# APIR: a universal FDR-control framework for boosting peptide identification power by aggregating multiple proteomics database search algorithms

Yiling Elaine Chen[1,†], Kyla Woyshner[2,†], MeiLu McDermott[2,3,†], Antigoni Manousopoulou[2], Scott Ficarro[4], Jarrod Marto[4], Xinzhou Ge[1], Leo David Wang[2,5,*], and Jingyi Jessica Li[1,6,7,8,9,*]

[1]Department of Statistics, University of California, Los Angeles, CA 90095
[2]Department of Immuno-Oncology, Beckman Research Institute, City of Hope National Medical Center, Duarte CA 91010
[3]The Quantitative and Computational Biology section, University of Southern California, Los Angeles, CA 90089
[4]Blais Proteomics Center, Dana-Farber Cancer Institute, Boston MA 02215
[5]Department of Pediatrics, Beckman Research Institute, City of Hope National Medical Center, Duarte CA 91010
[6]Interdepartmental Program in Bioinformatics, University of California, Los Angeles, CA 90095
[7]Department of Human Genetics, University of California, Los Angeles, CA 90095
[8]Department of Computational Medicine, University of California, Los Angeles, CA 90095
[9]Department of Biostatistics, University of California, Los Angeles, CA 90095
[†]These authors contributed equally to this work.
[*]To whom correspondence should be addressed. Email: jli@stat.ucla.edu (J.J.L.); leo.wang@coh.org (L.D.W.)

**Abstract**

Advances in mass spectrometry (MS) have enabled high-throughput analysis of proteomes in biological systems. The state-of-the-art MS data analysis relies on database search algorithms to quantify proteins by identifying peptide-spectrum matches (PSMs), which convert mass spectra to peptide sequences. Different database search algorithms use distinct search strategies and thus may identify unique PSMs. However, no existing approaches can aggregate all user-specified database search algorithms with guaranteed control on the false discovery rate (FDR) and guaranteed increase in the identified peptides. To fill in this gap, we propose a statistical framework, Aggregation of Peptide Identification Results (APIR), that is universally compatible with all database search algorithms. Notably, under a target FDR threshold, APIR is guaranteed to identify at least as many, if not more, peptides as individual database search algorithms do. Evaluation of APIR on a complex protein standard shows that APIR outpowers individual database search algorithms and guarantees the FDR control. Real data studies show that APIR can identify disease-related proteins and post-translational modifications missed by some individual database search algorithms. Note that the APIR framework is easily extendable to aggregating discoveries made by multiple algorithms in other high-throughput biomedical data analysis, e.g., differential gene expression analysis on RNA sequencing data.

## Introduction

Proteomics studies have discovered essential roles of proteins in complex disease such as neurodegenerative disease [1] and cancer [2–4]. These studies have demonstrated the potential of using proteomics to identify clinical biomarkers for disease diagnosis and therapeutic targets for disease treatment. In recent years, proteomics analytical technologies, particularly tandem mass spectrometry (MS)-based shotgun proteomics, have advanced immensely, thus enabling high-throughput identification and quantification of proteins in biological samples. Compared to prior technologies, shotgun proteomics has

simplified sample preparation and protein separation, reduced time and cost, and saved procedures that may result in sample degradation and loss [5]. In a typical shotgun proteomics experiment, a protein mixture is first enzymatically digested into peptides, i.e., short amino acid chains up to approximately $40$-residue long; the resulting peptide mixture is then separated and measured by tandem MS into tens of thousands of mass spectra. Each mass spectrum encodes the chemical composition of a peptide; thus, the spectrum can be used to identify the peptide's amino acid sequence and post-translational modifications, as well as to quantify the peptide's abundance with additional weight information (Fig. 1a).

Since the development of shotgun proteomics, numerous database search algorithms have been developed to automatically convert mass spectra into peptide sequences. Popular database search algorithms include SEQUEST [6], Mascot [7], MaxQuant [8], Byonic [9], and MS-GF+ [10], among many others. A database search algorithm takes as input the mass spectra from a shotgun proteomics experiment and a protein database that contains known protein sequences. For each mass spectrum, the algorithm identifies the best matching peptide sequence, a subsequence of a protein sequence, from the database; we call this process "peptide identification," whose result is a "peptide-spectrum match" (PSM). However, due to data imperfection (such as low-quality mass spectra, data processing mistakes, and protein database incompleteness), the identified PSMs often consist of many false PSMs, causing issues in the downstream system-wide identification and quantification of proteins [11].

To ensure the accuracy of PSMs, the false discovery rate (FDR) has been used as the most popular statistical criterion [12–21]. Technically, the FDR is defined as the expected proportion of false PSMs among the identified PSMs; in other words, a small FDR indicates good accuracy of PSMs. However, controlling the FDR is only one side of the story. Because shotgun proteomics experiments are costly, a common goal of database search algorithms is to identify as many true PSMs as possible to maximize the experimental output, in other words, to maximize the identification power given a target, user-specified FDR threshold (e.g., $1\%$ or $5\%$).

It has been observed that, with the same input mass spectra and FDR threshold, different database search algorithms often find largely distinct sets of PSMs [22–26]. There are two possible reasons for this phenomenon. One is that different algorithms find different sets of true PSMs by design. The other is that some algorithms have identified excessive false PSMs due to failed FDR control. It is important to disentangle these two reasons because, if the former is true, we may aggregate the distinct sets of identified PSMs to increase the peptide identification power; otherwise, before performing the aggregation, we must refine the outputs of the algorithms that have failed to control the FDR.

To investigate this question, in this study, we generated the first publicly available complex proteomics standard dataset from *Pyrococcus Furiosus (Pfu)* to approach the dynamic range of a typical proteomics experiment. We used this standard dataset for benchmarking five database search algorithms: SEQUEST [6], Mascot [7], MaxQuant [8], Byonic [9], and MS-GF+ [10]. Our results confirm that, while some database search algorithms fail to control the FDR, they are indeed designed to capture unique sets of PSMs (see Results for details). This result justifies the need for algorithm aggregation to boost the power of identifying peptides from shotgun proteomics data.

In the proteomics field, existing aggregation methods include Scaffold [25], MSblender [21], FDRAnalysis [27], iProphet [20], ConsensusID [19], and PepArML [12]. Among these six methods, except FDRAnalysis that has been shown infeasible for high-throughput proteomics [22], the rest have at least one of the two major drawbacks: (1) limited compatibility with database search algorithms and (2) lack of guarantee for identifying more peptides under the same FDR threshold. For the first drawback, except ConsensusID, the other five aggregation methods unanimously limit the choices of database search algorithms. As for the second drawback, although empirical evidence shows that, on some datasets, these aggregation methods may identify more peptides than those identified by individual database

search algorithms, none of these aggregation methods is guaranteed to do so by algorithm design.

In addition to the above aggregation methods developed for proteomics data, generic statistical methods developed for aggregating rank lists are in theory applicable to aggregating the PSM lists output by database search algorithms. However, none of these generic methods have been developed into software packages compatible with database search algorithms, nor are they guaranteed to identify more peptides given an FDR threshold. (Note that many generic methods aggregate rank lists without FDR control.) Therefore, the field calls for a robust, powerful, and flexible aggregation method that allows researchers to reap the benefits of the diverse and ever-growing database search algorithms.

Here we develop Aggregate Peptide Identification Results (APIR), a statistical framework that aggregates peptide identification results from multiple database search algorithms with FDR control. APIR is the first statistical framework universally adaptive to database search algorithms that output PSMs with scores (e.g., q-values or posterior error probabilities (PEPs)) and is guaranteed to identify at least as many as, if not more, peptides than individual database search algorithms do. APIR is a robust, flexible framework that enhances the power while controlling the FDR of peptide identification from shotgun proteomics data.

Note that the framework of APIR could be easily extended to aggregate discoveries made by multiple algorithms in other high-throughput biomedical data analysis, such as differential gene expression analysis on RNA sequencing data.

## Results

APIR aims to combine the identified PSMs of multiple database search algorithms with valid FDR control. Aside from an FDR threshold $q$ (e.g., $5\%$), APIR inputs the target-decoy search outputs (see Methods) of the database search algorithms users would like aggregate. APIR is a sequential FDR control framework that relies on APIR-adjust, its core component, to control the FDR in each step. Below we briefly describe APIR by first introducing APIR-adjust and then the sequential framework that aggregates the PSMs identified by any of pre-specified database search algorithms.

The core component of APIR is APIR-adjust, an FDR control method that identifies PSMs, under an FDR threshold $q$, from the output of a single database search algorithm; the output includes target and decoy PSMs with matching scores (see Methods). APIR-adjust is partly based on Clipper, a p-value-free FDR control framework [28] (see Methods).

