Deep Coverage of the *Escherichia coli* Proteome Enables the Assessment of False Discovery Rates in Simple Proteogenomic Experiments*[S]

Karsten Krug‡, Alejandro Carpy‡, Gesa Behrends‡, Katarina Matic‡, Nelson C. Soares‡, and Boris Macek‡§

Recent advances in mass spectrometry (MS) have led to increased applications of shotgun proteomics to the refinement of genome annotation. The typical “proteogenomic” workflows rely on the mapping of peptide MS/MS spectra onto databases derived via six-frame translation of the genome sequence. These databases contain a large proportion of spurious protein sequences which make the statistical confidence of the resulting peptide spectrum matches difficult to assess. Here we performed a comprehensive analysis of the *Escherichia coli* proteome using LTQ-Orbitrap MS and mapped the corresponding MS/MS spectra onto a six-frame translation of the *E. coli* genome. We hypothesized that the protein-coding part of the *E. coli* genome approaches complete annotation and that the majority of six frame-specific (novel) peptide spectrum matches can be considered as false positive identifications. We confirm our hypothesis by showing that the posterior error probability distribution of novel hits is almost identical to that of reversed (decoy) hits; this enables us to estimate the sensitivity, specificity, accuracy, and false discovery rate in a typical bacterial proteo-genomic dataset. We use two complementary computational frameworks for processing and statistical assessment of MS/MS data: MaxQuant and Trans-Proteomic Pipeline. We show that MaxQuant achieves a more sensitive six-frame database search with an acceptable false discovery rate and is therefore well suited for global genome reannotation applications, whereas the Trans-Proteomic Pipeline achieves higher specificity and is well suited for high-confidence validation. The use of a small and well-annotated bacterial genome enables us to address genome coverage achieved in state-of-the-art bacterial proteomics: identified peptide sequences mapped to all expressed *E. coli* proteins but covered 31.7% of the protein-coding genome sequence. Our results show that false discovery rates can be substantially underestimated even in “simple” proteo-genomic experiments obtained by means of high-accuracy MS and point to the necessity of further improvements concerning the coverage of peptide sequences by MS-based methods. Molecular & Cellular Proteomics 12: 10.1074/mcp.M113.029165, 3420–3430, 2013.

MS-based proteomics has become an indispensable tool for studying *in vivo* protein expression on a global scale (1). Briefly, in a typical “shotgun” proteomic experiment, the whole proteome of an organism is extracted and digested by a protease (e.g. trypsin). The resulting complex peptide mixtures are usually further fractionated and separated via liquid chromatography (LC) before ionization and analysis in the mass spectrometer. Recent innovations in MS technology (2–4) enable high peptide sequencing rates with high mass accuracy and sensitivity, placing the routine analysis of entire proteomes within reach (5, 6).

Modern genome annotation uses computational *ab initio* approaches to predict coding regions and gene models from raw sequencing data (7, 8). As the ultimate evidence of gene expression is the detection of its product, transcriptomic data are commonly used to train gene prediction algorithms (9). Similarly, MS-based proteomics is increasingly used in genome annotation. In a typical proteo-genomics experiment, MS/MS spectra of peptides are searched against databases derived via *in silico* six-frame translation of the whole genome sequence (10–14). This approach has been applied, alone or in combination with transcriptomic data, in order to refine genome annotation in several organisms, including *C. elegans* (15), *P. pacificus* (16), *S. cerevisiae* (17), *S. pombe* (18), *A. thaliana* (19), *S. nodorum* (20), *T. gondii* (21), *A. gambiae* (22), mouse (23), and human (24, 25). Bacteria are especially well suited for MS-assisted genome annotation because of their relatively simple genome structures and small genome sizes, which lead to overall better sequence coverage in a typical proteomics experiment (26–33).

The use of six-frame databases in proteo-genomics experiments is challenging because of their large sizes, which
increase the search space as well as affect the sensitivity of database searches (34). Additionally, these databases contain a high proportion of artificial sequences resulting from frames that are not transcribed (13, 35). These spurious protein sequences are difficult to discriminate from the true protein sequences, which makes the statistical confidence of the resulting peptide spectrum matches (PSMs)1 difficult to calculate.

Here we take advantage of the small size (4.6 Mb), simple architecture, and high annotation level of the *Escherichia coli* genome and use it as a benchmark model for proteo-genomic data interpretation. We derive a comprehensive dataset of proteins expressed in the exponential growth of *Escherichia coli* and map the corresponding MS/MS spectra onto a six-frame translation of the *E. coli* genome. We hypothesize that the protein-coding part of the *E. coli* genome approaches complete annotation, and we consider six frame-specific (novel) PSMs as wrongly identified. This enables us to estimate the factual false discovery rate in a simple proteo-genomic experiment. We show that the posterior error probability (PEP) distribution of novel peptides is almost identical to that of decoy (reversed) hits, which validates our assumption only 31.7% of the protein-coding genome sequence.

