Moderated Test Statistics to Detect Differential Deuteration in Hydrogen/Deuterium Exchange Mass Spectrometry Experiments

DOI: 10.1021/acs.analchem.1c02346
Handle: http://hdl.handle.net/1942/36962

Moderated test statistics to detect differential deuteration in HDX-MS experiments.

Jürgen Claesen[1,2], Srinath Krishnamurthy[3], Andy M Lau[4], Anastassios Economou[3]

[1] Department of Epidemiology and Data Science, Amsterdam UMC, VU University Amsterdam, 1081 HV Amsterdam, The Netherlands

[2] I-BioStat, Data Science Institute, Hasselt University, 3500 Hasselt, Belgium

[3] Laboratory of Molecular Bacteriology, Rega Institute of Medical Research, KU Leuven, 3000 Leuven, Belgium

[4] Department of Computer Science, University College London, London WC1E 6BT, UK

Corresponding author: j.claesen@amsterdamumc.nl

Abstract

With differential Hydrogen/Deuterium exchange, differences in the structure and dynamics of protein states can be studied. Detecting statistically significant differentially deuterated peptides is crucial to draw meaningful conclusions about the distinct conformations and dynamics of the protein under study. Here, we introduced a linear model in combination with an empirical Bayes approach to detect differentially deuterated peptides. Using a linear model allows to test for differences in deuteration between two (two-sample t-test) or more groups (F-statistic), while potentially controlling for the effects of other variables that are not of interest. The empirical Bayes approach improves the estimation of deuteration-level variances, especially in experiments with a low number of replicates. As a consequence, the two sample t-test and the F-statistic become moderated, resulting in a lower number of false positive and false negative findings. Furthermore, we introduce a thresholded moderated t-statistic to test if the observed deuteration differences are larger than a specified, biologically relevant difference. Finally, we underline the importance of having a sufficient number of replicates, and the effect of the number of replicates on the power of the statistical significance tests. The R-code for the proposed moderated test-statistics is available upon request.

1 Introduction

Hydrogen/Deuterium exchange (HX) combined with mass spectrometry (MS) is an analytical technique to study the conformation and dynamics of proteins [1]. Among various experimental approaches, differential HX-MS (ΔHX-MS) is commonly applied. It allows one to identify differences in structure and/or dynamics between multiple conditions or protein states, by comparing deuteration levels [2]. Typically, a bottom-up workflow is used in ΔHX-MS, i.e., the deuterated protein is digested into many peptides whose deuteration-levels are then determined by MS. To detect peptides that are significantly differentially deuterated between two or more conditions, a fixed threshold for the differences in deuteration-levels is sometimes used. Commonly, such empirical

thresholds require that deuteration differences between peptides are larger than 0.5Da. However, specifying a threshold ignores the variability of the peptide-deuteration-levels, and therefore leads to false positives (FP). Additionally, imposing strict thresholds to reduce the number of FPs, leads to ignoring small and moderate, but biologically relevant differences in deuteration, i.e., false negatives (FN). Statistical tests should be used instead, as they offer better control of FPs and FNs. However, the limited number of replicates in complex HX-MS experiments in combination with a moderately large number of peptides poses several challenges. When testing many hypotheses, the probability of finding FPs is large. In order to limit the number of false positives, family-wise error rate (FWER) and false discovery rate (FDR) controlling procedures are proposed [3]. Classical FWER controlling procedures are known to substantially reduce the power of statistical tests [4]. For example, in case of Bonferroni's procedure, the significance level, α, decreases fast when the number of tests increases, $\alpha^{\text{corrected}} = \alpha/(\text{number of tests})$. As a consequence, it becomes more difficult to reject a null hypothesis. FDR controlling procedures focus on the number of erroneously rejected null hypotheses among the rejected hypotheses. Controlling this proportion is less stringent than controlling the probability of finding one FP [4]. Additionally, standard univariate statistical tests, such as the t-test or Wilcoxon test, are less suited for high-dimensional settings where the number of variables to be tested is much larger than the number of observations. In such a setting, these tests are less precise and, as a result, high FDRs are found.

Comparing many variables between two or more conditions is not exclusively linked to ΔHX-MS. It also occurs in, for instance, quantitative proteomics and transcriptomics, where, respectively, protein abundance and gene expression are of interest. For differential gene expression experiments, statistical tests that borrow information between genes using empirical Bayes approaches have been developed [5–8]. In this article, we introduce the empirical Bayes approach and moderated test statistics in the setting of ΔHX-MS. We illustrate the performance of the proposed methods by means of simulated datasets and a real-life dataset.

2 Methodology

In a ΔHX-MS experiment, differences in deuteration between conditions are of interest. The t-test and confidence intervals based on the t-test are commonly used to identify statistically significant differences in deuteration-levels (see Supporting Information). Recently, an approach that combines a global confidence interval with a two-sample t-test has been proposed [9] (see Supporting Information). In this section, we introduce a linear model for statistical testing. Additionally, we present the empirical Bayes approach and the moderate F- and t-statistic.

2.1 Linear model

For a peptide, p, the deuteration-values can be described with the following linear model:

$$y_{pi} = \mathbf{X}\boldsymbol{\beta}_p + \varepsilon_{pi}, \tag{1}$$

where $\boldsymbol{y}_p$ is a vector of n deuteration values, $\boldsymbol{y}_p = [y_{p1}, \ldots, y_{pn}]^T$, $\boldsymbol{\beta}_p$ is a vector of m unknown coefficients, $\mathbf{X}$ a $(n \times p)$-matrix reflecting the experimental design, $\varepsilon_{pi} \sim \mathcal{N}(0, \sigma_p^2)$, and σ_p^2 is the residual variance. The deuteration values $\boldsymbol{y}_p$ can be computed by subtracting the average masses of the undeuterated peptide from the deuterated average masses, or by deconvoluting the isotope distribution of the (partially) deuterated peptide.

For example, consider a ΔHX-MS experiment with two conditions, $c1$ and $c2$, two labeling times, $t1$ and $t2$, and two replicates. The design matrix of this experiment is:

$$\mathbf{X}^T = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

with the number of columns of $\mathbf{X}^T$ corresponding to the number of observations, and the number of

rows corresponding to the number of unknown coefficients $\boldsymbol{\beta_p} = (\beta_{p,(c1,t1)}, \beta_{p,(c1,t2)}, \beta_{p,(c2,t1)}, \beta_{p,(c2,t2)})$.

