Supporting Information

BayesENproteomics: Bayesian elastic nets for quantification of peptidoforms in complex samples

Venkatesh Mallikarjun[1, 2, 3,*], Stephen M. Richardson[2] and Joe Swift[1, 2, *]

(1) Wellcome Centre for Cell-Matrix Research, University of Manchester, Oxford Road, Manchester, M13 9PT, UK.

(2) Division of Cell Matrix Biology and Regenerative Medicine, School of Biological Sciences, Faculty of Biology, Medicine and Health, Manchester Academic Health Science Centre, University of Manchester, Oxford Road, Manchester, M13 9PL, UK.

(3) Current address: University of Virginia, MR5, University Health System, 415 Lane Road Charlottesville, VA 22908, USA.

* Correspondence: V.M., e-mail, vm9uq@virginia.edu; J.S., e-mail, joe.swift@manchester.ac.uk, telephone, +44 (0) 161 275 1162.

Supporting Information

S-3 **Supplementary Materials and Methods**

S-10 **Supplementary References**

S-13 **Supplementary Figure 1**

Histogram of $max(\hat{R})$ values for all models created with BayesENproteomics with AMI.

S-14 **Supplementary Figure 2**

TPR-FDP comparisons 10x and 1/10x priors in BayesENproteomics analysis of *E. coli*:Human benchmark dataset.

S-15 **Supplementary Figure 3**

Current "gold-standard" methods performance on mouse:human mixed species dataset and analysis of single technical replicate negative control dataset.

S-17 **Supplementary Figure 4**

Heatmaps for proteins with incorrect fold change estimates in mixed mouse:human dataset.

S-18 **Supplementary Figure 5**

PNGase-F-treated dataset analysed using BayesENproteomics with either DGD imputation or without observation weights.

S-19 **Supplementary Figure 6**

Conventional pathway analysis of protein fold changes from a comparison of MSCs isolated from young vs old donors.

.xlsx files **Supplementary Tables S1-3**

Enrichments of Reactome pathways analysed by PANTHER.

Supplementary Materials and Methods

BayesENproteomics:

This method utilises a novel Bayesian linear regression algorithm with elastic net regularisation based on the hierarchical model detailed previously[1] to fit the model detailed in (2). Bayesian methods employ regularisation based on the prior distribution parameters were estimated from, with elastic net regularisation being equivalent to sampling from an intermediate Gaussian/Laplacian prior. The full hierarchical model for a single protein is detailed in equations (1)-(5).

$$\beta|\sigma^2,\tau^2,\lambda_2,\widetilde{\boldsymbol{X}},\widetilde{\boldsymbol{y}} \sim MVN\left(\begin{array}{c}\left(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}}+\boldsymbol{I}_p(\tau^{-2}+\lambda_2)\right)^{-1}\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{y}}, \\ \sigma^2\left(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}}+\boldsymbol{I}_p(\tau^{-2}+\lambda_2)\right)^{-1}\end{array}\right); \quad (1)$$

$$\sigma^2|\beta,\tau^2 \sim IG\left(\frac{(n-1+p)}{2},\ t+\frac{R^TR}{2}+\frac{(\lambda_1\beta)D_{\tau^2}^{-1}\beta}{2}+\frac{\lambda_2\beta^T\beta}{2}\right), t=0.01; \quad (2)$$

$$\tau_j^{-2}|\beta_j,\sigma^2,\lambda_1 \sim IGauss\left(\sqrt{\frac{\lambda_1^2\sigma^2}{\beta_j^2}},\ \lambda_1^2\right). \quad (3)$$

Let $\widetilde{\boldsymbol{X}}$ and $\widetilde{\boldsymbol{y}}$ represent the weighted design matrix and response variable (detailed further in a later section). Let τ^{-2} represent a vector of latent variables (τ_1^{-2} ... τ_p^{-2}) – sampled from the inverse Gaussian (IGauss) distribution in (1) – for each β_j such that larger values of τ_j^{-2} (j = 1…, p) result in β_j being shrunk towards zero. $D_{\tau^2}^{-1}$ denotes a diagonal matrix with elements τ_j^{-2}, j = 1..., p. Residual variance (σ^2) is sampled from an inverse gamma (IG) distribution in (2). To enforce sparsity, we employ two Regularisation hyperparameters, λ_1 and λ_2, with different conditional distributions, specifying LASSO (4) and ridge hyperparameters (5), respectively. Notably, while overall covariance is controlled by λ_1 through its effect on τ_j^{-2}, each β_j is given its own L_2 Regularisation hyperparameter, λ_{2_j}, similar to the LASSO-like "horseshoe" estimator[2]. This leads to smaller coefficients (i.e. "noise") being more aggressively shrunk towards zero compared to larger coefficients (i.e. "signal"), compared to regression using scalar regularisation

hyperparameters that lead to constant shrinkage across all βs. λ_1^2 and λ_2 are sampled from gamma distributions of the form $Gamma(a, b)$, with a posterior mean of $\frac{a}{b}$. All β_j are subject to regularisation.

$$\lambda_1^2|\,\tau^2 \sim Gamma\left(p + a_1,\ b_1 + \frac{1}{2}\sum_{j=1}^{p}\tau_j^2\right), b_1 = 1; \quad (4)$$

$$\lambda_{2_j}|\beta_j, \sigma^2 \sim Gamma\left(1 + a_2,\ b_2 + \frac{\beta_j^2}{2\sigma^2}\right), b_2 = 3. \quad (5)$$

Estimates of parameters were taken as means of Gibbs-sampled posterior distributions made from post-burn-in iterations. Constant values, t and $b_{1,2}$ in (2), (4) and (5) denote priors necessary to prevent variance estimates from approaching zero or infinity. t was set to 0.01 denoting an uninformative prior so as to allow the data to mostly determine variance estimates, whereas $b_{1,2}$ were set to 3 and $a_{1,2}$ were set to equal the number of different types of interactions minus 1 (e.g. specifying a model with both peptide:treatment and peptide:donor interactions will mean $a_{1,2} = 2 - 1 = 1$), increasing the strength of regularisation for particularly complex models and ensuring that even simpler models are still subject to regularisation. These values do not need to be set by the user and have been shown to work for a variety of datasets used here and elsewhere[3,4]. The Python3 version of BayesENproteomics allows the user to specify additional main and interaction β effects as desired. Unless otherwise stated, results are shown using DGD imputation of missing values.