The minimal medium containing 0.25 mg/ml lysine. The cell cultures were grown at 37 °C under vigorous shaking to an *A*~600~ of 1.9, then 1 ml of the stationary culture was spun down at 260 × g for 10 min in order to remove any remaining from the Luria/Miller medium. The bacterial cells were washed twice with M9 minimal medium consisting of M9 salts (6.78 g/l Na2HPO4, 3 g/l KH2PO4, 0.5 g/l NaCl, 1 g/l NH4Cl, Sigma-Aldrich) supplemented with additional 0.5% (w/v) glucose, 33 μM thiamine, 1 mM MgSO4, 0.1 mM CaCl2. Next, the resultant pellet was resuspended in a final volume of 1 ml M9. Immediately after, 5 μl of this culture were used to inoculate 5 ml of fresh M9 medium containing 0.25 mg/ml of lysine (Sigma-Aldrich). Overnight, minimal medium cell cultures were grown at 37 °C under vigorous shaking at an *A*~600~ of 0.5 and used to inoculate (1:100 dilution) 125 ml of fresh M9 medium containing 0.25 mg/ml of lysine. The cell cultures were grown to an *A*~600~ of 0.5, harvested via centrifugation at 3345 × g for 10 min. The crude protein extract was methanol/chloroform precipitated, and the protein precipitates were redissolved in denaturation buffer containing 6 mM urea/2 M thiouria in 10 mM Tris buffer. For estimation of the protein concentration, each extract was measured via Bradford assay (Bio-Rad).

**SDS-PAGE and In-gel Digestion**—In-gel digestion was performed as previously described (16). Briefly, extracted proteins were separated on a NuPage Bis-Tris 4–12% gradient gel (Invitrogen). The gel was stained with Coomassie Blue and subsequently cut into 15 slices. The resulting gel pieces were destained by being washed three times with 10 mM ammonium bicarbonate (ABC) and acetonitrile (ACN) (1:1, v/v). Proteins were then reduced with 10 mM dithiothreitol (DTT) in 20 mM ABC for 45 min at 56 °C and alkylated with 55 mM iodoacetamide in 20 mM ABC for 30 min at room temperature in the dark. After being washed two times with 5 mM ABC and one time with ACN, the gel pieces were dehydrated in a vacuum centrifuge. Proteins were digested with either trypsin (Promega Fitchburg, WI) or Lys-C (Wako Neuss, Germany) (12.5 ng/μl in 20 mM ABC) at 37 °C overnight. The resulting peptides were extracted in three subsequent steps with the following solutions: (i) 3% TFA in 30% ACN, (ii) 0.5% acetic acid in 100% ACN, and (iii) 100% ACN. After evaporation of the ACN in a vacuum centrifuge, peptide fractions were desalted using Stage-Tips (38).

**In-solution Digestion**—Protein extracts were reduced for 1 h at room temperature with 1 mM DTT and subsequently alkylated with 1 mM iodoacetamide for 1 h at room temperature in the dark. Proteins were pre-digested with Lys-C (1:100 w/w) for 3 h at room temperature. After dilution with 4 volumes of 20 mM ABC, proteins were digested overnight at room temperature with either trypsin (1:100 w/w) or Lys-C (1:100 w/w).

**Off-gel Isoelectric Focusing**—Peptides derived from the in-solution digestion were separated according to their isoelectric point using the 3100 OFFGel fractionator (Agilent Santa Clara, CA) following the manufacturer’s instructions. Peptide mixtures were separated into 12 fractions using 13-cm Immobiline DryStrips with a pH 3–10 gradient (GE Healthcare). Separation was performed at a maximum current of 50 μA until 50 kWh were reached. Peptide fractions were acidified with acidic solution (30% ACN, 5% acetic acid, and 10% TFA in water) and desalted using Stage-Tips.

**Strong Anion Exchange Chromatography**—Peptides from the in-solution digestion were desalted using solid phase extraction. Strong anion exchange chromatography was performed as described elsewhere (39). Briefly, desalted peptides were loaded at pH 11 onto an anion exchange column containing six layers of Empore/Disk Anion Exchange (3m, St. Paul MN) in a 200–μl pipette tip. For conditioning and elution, Britton & Robinson Universal Buffer (0.02 M CH3COOH, 0.02 M H2PO4, and 0.02 M H3BO3) at pH 3, 4, 5, 6, 8, and 11 was prepared. The column was activated with methanol and conditioned with 1 M NaOH followed by buffer (pH 11). The flow-through was acidified with acidic solution and loaded on a Stage-Tips. Peptides were eluted at pH 8, 6, 5, 4, and 3, acidified with acidic solution, and desalted using Stage-Tips.

**Nano-LC-MS/MS Analysis**—All peptide fractions were measured on an EASY-nLC II nano-LC (Proxeon Biosystems Odense, Denmark) coupled to an Orbitrap Velos mass spectrometer (Thermo Fisher Scientific). Chromatographic separation was done on a 15-cm Pic-oTip fused silica emitter with an inner diameter of 75 μm and an 8-μm
tip inner diameter (New Objective Woburn, MA) packed in-house with reversed-phase ReproSil-Pur C18-AQ 3-µm resin (Dr. Maisch GmbH Ammerbuch-Entringen, Germany). Peptides were injected into the column with solvent A (0.5% acetic acid) at 700 nl/min using a maximum pressure of 280 bar. Peptides were then eluted using an 81-min or a 221-min segmented gradient of 5%–50% solvent B (80% ACN in 0.5% acetic acid) at a flow rate of 2000 nl/min. The mass spectrometer was operated in data-dependent mode. Survey full scans for the MS spectra were recorded between 300 and 2000 Thompson at a resolution of 60,000 with a target value of 1E6 charges. The 15 most intense peaks from the survey scans were selected for fragmentation with collision-induced dissociation at a target value of 5000 charges. The fragment spectra were recorded in the linear ion trap. Selected masses were included in a dynamic exclusion list for 90 s.

