Bioinformatics, 36(26), 2020, i745–i753
doi: 10.1093/bioinformatics/btaa807
ECCB2020

Proteins

New mixture models for decoy-free false discovery rate estimation in mass spectrometry proteomics

Yisu Peng[1,†], Shantanu Jain[1,†], Yong Fuga Li[2], Michal Gregus̆[3,4], Alexander R. Ivanov[3,4], Olga Vitek[1,4] and Predrag Radivojac[1,3,4,*]

[1]Khoury College of Computer Sciences, Northeastern University, Boston, MA 02115, USA, [2]Illumina Inc., San Diego, CA 92122, USA, [3]Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA 02115, USA and [4]Barnett Institute of Chemical and Biological Analysis, Northeastern University, Boston, MA 02115, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Abstract

Motivation: Accurate estimation of false discovery rate (FDR) of spectral identification is a central problem in mass spectrometry-based proteomics. Over the past two decades, target-decoy approaches (TDAs) and decoy-free approaches (DFAs) have been widely used to estimate FDR. TDAs use a database of decoy species to faithfully model score distributions of incorrect peptide-spectrum matches (PSMs). DFAs, on the other hand, fit two-component mixture models to learn the parameters of correct and incorrect PSM score distributions. While conceptually straightforward, both approaches lead to problems in practice, particularly in experiments that push instrumentation to the limit and generate low fragmentation-efficiency and low signal-to-noise-ratio spectra.

Results: We introduce a new decoy-free framework for FDR estimation that generalizes present DFAs while exploiting more search data in a manner similar to TDAs. Our approach relies on multi-component mixtures, in which score distributions corresponding to the correct PSMs, best incorrect PSMs and second-best incorrect PSMs are modeled by the skew normal family. We derive EM algorithms to estimate parameters of these distributions from the scores of best and second-best PSMs associated with each experimental spectrum. We evaluate our models on multiple proteomics datasets and a HeLa cell digest case study consisting of more than a million spectra in total. We provide evidence of improved performance over existing DFAs and improved stability and speed over TDAs without any performance degradation. We propose that the new strategy has the potential to extend beyond peptide identification and reduce the need for TDA on all analytical platforms.

Availabilityand implementation: https://github.com/shawn-peng/FDR-estimation.

Contact: predrag@northeastern.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A typical bottom-up proteomics pipeline consists of several experimental and computational steps, combined to interrogate the presence, quantity, form and function of proteins in the biological mixture (Aebersold and Mann, 2003; Choudhary and Mann, 2010; Gingras *et al.*, 2007; Steen and Mann, 2004). Central to all these challenges is the task of accurately establishing the presence of peptide species in the sample (Hubler *et al.*, 2020; Kall *et al.*, 2008b), a step that relies on computational and statistical techniques to map spectra from the mass spectrometer to peptide sequences and assign confidence scores to the resulting peptide-spectrum matches (PSMs). Peptide identification is often performed via a search algorithm, where experimental spectra are scored against the theoretical spectra derived from a selected group of candidate peptides (Kim and Pevzner, 2014; Kong *et al.*, 2017; Perkins *et al.*, 1999; Tabb *et al.*, 2007; Yates *et al.*, 1995) or *de novo*, when restricting the set of candidate peptides is problematic (Dancik *et al.*, 1999; Frank and Pevzner, 2005).

Despite methodological variability in practice, the core of any peptide identification protocol is the scoring of PSMs that is intended to reflect their likelihood of being correct assignments (Hubler *et al.*, 2020; Li *et al.*, 2012). These schemes must meet both local and global requirements in that the ranking of PSMs for a given experimental spectrum must prioritize the most likely peptide assignments and that the scoring of those top-ranked PSMs over all experimental spectra must be calibrated so that the global ranking of top-ranked PSMs is meaningful (Keich and Noble, 2015). Well-performing search engines generally meet these requirements, in which case the set of identified or accepted PSMs can be reliably determined from the ranked list of top-scoring PSMs based on a score threshold. The list of identified PSMs ideally contains a large fraction of correct identifications (spectra matched to peptides they originated from) and not more than a small fraction of incorrect identifications (spectra matched to peptides they did not originate from).

False discovery rate (FDR) is defined as the expected proportion of incorrect identifications among reported identifications (Burger,

2018; Choi and Nesvizhskii, 2008; Storey, 2002). Over the past two decades, two major approaches for estimating FDR have emerged; i.e. target-decoy approaches (TDAs) and decoy-free approaches (DFAs). Target-decoy techniques search both the set of peptides possibly present in the sample (target database) and a set of peptides that are not in the sample (decoy database), where the role of the decoy database is to faithfully model the score distribution of incorrect top-scoring PSMs from the target database and thus facilitate FDR estimation (Elias and Gygi, 2007). TDAs differ in the construction of decoy sequences and search strategies such as separately or combined with target sequences (Jeong *et al.*, 2012). Decoy-free techniques, on the other hand, search only the target database and fit a generative two-component model to the set of scores corresponding to all top-scoring PSMs. The two-component model the correct and incorrect score distributions, typically using some combination of Gaussian, Gumbel and Gamma distributions. For example, Keller *et al.* (2002) model the score distribution of the correct top PSMs using a Gaussian distribution and incorrect top PSMs using a Gamma distribution. An expectation-maximization (EM) algorithm is applied to estimate the parameters of these distributions (Dempster *et al.*, 1977).

Each search strategy comes with pros and cons. Owing to its simplicity, TDA with a concatenated database search has dominated bottom-up proteomics, even if the benefits of competing decoy peptides with target peptides for experimental spectra are incompletely understood. In fact, the usefulness of TDA has been continuously challenged on several grounds (Cooper, 2011, 2012; Danilova *et al.*, 2019; Gupta *et al.*, 2011; Kall *et al.*, 2008a; Kim *et al.*, 2008), including the construction of decoy sequences, choice of FDR estimators and run time. Current practices generally rely on peptide reversal within each protein to construct decoys, based on empirical characterizations against the alternatives (Elias and Gygi, 2007). TDAs estimate FDR as the fraction of the number of decoy top PSMs and the number of target top PSMs above the threshold. While this approach is reasonable with large datasets, it is theoretically problematic as it can lead to FDR estimates above 1 and possibly even infinity. TDAs also consider protein databases twice in size, which can be computationally expensive for identifying post-translationally modified peptides or cross-linked peptides (Ji *et al.*, 2016; Rinner *et al.*, 2008). On the other hand, DFAs are not without problems either. While theoretically pleasing, these methods suffer from restrictive modeling assumptions as well as difficulties in resolving overlapping score distributions, especially when the fraction of correct PSMs is small (Ma *et al.*, 2012). They also lead to inconsistencies, such as ones where Gaussian-Gamma distributions give best fits on average yet the component densities have different supports and can lead to pathological situations; e.g. low-scoring PSMs might have a probability of 1 to be correct (Li, 2008). This is particularly problematic in experiments where distinguishing correct and incorrect PSMs is challenging.

The objective of this study is to introduce and explore new decoy-free FDR estimation procedures that combine the strengths of TDAs and DFAs. Specifically, we consider a two-sample approach, where the top or best-scoring PSMs are used in a manner similar to conventional DFA searches, and the second-best PSMs, much like decoy PSMs, are used to improve modeling of the incorrect top PSMs. We model the set of component densities using a relatively new family of skew normal distributions that offer desirable flexibility within the unimodal family yet provide elegant update rules for an EM-based optimization. We evaluate the new systems against both TDAs and DFAs on NIST spectral libraries from four species, ten additional PRIDE datasets from six species as well as an in-house case study using nanogram levels of total HeLa cell digest to demonstrate the potential for applications in high-sensitivity proteomics profiling. We demonstrate that leveraging the extra search information increases the accuracy and the stability of estimates, in particular in experiments where low amounts of biological material limit the quality and the number of spectra (Budnik *et al.*, 2018; Li *et al.*, 2015). Overall, we believe that the new algorithms have a potential to generalize beyond peptide identification to all types of search problems involving analytical platforms.

2 Background

2.1 Terminology and notation

Let $\mathcal{X} = \{x_i\}$ be a set of spectra collected from a mass spectrometer and $\mathcal{P} = \{p_j\}$ a set of candidate peptides that are possibly present in the biological sample. A search engine produces a set of triplets $(x, p, s) \in \mathcal{X} \times \mathcal{P} \times \mathbb{R}$, where s is the score assigned to the PSM (x, p). The higher the score, the more likely that the spectrum x was generated from p.

