MSqRob takes the missing hurdle: uniting intensity- and count-based proteomics

Ludger J.E. Goeminne[1,2,3,4†], Adriaan Sticker[1,2,3,4], Lennart Martens[2,3,4], Kris Gevaert[2,3] and Lieven Clement[1,4*]

E-mail: ludger.goeminne@UGent.vib.be; adriaan.sticker@UGent.vib.be; lennart.martens@UGent.vib.be; kris.gevaert@UGent.vib.be; lieven.clement@UGent.be

[1] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Krijgslaan 281 – S9, B9000 Ghent, Belgium

[2] VIB-UGent Center for Medical Biotechnology, VIB, Albert Baertsoenkaai 3, B9000 Ghent, Belgium

[3] Department of Biomolecular Medicine, Ghent University, Albert Baertsoenkaai 3, B9000 Ghent, Belgium

[4] Bioinformatics Institute Ghent, Ghent University, Belgium

ABSTRACT: Missing values are a major issue in quantitative data-dependent mass spectrometry-based proteomics. We therefore present an innovative solution to this key issue by introducing a hurdle model, which is a mixture between a binomial peptide count and a peptide intensity-based model component. It enables dramatically enhanced quantification of proteins with many missing values without having to resort to harmful assumptions for missingness. We demonstrate the superior performance of our method by comparing it with state-of-the-art methods in the field.

Label-free data-dependent quantitative mass spectrometry (MS) is the preferred method for deep and high-throughput identification and quantification of thousands of proteins in a single analysis[1]. However, this approach suffers from many missing values, which strongly reduce the amount of quantifiable proteins[2, 3]. There are three common causes for this missingness: (1) true absence of signal, or signal below detection limit in the MS1 spectrum; (2) lack of fragmentation and hence missed identification of the MS1 peak; and (3) failed identification of the acquired fragmentation spectrum. As a result, missingness is more likely to occur for low-abundant proteins and/or poorly ionizing peptides. However, missingness may also extend to mid- and even high-range intensities, e.g. when co-eluting peptides suppress an MS1 signal or when poor quality of an MS2 spectrum interferes with correct identification. Missingness due to lack of fragmentation can be mitigated by "matching between runs", where unidentified MS1 peaks in one run are aligned to identified peaks in another run in narrow retention time and mass-over-charge (*m/z*) windows[4]. Nevertheless, missing values remain widespread; a survey of 73 recent public proteomics data sets from the PRIDE database demonstrates an average of 44% missing values (Fig. 1 a, Supplementary Table 1).
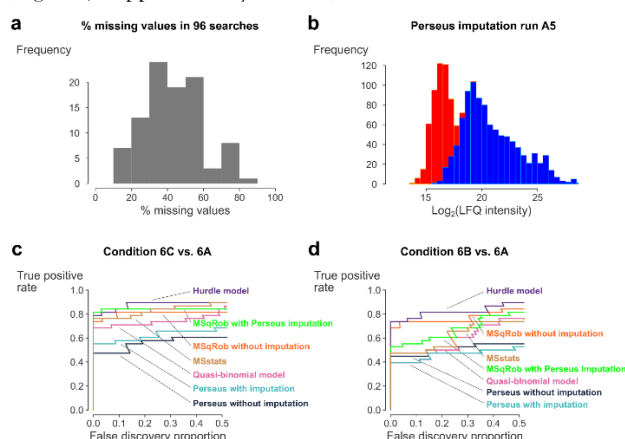

Figure 1. The impact of missing values and the superior performance of the hurdle model. (a) Missing values are highly prevalent in recent proteomics data. The histogram shows the distribution of the percentage of missing values in 96 peptides.txt files from 73 PRIDE projects with a PRIDE publication date in 2017 that applied shotgun proteomics to full or partial proteomes. Missingness ranges from 16 to 82%, with an average of 44% missing values. (b) Perseus imputation fails when too many missing values are present: the frequency distribution of the log₂-transformed LFQ intensities in the CPTAC dataset becomes bimodal after Perseus imputation due to the many imputed values, as exemplified by run A5. Observed values (978 LFQ intensities) are colored in blue, imputed values (484 LFQ intensities) are colored in red. Similar results for the other runs can be seen in Supplementary Fig. 1. (c, d) In the CPTAC dataset (38 UPS1 and 1,343 yeast proteins), the hurdle model outperforms

other methods. The fraction of true positive UPS1 proteins flagged as DA in conditions C vs. A and B vs. A is plotted as a function of the fraction of false positive yeast proteins in the total number of DA proteins.

A common solution for missingness is to impute the missing values. But, this comes at the expense of additional assumptions, e.g. k-nearest neighbors (kNN) assumes intensity-independent missingness, and the popular software packages Perseus[5] ("Perseus imputation", PI) and MSstats[6] assume missingness by low intensity. In reality, however, missing values originate from a mix of intensity-dependent and -independent mechanisms, which can moreover be strongly data set-specific[3, 7]. As a result, state-of-the-art imputation profoundly changes the distribution of protein-level intensities as it typically produces a second mode due to many imputed values, (shown for MaxQuant/Perseus in Fig. 1 b and Supplementary Fig. 1 – 2; and for kNN in Supplementary Fig. 3 – 4). This can have a deep impact on the downstream differential analysis. We demonstrate these effects on the widely studied CPTAC dataset[8], where more than 47% of peptide data are missing. Because the spike-in concentrations in this dataset cover a wide range, the perils of imputation can be clearly shown. Indeed, while our MSqRob[9] quantification tool shows excellent performance when combined with PI for comparison C (high spike-in concentration: 2.2 fmol/µL) vs. A (lowest spike in concentration: 0.25 fmol/µL) (Fig. 1 c), MSqRob with PI performs poorly for comparisons B (intermediate amount of spike-in: 0.74 fmol/µL) vs. A (Fig. 1 d) and C vs. B, where the differences in spike-in concentration are smaller (Supplementary Fig. 5, 6), thus rapidly accumulating false positives. MSqRob therefore omits imputation by default to avoid a severe backlash in performance[10]. However, without imputation, intensity-based methods cannot cope with complete missingness in one condition as it is impossible to calculate a fold change (FC) relative to a missing value. Thus, potentially interesting cases such as strong protein synthesis/stabilization or protein degradation remain undetected.

Conversely, peptide counting approaches naturally handle missing peptides in a run by a zero count. Yet relative quantification by peptide counting generally performs poorly (pink line in Fig. 1 c, d). This because counting disregards the inherent abundance-intensity relationship (within a given dynamic range) for each peptide[11, 12]. However, differentially abundant (DA) proteins for which no FC can be estimated due to missingness do differ in peptide counts (Supplementary Fig. 7) and proteins with a higher concentration have on average both higher peptide ion intensities and higher peptide counts, as it is more likely that even their poorly ionizing peptides are detected.

Combining intensity-based methods with counting-based methods to exploit this complementary information therefore seems promising. Webb-Robertson *et al.* combine intensity- and count-based statistics to filter out

peptides prior to differential analysis[13]. ProPCA uses principal components to summarize intensities and spectral counts into one value[14] and IDPQuantify uses Fisher's method to combine the p-values from a two-sample t-test with those of a quasi-Poisson regression on spectral counts[15]. However, ProPCA's combined metric lacks an intuitive interpretation, and does not include any downstream statistical analysis. Like Perseus, IDPQuantify only handles simple pairwise comparisons without any possibility to correct for confounding effects. Moreover, none of these methods is able to pinpoint whether the statistical significance is driven by the count component, the intensity component, or both, thus making it impossible to interpret the results in terms of FCs.

