context[11]. We performed a GO analysis on a reduced set of GO categories (GO-slim) on all predicted *D. melanogaster* gene models and the experimentally identified gene models (**Supplementary Table 3c–e** online). The experimentally identified gene models were over-represented in the 'known' GO categories (GO^+), confirming the results of the Pfam analysis. Conversely, under-represented gene model annotations included molecular function categories such as receptor activity, ion channel activity and transcription factor activity.

Finally, we performed a statistical analysis of the gene models according to average protein length and GO classification using a χ^2 test. All gene models were subdivided into those with (GO^+) or without a GO annotation (GO^-) and their length distribution was compared with that of all experimentally identified gene models. This comparison was performed within defined-length classes, thereby making it independent of the shotgun-induced length bias. It revealed a statistically significant (*P*-values ranging from 2 × 10E-34 to 3 × 10E-2) under-representation of GO^- gene models encoding proteins in the length classes of <500 amino acids (aa; **Fig. 4c,d** and **Supplementary Table 3f** online). As the shotgun approach itself should be unbiased for the presence or absence of GO annotations within a given length class, we suggest that the strong under-representation of the GO^- class in the observed gene models likely indicates a substantial proportion of them to be a result of incorrect gene predictions.

Proteomics-based genome annotation

The proteomics data set described here provides experimental confirmation for many of the computer-predicted protein sequences. Furthermore, the data set contains additional information in the form of high-quality uninterpreted spectra[12] that can be used to

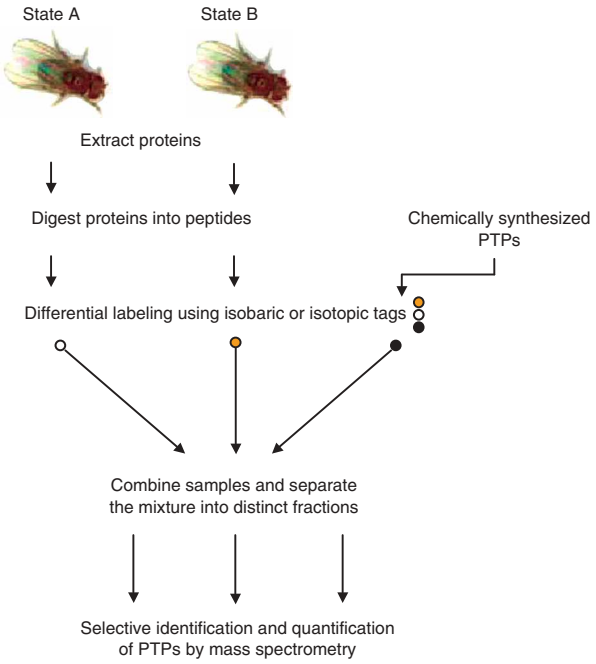

Figure 6 Outline of a typical PTP experiment. Proteins of two samples to be compared (for example, flies or cell lines) are extracted and digested with trypsin. The samples are differentially labeled using isotopic or isobaric tags and then combined. The resulting peptide mixture is subsequently spiked with PTPs of known amount that are tagged with a third label. The known amount of PTPs allows for an absolute quantification of the proteins in a complex mixture. The mixture is further separated into subfractions and every fraction is selectively scored for the presence of specific PTPs (proteins) by mass spectrometry. Quantification of these PTPs is achieved by the determination of isotope ratios. The PTPs required for such experiments are generated by chemical synthesis or expressed in bacteria as an array of PTPs that can be released from the polypeptide chain by a tryptic digest. The PTPs that are specific for the proteins of interest are selected from the PTP library: Most of the *D. melanogaster* proteins can be identified using at least one unique tryptic peptide. However, not all sequences are suited for proteomic approaches, and thus a library of PTPs needs to be defined within the total pool of unique sequences by computational or in most cases by experimental means. Critical constraints for their definition are the enzyme selected for proteolytic cleavage, the dynamic range of the used mass spectrometer as well as the absence of post-translational modifications on the mature peptide. This last criterion is of great importance because post-translational modifications are often reversible (dynamic) and thus again introduce ambiguity into the system.