**MS Data Processing—** Acquired MS data were preprocessed by MaxQuant (v.1.2.2.9) (40) in order to generate peak lists that could be submitted to a database search. Derived peak lists were submitted to the Andromeda (41) and Mascot v2.2.0 (Matrix Science, London, UK) search engines to query the genome database translated into all six reading frames. The genome sequence of *E. coli* (42, 43) was downloaded from the NCBI homepage (accession number NC 390913.2). The translation into all six reading frames was done from stop codon to stop codon by applying the bacterial and plant plasmid code (translation Table XI) using the Transseq tool that is part of the Emboss software package (44). We required a minimal length of six amino acids for each resulting putative open reading frame (ORF), which corresponds to the minimal peptide length that we required in the database search. To that database we added decoy sequences using the SequenceReverse.exe tool shipped with MaxQuant software. The resulting database consisted of 263,159 putative ORFs, 248 commonly observed lab contaminants, and 263,407 reversed sequences.

A database search was performed with the precursor mass tolerance set to 6 and 7 ppm for Andromeda and Mascot database searches, respectively. The fragment ion mass tolerance was set to 0.5 Da for both search engines. Full enzyme specificity for trypsin and Lys-C was required, and up to two missed cleavages were allowed. Oxidation of methionine and protein N-terminal acetylation were defined as variable modifications, and carbamidomethylation of cysteine was defined as a fixed modification.

The resulting lists of PSMs were further processed by MaxQuant and Trans-Proteomic Pipeline (v4.5 RAPTURE rev 0) (45). Andromeda database scores calculated by MaxQuant were converted to PEPs as described in Ref. 41. We calculated q-values by sorting the PSMs by their PEPs in ascending order. For each PSM we calculated the ratio between the number of decoy hits and the number of target PSMs having PEPs below the PEP of the actual PSM. Mascot result (.dat) files were converted to pepXML format and further processed by the PeptideProphet (45) module as part of the Trans-Proteomic Pipeline (TPP). We used the accurate mass binning option, excluded singly charged peptides, and used decoy hits to model the score distribution of false positives for semi-supervised mixture modeling. The false discovery rate (FDR) was controlled by filtering PSMs according to the probability assigned by PeptideProphet. The corresponding probability threshold was calculated by the calc TPPstat.pl perl script as part of the TPP, and the “Approx. P threshold for FDR” was used to filter the list of PSMs.

Acquired MS data were additionally searched against a recent annotation of the *E. coli* genome (UniProt reference proteome set; downloaded on January 18, 2012; 4309 protein entries) using MaxQuant v1.2.2.9 operating with the same database search parameters as described above. FDRs on peptide and protein group levels were set at 1%.

**Proteo-genomic Workflow—** Detected peptide sequences that resulted from searching the six-frame database were matched to the UniProt *E. coli* proteome database using BLASTP (Blast 2.2.25 +) (46, 47) to check whether they mapped to annotated proteins. We chose UniProt as the database because it offers a comprehensive and unified resource of protein sequences and all 4309 *E. coli* proteins are part of the Swiss-Prot section, which is manually annotated and reviewed. Therefore, this database should represent a high-quality annotation of the theoretical *E. coli* proteome. All peptides that produced a perfect match in the UniProt *E. coli* database were considered as annotated. In order to retrieve the genomic coordinates of detected peptides, we mapped their sequences to the genome database using TBLASTN. Because the BLAST algorithms are not optimized to find all occurrences of small sequences, we set the maximal E-value to 1E-4 and the number of alignments to 20 in order to ensure that the typically short peptide sequences could be found in the genome and proteome databases. To map these peptides unambiguously, we required a full-length alignment and 100% similarity. Multiple occurrences of the same peptide in the genome or proteome were considered separately. All peptides that did not produce a perfect match in the proteome database were defined as initial candidates in the list of novel peptides. To address the ambiguity of leucine and isoleucine, the initial set of novel peptides was checked once again using regular expression matching. Peptide sequences that could not be found in the proteome database because of any isobaric amino acids were removed from the initial set of novel peptides. In a second BLAST iteration, all six-frame ORFs that were detected by one or more novel peptides were matched to the proteome database as well as the non-redundant protein database (NCBI nr). In addition, we resubmitted the spectra of novel peptides to query the NCBI nr database using the Mascot search engine in order to check the consistency between PSMs derived from searching the six-frame translation and NCBI nr database. Together with the genome coordinates of the peptides and the annotated proteins, we used this information to classify the novel peptides into different types of annotation conflicts.

The proteo-genomic pipeline and further downstream data analysis were implemented in R v2.13.1 (48).

**Calculation of Protein Abundances—** We implemented the exponentially modified protein abundance index (PAI) (49) to estimate protein abundance. Briefly, the exponentially modified PAI is defined as $10^{PAI - 1}$, with the PAI (50) being the number of observed peptides divided by the number of observable peptides per protein. For the calculation of exponentially modified PAI values for our dataset, we focused on the tryptic part of our dataset, which comprised all detected proteins. The mass range used to define observable peptides was set from 600 Da to 6000 Da.

**Functional Annotation Analysis—** Gene Ontology annotation of the *E. coli* K12 proteome was derived from the Gene Ontology Annotation (UniProt-GOA) database (51), downloaded on February 29, 2012. We applied a two-sided hypergeometrical test to see whether specific annotation terms were significantly enriched or depleted among the set of proteins of interest. Derived p values were further adjusted to address multiple hypothesis testing using the methods proposed by Benjamini and Hochberg (52). The adjusted p values were -log10 transformed and visualized by the function heatmap.2 that is part of the R package gplots.

