Group-walk, a rigorous approach to group-wise false discovery rate analysis by target-decoy competition

Jack Freestone[1], Temana Short[1], William Stafford Noble[2], Uri Keich[1]
[1]School of Mathematics and Statistics F07
University of Sydney
[2]Departments of Genome Sciences and of Computer Science and Engineering
University of Washington

April 15, 2022

Abstract

Target-decoy competition (TDC) is a commonly used method for false discovery rate (FDR) control in the analysis of tandem mass spectrometry data. This type of competition-based FDR control has recently gained significant popularity in other fields after Barber and Candès laid its theoretical foundation in a more general setting that included the feature selection problem. In both cases, the competition is based on a head-to-head comparison between an (observed) target score and a corresponding decoy (knockoff) score. However, the effectiveness of TDC depends on whether the data is homogeneous, which is often not the case: in many settings, the data consists of groups with different score profiles or different proportions of true nulls. In such cases, applying TDC while ignoring the group structure often yields imbalanced lists of discoveries, where some groups might include relatively many false discoveries and other groups include relatively very few. On the other hand, as we show, the alternative approach of applying TDC separately to each group does not rigorously control the FDR.

We developed *Group-walk*, a procedure that controls the FDR in the target-decoy / knockoff setting while taking into account a given group structure. Group-walk is derived from the recently developed AdaPT — a general framework for controlling the FDR with side-information. We show using simulated and real datasets that when the data naturally divides into groups with different characteristics Group-walk can deliver consistent power gains that in some cases are substantial. These groupings include the precursor charge state (4% more discovered peptides at 1% FDR threshold), the peptide length (3.6% increase) and the mass difference due to modifications (26% increase).

Group-walk is available at https://cran.r-project.org/web/packages/groupwalk/index.html

1 Introduction

Tandem mass spectrometry (MS/MS) currently provides the most efficient means of studying proteins in a high-throughput fashion. In a "shotgun proteomics" MS/MS experiment, proteins in a

complex biological sample are extracted and digested into peptides, each with an associated charge. These charged peptides, called "precursors", are measured by the mass spectrometer, and a subset of the precursors are then selected for further fragmentation into charged ions, which are detected and recorded by a second round of mass spectrometry [16, 26]. The resulting tandem fragmentation spectra, or spectra for short, are then subjected to computational analysis. A typical 30-minute MS/MS experiment will generate approximately 18,000 such spectra. Canonically, each observed spectrum is generated by a single peptide. Thus, the first goals of the downstream analysis are to identify which peptide generated each of the observed spectra (the spectrum-ID problem) and to determine which peptides and which proteins were present in the sample (the peptide/protein detection problems).

In each of those three problems the canonical approach to determining the list of discoveries is by controlling the false discovery rate (FDR) through some form of target-decoy competition (TDC). TDC has been widely practiced by the computational mass spectrometry community since it was first proposed by Elias and Gygi [9, 6, 17, 10, 14, 31].

To understand how TDC is applied in practice, consider for example the spectrum-ID problem. This analysis is typically initiated by using a search engine to scan each input spectrum against a target peptide database for its best matching peptide (in practice the spectrum is matched only against a subset of candidate peptides whose mass is within the measurement tolerance of the so-called precursor mass associated with the spectrum). Pioneered by SEQUEST [12], the search engine uses an elaborate score function to quantify the quality of the match between each of the database peptides and the observed spectrum, recording the optimal peptide-spectrum match (PSM) for the given spectrum along with its score Z_i [25]. In practice, many expected fragment ions will fail to be observed for any given spectrum, and the spectrum is also likely to contain a variety of additional, unexplained peaks [26]. Hence, sometimes the reported PSM is correct — the peptide assigned to the spectrum was present in the mass spectrometer when the spectrum was generated — and sometimes the PSM is incorrect. Ideally, we would report only the correct PSMs, but obviously we do not know which PSMs are correct and which are incorrect; all we have is the score of the PSM, indicating its quality. TDC allows us to report a thresholded list of top-scoring PSMs while controlling the list's FDR, as explained next.

First, the same search engine is used to assign each input spectrum a *decoy* PSM score, $\tilde{Z}_i$, by searching for the spectrum's best match in a decoy database of peptides obtained from the original database by randomly shuffling or reversing each peptide in the database. Each decoy score $\tilde{Z}_i$ then directly competes with its corresponding target score Z_i to determine the reported list of discoveries, i.e., we only report target PSMs that win their competition: $Z_i > \tilde{Z}_i$. Additionally, the number of decoy wins ($\tilde{Z}_i > Z_i$) in the top k scoring PSMs is used to estimate the number of false discoveries in the target wins among the same top k PSMs. Thus, the ratio between the number of decoy wins and the number of target wins yields an estimate of the FDR among the target wins in the top k PSMs. To control the FDR at level α, the TDC procedure chooses the largest k for which the estimated FDR is still $\leq \alpha$, and it reports all target wins among those top k PSMs. It was recently shown that, assuming that *incorrect PSMs are independently equally likely to come from a target or a decoy match*, and provided we add 1 to the number of decoy wins before dividing by the number of target wins, this procedure rigorously controls the FDR [15].

As first noted by Efron in the context of the canonical p-value-based FDR analysis, the implicit assumption that FDR control relies on is that the data is essentially homogeneous. When, as often happens, the data is made of groups of hypotheses with distinct group-wise characteristics (e.g.,

different proportion of true nulls), performing a global FDR analysis can be problematic [8].

Efron's observation is even more important in the context of TDC, where the issue of score calibration is known to play a significant role [17, 20]. For example, Baker, Medzihradszky, and Chalkley noted that spectra characteristics vary with the precursor's charge state [1]. Specifically, they noted that by separately applying TDC to each charge state they obtain more discoveries at the 1% FDR threshold than they get when applying TDC to the combined dataset [1, Table V]. Arguably even more disturbing is the fact that, because their group-specific 1% FDR thresholds vary significantly [1, Table IV], inevitably when a single cutoff is used for the entire set, the discoveries from some groups will be overly conservative while for others they will be overly liberal.

Baker et al. tried to address the specific charge-state structure by redesigning their score function so it is better calibrated with respect to charge state. However, they also noted a similar problem of group-wise imbalance when considering phosphopeptides vs. unmodified ones: ignoring the distinction between the two groups meant that, in their case, the FDR in the phosphopeptides group was significantly underestimated [1, Table VI]. Their suggestion in this case coincided with Efron's solution, which is to apply a separate FDR analysis to each group. Clearly, the latter is more attractive as it does not require us to redesign our score function to address any potential group structure. However, as Efron noted, applying a separate FDR analysis to each group can fail to control the FDR in the finite sample setting. Indeed, we show here that separately applying TDC to each group can significantly underestimate the true FDR.

Instead, we offer Group-walk a new procedure that shares the essence of separate TDC analyses but rigorously controls the FDR even for finite samples. Group-walk derives its theoretical guarantees from a recently published paper that developed a general framework for the problem of multiple hypothesis testing with generic side information; i.e., associated with each hypothesis is a p-value, as well as a predictor encoding contextual information about the hypothesis [22]. Of course, a group ID, as in our case, is a special case of such side information, but to the best of our knowledge we are the first to effectively adapt the general framework of Lei and Fithian to the context of group-wise TDC analysis.

We demonstrate the potential power of Group-walk using simulated as well as real datasets. The simulated data allows us to examine how Group-walk takes advantage of the different ways in which groups may vary. We complement the simulations with investigating multiple, naturally defined group structures in randomly selected real datasets. Specifically, we compare the number of peptides discovered from mass spectrometry data using TDC (ignoring the group structure) and using Group-walk, taking into account the charge state, peptide length and the mass difference due to post-translational modifications. In those analyses Group-walk overall delivers more discoveries across the entire range of examined FDR thresholds, varying, depending on the context, from a median increase of 3.6% all the way to a median increase of 26%.