Given the PSMs identified by APIR-adjust from the outputs of multiple database search algorithms, the sequential framework of APIR combines these PSMs based on a mathematical fact: if disjoint sets of discoveries all have the false discovery proportion (FDP; also known as the empirical FDR) under $q$, then their union set also has the FDP under $q$. Hence, the sequential framework of APIR is designed to find disjoint sets of PSMs from search algorithms' outputs. The final output of APIR is the union of these disjoint sets, which is guaranteed to contain more unique peptides than what could be identified by any search algorithm.

We evaluated five popular database search algorithms and benchmarked APIR against two aggregation methods on our *Pfu* proteomics standard dataset. We also designed simulation studies to benchmark APIR against two naïve aggregation approaches: intersection and union of different database search algorithms' identified PSM sets. To demonstrate the power of APIR, we applied it to five real datasets, including our proteomics standard dataset, three acute myeloid leukemia (AML) datasets, and a triple-negative breast cancer (TNBC) dataset. Notably, we generated two of the three AML datasets from bone marrow samples of AML patients with either enriched or depleted leukemia stem cells (LSCs) for studying the disease mechanisms of AML.

## Byonic, Mascot, SEQUEST, MaxQuant and MS-GF+ capture unique true PSMs in the *Pfu* proteomics standard dataset, and MaxQuant fails to control the FDR

We first benchmarked five popular database search algorithms—Byonic [9], Mascot [7], SEQUEST [6], MaxQuant [8], and MS-GF+ [10]—on the proteomics standard dataset. Specifically, we ran tandem MS analysis on a *Pfu* protein standard sample and obtained $49{,}303$ mass spectra (see Methods). We then constructed a reference database by concatenating the *Pfu* database [29], the Uniprot Human database [29], and two contaminant databases: one for affinity purification (the CRAPome) [30] and the other from MaxQuant. (The inclusion of two contaminant databases followed the convention in proteomics data analysis.) In the reference database construction, we removed human proteins that contain *Pfu* peptides (via *in silico* trypsin digestion). Finally, we input the $49{,}303$ mass spectra and the reference database into the five database search algorithms (see Methods). To evaluate a database search algorithm, we consider its output PSMs, peptides, and master proteins as true if and only if they belong to either *Pfu* or the two contaminant databases.

Our evaluation results in Fig. 1b show that the five individual database search algorithms indeed capture unique true PSMs in this proteomics standard dataset at FDR thresholds $q = 1\%$ and $5\%$. Notably, at $q = 1\%$, the number of true PSMs identified by Byonic alone $(2{,}720)$ is nearly four times the number of true PSMs identified by all five algorithms $(727)$. At $q = 5\%$, Byonic again identifies more unique true PSMs $(1{,}903)$ than the true PSMs identified by all five algorithms $(1{,}416)$. Moreover, MaxQuant and MS-GF+ also demonstrate distinctive advantages: MaxQuant identifies $147$ and $520$ unique true PSMs, while MS-GF+ identifies $153$ and $218$ at $q = 1\%$ and $5\%$, respectively. In contrast, SEQUEST and Mascot show little advantage in the presence of Byonic: their identified true PSMs are nearly all identified by Byonic (Fig. S1). Our results confirm that these five database search algorithms have distinctive advantages in identifying unique PSMs, an observation that aligns well with existing literature [22–26, 31].

In terms of FDR control, four database search algorithms—Byonic, Mascot, SEQUEST, and MS-GF+—demonstrate robust FDR control as they keep the FDPs on the benchmark data under the FDR thresholds $q \in \{1\%, \ldots, 10\%\}$. In contrast, except at small values of $q$ such as $1\%$ or $2\%$, MaxQuant fails the FDR control by a large margin (Fig. 1c).

## For individual database search algorithms, APIR-adjust shows robust FDR control and power advantage on the *Pfu* proteomics standard dataset

To demonstrate the use of APIR-adjust, we applied it as a FDR-control add-on to the five database search algorithms for identifying PSMs from their outputs. On the *Pfu* proteomics standard dataset, we examined the FDPs and power of APIR-adjust for a range of FDR thresholds: $q \in \{1\%, \ldots, 10\%\}$. Our results in Fig. 1c show that APIR-adjust achieves the FDR control and similar power (to that of the default q-value/PEP thresholding) when applied to the outputs of Byonic, Mascot, SEQUEST, and MS-GF+. As for MaxQuant, APIR-adjust alleviates the FDR control issue by reducing the FDPs to be closer to the FDR thresholds, and it achieves the FDR control when $q > 5\%$.

Even with valid q-values or PEPs, q-value/PEP thresholding cannot guarantee to control the FDR unless all PSMs with q-values/PEPs $\leq q$ are present in the output of a database search algorithm. (The reason is that, if all true PSMs with q-values/PEPs $\leq q$ are removed from the output, then the FDP would be $1$ for calling the remaining PSMs with q-values/PEPs $\leq q$ as discoveries.) In other words, q-value/PEP thresholding no longer guarantees to control the FDR after a subset of PSMs are removed. In contrast, APIR-adjust does not have this constraint, and its FDR control still works for a subset of PSMs because APIR-adjust considers q-values/PEPs as scores in its internal FDR control

(see Methods), which is robust to missing PSMs.

To verify the FDR control performance of q-value/PEP thresholding and APIR-adjust (described above) on the proteomics standard dataset, we applied q-value/PEP thresholding after excluding from each database search algorithm's output the $1{,}416$ shared true PSMs identified by all five algorithms at the FDR threshold $q = 5\%$. Our results in Fig. S2 show that thresholding the q-values of MS-GF+ no longer controls the FDR. In contrast, APIR-adjust demonstrates a robust control of the FDR even with missing PSMs.

## Set union and intersection operations do not guarantee to control the FDR on the aggregated discoveries, while APIR does

In data analysis, a popular intuition is that, if multiple algorithms designed for the same purpose are applied to the same dataset to make discoveries, and that all algorithms have FDRs under $q$, then the intersection of the discoveries (i.e., the discoveries found by all algorithms) should have the FDR under $q$ [12]. However, this intuition does not hold in general. The reason is that, if all algorithms find different true discoveries, then their common discoveries (i.e., the intersection) could be enriched with false discoveries and thus have the FDR much higher than $q$. To demonstrate this, we designed a simulation study called the shared-false-PSMs scenario, where the set intersection operation fails to control the FDR. In parallel, we designed another simulation study called the shared-true-PSMs scenario, where the set union operation fails to control the FDR. (Although intuition says that the set union operation may not control the FDR, we include it here for completeness.)

Under the shared-true-PSMs scenario, we designed three toy database search algorithms that tend to identify overlapping true PSMs but non-overlapping false PSMs (Fig. 2a top). In contrast, under the shared-false-PSMs scenario, we designed another three toy database search algorithms that tend to identify overlapping false PSMs but non-overlapping true PSMs (Fig. 2a bottom) (see Methods). Under both scenarios, we first applied APIR-adjust to each toy database search algorithm's output. Then we aggregated APIR-adjust's identified PSMs from the three algorithms under each scenario using set intersection, set union, or APIR, and we evaluated the FDR of each aggregated PSM set. Fig. 2b shows that, while set union fails to control the FDR in the shared-true-PSMs scenario and set intersection fails in the shared-false-PSMs scenario, APIR controls the FDR in both scenarios.

## For aggregating multiple database search algorithms, APIR has verified FDR control and outpowers Scaffold and ConsensusID on the *Pfu* proteomics standard dataset

We demonstrate that APIR controls the FDR after aggregating individual search algorithms on the *Pfu* proteomics standard. As expected, APIR improves the power of identifying PSMs over that of individual search algorithms, under the same FDR threshold. Next, we benchmarked APIR against two existing aggregation methods, Scaffold and ConsensusID, because they are the only two aggregation methods compatible with the five database search algorithms we used: Byonic, Mascot, SEQUEST, MaxQuant, and MS-GF+. Because database search algorithms are typically time-consuming to run, we expect that users are unlikely to aggregate more than three database search algorithms in practice. Therefore, we examined $20$ combinations of the five algorithms, including $10$ combinations of any two algorithms and $10$ combinations of any three algorithms.