Linear regression comparison implementation. Peptide-based linear regression modelling has previously been shown to possess greater statistical power and accuracy than summarization models when detecting differentially abundant proteins[5]. We compare BayesENproteomics to other peptide-level regression models (detailed below) for calculating differentially abundant proteins and PTMs. Furthermore, while our main aim was to compare different regression implementations, to provide some external reference we also include current "gold-standard" quantification methods, namely protein-level summarization using Tukey's Median Polish as implemented in MSStats[6] (version 3.16.0), and peptide-level quantification using MsqRob[7]. However, specific implementation details may mean that these methods are not completely comparable. For instance, MsqRob does not implement any form of imputation whereas MSStats uses its own "MBimpute" option which uses an Accelerated Failure Time. We show results for MSStats and MsqRob with these default options and with DGD imputation.

Ordinary least squares (OLS):

Differential protein abundance was calculated using the simple linear model shown in (1), using the fitlm Matlab function. For calculating differential PTM abundance, the $\log_2$ fold change for each PTM'd peptide was normalised to the $\log_2$ fold change calculated for parent protein abundance. In cases where a single PTM site was shared by 2 or more peptides (i.e. missed cleavages), the most abundant one was used. Results are shown following DGD imputation of missing values.

Linear mixed-effects models with Huber residual weights (MsqRob and LME-H):

Firstly, we tested the original MsqRob algorithm (on dataframes created by MSStats) as implemented in MsqRobSum[8] (https://github.com/statOmics/MSqRobSum). This ridge regression/mixed-effects algorithm[9], modelled peptides as random effects (i.e. they were assumed to be randomly sampled from a larger population and that they accurately modelled the variance of that population) using the default model in (1) with the addition of sample/run-level βs. MsqRob[7] exploits the link between ridge regression and mixed effects models to assign each β_j a specific penalty, λ_j (where $\lambda_j = \sigma_i^2/\sigma_{ij}^2$, σ_i^2 = residual variance of protein i, σ_{ij}^2 = variance of coefficient β_j for protein $i, j = 1, 2 \ldots p$). Secondly, we recapitulated this algorithm using the fitglme Matlab function, including Huber weighting of residuals to fit the more complex model shown in

equation (2), referred to as LME-H from here on. To calculate changes in PTM abundances, peptide:group and peptide:donor interaction effects ($\beta_{f:g}$, and $\beta_{f:d}$ in (2), respectively) were included as random effects with the size of the resulting interaction βs denoting changes in the abundance of that peptide in response to treatment or donor effects, respectively. Results are shown following DGD imputation of missing values.

Weighting of residuals based on confidence of peptide identification

Identification of PTM'd peptides was performed by the inclusion of variable modifications during peptide database searching. Inclusion of multiple variable modifications was found to increase the number of false positive peptide identifications. The number of false-positive identifications was reduced by discarding peptides with low Mascot scores using a standard FDR cut-off based on identification p-values. We employ a Benjamini-Hochberg FDR[10] cut-off of < 0.2.

BayesENproteomics also employed a novel heuristic outlier weighting scheme that incorporates peptide identification confidence and residual size to identify potentially biologically relevant outliers compared to miss-identified peptides. This effectively weighted against outlier peptides (which may possess biologically relevant PTMs), particularly if confidence in their identification was low. In this case we used Mascot scores as our indicator for peptide identification confidence, similar to the weighting using the Posterior Error Probability implemented in Triqler[11] or the random digest effect in BayesProt[12]. Other Bayesian algorithms previously developed for labelled experiments[13,14] utilise different properties to weight peptides, such as isolation specificity and/or summed MS signal/noise ratios. Unfortunately, these methods are difficult to generalise to label-free experiments using different pre-processing pipelines that may not output the same variables. To increase the generalisability of BayesENproteomics, we constrained our analysis to variables output by most common pre-processing pipelines, namely peptide identification confidence (in this case Mascot scores). Firstly, Mascot scores, S_1..., S_n were scaled by dividing them by a modified Bonferroni-like cut-off (similar to that described on the Mascot website,

http://www.matrixscience.com/help/interpretation_help.html, accessed 23/10/17) and adjusted so that all the highest scoring peptides were weighted equally, as in (6) with values between 0 and 1.

$$\tilde{S} = min\left\{1,\ \frac{S}{(-10(\log_{10}(1/(20\mathrm{N}))) - 13)}\right\},\ \mathrm{N} = \text{total peptides in dataset.} \quad (6)$$

During each iteration of the Gibbs sampler, an n-dimensional vector of weights, w with elements $w_1 ..., w_n$, was calculated using a variation of the automatic outlier detection and weighting method by Ting et al.[15]. Initially, there was no *a priori* reason to exclude - or diminish the influence of - peptides that have passed the initial FDR screen. Instead, we opted to weight in favour of those peptides that either have high Mascot scores or low residuals (7). Transformed Mascot scores in $\tilde{S}$ were used to parameterize a binomial distribution giving $\hat{S}$ that would determine if observations from that were favourably weighted each Gibbs sampler iteration (8).