Differences in deuteration levels between conditions and/or labeling times can be tested by specifying contrasts between the coefficients $\boldsymbol{\beta}_p$, $\boldsymbol{\gamma}_p = \mathbf{C}^T\boldsymbol{\beta}_p$ where $\mathbf{C}$ denotes a contrast matrix. The t-statistic for contrast j is of following form:

$$t_{\hat{\gamma}_{pj}} = \frac{\hat{\gamma}_{pj}}{\text{s.e.}(\hat{\gamma}_{pj})} \tag{2}$$

where s.e.$(\hat{\gamma}_{pj})$ is the standard error of the estimated value $\hat{\gamma}_{pj}$. The standard error is derived from the sample residual variance, s_p^2, that follows, conditionally on σ_p^2, a scaled chi-square distribution:

$$s_p^2|\sigma_p^2 \sim \frac{\sigma_p^2}{d_p}\chi^2_{d_p} \tag{3}$$

where d_p is the degrees of freedom for the linear model for peptide p.

In our example, we are interested in the differences in deuteration between the two conditions at each time point:

$$\boldsymbol{\gamma}_p = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix} \times \begin{bmatrix} \beta_{p,(c1,t1)} \\ \beta_{p,(c1,t2)} \\ \beta_{p,(c2,t1)} \\ \beta_{p,(c2,t2)} \end{bmatrix} = \begin{bmatrix} \beta_{p,(c1,t1)} - \beta_{p,(c2,t1)} \\ \beta_{p,(c1,t2)} - \beta_{p,(c2,t2)} \end{bmatrix}.$$

2.2 Empirical Bayes and moderated test statistics

The residual variance, σ_p^2, cannot be precisely estimated when the sample size is small. Hence, using the t-statistic (2) to identify statistically significant differences in deuteration-levels will lead to a higher number of false findings than is desirable. As the same linear model (1) is fitted for each peptide, one can borrow information from the whole group of peptides. Using a pooled residual standard deviation, as proposed by [10, 11], is a possibility but ignores the peptide-specific variability. The variance of peptides with low variability will be overestimated, while the variance

of peptides with high variability will be underestimated. As an alternative, the residual variance of peptides can be moderated with empirical Bayes [12–14]. Empirical Bayes methods use a Bayesian hierarchical model to estimate the variance of the peptides. This model describes how the residual variance varies across peptides with an inverse-χ^2 prior located at prior estimate s_0^2 with d_0 degrees of freedom:

$$\sigma_p^2 \sim \frac{s_0^2 d_0}{\chi_{d_0}^2}. \tag{4}$$

The prior estimate, s_0^2, and the degrees of freedom, d_0, are estimated from the whole set of peptides [8]. Based on the prior distribution (4) and the observed distribution (3), the posterior variance, given s_p^2, is:

$$\tilde{s}_p^2 = \frac{d_0 s_0^2 + d_p s_p^2}{d_0 + d_p}. \tag{5}$$

The posterior variances shrink the observed variances, s_p^2, towards the prior estimate, s_0^2, with the level of shrinkage depending on the relative sizes of the observed and prior degrees of freedom.

The moderated posterior variances $\tilde{s}_p^2$ can be converted to moderated standard errors, $\widetilde{\text{s.e.}}(\hat{\gamma}_{pj})$, and replace s.e.$(\hat{\gamma}_{pj})$ in the classical t-statistic [8]:

$$\widetilde{t}_{\hat{\gamma}_{pj}} = \frac{\hat{\gamma}_{pj}}{\widetilde{\text{s.e.}}(\hat{\gamma}_{pj})}. \tag{6}$$

Under the null hypothesis, H_0: $\gamma_{pj} = 0$, the moderated t-statistic follows a t-distribution with $d_p + d_0$ degrees of freedom [8].

Similarly, the F-statistic, which can be used to test all contrasts simultaneously, can be converted to a moderated F-statistic by replacing the variance by the moderated posterior variance. The degrees of freedom of the F-distribution also increase from d_p to $d_p + d_0$, reflecting the extra information from all peptides.

2.3 Testing against a threshold

The t-statistic (2) and the moderated t-statistic (6) are used to test the null hypothesis H_0: $\gamma_{pj} = 0$ against the alternative hypothesis H_1: $\gamma_{pj} \neq 0$. In order to avoid considering biologically meaningless differences in deuteration to be statistically significant, hypothesis testing can be combined with specifying a threshold value, δ. The threshold is used to remove peptides with too small deuteration differences after testing. This approach ignores the variability of the deuteration-values, and results in an FDR that is lower than specified. A more appropriate approach is to reformulate the tested hypotheses, i.e., H_0: $|\gamma_{pj}| \leq \delta$ against H_1: $|\gamma_{pj}| > \delta$. McCarthy and Smith [15] introduced a method to determine p-values for any δ when using the moderated t-statistic:

$$\begin{aligned} p\text{-value} &= P(t_{pj}(\gamma_0) < -t_{pj}^{obs} - \delta | \gamma_{pj} = \gamma_0, \tilde{s}_p) \\ &+ P(t_{pj}(\gamma_0) > t_{pj}^{obs} - \delta | \gamma_{pj} = \gamma_0, \tilde{s}_p) \\ &= F(t_{pj}^{obs} + \delta) + F(t_{pj}^{obs} - \delta), \end{aligned}$$

where $\gamma_{pj} = \gamma_0$ is a conservative approximation of the null hypothesis, t_{pj}^{obs} is the observed moderated t-statistic and $F()$ is the uppertail probability function of the t-distribution with $d_0 + d_p$ degrees of freedom.

3 Data

We evaluated the performance of the proposed moderated test-statistics with simulated data and real-life data.

3.1 Simulated data

We simulated data reflecting a real-life experiment with 108 different peptides, measured in three different states ('0', '1' and '2'). Each peptide was exposed to deuterium for 1min, 10mins, 25mins, 60mins and 1440mins. For each peptide i, we generated deuteration values, D_i, from a normal distribution with mean μ_i and variance σ_i^2. The values for μ_i (Tables S1-S3) are based on the NucB2 dataset of the HaDeX R-package [16].

We considered four possible settings for the variances of the peptides, i.e., very small variation ($\sigma_i^2 \sim \mathcal{U}(0.001, 0.05)$), small variation ($\sigma_i^2 \sim \mathcal{U}(0.001, 0.2)$), medium variation ($\sigma_i^2 \sim \mathcal{U}(0.001, 0.8)$), and large variation ($\sigma_i^2 \sim \mathcal{U}(0.001, 1.6)$). Note that the term "variation" is used here to describe the range of the variances of the deuteration values. Therefore, in the "large variation" setting, the variances range from 0.001 to 1.6, while in the "very small variation setting", the variances range from 0.001 to 0.05. In case of the "very small variation"-setting, the variances are also small. For the other settings, a combination of small and medium, or small, medium and large variances can be found. Additionally, we varied the number of replicates, i.e., three, five and ten. For each of the 12 different scenarios, we generated 1,000 datasets.