Because of the trade-off between FDR and power, power comparison is valid only when FDR is controlled. Hence, for the three aggregation methods, APIR, Scaffold, and ConsensusID, we compared

them in terms of both their FDPs and power on the *Pfu* proteomics standard dataset. Regarding the power increase of each aggregation method over individual database search algorithms, we computed the percentage increases in the aggregated true PSMs, peptides, and proteins, by treating as baselines the maximal numbers of true PSMs, peptides, and proteins identified by the five database search algorithms (MaxQuant was used with APIR-adjust to control the FDR; see Fig. 1c). For example, to aggregate Byonic and MaxQuant, we would calculate the percentage increase in identified true PSMs by treating as the baseline the larger of two numbers: the numbers of true PSMs identified by Byonic and MaxQuant.

Our results in Fig. 3 and Fig. S4 show that, at both FDR thresholds $q = 5\%$ and $1\%$, APIR achieves consistent FDR control and power improvement over individual database search algorithms. In contrast, Scaffold controls the FDR but shows highly inconsistent power improvement, while ConsensusID neither controls the FDR nor has power improvement. Specifically, in terms of FDR control, although ConsensusID controls the FDR at $q = 1\%$, its FDPs exceed the FDR threshold by a large margin at $q = 5\%$: they rise above $15\%$ in $9$ out $20$ combinations. In summary, only APIR achieves power increase over individual database search algorithms consistently across the $20$ algorithm combinations, an advantage that neither Scaffold nor ConsensusID realizes.

A technical note is that Scaffold cannot control the FDR of aggregated PSMs; instead, it controls the FDRs of aggregated peptides and proteins, and it requires the target FDR thresholds to be input for both. Hence, strictly speaking, Scaffold is not directly comparable with APIR in terms of FDR control because APIR controls the FDR of aggregated PSMs. For a fair comparison, we implemented a variant of Scaffold, which, compared with the default Scaffold, has an advantage in power at the cost of an inflated FDR (see Methods). Our results in Fig. S5a and Fig. S6a show that at both FDR thresholds $q = 5\%$ and $1\%$, this Scaffold variant demonstrates a slightly inflated FDP in $11$ combinations at $q = 1\%$ and in $5$ combinations at $q = 5\%$. In terms of power, this Scaffold variant still fails to outperform the most powerful individual database search algorithm in $10$ combinations at $q = 1\%$ (Fig. S5b) and in $8$ combinations at $q = 5\%$ (Fig. S6b).

## APIR empowers peptide identification by aggregating the search results from Byonic, Mascot, SEQUEST, MaxQuant, and MS-GF+ on four real datasets

We next applied APIR with the aforementioned $20$ algorithm combinations to four real datasets: two in-house phospho-proteomics (explained below) AML datasets ("phospho AML1" and "phospho AML2") we generated for studying the properties of LSCs in AML pateints; a published nonphospho-proteomics AML dataset ("nonphospho AML") that also compares the stem cells with non-stem cells in AML patients [32]; and a published phospho-proteomics TNBC dataset that studies the drug genistein's effect on breast cancer [33]. Phospho-proteomics is a branch of proteomics; while traditional proteomics aims to capture all peptides in a sample, phospho-proteomics focuses on phosphorylated peptides, also called phosphopeptides, because phosphorylation regulates essentially all cellular processes [34]. On each dataset, we applied APIR at two FDR thresholds $q = 1\%$ and $5\%$ and examined its performance at four levels: the percentage increases in PSMs, peptides, peptides with modifications, and proteins, which we calculated in the same way as what we did for the proteomics standard dataset.

Our results in Fig. 4 ($q = 5\%$) and Fig. S7 ($q = 1\%$) show that APIR has positive percentage increases at two levels (PSMs and peptides) on all four datasets, confirming APIR's guarantee for identifying more peptides than individual algorithms do. At the peptide-with-modification level, APIR also achieves positive percentage increases across $20$ combinations on all four datasets with only one exception: APIR falls short by a negligible $0.1\%$ when aggregating the outputs of Byonic, Mascot, and SEQUEST on the TNBC dataset at $q = 5\%$. At the protein level, APIR still manages to outperform

single database search algorithms for all $20$ combinations on both phospho-proteomics AML datasets and for more than half of the combinations on the TNBC and nonphospho-proteomics AML datasets. Our results demonstrate that APIR could boost the usage efficiency of shotgun proteomics data.

## APIR identifies biologically meaningful proteins from AML and TNBC datasets

Next, we investigated the biological functions of the proteins missed by individual database search algorithms but recovered by APIR from the phospho AML and TNBC datasets.

On both phospho AML1 and AML2 datasets containing patient samples with enriched or depleted LSCs, APIR identified biologically meaningful proteins that were missed by individual database search algorithms. On phospho AML1, APIR identified across the $20$ combinations $80$ additional proteins (the union of the additional proteins APIR identified from the combinations) at the FDR threshold $q = 1\%$ and $121$ additional proteins at the FDR threshold $q = 5\%$. These two sets of additional proteins recovered by APIR include some well-known proteins, such as transcription intermediary factor 1-alpha (TIF1$\alpha$), phosphatidylinositol 4,5-bisphosphate 5-phosphatase A (PIB5PA), sterile alpha motif domain containing protein 3 (SAMD3), homeobox protein Hox-B5 (HOXB5), small ubiquitin-related modifier 2 (SUMO-2), transcription factor jun-D (JUND), glypican-2 (GPC2), dnaJ homolog subfamily C member 21 (DNAJC21), mRNA decay activator protein ZFP36L2, leucine-rich repeats and immunoglobulin-like domains protein 1 (LRIG-1), and mitochondrial intermembrane space import and assembly protein 40 (CHCHD4). Here we summarize the tumor-related functions of these well-known proteins. High levels of TIF1$\alpha$ are associated with oncogenesis and disease progression in a variety of cancer lineages such as AML [35–41]. PIB5PA has been shown to have a tumor-suppressive role in human melanoma [42]. Its high expression has been correlated with limited tumor progression and better prognosis in breast cancer patients [43]. SMAD3 is known to play key roles in the development and progression of various types of tumor [44–49]. HOXB5 is among the most affected transcription factors by the genetic mutations that initiate AML [50–52]. SUMO-2 has been found to play a key role in regulating CBX2, which is overexpressed in several human tumors (e.g., leukemia) and whose expression is correlated with lower overall survival [53]. JUND has been shown to play a central role in the oncogenic process leading to adult T-cell leukemia [54]. GPC2 has been identified as an oncoprotein and a candidate immunotherapeutic target in high-risk neuroblastoma [55]. DNAJC21 mutations have been linked to cancer-prone bone marrow failure syndrome [56]. ZFP36L2 has been found to induce AML cell apoptosis and inhibit cell proliferation [57]; its mutation has been associated with the pathogenesis of acute leukemia [58]. LRIG-1 has been found to regulate the self-renewing ability of LSCs in AML [59]. CHCHD4 plays key roles in regulating tumor proliferation [60]. On phospho AML2, APIR has identified $62$ additional proteins at FDR $1\%$ and $19$ additional proteins at FDR $5\%$, including JUND and myeloperoxidase (MPO). MPO is expressed in hematopoietic progenitor cells in prenatal bone marrow, which are considered initial targets for the development of leukemia [61–63].

On the TNBC dataset, APIR identified $92$ additional proteins missed by individual database search algorithms at the FDR threshold $q = 1\%$ and $69$ additional proteins at $q = 5\%$. In particular, at $q = 1\%$, APIR has uniquely identified breast cancer type 2 susceptibility protein (BRCA2), and Fanconi anemia complementation group E (FANCE). BRCA2 is a well-known breast cancer susceptibility gene; an inherited genetic mutation inactivating the BRCA2 gene can be found in people with TNBC [64–69]. The FANC-BRCA pathway, including FANCE and BRCA2, is known for its roles in DNA damage response. Inactivation of the FANC–BRCA pathway has been identified in ovarian cancer cell lines and sporadic primary tumor tissues [70, 71]. Additionally, at both $q = 1\%$ and $5\%$, we identified JUND and roundabout guidance receptor 4 (ROBO4); the latter regulates tumor growth and metastasis in multiple types of cancer, including breast cancer [72–75].

Our results, summarized in Table 1, demonstrate APIR's strong potential in identifying novel disease-related proteins.

### APIR empowers the identification of differentially expressed peptides

An important use of proteomics data is the differential expression (DE) analysis, which aims to identify proteins whose expression levels change between two conditions. Protein is the ideal unit of measurements; however, due to the difficulties in quantifying protein levels from tandem MS data, an alternative approach has been proposed and used, which first identifies differentially expressed peptides and then investigates their corresponding proteins along with their modifications. Because it is less error-prone to quantify peptides than proteins, doing so would dramatically reduce errors in the DE analysis.