Our motivation and our demonstrations are focused on the analysis of mass spectrometry data. However, Group-walk is in principle applicable in the broader context of competition-based approach to FDR, including its application to variable selection proposed by Barber and Candès [2]: if the data (variables) naturally partitions into groups with different characteristic the rigorous group-wise FDR analysis that Group-walk offers can be beneficial.

2 Background

The TDC procedure uses competition to control the FDR. The approach can be phrased in more general terms as follows. Let H_i ($i = 1, \ldots, m$) denote our m null hypotheses, e.g., in the spectrum-ID problem H_i is "the ith PSM is incorrect" and in the peptide detection problem H_i is "the ith peptide is present in the sample." Associated with each H_i are two competing scores: a target score Z_i (the higher the score the less likely H_i is) and a decoy score $\tilde{Z}_i$. For example, in spectrum-ID Z_i ($\tilde{Z}_i$) is the score of the optimal target (decoy) peptide match to the ith spectrum, whereas in peptide detection Z_i corresponds to the maximal score of all PSMs that are optimally matched to the ith target peptide (and $\tilde{Z}_i$ is similarly defined for the ith decoy peptide).

Adopting the notation of [11] we associate with each hypothesis a "winning" score W_i and a target/decoy-win label L_i. For the problems we consider here $W_i = Z_i \vee \tilde{Z}_i$ (where $x \vee y = \max\{x, y\}$), but other functions can be considered as well [2]. As for L_i, we define it as 1 if H_i corresponds to a target win and as -1 if H_i corresponds to a decoy win. Here we assume that ties ($Z_i = \tilde{Z}_i$) are broken randomly, but ignoring hypotheses with ties is also valid.

Let $N = \{i : H_i \text{ is a true null}\}$ and note that while typically in the context of hypotheses testing N is a constant, albeit unknown set, here we allow N to be a random set. The fundamental assumption that TDC relies on is

Assumption 1. *Conditional on all the scores $\{W_i\}_i$ and all the false null labels $\{L_i : i \notin N\}$, the true nulls are independently equally likely to be a target or a decoy win, i.e., the random variables (RVs) $\{L_i : i \in N\}$ are conditionally independent uniform ± 1 RVs.*

Clearly, if the assumption holds then $\{L_i : i \in N\}$ are still independent uniform ± 1 RVs after ordering the hypotheses in increasing order (ties broken randomly) so that $W_1 \leq W_2 \leq \cdots \leq W_m$. Hence, without loss of generality the scores are assumed to be in increasing order.

Let $D(k) = \sum_{i=k}^{m} 1_{\{L_i=-1\}}$ (the number of decoy wins with rank $\geq k$), and let $R(k) = \sum_{i=k}^{m} 1_{\{L_i=1\}}$ (the corresponding number of target wins). TDC defines its discovery list as $\{i \geq K : L_i = 1\}$, i.e., all the target wins with rank $\geq K$, where $K = K(\alpha) = \min\{k : (D(k) + 1)/(R(k) \vee 1) \leq \alpha\}$ is TDC's rejection threshold.

It was proved by He et al. that TDC controls the FDR in the finite sample setting [15]. That is, with $V(k) := \sum_{j=k}^{m} 1_{\{L_j=1, j \in N\}}$ (the number of true null target wins with rank $\geq k$), the FDP among the corresponding target wins is $Q_k = V(k)/(R(k) \vee 1)$, and $E(Q_K) \leq \alpha$. The expectation is taken with respect to the labels of the true null hypotheses conditional on the false null labels and on all the given scores.

At essentially the same time Barber and Candès established the same result through a sweeping generalization that deals with sequential hypothesis testing. In that setup the order of the hypotheses is predetermined, and a p-value is associated with each hypothesis so that given the order, as well as the p-values of the false nulls, the p-values of the true null hypotheses are indendent and identically distributed with a distribution that stochastically dominates the uniform distribution (so they are valid p-values). In our case we order the hypotheses by the scores W_i, and the p-values are defined as a "1-bit p-value": 1 for a decoy win and 1/2 for a target win. It is easy to see that Assumption 1 guarantees that the conditions of the sequential hypothesis testing are satisfied so it indeed generalizes TDC's setup.

One reason TDC was adopted in the mass spectrometry context, rather than relying on standard methods for control of the FDR such as the procedures by Benjamini and Hochberg [3] or

Storey [29], is that the latter require sufficiently informative p-values and, initially, no such p-values were computed in this context (as noted above, using the decoys we can always assign a 1-bit p-value to the hypotheses, but those are not informative enough to obtain effective results using the latter procedures). Moreover, the proteomics dataset typically consist of both "native" spectra (those for which their generating peptide is in the target database) and "foreign" spectra (those for which it is not). These two types of spectra create different types of false positives, implying that we typically cannot apply the standard FDR controlling procedures to the spectrum-ID problem even if we are able to compute p-values [18].

3 Methods

3.1 Group-walk

Group-walk is our novel procedure designed to analyze the same target-decoy data that TDC does while taking advantage of the data's group structure. Specifically, in addition to the label L_i and the score W_i, each hypothesis has a group identifier $G_i \in \{1, 2, \ldots, n_g\}$ where $n_g \geq 1$ is the number of groups. Accordingly, we revise Assumption 1 to account for the group information:

Assumption 2. *Conditional on all the scores $\{W_i\}_i$, the group IDs $\{G_i\}_i$ and all the false null labels $\{L_i : i \notin N\}$, the true nulls are independently equally likely to be a target or a decoy win.*

To understand how Group-walk works it is instructive to revisit TDC. Assuming the scores W_i are in increasing order, we define the "front" as the index k that we are currently considering as the rejection threshold. Starting with the front set to $k = 1$, TDC iteratively checks if the FDR estimated from all the hypotheses that the front is yet to pass over is below the threshold α (i.e., if $(D(k)+1)/(R(k) \vee 1) \leq \alpha$). If it is, then TDC stops and reports the location of the front (implying that all the target-winning hypotheses that the front is yet to pass over are rejected, or considered discoveries); otherwise, TDC advances the front by one step $k := k + 1$.

TDC considers a single group, and therefore a 1-dimensional front, whereas Group-walk's front is an n_g-dimensional vector $\mathbf{k} = (k_1, \ldots, k_{n_g})$, where each $k_g \in \{1, 2, \ldots, m_g\}$ and m_g is the size of the gth group. Assuming the scores $\{W_i^g\}_{i=1}^{m_g}$ in each group are sorted in increasing order (with corresponding labels L_i^g), Group-walk starts with the trivial front $\mathbf{k} = (1, \ldots, 1)$, i.e., the first index (lowest scoring hypothesis) in each of the groups. Just like TDC, it then iteratively checks if the FDR estimated from all the hypotheses that the front is yet to pass over is below the threshold α. For Group-walk this translates to checking whether $(1 + \sum_{g=1}^{n_g} D_g(k_g))/(1 \vee \sum_{g=1}^{n_g} R_g(k_g)) \leq \alpha$, where $D_g(k) = \sum_{i=k}^{m_g} 1_{\{L_i^g = -1\}}$ (the number of decoy wins in group g with rank $\geq k$) and $R_g(k) = \sum_{i=k}^{m_g} 1_{\{L_i^g = 1\}}$ (the corresponding number of target wins).

If that estimated FDR is $\leq \alpha$, then Group-walk stops and returns its current front (implying all the target-winning hypotheses that the front is yet to pass over are rejected); otherwise, Group-walk advances the front by one step along a single coordinate. This is where Group-walk fundamentally differs from TDC: its front is n_g-dimensional so it needs to choose which of the g groups should be advanced.