© 2007 Nature Publishing Group http://www.nature.com/naturebiotechnology

identify novel genes or exons. To identify so-far-unannotated protein sequences, we performed cross-comparative database searches for seven representative large experiments (**Supplementary Table 2** and **Supplementary Results**), which identify peptides that suggest the presence of an open reading-frame (ORF) in what was predicted to be a noncoding region (**Supplementary Table 4** online). Two classes of novel coding regions could be identified. The first group uncovered novel genes, an example of which is depicted in **Figure 5a**. Database searches established this previously unannotated ORF as the fruitfly homolog of the human ubiquitin-fold modifier-1-precursor (*UFM1*), the product of which is ligated to target proteins in a manner similar to ubiquitin[13]. It belongs to a growing family of conserved, very short ubiquitin-like proteins, which regulate a wide range of cellular functions, including cell-cycle progression, DNA repair and apoptosis.

The second group of peptides map to genomic regions of previously annotated genes. Although all coding regions implied by these peptides were missed in previous annotations due to erroneous or incomplete splice-site predictions, these novel coding regions do not uniformly define alternatively spliced isoforms, but instead constitute three subclasses. Incorrect or incomplete splice-site predictions may lead to a gene fusion (**Fig. 5b**), identification of missed splice variants (**Fig. 5c**) or extensions of existing exons (**Fig. 5d**).

These examples illustrate the shortcomings of current gene prediction algorithms and underscore the importance and usefulness of a comprehensive proteome catalog for genome annotation. Extrapolating from the pilot study results, we expect that the whole proteomic data set will contribute considerably to the annotation of the *D. melanogaster* genome. However, as for the gene models described above, this will require manual validation, confirmation and inclusion into FlyBase. This protracted, manual process is in progress in collaboration with professional FlyBase curators, and validated corrected or new gene models will be added to the public domain.

PTP library for targeted, low-redundancy proteome analysis

The shotgun approach has become widely used, especially for comparative and quantitative proteomics experiments[14–16]. However, besides the many advantages of this approach, such as high throughput, minimized sample handling and identification of low-abundance proteins, there are two inherent problems. First, the loss of association between a protein and the peptides generated after tryptic digestion poses substantial computational challenges to assign peptide sequences back to their cognate protein(s)[12]. Second, the shotgun strategy is not designed to selectively collect information for specific proteins. Although present in a sample, a protein may remain undetected owing to undersampling effects in the mass spectrometer[3]. One way to circumvent this problem is to reduce sample complexity. However, considering the dynamic range of proteins in a sample, a more suitable method would be to selectively detect specific proteins. A promising concept that addresses these two main shortcomings of the shotgun method relies on so-called PTPs. Within a single experiment, information on multiple proteins can be selectively retrieved and where appropriate, protein levels can be determined by absolute means (**Fig. 6**). Our experiments have uncovered PTPs for 6,980 *D. melanogaster* proteins derived from >50% of the current gene models. These PTPs, which are 6–55 aa long and generated by tryptic digestion, are in the measurable range of most mass spectrometers. To generate the list of experimentally observed PTPs, all 321,297 unique *in silico* fully tryptic peptides were aligned with the 72,281 distinct peptides identified by MS/MS. This generated 37,279 observed PTPs (**Supplementary Table 5** online). The number of observed PTPs per protein ranged from 1 to a total of 127 PTPs (on average five PTPs per protein). Of the latter, identified for the retinoid- and fatty acid–binding glycoprotein CG11064-PA, a subset was preferentially identified.

Our *D. melanogaster* PTP library will enable researchers to selectively analyze and quantify a high percentage of fruitfly proteins in non-gel-based comparative proteomics experiments without the need for further technology development (**Fig. 6**).

