Journal of proteome research

Perspective

pubs.acs.org/jpr

Solution to Statistical Challenges in Proteomics Is More Statistics, Not Less

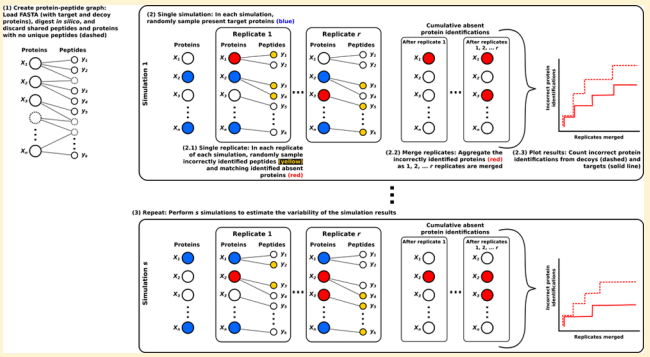
Oliver Serang*,†,‡ and Lukas Käll*,§

†Department of Informatik, Freie Universität Berlin, Takustr. 9, Berlin 14195, Germany

‡Leibniz-Institute for Freshwater Ecology and Inland Fisheries (IGB), Müggelseedamm 310, Berlin 12587, Germany

§Science for Life Laboratory, School of Biotechnology, Royal Institute of Technology − KTH, Tomtebodavägen 23A, Solna SE-171 21, Sweden

ABSTRACT: In any high-throughput scientific study, it is often essential to estimate the percent of findings that are actually incorrect. This percentage is called the false discovery rate (abbreviated "FDR"), and it is an invariant (albeit, often unknown) quantity for any well-formed study. In proteomics, it has become common practice to incorrectly conflate the protein FDR (the percent of identified proteins that are actually absent) with protein-level target-decoy, a particular method for estimating the protein-level FDR. In this manner, the challenges of one approach have been used as the basis for an argument that the field should abstain from protein-level FDR analysis altogether or even the suggestion that the very notion of a protein FDR is flawed. As we demonstrate in simple but accurate simulations, not only is the protein-level FDR an invariant concept, when analyzing large data sets, the failure to properly acknowledge it or to correct for multiple testing can result in large, unrecognized errors, whereby thousands of absent proteins (and, potentially every protein in the FASTA database being considered) can be incorrectly identified.



KEYWORDS: *protein identification, false discovery rate (FDR), simulation, multiple testing, human proteome, statistics*

Every well-formed question implies a corresponding null hypothesis, a perfect opposite to the question being asked. When we identify peptides in mass spectrometry, the naturally implied null is the set of absent peptides considered.[1,2] When we identify proteins, the naturally implied null is the set of absent proteins considered.[3] When we investigate differential quantification between two treatments, the null is defined by the set of proteins that do not differ significantly in quantity.[4]

Likewise, for a given question and its corresponding null hypothesis, the false discovery rate (FDR)[5] is a fixed mathematical quantity. If we propose a list of putative discoveries for our question, then the FDR is simply the percent of those discoveries that are incorrect (i.e., the percent of the proposed discoveries that result from the null hypothesis). Once we propose a question and a list of putative discoveries for that question, then the FDR is a concrete quantity as invariant as the number of entries in the list.

Of course, as concrete and invariant as the true FDR may be, we may still not know its exact value, and for this reason it is necessary to estimate the FDR. In proteomics, a variety of methods have been proposed for estimating the protein identification FDR: by using protein-level target-decoy search,[6] by adjusting protein-level target-decoy search results to account for the random scattering of false-positive peptides,[7] by analyzing decoy-free peptide search results and using unique peptides to assign protein identification FDRs,[8] or by using generative parametric models to compute probabilities that each protein is present given the spectral evidence of shared and unique peptides.[9−13] Creating empirical generative models in proteomics (i.e., modeling the process by which proteins become peptides and peptides become spectra, and spectra are sampled) is not very difficult, and so these approaches are primarily concerned with balancing accuracy of the model with the ability to perform efficient Bayesian inference. In contrast, nongenerative models directly model the backward inference process from spectra to peptides to proteins; these models are easily made efficient, but they can likewise hold less of a resemblence to the causal processes by which data are created.[14] Note that estimated protein-level probabilities can subsequently be used to compute estimates of protein identification FDRs by summing up the fractional expected number of absent proteins: If a protein has a posterior of 95%, then identifying the protein contributes an expected fractional quantity of 0.05 false protein discoveries.[15]