Here, we introduce a hurdle model that unites the advantages of MSqRob with the complementary information present in peptide counts, that avoids unrealistic imputation assumptions, and that provides interpretable results. This hurdle model takes peptide-level information as input and consists of a mixture model of two components: (1) a binary component that distinguishes between log_2-transformed peptide intensities that are either missing or observed; and (2) an MSqRob-based component to model the magnitude of log_2 peptide intensities passing the detection hurdle. Inference on the parameters of both model components has an intuitive interpretation: the binary component can be used to assess *differential detection (DD)* and returns *log odds ratios* (log ORs), while the MSqRob component allows to test for *differential abundance (DA)* of a protein and returns *log_2 fold changes* (log_2 FCs). When peptides are completely missing in one condition, the hurdle model reduces to a binomial model. However, when peptides are present in both conditions, the hurdle approach allows to combine information in the OR and FC test statistics, and can be used to infer on DD, DA, or both in a post-hoc analysis. Note, that the peptide counts can be overdispersed, which is why the variance component is estimated via quasi-likelihood and the model is termed a quasi-binomial model.

EXPERIMENTAL SECTION

In this section, we first demonstrate how we analyzed the frequency of missing values in recent PRIDE projects. Then, we discuss the nature and the origin of the CPTAC and the HEART datasets, followed by an overview of how these datasets were preprocessed and imputed with different strategies. We then introduce the Perseus, MSstats, MSqRob, the quasi-binomial model and the hurdle model as methods for statistical inference. Finally, we refer to our GitHub page which enables to reproduce all the analyses in this publication.

Missing values in recent PRIDE projects

For our analysis of the frequency of missing values of recent datasets in PRIDE, we downloaded 96

peptides.txt files corresponding to 73 PRIDE projects that adhere to the following conditions:

- label-free shotgun proteomics,

- full proteomes or enriched subsets of the proteome (i.e. no projects where there was an enrichment step for a chemical modification and no projects investigating protein-protein interactions),

- published in PRIDE in 2017 and

- searched with MaxQuant and peptides.txt file available from PRIDE.

An overview of the projects and peptides.txt files with their number of missing values, number of observed values and percentage of missing values is given in Supplementary Table 1.

Data availability

We made use of two example datasets.

1. In the benchmark CPTAC study 6, which can be downloaded from https://cptac-data-portal.georgetown.edu/cptac/dataPublic/list?currentPath=%2FPhase_I_Data%2FStudy6&nonav=true, the human UPS1 standard is spiked in 5 different concentrations in a yeast proteome background[8]. This allows to make comparisons for which the ground truth is known: when comparing two different spike-in conditions, only the UPS1 proteins are truly differentially abundant (DA), while the yeast proteins are not.

The three lowest spike-in conditions (A, B and C) from LTQ-Orbitrap at site 86, LTQ-Orbitrap O at site 65 and LTQ-Orbitrap W at site 56 were searched with MaxQuant 1.6.1.0 against a database containing 6,718 reviewed *Saccharomyces cerevisiae* (strain ATCC 204508/S288c) proteins downloaded from UniProt on September 14, 2017, supplemented with the 48 human UPS1 protein sequences provided by Sigma Aldrich. Carbamidomethylcysteine was set as a fixed modification and methionine oxidation, protein N-terminal acetylation and N-terminal glutamine to pyroglutamate conversion were set as variable modifications. Detailed search settings are described in Supplementary Material.

We only performed pairwise comparisons between the 3 lowest spike-in concentrations because of the huge ionization competition effects in the higher spike-in concentrations, as described earlier[9].

2. The HEART dataset originates from a large-scale proteomics study of the human heart, where 16 different regions and 3 different cell types from three healthy human adult hearts were studied[16].

For our analysis, we made use of the peptides.txt and proteinGroups.txt files made available by Doll *et al.* in ProteomeXchange via the PRIDE Archive repository under the identifier PXD006675. Here, we limited ourselves to the data from 6 regions: the atrial and

ventricular septa (SepA and SepV), the left and right atrium (LA and RA) and the left and right ventricle (LV and RV). We compare the atrial regions (LA, RA and SepA) to the ventricular regions (LV, RV and SepV), as well as LA vs. RA and LV vs. RV.

Preprocessing for MSqRob and the quasi-binomial model

Peptide intensities obtained from MaxQuant's peptides.txt file were $\log_2$-transformed and quantile normalized. Next, potential contaminants, reverse sequences and proteins that were only identified by peptides carrying a modification were removed from the data. Finally, proteins identified by only a single peptide were also removed.

Imputation methods

Missing values were either not imputed or imputed with kNN or Perseus imputation.

No imputation

For MSqRob without imputation, we additionally removed peptides that are only identified in a single run. Indeed, MSqRob directly models peptide intensities and the peptide-specific effect for peptides with one identification cannot be estimated because no replicates are available. This additional filtering step is not required for MSqRob with imputation, because missing values in the peptide intensity matrix are imputed first.

kNN imputation

k-nearest neighbors (kNN) imputation calculates a Euclidean distance metric on all peptide intensities to find the k most similar peptides and imputes the missing value with the average of the corresponding values from the k neighbors[17]. We used the default value of $k = 10$ neighbors.

Perseus imputation

We call "Perseus imputation (PI)" the standard imputation approach from the popular proteomics computational platform Perseus[5]. The imputation is achieved by using the "Replace missing values from normal distribution" function in Perseus 1.0.6.7. We applied PI on peptide intensities for MSqRob with PI and on LFQ intensities for Perseus with PI.

PI first constructs a rescaled normal distribution with: (1) a downshifted mean equal to the average of all the observed data minus d times the standard deviation of the observed data and (2) a standard deviation equal to w times the standard deviation of the observed data. Missing values are then imputed with random draws from this rescaled distribution. We adopt Perseus' default values for w (0.3) and d (1.8).

Statistical inference

In this subsection, we discuss differential protein analysis with Perseus and MSstats, followed by MSqRob, the quasi-binomial model and our hurdle model. All methods mentioned below model the data protein by protein. To improve readability, we will suppress the protein indicator in the remainder of this section.

Perseus

For the CPTAC dataset, LFQ intensities were imported into Perseus 1.6.0.7. Potential contaminants, reversed sequences and proteins only identified by peptides carrying modification sites were removed from the data. Next, empty columns were removed and via "Categorical annotation rows", runs were sorted according to their spike-in conditions (A, B or C). Data were either imputed with PI ("Perseus with imputation") or not ("Perseus without imputation"). Finally, we performed two-sample t-tests for each of the three comparisons.

For the atrial vs. ventricular comparison in the HEART dataset, Doll *et al.* provided the results of their Perseus with imputation approach in their Supplementary Data 5 file, available from https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-017-01747-2/MediaObjects/41467_2017_1747_MOESM7_ESM.xlsx.

