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

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

(1) Wellcome Trust Centre for Cell-Matrix Research;

(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, M13 9PL, UK.

* Correspondence: joe.swift@manchester.ac.uk

Running title: *Bayesian elastic nets for proteomics*

Key words: Label-free proteomics / post translational modification / Bayesian elastic nets / primary cells and tissue

Abstract

Multivariate regression modelling provides a statistically powerful means of quantifying the effects of a given treatment while compensating for sources of variation and noise, such as variability between human donors and the behaviour of different peptides during mass spectrometry. However, methods to quantify endogenous post-translational modifications (PTMs) are typically reliant on summary statistical methods that fail to consider sources of variability such as changes in levels of the parent protein. Here, we compare three multivariate regression methods, including a novel Bayesian elastic net algorithm (BayesENproteomics) that enables assessment of relative protein abundances while also quantifying identified PTMs for each protein. We tested the ability of these methods to accurately quantify expression of proteins in a mixed-species benchmark experiment, and to quantify synthetic PTMs induced by stable isotope labelling. Finally, we extended our regression pipeline to calculate fold changes at the pathway level, providing a complement to commonly used enrichment analysis. Our results show that BayesENproteomics can quantify changes to protein levels across a broad dynamic range while also accurately quantifying PTM and pathway-level fold changes.

Introduction

Spiralling costs of drug/therapeutic development and a low probability of success have driven an increase in the use of models based on patient-derived material in an attempt to determine translational potential prior to clinical studies. However, unlike samples from genetically homogeneous, completely inbred model organisms, patient samples can possess substantial differences between individuals. This variability between samples can make discerning a statistically significant effect extremely difficult using the standard statistical methods commonly used in biology. The lack of statistical power caused by poor signal-to-noise ratios in primary samples and the difficulty in obtaining large donor cohorts can compromise potentially promising biological studies, resulting in wasted time and prohibitive expenditure.

The problem of high inter-individual differences is rarely more acute than in high-throughput omics experiments. The difference between patient samples is particularly evident in omics experiments where subtle differences in the levels of individual features within samples (e.g. transcripts, peptides derived from proteins, etc.) are compounded over many such features, demonstrating quantifiable inter-individual differences even between genetically identical twins (Brodin et al, 2015). This is especially relevant to typical mass spectrometry (MS) proteomic methods wherein quantities associated with multiple features (i.e. peptides) are used in the calculation of a fold change for a single protein.

In bottom-up MS-based proteomics, proteins are enzymatically digested into peptides and the abundances of these individual peptides are used to derive protein quantifications. A core assumption is that peptide abundance is proportional to protein abundance. However, each peptide has different physico-chemical properties that mean they each behave differently during sample preparation or within the mass spectrometer itself. For instance, residues surrounding cleavage sites can affect efficiency of enzymatic digestion (Rodriguez et al, 2008); some peptides

also ionise less efficiently than others (Abaye et al, 2011) and in data-dependent analysis, peptides of similar mass-to-charge (m/z) ratios may compete during co-elution or ionisation (Schliekelman & Liu, 2014), potentially biasing the peptides finally detected. These discrepancies can be further exacerbated by biologically relevant post-translational modifications (PTMs) that alter the behaviour of individual peptides, resulting in substantial differences in measured intensities for peptides that belong to the same protein, even in purified protein samples. Even with normalisation, differences in the behaviour of individual features between donors and experimental treatments are difficult to account for.

It was recently shown that combining MS datasets from multiple fractions and extraction conditions could achieve near-complete coverage of the human cell proteome, including >14000 protein isoforms (Bekker-Jensen et al, 2017). Interestingly, the same study also identified ~10000 phosphorylation sites and ~7000 acetylation sites without specific enrichment. New sample preparation techniques and the increasing sensitivity of MS instrumentation has increased coverage of PTMs. However, with this increase in sensitivity comes the problem of resolving differential regulation of different forms of the same protein (proteoforms), as not all peptides belonging to a given protein respond in the same way to a given treatment, an assumption made by many commonly used quantification methods.

Common statistical methods for dealing with variability include the summarisation of peptide intensities using averages, sums or medians to give subject-level protein quantification (Goeminne et al, 2015). However, these subject-level summary models fail to account for differences in behaviour of peptides belonging to the same protein during analysis, and thus are prone to perturbation by outlier peptides. They are also highly dependent on obtaining large numbers of peptides from each protein and thus can suffer from low statistical power when analysing less extreme fold changes in small or low abundance proteins (e.g. transcription

factors). To address this, previous statistical models have focused on exclusion of outlier peptides to improve protein quantification (Forshed, 2013; Swift et al, 2013a), at the cost of a loss of information relating to different proteoforms for a given protein; or have attempted to deduce differences in the abundance of peptides belonging to different sub-populations of the same protein (Henao et al, 2013; Webb-Robertson et al, 2014; Zeng et al, 2017).

For proteomics, linear regression models represent a powerful means of teasing apart different, known sources of variability in peptide intensities, with the aim of returning the variability caused by a given experimental condition. Linear modelling has become the standard for transcriptomic analysis thanks to the widely used LIMMA package (Smyth, 2004) and has been successfully adapted for MS data analysis. For example, in the peptide-based model proposed by Choi et al (2014) with donor effects added:

$$y_{fgd} = \beta_0 + X_f\beta_f + X_g\beta_g + X_d\beta_d + \varepsilon_{fgd} \quad (1)$$

Where, for a given protein, y_{fgd} (the **response variable**) corresponds to the observed $\log_2$ intensity for peptide *f* of that protein under experimental condition *g* from donor *d*. **Predictor variables** β_f, β_g and β_d correspond to the effect sizes of peptide *f*, group *g* and donor *d*, respectively. β_0 represents the intercept term. ε corresponds to a Gaussian error term centred on 0 with a variance (σ^2) specific to each protein. The purpose of the regression calculation was to find how much the slope of the resulting line differed for a given experimental treatment (i.e. the effect size β_g caused by the experimental treatment of interest *g*). In this instance the design matrix $\boldsymbol{X}$ consisted solely of categorical variables encoded as binary variables of size $n \times p$ (where n = number of observations and p = number of βs).

However, peptide-based linear models can be prone to overfitting, wherein predictive power is lost as the model attempts to fit the noise within the data rather than any overall trends. This may be due to the effects of outlier peptides, such as those that possess a biologically relevant PTM with a fold change different to that of its parent protein, or simply due to peptide mis-identification. The effects of outlier peptides on the final linear model fit can be mediated by various weighting strategies (e.g. see Goeminne et al, 2016). However, for modelling changes in PTM abundance, it would be useful to obtain effect sizes for peptides that specifically interact with a given treatment and separate them from effect sizes for other parameters that may be interacting with these peptides. Linear modelling is more flexible than the name suggests and can also handle interactions between predictor variables; for example, where an experimental condition has an effect on the intensity of an individual peptide – independent of its effect on the protein as a whole – possibly due to some biologically important PTM of that specific peptide. For example, we can extend the model in (1) to allow separation of peptide:treatment and peptide:donor effects in studies using primary human samples, as in (2) (see also the model in Clough et al, 2009).

$$y_{fgd} = \beta_0 + X_f\beta_f + X_g\beta_g + X_d\beta_d + X_fX_g\beta_{f:g} + X_fX_d\beta_{f:d} + \varepsilon_{fgd} \qquad (2)$$

Wherein y_{fgd} corresponds to the intensity of peptide *f*, in experimental group *g*, from protein *i*, extracted from technical replicate *r* of donor *d*. The βs correspond to the fitted coefficients for each predictor variable. Note the interactions coefficients $\beta_{f:g}$, and $\beta_{f:d}$ in (2), denoting variability in individual peptide intensities, *f*, caused by the experimental treatment *g* or largely unknown features of the individual donor *d*. Unfortunately, overfitting becomes particularly acute in models that attempt to deal with many potential sources of variability, some of which may interact or correlate with one-another, and where the number of parameters to be estimated exceeds the number of observations. The inclusion of interaction terms in a model has potential to provide

important biological insight. However, interaction terms may not be necessary for all proteins in a dataset and the decision whether or not to include them would need to be made on a protein-by-protein basis.

Simpler models (e.g. those used by Choi et al, 2014, and Goeminne et al, 2016) minimise the potential for overfitting found in more complex models. However, when many potential sources of variability exist it may be necessary to fit increasingly complex models wherein the choice of terms included, particularly any interaction terms, may need to be decided on a protein-by-protein basis. Failure to select the correct terms could result in models failing to fit correctly, or even at all, especially if complex interaction terms are included when not needed. However, not including these terms may mean that subtle differences in peptide behaviour that occur only in specific proteins may go unaccounted for, resulting in a loss of accuracy. Thus, a means of automatically selecting appropriate features from a more complex model for use in fitting a final, sparse model is necessary.