**RESULTS**

We derived a comprehensive dataset of *E. coli* proteins by harvesting the cells in the exponential phase of growth, extracting the proteome, and applying three separation methods (strong anion exchange chromatography, off-gel isoelectric focusing, and gel-based LC-MS) in combination with protein digestion using two proteases, trypsin and Lys-C. We ana-
MS/MS spectra and peptide sequences identified after searching of a six-frame translation of E. coli chromosome using two MS data processing workflows

| Workflow   | MS/MS spectra identified (A) | MS/MS spectra identified (%) | Peptide sequences | Novel peptides | Decoy peptides | Lab contaminant peptides | Annotated E. coli proteins |
|------------|-----------------------------|-----------------------------|------------------|---------------|---------------|------------------------|----------------------------|
| MaxQuant   | 370,231                     | 19.1                        | 33,964           | 263           | 336           | 306                    | 2653                       |
| TPP        | 162,028                     | 8.3                         | 25,724           | 59            | 0             | 209                    | 2524                       |

False Positive Identifications Accumulate among Novel Peptide Hits—We investigated whether identified peptide sequences were present in the annotated protein-coding portion of the genome. Peptides that could not be assigned to any annotated protein in the UniProt E. coli database were referred to as six-frame-specific or novel peptides. Assuming that the protein-coding part of the E. coli genome approaches complete and correct annotation, these peptides can be considered as false positives and can be used to assess the performance of the applied proteo-genomic search strategies. In order to validate this hypothesis, we processed the MS data without any control of the FDR and classified the resulting peptide sequences according to whether they were annotated (“target,” 44,872 hits), reversed hits (“decoy,” 35,370 hits), or novel peptides (“novel,” 31,075 hits) (Fig. 1A). The PEP values of peptides that could be assigned to annotated proteins followed a very tight distribution around a median PEP of 3.26e-6, indicating that a high percentage were true positive identifications. Notably, the absolute numbers of decoy and novel hits were very similar, and the corresponding PEP values followed almost the same distribution, with median PEP values of 0.790 (decoy) and 0.787 (novel) (Fig. 1B). The PEP distribution of the novel hits resembled that of PSMs to a search space that contained a small fraction of correct hits, which pointed to the fact that very few true positive hits were expected to be found among the novel hits in E. coli (i.e. the genome is almost completely annotated). We postulated that if there were still many novel genes/peptides to discover, the proportion of correct hits among the novel peptides would be larger and the PEP distribution of novel and decoy hits would be significantly different. To demonstrate this, we considered a hypothetical case of partial genome annotation in which 20% or 50% of E. coli genes were unknown (i.e. not annotated). We then randomly sampled a subset (80% and 50%, respectively) of the proteins contained in the UniProt E. coli database to define the corresponding “partially annotated” E. coli proteome. We then re-classified the detected peptide sequences into “target” and “novel” (as shown in Fig. 1A) according to the newly defined E. coli “annotation” and visualized their PEP distributions. As expected, we obtained significantly different PEP distributions of novel hits and decoy hits (supplemental Fig. S3). Application of the TPP workflow further confirmed these results (supplemental Fig. S4). Taken together, these findings support our initial hypothesis of almost complete genome annotation of the protein-coding
part of the *E. coli* genome and point to the fact that the vast majority of novel peptides are in fact false positive hits. These peptides can be used as decoy hits to estimate an FDR according to the TDA. This does not mean that the novel peptides do not contain some true positives (see below); however, their number is negligible compared with the overall number of PSMs and will not significantly affect the overall outcome of our analysis. We calculated q-values (57) for all peptides that could be assigned to the existing protein database using the novel peptides as decoy hits and correlated the calculated values to “standard” q-values (Fig. 1C). Overall there was a very high correlation ($r = 0.99997$) of q-values calculated based on decoy peptides (x-axis) and novel peptides (y-axis). The high correlation between the two distributions of q-values shown in Fig. 1C pointed to the underestimation of target-decoy FDRs, as the novel peptides were not considered in the FDR estimation.

In total, 313 peptide sequences passing the default constraint of 1% FDR were specifically found in the genomic six-frame translation in a target-decoy database design. The PEP distributions of novel and decoy peptides were nearly identical, but they differed markedly from the PEP distribution of all target peptides present in the UniProt proteome database of *E. coli*. B, quantile-quantile plot of PEPs resulting from decoy (x-axis) and novel peptides (y-axis) demonstrating nearly identical PEP distributions of novel and decoy peptides. C, correlation of estimated false discovery rates (q-values) derived from decoy (x-axis) and novel peptide sequences (y-axis). Colors correspond to PSM populations shown in A.