Let now x be generated from some (unknown) peptide q and let $((x, p_1, s_1), (x, p_2, s_2), \ldots)$ be a ranked list of PSMs from a search engine for x such that $s_1 \geq s_2 \geq \ldots$ A PSM (x, p) for which $p = q$ is called the *correct match*, whereas all other PSMs involving x are called *incorrect matches*. Furthermore, given the list $((x, p_1, s_1), (x, p_2, s_2), \ldots)$, the PSM with the highest score, (x, p_1), is called the top, first or best-scoring PSM, the second-ranked PSM, (x, p_2), is called the second PSM, etc. Finally, we also distinguish among incorrect PSMs. The highest-scoring incorrect PSM for x will be referred to as the top, first or best incorrect PSM, whereas the second-best incorrect PSM will be referred to as the second incorrect PSM.

To reduce complexity, an MS/MS analysis pipeline often keeps only top PSMs for the set of spectra $\mathcal{X}$; i.e. only the top-scoring PSM for each spectrum x. It then determines a threshold τ such that the peptide p from each top hit (x, p) is considered *identified* when the score s from (x, p, s) satisfies $s \geq \tau$. If, further, $p = q$, p is considered to be the correct identification. The threshold τ can be set based on experience with particular search engines although the most rigorous approach is to estimate FDR for the set of identified peptides obtained by thresholding at τ. Current approaches restrict the analysis to top-scoring PSMs for each experimental spectrum. In this study, we remove this restriction and include both top PSMs and second-best PSMs to more confidently model the data distributions.

2.2 Skew normal family

The Gaussian family is widely used in many applications to model real-world data. However, the symmetry of the Gaussian density makes it an inferior choice for modeling skewed data. One approach to account for the skewness is to use a mixture of Gaussian distributions; however, finite Gaussian mixtures are ill-equipped to model the skewness, especially when the data is expected to be unimodal (Jain *et al.*, 2019). In such cases one may choose from one of the many skewed families such as Gumbel, Gamma, Weibull and skew normal. The use of Gumbel and Gamma distributions in the context of FDR estimation has been extensively studied (Li, 2008). In this article, we explore the appropriateness of the skew normal family for FDR estimation. Skew normal family is an appealing choice for modeling competition since the density of the maximum of two identically distributed Gaussian random variables is exactly skew normal (Arellano-Valle *et al.*, 2006).

The univariate skew normal (SN) family was introduced as a generalization of the normal family (Azzalini, 1985). It has a location (μ), a scale (ω) and a shape (λ) parameter, where λ controls the direction and degree of skewness. The distribution is right-skewed when $\lambda > 0$, left-skewed when $\lambda < 0$ and reduces to a normal distribution when $\lambda = 0$. The probability density function (pdf) of a random variable $X \sim \mathrm{SN}(\mu, \omega, \lambda)$ is given by

$$f_{\mathrm{SN}}(x; \mu, \omega, \lambda) = \frac{2}{\omega}\phi\left(\frac{x-\mu}{\omega}\right)\Phi\left(\frac{\lambda(x-\mu)}{\omega}\right), \; x \in \mathbb{R},$$

where $\mu, \lambda \in \mathbb{R}$, $\omega \in \mathbb{R}^+$, ϕ and Φ are the pdf and the cumulative distribution function (cdf) of the standard normal distribution $\mathrm{N}(0, 1)$, respectively. The cumulative distribution function of X is given by

$$F_{\mathrm{SN}}(x; \mu, \omega, \lambda) = \Phi\left(\frac{x-\mu}{\omega}\right) - 2\mathcal{T}\left(\frac{x-\mu}{\omega}, \lambda\right), \; x \in \mathbb{R},$$

where $\mathcal{T}(h, a)$ is Owen's T function (Young and Minder, 1974). The SN family can be alternatively parameterized by Δ and Γ instead of

λ and ω, as defined in Table 1. The alternate parametrization naturally arises in the stochastic representation of a SN random variable:

$$X \sim \mathrm{SN}(\mu, \omega, \lambda) \;\Rightarrow\; X \overset{d}{=} \mu + \Delta T + \Gamma^{1/2} U, \tag{1}$$

where $T \sim \mathrm{TN}(0, 1, \mathbb{R}_+)$, the standard normal distribution truncated below 0; $U \sim \mathrm{N}(0, 1)$, the standard normal distribution; and $\overset{d}{=}$ reads as 'equal in distribution'. The stochastic representation is useful for deriving many properties of the skew normal distribution and is also used in an EM-based maximum likelihood estimation (Lin *et al.*, 2007). The algorithms for the skew normal mixture models derived in this article also exploit this stochastic representation.

3 Materials and methods

In this section, we introduce two generative models and derive corresponding EM algorithms for parameter estimation. Let $\mathbb{S}_1$ denote the set of the first scores and $\mathbb{S}_2$ denote the set of the second scores of a tandem mass spectrometry (MS/MS) search. The first model relies solely on the score distributions of the top PSMs and thus only $\mathbb{S}_1$ is used for parameter estimation. The second model is an extension when first and second PSMs are both considered and uses $\mathbb{S}_1$ and $\mathbb{S}_2$ to estimate the parameters. The dataset sizes $|\mathbb{S}_1|$ and $|\mathbb{S}_2|$ need not be equal.

We assume in both models that the scores corresponding to a correct match and all incorrect matches follow skew normal distributions. Technically, we introduce C, I_1 and I_2 to denote the random variables corresponding to the scores of the correct match, the first incorrect match and the second incorrect match, respectively, as

$$C \sim \mathrm{SN}(\theta_c) \qquad I_1 \sim \mathrm{SN}(\theta_1), \qquad I_2 \sim \mathrm{SN}(\theta_2), \tag{2}$$

where θ denotes the skew normal parameters μ, ω and λ.

Sections 3.1 and 3.2 present only update rules of the proposed EM algorithms. We direct the reader to Supplementary Materials for additional details. Specifically, Supplementary Section S2 of Supplementary Materials shows the derivation of the algorithms and Supplementary Section S1 gives proofs of the supporting lemmas.

3.1 Top score skew normal mixture

The top score skew normal mixture, referred to as 1SMix model, is the conventional decoy-free model in which both component distributions are in the skew normal family. More formally, we model the first score S_1 as a mixture of the correct and first incorrect scores, each being a skew normal random variable; i.e.

$$S_1 \sim \alpha \mathrm{SN}(\theta_c) + (1 - \alpha)\mathrm{SN}(\theta_1).$$

The triple $\zeta = (\alpha, \theta_c, \theta_1)$ gives the parameters of the model. We obtain the maximum likelihood estimates of ζ from $\mathbb{S}_1$ using the EM algorithm for finite skew normal mixture estimation in Lin *et al.* (2007). For completeness, we give a derivation of the algorithm for the two-component mixture case in Supplementary Section S3. Using $\ddot{}$ and $\bar{}$ to accent the new and old parameters, respectively, the parameter update equations of the EM algorithm are as follows:

Table 1. Alternate parametrization for the skew normal distribution

Alternate parametrization		Related quantities
Canonical → alternate	Alternate → canonical	
$\Delta = \omega\delta$	$\lambda = \mathrm{sign}(\Delta)\sqrt{\Delta^2/\Gamma}$	$\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$
$\Gamma = \omega^2 - \Delta^2$	$\omega = \sqrt{\Gamma + \Delta^2}$	

Note: Update equations of the algorithm are better formulated in terms of the alternate parameters. The table gives the relationship between the alternate and the canonical parameters as well as additional related quantities.