$$w_i - 1 - \widehat{S}_i|\sigma^2,\ \widehat{S}_i \sim Gamma\left(\widehat{S}_i + \frac{1}{2},\ \frac{1}{2} + \frac{1}{2\sigma^2}R_i^2\right), i = 1 \ldots n; \quad (7)$$

$$\widehat{S}_i|\tilde{S} \sim Binom(1,\ \tilde{S}). \quad (8)$$

Where $\boldsymbol{R}$ is a vector of residuals with elements $R_1 \ldots R_n$, $R = \boldsymbol{y} - \boldsymbol{X}\beta$. Observations were then weighted by multiplying each row of $\boldsymbol{X}$ and each value of $\boldsymbol{y}$ by their respective weight calculated in (7), (9) and (10).

$$\widetilde{\boldsymbol{X}}_{ij} = \boldsymbol{X}_{ij} w_i,\ i = 1 \ldots n,\ j = 1 \ldots p; \quad (9)$$

$$\widetilde{\boldsymbol{y}}_i = \boldsymbol{y}_i w_i,\ i = 1 \ldots n. \quad (10)$$

Where $\widetilde{\boldsymbol{X}}$ and $\widetilde{\boldsymbol{y}}$ are the weighted design matrix and response vector, respectively.

Adaptive Multiple Imputation (AMI)

The proportion of MNR to MAR values is strongly dataset- and even protein-dependent[16], meaning that the optimal choice of distribution to impute from likely differs between datasets and proteins. Many state-of-the-art imputation methods rely on the user deciding whether they think that all missing values in a dataset are MAR or MNR and selecting an imputation method that performs well under these conditions[16,17]. The nature of peptide missingness may vary between proteins, so it is unlikely that one method is suitable for all datasets and all proteins within them. To address this, we employed an adaptive multiple imputation (AMI) strategy, within the Gibbs sampler described in Figure 2, where a logistic regression determines if specific peptides or treatments positively correlate with missingness. Those missing values associated with parameters that showed higher than average, positive correlation with missingness were deemed to be MNR and imputed from a truncated Gaussian distribution, or MAR and imputed from a Gaussian distribution otherwise within the main Gibbs sampler (similar to the model-based imputation previously described[18,19]). AMI uses a logistic regression to determine whether missing values are MAR or MNR and imputes from appropriate conditional distributions. Logistic regression, similar to that employed previously[19], was used to discern whether a given missing value was MAR or MNR as in (11).

$$\mathrm{Z}_{qi} = \log\left(\frac{R_{qi}}{1-R_{qi}}\right) = \theta_0 + X_f\theta_f + X_g\theta_g + \varepsilon_{qfg}. \quad (11)$$

Where R_q represents a binary vector for protein *q* with elements $r_1, \ldots r_n$ denoting whether an observation was missing ($r_i = 1$) or not ($r_i = 0$) and Z_q represents the logit transform of R_q to enable estimation of regressor coefficients θ using linear regression. As $\mathrm{Z}_{qi} \in \{-\infty, +\infty\}$, we set minimum and maximum values for Z_q to -10 (corresponding to a probability for observation i being missing, $r_i < 0.00005$) and 10 ($r_i > 0.99995$). X_f and X_g represent binary design matrices denoting the peptide (*f*) and treatment group (*g*) from which a given observation is derived. θ_j ($j = 1\ldots,(p_f + p_g + 1)$) represents a $1 \times (p_f + p_g + 1)$ vector of regressor coefficients; p_f and p_g represent the number of elements in θ_f and θ_g, respectively. θ_f and θ_g denote whether observations from a given peptide or treatment correlate with values in R_q (i.e. whether

probability of "missingness" increases when looking at intensities from particular peptides or from particular experimental treatment groups). The intercept term, θ_0 denotes the intrinsic probability of missingness for that protein. If $\theta_j > 0$ we inferred that missing observations associated with θ_j were MNR, and MAR otherwise. MAR and MNR missing values were imputed as part of each Gibbs sampler iteration as in (12) and (13).

$$\frac{\tilde{\boldsymbol{y}}_{(MAR)iq}}{w_{(MAR)i-1}} \sim MVN\left(\boldsymbol{X}_{(MAR)}\beta_{i-1}^{T},\ \sigma_{i-1}^{2}\left(\boldsymbol{X}_{(MAR)}\boldsymbol{X}_{(MAR)}^{T}\right)^{-1}\right); \tag{12}$$

$$\frac{\tilde{\boldsymbol{y}}_{(MNR)iq}}{w_{(MNR)i-1}} \sim TGauss\begin{pmatrix}\boldsymbol{X}_{(MNR)}\beta_{i-1}^{T},\ \sigma_{i-1}^{2}\left(\boldsymbol{X}_{(MNR)}{\boldsymbol{X}_{(MNR)}}^{T}\right)^{-1}, \\ a = min\{\boldsymbol{y}_{observed}\} - 2, b = \boldsymbol{y}_{\left\lfloor\frac{MNR}{n}\right\rfloor}\end{pmatrix}. \tag{13}$$

Where $\tilde{\boldsymbol{y}}_{(MAR)iq}$ represents a vector of score-weighted (see below) MAR $\log_2$(intensity) values for protein *q* at the *i*th iteration of the Gibbs sampler. β_{i-1}^{T} is the vector of parameter estimates and σ_{i-1}^{2} is the residual variance from the *i-1*th iteration of the Gibbs sampler, respectively. Missing values in $\tilde{\boldsymbol{y}}_{iq}$ are sampled according to the multivariate normal distribution (MVN) described in (12)[17]. MNR missing values were imputed from a truncated Gaussian (TGauss) with an upper limit determined by the percentile of observed $\log_2$ intensity values corresponding to the fraction of values that are deemed MNR ($\boldsymbol{y}_{\left\lfloor\frac{MNR}{n}\right\rfloor}$) from the missingness regression model in (13). In (13), a and b represent the lower and upper limits of the TGauss distribution. Thus, if 5% of observations for a given protein are determined to be MNR according to (11), then the cut-off for the truncated Gaussian is equal to the 5th percentile of observed $\log_2$ intensity values. New missing values were imputed for each iteration of the Gibbs sampler. The multiple imputation strategy implemented in BayesENproteomics thus accounted for the inherent uncertainty in imputation by basing the resulting β estimates (and subsequent hypothesis testing) on distributional estimates of missing values[20] rather than fixed point estimates as in OLS and LME-H, where single random samples could strongly influence individual protein fold change estimates.

