Dowsey, A. W. (2017). The need for statistical contributions to bioinformatics at scale, with illustration to mass spectrometry. *Statistical Modelling*, *17*(4-5), 290-299. https://doi.org/10.1177/1471082X17708519

# The need for statistical contributions to bioinformatics at scale, with illustration to mass spectrometry

**Andrew W. Dowsey** [1]

[1] School of Social & Community Medicine and School of Veterinary Sciences, Faculty of Health Sciences, University of Bristol, United Kingdom

---

**Address for correspondence:** Andrew W. Dowsey, School of Social & Community Medicine, Oakfield House, Oakfield Grove, Bristol BS8 2BN, United Kingdom.

**E-mail:** andrew.dowsey@bristol.ac.uk.

**Phone:** (+44) 117 33 10051.

**Fax:** (+44) 117 92 89582.

---

**Abstract:** In their article, Morris and Baladandayuthapani clearly evidence the influence of statisticians in recent methodological advances throughout the bioinformatics pipeline, and advocate for the expansion of this role. The latest acquisition platforms, such as next generation sequencing (genomics/transcriptomics) and hyphenated mass spectrometry (proteomics/metabolomics), output raw datasets in the order of gigabytes; it is not unusual to acquire a terabyte or more of data per study. The increasing computational burden this brings is a further impediment against the use of statistically rigorous methodology in the pre-processing stages of the bioinfor-

matics pipeline. In this article I describe the mass spectrometry pipeline and use it as an example to show that beneath this challenge lies a two-fold opportunity: (i) Biological complexity and dynamic range is still well beyond what is captured by current processing methodology, hence potential biomarkers and mechanistic insights are consistently missed; (ii) Statistical science could play a larger role in optimizing the acquisition process itself. Data rates will continue to increase as routine clinical omics analysis moves to large-scale facilities with systematic, standardized protocols. Key inferential gains will be achieved by borrowing strength across the sum total of all analyzed studies, a task best underpinned by appropriate statistical modeling.

---

**Key words:** Computational Statistics; Sparse Signal Processing; Mass Spectrometry; Proteomics; Metabolomics

# 1 Introduction

Largely replacing the use of 2-D gel electrophoresis, mass spectrometry (MS) interfaced to liquid chromatography (LC) or gas chromatography (GC) is now a key multi-billion dollar industry for profiling the protein and metabolite content of tissues and bodily fluids. Metabolites (small molecules) are the intermediates of biochemical reactions. Their levels are determined by the properties and concentration of protein enzymes, and hence are a consequence of many regulatory processes involving the genome, transcriptome and proteome. While reflecting the underlying transcriptome to some degree, proteins are also modulated by many factors such as post-translational modifications and protein-protein interactions. Disease processes either cause or effect

perturbations of the proteome and metabolome, hence their study is key to disease understanding. MS is therefore utilized at all stages of biological science, translational medicine and drug development to increase confidence in rationale.

Targeted MS approaches are utilized pervasively for drug metabolite identification and to monitor safety markers in toxicology. More recently, untargeted global profiling of the proteome and metabolome with MS has gained significant traction for discovery of disease mechanism and drug targets by generating pre-clinical and clinical evidence of their relevance to pathophysiology and phenotype, and their effect on disease and safety implicated biological pathways when modulated. In the systems medicine paradigm, the omics data (from MS and genomics/transcriptomics approaches such as next-generation sequencing) feed into knowledge-base guided pathway analysis and computational biology tools, supporting expert-guided literature review to establish the evidence base. Complementary to this, and critical to the success of modern drug development and diagnostics programs, untargeted MS is also key to the characterization of predictive, prognostic and mechanistic clinical biomarkers that realize early detection of a disease process or enable patient stratification based on drug safety or efficacy. The ultimate aim is the generation of specific assays that can accurately and reproducibly measure the levels of biomarkers in patient populations, compatible with use in a busy clinical environment. To date, most protein assays are generally based around antibodies, despite recent high-profile concerns with repeatability (Baker, 2015), 2015).Targeted MS offers significant advantages and has a clear role either to validate protein biomarkers in large clinical cohorts prior to costly antibody-based assay development, or for the clinical assay itself, as is already the case for small molecule analysis.