Here we compared APIR with MaxQuant and MS-GF+ by performing DE analysis on the phospho AML1 dataset. We focused on this dataset instead of the TNBC dataset or the nonphospho AML dataset because the phosphos AML datasets were generated for our in-house study and thus may yield new discoveries. The phospho AML1 dataset contains six bone marrow samples: three enriched with LSCs, two depleted of LSCs, and one control. To simplify our DE analysis, we selected two pairs of enriched and depleted samples as shown in Fig. 5a. Specifically, we first applied APIR to aggregate the outputs of MaxQuant and MS-GF+ on the phospho AML1 dataset using all six samples. Then we applied DESeq2 to identify DE peptides from the aggregated peptides of APIR, APIR-adjusted MaxQuant, and APIR-adjusted MS-GF+ using the four selected samples. Our results in Fig. 5 show that at the FDR threshold $q = 5\%$, we identified $318$ DE peptides from $224$ proteins based on APIR, $251$ DE peptides from $180$ proteins based on MaxQuant, and $242$ DE peptides from $190$ proteins based on MS-GF+, respectively. In particular, APIR identified $6$ leukemia-related proteins: the promyelocytic leukemia zinc finger (PLZF), serine/threonine-protein kinase B-raf (B-raf), signal transducer and activator of transcription 5B (STAT5B), promyelocytic leukemia protein (PML), cyclin-dependent kinase inhibitor 1B (CDKN1B), and retinoblastoma-associated protein (RB1), all of which belong to the AML KEGG pathway or the chronic myeloid leukemia KEGG pathway [76–78]. In particular, PLZF and CDKN1B were uniquely identified from the APIR aggregated results but not by either APIR-adjusted MaxQuant or APIR-adjusted MS-GF+.

We next investigated the phosphorylation on the identified DE peptides of PLZF or CDKN1B. With regard to PLZF, APIR identified phosphorylation at Threonine 282, which is known to activate cyclin-A2 [79], a core cell cycle regulator of which the deregulation seems to be closely related to chromosomal instability and tumor proliferation [80–82]. As for CDKN1B, APIR identified phosphorylation at Serine 140. Previous studies have revealed that ATM phosphorylation of CDKN1B at Serine 140 is important for stabilization and enforcement of the CDKN1B-mediated G1 checkpoint in response to DNA damage [83]. A recent study shows that inability to phosphorylate CDKN1B at Serine 140 is associated with enhanced cellular proliferation and colony formation [84]. Our results, summarized in Table 2, demonstrate that APIR can assist in discovering interesting proteins and relevant post-translational modifications.

## Discussion

We developed a statistical framework APIR to combine the power of distinct database search algorithms, by aggregating their identified PSMs from shotgun proteomics data with FDR control. The core component of APIR is APIR-adjust, an FDR-control method that re-identifies PSMs from a single database search algorithm's output without restrictive distribution assumptions. APIR offers a great advantage of flexibility: APIR is compatible with any database search algorithms. The reason lies in that APIR is a sequential approach based on a mathematical fact: given multiple disjoint sets of discoveries, each

with the FDP smaller than or equal to $q$, their union also has the FDP smaller than or equal to $q$. This sequential approach not only allows APIR to circumvent the need to impose restrictive distribution assumptions on each database search algorithm's output, but also ensures that APIR would identify at least as many, if not more, unique peptides as a single database search algorithm does.

By assessing APIR on the first publicly available complex proteomics standard dataset we generated, we verified that APIR consistently improves the power of peptide identification with the FDR controlled on the identified PSMs. Our extensive studies on AML and TNBC data suggest that APIR can discover additional disease-relevant peptides and proteins that are otherwise missed by individual database search algorithms.

The current implementation of APIR controls the FDR at the PSM level. However, in shotgun proteomics experiments, PSMs serve merely as an intermediate to identify peptides and then proteins, the real molecules of biological interest; thus, an ideal FDR control should occur at the protein level. A fact is that FDR control at the PSM level does not entail FDR control at the protein level because multiple PSMs may correspond to the same peptide sequence, and multiple peptides may correspond to the same protein. To realize the FDR control on the identified proteins, APIR-adjust needs to be carefully modified. A possible modification would be to construct a matching score for each protein from the matching scores of the PSMs that correspond to this protein's peptides. Future studies are needed to explore possible ways of constructing proteins' matching scores. Once we modify APIR-adjust to control the FDR at the protein level, the current sequential approach of APIR still applies: applying the modified APIR-adjust to sequentially identify disjoint sets of proteins from individual database search algorithms' outputs; outputting the union of these disjoint sets as discoveries.

Although APIR is designed for proteomics data, its framework is general and extendable to aggregating discoveries in other popular high-throughput biomedical data analyses, including peak calling from ChIP-seq data, differentially expressed gene (DEG) identification from bulk or single-cell RNA sequencing data, and differentially interacting chromatin region identification from Hi-C data [28]. For example, an extended APIR may aggregate discoveries made by popular DEG identification methods, such as DESeq2 [85], edgeR [86], and limma [87], to increase the power while maintaining the FDR control.

# Methods

## 1. APIR

### 1.1. Target-decoy search strategy

The key idea of the target-decoy search strategy is to generate a negative control of PSMs by matching mass spectra against artificially created, false protein sequences, called "decoy" sequences. Decoy sequences can be created in multiple ways, and a typical way is to reverse each protein sequence to obtain a corresponding decoy sequence. Given the decoy sequences, the target-decoy search strategy can be implemented as the concatenated search or parallel search.

In the concatenated search, a concatenated protein database is created by pooling original protein sequences, called "target" sequences, with decoy sequences; then a database search algorithm uses the concatenated protein database to find PSMs; consequently, each mass spectra is matched to either a target sequence or a decoy sequence with only one matching score (Fig. S8a).

In the parallel search, a database search algorithm conducts two parallel searches: a target search where each mass spectrum is matched to target sequences and a decoy search where the mass spectrum is matched to decoy sequences; consequently, each mass spectrum receives two matching scores for the two searches (Fig. S8b).

In both implementations, a PSM is called a target PSM or simply a PSM if it contains a target sequence; otherwise, it is called a decoy PSM. Finally, a database search algorithm uses the decoy PSMs, i.e., the PSMs known to be false, to estimate the FDR [11, 16]. In technical terms, each target PSM receives a q-value from an algorithm such as Byonic, Mascot, SEQUEST, and MS-GF+ [6, 7, 9, 10] or a posterior error probability (PEP) from an algorithm such as MaxQuant [8] (see Methods). Both q-value and PEP are related to the FDR so that users can control the FDR under a threshold $q$ if they keep only the target PSMs with q-values or PEPs not exceeding $q$; however, the FDR control is only guaranteed when the q-values and PEPs are valid [15].

### 1.2. APIR methodology

APIR aims to aggregate the PSMs identified from multiple database search algorithms with FDR control. Aside from an FDR threshold $q$ (e.g., $5\%$), APIR requires as input a list of target PSMs with scores and a list of decoy PSMs with scores from each database search algorithm. To maximize power, we recommend users to input the entire list of target PSMs and decoy PSMs by setting the internal FDR of each database search algorithm to $100\%$. Because nearly all database search algorithms output q-values or PEPs for target and decoy PSMs, we recommend users to use $-\log_{10}$-transformed q-values or $-\log_{10}$-transformed PEPs as the input of scores. Unless specified otherwise, in this manuscript, the results of APIR are generated based on $-\log_{10}$-transformed PEPs from MaxQuant and $-\log_{10}$-transformed q-values from the other four database search algorithms. Below we introduce the details of APIR by first introducing APIR-adjust and then the general framework based on APIR-adjust for aggregating search results.

#### 1.2.1. APIR-adjust: FDR control on the target PSMs identified by any individual search algorithm

The core component of APIR is APIR-adjust, an FDR-control method that re-identifies target PSMs from a single database search algorithm. APIR-adjust takes as input an FDR threshold $q$, a list of target PSMs with scores, and a list of decoy PSMs with scores. APIR-adjust then outputs identified target PSMs.

We first define the target coverage proportion as the proportion of target PSMs whose mass spectra also appear among the decoy PSMs. Depending on the database search algorithms and the implementation of target-decoy search strategy (concatenated or parallel), the target coverage proportion could vary from $0$ to $1$. When the target coverage proportion is high, most of the target PSMs could be one-to-one paired with decoy PSMs by their mass spectra so that in each pair, the decoy PSM score serves as a negative control for the target PSM score. When the proportion is low, we cannot form many pair-decoy score pairs but use decoy PSM scores collectively as a negative control. We thus design two approaches, tailored specifically for these two scenarios, into APIR-adjust.