$$\ddot{\alpha} = \frac{1}{|\mathbb{S}_1|}\sum_{s_1 \in \mathbb{S}_1} \bar{p}_1(s_1),$$

$$\ddot{\mu}_c = \frac{\sum_{s_1 \in \mathbb{S}_1} \bar{p}_c(s_1)\bar{m}_c(s_1, \bar{\Delta}_c)}{\sum_{s_1 \in \mathbb{S}_1} \bar{p}_c(s_1)},$$

$$\ddot{\mu}_1 = \frac{\sum_{s_1 \in \mathbb{S}_1} \bar{p}_1(s_1)\bar{m}_1(s_1, \bar{\Delta}_1)}{\sum_{s_1 \in \mathbb{S}_1} \bar{p}_1(s_1)},$$

$$\ddot{\Delta}_c = \frac{\sum_{s_1 \in \mathbb{S}_1} \bar{p}_c(s_1)\bar{d}_c(s_1, \ddot{\mu}_c)}{\sum_{s_1 \in \mathbb{S}_1} \bar{p}_c(s_1)\bar{w}(s_1, \theta_c)},$$

$$\ddot{\Delta}_1 = \frac{\sum_{s_1 \in \mathbb{S}_1} \bar{p}_1(s_1)\bar{d}_1(s_1, \ddot{\mu}_1)}{\sum_{s_1 \in \mathbb{S}_1} \bar{p}_1(s_1)\bar{w}(s_1, \theta_1)},$$

$$\ddot{\Gamma}_c = \frac{\sum_{s_1 \in \mathbb{S}_1} \bar{p}_c(s_1)\bar{g}_c(s_1, \ddot{\mu}_c, \ddot{\Delta}_c)}{\sum_{s_1 \in \mathbb{S}_1} \bar{p}_c(s_1)},$$

$$\ddot{\Gamma}_1 = \frac{\sum_{s_1 \in \mathbb{S}_1} \bar{p}_1(s_1)\bar{g}_1(s_1, \ddot{\mu}_1, \ddot{\Delta}_1)}{\sum_{s_1 \in \mathbb{S}_1} \bar{p}_1(s_1)},$$

where $\bar{m}_*$, $\bar{d}_*$, $\bar{g}_*$ and $\bar{w}_*$ ($* = c$ or 1) are as defined in Table 2. Quantities $\bar{p}_C$ and $\bar{p}_1$ are defined as

$$\begin{aligned} \bar{p}_c(s_1) &= \frac{\bar{\alpha} f_{\mathrm{SN}}(s_1; \bar{\theta}_c)}{\bar{\alpha} f_{\mathrm{SN}}(s_1; \bar{\theta}_c) + (1 - \bar{\alpha}) f_{\mathrm{SN}}(s_1; \bar{\theta}_1)}, \\ \bar{p}_1(s_1) &= \frac{(1 - \bar{\alpha}) f_{\mathrm{SN}}(s_1; \bar{\theta}_1)}{\bar{\alpha} f_{\mathrm{SN}}(s_1; \bar{\theta}_c) + (1 - \bar{\alpha}) f_{\mathrm{SN}}(s_1; \bar{\theta}_1)}. \end{aligned} \tag{3}$$

The algorithm stops when the log-likelihood (Supplementary Materials) difference per data point falls under 10^{-8}. FDR at a threshold value τ is thereafter estimated as

$$\begin{aligned} \mathrm{FDR}(\tau) &= \frac{(1 - \alpha)p(I_1 > \tau)}{p(S_1 > \tau)} \\ &\overset{est}{=} \frac{(1 - \alpha)(1 - F_{\mathrm{SN}}(\tau; \theta_1))}{\alpha(1 - F_{\mathrm{SN}}(\tau; \theta_c)) + (1 - \alpha)(1 - F_{\mathrm{SN}}(\tau; \theta_1))}. \end{aligned} \tag{4}$$

To practically compute $F_{\mathrm{SN}}(\tau; \theta)$, we use an approximation of Owen's T function by Young and Minder (1974).

3.1.1 Parameter initialization

The initial parameters for the EM algorithm are estimated by partitioning the data and using the method of moments estimators for SN distributions (Supplementary Materials). Precisely, $\mathbb{S}_1$ is first partitioned into two sets separated by its median. The points below the median are then used to obtain a method of moments estimator of θ_1 and the points above the median are used for θ_c. Empirically, we observed that the signs of Δ_1 and Δ_c do not change during the execution of the algorithm. To ensure that the entire parameter space is searched for an optimal fit, we run the algorithm four times covering all possible combinations of signs of Δ_1 and Δ_c, with the best fit chosen according to the value of the likelihood function. Parameter α is initialized at 0.5.

3.2 Top-two score skew normal mixture

In the top-two score approach, referred to as 2SMix model, we model both first and second PSM score distributions as skew normal mixtures. Since the second score, S_2, can come from the correct, first incorrect or second incorrect match, we model its density as a three-component mixture. The complete model is specified as follows.

$$\begin{aligned} S_1 &\sim \alpha \mathrm{SN}(\theta_c) + (1 - \alpha)\mathrm{SN}(\theta_1), \\ S_2 &\sim \alpha \mathrm{SN}(\theta_1) + (1 - \alpha - \beta)\mathrm{SN}(\theta_2) + \beta \mathrm{SN}(\theta_c), \end{aligned}$$

where $\alpha, \beta \in [0, 1]$ and $\alpha + \beta \leq 1$. The quintuple $\zeta = (\alpha, \beta, \theta_c, \theta_1, \theta_2)$ gives the parameters of the model. Observe that the two mixtures are tied via a shared parameter α because the fraction of the first incorrect PSMs in $\mathbb{S}_2$ must be identical to the fraction of correct PSMs in $\mathbb{S}_1$. The fractions of correct PSMs in $\mathbb{S}_1$ and $\mathbb{S}_2$ are further

Table 2. Useful quantities

Quantities
$\overline{m}_*(x,\Delta) = x - v(x,\overline{\theta}_*)\Delta$
$\overline{d}_*(x,\mu) = v(x,\overline{\theta}_*)(x-\mu)$
$\overline{g}_*(x,\mu,\Delta) = (x-\mu)^2 - 2\Delta v(x,\overline{\theta}_*)(x-\mu) + \Delta^2 w(x,\overline{\theta}_*)$
$v(x,\theta) = \mathbb{E}[T_x]$
$w(x,\theta) = \mathbb{E}[T_x^2]$
$T_x \sim \mathrm{TN}(\delta/\omega(x-\mu), 1-\delta^2, \mathbb{R}^+)$

Note: The parameter update equations are given in terms quantities defined below. The quantities accented with $\bar{}$ have $\overline{\zeta}$, the current estimate of the model parameters, as an implicit parameter. $\overline{\zeta}$ contains all the model parameters: α and/or β and the parameters for the skew normal components, θ_*; depending upon the model, * can take values c, 1 and 2. θ contains skew normal parameters μ, ω and λ. Parameters δ, Δ and Γ are related to ω and λ as per Table 1. $\mathrm{TN}(\mu, \sigma^2, \mathbb{R}^+)$ represents truncated normal distribution truncated below 0. $\mathbb{E}$ represents the expectation operator. The expectations of the first two moments of the TN random variable can be computed as shown in Supplementary Lemma S1 in Supplementary Materials.

restricted by the fact that the total number of correct PSMs cannot exceed the sample size; i.e. $\alpha + \beta \leq 1$.

Unlike the top score only model, the parameters for the two score model cannot be obtained by using the existing skew normal mixture estimation methods because of parameter sharing between the two mixtures. We derive a novel EM algorithm for the maximum likelihood estimation of ζ from $\mathbb{S}_1$ and $\mathbb{S}_2$. Using $\ddot{}$ and $\bar{}$ to accent the new and old parameter, respectively, the parameter update equations of the EM algorithm are as follows.

$$\ddot{\alpha} = \frac{\sum_{s_1\in\mathbb{S}_1}\overline{p}_c(s_1) + \sum_{s_2\in\mathbb{S}_2}\overline{r}_1(s_2)}{|\mathbb{S}_1| + |\mathbb{S}_2|},$$

$$\ddot{\beta} = \frac{\sum_{s_2\in\mathbb{S}_1}\overline{r}_c(s_2)}{|\mathbb{S}_2|},$$

$$\ddot{\mu}_c = \frac{\sum_{s_1\in\mathbb{S}_1}\overline{p}_c(s_1)\overline{m}_c(s_1,\overline{\Delta}_c) + \sum_{s_2\in\mathbb{S}_2}\overline{r}_c(s_2)\overline{m}_c(s_2,\overline{\Delta}_c)}{\sum_{s_1\in\mathbb{S}_1}\overline{p}_c(s_1) + \sum_{s_2\in\mathbb{S}_2}\overline{r}_c(s_2)},$$

$$\ddot{\mu}_1 = \frac{\sum_{s_1\in\mathbb{S}_1}\overline{p}_1(s_1)\overline{m}_1(s_1,\overline{\Delta}_1) + \sum_{s_2\in\mathbb{S}_2}\overline{r}_1(s_2)\overline{m}_1(s_2,\overline{\Delta}_1)}{\sum_{s_1\in\mathbb{S}_1}\overline{p}_1(s_1) + \sum_{s_2\in\mathbb{S}_2}\overline{r}_1(s_2)},$$

$$\ddot{\mu}_2 = \frac{\sum_{s_2\in\mathbb{S}_2}\overline{r}_2(s_2)\overline{m}_2(s_2,\overline{\Delta}_2)}{\sum_{s_2\in\mathbb{S}_2}\overline{r}_2(s_2)},$$