Regularization can be employed to minimise overfitting and perform feature selection. Regularization involves attaching penalty weights to each coefficient (β), minimising its contribution to the final fit. The nature of the penalty can differ depending on the type of regularization used. In the ridge regression algorithm used by Goeminne et al (2016) an "L_2" penalty is applied wherein βs are shrunk according to their squared norm, minimising the contribution of less important coefficients in explaining the total model variance. Although ridge regression shrinks coefficient estimates down to emphasise the most important βs, all βs remain non-zero. Another form or regularization is LASSO (least absolute shrinkage and selection operator), which shrinks βs according to their "L_1" absolute norm. With LASSO, some βs can be set to exactly zero, excluding them from the model and giving rise to so-called "sparse" solutions. Several variations of these methods have been proposed, including an "elastic net" variation that

combines properties of both ridge regression and LASSO to perform both model selection and shrinkage of remaining non-zero βs (Zou & Hastie, 2005). These frequentist regularization methods have Bayesian parallels depending on the prior distributions the β posteriors are sampled from, with L_2 regularization being equivalent to sampling of βs from a Gaussian distribution, L_1 being equivalent to sampling from a Laplace (double exponential) prior and the elastic net being equivalent to β sampling from an intermediate distribution (Bornn et al, 2010).

Here we develop a Bayesian elastic net algorithm to provide regularised protein-specific models (BayesENproteomics), taking into account potential donor variability and interactions between specific peptides and experimental conditions or donor effects. This has the advantage in implicitly allowing for any individual peptide to behave differently from its identified protein group in response to treatment or donor effects. This means that BayesENproteomics does not assume that peptides have been identified correctly – a point we enforce by incorporating observation weights based on peptide identification confidence. However, unlike previous Bayesian regression models that attempt to accommodate variable peptide behaviour within a protein (Henao et al, 2013; Webb-Robertson et al, 2014), BayesENproteomics can still provide fold change estimates for protein groups containing peptides with specific identifications, maintaining interpretability of subsequent results by presenting quantification of the “dominant” proteoform for that protein (separating different isoforms if unique peptides are available), whilst also presenting values for differentially abundant PTMs for further investigation. Finally, we extend our linear modelling pipeline to encompass pathway analysis as a complement to standard enrichment analysis.

Results

BayesENproteomics accurately quantifies protein fold changes in mixed-species samples

To produce a benchmark dataset with known (ground truth) fold changes, peptides prepared from primary human mesenchymal stem cell (MSC) lysates and female C57/BL6 mouse skin were mixed in ratios of 3:1, 1:1 and 1:3 and analysed with a Q Exactive HF mass spectrometer (Figure 1A). Mixed species datasets have been used previously to validate protein quantification methods (Swift et al, 2013a), but the method has practical applications, such as when quantifying protein content in xenograft models (Ivanovska et al, 2017; Swift et al, 2013b). Mixed mouse/human samples provide a difficult problem for prospective quantification algorithms as approximately 70% of tryptic peptides are shared between mouse and humans and thus have to be discarded to obtain estimates for specific proteins. This meant that the quantification algorithms examined here had a limited number of observations with which to work. Secondly, the dynamic range of absolute protein abundances was derived from actual biological samples and so provided an accurate report of the error associated with estimating fold changes in low abundance proteins. Finally, this experiment yielded twin datasets (mouse and human); one with pronounced donor-donor variability (human cells) and one where all animals had the same genotype and environment (mouse skin), and so could be adequately explained by the model in equation (2) without donor effects. We reasoned that an algorithm that could account for donor-donor variability should be as accurate on the human dataset as on the mouse dataset.

Logged signals from species-unique peptides were normalised to the medians of the opposite species-unique peptides to provide the observed $\log_2$ ratios used in quantification. To automatically select appropriate β parameters it was possible to employ regularization strategies such as ridge, LASSO or elastic net regression to select the best set of βs by shrinking unimportant βs to zero (Figure 1B).

As LASSO has been shown to lack consistency when selecting from several correlated βs, we chose to use elastic net-based regularization as the introduction of L_2 regularization has been shown to introduce consistency to LASSO estimates by inducing grouping of correlated βs (Zou & Hastie, 2005). However, the standard formula for calculating standard errors (SE) of β estimates ($\sigma^2\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, where σ^2 is the residual variance and $\boldsymbol{X}$ is the design matrix, with superscript T denoting the transpose of $\boldsymbol{X}$ and SEs for each β, lie on the main diagonal of the resulting square matrix) was not suitable for sparse regression models due to shrinking of variance estimates by regularisation. This resulted in underestimation of variance and incorrectly lower SEs and p-values. While there were no simple ways to calculate SEs of coefficients in sparse regularised models, bootstrapping has commonly been used, however this is computationally intensive to do for hundreds of proteins. Bootstrapped SE estimates also differ from iteration to iteration, due to the inherent randomness of bootstrapping, thus compromising reproducibility. Bayesian regression methods addressed this limitation by estimating parameters as distributions for which summary statistics (e.g. means, standard deviations and confidence intervals) were trivial to calculate. Bayesian regression methods were also more flexible in how the model parameters and hyperparameters could be defined. This made Bayesian regression methods generally better suited to fit non-linear models, such as in equation (2), which frequentist methods may have struggled to fit. To implement this non-linear regression we used a Monte Carlo Markov Chain (MCMC)-based Gibbs sampler to repeatedly sample from and update the prior conditional distributions for the parameters we wished to estimate (Figure 1C). Examples of the resulting chains are shown in Figure 1D.

MCMC-based methods can be used to deduce any unknown value, including missing values in proteomics experiments, provided appropriate conditional distributions can be provided. We used this property to sample from two prior distributions depending on whether a missing value was determined to be missing at random or missing non-randomly by logistic regression (see

Materials and Methods section for more details). As the majority of missing values in bottom-up proteomics are assumed to be left censored, it is common to perform imputation by sampling from a down-shifted Gaussian distribution. Missing values that were imputed separately for each protein by BayesENproteomics possessed a similar overall distribution to those imputed from a down-shifted Gaussian distribution as in Tyanova et al (2016), but with a greater number of values closer to the observed MS1 intensity mean, as expected (Figure 2A-C). This showed that BayesENproteomics, even when given the ability to decide otherwise, identified most missing values as missing non-randomly due to peptide abundances being below the limit of detection.

Figure 3A shows estimation of $\log_2$ fold changes for non-differentially abundant proteins in pairwise comparisons between three technical replicates from the same (mouse skin) sample, comparing the BayesENproteomics algorithm with ordinary least-squares (OLS) and linear mixed-effects (with Huber residual weights, LME-H) models (see Materials and Methods section). All methods correctly determined $\log_2$ fold change values clustered around zero and that no changes reached statistical significance (p-value of false discovery, with Benjamini-Hochberg correction, BHFDR < 0.05). In contrast, estimation of differentially abundant mouse (Figure 3B) and human (Figure 3C) proteins in our mixed human:mouse dataset showed larger variation in observed fold change values, with a number of proteins giving consistently incorrect fold change directions, regardless of the regression method. OLS and LME-H models had difficulty detecting any significantly differentially abundant proteins for expected fold changes < 3. In contrast, BayesENproteomics correctly identified more proteins as significantly differentially abundant in all comparisons (Figure 3B, C). Regularization and weighting of residuals in LME-H did not markedly improve accuracy of fold change estimates compared to standard OLS models, although this may be due to the more complex model (2) that LME-H was asked to fit. BayesENproteomics showed decreased coefficient of variance (CV) values for differentially abundant proteins in the mixed human:mouse dataset compared to both OLS and LME-H (Figure 3D).

BayesENproteomics correctly detects increase in stable isotope label following PNGase treatment in $H_2{}^{18}O$

Protein function and activity is strongly determined by diverse PTMs, giving rise to different proteoforms within a given protein population. MS-based proteomics represents a powerful means to systemically interrogate changes in PTM abundance. However, determining relative PTM fold changes from PTM'd peptide abundance is dependent on accurate protein quantification (Wu et al, 2011). While it is possible to calculate protein abundance from only unmodified peptides, this ignores the fact that some proteins are constitutively modified as part of their normal maturation pathway (e.g. many extracellular matrix proteins are heavily glycosylated prior to secretion) and excluding these PTM-containing peptides from protein quantification would exclude a large proportion of all proteoforms present for that protein. While methods have been developed to analyse individual proteoform abundance in multiplexed, labelled experiments (Malioutov et al, 2017), labelling is not always feasible and methods for quantifying differentially abundant PTMs in label-free experiments lag behind. Furthermore, while an ideal scenario would be to directly compare modified to unmodified peptide ratios (Tsai et al, 2017), this is not always achievable in label-free proteomics of complex samples. While linear regression modelling has been shown to impart greater power and accuracy to protein quantification compared to summary statistical methods (Goeminne et al, 2015), applying it to study PTMs is difficult due to the requirement for complex non-linear models that may be subject to overfitting.

To interrogate how well the three algorithms implemented here were able to detect differentially abundant PTMs, we developed a benchmark dataset using PNGase F to introduce a stable isotope (^{18}O) label on N-linked glycosylated residues (Figure 4A). We reasoned that PNGase F-treated peptides should show an increase in N-linked ^{18}O labelling compared to spontaneous deamidation observed in control peptides incubated in $H_2{}^{18}O$ buffer alone. To test if our PNGase F treatment of peptides had worked correctly (independently of any subsequent quantification

methods) we first searched treated and control samples separately and counted the total number of ^{18}O-labelled spectra identified (Figure 4B). This analysis showed that more ^{18}O-labelled spectra were identified in the PNGase F-treated samples than controls, which contained > 100 spectra presumably from spontaneous non-enzymatic deamidation. We expected that most spontaneously deamidated peptides would be unaffected by PNGase F treatment and that we would only observe an increase in relative ^{18}O-peptide abundance or parent ^{18}O-protein abundance for PNGase F-responsive peptides. Furthermore, as spontaneous deamidation can occur to both Q and N residues, resulting in ^{18}O-E and ^{18}O-D residues, respectively, while PNGase F catalyses only deamidation of N-linked glycosylation sites, comparison to fold changes of Q-linked deamidation provided an ideal negative control.