DISCUSSION

We used a directed, large-scale shotgun proteomics approach to generate a high-quality data set that enabled identification and confirmation of >60% of all fruitfly gene models. Elements essential to achieving this coverage were (i) the selection of diverse biological samples spanning the main developmental stages of *D. melanogaster*, (ii) a combination of multiple fractionation techniques at the cellular,

© 2007 Nature Publishing Group http://www.nature.com/naturebiotechnology

protein and peptide levels to reduce sample complexity and (iii) a feedback strategy that uses statistical analysis to integrate all results into design of subsequent proteomic experiments. Whereas brute-force, undirected, repeat analyses lead to saturation at a relatively low level, the directed analysis of complementary samples substantially increases proteome coverage. Although the ADE strategy is capable of precisely indicating under-represented or missed protein classes, subsequent experiments will cover only part of the targeted protein fraction. The pool of missed proteins will become smaller with time, but likewise, the yield of new peptides and proteins per experiment will decrease. This alternating feedback process of thorough data analysis and experimental redesign promises to enable extensive proteome coverage much faster than undirected approaches. The endpoint of full proteome coverage is difficult to define. Whereas in an average comparative gene expression experiment involving *D. melanogaster* the expression of roughly one-third of the gene models[17] can be detected, genome tiling arrays detected transcription from 93% of all annotated genes[18] in a set of diverse samples. It remains to be established whether the undetected genes are expressed at levels below the threshold of detection or whether they represent incorrect predictions. Although the fruitfly proteome catalog is not yet complete, it already contains sufficient information to address some of these issues. We demonstrated the usefulness of proteomics for confirming and improving genome annotation. Proteomics-based confirmation of many hypothetical proteins, amendment of existing gene models and identification of new genes are likely to lead to the development of improved gene prediction algorithms and eventually to an accurate and complete list of gene models.

The directed shotgun proteomics approach holds great promise for pharmaceutical applications: ~20% of the identified proteins are transmembrane proteins and ~10% are secreted. Both protein classes constitute the predominant targets for drug discovery. In particular, focusing analysis on PTPs enables selective targeting and quantification of proteins in very complex samples such as human serum. PTPs confer greater reliability on these protein identifications and, by enabling rapid screening of biological samples for biomarkers[19], hold immense potential for diagnostics. Sharing proteomics data sets and developing sophisticated algorithms for PTP prediction[3,20] will increase and eventually complete these libraries. This will open the field for global profiling of proteomes for which a platform has already been designed.

Our data set also provides an excellent resource for method development, to explore the temporal and spatial expression of proteins, and to identify overlaps and differences with large scale gene expression studies. Specific experiments to detect the type and exact position of post-translational modifications, such as sites of glycosylation[21] or phosphorylation[15,22], will supplement the continuously growing proteome catalog. Similar proteome annotation projects will supplement genome annotation, affect technology development and be fundamental to realizing the goals of the emerging field of systems biology.