Here we introduce notations to facilitate our discussion. Suppose a database search algorithm combination of an experimental design, a distributional family, and a background scenario outputs $m$ target PSMs with scores $T_1, \ldots, T_m$ and $n$ decoy PSMs with scores $D_1, \ldots, D_n$. Also, suppose that among the $m$ target PSMs, the first $s \leq \min(m, n)$ target PSMs can be paired one-to-one with decoy PSMs; accordingly, the target coverage proportion is $s/m$. Without loss of generality, we rearrange decoy PSM indices such that the $i$-th decoy PSM shares the same mass spectrum with the $i$-th target PSM for $1 \leq i \leq s$.

When the target coverage proportion is relatively high ($s/m \geq 40\%$), APIR-adjust identifies target PSMs using Clipper, a p-value-free statistical framework for FDR control on high-throughput data by contrasting two conditions, although a similar approach has been proposed in Couté, Bruley, and Burger [88]. Specifically, Clipper constructs a contrast score $C_i = T_i - D_i$ if $i = 1, \ldots, s$ and $C_i = 0$ if $i =$

$s+1,\ldots,m$; then it finds a cutoff $C_{\text{thre}} = \min\left\{t \in \{|C_i| : C_i \neq 0\} : \frac{|i:C_i \leq -t|+1}{\max(|i:C_i \geq t|,1)} \leq q\right\}$, and outputs $\{i : C_i \geq C_{\text{thre}}\}$ as the indices of identified target PSMs. Based on Clipper, APIR-adjust requires two assumptions to control the FDR: first, $T_1,\ldots,T_m,D_1,\ldots,D_n$ are mutually independent, and second, $T_i$ and $D_i$ are identically distributed if the $i$-th target PSM is false. See the original paper for detailed proofs that guarantee FDR control [28].

When the target coverage proportion is relatively low ($s/m < 40\%$), APIR-adjust uses a p-value-based approach to identify target PSMs. By assuming that the scores of decoy PSMs and false target PSMs are independently and identically distributed, the p-value-based approach constructs a null distribution by pooling $D_j, j = 1,\ldots,n$. Then APIR-adjust computes a p-value for the $i$-th target PSM as the tail probability right of $T_i$, i.e., $p_i = |\{j : D_j \geq T_i\}|/n, i = 1,\ldots,m$, and controls the FDR using the Benjamini-Horchberg procedure [89].

**1.2.2. APIR: a sequential framework for aggregating multiple search algorithms' identified target PSMs with FDR control** Suppose we are interested in aggregating $K$ algorithms. Let $W_k$ denote the set of target PSMs output by the $k$-th algorithm, $k = 1,\ldots,K$. APIR adopts a sequential approach that consists of $K$ rounds.

- In the first round, APIR applies APIR-adjust or q-value/PEP thresholding to each algorithm's output with the FDR threshold $q$. Denote the identified target PSMs from the $k$-th algorithm by $U_{1k} \subset W_k$. Define $J_1 \in \{1,\cdots,K\}$ to be the algorithm such that $U_{1J_1}$ contains the largest number of unique peptides among $U_{11},\ldots,U_{1K}$. We use the number of unique peptides rather than the number of PSMs because peptides are more biologically relevant than PSMs.

- In the second round, APIR first excludes all target PSMs output by $J_1$, identified or unidentified in the first round, i.e., $W_{J_1}$, from the outputs of the remaining database search algorithms, resulting in reduced sets of candidate target PSMs $W_1 \setminus W_{J_1}$, ..., $W_K \setminus W_{J_1}$. Then APIR applies APIR-adjust with FDR threshold $q$ to these reduced sets except $W_{J_1} \setminus W_{J_1} = \emptyset$. Denote the resulting sets of identified target PSMs by $U_{2k} \subset (W_k \setminus W_{J_1})$, $k \in \{1,\ldots,K\} \setminus \{J_1\}$. Again APIR finds the $J_2$-th algorithm such that $U_{2J_2}$ contains the most unique peptides.

- APIR repeats this in the subsequent rounds. In Round $\ell$ with $\ell \geq 2$, APIR first excludes all target PSMs output by the selected $\ell-1$ database search algorithms from the outputs of remaining database search algorithms and applies APIR-adjust. That is, APIR applies APIR-adjust with FDR threshold $q$ to identify a set of identified PSMs $U_{\ell k}$ from $W_k \setminus (W_{J_1} \cup \cdots \cup W_{J_{\ell-1}})$, the reduced candidate pool of algorithm $k$ after the previous $\ell-1$ rounds, for algorithms $k \in \{1,\cdots,K\} \setminus \{J_1,\cdots,J_{\ell-1}\}$. Then APIR finds the algorithm, which we denote by $J_\ell$, such that $U_{\ell J_\ell}$ contains the most unique peptides.

- Finally, APIR outputs $U_{1J_1} \cup \cdots \cup U_{KJ_K}$ as the identified target PSMs.

By adopting this sequential approach, APIR is guaranteed to identify at least as many, if not more, unique peptides as those identified by a single database search algorithm; under reasonable assumptions, APIR controls the FDR of the identified target PSMs under $q$. See Fig. 6b for graphical illustration and next section for the the theoretical guarantee of FDR control by APIR.

### 1.3. Post-processing

**Master protein recommendation** For a given PSM, database search algorithms may disagree on its master protein, causing difficulties in downstream analysis. APIR tackles this issue using a majority

vote. Specifically, APIR selects the most frequently reported master protein across database search algorithms for the given PSM. If there is a tie, APIR outputs all tied master proteins.

**Post-translational modification recommendation** For a given PSM, how APIR aggregates its modifications across database search algorithms depends on the type of modifications: static or variable. Static modifications occur universally at every instance of a specified amino acid residue or terminus. For example, tandem mass tags occur at every N-termimal. Since static modifications are known and could be specified in the database search process, different database search algorithms will agree in terms of the locations and types of static modifications. Therefore, for any PSM, APIR simply outputs its static modifications by any database search algorithm based on user specification. The default static modification used by APIR includes cysteine carbamidomethylation and tandem mass tags at N-terminal and lysine.

Unlike static modifications, variable modifications do not apply to all instances of an amino acid residue. For example, phosphorylation typically occurs at only one or few serines in a peptide with many serines. Because variable modifications are hard to detect, database search algorithms may disagree in the types (such as phosphorylation versus oxidation) and/or sites of modifications; however, they always agree on the number of modifications. Suppose database search algorithms report $M$ modifications for the given PSM. To handle these potential disagreements, APIR uses one of the two strategies to recommend variable modifications for a given PSM: PhosphoSitePlus (PSP)-free or PSP-based. In a PSP-free modification recommendation, APIR first counts the number of database search algorithms that report each modification—a combination of modification type and site. Then APIR reports the top $M$ most frequently reported variable modifications. A PSP-based modification strategy is similar to PSP-free except for the handling of tied phosphorylation sites. When there is a tie among phosphorylation sites, APIR reports the most frequently studied phosphorylation sites by searching the literature hits on PSP (`https://www.phosphosite.org/`), a manually curated and interactive resource for studying protein modifications. In particular, PSP has cataloged and counted existing literature and experiments by phosphorylation. Based on PSP, APIR reports the modification with the largest number of high-throughput literature hits if there is a tie between phosphorylations. If doing so fails to identify a unique modification, APIR compares their numbers of Cell Signaling Technology mass spectrometry studies that found the given phosphorylation and report the largest-numbered phosphorylation. If this fails to provide a unique modification, APIR will report the ties.

**Abundance aggregation** At the PSM level, APIR first averages a PSM's abundance across database search algorithms. Then APIR performs normalization by scaling $a_{ij}$, which denotes the averaged abundance of PSM $i$ in channel $j$, by $10^6/(\sum_i a_{ij})$ so that resulting normalized samples will have total abundance $10^6$.

To obtain the abundance at the peptide level, APIR averages the abundance of PSMs containing the same peptide and then performs a scaling across channels such that its cross-channel average equals $100$. Specifically, let $b_{ij}$ denotes the averaged abundance of peptide $i$ in sample $j$. The normalized abundance would be $100b_{ij}/(\sum_j b_{ij})$.

To obtain the abundance at the protein level, APIR averages the abundance of PSMs with the same recommended master protein and then performs the same row normalization as it does at the peptide level.

### 1.4. Theoretical results of APIR

To facilitate our discussion, we start with notations for the mathematical abstraction of APIR, followed by assumptions and proofs.