To detect differentially abundant PTMs we employed three statistical methods. The first was a simple summary statistical extension to the OLS protein quantification method wherein $\log_2$ fold changes for donor-normalised PTM'd peptide intensities were normalised to the $\log_2$ fold changes of their parent protein (calculated by OLS). The second involved using the ridge regression performed by LME-H to fit the non-linear model described in equation (2) where peptide:group and peptide:donor interaction βs were entered as random intercept terms to limit overfitting. Finally, we used BayesENproteomics to fit the non-linear model described in (2) to calculate PTM fold changes. From simple cell lysates, following PNGase F treatment without enrichment we found 261 ^{18}O-labelled sites (180 N-linked, 81 Q-linked) in 172 proteins. Histograms in Figure 4C showed that OLS (top-left) and LME-H (middle-left) could not discern an increase in N-linked ^{18}O-protein abundance following PNGase F treatment compared to Q-linked (all p-values > 0.05). In contrast, BayesENproteomics was able to correctly discern a positive skewing of the N-linked ^{18}O-protein histogram compared to that of the Q-linked proteins (all p-values < 0.05, Figure 4C, bottom-left).

Similarly, when measuring differentially abundant PTMs there was great disparity between the three methods. The summary statistical extension to OLS produced noisy estimates of PTM fold changes and was unable to discern an increase in N-linked ^{18}O-PTMs following PNGase F treatment compared to Q-linked (Figure 4C, top-right; 4D, top). LME-H displayed overly strong regularization of interaction terms, resulting in the majority of PTM fold change estimates being approximately zero (Figure 4C, mid-right; 4D, middle). BayesENproteomics estimates of N-linked ^{18}O-PTM fold changes were correctly positively skewed (Figure 4C, bottom-right; 4D, bottom) and significantly different from Q-linked ^{18}O sites. This suggested that BayesENproteomics could quantify PTM fold changes more accurately than either normalised summary statistics or mixed-effects models.

Pathway analysis using linear modelling (PALM)

We also applied a Bayesian regression model to examine treatment effects at the pathway level (Henao et al, 2013) (Figure 5A). Theoretically, linear regression modelling can have several advantages over classical pathway enrichment analysis. Firstly, the PALM approach proposed here utilised the entire dataset and so was not affected by potentially arbitrary statistical cut-offs used in defining 'interesting' and 'background' sets of enrichment analysis of proteins. Secondly, PALM used additional information regarding the direction and magnitude of fold changes and the error regarding their estimation to provide an overall estimate of how a given pathway behaved based on information available in a given dataset. Figures 5B and 5C show volcano plots assembled from pathway quantification from the human:mouse mixed species dataset based on protein-level output from OLS, LME-H and BayesENproteomics. While the magnitude of pathway-level $\log_2$ fold changes was typically not equal to their protein components due to being inferred from relatively fewer observations and down-weighted according to uncertainty in initial protein fold change estimates (always > 0), directionality was conserved leading to an estimate of

whether that pathway significantly increased or decreased. Pathway significance rose in accordance with the magnitude of their component protein fold changes (Figures 5B, C).

Discussion

Multivariate regression methods represent a powerful means to deconvolute sources of variance in complex experiments. Here we attempted to design a statistical method that was able to accurately quantify protein and PTM fold changes in label-free proteomics experiments performed on heterogeneous primary human samples. Others have previously shown the superiority of linear regression-based effect size estimates over summary statistics (Clough et al, 2009; Goeminne et al, 2015). Here we outlined a weighted Bayesian regression algorithm with elastic net regularisation to cope with differentially abundant PTMs as well as the high variability (and generally low availability) of samples from human donors, while still maintaining high interpretability of the resulting models. In particular, we utilised a novel weighing method that took into account confidence in peptide identification to improve quantification accuracy. We demonstrated the accuracy and specificity of the algorithm using serial mixed species datasets and stable isotope labelling using peptides obtained from multiple human donors. Characterisation of complex samples is notoriously complex due to ionisation suppression and detection masking by peptides with similar m/z ratios often leading to underestimation of larger fold changes (this study and Goeminne et al, 2016). The mixed species dataset used here represented an especially difficult case with most (~70%) peptides rendered unusable due to being shared between the two species. While all methods tended to underestimate fold change magnitude (likely due to conflicting peptides in data-dependent analysis, exacerbated in mixed species analysis), we showed that BayesENproteomics was better able to confidently identify differentially abundant proteins and PTMs than simpler regression methods.

Importantly, the use of Gaussian priors for estimating β parameters presumed that logged fold changes lay on a normal distribution. If it was suspected that this was not the case, BayesENproteomics may not be a suitable method of analysis. Furthermore, one limitation to our approach regarding PTM quantification was that each PTM was quantified independently which may have lead to an over-estimation of differences between co-occurring PTMs. Currently, BayesENproteomics does not utilise shared peptides, an approach that others have shown improves accuracy (Blein-Nicolas et al, 2012) at the cost of increased computation time. Use of empirical Bayes correction of variances (as in all methods tested here) can increase power by borrowing information from all proteins/PTMs without having to fit a model for all proteins simultaneously (as would be required to allow the use of shared peptides) for the purposes of calculating significance. MCMC samplers are known to be particularly computationally intensive. On an especially rich dataset, such as those generated by a modern Q Exactive HF mass spectrometer, analysis using BayesENproteomics can take several hours to process on a standard desktop computer. This is approximately three times longer than LME-H and several times longer than OLS regression.

Linear models are extremely customisable, and indeed, can require significant tailoring to best analyse complex multivariate experiments. Without sufficient background knowledge it can be very difficult to anticipate all potential sources of variability that need to be accounted for in any putative statistical model. Nevertheless, BayesENproteomics showed negligible perturbation by inclusion of extraneous parameters and so is well-suited for complex experimental setups that may include different batches, fractions, serial extractions, instruments and donors. Regularization hyperparameters for additional effects can be sampled from the conditional distributions detailed here.

Finally, we extended our linear modelling pipeline to encompass pathway analysis. In transcriptomics experiments it is common to achieve coverage of the entire transcriptome (> 20000 transcripts in a given human cell/tissue), where hundreds to thousands may be differentially regulated in response to any given treatment. These large numbers make enrichment analysis using χ^2 or Fisher's exact tests well-suited to analysing differentially regulated transcriptomic processes. In contrast, proteomics experiments typically cover only a fraction of the total proteome (< 3000 proteins out of a potential > 20000 in a given human cell/tissue, increased further when considering multiple proteoforms) where only a hundred or fewer proteins may be differentially regulated, making it difficult to discern statistically significant process enrichment and erroneously giving the impression that "nothing is happening", often despite obvious morphological or metabolic changes assessed by other methods. Attempting to model pathway-level changes as a function of component parts for transcriptomics has previously shown that incorporating fold change estimates can increase specificity for discriminating pathway-level changes (Ozerov et al, 2016). Although, it should be noted that the aim with PALM is not biomarker discovery (for which more stringent statistical cut-offs are needed) but rather to simply distil a complex dataset with many proteins down to fewer, more readily understandable pathways by assessing what processes are represented and what they do in response to a given treatment. Here we chose to use Reactome pathway annotations (Fabregat et al, 2018), but alternative protein-pathway annotations could be used.

With the improvement of MS technology, quantification of increasingly rare PTM'd peptides, and thus their parent proteoforms, has become possible even without enrichment. Many quantification methods have discarded or otherwise diminished the influence of these potentially relevant peptides in bottom-up protein quantification. Here, the use of regularised interaction coefficients facilitated the reporting of differentially abundant PTMs and the exclusion of potentially misidentified peptides from protein (and subsequent pathway) fold change estimates.

BayesENproteomics was able to account for differences in PTM'd peptide behaviour whilst increasing accuracy of protein quantification in benchmark datasets with limited observations and inter-donor variability. Importantly, interpretability of the resulting models was maintained by reporting individual fold changes for each protein, PTM or pathway represented in the dataset.

Materials and Methods

Cell culture. For the mixed species benchmark experiment, primary human mesenchymal stem cells (MSCs) were isolated from the bone marrow (knee and hip) of a single donor with informed consent and ethical approval, using established protocols (Strassburg et al, 2010). All experiments were performed in accordance with relevant guidelines and regulations, and with National Research Ethics Service and University of Manchester approvals. MSCs were cultured on tissue culture treated polystyrene (TCTP) in low-glucose DMEM with pyruvate (Thermo Fisher Scientific) supplemented with 10% fetal bovine serum (FBS, Labtech.com) and 1% penicillin/streptomycin cocktail (PS, Sigma-Aldrich).

Preparation of primary human MSCs lysates. Cells were lysed in 25 mM ammonium bicarbonate (AB) buffer containing 1.1% sodium dodecyl sulfate (SDS, Sigma), 0.3% sodium deoxycholate (Sigma), protease inhibitor cocktail (Sigma), phosphatase inhibitor cocktails (Sigma). Six 1.6 mm steel beads (Next Advance) were added to the tube and samples were homogenised with a Bullet Blender (Next Advance) at maximum speed for 2 minutes. Homogenates were cleared by centrifugation (12 °C, 10000 rpm, 5 minutes).