METHODS

Mass spectrometry. Nanoflow-LC-MS/MS was performed by coupling an UltiMate high-performance liquid chromatography (HPLC) system (LC-Packings/Dionex) in-line with a Probot (LC-Packings/Dionex) autosampler system and an LTQ ion trap (Thermo Electron). Samples were automatically injected into a 10-µl sample loop and loaded onto an analytical column (9 cm × 75 µm; packed in-house with Magic C18 AQ beads 5 µm, 100 Å (Microm)). Peptide mixtures were delivered to the analytical column at a flow rate of 300 nl/min of buffer A (5% acetonitrile, 0.2% formic acid) for 25 min and then eluted using a gradient of acetonitrile (10–45%; 0.5%/min) in 0.2% formic acid. Peptide ions were detected in a survey scan from 400 to 2,000 a.m.u. (1–2 µscans) followed by 3–6 data-dependent MS/MS scans (3 µscans each, isolation width 2 a.m.u., dynamic exclusion list 250, dynamic exclusion time 240 s) in an automated fashion. Alternatively, in a few cases an LTQ-FT or LCQ mass spectrometer was used (**Supplementary Table 2**). The analytical column was regenerated for 20 min with buffer A at 300 nl/min before loading the next sample. Alternative measurements: nanoflow-LC-MS/MS was performed by coupling an Agilent 1100 HPLC (Agilent Technologies), operated as described[23], to a 7-Tesla LTQ-FT mass spectrometer (Thermo) or an LTQ ion trap (Thermo Electron). For peptide LC, trapping columns (1 cm × 100 µm) and analytical columns (15 cm × 50 µm) were packed in-house with ReproSil-Pur C18- AQ, 3 µm (Dr. Maisch GmbH). Peptide mixtures were delivered at 3 µl/min on the trapping column for desalting. After flow-splitting down to ~150 nl/min, peptides were transferred to the analytical column and eluted in a gradient of acetonitrile (1%/min) in 0.1 M acetic acid. The eluate was sprayed via emitter tips (New Objective), butt-connected to the analytical column. Mass spectrometers were operated in data-dependent mode, automatically switching between MS and MS/MS acquisition for the three most abundant peaks in a given MS spectrum. In the LTQ-FT, full-scan MS spectra were acquired in the Fourier transform–ion cyclotron resonance (FT-ICR) at a target value of 5×10^6 with a resolution of 20,000. The three most intense ions were then isolated for accurate mass measurements by an FT-ICR–selected ion monitoring scan, which consisted of 10-Da mass range, at a resolution of 50,000. These ions were then fragmented in the linear ion trap. In the LTQ, MS scans were recorded in centroid mode at a target value of 30,000. Peptides were fragmented when the signal exceeded 2×10^4 counts by filling the ion trap at a target value of 10,000 with a maximum ion time of 100 ms.

Pfam analysis. The set of 16,743 distinct *D. melanogaster* protein sequences from the Berkeley Drosophila Genome Project BDGP3.2 database (**Supplementary Table 1**) was searched against the Pfam database of Hidden Markov Models (release 20, May 2006, 8,296 protein family models) using the HMMER software package (hmmer2.3.2, http://hmmer.janelia.org/). We carried out a stringent analysis and relied on the individual domain family gathering thresholds, as recommended in the HMMER user manual. Result files were parsed and the respective number of hits per Pfam domain were recorded for all proteins and for the subset of 9,124 experimentally identified proteins. Hypergeometric tests (Fisher's exact test) were used to assess the statistical significance of the over- or under-representation of certain Pfam domain families among the experimentally identified proteins compared to all distinct proteins. Only Pfam domains that have ten or more hits among all proteins (348 Pfam domain families) and that score $P < 0.01$ are shown in **Supplementary Table 3**. The expected number of false positives for this analysis in random data would be 3.5 (348 Pfam domain families with ten or more hits among all proteins, multiplied with the P value cutoff of 0.01). This indicates that 3–4 of the 62 Pfam families (listed in **Supplementary Table 3a**,**b**) could be in this list just by chance, as a result of multiple testing.

GO-slim analysis. GO-slim terms for the categories Cellular Component, Molecular Function, and Biological Process were mapped to all gene models (BDGP 3.2, 13,792 gene models) and to the 8,672 experimentally identified gene models using the GO Term mapper tool (http://go.princeton.edu/cgi-bin/GOTermMapper). This tool uses the map2slim.pl script (Chris Mungall, BDGP) to bin the submitted gene lists to a static set of broader, high-level parent generic GO-slim terms. The counts of the respective categories among all gene models and the experimentally identified gene models were calculated. Similar to the Pfam analysis, the statistical significance of an under- or over-representation of a certain GO-slim category was assessed with a hypergeometric test (Fisher's exact test, **Supplementary Table 3**). The expected number of false positives for this analysis in random data would be ~1.2 (119 GO-slim categories, multiplied with the P value cutoff of 0.01), indicating that one out of the significantly over- or under-represented categories could be among the significant categories just by chance. The entire list of the GO-slim categories (including those that are not statistically significant) is shown in **Supplementary Table 3c**,**e**.