MS of whole proteins is still challenging, hence to reduce spectrum complexity the dominant bottom-up methodology cuts proteins into smaller peptides prior to analysis, using an enzyme such as trypsin that cleaves at predictable positions. In untargeted proteomics and metabolomics, the biological sample is pre-separated by LC (or GC for volatile metabolites), to reduce complexity. Each small interval of elution out of the LC or GC column is ionized before mass analysis to determine the mass-to-charge ratio (m/z) of each of its constituent molecules. Each abundant biochemical appears as multiple mass spectral features with different charge states, adducts, losses or modifications, each with an intensity proportional to its abundance. Quantification of absolute abundance is non-trivial in untargeted MS as there is no accurate method to universally predict the scaling factor between a features intensity and its abundance. Nevertheless, since this coefficient is stable for each feature in well-controlled experimental conditions, comparison of relative intensity across biological samples can be used to quantify changes in relative abundance. Each feature is also observed as a series of peaks at increasing m/z due to the presence of natural isotopes such as 13C. The relative intensity of these peaks (the isotope distribution) is characteristic of the underlying molecular formula and can give a putative identification for that biochemical, but given that many metabolites and peptides share the same molecular formula, further mass analysis is necessary to identify definitively. In data-dependent acquisition (DDA), the most intense features are selected by a second low-resolution mass spectrometer for collision-induced fragmentation by a noble gas, with the m/z of these fragment ions captured by a third mass spectrometer. The resultant fragmentation patterns are predictable, enabling the routine identification of peptides by this method. Metabolite identification may require further rounds of fragmentation to increase specificity, and may include comparison with authentic biochemical standards.

DDA has been the workhorse approach for years, but limitations include biased and incomplete coverage due to intensity-dependent selection, and co-selection leading to chimeric fragmentation spectra that are difficult to interpret. In recent years, Data Independent Acquisition (DIA) strategies have emerged which mitigate these issues by using one or multiple broad selection windows that cover the mass range. In this approach fragmentation data is available for all successfully ionized and fragmented biochemicals, but due to the wide selection windows used many biochemicals are fragmented together. Unlike DDA, the fragment elution profiles are available, however, and so in theory fragments with matching elution profiles can be clustered back together by the bioinformatics pipeline.

## 2 Bioinformatics challenges

As stated by Morris and Baladandayuthapani, Driven by the computational challenges of high dimensionality and out of convenience, many commonly used standard analysis approaches are reductionistic (not modeling the entire data set), ad hoc and algorithmic (not model-based), stepwise and piecemeal (not integrating information together in a statistically efficient way or propagating uncertainty through to the final analysis). This succinctly describes the pervasive approach in proteomics and metabolomics MS. There are a number of commercial (e.g. Nonlinear Dynamics Progenesis, Waters Coorporation) and academic packages (e.g. MaxQuant (Cox and Mann, 2008) for proteomics, XCMS (Tautenhahn et al., 2012) for metabolomics) available. However, their approach is to take a series of deterministic, self-contained steps (Bessant, 2016) e.g. peak detection; clustering of peaks into quantitative features;

matching features across samples under chromatographic deformation; identification by pattern matching of fragments against databases of theoretically or experimentally derived spectra; grouping of peptides of a protein (protein inference), or adducts of a metabolite (metabolite inference); modeling associated quantifications to infer protein quantification; differential analysis across experimental conditions. Raw data is invariably converted to a symbolic representation at an early stage, so: (1) statistical evidence is lost from one stage to the next, thus errors propagate and amplify; (2) once detailed structure inherent in the raw data is gone, it is no longer possible to detect or quantify overlapping spectral signals; (3) since peaks are detected in isolation with no higher-level knowledge, numerous features are never even detected, despite informative evidence across the experiment as a whole.