Supplementary References

(1) Kyung, M.; Gilly, J.; Ghoshz, M.; Casellax, G. Penalized Regression, Standard Errors, and Bayesian Lassos. *Bayesian Anal.* **2010**, *5* (2), 369–412. https://doi.org/10.1214/10-BA607.

(2) Makalic, E.; Schmidt, D. F. A Simple Sampler for the Horseshoe Estimator. *IEEE Signal Process. Lett.* **2016**, 23 (1), 179–182. https://doi.org/10.1109/LSP.2015.2503725.

(3) Gilbert, H. T. J.; Mallikarjun, V.; Dobre, O.; Jackson, M. R.; Pedley, R.; Gilmore, A. P.; Richardson, S. M.; Swift, J. Nuclear Decoupling Is Part of a Rapid Protein-Level Cellular Response to High-Intensity Mechanical Loading. *Nat. Commun.* **2019**, *10* (1), 4149. https://doi.org/10.1038/s41467-019-11923-1.

(4) Herrera, J. A.; Mallikarjun, V.; Rosini, S.; Montero, M. A.; Warwood, S.; O'Caulian, R.; Knight, D.; Schwartz, M. A.; Swift, J. Laser Capture Microdissection Coupled Mass Spectrometry (LCM-MS) for Spatially Resolved Analysis of Formalin-Fixed and Stained Human Lung Tissues. *bioRxiv Prepr.* **2019**. https://doi.org/doi.org/10.1101/721373.

(5) Goeminne, L. J. E.; Argentini, A.; Martens, L.; Clement, L. Summarization vs Peptide-Based Models in Label-Free Quantitative Proteomics: Performance, Pitfalls, and Data Analysis Guidelines. *J. Proteome Res.* **2015**, *14* (6), 2457–2465. https://doi.org/10.1021/pr501223t.

(6) Choi, M.; Chang, C.; Clough, T.; Broudy, D.; Killeen, T.; MacLean, B.; Vitek, O. MSstats: An R Package for Statistical Analysis of Quantitative Mass Spectrometry-Based Proteomic Experiments. *Bioinformatics* **2014**, *30* (17), btu305. https://doi.org/10.1093/bioinformatics/btu305.

(7) Goeminne, L. J. E.; Gevaert, K.; Clement, L. Experimental Design and Data-Analysis in Label-Free Quantitative LC/MS Proteomics: A Tutorial with MSqRob. *J. Proteomics* **2018**, *171*, 23–36. https://doi.org/10.1016/j.jprot.2017.04.004.

(8) Sticker, A.; Goeminne, L. J. E.; Martens, L.; Clement, L. Robust Summarization and Inference in Proteome-Wide Label-Free Quantification. *bioRxiv* https://doi.org/10.1101/668863. https://doi.org/https://doi.org/10.1101/668863.

(9) Goeminne, L. J. E.; Gevaert, K.; Clement, L. Peptide-Level Robust Ridge Regression Improves Estimation, Sensitivity, and Specificity in Data-Dependent Quantitative Label-Free Shotgun

Proteomics. *Mol. Cell. Proteomics* **2016**, *15* (2), 657–668. https://doi.org/10.1074/mcp.M115.055897.

(10) Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **1995**, *57*, 289–300.

(11) The, M.; Käll, L. Integrated Identification and Quantification Error Probabilities for Shotgun Proteomics. *Mol. Cell. Proteomics* **2019**, *18* (3), 561–570. https://doi.org/10.1074/mcp.RA118.001018.

(12) Xu, J.; Patassini, S.; Rustogi, N.; Riba-Garcia, I.; Hale, B. D.; Phillips, A. M.; Waldvogel, H.; Haines, R.; Bradbury, P.; Stevens, A.; et al. Regional Protein Expression in Human Alzheimer's Brain Correlates with Disease Severity. *Commun. Biol.* **2019**, *2* (1), 43. https://doi.org/10.1038/s42003-018-0254-9.

(13) O'Brien, J. J.; O'Connell, J. D.; Paulo, J. A.; Thakurta, S.; Rose, C. M.; Weekes, M. P.; Huttlin, E. L.; Gygi, S. P. Compositional Proteomics: Effects of Spatial Constraints on Protein Quantification Utilizing Isobaric Tags. *J. Proteome Res.* **2018**, *17* (1), 590–599. https://doi.org/10.1021/acs.jproteome.7b00699.

(14) Peshkin, L.; Gupta, M.; Ryazanova, L.; Wühr, M. Bayesian Confidence Intervals for Multiplexed Proteomics Integrate Ion-Statistics with Peptide Quantification Concordance. *Mol. Cell. Proteomics* **2019**, *18* (10), 2108–2120. https://doi.org/10.1074/mcp.TIR119.001317.

(15) Ting, J. A.; D'Souza, A.; Schaal, S. Automatic Outlier Detection: A Bayesian Approach. *Proc. - IEEE Int. Conf. Robot. Autom.* **2007**, No. April, 2489–2494. https://doi.org/10.1109/ROBOT.2007.363693.

(16) Lazar, C.; Gatto, L.; Ferro, M.; Bruley, C.; Burger, T. Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. *J. Proteome Res.* **2016**, *15* (4), 1116–1125. https://doi.org/10.1021/acs.jproteome.5b00981.