Let $\Omega$ denote the set of all possible PSMs from a tandem MS experiment and $W_k \subset \Omega$ denote the set of target PSMs output by the $k$-th database search algorithm, $k = 1, \ldots, K$. Let $S_k$ denote the set containing scores of the PSMs in $W_k$. The exact definition of $S_k$ depends on the implementation of APIR-adjust: Clipper or the pooled approach. Specifically, if APIR-adjust adopts Clipper, $S_k = \{C_i : i \in W_k\} \subset \mathbb{R}$, where $C_i$ is the contrast score of Clipper (See the previous section). If APIR-adjust adopts the pooled approach, $S_k = \{p_i : i \in W_k\} \subset \mathbb{R}$, where $p_i$ is the p-value calculated using the pooled approach (see the previous section). We define $\mathcal{W} := \{W : W \subset \Omega\}$ to be the power set of $\Omega$ and $\mathcal{S} := \{S : S \subset \mathbb{R}\}$ to be the power set of $\mathbb{R}$.

Here we introduce the mathematical abstraction of APIR-adjust. Given an FDR threshold $q \in (0, 1)$ and a set of target PSMs $W$ with their scores $S$ from a single database search algorithm, we define $\mathcal{P}_q : \mathcal{W} \times \mathcal{S} \to \mathcal{W}$ as a *identification procedure* that takes $W$ and $S$ as input and outputs a subset of $W$ with FDR controlled under $q$.

Next, we introduce a *selection procedure*, denoted by $\mathcal{Q}$, that finds the index of the "best" set among multiple target PSM sets, where "best" in default APIR means having the most unique peptides. Specifically,

$$\mathcal{Q} : \underbrace{\mathcal{W} \times \cdots \times \mathcal{W}}_{\text{any finite number}} \to \{1, \ldots, K\}$$

takes as input multiple sets of target PSMs, each from a distinct database search algorithm, and outputs the index of the database search algorithm whose set is selected as the "best." In case the "best" set is not unique, $\mathcal{Q}$ randomly selects one of the "best" sets and outputs its index.

Then APIR consists of $K$ rounds:

$$\text{Round 1}: \left.\begin{aligned} U_{11} :=& \mathcal{P}_q\left(W_1, S_1\right); \\ &\vdots \\ U_{1K} :=& \mathcal{P}_q\left(W_K, S_K\right); \end{aligned}\right\} K \text{ sets of identified PSMs}$$
$$J_1 := \mathcal{Q}\left(\{U_{1k} : k = 1, \cdots, K\}\right); \text{ the index of the selected algorithm}$$
$$\vdots$$
$$\text{Round } \ell: \left.\begin{aligned} U_{\ell 1} :=& \mathcal{P}_q\left(W_1 \setminus \left(\cup_{k'=1}^{\ell-1} W_{J_{k'}}\right), S_1\right); \\ &\vdots \\ U_{\ell K} :=& \mathcal{P}_q\left(W_K \setminus \left(\cup_{k'=1}^{\ell-1} W_{J_{k'}}\right), S_K\right); \end{aligned}\right\} K \text{ sets of identified PSMs}$$
$$J_\ell = \mathcal{Q}\left(\{U_{\ell k} : k \neq J_1, \cdots, J_{\ell-1}\}\right); \text{the index of the selected algorithm}$$
$$\vdots$$
$$\text{Round } K: \left.\begin{aligned} U_{K1} :=& \mathcal{P}_q\left(W_1 \setminus \left(\cup_{k'=1}^{K-1} W_{J_{k'}}\right), S_1\right); \\ &\vdots \\ U_{KK} :=& \mathcal{P}_q\left(W_K \setminus \left(\cup_{k'=1}^{K-1} W_{J_{k'}}\right), S_K\right); \end{aligned}\right\} K \text{ sets of identified PSMs}$$
$$J_K = \mathcal{Q}\left(\{U_{Kk} : k \neq J_1, \cdots, J_{K-1}\}\right); \text{ the index of the selected algorithm}$$

and outputs $\cup_{\ell=1}^{K} U_{\ell J_\ell}$ as the final set of identified PSMs. Let $V_{\ell k}$ denote the number of false PSMs in $U_{\ell k}$, $\ell, k = 1, \ldots, K$. Because $U_{1J_1}, \ldots, U_{KJ_K}$ are mutually disjoint, to show FDR control we only need

to show that

$$\mathbb{E}\left[\frac{\sum_{\ell=1}^{K} V_{\ell J_\ell}}{\left(\sum_{\ell=1}^{K} |U_{\ell J_\ell}|\right) \vee 1}\right] \leq q\,, \tag{1}$$

where $a \vee b$ means $\max(a, b)$.

To facilitate our theoretical discussion, we would like to emphasize the source of randomness and how we represent them in our notations. First, both $\{W_k\}_{k=1}^K$ and $\{S_k\}_{k=1}^K$ are random because the shotgun proteomics technology is innately random. Consequently, the mass spectra from tandem MS experiments could vary in terms of numbers and quality, leading to random lists of target/decoy PSMs and random scores output by database search algorithms. By convention, we use capital letters "$W$" and "$S$" to represent a random set of PSMs and a random set of scores respectively. Second, $\mathcal{P}_q$ and $\mathcal{Q}$ could be random or deterministic functions. Third, $\{U_{\ell k} : \ell, k = 1, \cdots, K\}$ are random due to the random input of $\mathcal{P}_q$; therefore, they are also represented by capital $U$. For similar reasons, $\{J_1\}_{k=1}^K$ are also random. Lastly, as a result, notations such as $W_{J_\ell}$ and $S_{J_\ell}$ have two layers of randomness: random PSM sets and scores represented by $W$ and $S$ and random database search algorithm index $J_\ell$. Notably, although a capital letter, $K$ is deterministic because it represents the number of database search algorithms we want to aggregate.

Here we introduce assumptions of APIR for FDR control. Conditioning on $W_k$, let $\mu_k := \mathbb{E}[S_k]$ denote the set of expected scores output by algorithm $k$. We impose three sets of assumptions respectively on $\{W_k, \mu_k, S_k\}_{k=1}^K$, $\mathcal{P}_q$ and $\mathcal{Q}$.

As for $\{W_k, \mu_k, S_k\}_{k=1}^K$, we require

(A.1) conditioning on $\{W_k, \mu_k\}_{k=1}^K$, $S_1, \ldots, S_K$ are mutually independent.

As for $\mathcal{P}_q$, we assume the following.

(A.2) Conditioning on $\{W_k, \mu_k\}_{k=1}^K$ and given any subset $\widetilde{W}_k \subset W_k$ for any $k = 1, \ldots, K$, we obtain $\mathcal{P}_q(\widetilde{W}_k, S_k)$. Let $\widetilde{U}_k := \mathcal{P}_q(\widetilde{W}_k, S_k)$ and $\widetilde{V}_k$ denote the number of false PSMs in $\widetilde{U}_k$. Then $\mathbb{E}[\widetilde{V}_k/(|\widetilde{U}_k| \vee 1) \mid \{W_k, \mu_k\}_{k=1}^K] \leq q$ for all $k = 1, \ldots, K$. That is, $\mathcal{P}_q$ controls FDR when applied to any subset of target PSMs from each of the $K$ database search algorithms. Notably, this assumption is guaranteed for APIR-adjust if the assumptions in the previous section hold.

(A.3) Following (A.2), if we assume that $\widetilde{V}_k/(|\widetilde{U}_k| \vee 1)$ is independent of $|\widetilde{U}_k|/(\sum_{k=1}^K |\widetilde{U}_k| \vee 1)$ for $k = 1, \ldots, K$. That is, the FDP of the discoveries, i.e., identified PSMs from a subset of the target PSMs, from algorithm $k$ is independent of the proportion of the discoveries from algorithm $k$ among all discoveries.

Finally, we assume the following about $\mathcal{Q}$.

(A.4) Conditioning on $\{W_k, \mu_k\}_{k=1}^K$, $\{J_\ell\}_{\ell=1}^K$ is independent of $\{S_k\}_{k=1}^K$. That is, the output of procedure $\mathcal{Q}$ is conditionally independent of the randomness of the scores output by the $K$ algorithms.

We start our proof by first showing that conditioning on $\{W_k = w_k\}_{k=1}^K$, $\{\mu_k\}_{k=1}^K$ and $\{J_k = j_k\}_{k=1}^\ell$,

$$\{J_{\ell+1}, \cdots, J_K\} \perp \frac{V_{\ell j_\ell}}{|U_{\ell j_\ell}| \vee 1}. \tag{2}$$

Because $V_{\ell j_\ell}/(|U_{\ell j_\ell}| \vee 1)$ is the FDP of $\mathcal{P}_q(w_{j_\ell} \setminus (\cup_{k'=1}^{\ell-1} w_{j_{k'}}), S_{j_\ell})$, the randomness of $V_{\ell j_\ell}/(|U_{\ell j_\ell}| \vee 1)$ results solely from the randomness of scores in $S_{j_\ell}$. By (A.4), $S_{j_\ell}$ is independent of $\{J_{\ell+1}, \cdots, J_K\}$ conditioning on $\{W_k, \mu_k\}_{k=1}^K$ and $J_1, \ldots, J_\ell$. Equation (2) follows accordingly.