3.2 Real-life data

The differences in dynamics of a monomeric, dimeric and a C-tail removed variant of SecA, in bound and unbound state, were investigated with HX-MS [17]. For each of the six SecA variants, HX was recorded, in triplicate, at seven different time points, i.e., 10s, 30s, 1min, 2mins, 5mins 10mins and 30mins.

4 Results

4.1 Simulated data

We compared the moderated test statistics, i.e., the F-, t-, and thresholded t-statistic, with the standard t-statistic, the 'global threshold'-approach [10, 11], the hybrid significance test [9] and the 'manual thresholding'-approach. Additionally, we also checked the performance of the moderated t-statistic corrected with the Benjamini-Hochberg method for multiple testing. For each scenario and each method, we determined the number of FPs, FNs, TPs and TNs at $\alpha = 0.05$. Note that for the hybrid significance test, peptides for which the difference between the deuteration-values was smaller than the global threshold, the p-value was set to 1. For each method, the number of FPs, FNs, TPs and TNs were averaged over the 1,000 datasets.

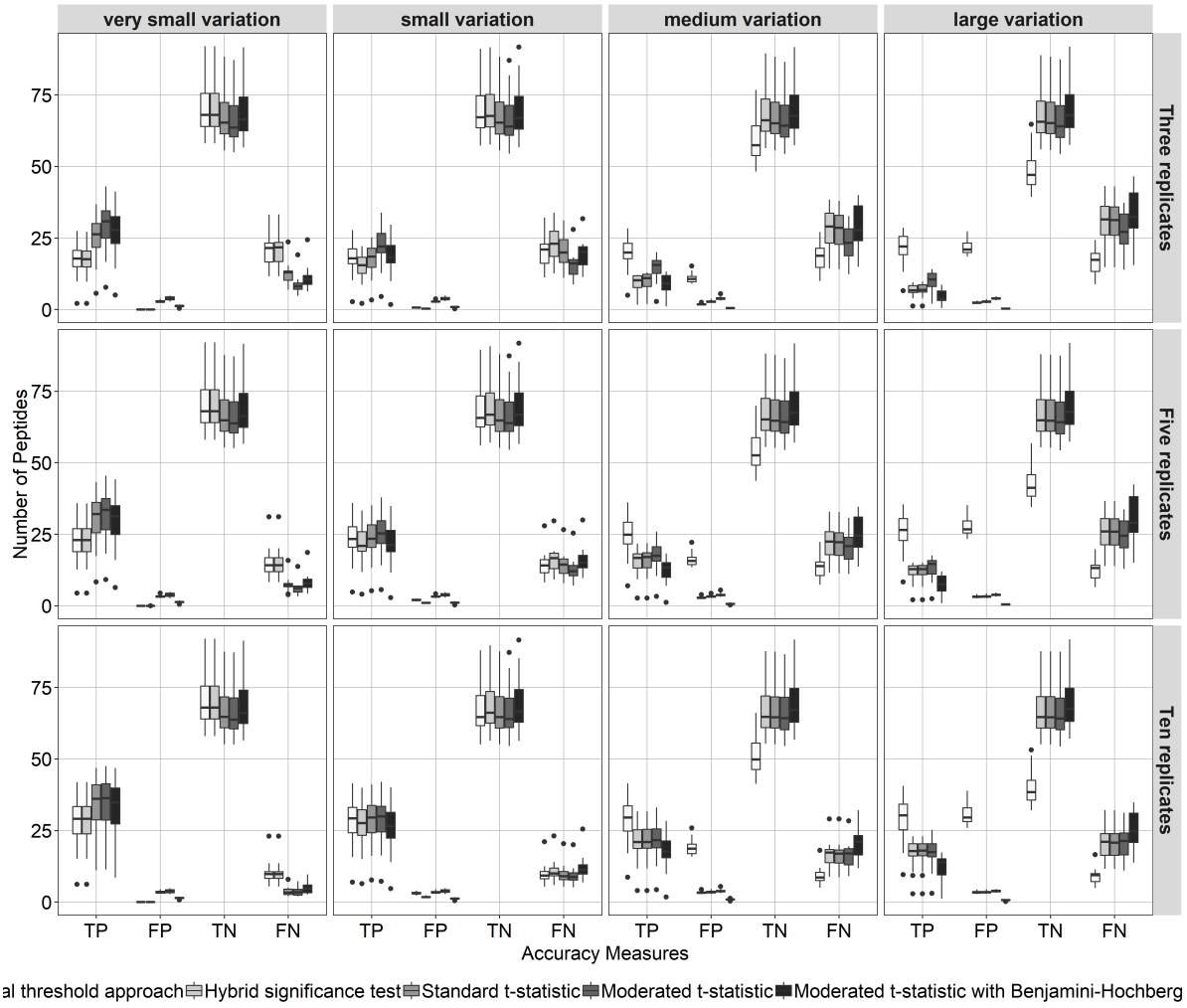
4.1.1 Pairwise comparisons

The moderated t-statistic, the standard t-statistic, the 'global threshold'-approach and the hybrid significance test allow to test for differences in the deuteration-levels of peptides between two groups or conditions.

As expected, the ability to identify statistically significant differences in deuteration-levels is a function of the number of replicates and of the variance of the deuteration-levels. The number of TPs and TNs increases and the number of FNs decreases when the number of replicates increases and/or when the variance of the deuteration-levels decreases (Figure 1). Additionally, as the variance of the deuteration-levels of a peptide can be estimated more accuractely with increasing numbers of replicates, the difference between the standard t-statistic and the moderated t-statistic becomes smaller, and eventually disappears.

For the "small variation" setting, the performance of all the tested methods for pairwise comparisons is comparable. The 'global threshold'-approach, the hybrid significance test and the moderated t-statistic corrected with the Benjamini-Hochberg procedure are more conservative than the moderated t-statistic but these differences are neglectable. The performance of the standard

Figure 1: Accuracy measures of the methods for pairwise comparisons at $\alpha = 0.05$.



t-statistic changes from similar to the 'global threshold'-approach to similar to the moderated t-statistic, as the number of replicates increases.