Group-walk employs an initial trivial strategy of iteratively advancing one group at a time until the front reaches a pre-determined index K in each group: $\mathbf{k} = (K, K, \ldots, K)$ (here we used $K = 40$). Note that in some rare cases Group-walk might stop before the front reaches this point,

and if the front needs to advance beyond the length of a group then that group is tossed away with no discoveries.

When Group-walk reaches the uniform front of $\mathbf{k} = (K, K, \ldots, K)$ it switches its strategy to try to predict which of the groups is most likely to have a decoy at the current location of the front. That is, considering $L^g_{k_g}$ for $g = 1, \ldots, n_g$, which one is most likely $= -1$? The motivation for this approach is that any hypothesis that the front passes over is "lost": it can no longer be discovered. Of course, hypotheses for which $L^g_i = -1$ cannot be discoveries because they are decoy wins, so we might as well sacrifice those. Another way of looking at it is that where the chances of $L^g_i = -1$ are higher it is less likely be a region of high quality target wins.

Group-walk attempts to solve the question of which $L^g_{k_g}$ is most likely $= -1$ by looking at the last K labels of each group and choosing the one that has the most decoy wins. The algorithmic description of Group-walk is provided in Algorithm 1 in the Supplement.

In Supplementary Section 6.4 we provide the details establishing that Group-walk controls the FDR in the finite-sample setting by showing it is a special case of (a slightly modified version of) the Adaptive p-value Thresholding (AdaPT) framework of Lei and Fithian [22]. AdaPT was developed to address the problem of controlling the FDR, where associated with each hypothesis is a p-value p_i as well as a predictor x_i encoding contextual information about the hypothesis. In our case the p-value is the 1-bit p-value and $x_i = (W_i, G_i)$.

3.2 Simulations

Groups of hypotheses can roughly differ along the following three attributes that might vary between the groups: the distribution of the null scores ("group null-effect"), the degree of separation between the true and false null scores ("group separation-effect"), and the proportion of false nulls ("group π-effect"). We use simulated data to look at how Group-walk exploits those effects by applying it to a group structure of two groups that differ with respect to some of those attributes while agreeing on the others. Additionally, we looked at how differences in group sizes can accentuate or moderate group effects.

We focused our simulations on the spectrum-ID problem using a previously published model [19, 18]. Briefly, each virtual "spectrum" σ_i is associated with three randomly drawn scores: the score X_i of the match between σ_i and its generating peptide, the score Y_i of the best match to σ_i in the target database minus the generating peptide (if σ_i is native), and the score $\tilde{Z}_i$ of the best match to σ_i in the decoy database. The three scores are drawn independently of one another as well as of the corresponding scores of all other spectra. More specifically, Y_i and $\tilde{Z}_i$ are sampled from a null distribution (which can be spectrum-specific), whereas X_i is sampled from an alternative distribution for a native σ_i, and X_i is set to the lowest possible score for a foreign σ_i. The target PSM score is $Z_i = X_i \vee Y_i$, and the decoy PSM score is $\tilde{Z}_i$. Finally, the PSM is incorrect when $Y_i \vee \tilde{Z}_i > X_i$. It is easy to see that conditional on the PSM being incorrect, $P(Y_i > \tilde{Z}_i) = P(Y_i < \tilde{Z}_i)$ independently of everything else; hence, the spectrum-ID model is indeed captured by Assumption 1.

In Supplementary Section 6.1 we describe in detail how given the sizes $m_1, m_2 \in \{1\text{e}4, 2\text{e}4, 4\text{e}4\}$ of the two groups of null hypotheses ("the PSM is incorrect, i.e., $Y_i \vee \tilde{Z}_i > X_i$"), the separation parameters $b_1, b_2 \in \{5, 50\}$, and the proportions of native spectra (which are related to but not identical to the proportion of false nulls, cf. Section 2) $\pi_1, \pi_2 \in \{0.3, 0.6, 0.9\}$ we draw a random dataset of two groups of scores with associated target/decoy win labels.

A score is "calibrated" if all true null scores are generated according to the same distribution. Hence, in order to examine a group null-effect we simulated uncalibrated scores. Specifically we simulated a group null-effect due to distinct precursor charge states where the first group simulates PSMs with a precursor charge state 3 and the second group simulates charge 2 (see Supplementary Section 6.1 for details).

To smooth out random effects due to the random nature of our draws we drew 40 random datasets ("runs") for each of the 628 parameter combinations we considered. We then applied Group-walk (with $K = 40$), as well as TDC to each drawn dataset, where TDC was applied while ignoring the group structure, and the ratio of discovery numbers was averaged over the 40 runs.

3.3 Analysis of mass spectra

To test Group-walk on real data sets we randomly sampled one spectrum file from each of 14 randomly selected projects from the Proteomics Identifications Database (PRIDE) [28]: 7 high-resolution and 7 low-resolution spectra files. We then used the Crux toolkit [27] to (a) create a randomly shuffled decoy database from each project's target peptide database, and (b) generate the PSMs by searching each spectrum against the concatenated target-decoy peptide database (for details see Supplementary Section 6.2).

When analyzing real data the spectrum-ID analysis can be problematic because the same peptide can generate multiple spectra, which can potentially break the true-nulls independence assumption and violate FDR control [15]. Thus, we chose to focus on the peptide-level analysis of the real data. In our peptide-level analysis the score Z_i ($\tilde{Z}_i$) we assign to the ith target (decoy) peptide is the maximal score of all PSMs matched to that peptide.

To reduce the random effects due to the random nature of our decoys we repeated the database construction 20 times with different seeds to create varied shuffled decoy peptides. We then applied TDC as well as Group-walk to the list of target-decoy scores where the latter was also given the relevant group structure as described below. We compared their power by looking at quantiles of ratios of Group-walk to TDC discovery numbers where the quantiles were taken with respect to the combined list of ratios made of 20 such ratios (one for each decoy) per each dataset. We used overall quantiles here because TDC occasionally reported 0 discoveries so taking an average of the 20 ratios per dataset would have been problematic.

3.3.1 Charge and peptide length in low-resolution MS/MS data

We examined how Group-walk can take advantage of the precursor charge state and the peptide length focusing on the 7 low-resolution MS/MS data sets, because the differences in calibration were empirically pronounced there and thus served as a viable application of Group-walk. These data sets are still relevant despite the fact that they are low-resolution as the oldest one was published in 2019.

To group by charge we partitioned the peptides into groups according to the charge state of the spectrum that was maximally matched with the peptide. Note that spectra with charge states +2 or +3 include the majority of the spectra, whereas spectra with charge states +1, +4 and +5 are less frequent. The problem with having small groups is that Group-walk is more likely to exhaust them before the estimated FDR reaches the threshold α, thus preventing us from making any discoveries in them. To get around this, we merged the smaller groups — those that comprise less than 10% of

the total number of peptides — with the previous charge state group (or, in the case of the lowest charge state group, with the subsequent group).

Grouping peptides via the length of the matched peptide sequences is more challenging given the much larger range of observed peptide lengths. However, as noted below, our simulation studies suggested that it is advantageous to partition the peptides into equally sized groups (Section 4.2). Additionally, we expect that groups of peptides with similar lengths will have a more homogenous score distribution than groups with significantly varying peptide lengths. Hence, a natural length-based partition of the peptides is to partition the range of peptide lengths into g roughly equal-sized groups each covering a contiguous range, or an interval of peptide lengths. Here we considered $g = 2, 3$.