We can then show FDR control in the $\ell$-th round conditioning on $\{W_k, \mu_k\}_{k=1}^K$ and $\{J_\ell\}_{\ell=1}^K$:

$$\begin{aligned}
&\mathbb{E}\left[\frac{V_{\ell J_\ell}}{|U_{\ell J_\ell}| \vee 1}\middle| W_1 = w_1, \cdots, W_K = w_K, \{\mu_k\}_{k=1}^K, J_1 = j_1, \cdots, J_K = j_K\right] \\
=&\mathbb{E}\left[\frac{V_{\ell j_\ell}}{|U_{\ell j_\ell}| \vee 1}\middle| W_1 = w_1, \cdots, W_K = w_K, \{\mu_k\}_{k=1}^K, J_1 = j_1, \cdots, J_K = j_K\right] \\
=&\mathbb{E}\left[\frac{V_{\ell j_\ell}}{|U_{\ell j_\ell}| \vee 1}\middle| W_1 = w_1, \cdots, W_K = w_K, \{\mu_k\}_{k=1}^K, J_1 = j_1, \cdots, J_\ell = j_\ell\right] \\
\leq& q\,,
\end{aligned}$$

where the last equality results from (2). The last inequality holds by (A.2).

Finally, we prove (1):

$$\begin{aligned}
&\mathbb{E}\left[\frac{\sum_{\ell=1}^K V_{\ell J_\ell}}{\left(\sum_{\ell=1}^K |U_{\ell J_\ell}|\right) \vee 1}\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[\sum_{\ell=1}^K \frac{|U_{\ell J_\ell}|}{(|U_{1J_1}| + \cdots + |U_{KJ_K}|) \vee 1} \frac{V_{\ell J_\ell}}{|U_{\ell J_\ell}| \vee 1}\middle| \{W_k, \mu_k, J_k\}_{k=1}^K\right]\right] \\
=&\mathbb{E}\left[\sum_{\ell=1}^K \mathbb{E}\left[\frac{|U_{\ell J_\ell}|}{(|U_{1J_1}| + \cdots + |U_{KJ_K}|) \vee 1}\middle| \{W_k, \mu_k, J_k\}_{k=1}^K\right] \mathbb{E}\left[\frac{V_{\ell J_\ell}}{|U_{\ell J_\ell}| \vee 1}\middle| \{W_k, \mu_k, J_k\}_{k=1}^K\right]\right] \qquad (3) \\
\leq& \sum_{\ell=1}^K \mathbb{E}\left[\mathbb{E}\left[\frac{|U_{\ell J_\ell}|}{(|U_{1J_1}| + \cdots + |U_{KJ_K}|) \vee 1}\middle| \{W_k, \mu_k, J_k\}_{k=1}^K\right]\right] \cdot q \\
\leq& q\,,
\end{aligned}$$

where (3) holds as a result of (A.3).

## 2. Simulation studies

Here we describe how we conducted the simulation studies. Suppose that we have a total of $10^4$ mass spectra and that target PSMs and decoy PSMs are ordered in such a way that the $i$-th target PSM shares the same mass spectrum as the $i$-th decoy PSM. Among the $10^4$ target PSMs, $1500$ are true PSMs, and the rest are false. Let $T_{ij}$ and $D_{ij}$ denote the score of the $i$-th target PSM and the score of the $i$-th decoy PSM by toy database search algorithm $j$, $j = 1, \ldots, 6$. In addition, we generated missing indices $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3 \subset \{1, 2, \ldots, 10^4\}$ by randomly sampling without replacement $1000$, $2000$, and $3000$ indices from $\{1, 2, \ldots, 10^4\}$. We set $\mathcal{M}_{j+3} = \mathcal{M}_j$ for $j = 1, 2, 3$. We generated $200$ simulated datasets under either the shared-true-PSMs scenario or the shared-false-PSMs scenario using the following procedures.

Under the shared-true-PSMs scenario, we generated the target and decoy output of toy search algorithms $1, 2, 3$ using the following procedure. If the $i$-th target PSM is true, we generated $X_i$ from the exponential distribution with mean $8$, $Y_i$ from the exponential distribution with mean $1$ and set $T_{i1} = T_{i2} = T_{i3} = X_i$ and $D_{i1} = D_{i2} = D_{i3} = Y_i$; if the $i$-th target PSM is false, we generated $T_{i1}$, $T_{i2}$, $T_{i3}$, $D_{i1}$, $D_{i2}$, $D_{i3}$ independently from exponential with mean $1$. Under the shared-false-PSMs scenario, we generated the target and decoy output of toy search algorithms $4, 5, 6$ using the following procedure. If the $i$-th target PSM is true, we generated $T_{i4}$, $T_{i5}$, $T_{i6}$ independently from exponential with mean $4$ and $D_{i4}$, $D_{i5}$, $D_{i6}$ independently from exponential with mean $1$; if the $i$-th target PSM is false, we first generated $X_i$ and $Y_i$ independently from the exponential distribution with mean $1$ and then set $T_{i4} = T_{i5} = T_{i6} = X_i$ and $D_{i4} = D_{i5} = D_{i6} = Y_i$. Under either scenario, we set $T_{ij}$ to be a missing

value if $i \in \mathcal{M}_j$ so that each algorithm captures unique target PSMs.

We examined the actual FDRs of APIR-adjust on each toy database search algorithm and of aggregation methods: union or intersection of the identified PSM sets from individual database search algorithms, and APIR at the FDR threshold $q = 5\%$. For each FDR-control method, we calculated an FDP—the proportion of identified PSMs that are false—on each simulated data and averaged those $200$ FDPs to compute the FDR. To obtain the FDP of APIR-adjust, we applied APIR-adjust with the FDR threshold $q = 5\%$ to each toy database search algorithm. To obtain the FDP of union/intersection, we took the union/intersection of the three sets of identified target PSMs by APIR-adjust, one per each toy database search algorithm. To obtain the FDP of APIR, we applied the default APIR to aggregate the three toy database search algorithms with the FDR threshold $q = 5\%$.

## 3. Data generation

### 3.1. Complex proteomics standard dataset generation

The complex proteomics standard (CPS) (part number 400510) was purchased by Agilent (Agilent, Santa Clara, CA, USA). CPS contains soluble proteins extracted from the archaeon *Pyrococcus furiosus* (*Pfu*). All other chemicals were purchased from Sigma Aldrich (Sigma Aldrich, St. Louis, MO, USA). The fully sequenced genome of *Pfu* encodes for approximately 2000 proteins that cover a wide range of size, pI, concentration levels, hydrophobic/hydrophilic character, etc. CPS (500ug total protein) was dissolved in 100uL of 0.5 M tri-etrhylammonium bicarbonate (TEAB) and 0.05% sodium dodecyl sulfate (SDS) solution. Proteins were reduced using tris(2-carboxyethyl)phosphine hydrochloride (TCEP) (4 uL of 50mM solution added in the protein mixture and sample incubated at 60$^\circ$C for 1hour) and alkylated using methyl methyl methanethiosulfonate (MMTS) (2 uL of 50mM solution added in the protein mixture and sample incubated at room temperature for 15 minutes). To enzymatically digest the proteins, 20ug trypsin dissolved 1:1 in ultrapure water was added in the sample and this was incubated overnight (16 hours) in dark at 37$^\circ$C. The tryptic peptides were cleaned with C-18 tips (part number 87784) from Thermo Fisher Scientific (Thermo Fisher Scientific, Waltham, MA, USA) following the manufacturer's instructions. Peptides were LC-MS analysed using the Ultimate 3000 uPLC system (EASY-Spray column, part number ES803A, Thermo Fisher Scientific) hyphenated with the Orbitrap Fusion Lumos mass spectrometry instrument (Thermo Fisher Scientific). Peptides were fragmented using low energy CID and detected with the linear ion trap detector.