**Fig. 1.** Distribution of PSM confidence scores. A, distributions of posterior error probabilities (PEPs) of different PSM populations that resulted from searching a genomic six-frame translation in a target-decoy database design. The PEP distributions of novel and decoy peptides were nearly identical, but they differed markedly from the PEP distribution of all target peptides present in the UniProt proteome database of *E. coli*. B, quantile-quantile plot of PEPs resulting from decoy (x-axis) and novel peptides (y-axis) demonstrating nearly identical PEP distributions of novel and decoy peptides. C, correlation of estimated false discovery rates (q-values) derived from decoy (x-axis) and novel peptide sequences (y-axis). Colors correspond to PSM populations shown in A.
four possible contingencies (true positive, false positive, false negative, and true negative) of this comparison (supplemental Fig. S8B). Based on the derived contingency tables, we assessed sensitivity, specificity, accuracy, and the factual FDR as a function of the FDR utilized by both approaches (Fig. 3). To assess FDR_{fact}, we used the number of obvious false positive identifications (decoy plus novel peptides) as an estimate for the number of latent false positives among the list of all detected peptide sequences (true positives plus false positives), equivalent to the TDA. Both workflows demonstrated

**TABLE II**

Novel peptide sequences identified in both data processing workflows; accession numbers shown in brackets correspond to proteins of another organism or *E. coli* strain

| ID | Peptide sequence | Annotation conflict in *E. coli* | Remarks | UniProt accession number(s) |
|----|------------------|----------------------------------|---------|-----------------------------|
| A  | VGSESWWQSK       | Erroneous initiation (upstream peptide) | Known conflict | P13039 |
| B  | INQTSAMPEK       | Erroneous initiation (spanning peptide) | Known conflict | P32695 |
| C  | LAMPSGNQEP      | Erroneous initiation (spanning peptide) | Correct in *E. coli* O157:H7 | P0CB62 (Q8X3T3) |
| D  | MMQTVLAK         | Erroneous initiation (spanning peptide) | Known conflict (in *E. coli* O157:H7) | P00909 (Q8X7B7) |
| E  | CSEFGEAIENM     | Point mutation/sequencing error or deamidated version | Present in other *E. coli* strains (e.g. MS 116-1, DH1) | P13039 (D8AHK0) (E6P3Y4) |
| F  | GVALHAVK         | Not present | Present in *E. coli* strain MS 117-3 | (E9THR2) |
| G  | SLYSIALIR        | Not present | 78% similarity to a sequence in *Selaginella moellendorfii* | (D8RY37) |
| H  | GLSGPASQATVAAP   | Not present | Clear/false positive | |
| I  | LSIRIQPPK        | Not present | Unclear | |

**Fig. 2.** True positive and likely false positive examples from among the highest scoring novel peptides. A, schematic representation of the erroneous initiation of the *fes* gene. Annotated proteins are shown in blue, detected peptides are depicted in black, and six-frame ORFs are shown in green. ORFs that were hit by peptides are shown in dark green. The novel peptide (VGSESWQSK) was located upstream of the predicted protein N terminus. The corresponding six-frame ORF encompassed the complete sequence and employed the same reading frame as the *fes* gene. B, corresponding MS/MS spectrum of the novel peptide depicted in A annotated with a comprehensive series of b and y fragment ions. C, schematic representation of a dubious novel peptide identified at a 1% FDR by both data processing workflows used in this study. Although an adjacent cluster of peptides was detected that mapped to the *tref* gene, the novel peptide (LSIRIQPPK) utilized a different reading frame. D, MS/MS spectrum of the corresponding novel peptide shown in C poorly annotated with b and y fragment ions.
high specificity and accuracy, which are essential for discriminating false positive from true positive identifications. The sensitivity of the TPP-based workflow was consistently lower (on average 42.3% lower) than that of the MaxQuant workflow across different FDR thresholds. Strikingly, the factual FDR as a function of the decoy FDR utilized by MaxQuant increased linearly with a constant ratio FDRfact/FDRdecoy of about 3.5 in our particular study, pointing to the FDR underestimation that occurs when using the TDA. Conversely, the FDRfact did not approach the probability-based FDR used in the TPP workflow, confirming the conservative character of the MMA. We expect that these properties of the MMA and the TDA will be similar in other proteo-genomics datasets of similar size and complexity and even more pronounced in proteo-genomic analyses of larger and more complex genomes.

**The Expressed Proteome of E. coli in Exponential Growth Phase**—The dataset derived in this study represents one of the most comprehensive proteomics datasets of *E. coli*. In order to assess proteome coverage, we searched the acquired MS spectra against the UniProt proteome database using MaxQuant operating with default parameters. Resubmission of the 1.9 m spectra to the Andromeda search engine identified 42,780 non-redundant peptide sequences (supplemental Table S4) corresponding to 2626 distinct *E. coli* proteins (supplemental Tables S5 and S6) with an FDR of 1% at the protein level. A detailed summary of all sub-datasets concerning the different fractionation methods and the two enzymes used can be found in supplemental Table S7. Although 2626 proteins represent about 61% of the annotated proteome, a dataset of similar size was reported before (36). Therefore, we estimate that *E. coli* grown in batch culture under aerobic conditions does not express more than about 2700 proteins and conclude that our dataset approached full coverage of the *E. coli* proteome. A Gene Ontology (59) term enrichment/depletion analysis of the expressed proteome revealed an underrepresentation of functions related to motion (e.g. flagellum organization (p = 3.45e-7), motor activity (p = 4.14e-6)) and transposons (e.g. transposition (p = 2.33e-16), transposase activity (p = 7.19e-11)). Therefore, these functions characterize the part of the *E. coli* proteome that we did not identify in our dataset. Further details about the functional analysis of the expressed proteome can be found in supplemental Figs. S9C and S9D and supplemental Tables S8–S12, respectively.