In the beginning, these issues were more or less irrelevant since the mainstays of proteomics MS analysis (life sciences research) and metabolomics (organic chemistry) predominantly used simplified model systems defined and manipulated within tightly controlled experimental environments. Bioinformatics methodology for MS, however, continues to be fundamentally reliant on these characteristics. In biomedical research, despite stringent control of confounding factors in experimental design, a step-change in complexity and heterogeneity is evident within typical disease models and clinical samples. The field has acknowledged this challenge, but has tackled it mainly by improving the instrumentation. The introduction of the Orbitrap instrument (Hu et al., 2005),where ions are trapped electrostatically in orbit around a central electrode to induce a current capturing the frequency of oscillation of all ions simultaneously, yielded a step-change in performance due to its high resolution; in Orbitrap data less signals overlap, therefore more data can be interpreted by the bioinformatics tools. The resolution of other vendors instrumentation, mainly based

on time-of-flight analyzers (Chernushevich et al., 2001), where ions are subjected to an accelerating voltage and drift down a field-free flight tube in a duration quadratically related to m/z, has now narrowed or eliminated this gap - particularly when considered together with additional resolving power through extremely fast cycle times or third-dimension separations using ion mobility. Nevertheless, this comes at a cost; current-generation instrumentation retail from $0.5 to $1 million. Furthermore, three significant problems remain: (1) The rate of data output has increasing tremendously (5-20Gb per dataset is now not uncommon), which continues to discourage software developers from using the raw data for more than just simple peak detection; (2) Ionization efficiency has not improved with cycle time, hence less ions are recorded within each spectrum; nosier peaks are more difficult to extract. (1) The distribution of biomolecular signals across m/z is markedly non-uniform since the molecules are largely composed of protons and neutrons that are approximately one atomic mass unit in mass. Hence signals cluster together and often continue to overlap, while unfragmented molecular isomers will always overlap in MS. Morris and Baladandayuthapani assert that statisticians understand the profound effect of sampling design decisions and that Statistical expertise in the experimental design and low-level processing stages are equally if not more important than end stage-modeling. I am in agreement but would go further and suggest that statisticians could also have a strong influence on the design and operation of the instruments themselves. Spectral acquisition is optimized for visual interpretation, while existing bioinformatics tools aim to mirror that visual interpretation. There is the opportunity for statistical modelers to work with instrument engineers to optimize based on minimizing inferential uncertainty based on jointly-constructed generative models of MS formation.

Since few MS bioinformatics tools propagate uncertainty or borrow strength within

or across steps, inaccuracies and incorrect decisions cascade and amplify downstream. This contrasts with the development of probabilistic models for next-generation sequencing (Glaus et al., 2012; Li and Dewey, 2011), which has moved the field from differential gene expression towards the more challenging goal of robust comparative analysis of gene isoforms (transcripts). Probabilistic models are used to infer isoform-level abundances and their differences. However, in MS proteomics the analysis of protein isoforms (Webb-Robertson et al., 2014), which include protein species heterogeneity at DNA-level (allelic variations), RNA-level (alternative isoforms) and protein-level (post translational modifications - PTMs), is still in its infancy: Most protein inference methodology considers only the most consensus abundance pattern per protein, which is hoped to be representative (Serang and Kll, 2015). In studies on eukaryotes of higher complexity (most animals and plants for example), shared peptides are the norm rather than the exception, and in many cases peptides unique to a single protein isoform do not exist. This is a significant challenge for protein assay development by targeted MS, as peptides used for quantification need to be (i) unique in that type of biological sample, (ii) offer sufficiently sensitive detection on the MS, and (iii) be recoverable from the biological sample in a highly reliable and reproducible manner. Current resources for targeted MS curate useful peptide data regarding specificity of the peptide and its fragments. However, there is currently a lack of critical data on peptide recovery and quantitative reproducibility which is required for optimal peptide selection and rapid and robust assay development. Alongside these developments, in 2014 two teams published high-profile drafts of the human proteome (Kim et al., 2014; Wilhelm et al., 2014) potentially a gold mine for translational medicine. While the drafts attain wide protein identification coverage, they also underline the statistical challenges facing current methodology, which

is only just beginning to deal with false-discovery rate issues in studies of this scale (Serang and Noble, 2012). Moreover, at present the draft proteomes only include simple quantitative information, of limited utility to clinical biomarker discovery.