MSqRob

MSqRob has been described in detail in Goeminne *et al.* (2016)[9]. Briefly, for each protein, the $\log_2$-transformed peptide intensities y_{pr} for peptide $p = 1, \ldots, P$ in run $r = 1, \ldots, R$ are modeled as follows:

$$y_{pr} = \beta^0 + x_{pr}^{\mathrm{T}}\beta + \beta_p^{\mathrm{peptide}} + u_r^{\mathrm{run}} + \varepsilon_{pr}, \qquad \text{(Eq. 1)}$$

where β^0 is the intercept, $\beta_p^{\mathrm{peptide}}$ is the fixed effect of the individual peptide sequence p, u_r^{run} is a random run effect ($u_r^{\mathrm{run}} \sim \mathrm{N}(0, \sigma_u^2)$) that accounts for the correlation of peptide intensities y_{jr} and y_{kr} from the same protein within the same run r, $x_{pr}^{\mathrm{T}} = \left(x_{1,pr}, \ldots, x_{m,pr}\right)^{\mathrm{T}}$ is a vector with the covariate pattern of the m remaining predictors, $\beta = [\beta_1, \ldots, \beta_m]^{\mathrm{T}}$ is a vector of parameters modeling the effect of each predictor on the peptide intensity conditionally on the remaining covariates, and ε_{pr} is the error term that is assumed to be normally distributed ($\varepsilon_{pr} \sim \mathrm{N}(0, \sigma^2)$).

For the CPTAC and HEART datasets, the MSqRob model can be written as follows:

$$y_{pr} = \beta^0 + \sum_{t=1}^{T} x_{t,pr}^{\mathrm{treat}}\beta_t^{\mathrm{treat}} + \sum_{b=1}^{B} x_{b,pr}^{\mathrm{block}}\beta_b^{\mathrm{block}} + \beta_p^{\mathrm{peptide}} + u_r^{\mathrm{run}} + \varepsilon_{pr}, \qquad \text{(Eq. 2)}$$

where β^0 is the intercept, β_t^{treat} is the effect of interest ($t = 1$ to 3 corresponding to spike-in conditions A, B, C for CPTAC, $t = 1$ to 6 for the six cardiac regions in the

HEART dataset), β_b^{block} is a blocking factor (lab $b = 1$ to 3 for CPTAC, patient $b = 1$ to 3 for HEART), $x_{t,pr}^{\text{treat}}$ and $x_{b,pr}^{\text{block}}$ are dummy variables which are equal to 1 if run r corresponds to treatment t or block b, respectively, and 0 otherwise. β_p^{peptide} is the effect of the individual peptide sequence p, u_r^{run} is a random effect that accounts for the fact that peptides within each run r are correlated. Due to the parameterization of the model, the following restrictions apply: $\sum_{t=1}^{T} \hat{\beta}_t^{\text{treat}} = 0$, $\sum_{b=1}^{B} \hat{\beta}_b^{\text{block}} = 0$, $\sum_{p=1}^{P} \hat{\beta}_p^{\text{peptide}} = 0$ and $\sum_{r=1}^{R} \hat{u}_r^{\text{run}} = 0$.

The effect sizes $\beta_{(.)}^{(.)}$ (except for the intercept) are estimated using penalized regression. Distinct ridge penalties are used for the treatment, block and peptide parameters, respectively and the ridge penalties are tuned by exploiting the link between ridge regression and mixed models (see e.g. Ruppert *et al.* (2003), chapter 4[18]). Outliers are accounted for using M-estimation with Huber weights. Protein-wise degrees of freedom are now calculated in a less liberal way as: $R - (1 + (T - 1) + (B - 1))$. Variances are stabilized by borrowing information over proteins using limma's empirical Bayes approach, which results in a moderated t-test. Scripts to run MSqRob are provided on our GitHub page (see below under "Code availability").

MSstats

MSstats has been described by Choi *et al.* (2014)[6]. For our analysis, we used the default settings of MSstats version 3.12.2. During preprocessing, feature intensities are log$_2$-transformed and normalized by equalizing the run medians. Then, missing values are imputed with the default MSstats settings. Features are summarized to the protein level with Tukey's median polish method. Treatment effects are specified as "Condition" and blocking factors as "BioReplicate" in the MSstats workflow.

Quasi-binomial model

When assuming that the number of observed peptides n_r^{peptide} in each run r (after preprocessing) for a protein are binomially distributed

$$n_r^{\text{peptide}} | \mathbf{x}_r \sim \text{Binomial}(P, \pi_r), \quad \text{(Eq. 3)}$$

with P the total number of unique peptides observed over all runs $r = 1, \ldots, R$ for this protein, and π_r the probability to identify a peptide for this protein; the peptide counts can be modeled using logistic regression. However, they are often overdispersed with respect to the binomial distribution. We therefore adopt quasi-binomial regression (McCullagh and Nelder (1989), section 4.5[19]) where we model the first two moments (mean $E[n_r^{\text{peptide}}]$ and variance $\text{Var}[n_r^{\text{peptide}}]$) as follows:

$$E[n_r^{\text{peptide}}] = P\pi_r, \quad \text{(Eq. 4)}$$

$$\text{logit}(\pi_r) = \log\left(\frac{\pi_r}{1 - \pi_r}\right) = \mathbf{x}_r^{\text{T}} \boldsymbol{\beta}, \quad \text{(Eq. 5)}$$

$$\text{Var}[n_r^{\text{peptide}}] = \varphi P \pi_r (1 - \pi_r), \quad \text{(Eq. 6)}$$

with φ a dispersion parameter accommodating for a more flexible variance function than that of the binomial regression model. Similar to the original edgeR count model for gene expression, information is borrowed over the dispersion parameters of all models with the empirical Bayes procedure of the limma R package[20]. To avoid overly liberal results from inaccurate variance estimations, underdispersed φ estimates are set to 1. Note that the quasi-binomial approach models the log odds, i.e. the logarithm of the probability that a peptide is detected divided by the probability that a peptide is not detected, i.e. the odds on detection. Hence, the model will return log ORs when contrasting different treatments to each other, which can be used to infer on differential peptide detection for a specific protein.

Hurdle model

The normalized intensities for each peptide p in each run r are typically assumed to follow a log-normal distribution. Upon log$_2$-transformation, missing (zero) intensities are set at $-\infty$ and cannot be modeled with intensity-based methods such as Perseus, MSstats and MSqRob. Missing values are therefore either omitted or imputed.

Here, we consider a hurdle model that consists of two parts: a binary component z_{pr} that distinguishes between peptide intensities in run r that are missing ($z_{pr} = 0$) or observed ($z_{pr} = 1$) with detection probability π_r; and a normal component y_{pr} with mean μ_{pr} and variance σ^2 to model log$_2$ peptide intensities passing the detection hurdle. Note, that the detection probability π_r and the mean μ_{pr} can be further parameterized using peptide specific effects, a random run effect u_r^{run}, and additional covariates $\mathbf{x}_{pr}$. More formally, the hurdle model for log$_2$-transformed intensities y_{pr} for peptide $p = 1, \ldots, P$ in run $r = 1, \ldots, R$ can be specified as follows:

$$z_{pr} | \mathbf{x}_{pr} \sim \text{Bernoulli}(\pi_r) \quad \text{(Eq. 7)}$$

$$y_{pr} | z_{pr} = 1, \mathbf{x}_{pr}, u_r^{\text{run}} \sim N(\mu_{pr}, \sigma^2) \quad \text{(Eq. 8)}$$