Preparation of mouse skin lysates. Skin from 3 month old female C57/BL6 mice was a gift from the Hardman group (University of Manchester; all animal work was performed under UK Home Office and local ethics committee approval). Ventral skin was shaved and scraped, then dissected into 1 mm^3 pieces. 100 µL of 8 M urea (Fisher Scientific) in 25 mM AB containing 25 mM dithiothretol (DTT, Sigma), protease inhibitor cocktail (Sigma), sodium fluoride (Sigma) and sodium orthovanadate (Sigma) was added to the tissue sections with six 1.6 mm steel beads (Next Advance). Samples were then homogenised with a Bullet Blender (Next Advance) at maximum speed for 3 minutes. Resulting homogenates were cleared by centrifugation (12 °C, 10000 rpm, 5 minutes).

Preparation of tryptic peptides. Immobilised trypsin beads (Perfinity Biosciences) were suspended in 150 µL of digest buffer (1.33 mM $CaCl_2$, Sigma, in 25 mM AB) with 50 µL of cell or tissue lysate and shaken at 1400 rpm overnight at 37 °C in a thermocycler. The resulting digest was then reduced by the addition of 4 µL of 500 mM DTT (Sigma, in 25 mM AB; 10 min. shaking at 1400 rpm at 60 °C) and alkylated by the addition of 12 µL 500 mM iodoacetamide (Sigma, in 25 mM AB; 30 min. shaking in the dark at room temperature). Immobilised trypsin beads were removed by centrifugation at 10000 rpm for 10 min. Supernatant containing reduced, alkylated peptides were transferred to 1.5 mL 'LoBind' Eppendorf tubes and acidified by addition of 5 µL 10% trifluoroacetic acid (TFA, Riedel-de Haën) in water, and cleaned by two-phase extraction (3 x addition of 200 µL ethyl acetate, Sigma, followed by vortexing and aspiration of the organic layer). Peptides were desalted using POROS R3 beads (Thermo Fisher) in accordance with the manufacturer's protocol and lyophilized. Peptide concentrations (measured by Direct Detect spectrophotometer, Millipore) in injection solution (5% HPLC grade acetonitrile, Fisher Scientific, 0.1% TFA in deionized water) were adjusted to 200 ng/µL prior to MS analysis.

Preparation of serial mixtures of human and mouse peptides. Primary human MSCs and mouse skin sections were prepared separately as described above. Tryptic peptides from each preparation were mixed in ratios of 3:1, 1:1 and 1:3 (human:mouse, one series for each human donor:mouse), as described previously in (Swift et al, 2013a).

Stable isotope labelling of peptides. Desalted peptides were resuspended in 100 mM AB dissolved in $H_2{}^{18}O$ with or without 5 U PNGase F (Sigma), to a final volume of 10 µL. Peptides were then incubated in a thermocycler (1400 rpm, 37 °C) for 5 hours before drying down to a minimal volume in a vacuum centrifuge and resuspension in injection solution prior to MS analysis.

Mass spectrometry (MS). Digested samples were analysed by liquid chromatography (LC) coupled tandem MS (LC-MS/MS) using an UltiMate® 3000 Rapid Separation LC system (RSLC, Dionex Corporation) coupled to a Q Exactive HF (Thermo Fisher), peptide mixtures were separated using a multistep gradient from 95% A (0.1% formic acid, FA, Thermo Fisher, in water) and 5% B (0.1% FA in acetonitrile) to 7% B at 1 min, 18% B at 58 min, 27% B at 72 min and 60% B at 74 min at 300 nL/min, using a 75 mm x 250 µm i.d. 1.7 µm CSH C18, analytical column (Waters). Peptides were selected for fragmentation automatically by data dependent analysis.

Preliminary data analysis using Progenesis QI (Nonlinear Dynamics). Spectra from multiple samples were automatically aligned using Progenesis QI with manual placement of vectors where necessary. Peak picking sensitivity was set to 4/5 and features with a charge greater than +4 or fewer than 3 isotopes were excluded from further analysis. In the mouse/human mixed-species experiment, remaining features were searched using Mascot (Matrix Science), against the SwissProt and TREMBL pan-mammalian database (non-human and non-mouse identifications were removed). The peptide database was modified to search for alkylated cysteine residues (monoisotopic mass change, +57.021 Da) as a fixed modification, with oxidized methionine (+15.995 Da), hydroxylation of asparagine, aspartic acid, proline or lysine (+15.995 Da) and phosphorylation of serine, tyrosine, threonine (+79.966 Da) as variable modifications. For the PNGase F-treated experiment, the peptide database was modified to search for alkylated cysteine residues (+57.021 Da) as a fixed modification, with oxidized methionine (+15.995 Da), ^{18}O-deamidation of asparagine/glutamine (+2.988 Da) and ^{16}O-deamidation of asparagine/glutamine (+0.984 Da). A maximum of one missed cleavage was allowed. Peptide detection intensities were exported from Progenesis QI as Excel (Microsoft) spread sheets for further processing. Deamidated spectrum counts were performed in Scaffold (version 4, Proteome Software) using .DAT files from Mascot, following pre-processing in Progenesis QI as detailed above. We applied quantification calculations to proteins detected with >2 unique (by

sequence) peptides. Peptides assigned to proteins with 'unreviewed' status in the UniProt database were reassigned to the most abundant 'reviewed' protein with sequence identity in the dataset. Peptides shared between different protein groups were excluded from subsequent analysis.

Multi-variate regression modelling. To compare different types of regression analysis we used three different types of linear regression algorithms: (**A**) ordinary least squares (OLS); (**B**) Linear mixed effects models with Huber residual weights (LME-H); and (**C**) BayesENproteomics. Further details and references for each method are provided in the following sections. All analysis was performed using code written for Matlab R2015a (The MathWorks Inc).

Imputation of missing values

For OLS and LME-H methods, imputation of missing values was performed using a multiple imputation strategy as described by (Karpievitch et al, 2012), downshifted to accommodate previous observations showing that missing values are more likely to be missing due to low abundance (Goeminne et al, 2016). Briefly, for each peptide, non-missing peptide intensities were averaged to model the centre of a normal distribution with a width equal to 0.3 times that peptide's σ (standard deviation). The mean of centre of this normal distribution was negatively shifted by 1.6σ to model low abundance missingness, as described by Tyanova et al (2016) (albeit for protein-level imputation). Missing values were then randomly imputed from this downshifted normal distribution.

Previous observations, such as by Tyanova et al (2016), have shown that imputation from a down-shifted Gaussian distribution may be more suitable for imputation of missing peptide intensities from low abundance proteins, which are likely to be "missing non-randomly" (MNR), than for missing values from high abundance proteins, which are more likely to be "missing at random"

(MAR) (Goeminne et al, 2015). However, others have suggested that the type of "missingness" (MAR or MNR) that predominates within a given dataset should decide what imputation strategy is used. For BayesENproteomics, we employed a logistic regression-based model similar to that employed by Li et al (2011) to discern whether a given missing value was MAR or MNR. Briefly, a model of "missingness" was first created to differentiate observations that were MAR from those that were MNR as in (3).

$$\mathrm{Z}_{qi} = \log\left(\frac{R_{qi}}{1-R_{qi}}\right) = \theta_0 + X_f\theta_f + X_g\theta_g + \varepsilon_{qfg} \quad (3)$$

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. We assume that most missing values in any proteomics experiment are MNR, an assumption that others have shown is valid (Goeminne et al, 2016; Karpievitch et al, 2012). Therefore, if $\theta_j > \theta_0$ and $\theta_j > 0$ 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 (4) and (5).

$$\frac{\tilde{\boldsymbol{y}}_{(MAR)iq}}{w_{(MAR)i-1}} \sim MVN\left(\boldsymbol{X}_{(MAR)}\beta_{i-1}^{T},\ \sigma_{i-1}^{2}\left(\boldsymbol{X}_{(MAR)}\boldsymbol{X}_{(MAR)}^{T}\right)^{-1}\right) \quad (4)$$

$$\frac{\tilde{\boldsymbol{y}}_{(MNR)iq}}{w_{(MNR)i-1}} \sim TGauss\begin{pmatrix}\boldsymbol{X}_{(MNR)}\beta_{i-1}^{T},\ \sigma_{i-1}^{2}\left(\boldsymbol{X}_{(MNR)}\boldsymbol{X}_{(MNR)}{}^{T}\right)^{-1}, \\ a = min\{\boldsymbol{y}_{observed}\} - 2, b = \boldsymbol{y}_{\left\lceil\frac{MNR}{n}\right\rceil}\end{pmatrix} \quad (5)$$

Where $\tilde{\boldsymbol{y}}_{(MAR)iq}$ represents a vector of score-weighted (see below) MAR $\log_2$(intensity) values for protein *q* at the ith 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 $i\text{-}1^{th}$ iteration of the Gibbs sampler, respectively. Missing values in $\tilde{\boldsymbol{y}}_{iq}$ are sampled according to the multivariate normal distribution (MVN) described in (4) (see also section 3.1 in Zeng et al, 2017). 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 ($\boldsymbol{y}_{\left\lceil\frac{MNR}{n}\right\rceil}$) from the missingness regression model in (3). In (5), 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 (3), 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 (Schafer & Olsen, 1998) rather than fixed point estimates as in OLS and LME-H, where single random samples could strongly influence individual protein fold change estimates.