3.3.2 Open search

Canonically, each observed spectrum is not scored against the complete target-decoy database; instead, only a subset of the peptides in the database are scored — those whose mass is within some narrow window of tolerance about the observed spectrum mass. In practice, we can extend this window as large as we like, creating what is called an *open search*. Such a search allows us to identify spectra generated by peptides that undergo modifications which significantly change their mass. Specifically, in a standard (narrow) search, the mass of the unmodified (database) peptide does not fit within the tolerance window of such a modified precursor. Such modifications of the peptide mass may occur through several mechanisms including methylation and phosphorylation [4].

After identifying the spectra using an open search we proceed as in the narrow search by assigning each target/decoy peptide a score that corresponds to the score of its maximally matched PSM. At this point we can proceed with TDC; however, in doing so we would miss out on the information that is encoded in the mass difference between the unmodified peptide and the precursor. Specifically, we expect that the mass difference for an incorrect PSM will be randomly spread over the range of possible values, whereas the mass difference of a correct PSMs will concentrate on a few values including, for example, the mass difference due to phosphorylation.

To take advantage of that information we apply Group-walk to the peptides after splitting them into two groups as follows. For each identified target/decoy peptide in the database we determine the difference between its mass and the matching spectrum's precursor mass. We then bin the mass differences according to some regular intervals and assign the identified peptides whose corresponding mass difference falls into the top b most frequent intervals into one group (we used $b = 10$), and the remaining into a second group.

It is important to note that this grouping satisfies Assumption 2 required for theoretical FDR control. More specifically, knowing the mass differences between the matched peptides and the precursors (along with the winning scores and labels of correctly matched peptides), we still expect the *incorrect* matches to be independent and equally likely to be a target or a decoy win.

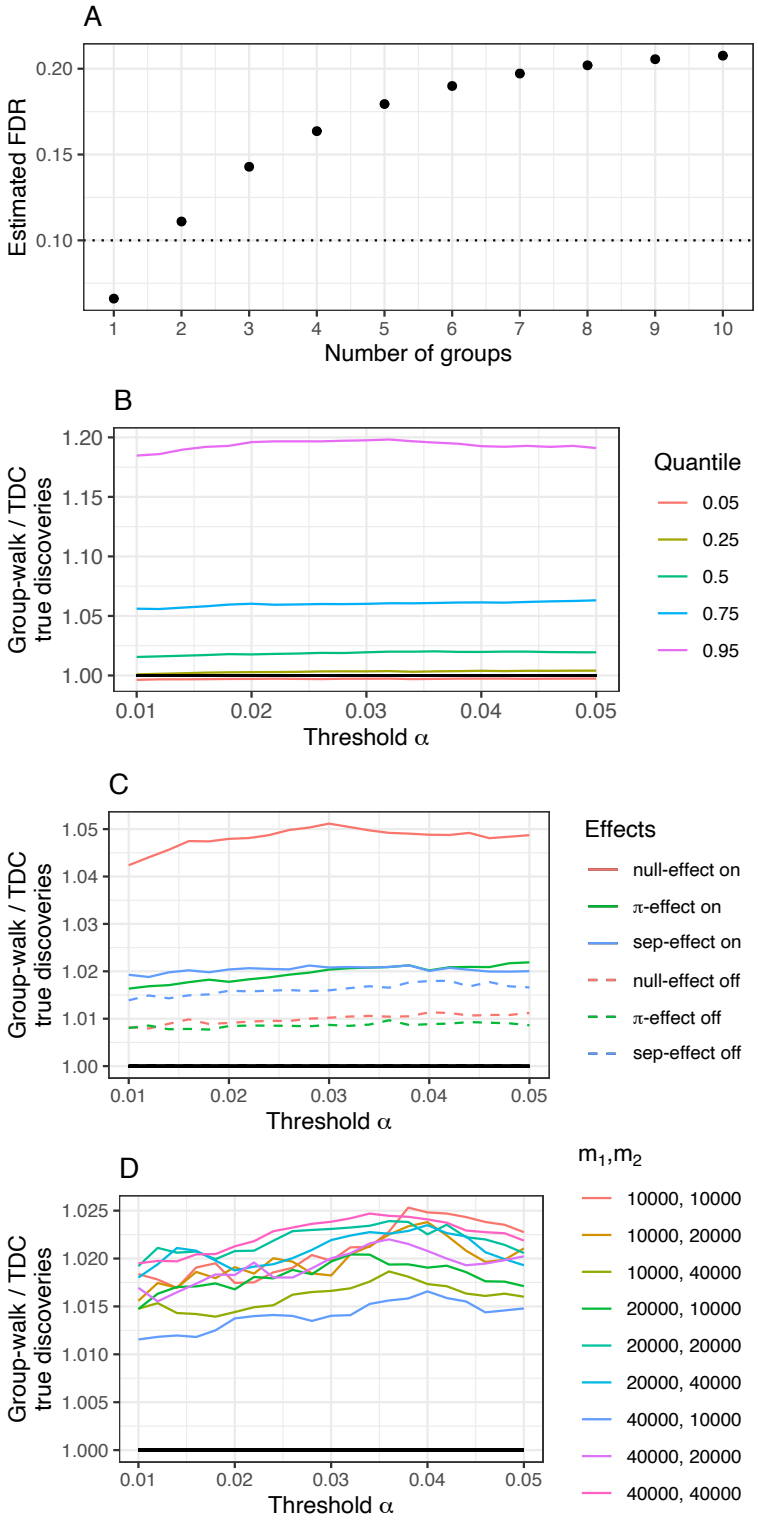


Figure 1: (A) The estimated FDR as a function of the number of groups. TDC is applied separately to each group and the overall FDR is averaged over $100,000$ runs where the setup for each run is described in Supplementary Section 6.3. The estimated standard errors for each estimated FDR are too small to be seen in the figure. (B-D) Quantiles of the average ratio between the true discoveries from Group-walk and TDC in simulated spectrum-ID data. The average is taken over 40 runs for each of the 628 parameter combinations. (B) some primary quantiles of the 628 averaged ratios. (C) the median of the averaged ratios when we split the parameter combinations into two equal sized groups according to whether each of the three considered group effects is present (solid) or absent (dashed). (D) the median of the averaged ratio for all parameter combinations with the prescribed groups sizes (m_1, m_2).

4 Results

4.1 Separate TDC analysis does not generally control the FDR

As motivation for creating our Group-walk procedure, we begin by demonstrating that naively applying TDC separately to each group can fail to control the FDR. In Supplementary Section 6.3 we provide the details of a simulation we carried out to illustrate this effect. Specifically, our simulations demonstrate that the actual FDR in this setup (estimated from 100K randomly drawn sets) rapidly increases beyond the FDR threshold as we increase the number of groups (Figure 1A). For example, with just three groups the actual FDR of 0.143 is already >40% higher than the nominal $\alpha = 0.1$, and with eight groups it is >100% higher (at 0.202).

4.2 Simulation: the spectrum-ID problem

In Figure 1B we look at quantiles of the average Group-walk-to-TDC discovery ratio using our simulated spectrum-ID scenario. The ratio's average is taken with respect to the 40 runs, and the quantiles are with respect to the 628 parameter combinations. The plot suggests that in the context of our simulations at 1% FDR we observe a typical gain of about 1.6% more discoveries using Group-walk (median, green). That gain often grows to over 5.6% (upper quartile, blue) and occasionally to more than 18% (0.95 quantile, violet). Notably, there is typically very little loss in using Group-walk: the lower quartile still shows a very marginal gain (0.1%), and the 0.05 quantile corresponds to less than 0.4% fewer Group-walk discoveries.

In panel C we examine how Group-walk takes advantage of the three types of group effects by looking at the median of the (average) Group-walk-to-TDC discovery ratio when the data is split three times into two sets (each made of 324 parameter combinations) according to whether or not the corresponding group effect was present in the data. While Group-walk benefits from all three group-effects, the null-group effect seems to be the most prominent one in our setup.