In clinical samples there are also a number of confounding factors that are difficult to avoid and challenging to downstream statistical modeling and Intregromics (Morris and Baladandayuthapani, section 6.2), such as (i) interaction between multiple disease processes, (ii) decoupling between disease sites and sampling sites such as biofluids, and (iii) the aggregation of multiple latent endotypes exhibiting similar phenotype, such as in asthma. The integration of quantitative outputs from proteomics and metabolomics into rigorous statistical models for pre-clinical and clinical trial design (e.g. mixed-effects models, survival/frailty models), clinical prediction models (e.g. logistic regression) and latent endotype models (e.g. factor analysis, Bayesian networks) is reliant on their appropriate and accurate statistical handling. Unfortunately, this is exceptionally problematic with current pipelines, since multiple confounding effects are coalesced into single measurements with complex error distributions, non-linearity and large amounts of outliers and data missing not at random. There are two main mechanisms for this missingness: (i) Low intensity features are much more likely to be censored due to insensitivity in the feature detection method; (ii) Features can be missed at random, due to a combination of technical and informatics issues. Unfortunately these effects are mostly ignored, but mixed-effects methodology has been presented that can compensate for these missingness mechanisms and optionally impute the missing data (Karpievitch et al., 2009a). Nevertheless, improved upstream processing with uncertainty propagation would be a much more optimal consideration.

# 3 State of play

One side effect of presenting pre-processed feature data for downstream modeling is that the true generative model of the raw data has become obfuscated. While it is universally acknowledged by bioinformaticians that next-generation sequencing generates count data, it is less well understood that the same is true for most MS instruments nb. Orbitrap and related devices output a signal that aims to be proportional to the underlying ion counts. Studies have shown that the Poisson distribution is a good fit to mass spectral noise, dwarfing other components, while isotope distributions exhibit multinomial variation (Du et al., 2008). Nevertheless, MS feature detection approaches universally assume Gaussian noise. A Poisson model of ion-counting statistics immediately offers two advantages: Firstly, negative counts are unattainable. Secondly, it supports the observation that variance appears to drop as m/z increases, which has led to approaches to estimate variance in each segment of a spectrum (Wang et al., 2008). However, since peak width tends to increase as m/z rises, peaks will on average decrease in intensity, assuming average peak volumes remain constant. A Poisson explanation would predict this heteroscedastic relationship between intensity and variance.

We have recently demonstrated a biomarker discovery pipeline that markedly improves sensitivity over state-of-the-art commercial software by operating at the raw data level throughout, using a pipeline of (i) sparse regression with a Poisson likelihood, (ii) image registration to account for chromatogram alignment, and (iii) Bayesian functional wavelet mixed-effects modeling (Liao et al., 2014). Nevertheless, much more needs to be done. Firstly, functional regression does not remove the requirement for feature detection, since the derived changes need to be assigned to

the proteins or metabolites they relate to. Moreover, while functional regression can have a significant effect on feature detection performance simply because the derived differential features are much less likely to overlap than all features together, feature extraction approaches must continue to be improved, as cataloging the identity of all proteins/metabolites in a dataset will always remain an important use case. Given the scale of the raw data, from a computational viewpoint the use of High Performance Computing clusters and/or Graphical Processing Units is pertinent. From a statistical modeling viewpoint the debate is whether Bayesian computational approaches are necessary or whether point-based methods would be sufficient, given that information on uncertainty is already duplicated across the dataset by the sets of features relating to the same protein or metabolite. The efficacy of the later has already been demonstrated under the prior assumption that the spatial distribution of features is sparse e.g. The NITPICK method (Renard et al., 2008) utilizes a non-negative extension of the LASSO to extract whole features from MS data. More recently, an analogous approach based on mixture modeling has also been proposed (Browne et al., 2010).

Secondly, while we found substantial amounts of significant novel differential expression on clinical DDA data, the majority of these candidate biomarkers were expressed at too low a level to be selected for fragmentation. DIA has significant potential to mitigate this issue, but post-hoc clustering to assign fragments to their unfragmented parents has only been partially effective; at present identification coverage has not greatly surpassed DDA (Tsou et al., 2015). Nevertheless, existing methodology does not consider uncertainty. In metabolomics, the equivalent problem of clustering together metabolite adducts (metabolite modifications generated during ionization) has been tackled successfully with a hierarchical Bayesian mixture model (Suvitaival et al., 2014). In this approach, the chromatographic profiles of each adduct are extracted

and their pairwise correlation calculated for input into the model.