The log-likelihood for $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_R]^{\text{T}}$, $\boldsymbol{\mu} = [\mu_{11}, \ldots, \mu_{PR}]^{\text{T}}$ and σ^2 given $\mathbf{y} = [y_{11}, \ldots, y_{PR}]^{\text{T}}$, $\mathbf{z} = [z_{11}, \ldots, z_{PR}]^{\text{T}}$, $\mathbf{u}^{run} = [u_1, \ldots, u_R]^{\text{T}}$, and $\mathbf{X}$ the matrix with rows $\mathbf{x}_{pr}^{\text{T}}$ can then be written as:

$$l(\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma^2 | \mathbf{y}, \mathbf{z}, \mathbf{X}, \mathbf{u})$$
$$= \sum_{pr} (1 - z_{pr}) \log(1 - \pi_r) + \sum_{pr} z_{pr} \log(\pi_r)$$
$$+ \sum_{pr} z_{pr} \log[N(\mu_{pr}, \sigma^2)] \quad \text{(Eq. 9)}$$

Note, that the log-likelihood implies an estimation orthogonality between π_r and μ_{pr} and that the first two terms in the equation are equivalent to the log-likelihood of a Bernoulli process. Further, we omit a peptide-

specific effect for the parameterization of π_r because this leads to complete separation for too many proteins. The detection probability is thus considered constant for all peptides of a particular protein in run r. When summing over the peptides $p = 1, \ldots, P$, peptide counts $n_r^{\text{peptide}} = \sum_p z_{pr}$ are obtained and it is computationally more efficient to estimate π_r using a binomial model for the peptide counts. To account for overdispersion in the counts, we will again estimate the detection probability π_r using quasi-binomial regression with Model (Eq. $4 - 6$). The last term in the log-likelihood corresponds to a normal log-likelihood and we propose to estimate μ_{pr} and σ^2 using the MSqRob Model (Eq. 2).

Both model components model their means using the same covariate pattern $\boldsymbol{x}_{pr}$ and they will allow us to assess the same contrast of interest to infer on the log OR on detection and a $\log_2$ FC between conditions, respectively. Assuming independence between both statistics under the null hypothesis (Supplementary Table 5, Supplementary Figure 14), the hurdle model allows to assess the omnibus null hypothesis of no differential detection and no differential expression by combining the p-values of MSqRob without imputation and of the quasi-binomial model component. To this end, we first transform the p-values to z-values and combine them in a chi-square statistic:

$$\chi_{\text{hurdle}}^2 = z_{\text{MSqRob}}^2 + z_{\text{quasi-binomial}}^2 \qquad \text{(Eq. 10)}$$

If an MSqRob $\log_2$ FC can be estimated the chi-square statistic follows a chi-square distribution with 2 degrees of freedom under the omnibus null hypothesis, otherwise the corresponding p-value is equivalent to that of the quasi-binomial model.

Two-stage inference

We used stageR version 1.3.29 to implement a two-stage inference procedure[21]. It is a two-stage testing paradigm that leverages power of aggregating multiple tests per protein (here a test for differential detection, DD and differential abundance, DA) in the screening stage. Upon rejection of the omnibus null hypothesis, stageR performs post-hoc tests to assess on DD and DA. We adopt the Benjamini-Hochberg False Discovery Rate (FDR) procedure on the aggregated tests in the first stage. In the post-hoc analysis, we use the modified Holm procedure implemented in stageR to control the Family Wise Error Rate for the DD and DA tests within a protein at the FDR-adjusted significance level of the first stage.

Multiple testing correction

All p-values for MSstats, MSqRob, the quasi-binomial and the hurdle model were corrected for multiple testing using the Benjamini-Hochberg FDR correction. Perseus uses a permutation-based FDR based on 250 iterations.

Code availability

All scripts to reproduce the results in this contribution and in supplementary material are available at https://github.com/statOmics/MSqRobHurdlePaper.

RESULTS AND DISCUSSION

We use the CPTAC study 6 dataset[8] to compare our hurdle approach to the quasi-binomial model alone, to MSqRob alone, to Perseus (with and without imputation), and to MSstats. Fig. 1 c, d demonstrates the superior performance of the hurdle model: it consistently outperforms the other algorithms. Moreover, the hurdle model also always outperforms MSqRob with imputation under a low abundance assumption such as PI (Supplementary Fig. 5, 6). In the lowest spike-in condition, A, missing values in the UPS1 proteins are mainly caused by low abundance, allowing the count-component of the hurdle model to add to the strength of the intensity-component. In comparison C vs. B, the UPS1 spike-in concentrations are higher, reducing the difference in peptide counts between both conditions, and mainly driving significance of the hurdle model by the difference in average intensities. In this case, the hurdle model performs on par with MSqRob without imputation, as is expected, while PI approaches still perform poorly (Supplementary Fig. 5, 6).

Next, we assess the human heart dataset of Doll *et al.* (2017)[16]. For the 7,822 gene identifiers in common after preprocessing, we assessed the overlap between the 1,500 most significantly regulated identifiers in the Doll *et al.*, hurdle, and MSqRob comparisons between the atrial to the ventricular proteome (Fig. 2 a, Supplementary Table 2). The 671 identifiers shared between all methods correspond to proteins with a strong DA and many identified peptides (Supplementary Fig. 8). The 158 identifiers shared between Perseus and hurdle mostly have strong DD and only few peptides in one of the heart chambers (Supplementary Fig. 9). The 164 identifiers unique to hurdle differ strongly in peptide counts, but the $\log_2$ FCs are close to one. (Supplementary Fig. 10). The 596 identifiers unique to Perseus often show very few peptide identifications or very small FCs, making these results more questionable (Supplementary Fig. 11). Indeed, the imputation strategy again has a profound impact on the results: there is a 22% non-overlap between Perseus with and without the PI strategy. Moreover, because of the stochastic nature of the PI imputation, on average 4.5% of the first 1,500 proteins declared DA did not overlap between two repeated PI analyses. The 75 identifiers shared between MSqRob and Perseus, finally, mainly have many peptides, but a relatively small FC (Supplementary Fig. 12). These are not included in the top 1,500 differential proteins by the hurdle model as their place in the ranking is taken by identifiers with a strong DD. A similar analysis for the

500 and 1,000 most differential proteins for each method is given in Supplementary Fig. 13.

For our novel approach we also observe a strong association between the OR and FC estimates (Fig. 2 b). Out of the 1,765 hurdle identifiers that are significant at the 5% FDR level, 397 were significantly DD, 815 significantly DA, 440 both significantly DD and DA, and 113 could not be attributed to either DA or DD in our *post hoc* analysis. All 440 identifiers declared both DD and DA have a FC and OR estimate in the same direction (i.e. both up or both down).
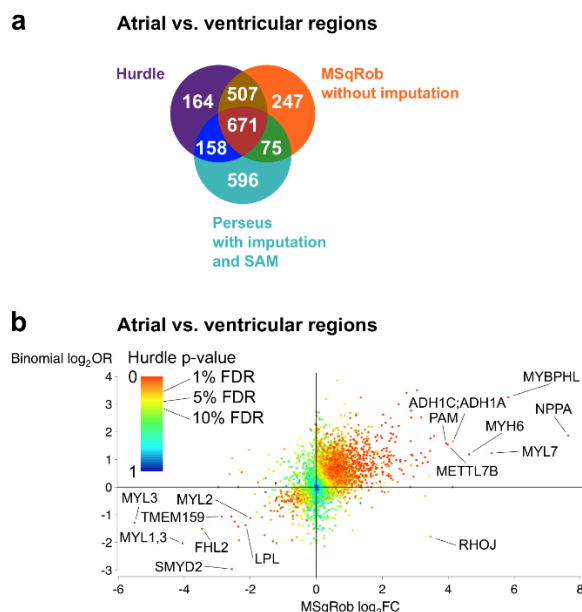
Figure 2. The hurdle model detects both DD and DA proteins. (a) The Venn diagram shows the overlaps in the 1,500 most significant proteins when contrasting atrial (average of left atrium, LA; right atrium, RA; atrial septum, SepA) to ventricular regions (average of left ventricle, LV; right ventricle, RV; ventricular septum, SepV) in the HEART dataset between hurdle, MSqRob without prior imputation, and Perseus with imputation and SAM as implemented by Doll *et al.* (b) Proteins that are significant with the hurdle model often display a strong positive association between FC and OR. Here, $\log_2$ FCs are plotted as a function of $\log_2$ ORs in the atrial vs. ventricular comparison for all 6,489 proteins with an FC estimate. Proteins are colored by p-value. Common FDR cut-offs are indicated on the color bar.