Finally, in panel D we look at the median of the (average) Group-walk-to-TDC discovery ratio when the data is split according to the size of the two present groups (m_i). Notably, Group-walk performs relatively better when $m_1 = m_2$, and its advantage decreases as the difference between the group sizes increases (blue and olive curves). This makes sense intuitively if you think about a limiting case where, say $m_1 \ll m_2$, in which case there will be little difference between applying TDC to the combined list and to group 2 by itself.

4.3 Charge and peptide length in low-resolution MS/MS data

It is well understood that charge contributes to the lack of calibration for some score functions, like SEQUEST's XCorr score [20]. Additionally, we observe a similar behaviour relating to peptide length. Here we demonstrate how the usual TDC procedure can be improved upon when applying Group-walk to such uncalibrated scores. The peptides were partitioned into 2 or more groups as described in Section 3.3.1, and again we compared the average ratio of Group-walk to TDC discoveries.

Figure 2A shows that for both charge- (2 or more groups) and length-based (2 groups) Group-walk typically offers a modest gain in power across all FDR thresholds in the considered range ($[0.01, 0.05]$). Specifically, at the 1% FDR level the median power, when grouping by charge,

increases by charge 4%, and 3.6% when grouping by length. Moreover, considering the upper quartiles we see that 25% of the time Group-walk reports at least 19% more discoveries when grouping by charge and at least 11% more when grouping by length. Examining the lower quartiles we find Group-walk still reports more discoveries than TDC: 1.4% more when grouping by charge and 0.5% more when grouping by length.

It is worth noting that when our peptide length-based partition uses 3 groups the results are quite similar to when we partition into 2 groups as reported above (Supplementary Figure 3).

4.4 Open search

Comparing the performance of Group-walk and TDC in open search we see that regardless of whether the score is calibrated (Tailor) or not (XCorr) Group-walk typically delivers a fair bit more discoveries than TDC across the entire range of considered FDR thresholds (Figure 2B).

In particular, at the 1% FDR threshold Group-walk typically reports 12% more discoveries than TDC when using Tailor (solid red) and 24% more when using XCorr (solid blue). Moreover, considering the upper quartiles we see that 25% of the time Group-walk reports at least 20% more discoveries using Tailor and at least 50% more using XCorr. Examining the lower quartiles we find Group-walk still reports more discoveries: 6% more with Tailor and 15% more using XCorr.

Considering that Tailor scores are relatively well calibrated, this suggests that, as predicted, Group-walk is leveraging the fact that different groups exhibit a different proportion of correct peptide identifications and not just a difference in calibration between groups.

5 Discussion

TDC is the common approach to FDR control in the analysis of tandem mass spectrometry data. The implicit assumption that TDC relies on is that the data is essentially homogeneous. However, in practice the data is often made of groups of hypotheses (PSMs or peptides) with distinct group-wise characteristics. Applying TDC while ignoring this structure creates imbalanced lists of discoveries when viewed with the correct group structure in mind: the discoveries from some groups will be overly conservative while for others they will be overly liberal.

While applying a separate TDC analysis to each group addresses this problem, it might fail to control the FDR in the combined list of discoveries, and the examples we provide show that this failure can be substantial. Instead, we offer Group-walk, a novel procedure that shares some of the characteristics of separate analyses while still controlling the combined FDR. It is worth noting that, particularly when using the common 1% FDR threshold, Group-walk can possibly deliver even more discoveries than the, non-rigorous, separate analyses. Indeed, the +1 "penalty" in the estimated FDR is applied to each group in the latter case, but Group-walk uses a combined FDR estimate so in its case there is a single penalty.

Group-walk can be considered as a specialization of a slightly modified version of AdaPT, which is a recently published general framework for the problem of multiple hypothesis testing with generic side information [22]. Note that as a general framework AdaPT does not specify how to advance its analogue of what we call the front, leaving it instead as a generic function called `UPDATE` (a specialized version of `UPDATE` is offered in [22], but it only applies for the classical

p-value setup rather than the competition-based setup we study here). Still, showing that Group-walk is consistent with the general requirements of AdaPT's `UPDATE` allows us to establish that Group-walk controls the FDR in the finite sample setting.

To be effective, Group-walk requires typically at least K hypotheses in each group (throughout this work we used $K = 40$). Some other approaches in the spirit of separate TDC analyses have been offered for cases where some groups are small, but no finite-sample FDR control was established (e.g. [13]).

We demonstrated using simulated data how Group-walk can benefit from the varying types of group structures: null-effect, separation-effect, and π-effect. It should be noted that both the separation- and π-effects have general trends that affect how well Group-walk can take advantage of them. For example, when μ is large enough so the alternative distribution becomes sufficiently separated from the null distribution, the problem becomes trivial, and TDC will do just as well as Group-walk even if the data exhibits a significant group π-effect.

This raises the important question of how do we know when should we use Group-walk. For now we leave this for future research, noting that our results show that the penalty of using Group-walk in the absence of group structure seems rather marginal, particularly when compared with the significant gains it potentially offers.

A particularly appealing property of Group-walk is that it frees us from worrying about whether our score is sufficiently calibrated with respect to some property of the data, such as the precursor's charge state, or put differently, whether our data exhibits a group null-effect. Indeed, as we saw in the discovery gains Group-walk offered when grouping the low-resolution PRIDE datasets by charge, all we have to do is partition the data into groups according to the considered feature and apply Group-walk which will then take advantage of the null-effect.

While we also saw gains when analyzing the PRIDE data sets according to the peptide length and to the mass difference in open search, we note that it is not always obvious how to partition the hypotheses into groups. This is even more the case if the feature we consider is not discrete. This suggests another promising direction for future research, which is to go back to AdaPT's view of having more general side information rather than just a group ID. Note that again, it is not obvious what needs to be done, as such an approach would still require specializing AdaPT's generic `UPDATE` function for our competition-based setup.

While TDC has been in use for some time now in computational mass spectrometry, Barber and Candès recently used the same principle in their knockoff+ procedure to control the FDR in feature selection in a classical linear regression model [2]. Moreover, following their work and the introduction of a more flexible formulation of the variable selection problem in the model-X framework of Candès et al. [5], competition-based FDR control has gained a lot of interest in the statistical and machine learning communities. Knockoffs and decoys serve the same purpose in competition-based FDR control. In particular, Group-walk is applicable to the knockoff context just as well as it is to TDC. That said, in the kind of feature selection problems considered by Barber and Candès the fraction of false nulls (relevant features) is often too low to effectively employ Group-walk.

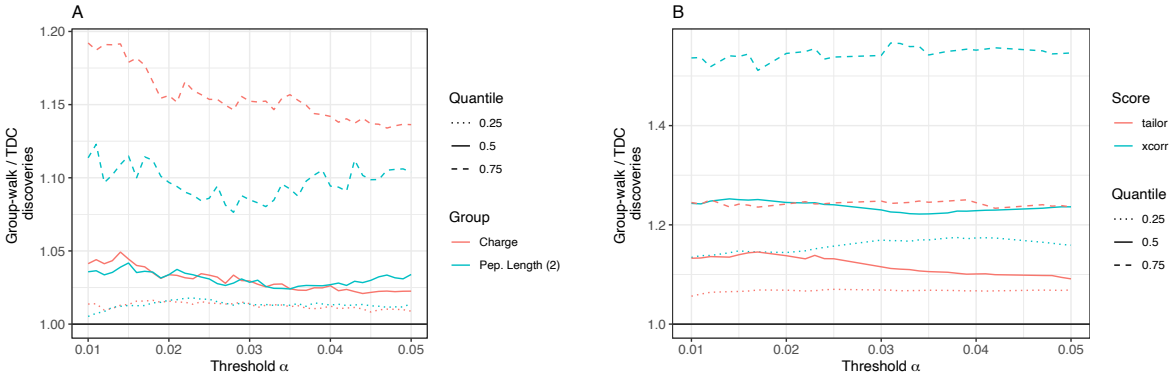


Figure 2: **The quartiles of the ratios of Group-walk to TDC discoveries in real data.** (A) Group-walk was applied to partitions of the peptides according to the charge state of the spectrum using 2 or more groups (red), as well as into 2 groups according to the length of the matching peptide (blue). (B) Results from open-searches using the XCorr score (blue) and Tailor score (red). The peptides are partitioned into two groups: the peptides whose mass difference to the spectrum's precursor mass falls into the top $b = 10$ most abundant bins in one group, and the rest in the second group.