If you consider solely statistical contributions operating on pre-processed data, there is a healthy amount of activity in the field. To briefly illustrate a subset, chromatogram alignment has been demonstrated using a Bayesian mixture model with a Dirichlet Process prior (Benjamin et al., 2013), or Hidden Markov Models with inference provided by either Expectation-Maximization (Listgarten et al., 2007) or Markov-Chain Monte Carlo (MCMC) (Befekadu et al., 2011). Time-dependent instrumental drift and batch effects have been investigated using Surrogate Variable Analysis (Karpievitch et al., 2009b) and corrected for by a hierarchical Bayesian model incorporating Gaussian Process priors (Ranjbar et al., 2015). In metabolomics, Latent Dirichlet Allocation has been utilized as a topic model to extract and cluster common sub-patterns from fragmentation spectra (Hooft et al., 2016). These common sub-patterns are indicative of shared molecular substructures that enable some mechanistic interpretation where a full identification is infeasible. In proteomics, feature-based mixed-effects modeling is popular, and is used for integrated modeling of the study design together with relationships between feature-level and protein-level quantification. Models have been fitted using Restricted Maximum Likelihood (Choi et al., 2014), and extended to include both M-estimation for robustness to outliers and empirical Bayes to borrow strength across proteins (Goeminne et al., 2015). A recent Bayesian MCMC solution includes modelling of individual peptide variance, down-weighting the influence of unreliable peptides on the result (Freeman et al., 2015).

One area that could see increased research focus is structured learning incorporating protein and metabolite pathway information. Both Protein-Set Enrichment Analysis

(Lavalle-Adam et al., 2014) and Metabolite Pathway Enrichment Analysis (Kankainen et al., 2011) have been presented, which utilize biological knowledge-base information in a manner analogous to GSEA (Morris and Baladandayuthapani, Section 6.1). However, it is vital to consider such approaches using early or intermediate data rather than end-stage protein/metabolites quantifications, in order to handle the significant uncertainties inherent in MS data, both in terms of quantifications and, in particular for metabolomics, identifications (Rogers et al., 2009). Data-driven structure learning and integromics has seen limited research; a notable contribution is the technique of Latent Protein Trees, which offers a Bayesian latent factor model that establishes a binary tree of correlations that could represent the hierarchy of peptides, proteins and biological pathways (Henao et al., 2013). Its integration with gene expression data was discussed in (Carin et al., 2012).

# 4 Conclusion

In this article I have illustrated the landscape of current bioinformatics tools for MS, and have shown that despite the richness of information contained in modern MS datasets, much is thrown away at the early stage of the bioinformatics pipeline before statistical modeling is performed. There are significant opportunities to impact the field through the development of scalable modeling techniques for pre-processing MS data, firstly to extract far more of the biological complexity and dynamic range that remains hidden, and secondly to optimize the acquisition process itself, which is unnecessarily detailed for some regions of the proteome, whilst lacking for others. Methods such as these will significantly increase our understanding of underlying

variations in MS experiments in a clinical setting and provide an enabling pathway ultimately leading to enhanced sensitivity and robustness of these technologies in translational and clinical research.

The promise of clinical proteomics and metabolomics has led to the establishment of several large-scale biomarker discovery facilities worldwide, including the Stoller Biomarker Discovery Centre (SBDC) for proteomics at The University of Manchester, and the UK Phenome Centres for metabolic phenotyping at Imperial College London and the University of Birmingham. The SDBC facility, for example, combines the latest instrumentation and techniques for robust discovery proteomics, candidate set validation on thousands of samples, through to targeted MS of individual biomarkers for clinical assay development. Crucially, as these centers are underpinned by standard operating procedures controlling collection, preparation and analysis across thousands of samples, they will provides us with the unique opportunity to develop statistical modeling tools that borrow strength over the sum complement of all studies analyzed, in order to characterize the biological and technical variation key to biomarker validation and clinical assay development, across the spectrum of health and disease, with technical bias minimized and comparability maximized.

## Acknowledgements

I wish to thank Garth Cooper, Richard Unwin, Simon Hubbard (University of Manchester), Andy Jones, Robert Beynon, (University of Liverpool) and Warwick Dunn (University of Birmingham) for discussions. This contribution was supported by Bilateral BBSRC-NSF/BIO grant BB/M024954/1 and MRC Methodology Research

Programme grants MR/L011093/1 and MR/N028457/1.

## References

Baker, M. (2015). Reproducibility crisis: Blame it on the antibodies. *Nature*, **521** (7552), 274–276. ISSN 0028-0836, 1476-4687. doi: 10.1038/521274a. URL http://www.nature.com/doifinder/10.1038/521274a.