$$\ddot{\Delta}_c = \frac{\sum_{s_1\in\mathbb{S}_1}\overline{p}_c(s_1)\overline{d}_c(s_1,\ddot{\mu}_c) + \sum_{s_2\in\mathbb{S}_2}\overline{r}_c(s_2)\overline{d}_c(s_2,\ddot{\mu}_c)}{\sum_{s_1\in\mathbb{S}_1}\overline{p}_c(s_1)\overline{w}(s_1,\theta_c) + \sum_{s_2\in\mathbb{S}_2}\overline{r}_c(s_2)\overline{w}(s_2,\theta_c)},$$

$$\ddot{\Delta}_1 = \frac{\sum_{s_1\in\mathbb{S}_1}\overline{p}_1(s_1)\overline{d}_1(s_1,\ddot{\mu}_1) + \sum_{s_2\in\mathbb{S}_2}\overline{r}_1(s_2)\overline{d}_1(s_2,\ddot{\mu}_1)}{\sum_{s_1\in\mathbb{S}_1}\overline{p}_1(s_1)\overline{w}(s_1,\theta_1) + \sum_{s_2\in\mathbb{S}_2}\overline{r}_1(s_2)\overline{w}(s_2,\theta_1)},$$

$$\ddot{\Delta}_2 = \frac{\sum_{s_2\in\mathbb{S}_2}\overline{r}_2(s_2)\overline{d}_2(s_2,\ddot{\mu}_2)}{\sum_{s_2\in\mathbb{S}_2}\overline{r}_2(s_2)\overline{w}(s_2,\theta_2)},$$

$$\ddot{\Gamma}_c = \frac{\sum_{s_1\in\mathbb{S}_1}\overline{p}_c(s_1)\overline{g}_c(s_1,\ddot{\mu}_c,\ddot{\Delta}_c) + \sum_{s_2\in\mathbb{S}_2}\overline{r}_c(s_2)\overline{g}_c(s_2,\ddot{\mu}_c,\ddot{\Delta}_c)}{\sum_{s_1\in\mathbb{S}_1}\overline{p}_c(s_1) + \sum_{s_2\in\mathbb{S}_2}\overline{r}_c(s_2)},$$

$$\ddot{\Gamma}_1 = \frac{\sum_{s_1\in\mathbb{S}_1}\overline{p}_1(s_1)\overline{g}_1(s_1,\ddot{\mu}_1,\ddot{\Delta}_1) + \sum_{s_2\in\mathbb{S}_2}\overline{r}_1(s_2)\overline{g}_1(s_2,\ddot{\mu}_1,\ddot{\Delta}_1)}{\sum_{s_1\in\mathbb{S}_1}\overline{p}_1(s_1) + \sum_{s_2\in\mathbb{S}_2}\overline{r}_1(s_2)},$$

$$\ddot{\Gamma}_2 = \frac{\sum_{s_2\in\mathbb{S}_2}\overline{r}_2(s_2)\overline{g}_2(s_2,\ddot{\mu}_2,\ddot{\Delta}_2)}{\sum_{s_2\in\mathbb{S}_2}\overline{r}_2(s_2)},$$

where quantities $\overline{m}_*, \overline{d}_*, \overline{g}_*$ and $\overline{w}$ ($* = c, 1$ or 2) are as defined in Table 2; $\overline{p}_c, \overline{p}_1$ are the same as those defined in Equation 3 and $\overline{r}_c, \overline{r}_1$ and $\overline{r}_2$ are as defined below.

$$\overline{r}_c(s_2) = \frac{\overline{\beta} f_{\mathrm{SN}}(s_2;\overline{\theta}_c)}{\overline{\alpha} f_{\mathrm{SN}}(s_2;\overline{\theta}_1) + (1-\overline{\alpha}-\overline{\beta}) f_{\mathrm{SN}}(s_2;\overline{\theta}_2) + \overline{\beta} f_{\mathrm{SN}}(s_2;\overline{\theta}_c)},$$

$$\overline{r}_1(s_2) = \frac{\overline{\alpha} f_{\mathrm{SN}}(s_2;\overline{\theta}_1)}{\overline{\alpha} f_{\mathrm{SN}}(s_2;\overline{\theta}_1) + (1-\overline{\alpha}-\overline{\beta}) f_{\mathrm{SN}}(s_2;\overline{\theta}_2) + \overline{\beta} f_{\mathrm{SN}}(s_2;\overline{\theta}_c)},$$

$$\overline{r}_2(s_2) = \frac{(1-\overline{\alpha}-\overline{\beta}) f_{\mathrm{SN}}(s_2;\overline{\theta}_2)}{\overline{\alpha} f_{\mathrm{SN}}(s_2;\overline{\theta}_1) + (1-\overline{\alpha}-\overline{\beta}) f_{\mathrm{SN}}(s_2;\overline{\theta}_2) + \overline{\beta} f_{\mathrm{SN}}(s_2;\overline{\theta}_c)}.$$

As before, FDR is estimated according to Equation 4.

3.2.1 Parameter initialization

Similar to the parameter initialization for the top score mixture model, the top score is partitioned into two sets separated by its median and the points below the median are used to obtain a method of moments estimator of θ_1 and the points above the median are used for θ_c. The points of the second score corresponding to the top scores below the median, are used to obtain the initial estimate of θ_2. To ensure that the entire parameter space is searched for an optimal fit, we run the algorithm eight times covering all possible combinations of signs of Δ_c, Δ_1 and Δ_2, with the final fit selected based on the value of the likelihood function. Parameters α and β are both initialized at 0.5.

4 Experiments and results

The experiments in this study were designed to investigate the properties and performance of the new methods. We first look at the accuracy of FDR estimation using the spectral libraries from NIST. We further use the libraries from NIST and datasets from PRIDE to evaluate the quality of the fit of the generative models and quantify the stability of FDR estimation. Finally, we use an in-house experiment with diluted lysate of HeLa cells, with the total amount of digested protein ranging from 0.1 to 100 ng per analysis, to assess the robustness of FDR estimation to uncertainty and noise resulting from reduced levels of biological material and reduced levels of analytes.

4.1 Datasets

We used public and in-house data for model evaluation. The public data consisted of 4 ion trap datasets across 4 species from NIST spectral libraries (Stein, 1990) and 10 datasets across 6 species from the PRIDE database (Vizcaino *et al.*, 2016). All datasets are summarized in Table 3. The protocols for generating in-house data and all relevant experimental details are described in Section 4.6.

4.2 Database search

All searches were carried out using MS-GF+ (Kim and Pevzner, 2014), with search parameters identical to those from the publications associated with each dataset. Each dataset was searched against the corresponding species' proteomics database downloaded from UniProtKB (Bairoch, 2004). We carried out two searches. The first run was a TDA, where the decoy database was constructed by reversing tryptic peptides as proposed by Elias and Gygi (2007) and then concatenating these peptides to the target database. FDR at a score threshold τ was estimated as $\mathrm{FDR}(\tau) = \frac{n_D(\tau)}{n_T(\tau)}$, where $n_D(\tau)$ is the number of top-scoring PSMs above τ that came from the decoy database and $n_T(\tau)$ is the number of top-scoring PSMs above τ that came from the target database. The second search was performed using the target database only and retaining up to 10 highest-scoring PSMs for each experimental spectrum. The results of these searches were used for decoy-free FDR estimation, as described in Section 3.

Table 3. Datasets from *Arabidopsis thaliana, Drosophila melanogaster, Escherichia coli, Homo sapiens, Mus musculus, Saccharomyces cerevisiae* and *Caenorhabditis elegans* used for evaluation

Dataset PXD	Species	Spectra	PSM	Precursor tolerance	Instrument	Fragmentation method	Missed cleavages
PXD001179	*A.thaliana*	116 487	80 894	10 ppm	LCQ/LTQ	CID or by detection	1
PXD006080	*D.melanogaster*	181 749	72 240	25 ppm	Orbitrap/FTICR/Lumos	CID or by detection	1
PXD001481	*E.coli*	59 765	43 217	10 ppm	LCQ/LTQ	CID or by detection	1
PXD012755	*H.sapiens*	48 754	48 451	25 ppm	Orbitrap/FTICR/Lumos	CID or by detection	1
PXD011988	*H.sapiens*	35 358	35 176	25 ppm	Orbitrap/FTICR/Lumos	CID or by detection	1
PXD013092	*M.musculus*	86 139	55 312	15 ppm	Q-Exactive	HCD	2
PXD001054	*M.musculus*	69 198	66 113	15 ppm	Q-Exactive	HCD	2
PXD001054	*M.musculus*	57 701	55 312	15 ppm	Q-Exactive	HCD	2
PXD001928	*S.cerevisiae*	39 284	38 890	10 ppm	Q-Exactive	CID or by detection	2
PXD001928	*S.cerevisiae*	37 087	36 402	10 ppm	Q-Exactive	CID or by detection	2
NIST Ion Trap	*C.elegans*	67 470	67 308	25 ppm	LCQ/LTQ	CID or by detection	2
	H.sapiens	340 351	339 857	25 ppm	LCQ/LTQ	CID or by detection	2
	M.musculus	149 453	149 325	25 ppm	LCQ/LTQ	CID or by detection	2
	S.cerevisiae	92 608	92 507	25 ppm	LCQ/LTQ	CID or by detection	2

Note: MS-GF+ automatically sets the fragment ion tolerance based the chosen fragmentation method.