(17) Webb-Robertson, B.-J. M.; Wiberg, H. K.; Matzke, M. M.; Brown, J. N.; Wang, J.; McDermott, J. E.; Smith, R. D.; Rodland, K. D.; Metz, T. O.; Pounds, J. G.; et al. Review, Evaluation, and

Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry-Based Label-Free Global Proteomics. *J. Proteome Res.* **2015**, *14* (5), 1993–2001. https://doi.org/10.1021/pr501138h.

(18) Karpievitch, Y.; Stanley, J.; Taverner, T.; Huang, J.; Adkins, J. N.; Ansong, C.; Heffron, F.; Metz, T. O.; Qian, W.-J.; Yoon, H.; et al. A Statistical Framework for Protein Quantitation in Bottom-up MS-Based Proteomics. *Bioinformatics* **2009**, *25* (16), 2028–2034. https://doi.org/10.1093/bioinformatics/btp362.

(19) Li, F.; Nie, L.; Wu, G.; Qiao, J.; Zhang, W. Prediction and Characterization of Missing Proteomic Data in Desulfovibrio Vulgaris. *Comp. Funct. Genomics* **2011**, *2011*, 780973. https://doi.org/10.1155/2011/780973.

(20) Schafer, J. L.; Olsen, M. K. Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. *Multivariate Behav. Res.* **1998**, *33* (4), 545–571. https://doi.org/10.1207/s15327906mbr3304_5.

Supplementary Figure 1

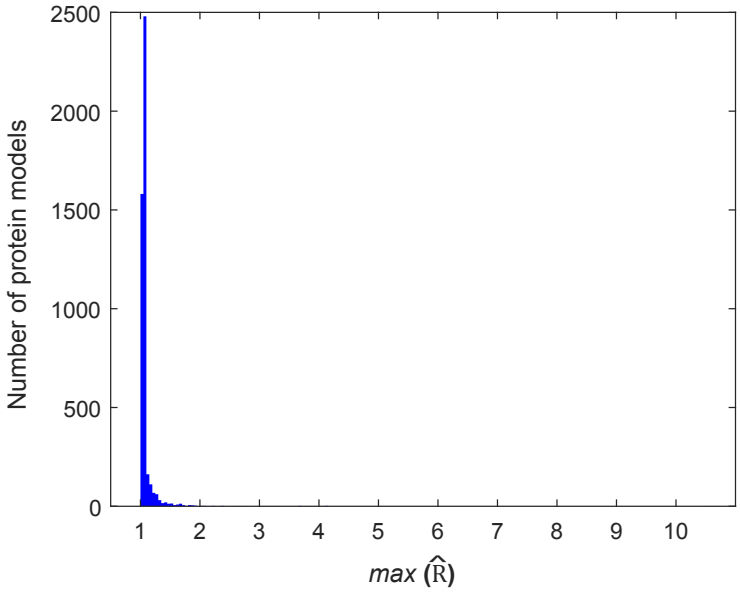

Figure S1: Histogram of $max(\widehat{R})$ values for all models created with BayesENproteomics with AMI. Majority of values were ~1. Maximum value observed among all datasets tested in this manuscript was 10.6.