**Coverage of the *E. coli* Genome with Identified Peptides**—This comprehensive dataset enabled us to address general features of bacterial proteomics experiments, especially in the context of the coverage of the genome sequence by detected peptides. We first defined the protein-coding part of the genome by mapping the 4309 proteins present in the UniProt *E. coli* database onto the chromosome (4.6 Mb) (Fig. 4A). This analysis revealed that 86.8% (4.0 Mb) of the genome is annotated in the protein database and therefore protein coding. We next used sequences of all 2626 proteins identified in our dataset to estimate the size of the expressed part of the genome, which corresponded to 65.4% (2.6 Mb) of protein-coding genome regions. Finally, mapping of the detected peptide sequences onto the chromosome captured 1.27 Mb of the raw genome sequence, matching 31.7% of the protein-coding part of the genome (Fig. 4B). The number of MS/MS events with which each nucleotide was represented ranged from 1 to 1344 with an average coverage of 20 MS/MS and median number of 7 MS/MS events per nucleotide (Fig. 4C).
The performance of different search strategies in proteogenomic applications was explored in a number of previous studies. For example, the application of TDA to searches of protein databases derived via six-frame translation has been assessed recently (35), and previous reports point to important general considerations for the application of this approach in database searches (60, 61). There is a global consensus that the increased size of the databases obtained via six-frame translation decreases the sensitivity and specificity of database searches and that the spurious protein sequences present in such databases make the statistical confidence of the resulting PSMs difficult to assess (13, 62). To circumvent this issue, we hypothesized that the annotation of the protein-coding part of the \textit{E. coli} genome approaches completeness. If correct, this would enable us to consider novel PSMs as false positive identifications and to assess general features of a typical bacterial proteo-genomic data-set, such as sensitivity, specificity, accuracy, and factual false discovery rate. Our hypothesis was confirmed by almost identical distributions of the PEP values of the novel and decoy hits, as well as by the low number of detected novel peptides. We note that several of the detected novel peptides were true positive hits, but because of their extremely low number we expect their effect on the reported values to be minimal.

We used two complementary MS/MS data processing frameworks: MaxQuant implementing the target-decoy approach, and Trans-Proteomic Pipeline/Peptide Prophet using the mixture model approach for FDR assessment. Although both achieved deep proteome coverage, the MaxQuant-based workflow identified a significantly greater number of peptides, whereas the TPP-based workflow had a significantly lower factual FDR. However, TPP also led to decreased sensitivity, resulting in a smaller number of identified spectra, which can significantly affect the coverage of a genome sequence by detected peptides. In our view, the MaxQuant workflow led to a better tradeoff between maximal peptide identification rates (sensitivity) desired in proteo-genomic studies and FDR and is better suited for global genome re-annotation studies; in contrast, the MMA and semi-supervised model are better suited for applications that require high specificity (e.g. the detection of splice variants or single nucleotide variation (SNVs)).

Somewhat surprisingly, our data point to a substantial underestimation of the FDR even in “simple” proteo-genomic experiments utilizing high-accuracy mass spectrometry. Although we cannot perform the same assessment for proteogenomics experiments in higher organisms (because of the lower quality of their annotation), we expect that these issues will be even more pronounced due to the size and complexity of their genomes (e.g. effects of alternative splicing). Therefore, special care should be taken to control the search space and calculate the FDR as accurately as possible. Several strategies for decreasing the search space in proteo-genomic databases have been proposed (63); we argue that the use of high accuracy is one of the most effective ways to achieve this. In this context, the use of “high-high” acquisition methods (ones in which the survey and MS/MS scans are acquired at high (ppm to sub-ppm) accuracy) (2, 3) will further improve the confidence of detected PSMs and become indispensable in future proteo-genomic experiments. These acquisition strategies still come with a prolonged duty cycle (i.e. lower number of acquired MS/MS events) on most of the currently used MS platforms, but recent advances in several high-accuracy analyzers (64) will soon enable the routine acquisition of high-accuracy MS data at all levels. However, novel
peptides detected in such experiments should still undergo thorough investigation before being treated as true positive identifications, regardless of the acquisition method or proteo-genomic workflow used.

The comprehensive proteome dataset derived for the purpose of this study enabled us to assess another important aspect of a proteo-genomics experiment: coverage of the genome sequence by identified peptide sequences. The field of proteomics is getting to the remarkable stage of the identification and quantification of all gene products expressed under specific conditions, and this especially applies to organisms with small and relatively simple genomes, such as bacteria and yeast (5). In addition to the detection of a gene product, genome re-annotation also requires high coverage of the genome sequence. In our study, we achieved comprehensive detection of the expressed E. coli proteome that was in agreement with previous studies (36); however, the identified peptide sequences covered 48.5% of the estimated expressed, 31.7% of all protein-coding, and only 27.5% of the total genome sequence. Interestingly, in the part of the E. coli genome covered by peptide sequences, each nucleotide was detected by an average of 20 MS/MS scans (median: 7 MS/MS scans), which corresponds to 20-fold base coverage in genomic terminology. As next-generation sequencing studies routinely achieve up to 50-fold base coverage of 99.9% of the genome sequence (65, 66), our results demonstrate the limitation of using MS-based proteomics for the sole purpose of genome annotation. Despite the fact that improvements are constantly being made in MS technology, it is hard to see how the genome sequence coverage by detected peptides will be improved to the level achieved by the next-generation sequencing technology. In addition, proteomics can obviously address only the protein-coding part of the genome, which will be a major problem in large genomes such as those in humans, in which only about 1% of the total genome sequence is protein coding. Therefore, we believe that the major impact of proteogenomics will be not in genome re-annotation, but in the analysis of features that are beyond the reach of genomics, such as posttranslational modifications of proteins in the context of individualized protein databases derived via next-generation sequencing. However, the routine application of proteomics in these areas will require further substantial improvements aimed at increasing the sequencing speed/coverage (MS level) and specificity/sensitivity (bioinformatic workflows).