Cross-comparative searches for genome annotation. A representative subset of ~1.3 Mio. MS/MS spectra (13% of total) were searched both against our standard protein database (BDGP release 3.2) and against a

six frame–translated genome sequence database (BDGP release 4.2), generating two TurboSequest outputs for every spectrum. In principle, the sequence of a peptide that is identified with a high PeptideProphet probability score in the genomic database search, but a low score in the protein database ($P < 0.5$) search could represent an interesting hit. These hits could be explained if the respective protein sequence entry is missing in the protein database, either through erroneous or incomplete gene prediction, or if an entire protein category (e.g., small ORFs) is not included in the database.

Generation of the *D. melanogaster* PTP library. The large amount of experimental data allowed us to generate a comprehensive PTP library for roughly 50% of all *D. melanogaster* gene models (**Supplementary Table 5**, or publicly available through PeptideAtlas), comprising only experimentally confirmed and statistically validated peptides (PeptideProphet $P > 0.9$). For *D. melanogaster*, we have calculated a total number of 321,297 candidate (putative) PTPs, assuming a complete cleavage by trypsin and setting a length constraint for the peptides of 6–55 aa, reflecting the experimentally determined dynamic range of the tandem mass spectrometer (unique peptide assignments including missed cleavage sites were excluded from the list). To generate the list of experimentally confirmed PTPs, the complete set of candidate PTPs was aligned with the 72,281 distinct peptides that have been identified by MS/MS. This resulted in a total of 37,279 observed PTPs (**Supplementary Table 5**). The number of PTPs per protein ranged from 1 to a total of 127 PTPs of which usually a subset was preferentially identified.

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

ACKNOWLEDGMENTS

We thank Bernd Roschitzki, Bertran Gerrits, Eva Niederer, Marko Jovanovic, Cristian Köpfli and Michael Walser for technical help, Hans Jespersen and Soeren Schandorff from Proxeon Bioinformatics for discussions regarding the proteotypic peptide data analysis and Hubert K. Rehrauer for help with statistical analysis. The project was funded by the University Research Priority Program Systems Biology/Functional Genomics of the University of Zurich. E.B., S.M., S.S. and S.L. are members of the Center for Model Organism Proteomes (C-MOP) which is funded by the University of Zurich (http://www.mop.unizh.ch). S.L. was supported by a Career Development Award of the University of Zurich. This work was also supported in part by a UBS grant to E.B. and K. Basler, and with federal funds from the US National Heart, Lung, and Blood Institute, National Institutes of Health under contract No. N01-HV-28179.

AUTHOR CONTRIBUTIONS

E.B. and S.M. conducted most of the experimental work; E.B. performed most of the LTQ measurements, coordinated the interdisciplinary project and carried out the proteomics-based genome annotation work. S.M. performed the GO analyses; C.H.A. coordinated and carried out the bioinformatic analyses of the data set (Pfam analysis with C.P.), and conceived the ADE strategy; H.B. established the necessary statistical infrastructure for the project, carried out statistical analyses together with C.H.A. and performed the simulations; S.L. implemented the SBEAMS database, generated the *Drosophila* Peptide Atlas (supported by E.W.D.) and supported E.B with the proteomics-based genome annotation work; F.P. and C.P. (supported by E.W.D.) implemented and maintained the computational infrastructure at the Functional Genomics Center Zurich (FGCZ); U.L. provided the protein parameter computations; O.R. performed the gel-filtration experiments; H.L. helped with the setup of the LTQ instrument and LTQ MS analyses; P.G.A.P. generated the software for FCF calculations; J.M. set up the Free Flow Electrophoresis system and helped with experiments; K.K. and S.S. helped with selected experiments; F.K. helped with LCQ, and the initial experimental strategy; J.K. shared large data sets and performed the LTQ and LTQ-FT-ICR measurements in the lab of A.J.R.H.; R.S. supported the project with FGCZ resources; E.H. and R.A initiated the project and provided intellectual and financial support; R.A carried senior authorship responsibility.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