When the variance of D_i is very small ("very small variation"), the performance of the 'global threshold'-approach is identical to the hybrid significance test, but worse than the performance of the (corrected) moderated t-statistic and the standard t-statistic. A substantially lower number of TPs and higher number of FNs are found, while there is no or a neglectible difference in the number of TNs and a marginally lower number of FPs. A low number of TPs and no FPs in combination with many FNs indicates that, in this setting, the 'global threshold'-approach overestimates the deuteration-variances of many peptides. The hybrid significance test, i.e., applying Welch's t-test

on the set of peptides selected by the 'global threshold'-approach, to confirm that these peptides are differentially deuterated, is not able to correct for this overestimation.

In case of the "medium variation" and "large variation" settings, the 'global threshold'-approach identifies more TPs, but has a substantial higher number of FPs, and a considerable smaller number of TNs and FNs. The hybrid significance test has a similar number of FPs as the (moderated) t-statistic, indicating that, for these settings, the 'global threshold'-approach underestimates the deuteration-variances of several peptides, and that this underestimation-issue can be corrected by the hybrid significance test.

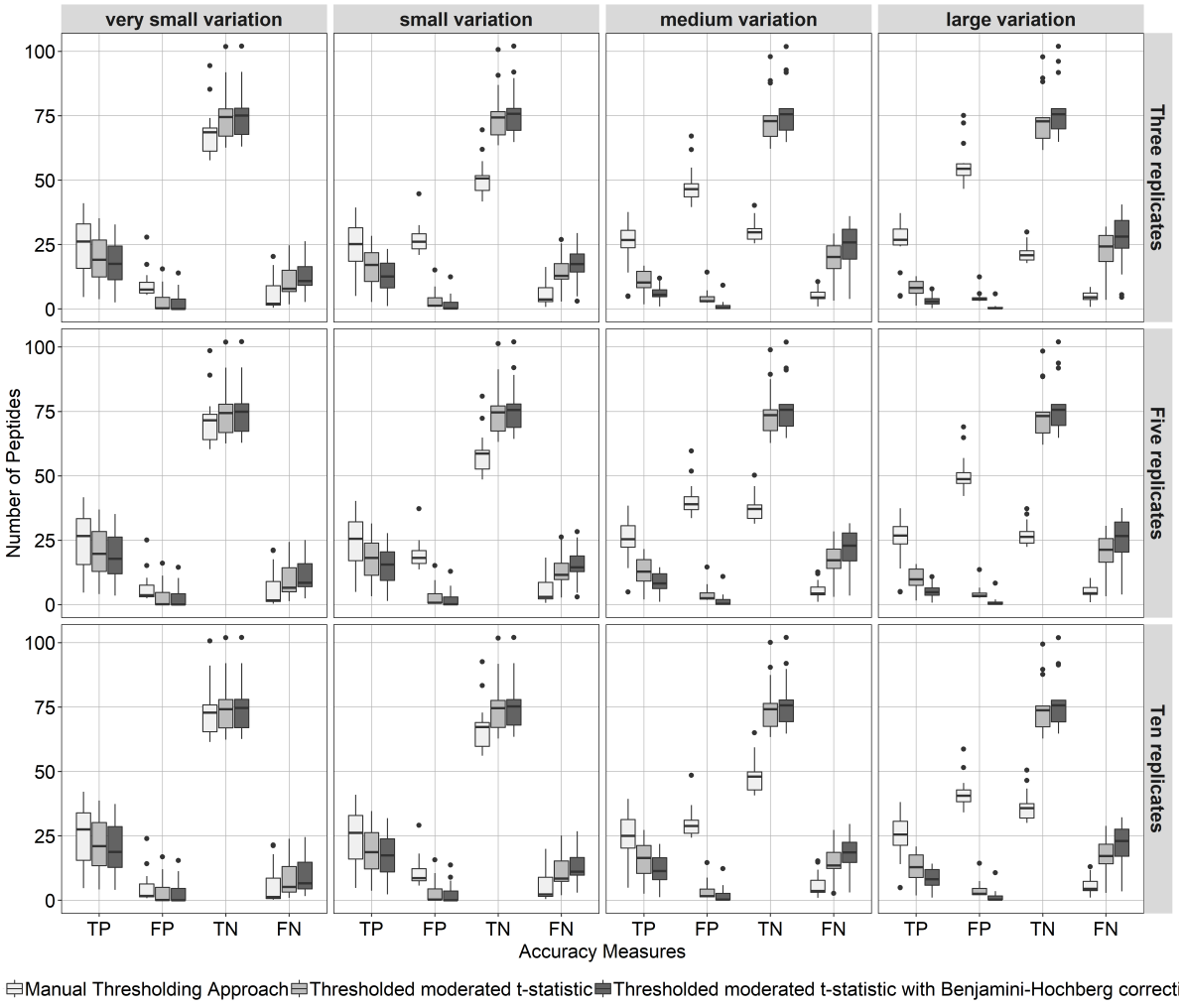
Similar results were found when using a significance level of 1% or 10% (results not shown). Additionally, we would like to point out that the magnitude of the problem with the under- and overestimation of the deuteration-variance of the 'global treshold'-approach is dependent on the distribution of the variances. With the uniform distribution, where each value in the specified interval is equally likely to occur, the magnitude of over- and underestimation is greater than when the variances follow a normal distribution. For the normal distribution, values close to the average have a higher probability than values at the tails of the distribution. The pooled residual standard deviation, as estimated by the 'global threshold'-approach, is close to the average of the normal distribution. As a result, the occurrence of under- and overestimated variances is limited, and as a consequence, the performance of the 'global threshold'-approach and the moderated t-statistic are comparable. On the other hand, if the variance follows a bimodal or multimodal distribution, or a distribution where the extremes are more likely than the centre of the distribution, the magnitude of over- and underestimation will be larger than shown here.

4.1.2 Thresholded pairwise comparisons

The thresholded moderated t-statistic and the 'manual threshold'-approach allow to specify a minimum difference that is considered biologically meaningful. While the thresholded moderated t-statistic tests if the observed difference in deuteration is larger than the specified threshold, the 'manual threshold'-approach only checks if the difference is larger or smaller than the threshold.

Here, we defined a threshold of 0.25Da to illustrate the performance of both methods.

Figure 2: Accuracy measures for thresholded pairwise comparisons at $\alpha = 0.05$.