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 peptide false discovery rate (FDR). The number of false-positive identifications was reduced by discarding peptides with low Mascot scores using a standard FDR cut-offs based on identification p-values. We employ a Benjamini-Hochberg FDR (Benjamini & Hochberg, 1995) cut-off of < 0.2.

BayesENproteomics also employed a novel heuristic outlier weighting scheme that 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. 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.

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

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 in Ting et al (2007). 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).

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

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

Where $\boldsymbol{R}$ is a vector of residuals with elements $R_1 \dots R_n$, $R = \boldsymbol{y} - \boldsymbol{X}\beta$. 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).

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

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

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

Linear regression 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 (Goeminne et al, 2015). Here we compare three different models for calculating differentially abundant proteins and PTMs:

Ordinary least squares (OLS): differential protein abundance was calculated using the simple linear model (Choi et al, 2014; Clough et al, 2009) 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.

Linear mixed-effects models with Huber residual weights (LME-H): the ridge regression/mixed-effects algorithm developed by Goeminne et al (2016), wherein peptides were modelled 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 more complex model shown in equation (2). Goeminne et al (2016) exploited 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 $i, j = 1, 2 \ldots p$). We recapitulated this algorithm using the fitglme Matlab function, including Huber weighting of residuals. 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.

BayesENproteomics: this method utilises a novel Bayesian linear regression algorithm with elastic net regularization based on the hierarchical model detailed by Kyung et al (2010). Bayesian methods employ regularization based on the prior distribution parameters were estimated from, with elastic net regularization being equivalent to sampling from an intermediate Gaussian/Laplacian prior. Sampling was performed using a Gibbs sampler with the maximum number of iterations for each protein set at $25 \times runs + p_f$ (*runs* = total number of MS runs to be analysed, p_f = number of peptides), or 1,000 – whichever was higher – with half of these as burn-in iterations. The full hierarchical model for a single protein is detailed in equations (11)-(15).

$$\beta|\sigma^2, \tau^2, \lambda_2, \widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{y}} \sim MVN\begin{pmatrix} \left(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} + \boldsymbol{I}_p(\tau^{-2} + \lambda_2)\right)^{-1}\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{y}}, \\ \sigma^2\left(\widetilde{\boldsymbol{X}}^T\widetilde{\boldsymbol{X}} + \boldsymbol{I}_p(\tau^{-2} + \lambda_2)\right)^{-1} \end{pmatrix} \quad (11)$$

$$\sigma^2|\beta,\tau^2 \sim IG\left(\frac{(n-1+p)}{2},\ \frac{1}{100}+\frac{R^TR}{2}+\frac{(\lambda_1\beta)D_{\tau^2}^{-1}\beta}{2}+\frac{\lambda_2\beta^T\beta}{2}\right) \quad (12)$$

$$\tau_j^{-2}|\beta_j,\sigma^2,\lambda_1 \sim GIG\left(-\frac{1}{2},\ \sqrt{\frac{\lambda_1^2\sigma^2}{\beta_j^2}},\ \lambda_1^2\right) \quad (13)$$

Let τ^{-2} represent a vector of latent variables ($\tau_1^{-2} \ldots \tau_p^{-2}$) – sampled from the generalised inverse Gaussian (GIG) distribution in (13) using the sampler in Makalic and Schmidt (2016) – 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 (11). To enforce sparsity, we employ two regularization hyperparameters, λ_1 and λ_2, with different conditional distributions, specifying LASSO (14) and ridge hyperparameters (15), 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 regularization hyperparameter, $\lambda_{2\,j}$, similar to the LASSO-like "horseshoe" estimator (Makalic & Schmidt, 2016). 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 regularization hyperparameters that lead to constant shrinkage across all βs. λ_1 and λ_2 are sampled from gamma distributions of the form $Gamma(a,b)$, with a posterior mean of $\frac{a}{b}$.

$$\lambda_1^2|\,\tau^2 \sim Gamma\left(p,\ 1+\frac{1}{2}\sum_{j=1}^{p}\tau_j^2\right) \quad (14)$$

$$\lambda_{2\,j}|\beta_j,\sigma^2 \sim Gamma\left(1,\ 3+\frac{\beta_j^2}{2\sigma^2}\right) \quad (15)$$

Estimates of parameters were taken as means of Gibbs-sampled posterior distributions made from post-burn-in iterations.

Pathway analysis using linear modelling (PALM). PALM using Reactome (Fabregat et al, 2018; Milacic et al, 2012) pathway annotations was performed as described above using logged fold changes (calculated from the three methods detailed above) as the response variable ($\boldsymbol{y}$) according to the model for a given Reactome pathway:

$$y_{gp} = \beta_0 + X_g\beta_g + X_p\beta_p + \varepsilon_{gp} \quad (16)$$

Where β_g and β_p denote effect sizes due to experimental treatment *g* and protein *p*. Residual weights (w) were set to $min\left(1, \frac{|\boldsymbol{y}_i|}{SE_i}\right)$, where SE_i ($i = 1 \ldots n$) is the standard error of that protein fold change estimate, $\boldsymbol{y}_i$. The code allows the user to limit estimation to only those pathways with a minimum observed number of proteins present in a given dataset; here, we set the minimum to be five.

Empirical Bayes correction of protein variances. All methods included Empirical Bayes correction for stabilising variance estimation (Kammers et al, 2015; Smyth, 2004). Empirical Bayes-modified t-tests were used to calculate p-values for determining significance of differentially abundant proteins and PTMs.

Data availability

Associated Mascot files and Progenesis QI peptide ion output used in this study, along with BayesENproteomics Matlab code used here (with protein/pathway-level output files as .csv) are freely available from www.github.com/VenkMallikarjun/BayesENproteomics.

Acknowledgements

VM was supported by a studentship from the Sir Richard Stapley Educational Trust. JS was funded by a Biotechnology and Biological Sciences Research Council (BBSRC) David Phillips Fellowship (BB/L024551/1). Mass spectrometry was carried out at the Wellcome Trust Centre for Cell-Matrix Research (WTCCMR; 203128/Z/16/Z) Proteomics Core Facility. The authors thank: Prof. Tim Board (Wrightington Hospital) for provision of tissue samples for MSC isolation; Dr. Charis Saville (University of Manchester) for provision of mouse tissue.

Author contributions

VM designed and implemented the algorithm and developed the benchmark datasets used in this manuscript. VM, SMR and JS wrote the paper.

Conflicts of interest

The authors declare no competing interests.

References

Abaye DA, Pullen FS, Nielsen BV (2011) Peptide polarity and the position of arginine as sources of selectivity during positive electrospray ionisation mass spectrometry. *Rapid Commun Mass Spectrom* **25:** 3597-3608

Bekker-Jensen DB, Kelstrup CD, Batth TS, Larsen SC, Haldrup C, Bramsen JB, Sorensen KD, Hoyer S, Orntoft TF, Andersen CL, Nielsen ML, Olsen JV (2017) An optimized shotgun strategy for the rapid generation of comprehensive human proteomes. *Cell Syst* **4:** 587-+

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

Blein-Nicolas M, Xu H, de Vienne D, Giraud C, Huet S, Zivy M (2012) Including shared peptides for estimating protein abundances: A significant improvement for quantitative proteomics. *Proteomics* **12:** 2797-2801

Bornn L, Gottardo R, Doucet A (2010) Grouping priors and the Bayesian elastic net. *ArXiv***:** 1001.4083

Brodin P, Jojic V, Gao TX, Bhattacharya S, Angel CJL, Furman D, Shen-Orr S, Dekker CL, Swan GE, Butte AJ, Maecker HT, Davis MM (2015) Variation in the human immune system is largely driven by non-heritable influences. *Cell* **160:** 37-47

Choi M, Chang CY, Clough T, Broudy D, Killeen T, MacLean B, Vitek O (2014) MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* **30:** 2524-2526

Clough T, Key M, Ott I, Ragg S, Schadow G, Vitek O (2009) Protein quantification in label-free LC-MS experiments. *J Proteome Res* **8:** 5275-5284

Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B, Milacic M, Roca CD, Rothfels K, Sevilla C, Shamovsky V, Shorser S, Varusai T, Viteri G, Weiser J, Wu GM et al (2018) The Reactome pathway knowledgebase. *Nucl Acids Res* **46:** D649-D655

Forshed J (2013) Protein quantification by peptide quality control (PQPQ) of shotgun proteomics data. In *The Low Molecular Weight Proteome. Methods in Molecular Biology (Methods and Protocols)*, Bäckvall H, Lehtiö J (eds), Vol. 1023, pp 149-158. Springer, New York, NY

Goeminne LJE, Argentini A, Martens L, Clement L (2015) Summarization vs peptide-based models in label-free quantitative proteomics: performance, pitfalls, and data analysis guidelines. *J Proteome Res* **14:** 2457-2465

Goeminne LJE, Gevaert K, Clement L (2016) Peptide-level robust ridge regression improves estimation, sensitivity, and specificity in data-dependent quantitative label-free shotgun proteomics. *Mol Cell Proteomics* **15:** 657-668

Henao R, Thompson JW, Moseley MA, Ginsburg GS, Carin L, Lucas JE (2013) Latent protein trees. *Ann Appl Stat* **7:** 691-713