Acknowledgments—We thank Dr. Bourmedine Soufi for comments regarding the experimental setup and manuscript.

Data access: All raw files, peak lists, FASTA databases, and result tables obtained from MaxQuant and TPP were submitted to the PeptideAtlas data repository. The data can be accessed with the following ID and password:

ID: PASS00147
Password: GW584mr

* This work was supported by the Juniorprofessoren-Programm of the BW Stiftung, SFB766, and PRIME-XS.

S This article contains supplemental material.

§ To whom correspondence should be addressed: Prof. Dr. Boris Macek, Proteome Center Tuebingen, Interfaculty Institute for Cell Biology, University of Tuebingen, Auf der Morgenstelle 15, 72076 Tuebingen, Germany. Tel.: +49-(0)7071-29-5779; Fax: +49-(0)7071-29-5779; E-mail: boris.macek@uni-tuebingen.de.

Author contributions: K.K. and B.M. designed the experiments; K.K. performed the bioinformatic analysis; A.C., G.B., K.M., and N.C.S. performed sample preparation and MS measurements; K.K. and B.M. wrote the manuscript.

REFERENCES

1. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. Nature 422, 198–207
2. Michalski, A., Damoc, E., Hauschild, J. P., Lange, O., Wieghaus, A., Makarov, A., Nagaraj, N., Cox, J., Mann, M., and Horning, S. (2011) Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. Mol. Cell. Proteomics 10, M111.011015
3. Michalski, A., Damoc, E., Lange, O., Denisov, E., Nolting, D., Muller, M., Viner, R., Schwartz, J., Remes, P., Belford, M., Dunyach, J. J., Cox, J., Horning, S., Mann, M., and Makarov, A. (2012) Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. Mol. Cell. Proteomics 11, O111.013698
4. Olsen, J. V., Schwartz, J. C., Griep-Raming, J., Nielsen, M. L., Damoc, E., Denisov, E., Lange, O., Remes, P., Taylor, D., Splendore, M., Wouters, E. R., Senko, M., Makarov, A., Mann, M., and Horning, S. (2009) A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. Mol. Cell. Proteomics 8, 2759–2769
5. de Godoy, L. M., Olsen, J. V., Cox, J., Nielsen, M. L., Hubner, N. C., Frohlich, F., Walther, T. C., and Mann, M. (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature 455, 1251–1254
6. Picotti, P., Bodenmiller, M., Mueller, L. N., Domon, B., and Aebersold, R. (2009) Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. Cell 138, 795–806
7. Frishman, D., and Valencia, A. (2009) Modern Genome Annotation: The BioSapiens Network, Springer, New York
8. Brent, M. R. (2005) Genome annotation past, present, and future: how to define an ORF at each locus. Genome Res. 15, 1777–1786
9. Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24, 637–644
10. Kuster, B., Mortensen, P., Andersen, J. S., and Mann, M. (2001) Mass spectrometry allows direct identification of proteins in large genomes. Proteomics 1, 641–650
11. Armengaud, J. (2010) Proteo-genomics and systems biology: quest for the ultimate missing parts. Expert Rev. Proteomics 7, 65–77
12. Yates, J. R., 3rd, Eng, J. K., and McCormack, A. L. (1995) Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. Anal. Chem. 67, 3202–3210
13. Castellana, N., and Bafna, V. (2010) Proteo-genomics to discover the full coding content of genomes: a computational perspective. J. Proteomics 73, 2124–2135
14. Tanner, S., Shen, Z. X., Ng, J., Florea, L., Guigo, R., Briggs, S. P., and Bafna, V. (2007) Improving gene annotation using peptide mass spectrometry. Genome Res. 17, 231–239
15. Merrihew, G. E., Davis, C., Ewing, B., Williams, G., Kall, 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
16. Borchert, N., Dieterich, C., Krug, K., Schultz, W., Jung, S., Nordheim, A., Sommer, R. J., and Macek, B. (2010) Proteo-genomics of Pristionchus pacificus reveals distinct proteome structure of nematode models. Genome Res. 20, 837–846
17. Oshiro, G., Wodicka, L. M., Washburn, M. P., Yates, J. R., Lockhart, D. J., and Winzeler, E. A. (2002) Parallel identification of new genes in Saccharomyces cerevisiae. Genome Res. 12, 1210–1220
18. Bitton, D. A., Wood, V., Scutt, P. J., Grallert, A., Yates, T., Smith, D. L.,
Hagan, I. M., and Miller, C. J. (2011) Augmented annotation of the Schizosaccharomyces pombe genome reveals additional genes required for growth and viability. Genetics 187, 1207–1217

19. Castellana, N. E., Payne, S. H., Shen, Z. X., Stanke, M., Bafna, V., and Briggs, S. P. (2008) Discovery and revision of Arabidopsis genes by proteo-genomics. Proc. Natl. Acad. Sci. U.S.A. 105, 21034–21038

20. Bringas, S., Hane, J. K., Case, T., Tan, K. C., Lipscombe, R., Solomon, P. S., and Oliver, R. P. (2008) Deep proteo-genomics; high throughput gene validation by multidimensional liquid chromatography and mass spectrometry of proteins from the fungal wheat pathogen Stagonospora nodorum. BMC Bioinformatics 10, 301