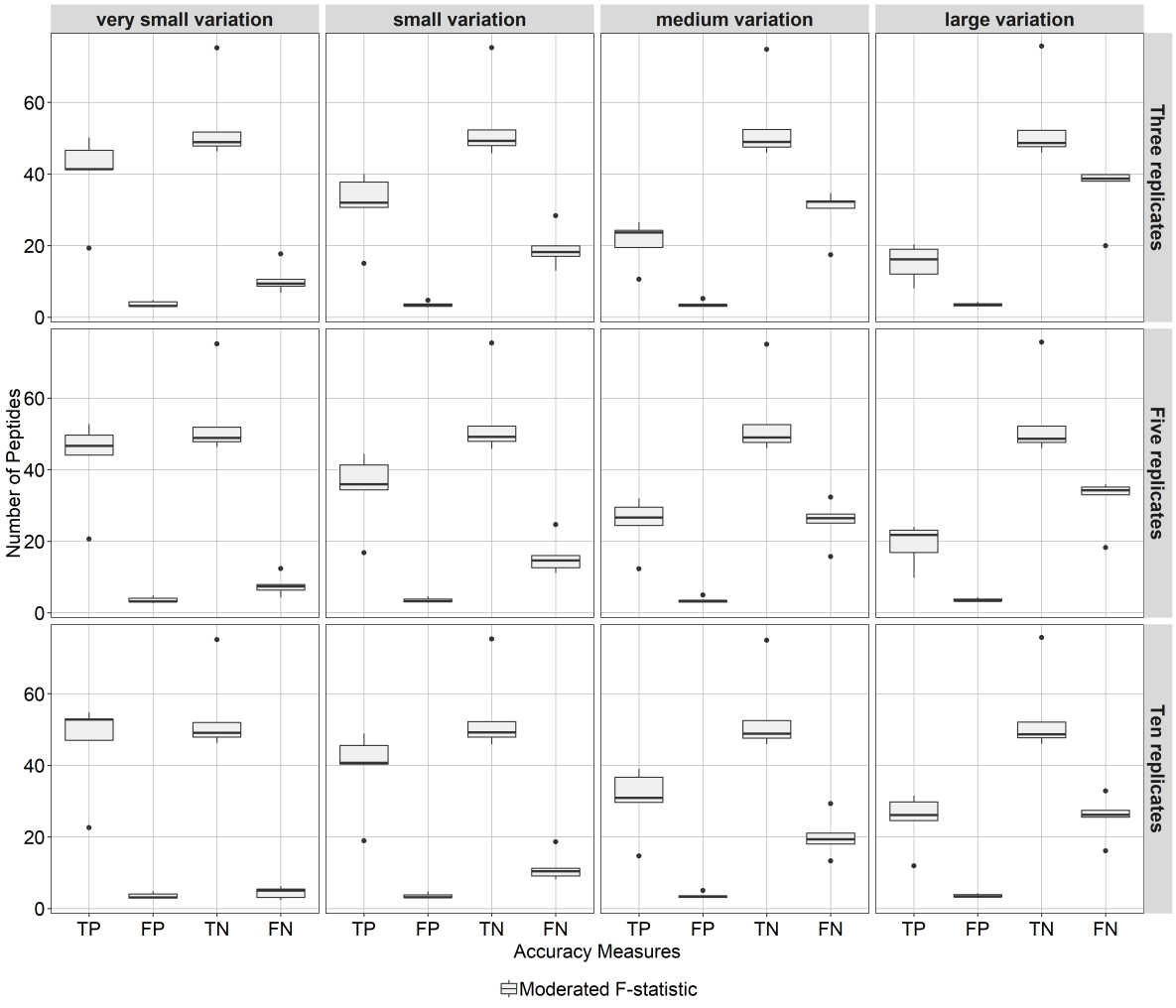
In contrast to pairwise comparisons, the number of TPs and TNs does not change when the number of replicates increases (Figure 2). The number of FNs decreases when the number of replicates increases. For the (corrected) thresholded moderated t-statistic, higher variability of D_i-values leads to a lower number of TPs while the number of FPs and TPs remains constant. In case of the 'manual threshold'-approach the opposite can be observed: the number of TPs does not change, but the number of FPs increases and the number of TNs decreases considerably with increasing levels of variability. The substantial increase of FPs is a consequence of ignoring the variance of the D_i-values. For the "very small variation" setting, the variances are too small to negatively

affect the performance of the 'manual threshold'-approach. In this setting, the performance of the 'manual threshold'-approach and the (corrected) thresholded moderated t-statistic are comparable.

4.1.3 Moderated F-statistic

The moderated F-statistic allows to detect differences between two or more groups or conditions. Similarly to the moderated test statistics for pairwise comparisons, the power of the F-statistic increases when the number of replicates increases, i.e., more TPs, and less FNs (Figure 3). With increasing variability, the number of TPs decreases and the number of FNs increases, while the number of FPs remains constant.

Figure 3: Accuracy measures for the moderated F-statistic at $\alpha = 0.05$.



4.2 Real-life data

4.2.1 Preprocessing

For each sample of the SecA-variants, the 'cluster' data-file from Waters DynamX software was converted to deuteration uptake values, D:

$$D_{i,t,s} = m/z_{i,t,s} \times z_{i,t,s} - m/z_{i,t_0,s} \times z_{i,t_0,s}$$

where $m/z_{i,t,s}$ and $z_{i,t,s}$ are the mass-to-charge and charge of peptide i at labeling time t in sample s, respectively. For peptides with multiple charges at labeling time t, the average D_t-value was calculated. Peptides which were not present in each sample and at each labeling time were removed.