It is unsurprising that each of these methods for estimating protein identification FDR has its own set of assumptions; therefore, each has its own strengths and weaknesses; however, it is incorrect to conflate the true FDR with one particular method for estimating the FDR (e.g., protein-level target-decoy

Received: June 18, 2015
Published: August 10, 2015

DOI: 10.1021/acs.jproteome.5b00568
J. Proteome Res. 2015, 14, 4099−4103

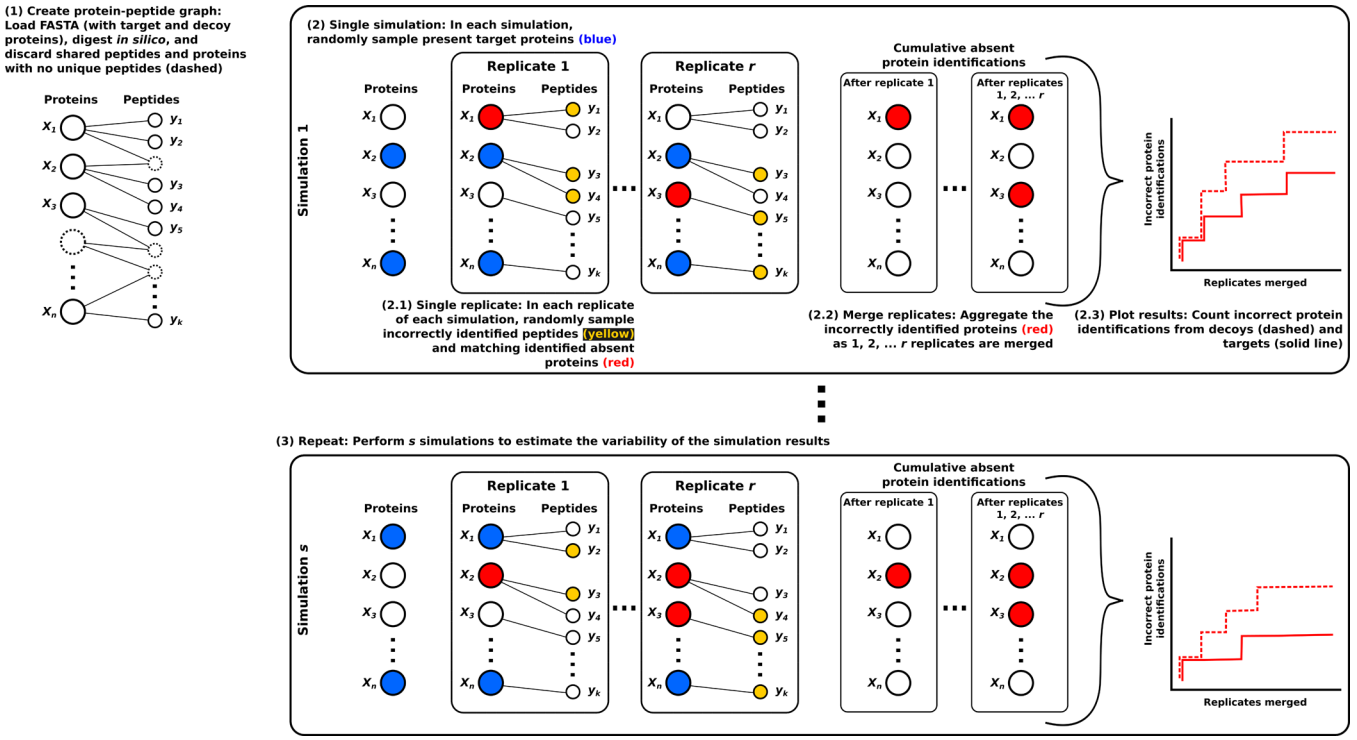


Figure 1. Empirical, decoy-free simulation of multiple testing effect. (1) Human proteome FASTA file is loaded and concatenated with reversed decoy proteins. The result is trypsin-digested in silico, ignoring shared peptides and proteins without any unique peptides. (Ignored proteins and peptides are drawn with dashed lines.) (2) In each simulation, $(1 - \pi_0)n$ present proteins are sampled without replacement from the target proteins (blue indicates present). (2.1) In each replicate data set of each simulation, αk absent target peptides are sampled without replacement; any absent protein matching one of these peptides is an incorrect identification for that replicate. (2.2) Cumulatively merge the absent protein identifications. In simulation 1, at least one incorrect protein identification is shown from replicate 1 (X_1). After merging replicates 1, 2, ... r, at least two absent protein identifications are found (X_1 and X_3). (2.3) The counts of the cumulative incorrect protein identifications from absent target proteins and the cumulative incorrect protein identifications from decoy proteins are plotted against the number of replicates merged. (3) Several full simulations are performed (each simulation using its own set of present proteins) to illustrate the low variability of the plotted result even when different sets of proteins are present.

search) and thereby denounce the true FDR as having the weaknesses of the method used to estimate it. This type of reasoning, whereby an entire theoretical notion is ruled out because of fears about the current implementation, is a fallacy frequently found on the wrong side of history: It was this very same flavor of reasoning employed in Prof. Paul Ehrlich's now infamous 1968 book *The Population Bomb*, wherein he incorrectly claimed that the world could not support more than the then-current population of 3.6 billion people because the world would run out of a small number of precious commodities[16] (e.g., tin, which had been one of the key