Supplementary Figure 2

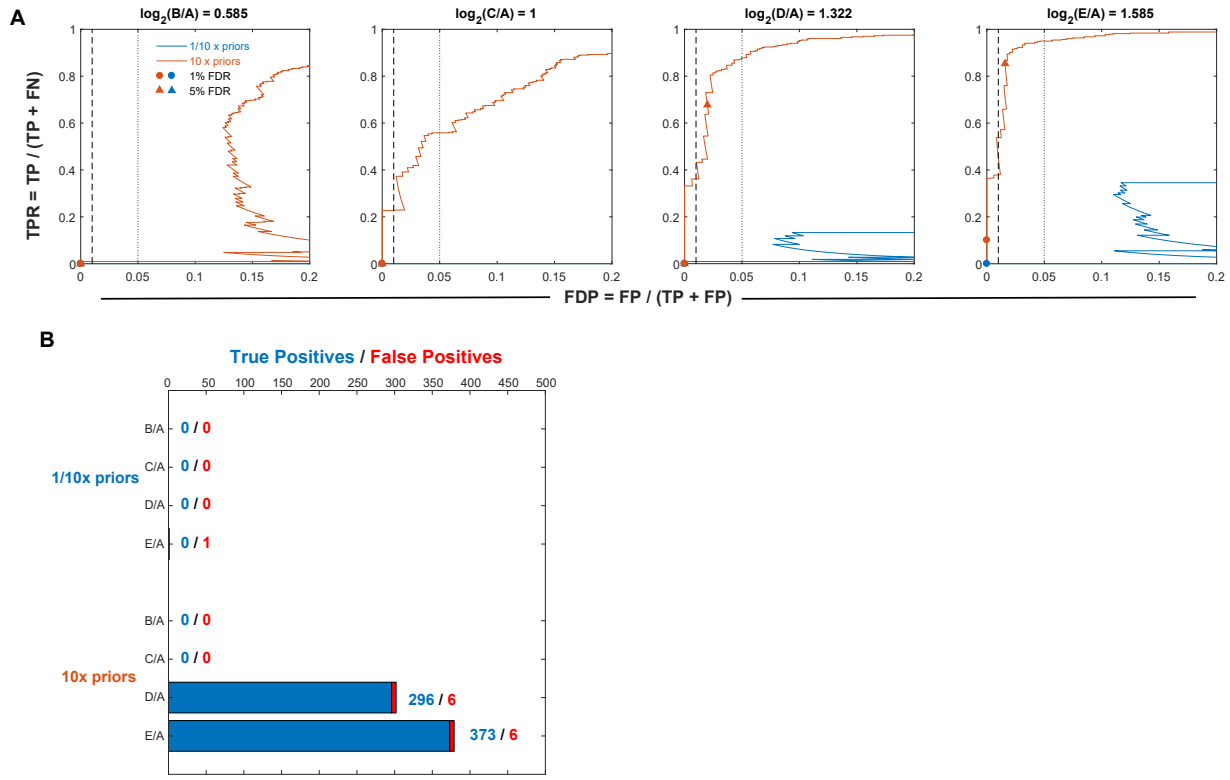

Figure S2: ROC curves for 10x and 1/10x prior values in BayesENproteomics analysis of E. coli:Human benchmark dataset. (**A**) TPR-FDP curves for the indicated comparison created using BayesENproteomics with either $t = 0.001$, $b_{1,2} = 0.3$ and $a_{1,2} = 0.1$ (1/10 x priors) or $t = 0.1$, $b_{1,2} = 30$ and $a_{1,2} = 10$ (10 x priors) with BHFDR-adjusted p-values. Vertical lines indicate FDP = 0.01 (dashed) or 0.05 (dotted). Circles and triangles denote empirical false discovery rate (FDR) when significance threshold (α) is set to 0.01 or 0.05, respectively. FDR can be said to be properly controlled if empirical FDRs are behind the 0.01 or 0.05 FDP verticlal lines. Absent symbols indicate an empirical FDR > 0.2 for that method in that comparison. BayesENproteomics correctly controlled FDR in all but the E/A comparision but still outperformed other methods. However, performance decreased when observation weights were removed. (**B**) Absolute numbers of true and false positives for the indicated comparisons for adjusted p-values < 0.05.

Supplementary Figure 3

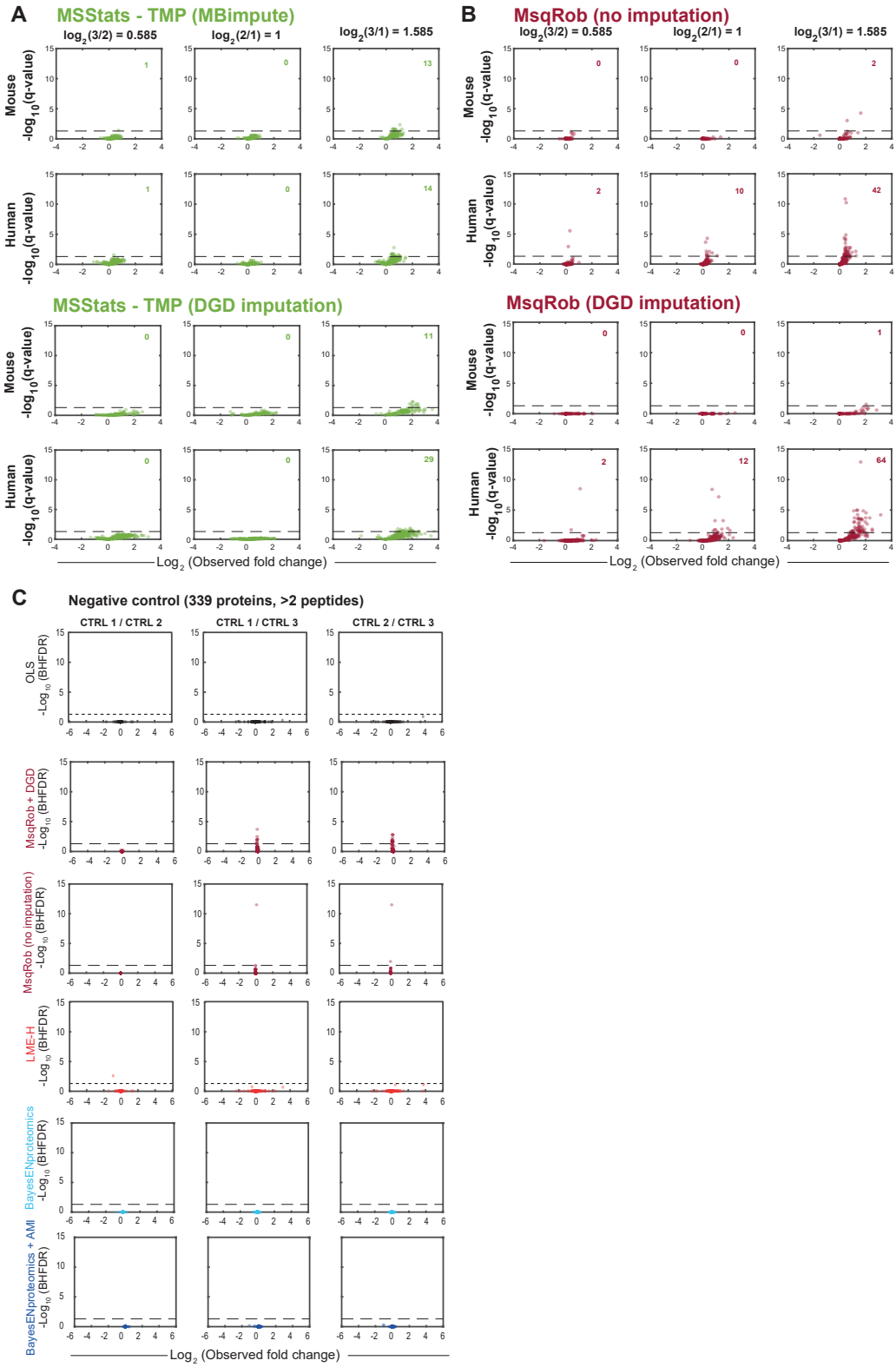

Figure S3: Current "gold-standard" methods performance on mouse:human mixed species dataset and analysis of single technical replicate negative control dataset. (**A, B**) Volcano plots differentially abundant proteins identified by MSStats (TMP) and MsqRob, respectively, using the indicated imputation method. (**C**) Volcano plots showing differentially abundant proteins identified in mouse skin technical replicate negative control dataset. All results use DGD imputation unless otherwise stated.