Ivanovska IL, Swift J, Spinler K, Dingal D, Cho S, Discher DE (2017) Cross-linked matrix rigidity and soluble retinoids synergize in nuclear lamina regulation of stem cell differentiation. *Mol Biol Cell* **28:** 2010-2022

Kammers K, Cole RN, Tiengwe C, Ruczinski I (2015) Detecting significant changes in protein abundance. *EuPA Open Proteom* **7:** 11-19

Karpievitch YV, Dabney AR, Smith RD (2012) Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics* **13:** 9

Kyung M, Gilly J, Ghoshz M, Casellax G (2010) Penalized regression, standard errors, and Bayesian lassos. *Bayesian Anal* **5:** 369-412

Li F, Nie L, Wu G, Qiao JJ, Zhang WW (2011) Prediction and characterization of missing proteomic data in Desulfovibrio vulgaris. *Compar Funct Genom* **2011:** 780973

Makalic E, Schmidt DF (2016) A simple sampler for the horseshoe estimator. *IEEE Signal Process Lett* **23:** 179-182

Malioutov D, Chen T, Jaffe J, Airoldi E, Carr S, Budnik B, Slavov N (2017) Quantifying homologous proteins and proteoforms. *BioRxiv*: 10.1101/168765

Milacic M, Haw R, Rothfels K, Wu G, Croft D, Hermjakob H, D'Eustachio P, Stein L (2012) Annotating cancer variants and anti-cancer therapeutics in reactome. *Cancers (Basel)* **4:** 1180-1211

Ozerov IV, Lezhnina KV, Izumchenko E, Artemov AV, Medintsev S, Vanhaelen Q, Aliper A, Vijg J, Osipov AN, Labat I, West MD, Buzdin A, Cantor CR, Nikolsky Y, Borisov N, Irincheeva I, Khokhlovich E, Sidransky D, Camargo ML, Zhavoronkov A (2016) In silico Pathway Activation Network Decomposition Analysis (iPANDA) as a method for biomarker development. *Nat Commun* **7:** 11

Rodriguez J, Gupta N, Smith RD, Pevzner PA (2008) Does trypsin cut before proline? *J Proteome Res* **7:** 300-305

Schafer JL, Olsen MK (1998) Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behav Res* **33:** 545-571

Schliekelman P, Liu SB (2014) Quantifying the effect of competition for detection between coeluting peptides on detection probabilities in mass-spectrometry-based proteomics. *J Proteome Res* **13:** 348-361

Smyth GK (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3:** 1-25

Strassburg S, Richardson SM, Freemont AJ, Hoyland JA (2010) Co-culture induces mesenchymal stem cell differentiation and modulation of the degenerate human nucleus pulposus cell phenotype. *Regen Med* **5:** 701-711

Swift J, Harada T, Buxboim A, Shin J-W, Tang H-Y, Speicher DW, Discher DE (2013a) Label-free mass spectrometry exploits dozens of detected peptides to quantify lamins in wildtype and knockdown cells. *Nucleus* **4:** 450-459

Swift J, Ivanovska IL, Buxboim A, Harada T, Dingal PCDP, Pinter J, Pajerowski JD, Spinler KR, Shin J-W, Tewari M, Rehfeldt F, Speicher DW, Discher DE (2013b) Nuclear lamin-A scales with tissue stiffness and enhances matrix-directed differentiation. *Science* **341:** 1240104

Ting JA, D'Souza A, Schaal S (2007) Automatic outlier detection: A Bayesian approach. *Proc - IEEE Int Conf Robot Autom***:** 2489–2494

Tsai TH, Hao ZQ, Hong QT, Moore B, Stella C, Zhang JH, Chen Y, Kim M, Koulis T, Ryslik GA, Verschueren E, Jacobson F, Haskins WE, Vitek O (2017) Statistical characterization of therapeutic protein modifications. *Sci Rep* **7:** 13

Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, Mann M, Cox J (2016) The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods* **13:** 731-740

Webb-Robertson BJM, Matzke MM, Datta S, Payne SH, Kang JY, Bramer LM, Nicora CD, Shukla AK, Metz TO, Rodland KD, Smith RD, Tardiff MF, McDermott JE, Pounds JG, Waters KM (2014) Bayesian proteoform modeling improves protein quantification of global proteomic measurements. *Mol Cell Proteomics* **13:** 3639-3646

Wu RH, Dephoure N, Haas W, Huttlin EL, Zhai B, Sowa ME, Gygi SP (2011) Correct interpretation of comprehensive phosphorylation dynamics requires normalization by protein expression changes. *Mol Cell Proteomics* **10:** 12

Zeng IS, Lumley T, Ruggiero K, Middleditch M (2017) A Bayesian approach to multivariate and multilevel modelling with non-random missingness for hierarchical clinical proteomics data. *BioRxiv***:** 10.1101/153049

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J R Stat Soc Ser B-Stat Methodol* **67:** 301-320

Figures

Figure 1.

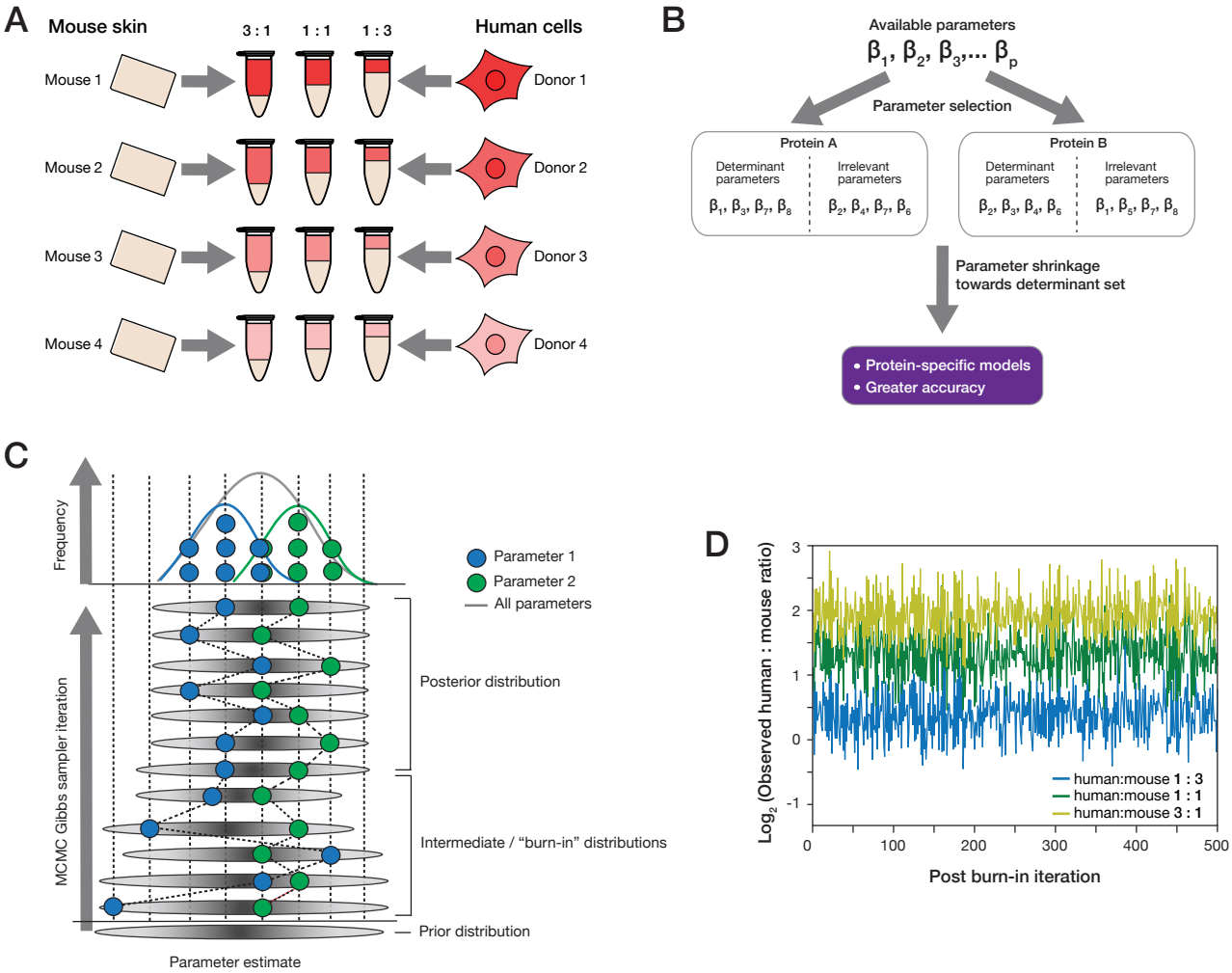

Figure 1. Extraction and preparation of mouse and human proteins for use in label-free quantification using multiple regression models. (**A**) Schematic diagram of the protocol used to prepare samples for the mixed species benchmark experiment. (**B**) Flow diagram showing the effect of regularization on estimated parameters used in fitting models. (**C**) Schematic representation of how the Gibbs sampler arrives at the final (posterior) distribution for a single parameter. As the sampler proceeds, it will ideally converge on a single posterior distribution, as progressive iterations are biased towards areas of higher likelihood. Iterations from before this point are typically discarded ("burn-in" iterations) as those samples will not come from the target posterior distribution. Each parameter is then summarized from its respective estimated distribution (green and blue distributions) and all parameters lie on the target posterior distribution (grey distribution). (**D**) Example traces from post-burn-in samples to build posterior β distributions corresponding to $\log_2$ (protein abundance) estimates.