4.3 Quality of FDR estimates

We searched NIST spectral libraries to establish the accuracy of FDR estimation. For each species and instrument platform, a NIST library consists of a set of consensus spectra, each associated with a peptide sequence, that can be considered as ground truth for our evaluation. After completing a search for which we estimated FDR, we computed the fraction of identified PSMs that did not match peptides from the NIST database as the true FDR and compared the two FDR values. This approach, however, has limitations. First, some peptides from NIST were not present in UniProtKB ensuring incorrect identifications in our searches whenever such a peptide received a sufficiently high score. Second, a peptide-spectrum pair in the NIST library may not always be a correct assignment in the first place because MS/MS searches may repeatedly lead to the same incorrect identifications due to database issues, peculiarities of the search parameters and software, or random chance. Third, we used the precursor mass tolerance of 25 ppm that may be too stringent for the instrument types. This precursor tolerance was chosen to demonstrate the proof-of-principle of the developed approaches and show its potential applicability to data generated by different types of mass analyzers. Additionally, in some cases k different peptides may be tied for the top score. We counted a $\frac{k-1}{k}$ fractional error in these cases if the correct peptide was among the k peptides; otherwise, we counted a full error, regardless of the presence of the correct peptide in UniProtKB. An example of such a situation are peptides with leucine-to-isoleucine substitutions.

Figure 1 shows the estimated versus true FDR averaged over four species from NIST in logarithmic and linear scale. We observe that the one-sample DFAs underestimate FDR, whereas the TDA and the two-sample DFA (2SMix) generates a curve closer to the diagonal line. Based on these results we conclude that the performance of TDA and the two-sample DFA is comparable, with the two-sample DFA having a slightly better performance in the low FDR range (0.001–0.01) and TDA having a slightly better performance in the high FDR range (0.01-0.1).

4.4 Quality of the fit

Spectral libraries from NIST were also used to evaluate quality of the fit of the three DFAs. To do so, we plot the estimated pdfs against the empirical score distributions in Figure 2. For each dataset, we evaluate the log-likelihood of the mixture sample $\mathbb{S}_1$ and measure the cumulative distribution function (cdf) fit by computing δ_{CDF} as the unnormalized distance by Yang *et al.* (2019), with $p = 1$, between the empirical and estimated cdfs. For the two-sample DFA, we also evaluate the log-likelihood of the combined samples $\mathbb{S}_1$ and $\mathbb{S}_2$ and additionally compute δ_{CDF} for $\mathbb{S}_2$. The distance between two cdfs was computed using the discrete cdf vectors of length $|\mathbb{S}_1|$ or $|\mathbb{S}_2|$, as applicable.

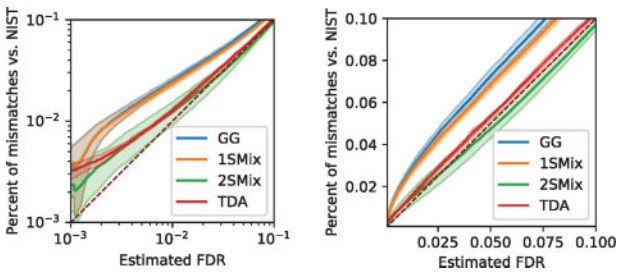

Fig. 1. Fraction of mismatches in NIST library versus estimated FDR. The closer to the identity line, the more accurate the estimation. Each curve is averaged over four NIST datasets, with the bands showing 68% confidence intervals. On the left, we show the log-scale to emphasize the range of more practical interest, while on the right we use linear scale

One-sample skew normal DFA improved the quality of the fit over Gamma-Gaussian DFA both in terms of log-likelihood and δ_{CDF} (Supplementary Materials). The log-likelihood values have been normalized by the sample size thus making the differences appear smaller than they are, whereas the δ_{CDF} measure appeared to be more in line with the visual inspection of the pdf fit. The two-sample skew normal DFA has somewhat reduced quality on $\mathbb{S}_1$ compared to the one-sample skew normal DFA in both measures, but the high-quality fitting on $\mathbb{S}_2$ compensates for the difference. In addition, the quality of the fit of the second scores suggests that $\mathbb{S}_2$ indeed plays a role similar to that of the decoy database.

Datasets from PRIDE were additionally used to evaluate quality of the fit of the DFAs and to compare the cutoff values with TDA. The results of these experiments are summarized in Supplementary Materials for each of the 10 PRIDE datasets. Supplementary Table S1 gives summaries over these datasets. The findings on these datasets mirror those from NIST spectral libraries and increase confidence in strong performance of the two-sample DFA.

4.5 Stability of FDR estimates

The stability of the FDR estimates was investigated using bootstrapping (Efron and Tibshirani, 1986). In each of the $B = 200$ bootstrap iterations, the spectra entering the search were sampled with replacement into an equal-sized set. After the database search, the 1% FDR score threshold τ was estimated for each bootstrapped set using

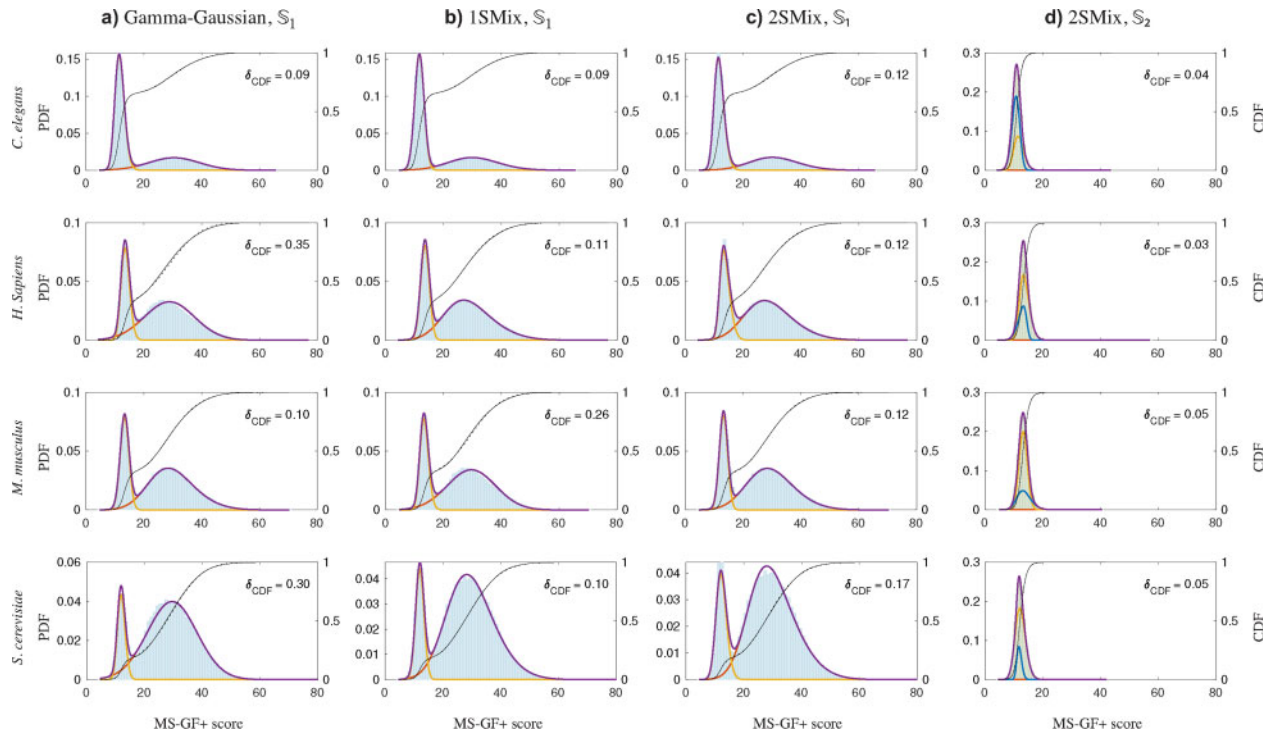

Fig. 2. Model fitting on four NIST datasets. (**a**) One-sample Gamma-Gaussian DFA estimation as proposed by Keller *et al.* (2002), (**b**) one-sample skew normal mixture 1SMix and (**c, d**) two-sample skew normal mixture 2SMix. Histograms show score distributions $\mathbb{S}_1$ (light blue) and $\mathbb{S}_2$ (light green), as a function of *E*-value. Purple densities superimpose estimated mixtures and their component distributions (yellow = top incorrect, blue = second-best incorrect orange = correct). Estimated cdfs are shown in dotted black lines which that are mostly overlapping with the empirical cdfs shown in solid black lines. Distances δ_{CDF}, log-likelihoods and 1% FDR thresholds are summarized in Supplementary Table S1, Supplementary Materials