References

[1] P. R. Baker, K. F. Medzihradszky, and R. J. Chalkley. "Improving software performance for peptide electron transfer dissociation data analysis by implementation of charge state-and sequence-dependent scoring". In: *Molecular & Cellular Proteomics* 9.9 (2010), pp. 1795–1803.

[2] R. F. Barber and E. J. Candès. "Controlling the false discovery rate via knockoffs". In: *The Annals of Statistics* 43.5 (2015), pp. 2055–2085.

[3] Y. Benjamini and Y. Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[4] B. Bogdanow, H. Zauber, and M. Selbach. "Systematic errors in peptide and protein identification and quantification by modified peptides". In: *Molecular & Cellular Proteomics* 15.8 (2016), pp. 2791–2801.

[5] E. Candes, Y. Fan, L. Janson, and J. Lv. "Panning for gold:'model-X'knockoffs for high dimensional controlled variable selection". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.3 (2018), pp. 551–577.

[6] F. R. Cerqueira, A. Graber, B. Schwikowski, and C. Baumgartner. "MUDE: a new approach for optimizing sensitivity in the target-decoy search strategy for large-scale peptide/protein identification". In: *Journal of proteome research* 9.5 (2010), pp. 2265–2277.

[7] B. J. Diament and W. S. Noble. "Faster SEQUEST searching for peptide identification from tandem mass spectra". In: *Journal of proteome research* 10.9 (2011), pp. 3871–3879.

[8] B. Efron. "Simultaneous inference: When should hypothesis testing problems be combined?" In: *The annals of applied statistics* 2.1 (2008), pp. 197–223.

[9] J. E. Elias and S. P. Gygi. "Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry". In: *Nature methods* 4.3 (2007), pp. 207–214.

[10] J. E. Elias and S. P. Gygi. "Target-decoy search strategy for mass spectrometry-based proteomics". In: *Proteome bioinformatics*. Springer, 2010, pp. 55–71.

[11] K. Emery, S. Hasam, W. S. Noble, and U. Keich. "Multiple competition-based FDR control and its application to peptide detection". In: *International Conference on Research in Computational Molecular Biology*. Springer. 2020, pp. 54–71.

[12] J. K. Eng, A. L. McCormack, and J. R. Yates. "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database". In: *Journal of the american society for mass spectrometry* 5.11 (1994), pp. 976–989.

[13] Y. Fu and X. Qian. "Transferred subgroup false discovery rate for rare post-translational modifications detected by mass spectrometry". In: *Molecular & Cellular Proteomics* 13.5 (2014), pp. 1359–1368.

[14] V. Granholm, J. F. Navarro, W. S. Noble, and L. Käll. "Determining the calibration of confidence estimation procedures for unique peptides in shotgun proteomics". In: *Journal of proteomics* 80 (2013), pp. 123–131.

[15] K. He, Y. Fu, W.-F. Zeng, L. Luo, H. Chi, C. Liu, L.-Y. Qing, R.-X. Sun, and S.-M. He. "A theoretical foundation of the target-decoy search strategy for false discovery rate control in proteomics". In: *arXiv preprint arXiv:1501.00537* (2015).

[16] P. Hernandez, M. Müller, and R. D. Appel. "Automated protein identification by tandem mass spectrometry: issues and strategies". In: *Mass spectrometry reviews* 25.2 (2006), pp. 235–254.

[17] K. Jeong, S. Kim, and N. Bandeira. "False discovery rates in spectral identification". In: *BMC bioinformatics* 13.16 (2012), pp. 1–15.

[18] U. Keich and W. Noble. "Controlling the FDR in imperfect database matches applied to tandem mass spectrum identification". In: *Journal of the American Statistical Association* 10.01621459.2017 (2017), p. 1375931.

[19] U. Keich, A. Kertesz-Farkas, and W. S. Noble. "Improved false discovery rate estimation procedure for shotgun proteomics". In: *Journal of proteome research* 14.8 (2015), pp. 3148–3161.

[20] U. Keich and W. S. Noble. "On the importance of well-calibrated scores for identifying shotgun proteomics spectra". In: *Journal of proteome research* 14.2 (2015), pp. 1147–1160.

[21] U. Keich and W. S. Noble. "Progressive calibration and averaging for tandem mass spectrometry statistical confidence estimation: Why settle for a single decoy?" In: *International Conference on Research in Computational Molecular Biology*. Springer. 2017, pp. 99–116.

[22] L. Lei and W. Fithian. "AdaPT: an interactive procedure for multiple testing with side information". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.4 (2018), pp. 649–679.

[23] D. H. May, K. Tamura, and W. S. Noble. "Detecting modifications in proteomics experiments with Param-Medic". In: *Journal of proteome research* 18.4 (2019), pp. 1902–1906.

[24] D. H. May, K. Tamura, and W. S. Noble. "Param-Medic: A tool for improving MS/MS database search yield by optimizing parameter settings". In: *Journal of proteome research* 16.4 (2017), pp. 1817–1824.

[25] A. I. Nesvizhskii. “A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics”. In: *Journal of proteomics* 73.11 (2010), pp. 2092–2123.

[26] W. S. Noble and M. J. MacCoss. “Computational and statistical analysis of protein mass spectrometry data”. In: *PLOS computational biology* 8.1 (2012), e1002296.

[27] C. Y. Park, A. A. Klammer, L. Kall, M. J. MacCoss, and W. S. Noble. “Rapid and accurate peptide identification from tandem mass spectra”. In: *Journal of proteome research* 7.7 (2008), pp. 3022–3027.

[28] Y. Perez-Riverol, A. Csordas, J. Bai, M. Bernal-Llinares, S. Hewapathirana, D. J. Kundu, A. Inuganti, J. Griss, G. Mayer, M. Eisenacher, et al. “The PRIDE database and related tools and resources in 2019: improving support for quantification data”. In: *Nucleic acids research* 47.D1 (2019), pp. D442–D450.

[29] J. D. Storey. “A direct approach to false discovery rates”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.3 (2002), pp. 479–498.

[30] P. Sulimov and A. Kertész-Farkas. “Tailor: A nonparametric and rapid score calibration method for database search-based peptide identification in shotgun proteomics”. In: *Journal of proteome research* 19.4 (2020), pp. 1481–1490.

[31] M. The, A. Tasnim, and L. Käll. “How to talk about protein-level false discovery rates in shotgun proteomics”. In: *Proteomics* 16.18 (2016), pp. 2461–2469.