21. Xia, D., Sanderson, S. J., Jones, A. R., Prieto, J. H., Yates, J. R., Bromley, E., Tomley, F. M., Lai, K., Sinden, R. E., Brunk, B. P., Roos, D. S., and Wastling, J. M. (2008) The proteome of Toxoplasma gondii: integration with, and prospective use of, the genome sequence for insights into gene expression and annotation. Genome Biol. 9, R116

22. Kalume, D. E., Peri, S., Reddy, R., Zhong, J., Okulate, M., Kumar, N., and Pandey, A. (2005) Genome annotation of Anopheles gambiae using mass spectrometry-derived data. BMC Genomics 6, 128

23. Brouch, M., Saunders, G. I., Frankish, A., Collins, J. A., Yu, L., Wright, J., Verstraten, R., Adams, J. D., Harron, J., Choudhury, J. S., and Hubbard, T. P. (2007) A proteo-genomic analysis of discovery of novel protein-coding genes, alternative splicing, and “re resurrected” pseudogenes in the mouse genome. Genome Res. 21, 756–767

24. Bitton, D. A., Smith, D. L., Connolly, Y., Scutt, P. J., and Miller, C. J. (2010) An integrated mass-spectrometry pipeline identifies novel protein coding-regions in the human genome. PLoS One 5, e8849

25. Fermin, D., Allen, B. B., Blackwell, T. W., Menon, R., Adamski, M., Xu, Y., Ulrich, P. E., Bray, P., Schiavello, L., and Fleischhacker, W. O. (2008) New gene and gene model detection using a whole genome open reading frame analysis in proteomics. Genome Biol. 7, R35

26. Armengaud, J. (2012) Microbiology and proteomics, getting the best of both worlds! Environ. Microbiol. 15, 12–23

27. Armengaud, J. (2009) A perfect genome annotation is within reach with the proteomics and genomics alliance. Curr. Opin. Microbiol. 12, 292–300

28. Chen, W. B., Laidig, K. E., Park, Y., Park, K., Yates, J. R., Lamont, R. J., and Hackett, L. J. (2006) Discovery of Porphyromonas gingivalis genome with peptide fragmentation mass spectra. Analyst 131, 52–57

29. Wang, R., Prince, J. T., and Marcotte, E. M. (2005) Mass spectrometry of the M. smegmatis proteome: protein expression levels correlate with function, operons, and codon bias. Genome Res. 15, 1118–1126

30. de Souza, G. A., Maen, H., Softeland, T., Seelenburg, G., Prasad, S., Jonassen, I., and Wikier, H. G. (2008) High accuracy mass spectrometry analysis as a valid tool to improve genome annotation. Nucleic Acids Res. 36, D396–D403

31. de Souza, G. A., Softeland, T., Koehler, C. J., Thiede, B., and Wiker, H. G. (2008) High accuracy mass spectrometry-derived data. Mol. Cell. Proteomics 7, 1265–1272

32. Krug, K., Nahnsen, S., and Macek, B. (2011) Mass spectrometry at the M111.011627 interface of proteomics and genomics. Trends Genet. 17, 276–277

33. Kelkar, D. S., Kumar, D., Kumar, P., Balakrishnan, L., Muthusamy, B., Iwasaki, M., Miwa, S., Ikegami, T., Tomita, M., Tanaka, N., and Ishihama, Y. (2011) Shotgun proteomics aids discovery of novel protein-coding genes. BMC Genomics 12, 3233–3243

34. Verstraten, R., Adams, D. J., Harrow, J., Choudhary, J. S., and Hubbard, T. P. (2007) Proteo-genomic analysis of the Schizosaccharomyces pombe proteome on a microarray scale. Anal. Chem. 82, 2616–2620

35. Blakely, S., Overton, I. M., and Hubbard, S. J. (2012) Addressing statistical biases in nucleotide-derived protein databases for proteo-genomic search strategies. J. Proteome Res. 11, 5221–5234

36. Iwasaki, M., Miwa, S., Ikegami, T., Tomita, M., Tanaka, N., and Ishihama, Y. (2010) One-dimensional capillary liquid chromatographic separation coupled with tandem mass spectrometry unvels the Escherichia coli proteome on a microarray scale. Anal. Chem. 82, 2616–2620

37. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L., and Mori, H. (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol. Syst. Biol. 2, 2006.0008
approach and false discovery rate: when things may go wrong. J. Am. Soc. Mass Spectrom. 22, 1111–1120
61. Cooper, B. (2012) The problem with peptide presumption and the downfall of target-decoy false discovery rates. Anal. Chem. 84, 9663–9667
62. Nesvizhskii, A. I. (2010) A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J. Proteomics 73, 2092–2123
63. Helmy, M., Tomita, M., and Ishihama, Y. (2012) Peptide identification by searching large-scale tandem mass spectra against large databases: bioinformatics methods in proteo-genomics. Genes, Genomes, and Genomics 6, 76–85
64. Srivatsan, A., Han, Y., Peng, J. L., Tehranchi, A. K., Gibbs, R., Wang, J. D., and Chen, R. (2008) High-precision, whole-genome sequencing of laboratory strains facilitates genetic studies. PLoS Genet. 4, e1000139
65. Metzker, M. L. (2010) Next generation technologies: basics and applications. Environ. Mol. Mutagen. 51, 691–691