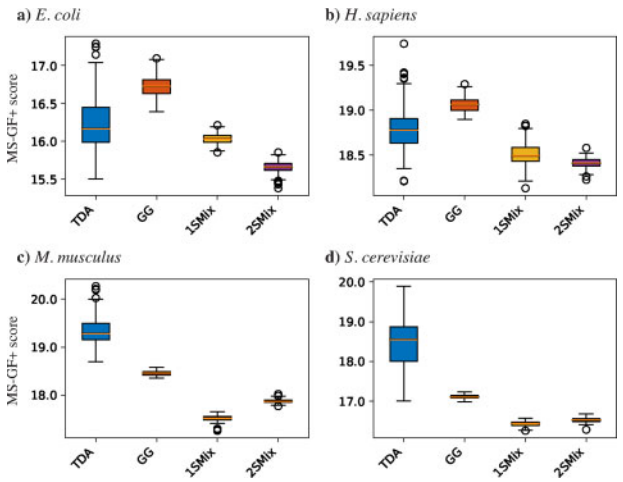

Fig. 3. Stability of FDR estimates on four select datasets from PRIDE. The stability of estimates was evaluated using 200 bootstrapping iterations and measuring the 1% FDR threshold in each of the iterations, as shown in the *y*-axis of each plot. The larger dispersion of established thresholds corresponds to lower stability of estimates

TDA and three DFAs. The variability in τ was then used to quantify stability of the estimates.

The stability of the four FDR estimation methods is compared in Figure 3 on four representative datasets from PRIDE. The results show that the TDA is generally less stable than any of the DFAs. This result is not entirely surprising given that the estimates of low FDR are often made based on a small number of decoy PSMs. Among DFAs, we find that one-sample DFAs were less stable than the two-sample DFA, suggesting that the two-sample DFA was able to capitalize on the existence of $\mathbb{S}_2$ to both improve and stabilize the estimate.

4.6 HeLa cell digest experiments

4.6.1 Experimental setting

To mimic the experiments requiring proteomic profiling of limited biomedical samples, we analyzed digested total lysate of cultured HeLa cells, which was selected as a representative high-complexity model sample. Sample aliquots were diluted to the desired concentration levels that corresponded to the total amount of digested protein ranging from 0.1 to 100 ng per analysis. The resulted specimens were analyzed using the conventional nano-flow liquid chromatography coupled with tandem mass spectrometry (nanoLC-MS/MS)-based approach, involving the separation conducted on a conventional 75 μm inner diameter (ID) in-house bead-packed column. According to our estimates, the injected sample amounts corresponded to approximately 1–1000 HeLa cells. The generated nanoLC-MS/MS data files were subjected to the analysis of spectral data, using the approach described next.

4.6.2 LC-MS/MS proteomics analysis

HeLa protein digest standard (P/N 88328, Thermo Fisher Scientific, Waltham, MA) was resuspended in 2% formic acid to desired concentration levels. 0.1, 1, 10, 50 and 100 ng of the HeLa digest aliquots were subjected to LC-MS/MS-based proteomics profiling. At least three technical replicates (i.e. replicate LC-MS/MS analyses of the same sample amount) were used across the whole study. The sample was loaded with the autosampler directly onto a self-packed column, which was made from a 75 μm ID 360 μm OD fused-silica capillary tubing (Molex, Polymicro Technologies, Phoenix, AZ) with a pulled tip filled with 20 cm of 1.9 μm ReproSil-Pur 120 C18-AQ (Dr. Maisch, Ammerbuch, Germany). Peptides were eluted at 150 nl/min from the column using an UltiMate 3000 HPLC system (Thermo Fisher Scientific) with a 60 min linear gradient from 1%

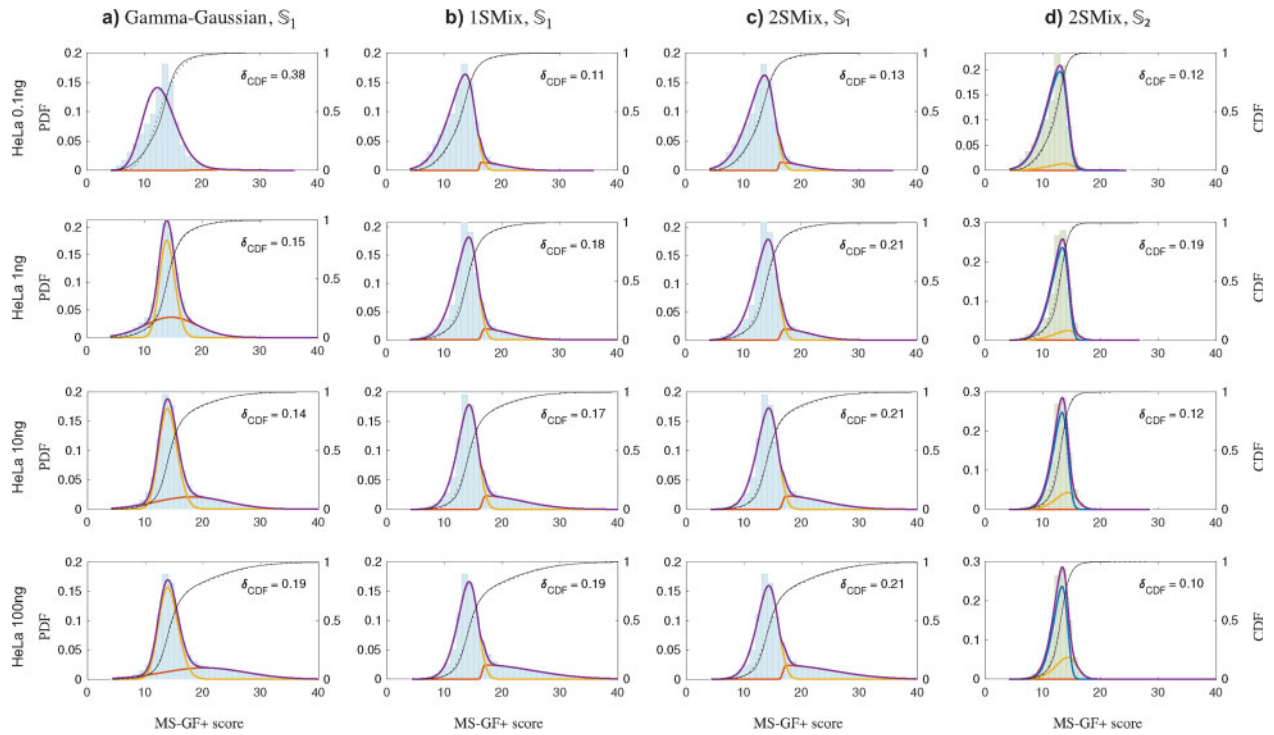

Fig. 4. Model fitting on four select HeLa cell datasets. (**a**) One-sample Gamma-Gaussian DFA estimation as proposed by Keller *et al.* (2002), (**b**) one-sample skew normal mixture 1SMix and (**c, d**) two-sample skew normal mixture 2SMix. Histograms show score distributions $\mathbb{S}_1$ (light blue) and $\mathbb{S}_2$ (light green), as a function of *E*-value. Purple densities superimpose estimated mixtures and their component distributions (yellow = top incorrect, blue = second-best incorrect orange = correct). Estimated cdfs are shown in dotted black lines which that are mostly overlapping with the empirical cdfs shown in solid black lines. Distances δ_{CDF}, log-likelihoods and 1% FDR thresholds are summarized in Supplementary Table S1, Supplementary Materials