6 Supplementary Material

Algorithm 1: Group-walk

Input: the FDR threshold $\alpha \in (0,1)$;
a window size $K \in \mathbb{Z}^+$;
a list of m_g labels $L_i^g = \pm 1$ for each group $g \in \{1,\dots,n_g\}$ where the indices are sorted so that the winning scores W_i^g in each group are increasing: $W_1^g \le W_2^g \le \cdots \le W_{m_g}^g$;
Output: A discovery list $\bigcup_g \{i \ge k_g \mid L_i^g = 1\}$
Initialize the threshold indices $k_g := 1$ for $g \in \{1,\dots,n_g\}$;
while *there exists* $g \in \{1,\dots,n_g\}$ *such that* $k_g \le m_g$ **do**
  $F\hat{D}P := \frac{1+\sum_g D_g(k_g)}{1 \vee \sum_g R_g(k_g)}$;
  if $F\hat{D}P \le \alpha$ **then**
    break;
  else
    if *there exists* $g \in \{1,\dots,n_g\}$ *such that* $k_g \le K$ **then**
      Randomly select l from $\{l \mid k_l = \min_g k_g\}$;
      and set $k_l := k_l + 1$;
    else
      for $g \in \{1,\dots,n_g\}$ **do**
        $w_g := D_g(k_g - K) - D_g(k_g)$
      end
      Randomly select l from $\{l \mid w_l = \max_g w_g\}$; and set $k_l := k_l + 1$;
    end
  end
end
return $\bigcup_g \{i \ge k_g \mid L_i^g = 1\}$

6.1 Simulation: the spectrum-ID model in detail

We associate with each (virtual) "spectrum", σ_i, a score X_i as follows. If σ_i is native (cf. Section 2) then X_i models the match between σ_i and the peptide that generated it, so we draw $X_i \sim 1 - \text{Beta}(a,b)$, where we fix $a = 0.05$ and we consider b as the *separation parameter* between true and false null scores. Otherwise, σ_i is foreign so we set X_i to the lowest possible score which is $X_i = 0$ here. All the draws described here are done independently.

In addition, assuming that associated with each spectrum are n candidate peptides (those within the tolerance of the measured precursor mass, here we used $n = 100$), we associate with each native spectrum a score Y_i corresponding to the best match to σ_i among the remaining $n-1$ candidate target peptides, whereas for a foreign spectrum Y_i corresponds to the best match among all n candidates. As all these target matches are considered random we model Y_i as the maximum of $n-1$ (or n) Unif$(0,1)$ scores, i.e., we sample $Y_i \sim \text{Beta}(n-1,1)$ ($Y_i \sim \text{Beta}(n,1)$). The target PSM score is then $Z_i = X_i \vee Y_i$.

We also associate with each spectrum a decoy PSM score, $\tilde{Z}_i$, modeling the best match to σ_i among n candidate decoy peptides. Consistent with our Unif $(0, 1)$ model of a random match we therefore draw $\tilde{Z}_i \sim$ Beta $(n, 1)$. It follows that Assumptions 1 and 2 are only approximately valid under this particular model: for native spectra there is a slightly larger chance a true null will be a decoy win. As this creates a slightly conservative bias it is not concerning as far as FDR control. Additionally, if σ_i is foreign then $Z_i \sim$ Beta $(n, 1)$ so for a foreign spectrum target and decoy wins are equally likely.

To make the score uncalibrated, we define the winning scores as $W_i^g := Q\left(\max\left(Z_i^g, \tilde{Z}_i^g\right), \mu_i^g, \beta_i^g\right)$, where $Q(\cdot, \mu, \beta)$ is the quantile function of the Gumbel distribution with parameters μ, β. The latter location and scale parameters are sampled from a population of such parameters generated by fitting Gumbel distributions to PSM scores generated from real yeast data, with a presumed charge state of 2 or 3. Further details can be found in [21].

Our simulated dataset of PSM scores consists of two groups ($n_g = 2$) where each group has m_g null hypotheses ("the PSM is incorrect, i.e., $Y_i \vee \tilde{Z}_i > X_i$"). A fraction of π_g of the target scores (Z_i^g) within each group, corresponding to the native spectra, are sampled as described above from the alternative distribution ($X_i \sim 1 -$ Beta (a, b)), and the remaining target scores, as well as all the decoy scores ($\tilde{Z}_i^g$), are sampled from the null distribution. All the draws are made independently.

We varied the group sizes $m_1, m_2 \in \{1\text{e}4, 2\text{e}4, 4\text{e}4\}$, the separation parameters $b_1, b_2 \in \{5, 50\}$ and the proportion of native spectra in each group $\pi_1, \pi_2 \in \{0.3, 0.6, 0.9\}$. In addition, to study the group null-effect we create two different types of datasets. In the first type the Gumbel parameters that induce the non-calibrated feature of the data, $\{\mu_i^g, \beta_i^g\}_{i,g}$ are all randomly drawn with replacement from the same population of location and scale parameters. Any such homogeneous dataset is uncalibrated but there is no group null-effect. The second data type is generated where the Gumbel parameters of the first group are drawn from the population of parameters that were originally fitted to PSMs with a presumed charge state of 3, while the Gumbel parameters of the second group are drawn from the set fitted to a presumed charge of 2. As noted in the Introduction, different charge states are known to be associated with a group null-effect.

Again, we generated 40 samples for each combination of parameters and examined the average ratio of Group-walk ($K = 40$) to TDC true discoveries number for each of the 628 distinct parameter combinations.

6.2 Supplementary Analysis of mass spectra

The following data sets that we used were sampled from projects available from the Proteomics Identifications Database (PRIDE) [28]. One spectrum file was randomly selected and downloaded from each of 14 randomly selected projects such that 7 were high-resolution spectra files and 7 were low-resolution spectra files. Each project was submitted to PRIDE no earlier than 2018. The 7 low-resolution spectrum files came from PXD012522, PXD013482, PXD014774, PXD015485, PXD017622, PXD022092 and PXD027867. The 7 high-resolution spectrum files came from PXD006856, PXD008996, PXD012611, PXD013274, PXD019186, PXD022257 and PXD025130. We also downloaded the target peptide database from PRIDE except for human data where we used the UniProt database (UP000005640) downloaded on 9/11/2021.

We used the Crux toolkit [27] to assign peptides to spectra as follows. First, for each PRIDE dataset, the Crux command tide-index [7] was used to construct the concatenated target-decoy

peptide database with the (Param-Medic) option `--auto-modifications-spectra` to account for variable modifications [23]. To reduce the random effects due to the random nature of our decoys we repeated the database construction 20 times with different seeds to create varied shuffled decoy peptides, which our results were aggregated over. Tide-search [7], with parameter settings optimized by Param-Medic [24], was then used to generate the PSM scores by searching the concatenated target-decoy peptide database with the exact options as follows:

- When grouping by the precursor charge state and the peptide length we used the following options in tide-search: `--auto-precursor-window warn`, `--auto-mz-bin -width warn`, `--concat T`.
- In the open search analysis we used: `--auto-mz-bin-width warn`, `--precursor-window-type mass`, `--precursor-window 100`, `--use-tailor-calibration T`, `--concat T`. In addition, we used both the XCorr score, and a calibrated version of XCorr score called *Tailor* score [30]. Note, that the option `--use-tailor-calibration T` returns both scores.

We restricted the analysis of grouping by charge and peptide length to the 7 low-resolution datasets whereas the open search analysis was done on all 14 datasets. To discretize the mass difference in the latter analysis we used multiples of the canonical value $\lambda = 1.0005079$ Da, as the breaks that define the intervals/bins of mass differences: $\{\ldots, -3\lambda/2, -\lambda/2, \lambda/2, 3\lambda/2, \ldots\}$.