As in the original publication, comparing the left to the right ventricle did not render any significant proteins. However, when comparing the left to the right atrium, we found eight significant proteins compared to only three in the original publication (Supplementary Table 3). The higher abundance of the muscle contraction protein myotilin[22] in the left atrium was confirmed by both methods, but the hurdle model additionally indicates the heavy myosin chain MYH7 as DA. Moreover, the hurdle model also flags the voltage-

dependent calcium channel CACNA2D2 and BMP10, an essential factor for embryonal cardiac development[23]. Finally, SERINC3 and PNMA1, which were declared DA by Doll *et al.* actually come with very limited evidence for DA and their significance is likely driven by the inherent randomness of PI (Supplementary Table 4).

CONCLUSIONS

In summary, we demonstrated the negative impact of imputation in protein quantification, and the potential flaws it induces in the downstream analysis when imputation assumptions are violated. We therefore propose a hurdle approach that can handle missingness without having to resort to imputation and its associated assumptions by leveraging both an intensity-based approach (MSqRob) and a count-based quasi-binomial approach. Our hurdle approach outcompetes state-of-the-art methods with and without imputation in the presence of both strong missingness as well as when missingness is limited. Moreover, our hurdle method continues to provide valid inference when peptides are absent in one condition. It is therefore a highly promising approach to detect the sudden appearance of post-translationally modified peptides in protein regulation alongside more traditional differential protein expression.

ASSOCIATED CONTENT

Supporting Information

supplementary_informationHurdle.pdf: an extended methods section and all supplementary figures. (PDF)

supplementary_table1.xlsx: Overview of the projects and files used in Fig. 1a together with the number of missing values, the number of observed values and the percentage of missing values in each file. (XLSX)

supplementary_table2.xlsx: Overview of the relevant protein-level output from the hurdle model (sheet 2), the Pereus with imputation and SAM approach used by Doll *et al.* (sheet 3), MSqRob (sheet 4) and the quasibinomial model (sheet 5) when comparing the atrial to the ventricular regions in the HEART dataset. (XLSX)

supplementary_table3.xlsx: Overview of the $\log_2$ FC, $\log_2$ OR, χ^2 statistics, p-values and Benjamini-Hochberg-corrected p-values (q-values) when comparing the left to the right atrium with the hurdle model. (XLSX)

supplementary_table4.xlsx: Overview of the data underlying proteins SERINC3, PNMA1, ACSM2A, PDE7B and PNPLA7. (XLSX)

supplementary_table5.xlsx: Overview of the mock treatment levels and the condition-lab combinations for each MS run in the mock analysis. (XLSX)

AUTHOR INFORMATION

Corresponding Author

* Tel.: +32 9 264 49 04. Fax: +32 264 49 95. E-mail: Lieven.Clement@UGent.be.

Present Addresses

†   Laboratory of Integrative and Systems Physiology, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland.

Author Contributions

L.G. contributed R code, analyzed the data and wrote the paper, A.S. contributed R code and wrote the paper, K.G. and L.M. wrote the paper. L.C. contributed R code, conceived the idea and wrote the paper.