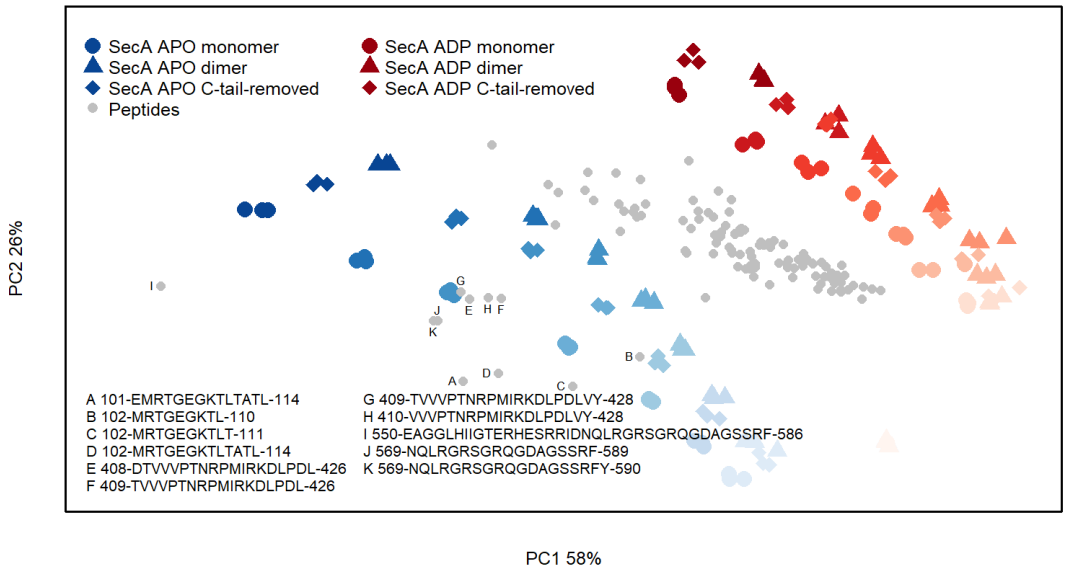
4.2.2 Exploratory data analysis

The deuteration-values of 112 peptides were inspected with variably weighted spectral map analysis [18, 19] (see Supporting Information), a multivariate method to identify clusters of samples and clusters of features, here, peptides (Figure 4). The three replicates of each variant of unbound SecA (SecA APO), and SecA in complex with ADP (SecA ADP), group closely together, indicating neglectible variability of the extracted deuteration-values within each variant. Additionally, all samples of SecA ADP (red) cluster together, and all samples of SecA APO (blue) cluster together. This suggests that the differences between SecA APO and SecA ADP are larger than, for instance, the differences between the monomeric and dimeric variants of SecA APO. Furthermore, the effect of the different labeling times can also be seen. Samples that are briefly exposed to deuterium, e.g. 10s , are located on the right hand side of the spectral map (light colors), while samples with longer exposure times (darker colors) gradually move to the center (SecA ADP) or to the left hand side (SecA APO). In case of SecA APO, the differences between the monomeric, dimeric and shortened variant get bigger when the labeling time increases.

The majority of the peptides (grey dots) can be found in the space between SecA APO and SecA

ADP samples, while eleven peptides are located to the left of the SecA APO samples. These peptides are part of the NBD1, NBD linker and NBD2 regions of SecA, which are crucial for ADP-binding. More specifically, the peptides map to helicase motif I (NBD1, 104-109), the NBD linker region (412-419), helicase motif Vb (NBD2, 556-558) and to helicase motif VI (NBD2, 570-577). Most likely, the differences in deuteration-values of these peptides are responsible for not clustering SecA APO and SecA ADP samples together.

Figure 4: Spectral map of the SecA APO and SecA ADP data. Each protein-state is represented by a series of points, i.e., one point per replicate and one point per labelling time. The color shading of the points correspond to the labelling time. The longer the labelling time, the darker the color. Each peptide is represented by one point (in grey).



4.2.3 Differential deuteration analysis

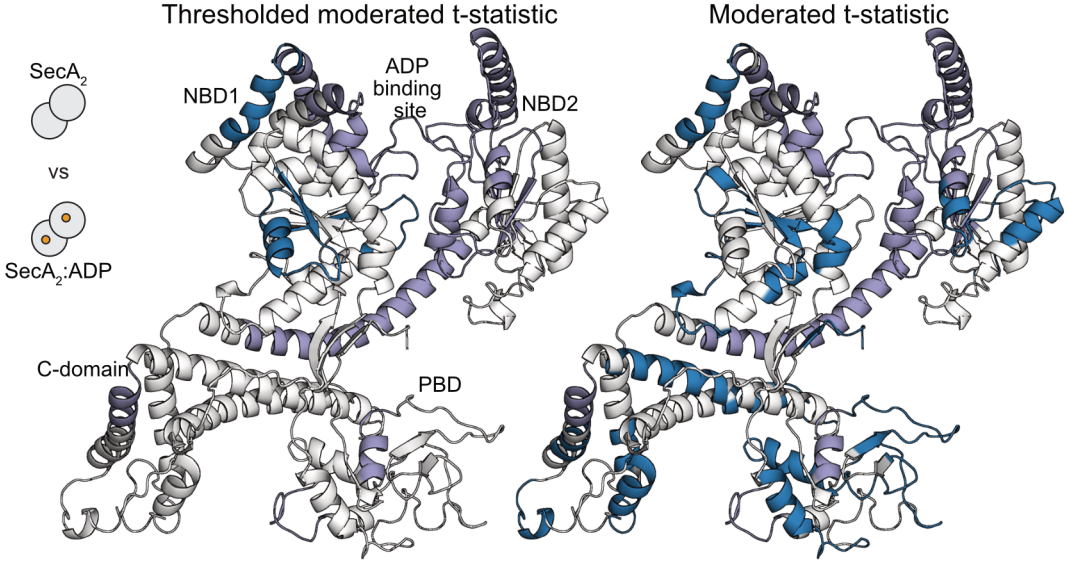
To formally test for differences in deuteration, we used the (thresholded) moderated t-statistic, the 'global thresholding'-approach and the hybrid significance test for pairwise comparisons and the moderated F-statistic to compare the three variants of SecA. In order to be considered statistically significant the p-values of the (thresholded) moderated t-statistic, adjusted with the Benjamini-Hochberg procedure for multiple testing [4], should be smaller than or equal to the 5% significance level. We also applied the "manual thresholding"-approach as used in [17], i.e, in order to be

considered a meaningful difference in deuteration, the difference has to be larger than 0.5Da and has to be at least 10% of the maximum deuteration-level of the peptide.

Differences between SecA APO and SecA ADP

ADP binding to the nucleotide cleft of SecA causes rigidification throughout the protein. As a consequence, many statistically significant differences in deuteration between SecA APO and SecA ADP for the monomeric, dimeric and C-tailed-removed variant were found (Figure S1). In comparison to the "manual thresholding"-approach, the 'global thresholding'-approach and the hybrid significance test, the moderated t-statistic and the thresholded moderated t-statistic ($\Delta = 0.50$) found more differentially deuterated peptides (Figure 5, Table S4). However, there are little or no differences in the number of differentially deuterated peptides between the global thresholding-approach and the hybrid significance test, indicating overestimation issues of the deuteration-variances.

Figure 5: Differentially deuterated peptides at 5mins labeling time when comparing SecA APO and SecA APD (PDB ID:2VDA). The regions in purple are found by the moderated test-statistics and the "manual thresholding"-approach, while the regions in blue are solely found by the thresholded moderated t-statistic (left) and/or by the moderated t-statistic (right).