Supplementary Figure 4

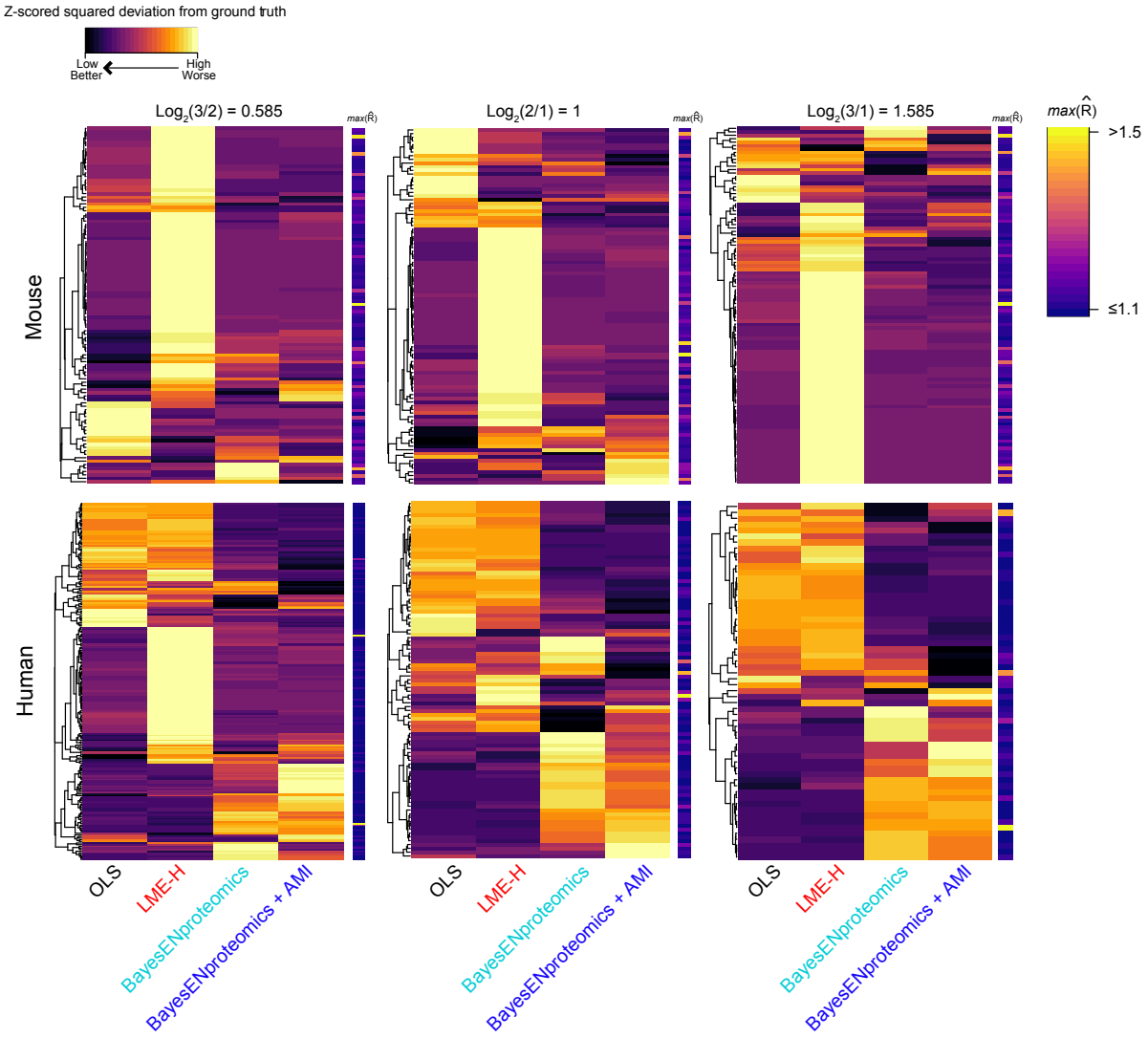

Figure S4: Heatmaps for proteins with incorrect fold change estimates in mixed mouse:human dataset. Heatmaps show z-scored squared deviation from ground truth fold changes for all proteins that gave fold changes with incorrect direction in the indicated comparison when using any of the indicated quantification methods. Columns on the right of each heatmap show $max(\hat{R})$ values for each protein.