Befekadu, G. K., Tadesse, M. G., Tsai, T., and Ressom, H. W. (2011). Probabilistic Mixture Regression Models for Alignment of LC-MS Data. *IEEE/ACM T. Comp. Biol. and Bioinform.*, **8**(5), 1417–1424. ISSN 1545-5963. doi: 10.1109/TCBB.2010.88.

Benjamin, A. M., Thompson, J. W., Soderblom, E. J., Geromanos, S. J., Henao, R., Kraus, V. B., Moseley, M. A., and Lucas, J. E. (2013). A flexible statistical model for alignment of label-free proteomics data incorporating ion mobility and product ion information. *BMC Bioinformatics*, **14**(1), 364. ISSN 1471-2105. doi: 10.1186/1471-2105-14-364. URL http://www.biomedcentral.com/1471-2105/14/364/abstract.

Bessant, C. (2016). *Proteome Informatics*. ISBN 978-1-78262-428-8. URL http://pubs.rsc.org/en/content/ebook/978-1-78262-428-8#!divbookcontent.

Browne, W. J., Dryden, I. L., Handley, K., Mian, S., and Schadendorf, D. (2010). Mixed effect modelling of proteomic mass spectrometry data by using Gaussian mixtures. *Journal of the Royal Statistical Society C*, **59**(4), 617–633. doi: 10.1111/j.1467-9876.2009.00706.x. URL http://dx.doi.org/10.1111/j.1467-9876.2009.00706.x.

Carin, L., Hero, A., Lucas, J., Dunson, D., Chen, M., Henao, R., Tibau-Piug, A., Zaas, A., Woods, C., and Ginsburg, G. (2012). High Dimensional Longitudinal Genomic Data: An analysis used for monitoring viral infections. *IEEE Signal Processing Magazine*, **29**(1), 108 –123. ISSN 1053-5888. doi: 10.1109/MSP.2011.943009.

Chernushevich, I. V., Loboda, A. V., and Thomson, B. A. (2001). An introduction to quadrupole-time-of-flight mass spectrometry. *Journal of Mass Spectrometry*, **36**(8), 849–865. doi: 10.1002/jms.207. URL http://dx.doi.org/10.1002/jms.207.

Choi, M., Chang, C.-Y., Clough, T., Broudy, D., Killeen, T., MacLean, B., and Vitek, O. (2014). MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics*, **30**(17), 2524–2526. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btu305. URL http://bioinformatics.oxfordjournals.org/content/30/17/2524.

Cox, J. and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotech*, **26**(12), 1367–1372. ISSN 1087-0156. doi: 10.1038/nbt.1511. URL http://dx.doi.org/10.1038/nbt.1511.

Du, P., Stolovitzky, G., Horvatovich, P., Bischoff, R., Lim, J., and Suits, F. (2008). A noise model for mass spectrometry based proteomics. *Bioinformatics*, **24**(8), 1070–1077. doi: 10.1093/bioinformatics/btn078. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/24/8/1070.

Freeman, O. J., Unwin, R. D., Dowsey, A. W., Begley, P., Ali, S., Hollywood, K. A., Rustogi, N., Petersen, R. S., Dunn, W. B., Cooper, G. J. S., and Gardiner, N. J.

(2015). Metabolic dysfunction is restricted to the sciatic nerve in experimental diabetic neuropathy. *Diabetes*, page db150835. ISSN 0012-1797, 1939-327X. doi: 10.2337/db15-0835. URL http://diabetes.diabetesjournals.org/content/early/2015/10/08/db15-0835.

Glaus, P., Honkela, A., and Rattray, M. (2012). Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics*, **28**(13), 1721–1728. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bts260. URL http://bioinformatics.oxfordjournals.org/content/28/13/1721.

Goeminne, L. J. E., Gevaert, K., and Clement, L. (2015). Peptide-level robust ridge regression improves estimation, sensitivity and specificity in data-dependent quantitative label-free shotgun proteomics. *Molecular & cellular proteomics: MCP*. ISSN 1535-9484. doi: 10.1074/mcp.M115.055897.

Henao, R., Thompson, J. W., Moseley, M. A., Ginsburg, G. S., Carin, L., and Lucas, J. E. (2013). Latent protein trees. *The Annals of Applied Statistics*, **7**(2), 691–713. ISSN 1932-6157. doi: 10.1214/13-AOAS639. URL http://projecteuclid.org/euclid.aoas/1372338464.