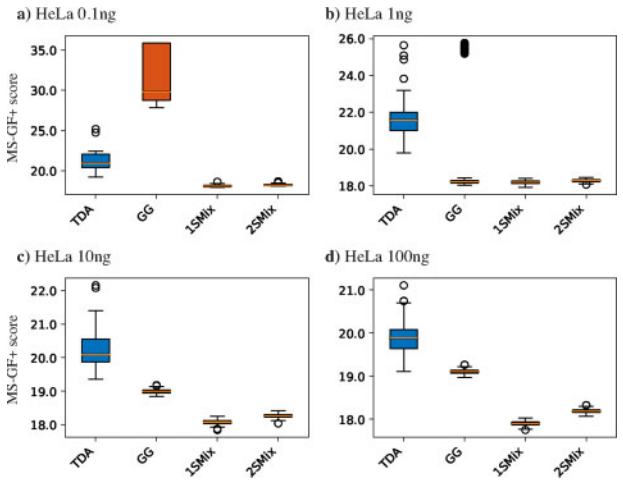

Fig. 5. Stability of FDR estimates on four select datasets from the HeLa cell experiments. The stability of estimates was evaluated using 200 bootstrapping iterations and measuring the 1% FDR threshold in each of the iterations, as shown in the *y*-axis of each plot. The larger dispersion of established thresholds corresponds to lower stability of estimates

solvent B to 20% solvent B (100% acetonitrile, 0.1% formic acid) mixed with solvent A (0.1% formic acid in water). The eluent composition was changed from 20 to 80% of solvent B over 2 min and held constant for 3 min. Finally, the elution solvent composition was changed from 80% solvent B to 99% solvent A over 1 min, and then held constant at 99% of solvent A for 15 min. The application of a 2.3 kV distal voltage electrosprayed the eluting peptides directly into an Orbitrap Fusion Lumos^{TM} mass spectrometer equipped with a Nanospray Flex Ion Source (both Thermo Fisher Scientific). Mass spectrometer-scanning functions and HPLC gradients were controlled by the Xcalibur software (Thermo Fisher Scientific, v.4.1.50). The temperature of the ion transfer tube was set to 275°C. The mass spectrometer was set to scan MS1 at 120 000 resolution at *m/z* 200 with an Automatic Gain Control (AGC) target set at 4e5 and for maximum injection time 50 ms. The RF lens was set to 30%. The scan range was *m/z* 375–1500. Monoisotopic precursor selection mode was set to 'Peptide.' For MS2, data-dependent acquisition mode was used. MS/MS spectra were acquired in the linear ion trap (rapid scan mode, HCD) with an AGC target of 3e4 and a maximum injection time (IT) at 35 ms. The highest abundance peaks were analyzed by MS2 for a cycle time of 3 s and injecting ions using parallelization mode. Peptides were isolated with an isolation window of *m/z* 1.6 and fragmented at higher-energy collisional dissociation energy of 28%. Only ions with a charge state of two through seven were considered for MS2. Dynamic exclusion was set at 30 s. The conversion of LC-MS .raw files to .mgf files was done using MSFileReader (v.2.2.62) and RawConverter v.1.1.0.23 (He *et al.*, 2015). The default conditions for conversion were used, with one exception, charge states from two through seven were used. The datasets were deposited in PRIDE (PXD020322).

4.6.3 Results on HeLa cell experiments

Figure 4 shows a significantly improved fit of one- and two-sample skew normal mixtures compared to the Gamma-Gaussian mixture. Figure 5 further visualizes stability of the 1% FDR threshold in a bootstrapping experiment (as described in Section 4.5), suggesting that the two-sample skew normal mixture (2SMix) offers an attractive combination of fit and stability. Finally, Figure 6 shows the number of identified PSMs as a function of estimated FDR in each of the experiments. It is worth noting here that the comparisons in Figure 6 are not straightforward because each method estimates its own FDR and does so with different accuracy. However, we have previously demonstrated that TDA and the 2SMix DFA have comparable quality of FDR estimates (Fig. 1). In that light, we can more confidently infer an increased number of PSM identifications for the

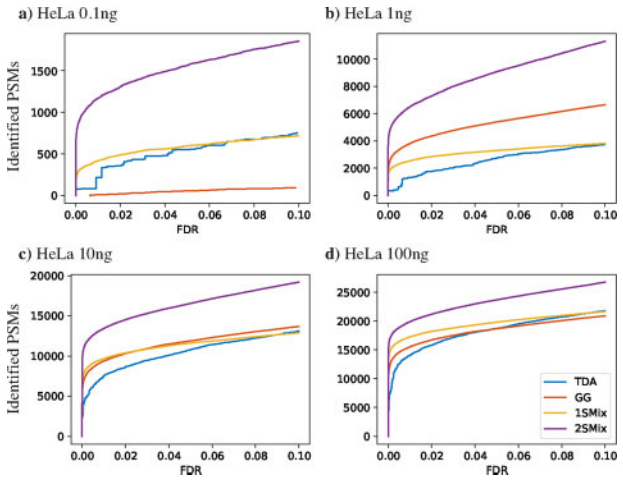

Fig. 6. The number of identified PSMs on the four select HeLa cell experiments at a specific FDR, separately estimated by each of the four individual methods

2SMix DFA compared to TDA. Specifically, 687 more identifications for 0.1 ng (+331%), 2309 for 1 ng (+168%), 3488 for 10 ng (+47%) and 2469 for 100 ng (+18%) when averaged over the three replicates of each experiment.

Deep proteomic profiling of scarce biological and clinical samples is still a major challenge. The ability to qualitatively and quantitatively characterize thousands of proteins and their post-translational modifications present in limited samples (e.g. rare cell populations, microneedle biopsies, microsampled liquid biopsies and even individually isolated single cells) is immensely important for getting new information in fundamental biology research and enabling novel diagnostic and prognostic studies (Huffman *et al.*, 2019; Li *et al.*, 2015, 2018; Lombard-Banek *et al.*, 2019; Shao *et al.*, 2018; Zhu *et al.*, 2018). However, the conventional nanoLC-MS/MS techniques fail to generate highly informative data at such sample levels. Since protein-derived analytes are at very low amounts in limited samples, the resulting MS and MS/MS spectra are generally sparse and low intensity. Interpretation of MS/MS fragmentation patterns resulting in correct peptide sequence identification and ultimately in in-depth protein and proteome characterization becomes a challenge using such low signal-to-noise-ratio and low fragmentation-efficiency spectra. Therefore, nanoLC-MS/MS analysis of limited samples typically results in a low conversion efficiency from tandem MS spectra to high-quality PSMs and a high FDR in peptide and protein identification, which in turn lead to limitations in quantitative analysis. We believe that the methodology proposed in this work improves the analysis of such samples.

5 Conclusions

Accurate FDR estimation has been one of the major computational challenges in bottom-up proteomics (Aggarwal and Yadav, 2016; Nesvizhskii, 2010) and is a key component of both peptide and protein identification (Li and Radivojac, 2012; Serang and Noble, 2012). Although several approaches have been widely evaluated and used (Elias and Gygi, 2007; Jeong *et al.*, 2012; Kall *et al.*, 2008a; Keller *et al.*, 2002), questions remain about their modeling assumptions, accuracy, stability, rigor and speed. The new types of experiments with low-amount analytes from limited samples, as the HeLa studies from our work, exemplify these challenges and require improved estimators. To address these challenges we proposed and evaluated new decoy-free methods for FDR estimation. Our methods rely on mixtures of skew normal distributions designed to model all component distributions. Importantly, our approaches eliminate the need to use a decoy database and, with it, the competition between peptides potentially present in the biological sample with those that are not. This is particularly evident in our two-sample DFA that relies on the score distribution of second-best PSMs associated with each spectrum and also models some level of dependence between first and second score distributions via parameter sharing and constraints.

The new mixture model methodology was extensively evaluated on public and in-house data. We show that one-sample DFAs are slightly inferior to TDA in terms of quality of FDR estimation, although they are faster and often more stable. On the other hand, our two-sample DFA offers an equivalent level of accuracy of FDR estimates as TDA, but with increased stability, improved speed and slightly reduced cutoff thresholds that result in an increased number of PSM identifications (Section 4). At the same time, the two-sample DFA retains methodological elegance of one-sample DFAs because skew normal distributions lend themselves to an efficient maximum likelihood optimization using expectation-maximization (Section 3). We believe that the new method will be applicable across a range of FDR estimation scenarios in bottom-up proteomics and beyond; e.g. with searches including post-translational modifications (Fu, 2012), cross-linked peptides (Walzthoeni *et al.*, 2012), semi-tryptic peptides (Alves *et al.*, 2008), *de novo* searches (Dancik *et al.*, 1999; Frank and Pevzner, 2005), small molecule searches (Scheubert *et al.*, 2017; Wang *et al.*, 2018).