Figure 2.

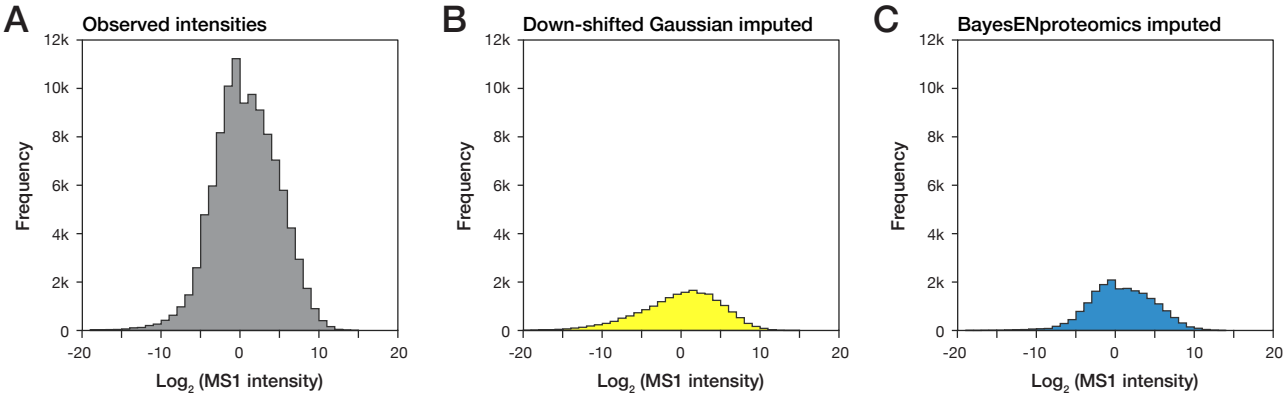

Figure 2. Comparison of missing value imputation strategies. Histograms allowing a comparison of distributions of (**A**) observed peptide intensities from both mouse and human proteins; (**B**) peptide intensities from presumed non-random missing value imputation (Tyanova et al, 2016); (**C**) peptide intensities from adaptive imputation with BayesENproteomics.

Figure 3.

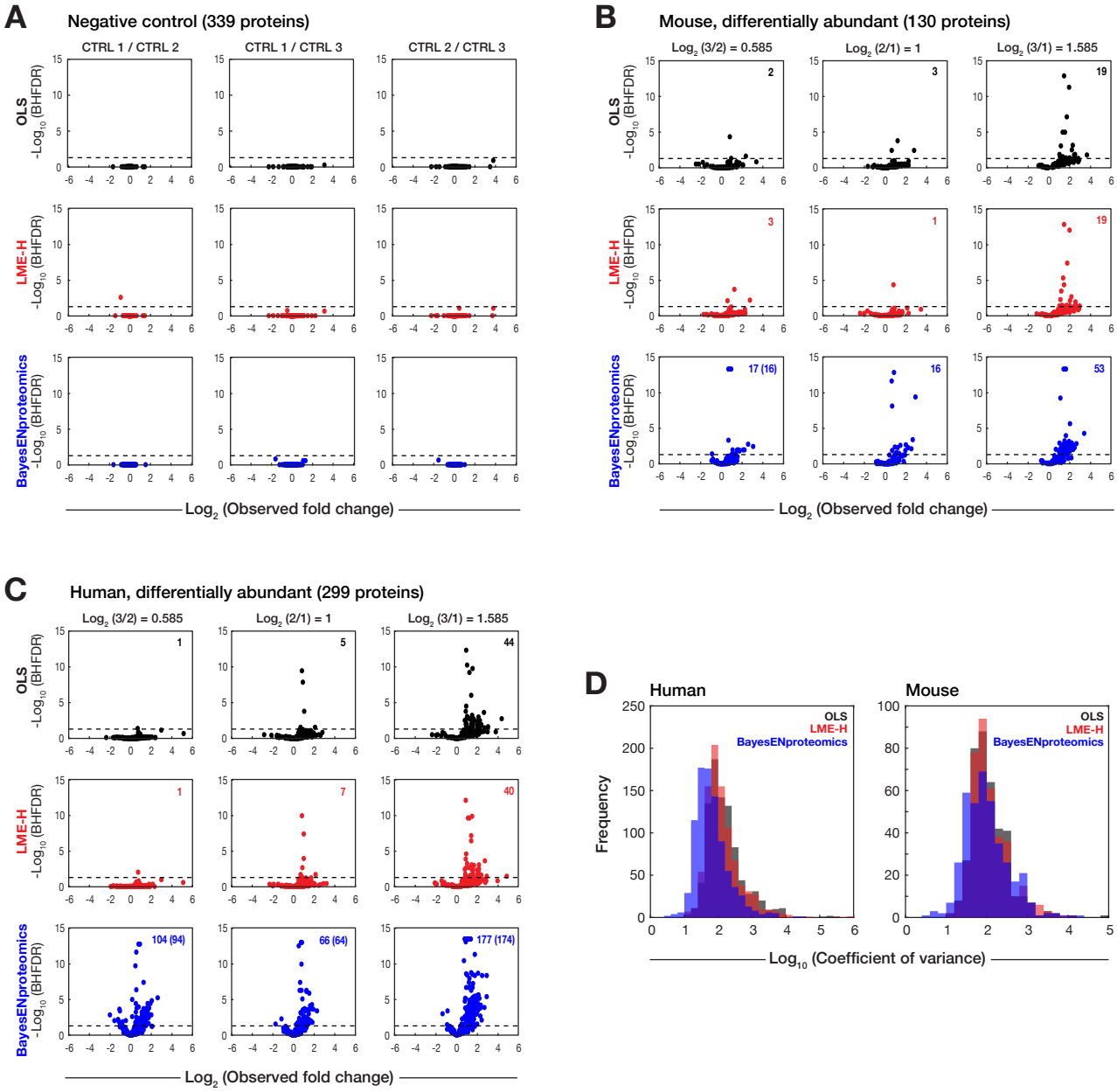
A
Negative control (339 proteins)
CTRL 1 / CTRL 2
CTRL 1 / CTRL 3
CTRL 2 / CTRL 3
OLS
LME-H
BayesENproteomics
-Log10 (BHFDR)
Log2 (Observed fold change)
B
Mouse, differentially abundant (130 proteins)
Log2 (3/2) = 0.585
Log2 (2/1) = 1
Log2 (3/1) = 1.585
2
3
19
3
1
19
17 (16)
16
53
C
Human, differentially abundant (299 proteins)
1
5
44
1
7
40
104 (94)
66 (64)
177 (174)
D
Human
Mouse
OLS
LME-H
BayesENproteomics
Frequency
Log10 (Coefficient of variance)

Figure 3. BayesENproteomics increases specificity compared to other multi-variate regression methods. (**A**) Volcano plots showing quantification of fold change estimates (x-axis) vs. significance (y-axis) observed in pairwise comparisons between three technical replicates. (**B**) Volcano plots showing quantification of mouse proteins across all comparisons in mixed human:mouse benchmark experiment. (**C**) Volcano plots showing quantification of human proteins across all comparisons in mixed human:mouse benchmark experiment. (**D**) Histograms showing coefficients of variation (CVs) for all differentially abundant proteins in the mixed human:mouse benchmark experiment. Dotted lines in (A)-(C) indicate Benjamini-Hochberg false discovery rate corrected p-value (BHFDR) = 0.05. OLS fit using simple model in equation (1), black; LME-H according to equation (2), red; BayesENproteomics as in (2), blue. n = 4 human donors, 4 mice. Donor effects in (1) and (2) were excluded in analysis of mouse proteins. Numbers of significantly differentially abundant proteins are indicated in the top right of each plot, with number in brackets denoting those following the direction of the expected fold change.

Figure 4.