Notes

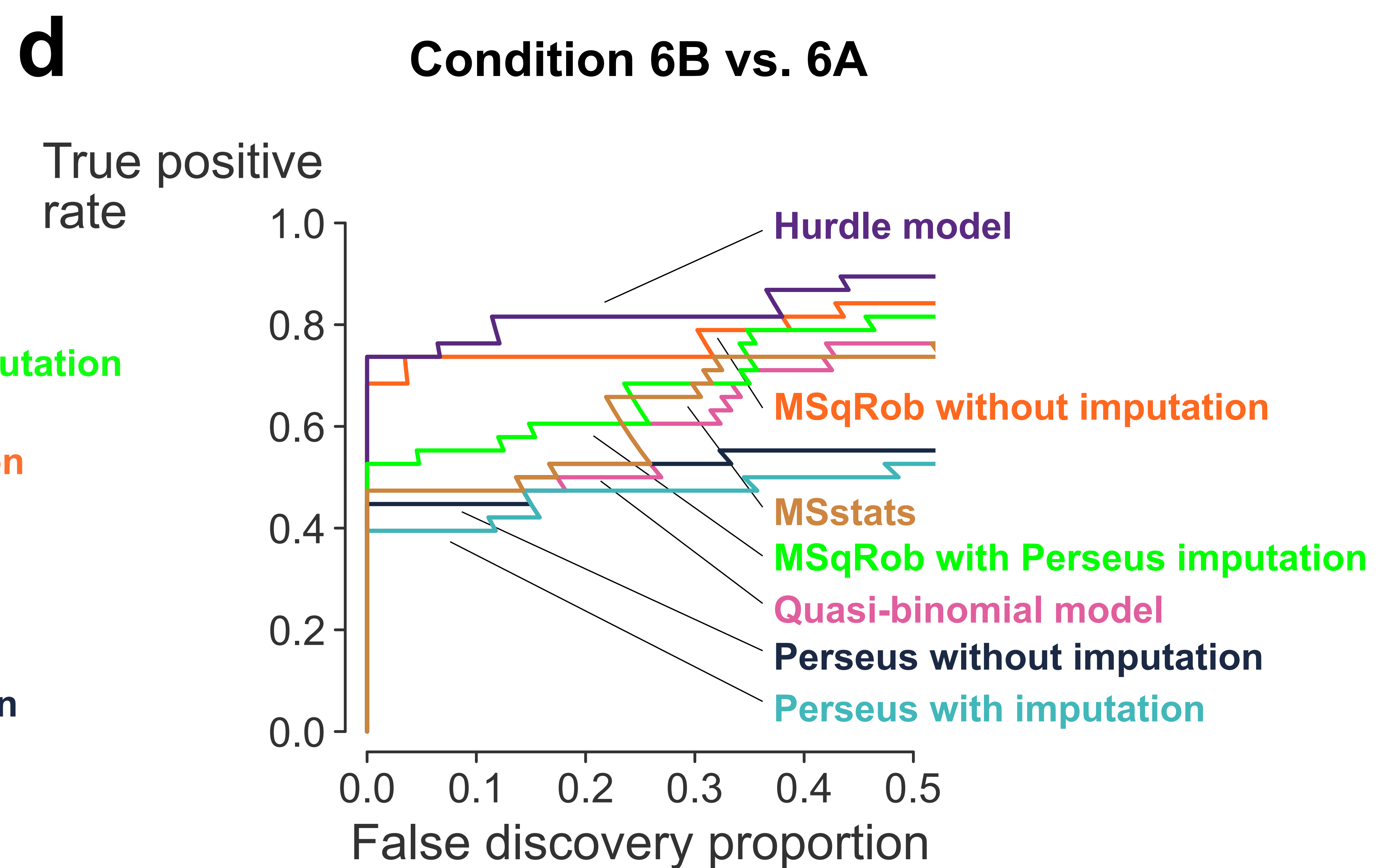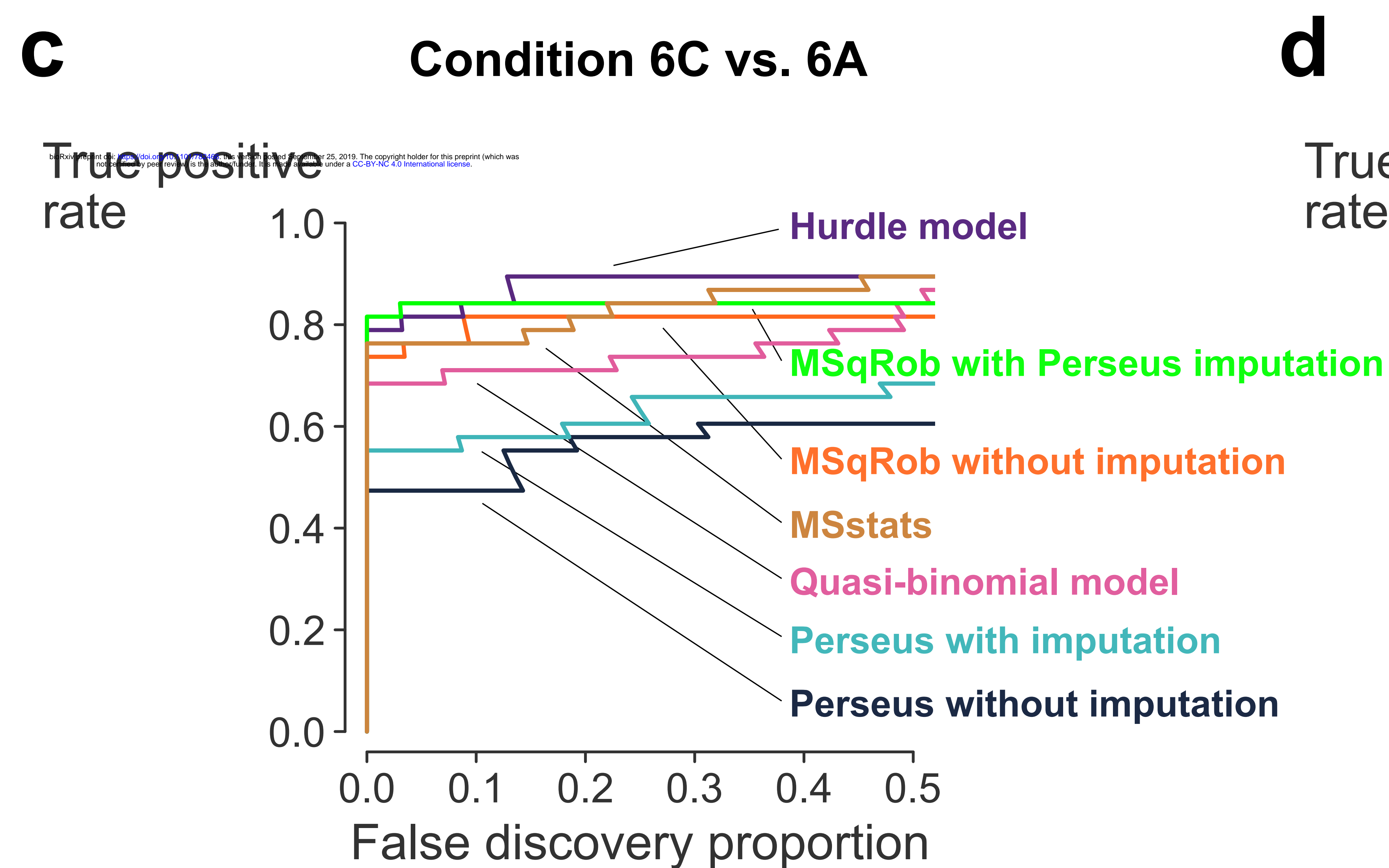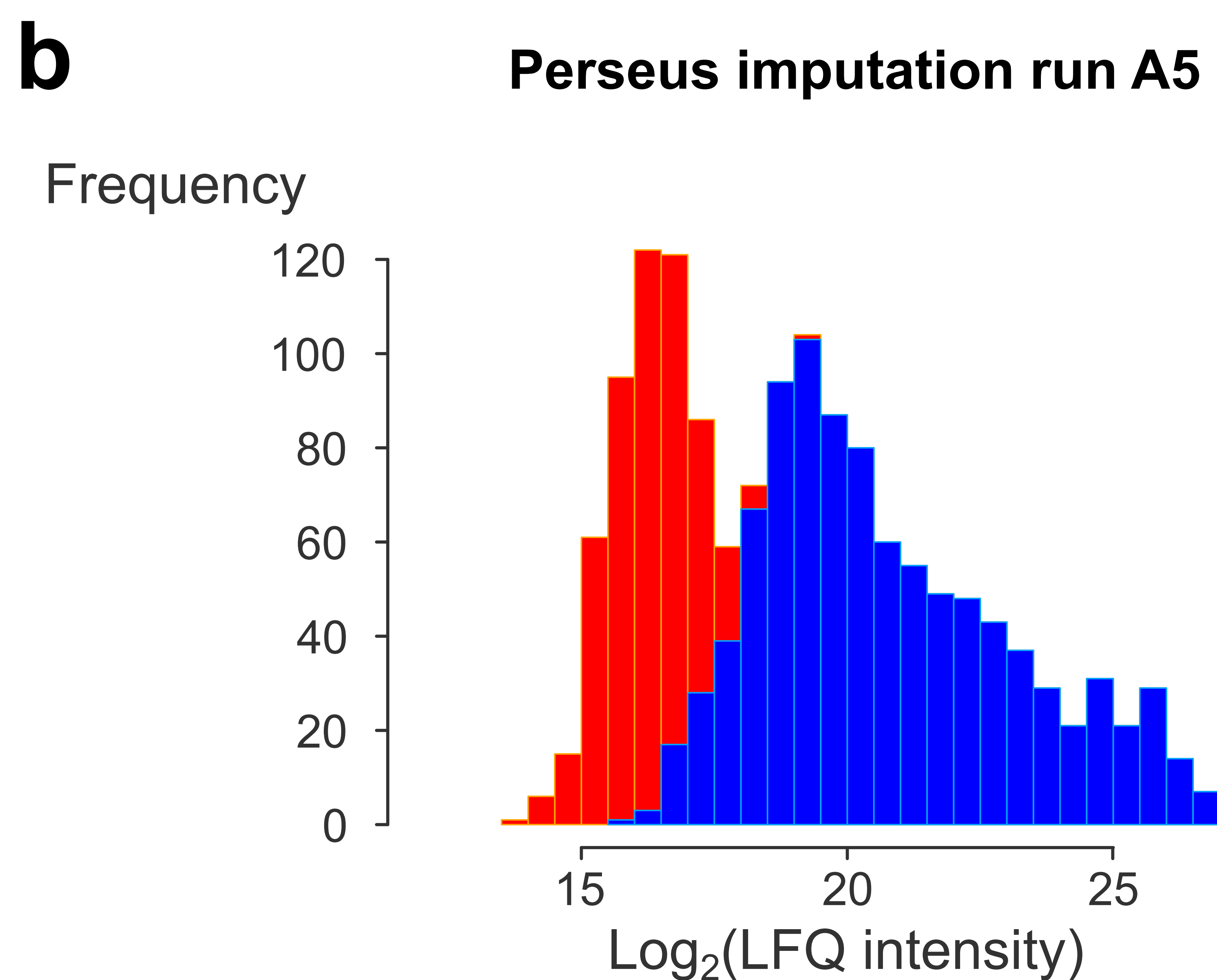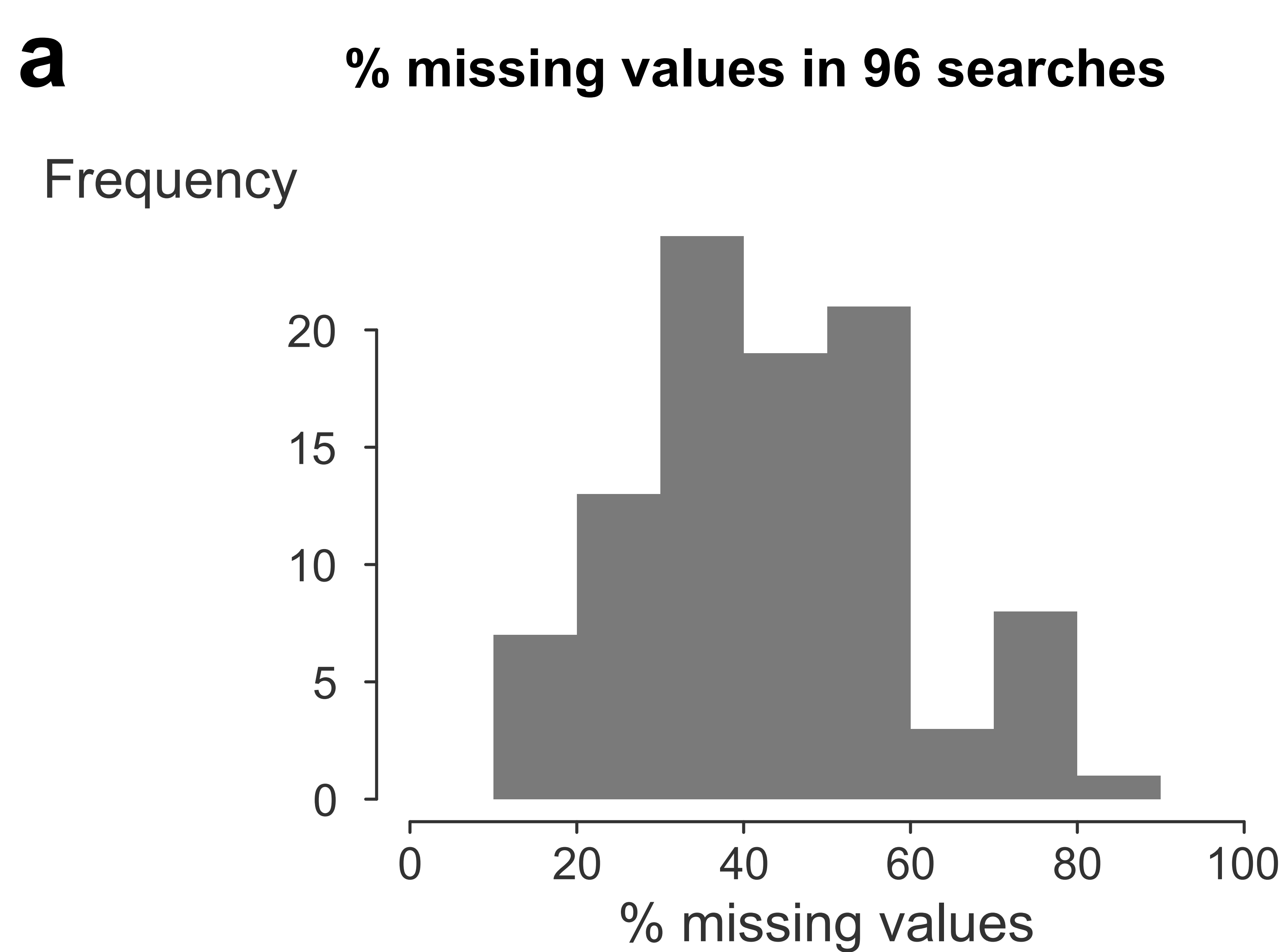The authors declare no competing interests.