The statistically significant differences in deuteration between SecA APO and SecA ADP were almost exclusively positive, i.e., larger deuteration-values for SecA APO (Figures S2 and S3). The largest differences in deuteration were found in the helicase motifs of the NBD1 and NBD2-regions

of SecA. These motifs are either directly involved in the binding of ADP or in conveying ADP-driven conformational changes to the rest of the protein [17]. The magnitude of these differences did not change with increasing labeling times, in contrast to peptides from the NBD1-NBD2-linker region. The differences in deuteration of these peptides became substantially larger with increasing labeling times. Allosteric effects of ADP binding became more prominent with increasing labeling times.

The differences in deuteration, identified by the moderated statistics as statistically significant and not by the "manual thresholding"-approach, are limited, but the majority of these differences were selected by the "manual thresholding"-approach at longer labeling times.

Differences between the monomeric, and dimeric and C-tail-removed variants

Next to the effect of ADP binding on the structure of SecA, we also investigated the differences between the monomeric, dimeric and C-tail-removed variant of SecA in the APO-state and in the ADP-bound state. For both states, the number of statistically significant differences, as indicated by the moderated F-statistic, (Figure S4) is smaller than the number of statistically significant differences when comparing the APO-state with the ADP-bound state (Figure S1). In case of the moderated t-statistic, corrected with the Benjamini-Hochberg procedure, more than 50% of the peptides are statistically significant when comparing the variants of the APO-state, while in the ADP-bound state, the number of statistically significant peptides is generally lower. Similarly to the comparisons between ADP and APO, no or litte differences between the number of differentially deuterated peptides as indicated by the 'global threshold'-approach and the by the hybrid significance test are seen (Table S5).

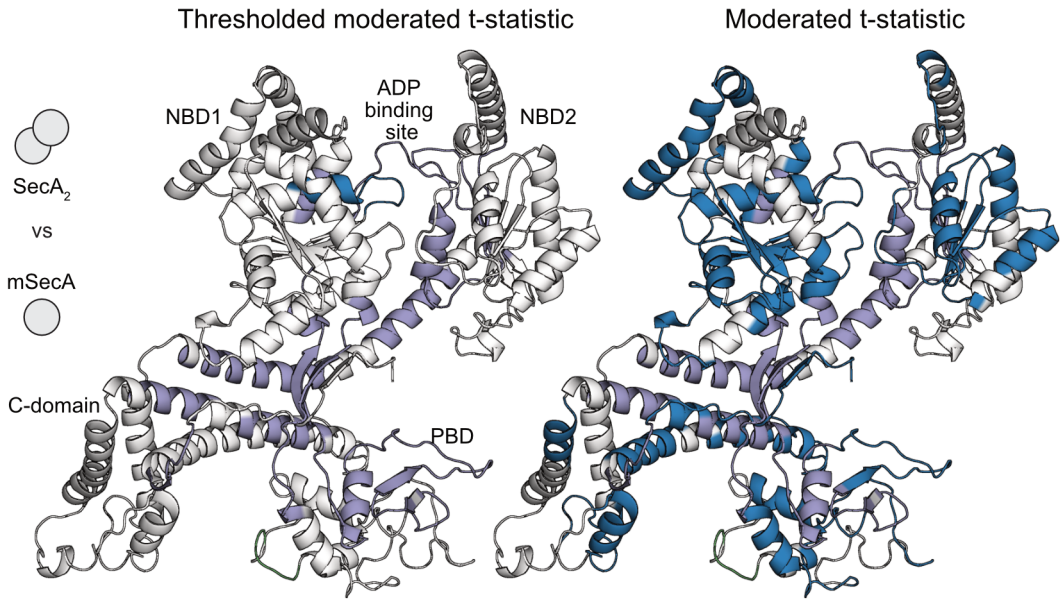
We focus here at the differences in deuteration between the monomeric and dimeric variant (Figure S5 and S6), and between the C-tail-removed and dimeric variant (Figure S7 and S8) identified by the (thresholded) moderated t-statistic.

When comparing the monomeric variant with the dimeric variant, the vast majority of the statistically significant peptides have positive differences in deuteration, indicating increased dynamics of the monomeric SecA variant. Many of these peptides are part of regions in SecA that are

involved in the dimer interface. A limited set of peptides, i.e., 286-LVKEGIMDEGESL-298, 285-LLVKEGIMDEGESL-298, 236-YKRVNKIIPHLIR-248, and 236-YKRVNKIIPHLIRQ-249, that are an integral part of the PBD-region of SecA are consistently statistically significant with more deuteration in the ADP-bound state. Although the magnitude of the statistically significant differences in the ADP-bound state is smaller than the differences in the APO state, a moderate overlap between the statistically significant peptides can be found between the APO and the ADP-bound state.

Similarly to the comparison between the bound and unbound SecA, all peptides identified by the "manual thresholding"-approach were found by the (thresholded) moderated t-statistics (Figure 6). The peptides that are found by the moderated test-statistics but not by the "manual thresholding"-approach are located in the NBD1- and NBD2-region of SecA.

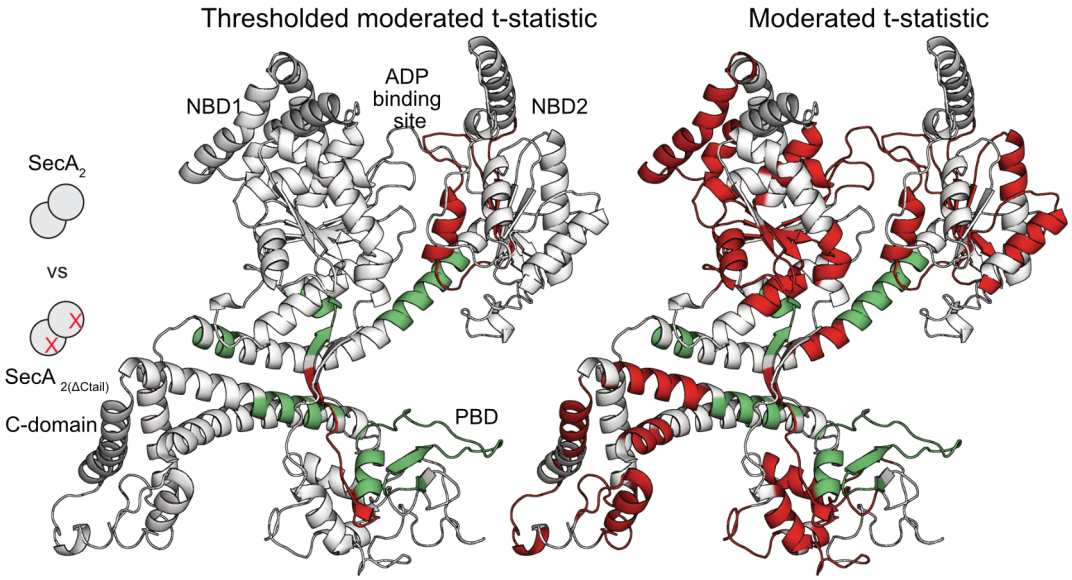
Figure 6: Differentially deuterated peptides at 5mins labeling time when comparing unbound dimeric SecA with unbound monomeric SecA (PDB ID:2VDA). The regions in purple (higher deuteration-levels in monomeric SecA) and green (higher deuteration-levels in dimeric SecA) are found by the moderated test-statistics and the "manual thresholding"-approach, while the regions in blue are solely found by the thresholded moderated t-statistic (left) and/or by the moderated t-statistic (right).