materials used for canning food); however, tin cans were but one mechanism for canning or packaging food, and the wide availability of aluminum, paper products, and plastics as substitutes during the latter half of the twentieth century rendered rising tin prices moot. In fact, only a few months before Ehrlich's book was published, the newly released film *The Graduate* saw this proclamation made to a young Dustin Hoffman: "Just one word...Plastics...There's a great future in plastics."

In a paper accompanying one of the two drafts of the human proteome, Wilhelm et al.[17] suggest multiple times that the very notion of protein identification FDR should be deprecated and left behind in mass spectrometry. For example, when identifying proteins on a data set of massive scale (the authors consider 71 million peptide-mass spectra), the authors surprisingly eschew FDR estimation, saying, "We refrained from calculating protein FDR measures as the concept of a protein FDR is problematic...protein FDRs are actually not very meaningful because proteomics measures peptides not proteins and the definition of a 'decoy protein' is quite problematic." By saying this, the authors implicitly equate the true, invariant FDR (i.e., the proportion of incorrectly identified proteins that are included in any particular proposed human proteome) with a single, imperfect implementation for estimating the FDR and then use this as an excuse to ignore the very concept of protein identification FDR, which is an invaluable concept in high-throughput science. Furthermore, the authors subsequently employ this reasoning in a circular manner, suggesting that their protein FDR is acceptable despite the extremely large number of decoy proteins identified and then use that as the justification for a novel FDR estimation method, which computes low protein FDR estimates for their draft human proteome.

The hazards of ignoring protein-level FDR are easily compounded, for once we exchange statistical methods for human intuition, we can be easily fooled. For instance, when Wilhelm et al.[17] merge across their r = 16 857 replicate experiments, they identify any protein that matches at least one high-scoring peptide in at least one experiment. To human intuition, this may sound quite reasonable, but when viewed through a more statistical lens, merging results in this manner can easily introduce optimistic biases from multiple testing. It is precisely these types of unrecognized multiple testing errors