Hooft, J. J. J. v. d., Wandy, J., Barrett, M. P., Burgess, K. E. V., and Rogers, S. (2016). Topic modeling for untargeted substructure exploration in metabolomics. *Proceedings of the National Academy of Sciences*, **113**(48), 13738–13743. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1608041113. URL http://www.pnas.org/content/113/48/13738.

Hu, Q., Noll, R. J., Li, H., Makarov, A., Hardman, M., and Graham Cooks, R. (2005). The Orbitrap: a new mass spectrometer. *Journal of mass spectrometry: JMS*, **40**

(4), 430–443. ISSN 1076-5174. doi: 10.1002/jms.856.

Kankainen, M., Gopalacharyulu, P., Holm, L., and Orei, M. (2011). MPEAmetabolite pathway enrichment analysis. *Bioinformatics*, **27**(13), 1878–1879. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr278. URL https://academic.oup.com/bioinformatics/article/27/13/1878/185056/MPEA-metabolite-pathway-enrichment-analysis.

Karpievitch, Y., Stanley, J., Taverner, T., Huang, J., Adkins, J. N., Ansong, C., Heffron, F., Metz, T. O., Qian, W.-J., Yoon, H., Smith, R. D., and Dabney, A. R. (2009a). A statistical framework for protein quantitation in bottom-up MS-based proteomics. *Bioinformatics*, **25**(16), 2028–2034. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btp362. URL http://bioinformatics.oxfordjournals.org/content/25/16/2028.

Karpievitch, Y. V., Taverner, T., Adkins, J. N., Callister, S. J., Anderson, G. A., Smith, R. D., and Dabney, A. R. (2009b). Normalization of peak intensities in bottom-up MS-based proteomics using singular value decomposition. *Bioinformatics*, **25**(19), 2573 –2580. doi: 10.1093/bioinformatics/btp426. URL http://bioinformatics.oxfordjournals.org/content/25/19/2573.abstract.

Kim, M.-S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., Madugundu, A. K., Kelkar, D. S., Isserlin, R., Jain, S., Thomas, J. K., Muthusamy, B., Leal-Rojas, P., Kumar, P., Sahasrabuddhe, N. A., Balakrishnan, L., Advani, J., George, B., Renuse, S., Selvan, L. D. N., Patil, A. H., Nanjappa, V., Radhakrishnan, A., Prasad, S., Subbannayya, T., Raju, R., Kumar, M., Sreenivasamurthy, S. K., Marimuthu, A., Sathe, G. J., Chavan, S., Datta, K. K., Subbannayya, Y., Sahu, A., Yelamanchi, S. D., Jayaram, S., Rajagopalan,

P., Sharma, J., Murthy, K. R., Syed, N., Goel, R., Khan, A. A., Ahmad, S., Dey, G., Mudgal, K., Chatterjee, A., Huang, T.-C., Zhong, J., Wu, X., Shaw, P. G., Freed, D., Zahari, M. S., Mukherjee, K. K., Shankar, S., Mahadevan, A., Lam, H., Mitchell, C. J., Shankar, S. K., Satishchandra, P., Schroeder, J. T., Sirdeshmukh, R., Maitra, A., Leach, S. D., Drake, C. G., Halushka, M. K., Prasad, T. S. K., Hruban, R. H., Kerr, C. L., Bader, G. D., Iacobuzio-Donahue, C. A., Gowda, H., and Pandey, A. (2014). A draft map of the human proteome. *Nature*, **509**(7502), 575–581. ISSN 0028-0836. doi: 10.1038/nature13302. URL http://www.nature.com/nature/journal/v509/n7502/full/nature13302.html.

Lavalle-Adam, M., Rauniyar, N., McClatchy, D. B., and Yates, J. R. (2014). PSEA-Quant: A Protein Set Enrichment Analysis on Label-Free and Label-Based Protein Quantification Data. *Journal of Proteome Research*, **13**(12), 5496–5509. ISSN 1535-3893. doi: 10.1021/pr500473n. URL http://dx.doi.org/10.1021/pr500473n.

Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, **12**(1), 323. URL http://www.biomedcentral.com/1471-2105/12/323/.