REFERENCES

1.   Hindupur, S. K.; Colombi, M.; Fuhs, S. R.; Matter, M. S.; Guri, Y.; Adam, K.; Cornu, M.; Piscuoglio, S.; Ng, C. K. Y.; Betz, C.; Liko, D.; Quagliata, L.; Moes, S.; Jenoe, P.; Terracciano, L. M.; Heim, M. H.; Hunter, T.; Hall, M. N., The protein histidine phosphatase LHPP is a tumour suppressor. *Nature* **2018,** *555*, 678.

2.   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.; Waters, K. M., Review, Evaluation, and Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry-Based Label-Free Global Proteomics. *Journal of Proteome Research* **2015,** *14* (5), 1993-2001.

3.   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. *Journal of Proteome Research* **2016,** *15* (4), 1116-1125.

4.   Tyanova, S.; Temu, T.; Cox, J., The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols* **2016,** *11* (12), 2301-2319.

5.   Tyanova, S.; Temu, T.; Sinitcyn, P.; Carlson, A.; Hein, M. Y.; Geiger, T.; Mann, M.; Cox, J., The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods* **2016,** *13*, 731.

6.   Choi, M.; Chang, C.-Y.; 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), 2524-2526.

7.   Karpievitch, Y. V.; Dabney, A. R.; Smith, R. D., Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics* **2012,** *13 Suppl 16*, S5.

8.   Paulovich, A. G.; Billheimer, D.; Ham, A.-J. L.; Vega-Montoto, L.; Rudnick, P. A.; Tabb, D. L.; Wang, P.; Blackman, R. K.; Bunk, D. M.; Cardasis, H. L.; Clauser, K. R.; Kinsinger, C. R.; Schilling, B.; Tegeler, T. J.; Variyath, A. M.; Wang, M.; Whiteaker, J. R.; Zimmerman, L. J.; Fenyo, D.; Carr, S. A.; Fisher, S. J.; Gibson, B. W.; Mesri, M.; Neubert, T. A.; Regnier, F. E.; Rodriguez, H.; Spiegelman, C.; Stein, S. E.; Tempst, P.; Liebler, D. C., Interlaboratory Study Characterizing a Yeast Performance Standard for Benchmarking LC-MS Platform Performance. *Molecular & Cellular Proteomics* **2010,** *9* (2), 242-254.

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. *Molecular & Cellular Proteomics* **2016,** *15* (2), 657-668.

10.   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. *Journal of Proteome Research* **2015,** *14* (6), 2457-2465.

11.   Blein-Nicolas, M.; Zivy, M., Thousand and one ways to quantify and compare protein abundances in label-free bottom-up proteomics. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **2016,** *1864* (8), 883-895.

12.   Liu, K.; Zhang, J.; Wang, J.; Zhao, L.; Peng, X.; Jia, W.; Ying, W.; Zhu, Y.; Xie, H.; He, F.; Qian, X., Relationship between Sample Loading Amount and Peptide Identification and Its Effects on Quantitative Proteomics. *Analytical Chemistry* **2009,** *81* (4), 1307-1314.