Comparison of the C-tail-removed variant and the dimeric variant of SecA shows that the C-tail-removed variant has, in general, higher deuteration-values than the dimeric variant (Figure S7 and S8). The differences in deuteration are small, indicating a minor increase in the dynamics of the C-tail-removed variant compared to dimeric SecA. Many differentially deuterated peptides were also found when comparing the monomeric variant with the dimeric variant of SecA. These two observations suggest that the removal of the C-tail has a similar but smaller effect on the structure and dynamics of SecA as switching from a dimeric state to a monomeric state.

Similarly to the other comparisons, bound versus unbound SecA and dimeric versus monomeric SecA, all peptides identified by the "manual thresholding"-approach were found by the (thresholded) moderated t-statistics (Figure 7).

Figure 7: Differentially deuterated peptides at 5mins labeling time when comparing unbound dimeric SecA (PDB ID:2VDA) with the C-tail removed dimeric variant (PDB ID: 1M6N). The regions in green are found by the moderated test-statistics and the "manual thresholding"-approach, while the regions in red are solely found by the thresholded moderated t-statistic (left) and/or by the moderated t-statistic (right).



5 Conclusion

The proposed moderated test statistics allow the identification of differentially deuterated peptides with ample power. They can be easily combined with statistical models, allowing the analysis of complex ΔHX-MS experiments. The moderated F-statistic can be used as a preliminary method to detect differentially deuterated peptides when comparing more than two groups. In contrast to the 'manual thresholding'-approach, the thresholded moderated t-statistic offers a statistical framework to test for biological relevant differences, and has, as a consequence, a smaller number of false positives. The moderated test statistics also allow to identify subtle differences in deuteration that are not picked up with 'manual thresholding'-approaches.

Although all test statistics allow to control the number of false positives, important differences between the moderated test statistics, the 'global threshold'-approach and the hybrid significance test can be seen. The 'global threshold'-approach suffers from under- and overestimation of the deuteration variances. As a result, in case of underestimation, the number of true positive findings is reduced or, in case of overestimation, more false positives are found. The hybrid significance test depends on the results of the global threshold approach. It can reduce the number of false positives when the deuteration-variance is underestimated, but it cannot increase the number of true positives when the variability of the deuteration-levels is overestimated. As a consequence, the true significance level of the global threshold approach and the hybrid significance test can be different from the specified significance level. The moderated test statistics can be combined with FWER- or FDR-controlling procedures, reducing the number of false positives but at the cost of having more false negatives.

Supporting Information

Additional details on methods, simulated data and results, including figures and tables, can be found in Supporting Information.

Acknowledgements

SK was a FWO [PEGASUS]^2 MSC fellow. This project has received funding from the Research Foundation – Flanders (FWO) and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreements No 665501

References

(1) Engen, J. *Anal. Chem.* **2009**, *81*, 7870–7875.

(2) Chalmers, M.; Busby, S.; Pascal, B.; West, G.; Griffin, P. *Expert Rev Proteomics* **2011**, *8*, 43–59.

(3) Dudoit, S.; van der Laan, M., *Multiple Testing Procedures with Applications to Genomics*; Springer Series in Statistics; Springer-Verlag New York: 2008.

(4) Benjamini, Y.; Hochberg, Y. *R Stat Soc Series B Stat Methodol.* **1995**, *57*, 289–300.

(5) Baldi, P.; Long, A. *Bioinformatics* **2001**, *17*, 509–519.

(6) Efron, B.; Tibshirani, R.; Storey, J.; Tusher, V. *J. Am. Stat. Assoc.* **2001**, *96*, 1151–1160.

(7) Lonnstedt, I.; Speed, T. *Stat. Sin.* **2002**, *12*, 31–46.

(8) Smyth, G. *Stat. Appl. Genet. Mol. Biol.* **2004**, *3*, Article 3.

(9) Hageman, T.; Weis, D. *Anal. Chem.* **2019**, *91*, 8008–8016.

(10) Houde, D.; Berkowitz, S.; Engen, J. *J. Pharm. Sci.* **2011**, *100*, 2071–2086.

(11) Weis, D. *J. Pharm. Sci.* **2018**, *108*, 807–810.

(12) Efron, B.; Morris, C. *J. Am. Stat. Assoc.* **1973**, *68*, 117–130.

(13) Morris, C. *J. Am. Stat. Assoc.* **1983**, *78*, 47–65.

(14) Casella, G. *Am. Stat.* **1985**, *39*, 83–87.

(15) McCarthy, D.; Smyth, G. *Bioinformatics* **2009**, *25*, 765–771.

(16) Puchala, W.; Burdukiewicz, M.; Kistowski, M.; Dabrowska, K.; Badaczewska-Dawid, A.; Cysewski, D.; Dadlez, M. *Bioinformatics* **2020**, *btaa587*.

(17) Krishnamurthy, S.; Eleftheriadis, N.; Karathanou, K.; Smit, J.; Portaliou, A.; Chatzi, K.; Karamanou, S.; Bondar, N.; Gouridis, G.; Economou, A. *Structure* **2021**, *29*, 846–858.e7.

(18) Lewi, P. *Arzneim.-Forsch.* **1976**, *26*, 1295–1300.

(19) Wouters, L.; Göhlmann, H.; Bijnens, L.; Kass, S.; Molenberghs, G.; Lewi, P. *Biometrics* **2003**, *59*, 1131–1139.

Figure 8: For Table of Contents Only