Acknowledgements

The authors acknowledge Thermo Fisher Scientific for their support through a technology alliance.

Funding

This work was supported by the National Institutes of Health awards [R01GM103725 to P.R., R01GM120272 to A.R.I., R01CA218500 to A.R.I., R35GM136421 to A.R.I.].

Conflict of Interest: none declared.

References

Aebersold,R. and Mann,M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207.

Aggarwal,S. and Yadav,A.K. (2016) False discovery rate estimation in proteomics. *Methods Mol. Biol.*, **1362**, 119–128.

Alves,P. *et al.* (2008) Fast and accurate identification of semi-tryptic peptides in shotgun proteomics. *Bioinformatics*, **24**, 102–109.

Arellano-Valle,R.B. *et al.* (2006) A unified view on skewed distributions arising from selections. *Can. J. Stat.*, **34**, 581–601.

Azzalini,A. (1985) A class of distributions which includes the normal ones. *Scand. J. Stat.*, **12**, 171–178.

Bairoch,A. (2004) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–159.

Budnik,B. *et al.* (2018) SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biol.*, **19**, 161.

Burger,T. (2018) Gentle introduction to the statistical foundations of false discovery rate in quantitative proteomics. *J. Proteome Res.*, **17**, 12–22.

Choi,H. and Nesvizhskii,A.I. (2008) False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *J. Proteome Res.*, **7**, 47–50.

Choudhary,C. and Mann,M. (2010) Decoding signalling networks by mass spectrometry-based proteomics. *Nat. Rev. Mol. Cell Biol.*, **11**, 427–439.

Cooper,B. (2011) The problem with peptide presumption and low Mascot scoring. *J. Proteome Res.*, **10**, 1432–1435.

Cooper,B. (2012) The problem with peptide presumption and the downfall of target-decoy false discovery rates. *Anal. Chem.*, **84**, 9963–9967.

Dancik,V. *et al.* (1999) De novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.*, **6**, 327–342.

Danilova,Y. *et al.* (2019) Bias in false discovery rate estimation in mass-spectrometry-based peptide identification. *J. Proteome Res.*, **18**, 2354–2358.

Dempster,A.P. *et al.* (1977) Maximum likelihood from data via the EM algorithm. *J. R. Stat. Soc. B*, **39**, 1–38.

Efron,B. and Tibshirani,R. (1986) Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.*, **1**, 54–77.

Elias,J.E. and Gygi,S.P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods*, **4**, 207–214.

Frank,A. and Pevzner,P. (2005) PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal. Chem.*, **77**, 964–973.

Fu,Y. (2012) Bayesian false discovery rates for post-translational modification proteomics. *Stat. Interface*, **5**, 47–59.

Gingras,A.C. *et al.* (2007) Analysis of protein complexes using mass spectrometry. *Nat. Rev. Mol. Cell Biol.*, **8**, 645–654.

Gupta,N. *et al.* (2011) Target-decoy approach and false discovery rate: when things may go wrong. *J. Am. Soc. Mass Spectrom.*, **22**, 1111–1120.

He,L. *et al.* (2015) Extracting accurate precursor information for tandem mass spectra by RawConverter. *Anal. Chem.*, **87**, 11361–11367.

Hubler,S.L. *et al.* (2020) Challenges in peptide-spectrum matching: a robust and reproducible statistical framework for removing low-accuracy, high-scoring hits. *J. Proteome Res.*, **19**, 161–173.

Huffman,R.G. *et al.* (2019) DO-MS: data-driven optimization of mass spectrometry methods. *J. Proteome Res.*, **18**, 2493–2500.

Jain,S. *et al.* (2019) Identifiability of two-component skew normal mixtures with one known component. *Scand. J. Stat.*, **46**, 955–986.

Jeong,K. *et al.* (2012) False discovery rates in spectral identification. *BMC Bioinformatics*, **13**, S2.

Ji,C. *et al.* (2016) XLSearch: a probabilistic database search algorithm for identifying cross-linked peptides. *J. Proteome Res.*, **15**, 1830–1841.

Kall,L. *et al.* (2008a) Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.*, **7**, 29–34.

Kall,L. *et al.* (2008b) Posterior error probabilities and false discovery rates: two sides of the same coin. *J. Proteome Res.*, **7**, 40–44.

Keich,U. and Noble,W.S. (2015) On the importance of well-calibrated scores for identifying shotgun proteomics spectra. *J. Proteome Res.*, **14**, 1147–1160.

Keller,A. *et al.* (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.*, **74**, 5383–5392.

Kim,S. *et al.* (2008) Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J. Proteome Res.*, **7**, 3354–3363.

Kim,S. and Pevzner,P.A. (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.*, **5**, 5277.

Kong,A.T. *et al.* (2017) MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods*, **14**, 513–520.

Li,Q. (2008) Statistical methods for peptide and protein identification using mass spectrometry. Ph.D. Thesis, University of Washington.

Li,Y.F. and Radivojac,P. (2012) Computational approaches to protein inference in shotgun proteomics. *BMC Bioinformatics*, **13**, S4.

Li,Y.F. *et al.* (2012) Protein identification problem from a Bayesian point of view. *Stat. Interface*, **5**, 21–38.

Li,S. *et al.* (2015) An integrated platform for isolation, processing, and mass spectrometry-based proteomic profiling of rare cells in whole blood. *Mol. Cell Proteomics*, **14**, 1672–1683.

Li,Z.Y. *et al.* (2018) Nanoliter-scale oil-air-droplet chip-based single cell proteomic analysis. *Anal. Chem.*, **90**, 5430–5438.

Lin,T.I. *et al.* (2007) Finite mixture modelling using the skew normal distribution. *Stat. Sin.*, **17**, 909–927.

Lombard-Banek,C. *et al.* (2019) Microsampling capillary electrophoresis mass spectrometry enables single-cell proteomics in complex tissues: developing cell clones in live *Xenopus laevis* and zebrafish embryos. *Anal. Chem.*, **91**, 4797–4805.

Ma,K. *et al.* (2012) A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. *BMC Bioinformatics*, **13**, S1.

Nesvizhskii,A.I. (2010) A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics*, **73**, 2092–2123.

Perkins,D.N. *et al.* (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.

Rinner,O. *et al.* (2008) Identification of cross-linked peptides from large sequence databases. *Nat. Methods*, **5**, 315–318.

Scheubert,K. *et al.* (2017) Significance estimation for large scale metabolomics annotations by spectral matching. *Nat. Commun.*, **8**, 1494.

Serang,O. and Noble,W. (2012) A review of statistical methods for protein identification using tandem mass spectrometry. *Stat. Interface*, **5**, 3–20.

Shao,X. *et al.* (2018) Integrated proteome analysis device for fast single-cell protein profiling. *Anal. Chem.*, **90**, 14003–14010.

Steen,H. and Mann,M. (2004) The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.*, **5**, 699–711.

Stein,S.E. (1990) National Institute of Standards and Technology (NIST) mass spectral database and software. *Version 3.02, USA*.

Storey,J.D. (2002) A direct approach to false discovery rate. *J. R. Stat. Soc. B*, **64**, 479–498.

Tabb,D.L. *et al.* (2007) MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. Proteome Res.*, **6**, 654–661.

Vizcaino,J.A. *et al.* (2016) 2016 update of the PRIDE database and related tools. *Nucleic Acids Res.*, **44**, D447–D456.

Walzthoeni,T. *et al.* (2012) False discovery rate estimation for cross-linked peptides identified by mass spectrometry. *Nat. Methods*, **9**, 901–903.

Wang,X. *et al.* (2018) Target-decoy-based false discovery rate estimation for large-scale metabolite identification. *J. Proteome Res.*, **17**, 2328–2334.

Yang,R. *et al.* (2019) A new class of metrics for learning on real-valued and structured data. *Data Min. Knowl. Disc.*, **33**, 995–1016.

Yates,J.R. *et al.* (1995) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem.*, **67**, 1426–1436.

Young,J.C. and Minder,C.E. (1974) Algorithm as 76: an integral useful in calculating non-central t and bivariate normal probabilities. *J. R. Stat. Soc. C*, **23**, 455–457.

Zhu,Y. *et al.* (2018) Nanodroplet processing platform for deep and quantitative proteome profiling of 10-100 mammalian cells. *Nat. Commun.*, **9**, 882.