6.3 Separate TDC analysis does not generally control the FDR

It is not difficult to find examples showing separate TDC analyses fail to control the overall FDR in a finite sample setting. For example, consider the following setup for two groups. For $g \in \{1, 2\}$, let m_g be the number of hypotheses in group g and π_g be the fraction of false nulls. We set $m_1 = 100$, $m_2 = 1000$ and $\pi_1 = \pi_2 = 0.1$. With Z_i^g and $\tilde{Z}_i^g$ denoting the target and decoy scores of the ith hypothesis in group g, respectively, we draw all scores independently as follows: $Z_i^g \sim \mathcal{N}(0, 1)$ if Z_i^g comes from a null hypothesis, and $Z_i^g \sim \mathcal{N}(1.5, 1)$ if Z_i^g comes from a false null hypothesis. The decoy score $\tilde{Z}_i^g$ is distributed $\mathcal{N}(0, 1)$. For each hypothesis, we determine the winning scores and labels as defined in Section 2 and apply the TDC procedure separately, resulting in two discovery lists. Repeating the above simulation for 100,000 runs at an FDR threshold of $\alpha = 0.25$, we found the *combined* empirical FDR was 0.258 with an estimated standard error of 0.0005; 16 standard errors above the threshold.

While the above failure to control the FDR is statistically significant, it is rather negligible in terms of the magnitude of the excess FDR. However, by increasing the number of groups we can readily construct examples where the actual FDR of separate TDC analyses is substantially higher than the nominal threshold.

Specifically, we consider n_g independent and identically distributed groups where each group is made of 22 hypotheses, 6 of which are true null hypotheses and 16 are false nulls. Moreover, we assume the range of possible scores for each hypothesis is such that the order of the hypotheses is deterministic. Specifically, when sorted in increasing order, first come 3 of the null hypotheses followed by 8 of the false nulls, then the remaining 3 nulls followed by the remaining 8 false nulls. For each false null hypothesis we assume the target always wins, whereas the labels of all true null hypotheses are independently and uniformly drawn.

We repeatedly and independently drew the sequence of labels for each of the n_g groups, applied TDC separately to each group at level $\alpha = 0.1$ and noted the actual FDP in the combined list of discoveries. We estimated the FDR in the combined list by averaging the actual FDP over 100,000 sets of draws for each value of n_g, as we increased it from 1 to 10. These estimated FDR values given in Figure 1A demonstrate how rapidly the excess FDR grows in this case, indicating that even for a small number of groups, separate analyses can provide a poor method of FDR control. For example, with just 3 groups at 0.143 the actual FDR is already over 40% higher than the nominal $\alpha = 0.1$ and with 8 groups at 0.202 it is more than 100% higher.

Let V_g denote the number of true null target wins discovered by TDC in group g, and let R_g be the corresponding number of all discoveries. It is not hard to show that as $n_g \to \infty$ the FDR of the separate TDC analyses will converge to the mFDR of each group which is defined as $E(V_g)/E(R_g)$. Given the size of the problem, it is straightforward to evaluate the ratio of expectations and find that in this case mFDR $= 1.15625/5.40625 \approx 0.214$; significantly over the nominal 0.1.

6.4 Group-walk is a specialized version of AdaPT

We establish Group-walk's FDR control by showing it is a special case of (a slightly modified version of) the Adaptive p-value Thresholding (AdaPT) framework of Lei and Fithian. AdaPT was developed to address the problem where associated with each hypothesis is a p-value p_i, as well as a predictor x_i encoding contextual information about the hypothesis [22]. In our case the p-value is the 1-bit p-value (1/2 for a target win and 1 for a decoy win), and $x_i = (W_i, G_i)$.

AdaPT progresses through iteratively defined rejection thresholds, $s_t(x)$, that are analogous to Group-walk's front. Indeed, s_t describes the front position at step t of the procedure and for each fixed x the rejection thresholds are monotone decreasing in t. To show that Group-walk is an adaptation of AdaPT we introduce t as the same kind of step counter to Group-walk, and we define $s_t(x)$ in that case as 0 for all the hypotheses that Group-walk's front has passed over by step t, and as 1/2 for all the other hypotheses: with $\mathbf{k}_t = (k_1, \ldots, k_{n_g})$ denoting Group-walk's front at step t, $s_t(W_i^g, g) = 0$ for all $1 \le i < k_g$ and $s_t(W_i^g, g) = 1/2$ for all $k_g \le i \le m_g$.

With this definition of s_t, and provided we slightly modify the definition of AdaPT's A_t to use a strict inequality:

$$A_t = |\{i \,:\, p_i > 1 - s_t(x_i)\}|, \tag{1}$$

A_t coincides with Group-walk's $\sum_{g=1}^{n_g} D_g(k_g)$. Similarly AdaPT's $R_t = |\{i \,:\, p_i \le s_t(x_i)\}|$ coincides with Group-walk's $\sum_{g=1}^{n_g} R_g(k_g)$. It follows that with this modification their estimated FDR, $(1 + A_t)/(1 \vee R_t)$, coincides with Group-walk's and hence AdaPT's rejection criterion coincides with Group-walk's in this case.

AdaPT does not specify how to adjust its front s_t except to note that the update needs to be a function of A_t, R_t, s_t as well as of all the contextual information ((W_i, G_i) in our case), and all its $\tilde{p}_{t,i}$ which are defined in (2) of [22]. Here is where we need to make our second change to AdaPT's original definitions. Specifically we change the second inequality in their definition (2) to a weak one:

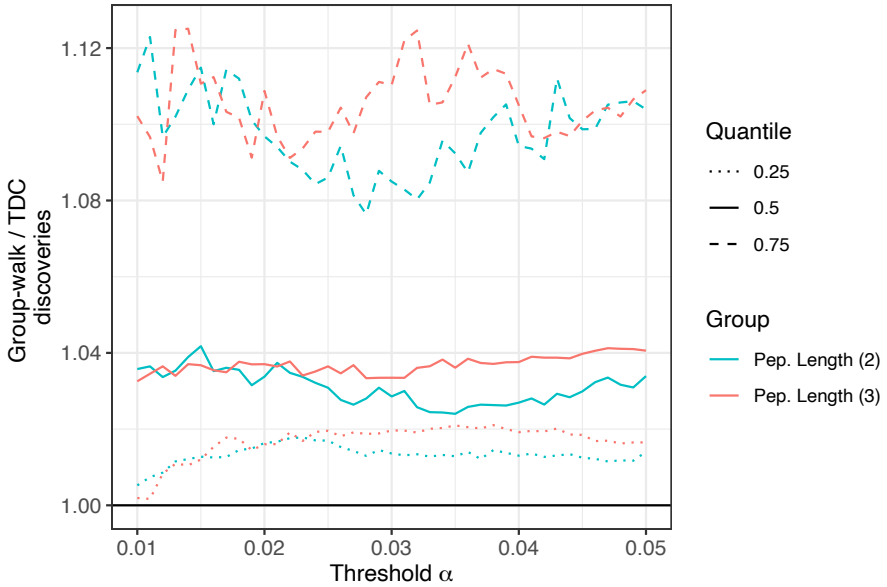
$$\tilde{p}_{t,i} = \begin{cases} p_i & s_t(x_i) < p_i \le 1 - s_t(x_i) \\ \{p_i, 1 - p_i\} & \text{otherwise} \end{cases}. \tag{2}$$

With the latter change it is easy to see $\tilde{p}_{t,i}$ resolves the target/decoy label, L_i, for any hypothesis that the front has passed over by step t and the label remains unresolved otherwise. It follows that Group-walk's update procedure, which moves the front by moving one of the indices forward, $k_g := k_g + 1$, based on the information gathered from the labels of hypotheses the front already passed over, as well as on the scores and the group assignments is consistent with AdaPT's generic update protocol.

Finally, we verified that Theorem 1 of [22], which proves that AdaPT controls the FDR in the finite sample setting, carries through with our two modifications above: (1) and (2). It follows that under Assumption 2 Group-walk also controls the FDR in the finite sample setting.

6.5 Supplementary Figures

Partitioning into 2 vs. 3 groups based on peptide length



s

Figure 3: The figure presents the quartiles of the ratios of Group-walk to TDC discoveries when partitioning the peptides via the the length of the matching peptide into either 2 (blue) or 3 (red) groups.