Supplementary Figure 5

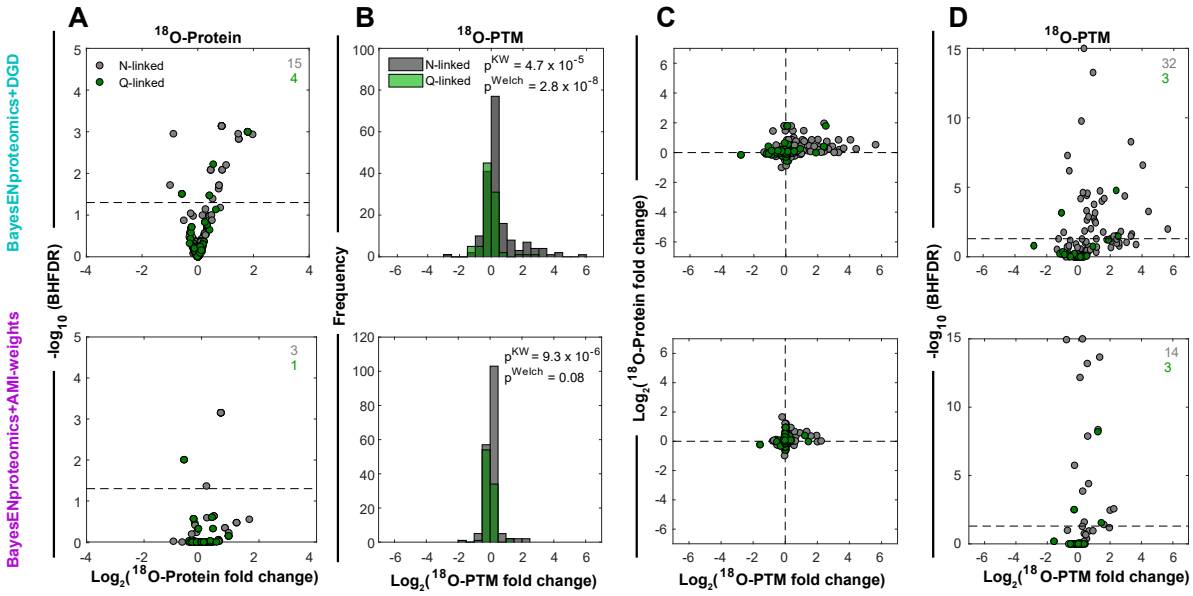

Figure S5: PNGase-F-treated dataset analysed using BayesENproteomics with either DGD imputation or without observation weights. (**A**) Volcano plots showing N-linked (grey) and Q-linked (green) proteins identified as differentially abundant following PNGase-F treatment with BayesENproteomics with DGD imputation (top) or with observation weights removed (bottom). (**B**) Histograms showing (lack of) skewing of N-linked PTMs compared to Q-linked. (**C**) Comparison of N-linked and Q-linked protein versus PTM $\log_2$(fold changes). (**D**) Volcano plots N-linked (grey) and Q-linked (green) PTMs identified as differentially abundant.

Supplementary Figure 6

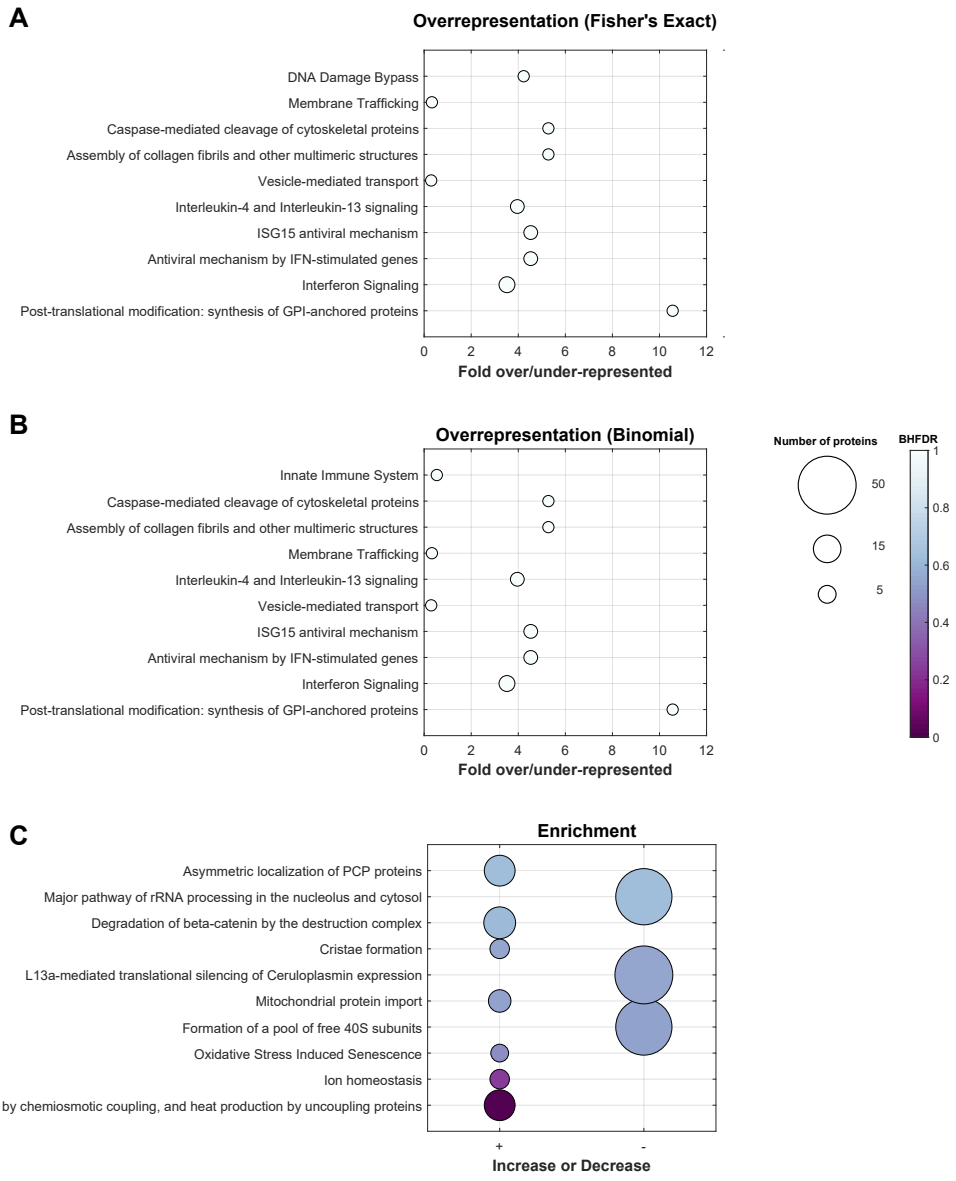

Figure S6: Conventional pathway analysis of protein fold changes from a comparison of MSCs isolated from young vs old donors. Dotplots show top 10 Reactome pathway terms. Size of dots denotes number of proteins assigned to a given pathway. Colour intensity denotes significance. All analyses performed using PantherDB (www.pantherdb.org) (**A**) Overrepresentation analysis using Fisher's Exact tests showed no significant terms. Terms sorted by unadjusted p-values (as BHFDR values all equal 1). (**B**) Overrepresentation analysis using Binomial tests showed no significant terms. Terms sorted by unadjusted p-values (as BHFDR values all equal 1). (**C**) Enrichment analysis. Terms sorted by BHFDR.