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

1. Pennisi, E. Searching for the genome's second code. *Science* **306**, 632–635 (2004).
2. Tupy, J.L. *et al.* Identification of putative noncoding polyadenylated transcripts in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* **102**, 5495–5500 (2005).
3. Kuster, B., Schirle, M., Mallick, P. & Aebersold, R. Scoring proteomes with proteotypic peptide probes. *Nat. Rev. Mol. Cell Biol.* **6**, 577–583 (2005).
4. Desiere, F. *et al.* Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.* **6**, R9 (2005).
5. Anderson, N.L. & Anderson, N.G. The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* **1**, 845–867 (2002).
6. de Godoy, L.M. *et al.* Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. *Genome Biol.* **7**, R50 (2006).
7. Duret, L. & Mouchiroud, D. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* **96**, 4482–4487 (1999).
8. 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).
9. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E.L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).
10. Bendtsen, J.D., Nielsen, H., von Heijne, G. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **340**, 783–795 (2004).
11. Giot, L. *et al.* A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727–1736 (2003).
12. Nesvizhskii, A.I. & Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteomics* **4**, 1419–1440 (2005).
13. Komatsu, M. *et al.* A novel protein-conjugating system for Ufm1, a ubiquitin-fold modifier. *EMBO J.* **23**, 1977–1986 (2004).
14. Washburn, M.P., Wolters, D. & Yates, J.R., III. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242–247 (2001).
15. Tao, W.A. *et al.* Quantitative phosphoproteome analysis using a dendrimer conjugation chemistry and tandem mass spectrometry. *Nat. Methods* **2**, 591–598 (2005).
16. Ong, S.E. & Mann, M. Mass spectrometry-based proteomics turns quantitative. *Nat. Chem. Biol.* **1**, 252–262 (2005).
17. Johansson, K.C., Metzendorf, C. & Soderhall, K. Microarray analysis of immune challenged *Drosophila* hemocytes. *Exp. Cell Res.* **305**, 145–155 (2005).
18. Stolc, V. *et al.* A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science* **306**, 655–660 (2004).
19. Pan, S. *et al.* High throughput proteome screening for biomarker detection. *Mol. Cell. Proteomics* **4**, 182–190 (2005).
20. Mallick, P. *et al.* Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.* **25**, 125–131 (2007).
21. Zhang, H. *et al.* High-throughput quantitative analysis of serum proteins using glycopeptide capture and liquid chromatography mass spectrometry. *Mol. Cell. Proteomics* **4**, 144–155 (2005).
22. Corthals, G.L., Aebersold, R., Goodlett, D.R. & Burlingame, A.L. Mass spectrometry: modified proteins and glycoconjugates. in *Methods in Enzymology* Vol. 405 (ed. Burlingame, A.L.) 66–81, (Academic Press, Boston, 2005).
23. Krijgsveld, J. *et al.* Metabolic labeling of *C. elegans* and *D. melanogaster* for quantitative proteomics. *Nat. Biotechnol.* **21**, 927–931 (2003).
24. Gygi, S.P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17**, 994–999 (1999).
25. Gopal, S. *et al.* Homology-based annotation yields 1,042 new candidate genes in the *Drosophila melanogaster* genome. *Nat. Genet.* **27**, 337–340 (2001).
26. Rubin, G.M. *et al.* A *Drosophila* complementary DNA resource. *Science* **287**, 2222–2224 (2000).

© 2007 Nature Publishing Group http://www.nature.com/naturebiotechnology