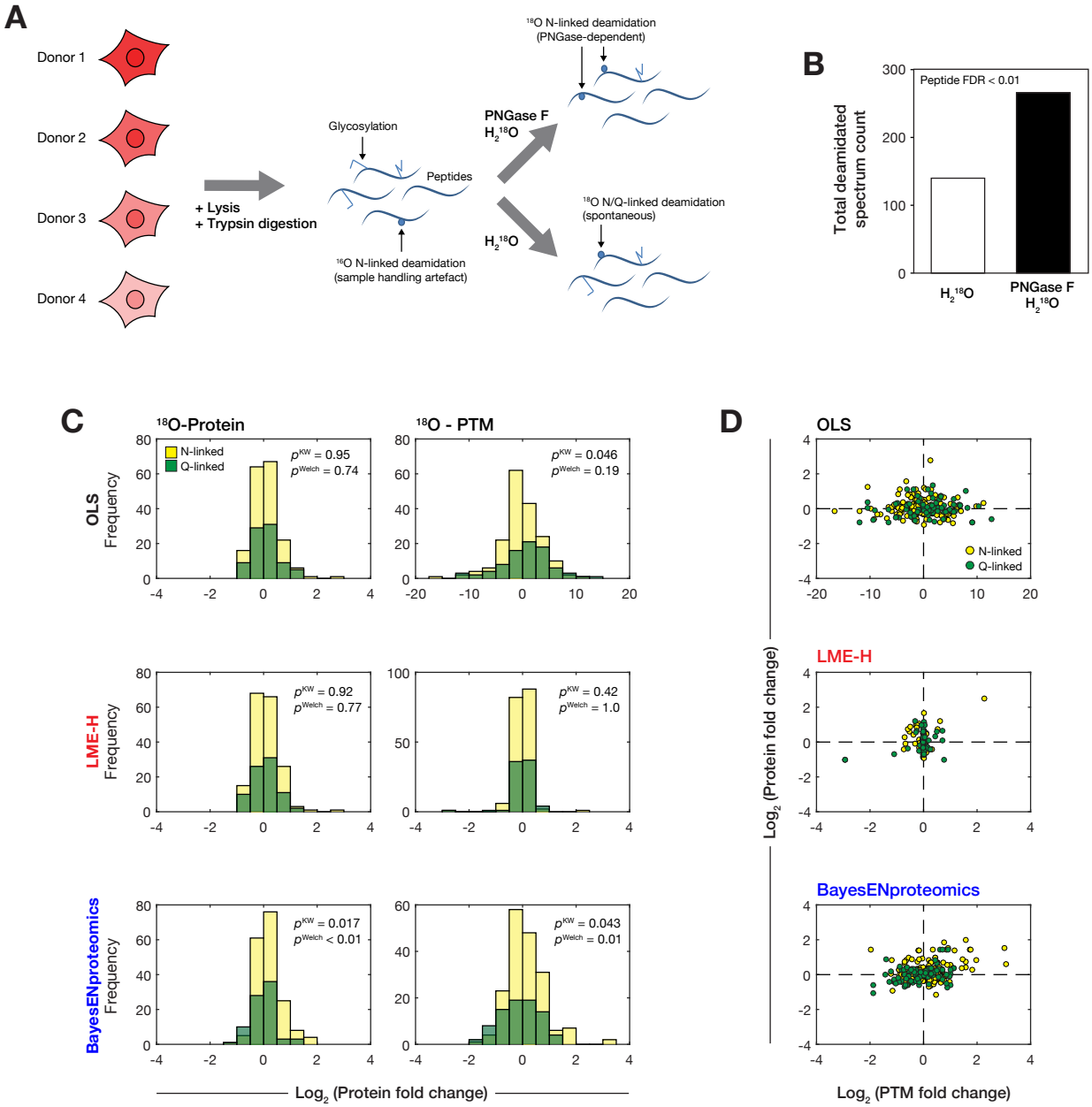
A
Donor 1
Donor 2
Donor 3
Donor 4
+ Lysis
+ Trypsin digestion
Glycosylation
Peptides
16O N-linked deamidation
(sample handling artefact)
PNGase F
H2 18O
18O N-linked deamidation
(PNGase-dependent)
H2 18O
18O N/Q-linked deamidation
(spontaneous)
B
Peptide FDR < 0.01
Total deamidated spectrum count
H2 18O
PNGase F
H2 18O
C
18O-Protein
18O - PTM
N-linked
Q-linked
OLS
LME-H
BayesENproteomics
Frequency
pKW = 0.95
pWelch = 0.74
pKW = 0.046
pWelch = 0.19
pKW = 0.92
pWelch = 0.77
pKW = 0.42
pWelch = 1.0
pKW = 0.017
pWelch < 0.01
pKW = 0.043
pWelch = 0.01
Log2 (Protein fold change)
D
OLS
LME-H
BayesENproteomics
N-linked
Q-linked
Log2 (Protein fold change)
Log2 (PTM fold change)

Figure 4. PNGase-catalysed stable isotope labelling quantified by summarization with normalisation and multivariate regression methods. (**A**) Schematic diagram illustrating experimental setup for the variable PTM benchmark dataset. (**B**) The number of deamidated peptides was increased following 5 hours PNGase F treatment. Only peptides from proteins identified by >1 peptide were counted. (**C**) Histograms showing fold changes in N-linked ^{18}O labelled N-linked deamidation (yellow) and ^{18}O labelled Q-linked deamidation (green). (**D**) Scatter plots comparing protein and PTM log_2 fold changes for the indicated quantification methods. n = 5 human donors. Indicated p-values show significance from comparison of N-linked (yellow) and Q-linked (green) histograms calculated by Welch's two-sample t-test (p^{Welch}) or Kruskal-Wallis test (p^{KW}).

Figure 5.

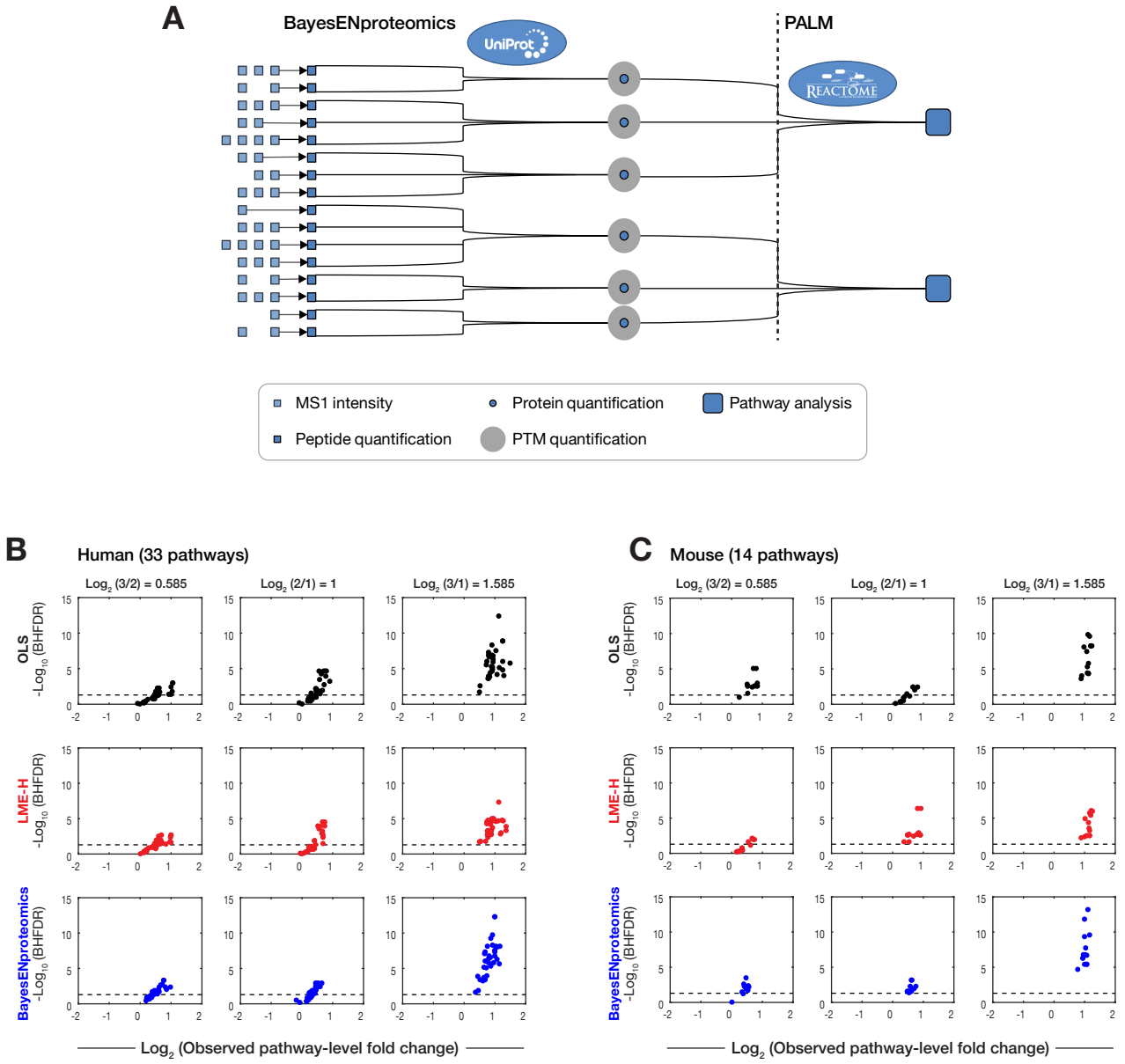
A
BayesENproteomics
UniProt
PALM
REACTOME
MS1 intensity
Peptide quantification
Protein quantification
PTM quantification
Pathway analysis
B
Human (33 pathways)
C
Mouse (14 pathways)
Log2 (3/2) = 0.585
Log2 (2/1) = 1
Log2 (3/1) = 1.585
OLS
LME-H
BayesENproteomics
-Log10 (BHFDR)
Log2 (Observed pathway-level fold change)

Figure 5. Pathway analysis using linear models (PALM) on mixed-species dataset. (**A**) Schematic diagram showing hierarchical tree structure for linear model assembly from MS1 intensities (far left) to pathways (far right). (**B**) and (**C**) Volcano plots showing significantly differentially regulated pathways calculated from mouse:human mixed species datasets from OLS (black), LME-H (red) and BayesENproteomics (blue) protein-level output with subsequent PALM pathway analysis. Each dot represent a single Reactome pathway. Pathways follow the expected direction of the component proteins, though confidence in original protein-level fold change estimates may affect exact magnitudes of pathway-level fold changes. Only pathways with ≥5 proteins present in the dataset are shown.