Liao, H., Moschidis, E., Riba-Garcia, I., Zhang, Y., Unwin, R. D., Morris, J. S., Graham, J., and Dowsey, A. W. (2014). A new paradigm for clinical mass spectrometry analysis based on biomedical image computing principles. Beijing, China, January 2014.

Listgarten, J., Neal, R. M., Roweis, S. T., Wong, P., and Emili, A. (2007). Difference detection in LC-MS data for protein biomarker discovery. *Bioinformatics*, **23**(2), e198–204. doi: 10.1093/bioinformatics/btl326. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/23/2/e198.

Ranjbar, M. R. N., Tadesse, M. G., Wang, Y., and Ressom, H. W. (2015). Bayesian Normalization Model for Label-Free Quantitative Analysis by LC-MS. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **12**(4), 914–927. ISSN 1545-5963. doi: 10.1109/TCBB.2014.2377723.

Renard, B., Kirchner, M., Steen, H., Steen, J., and Hamprecht, F. (2008). NITPICK: peak identification for mass spectrometry data. *BMC Bioinformatics*, **9** (1), 355. ISSN 1471-2105. doi: 10.1186/1471-2105-9-355. URL http://www.biomedcentral.com/1471-2105/9/355.

Rogers, S., Scheltema, R. A., Girolami, M., and Breitling, R. (2009). Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics*, **25**(4), 512–518. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btn642. URL http://bioinformatics.oxfordjournals.org/content/25/4/512.

Serang, O. and Kll, L. (2015). The solution to statistical challenges in proteomics is more statistics, not less. *Journal of Proteome Research*. ISSN 1535-3893. doi: 10.1021/acs.jproteome.5b00568. URL http://dx.doi.org/10.1021/acs.jproteome.5b00568.

Serang, O. and Noble, W. (2012). A review of statistical methods for protein identification using tandem mass spectrometry. *Statistics and its interface*, **5**(1), 3–20. ISSN 1938-7997.

Suvitaival, T., Rogers, S., and Kaski, S. (2014). Stronger findings for metabolomics through Bayesian modeling of multiple peaks and compound correlations. *Bioinformatics*, **30**(17), i461–i467. ISSN 1367-4803, 1460-2059. doi: 10.1093/

bioinformatics/btu455. URL http://bioinformatics.oxfordjournals.org/content/30/17/i461.

Tautenhahn, R., Patti, G. J., Rinehart, D., and Siuzdak, G. (2012). XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data. *Analytical Chemistry*, **84**(11), 5035–5039. ISSN 0003-2700. doi: 10.1021/ac300698c. URL http://dx.doi.org/10.1021/ac300698c.

Tsou, C.-C., Avtonomov, D., Larsen, B., Tucholska, M., Choi, H., Gingras, A.-C., and Nesvizhskii, A. I. (2015). DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nature Methods*, **12**(3), 258–264. ISSN 1548-7091. doi: 10.1038/nmeth.3255. URL http://www.nature.com/nmeth/journal/v12/n3/full/nmeth.3255.html.

Wang, Y., Zhou, X., Wang, H., Li, K., Yao, L., and Wong, S. T. (2008). Reversible jump MCMC approach for peak identification for stroke SELDI mass spectrometry using mixture model. *Bioinformatics*, **24**(13), i407–i413. doi: 10.1093/bioinformatics/btn143. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/24/13/i407.

Webb-Robertson, B.-J. M., Matzke, M. M., Datta, S., Payne, S. H., Kang, J., Bramer, L. M., Nicora, C. D., Shukla, A. K., Metz, T. O., Rodland, K. D., Smith, R. D., Tardiff, M. F., McDermott, J. E., Pounds, J. G., and Waters, K. M. (2014). Bayesian Proteoform Modeling Improves Protein Quantification of Global Proteomic Measurements. *Molecular & Cellular Proteomics*, page mcp.O113.030932. ISSN 1535-9476, 1535-9484. doi: 10.1074/mcp.O113.030932. URL http://www.mcponline.org/content/early/2014/08/16/mcp.O113.030932.

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., and Kuster, B. (2014). Mass-spectrometry-based draft of the human proteome. *Nature*, **509**(7502), 582–587. ISSN 0028-0836. doi: 10.1038/nature13319. URL http://www.nature.com/nature/journal/v509/n7502/full/nature13319.html.