DOI: 10.1021/acs.jproteome.5b00568
J. Proteome Res. 2015, 14, 4099−4103

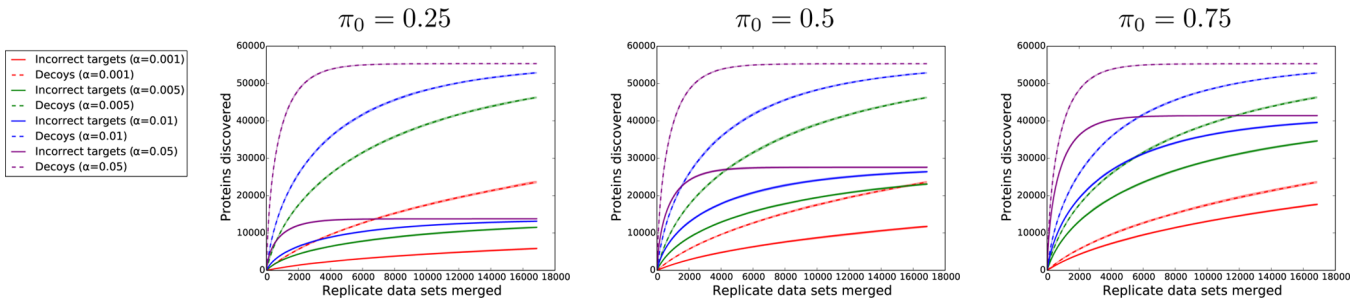


Figure 2. Simulation results, illustrating the hazard of multiple testing. For each value of π_0, the proportion of target proteins that are actually absent, and for each value of α, the peptide PEP, $s = 128$ simulations are performed, where each simulation merges the results of all $r = 16\,857$ replicate experiments. Although π_0 and α modulate the accumulation of false-positive target protein identifications, the accumulation from multiple testing is nonetheless severe, even when very conservative values are chosen. For each set of parameters, the dark line (dashed for decoys) represents the average over the 128 replicate simulations, and the shaded color represents the minimum and maximum bounds over all 128 such replicates.

that resulted in poor reproducibility for early genome-wide association studies and which continue to contribute to negative perceptions about that field.[18,19]

Here we perform simple simulations emulating the workflow of Wilhelm et al.,[17] which illustrate the effects of multiple testing on the protein false discoveries using the same FASTA database. (This database includes 82 092 proteins after excluding those containing amino acid wildcard characters.) This FASTA file is first loaded and then digested in silico to form the protein−peptide graph. Then, all shared peptides are removed from the protein−peptide graph, and subsequently any protein without a unique peptide is also removed.

In each simulation, a set of present proteins from the human proteome is chosen at random without replacement. The expected number of present proteins is $(1 - \pi_0)n$, where n target proteins are considered in the database (each must have at least one unique peptide to be considered) and π_0 is the percent of those target proteins that are absent.[20] For example, if (contaminant free) spectra from a UPS protein standard were searched against a FASTA database containing exactly the UPS proteins (no more proteins and no fewer) from the standard, then $\pi_0 = 0$, and a decoy count could overestimate the number of absent target proteins (which makes sense, as there can be no absent target proteins in that example). On the contrary, if spectra from human tissue were searched against a FASTA database from *Pyrococcus furiosus*, then $\pi_0 \approx 1$, meaning the decoy protein counts may closely resemble the number of absent target protein identifications. In practice, the value of π_0 is generally unknown but can be estimated using various asymptotics (by using the fact that the low-scoring targets are more decoy-like if the score is discriminative).[3,20]

The randomly sampled present and absent proteins are used in subsimulations (i.e., individual experimental replicates that will be aggregated). In each simulated experimental replicate, we sample incorrect peptide identifications with replacement because it is possible for the same peptide to be incorrectly identified twice (whereas the set of present proteins is sampled without replacement because they refer to presence−absence state rather than identifications). When k peptides are identified in a replicate experiment at peptide-level FDR α, αk incorrect peptide identifications are expected, and these incorrect peptide identifications are randomly distributed over the database of peptide sequences because it has been shown that false-positive peptides stratify randomly by Reiter et al.[7] For these experiments, we use $k = 10\,000$ (a conservative number of peptides identified in a single replicate[1]). As performed by Wilhelm et al.,[17] tryptic peptides with length of at least six amino acids are considered, resulting in 684 640 target peptide candidates (258 668 of which are unique peptides, from which the $\alpha k = \alpha$ 10 000 false-positive peptides are sampled) and 82 092 target proteins ($n = 55\,294$ of which contain at least one unique peptide). By definition, we can say that a protein is incorrectly identified (according to the one-peptide rule, as used by Wilhelm et al.[17]) when (1) that protein is absent from the sample and (2) the protein matches at least one incorrect peptide identification (Figure 1).