13.   Webb-Robertson, B.-J. M.; McCue, L. A.; Waters, K. M.; Matzke, M. M.; Jacobs, J. M.; Metz, T. O.; Varnum, S. M.; Pounds, J. G., Combined Statistical Analyses of Peptide Intensities and Peptide Occurrences Improves Identification of Significant Peptides from MS-Based Proteomics Data. *Journal of Proteome Research* **2010,** *9* (11), 5748-5756.

14.   Dicker, L.; Lin, X.; Ivanov, A. R., Increased Power for the Analysis of Label-free LC-MS/MS Proteomics Data by Combining Spectral Counts and Peptide Peak Attributes. *Molecular & Cellular Proteomics* **2010,** *9* (12), 2704-2718.

15.   Chen, Y.-Y.; Chambers, M. C.; Li, M.; Ham, A.-J. L.; Turner, J. L.; Zhang, B.; Tabb, D. L., IDPQuantify: Combining Precursor Intensity with Spectral Counts for Protein and Peptide Quantification. *Journal of Proteome Research* **2013,** *12* (9), 4111-4121.

16.   Doll, S.; Dreßen, M.; Geyer, P. E.; Itzhak, D. N.; Braun, C.; Doppler, S. A.; Meier, F.; Deutsch, M.-A.; Lahm, H.; Lange, R.; Krane, M.; Mann, M., Region and cell-type resolved quantitative proteomic map of the human heart. *Nature Communications* **2017,** *8* (1), 1469.

17.   Beretta, L.; Santaniello, A., Nearest neighbor imputation algorithms: a critical evaluation. *BMC Medical Informatics and Decision Making* **2016,** *16* (3), 74.

18.   Ruppert, D.; Wand, M. P.; Carroll, R. J., *Semiparametric Regression*. Cambridge University Press: 2003.

19.   McCullagh, P.; Nelder, J., *Generalized Linear Models SECOND EDITION*. Chapman and Hall/CRC 1989; p 532.

20.   Robinson, M. D.; McCarthy, D. J.; Smyth, G. K., edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010,** *26* (1), 139-140.
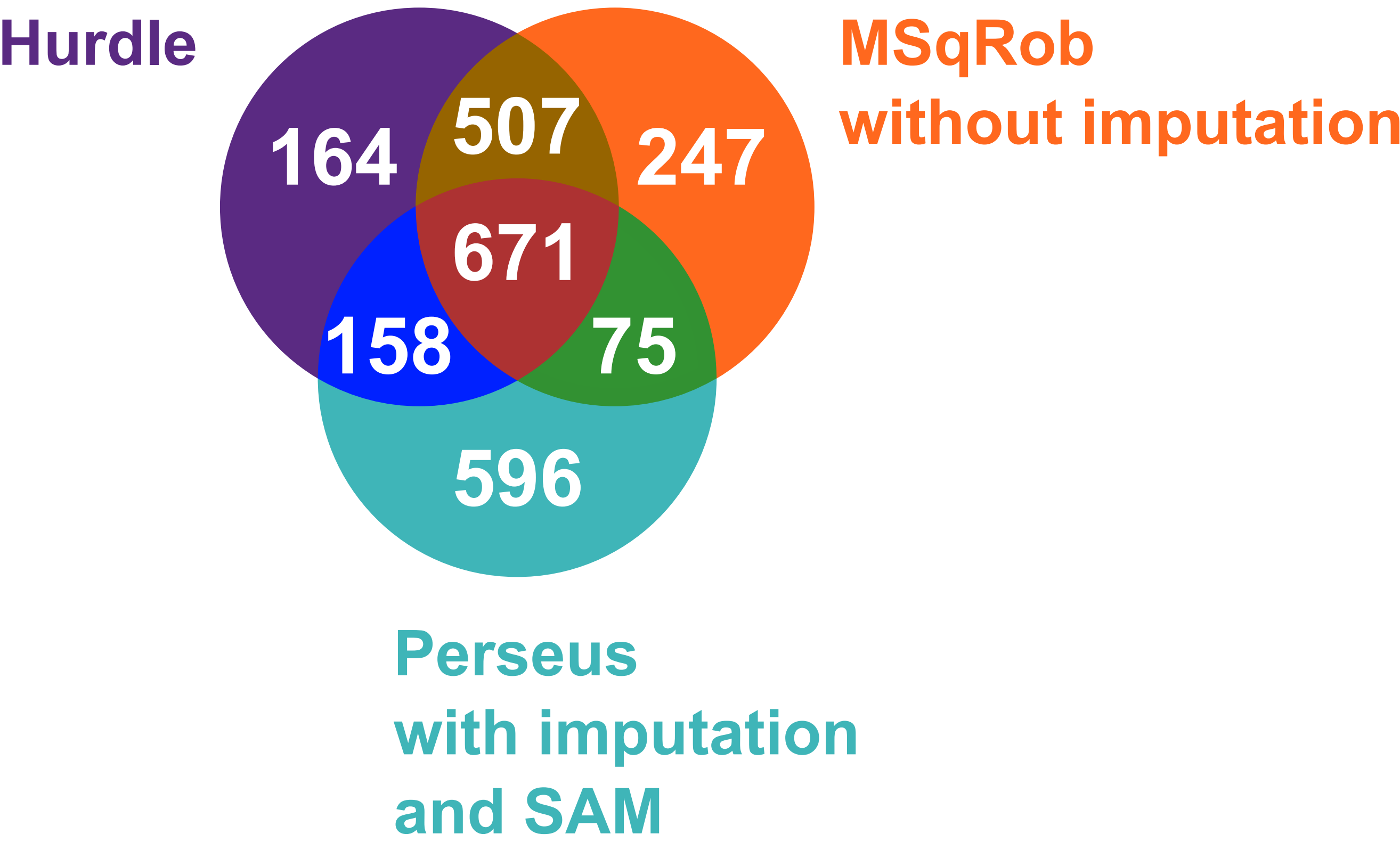
21.   Van den Berge, K.; Soneson, C.; Robinson, M. D.; Clement, L., stageR: a general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage. *Genome Biology* **2017,** *18* (1), 151.

22.   Wang, J.; Dube, D. K.; Mittal, B.; Sanger, J. M.; Sanger, J. W., Myotilin dynamics in cardiac and skeletal muscle cells. *Cytoskeleton* **2011,** *68* (12), 661-670.

23.   Huang, J.; Elicker, J.; Bowens, N.; Liu, X.; Cheng, L.; Cappola, T. P.; Zhu, X.; Parmacek, M. S., Myocardin regulates BMP10 expression and is required for heart development. *The Journal of Clinical Investigation* **2012,** *122* (10), 3678-3691.

a % missing values in 96 searches

Frequency

% missing values

b Perseus imputation run A5

Frequency

Log$_2$(LFQ intensity)

c Condition 6C vs. 6A

True positive rate

Hurdle model
MSqRob with Perseus imputation
MSqRob without imputation
MSstats
Quasi-binomial model
Perseus with imputation
Perseus without imputation

False discovery proportion

d Condition 6B vs. 6A

True positive rate

Hurdle model
MSqRob without imputation
MSstats
MSqRob with Perseus imputation
Quasi-binomial model
Perseus without imputation
Perseus with imputation

False discovery proportion

a   Atrial vs. ventricular regions

b   Atrial vs. ventricular regions

Binomial log$_2$OR   Hurdle p-value