Thus, by following the experimental design of Wilhelm et al.,[17] each subsimulation (i.e., each experimental replicate) poses yet another chance to incorrectly identify absent proteins (when π_0 and α are any nonzero values, i.e., as long as any target proteins are absent from the sample and as long as the peptide identifications are anything less than perfect). This leads to the cumulative aggregation of incorrect protein identifications. For example, if 25% of the target proteins are present (i.e., $\pi_0 = 0.75$) and a peptide FDR threshold of $\alpha = 0.01$ is used, then merging 2000 replicate data sets is expected to produce roughly 28 000 decoy protein identifications and roughly 7000 incorrect target protein identifications (blue dashed series and blue series, left panel of Figure 2, respectively). After merging 6000 replicate data sets, the decoy protein identifications have increased to roughly 40 000 and the incorrect target protein identifications have increased to roughly 10 000. Note that in these experiments, the ratio of the incorrect target protein identifications to decoy protein identifications is typically π_0 because a random decoy protein has a 100% chance of being absent, while the percentage of target proteins that are absent equals π_0; in comparison, after merging all 16 857 replicates, Wilhelm et al.[17] identify 12 323 decoy proteins in comparison with the 19 376 target proteins identified.

Because the Wilhelm et al.[17] design accumulates false-positive target protein identifications in this manner, it is likely that this contributes to proteins in the human proteome draft, including but not limited to the 100 olfactory receptors that were identified in unexpected tissues.[21] Note that even when the peptide-level FDR α is vanishingly small, k (the number of peptide hypotheses tested) is large enough that αk (the expected number of incorrectly identified peptides per replicate) is nonzero and can thus be amplified substantially when aggregating $r = 16\,857$ protein-level hypotheses. This is the quintessential issue of multiple testing that defies human intuition: Even very stringent thresholds in each replicate

DOI: 10.1021/acs.jproteome.5b00568
J. Proteome Res. 2015, 14, 4099−4103

experiment can be overcome by uncorrected merging of multiple replicate experiments in this manner.

In summary, when αkr is large (i.e., when the peptide-level FDR, the number of peptide identifications, and the number of merged replicates are all sufficiently larger than zero), the number of false-positive identifications (both target and decoy) will become large until it begins saturating the entire set of possible protein-level hypotheses. As more and more protein identifications are accumulated over several replicates, in these experiments the asymptotic ratio of decoy identifications to absent target identifications becomes π_0, the percentage of targets that are truly absent.

In accordance with our simulations, Wilhelm et al.[17] also find a large number of decoy protein identifications; however, they claim that the decoy protein identifications dramatically overestimate the false-positive target identifications (so much so, that they claim that the decoy measurement is useless); however, as long as $\alpha > 1/k = 1/10\,000$ (thereby indicating that a nonzero number of false-positive peptides is expected in each replicate experiment), the only manner by which the decoy protein FDR will dramatically overestimate the target FDR is when $\pi_0 \approx 0$ (i.e., nearly every possible human protein is present at currently detectable levels in every tissue, thus leaving no room for false positives by design and also rendering identification with high-throughput methods completely redundant). In short, it is essentially impossible for multiple testing to not be a factor when so many replicate experiments are merged in this manner; therefore, it is almost certain that the massive number of proteins identified by Wilhelm et al.[17] (19 376 proteins, roughly 22% of UniProt, and an even higher percentage of those proteins containing at least one unique peptide) is an overestimate of what can currently be detected.[21,22]

It should be said that such statistical errors in high-throughput fields like proteomics are quite explicable and are related to the often incorrect presumption of certainty in our field; even the language in proteomics favors phrases like "identified proteins", even when referring to inferences even though those inferences may, in fact, later be found to be erroneous. Words like "identified" do not put us in the right mindset to properly recognize inevitable uncertainty, nor do they let us anticipate the future adjustments that will be made to our conclusions. In light of this, it is little surprise that statisticians favor words like "hypothesize", which force us to directly confront our uncertainty. In high-throughput science, it is not the uncertainty itself that inevitably leads to incorrect conclusions; instead, it is the confluence of uncertainty and the presumption of certainty that together produce bad results.[23] And just as estimates for protein FDR improve, so too will the drafts of the human proteome continue to evolve; for the drafts of the human proteome to continue to evolve, it would benefit the community if the draft made available the unfiltered scores or probabilities from all peptides identified (including those of low quality and those from decoy searches[24]) to enable the use of other statistical approaches to complement the target-decoy FDR estimate and to infer whether the proteins included in the current draft are too optimistic (i.e., more than were genuinely present) or too conservative (i.e., fewer than were genuinely present).[3]

The use of meaningful and interpretable and concrete mathematical concepts (e.g., FDR and the proper accounting for effects of multiple testing) are deeply intertwined in the fabric of high-throughput science and are essential to preventing subtle, but profound errors. In the rush to publish the most far-reaching conclusions (e.g., the largest numbers of identified proteins) in the most prestigious journals, the burdens of statistical subtleties can easily be forgotten. It is imperative to note that computational proteomics is a community, and as such the responsibility for catching such subtleties does not lie solely with the authors but is a greater result of the values that the field projects and rewards. (Specifically, the number of identifications is often valued above all else.) In this manner, it is incumbent upon the field to place less emphasis on the number of identifications achieved in a particular paper (which places an implicit pressure on authors) and instead to value the work as a whole. Through this lens, the scale of the data processed by Wilhelm et al.[17] still represents a serious technical achievement in its own right, which in our view would merit publication in top journals even after eliminating a significant portion of the protein identifications.

Just as we challenge biologists and other scientists not to ignore statistics even though rigorous statistical analysis may sometimes decrease the number of proposed findings, we also challenge the next generation of computational and statistical researchers to create innovative methods for false discovery rate estimation and multiple testing correction. These types of statistical analyses are so essential to high-throughput science that creative new ideas and advances offer to make a wide scientific impact across myriad fields. To paraphrase that classic scene from *The Graduate*, we offer the following advice: "There's a great future in statistics."

AUTHOR INFORMATION

Corresponding Authors

*O.S.: Tel: +49 30 838 75870. E-mail: orserang@uw.edu.
*L.K.: Tel: +46 76 9425179. E-mail: lukas.kall@scilifelab.se.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Käll, L.; Canterbury, J.; Weston, J.; Noble, W. S.; MacCoss, M. J. A semi-supervised machine learning technique for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4*, 923−25.

(2) Granholm, V.; Noble, W. S.; Käll, L. On using samples of known protein content to assess the statistical calibration of scores assigned to peptide-spectrum matches in shotgun proteomics. *J. Proteome Res.* **2011**, *10* (5), 2671−2678.

(3) Serang, O.; Paulo, J.; Steen, H.; Steen, J. A. A non-parametric cutout index for robust evaluation of identified proteins. *Mol. Cell. Proteomics* **2013**, *12* (3), 807−812.

(4) Serang, O.; Cansizoglu, A. E.; Käll, L.; Steen, H.; Steen, J. A. Nonparametric bayesian evaluation of differential protein quantification. *J. Proteome Res.* **2013**, *12* (10), 4556−4565.

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

(6) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4* (3), 207−214.

(7) Reiter, L.; Claassen, M.; Schrimpf, S. P.; Jovanovic, M.; Schmidt, A.; Buhmann, J. M.; Hengartner, M. O.; Aebersold, R. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol. Cell. Proteomics* **2009**, *8* (11), 2405−2417.

DOI: 10.1021/acs.jproteome.5b00568
J. Proteome Res. 2015, 14, 4099−4103

(8) Kim, S.; Gupta, N.; Pevzner, P. A. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J. Proteome Res.* **2008**, *7*, 3354−3363.
(9) Li, Y. F., Arnold, R. J., Li, Y., Radivojac, P., Sheng, Q., Tang, H. A Bayesian approach to protein inference problem in shotgun proteomics. In *Proceedings of the Twelfth Annual International Conference on Computational Molecular Biology*; Vingron, M., Wong, L., Eds.; Lecture Notes in Bioinformatics *12*; Springer: Berlin, Germany, 2008; pp 167−180.
(10) Serang, O.; MacCoss, M. J.; Noble, W. S. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. *J. Proteome Res.* **2010**, *9* (10), 5346−5357.
(11) Serang, O.; Noble, W. S. Faster mass spectrometry-based protein inference: Junction trees are more efficient than sampling and marginalization by enumeration. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2012**, *9* (3), 809−817.
(12) Arnold, R. J.; Li, Y. F.; Radivojac, P.; Tang, H. Protein identification problem from a Bayesian point of view. *Statistics and Its Interface* **2012**, *5* (1), 21−37.
(13) Serang, O. The Probabilistic Convolution Tree: Efficient Exact Bayesian Inference for Faster LC-MS/MS Protein Inference. *PLoS One* **2014**, *9* (3), e91507.
(14) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **2003**, *75*, 4646−4658.
(15) Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. Posterior error probabilities and false discovery rates: two sides of the same coin. *J. Proteome Res.* **2008**, *7* (1), 40−44.
(16) Ehrlich, P. R. *The Population Bomb*; Ballantine: New York, 1968.
(17) Wilhelm, M.; Schlegl, J.; Hahne, H.; Gholami, A. M.; Lieberenz, M.; Savitski, M. M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; Mathieson, T.; Lemeer, S.; Schnatbaum, K.; Reimer, U.; Wenschuh, H.; Mollenhauer, M.; Slotta-Huspenina, J.; Boese, J.-H.; Bantscheff, M.; Gerstmair, A.; Faerber, F.; Kuster, B. Mass-spectrometry-based draft of the human proteome. *Nature* **2014**, *509*, 582−587.
(18) Cardon, L. R.; Palmer, L. J. Population stratification and spurious allelic association. *Lancet* **2003**, *361* (9357), 598−604.
(19) Pearson, T. A.; Manolio, T. A. How to interpret a genome-wide association study. *JAMA* **2008**, *299* (11), 1335−1344.
(20) Storey, J. D. The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value. *Annals of Statistics* **2003**, *31* (6), 2013−2035.
(21) Ezkurdia, I.; Vazquez, J.; Valencia, A.; Tress, M. Analyzing the first drafts of the human proteome. *J. Proteome Res.* **2014**, *13* (8), 3854−3855.
(22) Abascal, F.; Ezkurdia, I.; Rodriguez-Rivas, J.; Rodriguez, J. M.; del Pozo, A.; Vázquez, J.; Valencia, A.; Tress, M. L. Alternatively spliced homologous exons have ancient origins and are highly expressed at the protein level. *PLoS Comput. Biol.* **2015**, *11* (6), e1004325.
(23) Serang, O.; Moruz, L.; Hoopmann, M. R.; Käll, L. Recognizing uncertainty increases robustness and reproducibility of mass spectrometry-based protein inferences. *J. Proteome Res.* **2012**, *11* (12), 5586−91.
(24) Pevzner, P. A.; Kim, S.; Ng, J. Comment on "protein sequences from mastodon and tyrannosaurus rex revealed by mass spectrometry. *Science* **2008**, *321* (5892), 1040.

DOI: 10.1021/acs.jproteome.5b00568
J. Proteome Res. 2015, 14, 4099−4103