On this complex proteomics standard dataset, we benchmarked the five database search algorithms—SEQUEST [6], Mascot [7], MaxQuant [8], Byonic [9], and MS-GF+ [10]—in terms of identifying target PSMs. Specifically, we first generated a reference database by concatenating the Uniprot *Pyrococcus furiosus* (*Pfu*) database [29], the Uniprot Human database [29], and two contaminant databases: the CRAPome [30] and the contaminant databases from MaxQuant. During the process, we performed *in silico* digestion of *Pfu* proteins and removed human proteins that contained *Pfu* peptides from the reference database. We then input the *Pfu* mass spectra and the resulting database into a database search algorithm. We consider a target PSM as true if the database search algorithm reports its master protein as from *Pfu* or the two contaminants and false if from the human. The *in silico* digestion was performed in Python using the `pyteomics.parser` function from `pyteomics` with the following settings: Trypsin digestion, two allowed missed cleavages, minimum peptide length of six [90, 91]

### 3.2. Phospho-proteomics AML datasets generation

Frozen cell lysates were further diluted with 7.2 M guanidinium hydrochloride (GuHCl) with 100 mM ammonium bicarbonate, reduced with 10 mM DTT for 30 minutes at 56$^\circ$C, and alkylated with 22.5

mM iodoacetamide for 30 minutes protected from light. After diluting GuHCl to a concentration of 1 M, proteins were digested overnight with trypsin at 37°C. Peptides were desalted by C18, dried by vacuum centrifugation, labeled with TMT stable isotope labeling reagents (ThermoFisher Scientific, Madison, WI), combined, and dried by vacuum centrifugation. Combined labeled peptides were desalted to remove labeling by-products, and phosphopeptides were enriched by immobilized metal affinity chromatography as described (PMID: 27365422). Enriched phosphopeptides were analyzed by RP-SAX-RP at a depth of 13 fractions as described (PMID: 27365422).

## 4. Public data used in this study

- The raw MS data files of the TNBC dataset are available at the PRoteomics IDEntifications Database (PRIDE) with the dataset identifier PXD002735 [92].
- The raw MS data files of the nonphospho AML dataset are available at the (PRIDE) with the dataset identifier PXD008307 [92].

## 5. Implementation of database search algorithms

### 5.1. On the proteomics standard

**Byonic, SEQUEST, and Mascot** Byonic, SEQEUST, and Mascot were each run in Proteome Discoverer 2.3.0.523 (ThermoScientific). The following settings were used for all $5$ database search algorithms: $10$ ppm precursor tolerance; $0.6$ Da fragment tolerance; static modifications: methylthio (C); dynamic modifications: deamination (NQ), oxidation (M). Percolator was used in conjunction with both SEQUEST and Mascot, and the target decoy mode was set to separate. To acquire the total list of identified PSMs, peptides, and proteins, internal FDRs for all database search algorithms were set to $100\%$.

**MaxQuant** MaxQuant was implemented with the following settings: $10$ ppm match tolerance; $0.6$ Da fragment tolerance; static modifications: methylthio (C); dynamic modifications: deamination (NQ), oxidation (M); second peptide search: True. To acquire the total list of identified PSMs, peptides, and proteins, the internal FDR was set to $100\%$. MaxQuant outputs a posterior error probability (PEP) for each target PSM and decoy PSM.

**MS-GF+** MS-GF+ was implemented with the following settings: $10$ ppm match tolerance; static modifications: methylthio (C); dynamic modifications: deamination (NQ), oxidation (M). To acquire the total lists of identified PSMs, peptides, and proteins, the internal FDR was set to $100\%$.

### 5.2. On the phospho AML datasets

**Byonic, SEQUEST, and Mascot** The phospho AML spectra were searched with the following settings: $10$ ppm precursor tolerance; $0.02$ Da fragment tolerance; static modifications: TMT6plex (N-term, K), Carbamindomethyl (C); dynamic modifications: Oxidation (M), Phopho (STY).

**MaxQuant** MaxQuant was implemented with the following settings.

- Group-specific parameters:
  - type: reporter ion MS2, isobaric labels 6plex TMT, filter by PIF (minimal reported PIF $0.75$);

  - modifications: variable modifications including oxidation (M) and phosphorylation (Y), fixed modification carbamidomethyl (C), maximal number of modifications per peptide $5$;
  - instrument: orbitrap with default parameters except that the first search peptide tolerance is set to $10$ ppm;
  - digestion: enzyme Trypsin/P, missed cleavage $2$.

- Global parameters:

  - sequences: contaminants FALSE, minimal peptide length $6$, maximal peptide mass (Da) $10000$ Da, and the rest parameters are default;
  - advanced identification: use second peptides (default);
  - MS/MS-ITMS: all default parameters except that ITMS MS/MS match tolerance $0.6$ Da;
  - identification: PSM FDR and protein FDR $1$, with rest parameters set to default;
  - protein quantification: set "Use only unmodified peptides and..." and "Advanced ratio estimation" to false;
  - MS/MS FTMS: FTMS MS/MS match tolarance $10$ ppm and the rest parameters are set to default.

MaxQuant outputs a posterior error probability (PEP) for each target PSM and decoy PSM.

**MS-GF+** MS-GF+ was implemented with the following settings: $10$ ppm precursor tolerance; search decoy database: 1 (true); instrument ID: 1 (Orbitrap/FTICR/Lumos); Enzyme ID: 1 (Trypsin); protocol ID: 4(TMT); output additional features: 1 (true); maximum missed cleavages: 2; maximum number of variable modifications per peptide: 4; variable modifications including oxidation (M) and phosphorylation (Y); fixed modification carbamidomethyl (C), TMT6plex (K and N-term). To acquire the total lists of identified PSMs, peptides, and proteins, the internal FDR was set to $100\%$.

### 5.3. On the TNBC dataset

**Byonic, SEQUEST, and Mascot** The Genistein spectra were searched with the following settings: $20$ ppm precursor tolerance; $0.02$ Da fragment tolerance; static modifications: TMT6plex (N-term, K), Carbamindomethyl (C); dynamic modifications: Oxidation (M), Phopho (STY).

**MaxQuant** MaxQuant was implemented with the following settings.

- Group-specific parameters:

  - type: reporter ion MS2, isobaric labels 6plex TMT, filter by PIF (minimal reported PIF $0.75$);
  - modifications: variable modifications including oxidation (M) and phosphorylation (Y), fixed modification carbamidomethyl (C), maximal number of modifications per peptide $5$;
  - instrument: orbitrap with default parameters except that the first search peptide tolerance is set to $20$ ppm;
  - digestion: enzyme Trypsin/P, missed cleavage $2$.

- Global parameters:

  - sequences: contaminants FALSE, minimal peptide length $6$, maximal peptide mass (Da) $10000$ Da, and the rest parameters are default;

- advanced identification: use second peptides (default);
- MS/MS-ITMS: all default parameters except that ITMS MS/MS match tolerance $0.6$ Da;
- identification: PSM FDR and protein FDR $1$, with rest parameters set to default;
- protein quantification: set "Use only unmodified peptides and..." and "Advanced ratio estimation" to false;
- MS/MS FTMS: FTMS MS/MS match tolarance $20$ ppm and the rest parameters are set to default.

MaxQuant outputs a posterior error probability (PEP) for each target PSM and decoy PSM.

**MS-GF+** MS-GF+ was implemented with the following settings: $20$ ppm precursor tolerance; search decoy database: 1 (true); instrument ID: 1 (Orbitrap/FTICR/Lumos); Enzyme ID: 1 (Trypsin); protocol ID: 4(TMT); output additional features: 1 (true); maximum missed cleavages: 2; maximum number of variable modifications per peptide: 4; variable modifications including oxidation (M) and phosphorylation (Y); fixed modification carbamidomethyl (C), TMT6plex (K and N-term). To acquire the total lists of identified PSMs, peptides, and proteins, the internal FDR was set to $100\%$.

### 5.4. On the non-phospho AML dataset

**Byonic, SEQUEST, and Mascot** The nonphospho AML spectra were searched with the following settings: $10$ ppm precursor tolerance; $0.6$ Da fragment tolerance; digestion enzyme: Lys-c; static modifications: TMT6plex (N-term, K), carbamindomethyl (C); dynamic modifications: oxidation (M).

**MaxQuant** MaxQuant was implemented with the following settings.

- Group-specific parameters:
  - type: reporter ion MS2, isobaric labels 6plex TMT, filter by PIF (minimal reported PIF $0.75$);
  - modifications: variable modifications oxidation (M), fixed modification carbamidomethyl (C), maximal number of modifications per peptide $5$;
  - instrument: orbitrap with default parameters except that the first search peptide tolerance is set to $10$ ppm;
  - digestion: enzyme Lysc/P, missed cleavage $2$.
- Global parameters:
  - sequences: contaminants FALSE, minimal peptide length $6$, maximal peptide mass (Da) $10000$ Da, and the rest parameters are default;
  - advanced identification: use second peptides (default);
  - MS/MS-ITMS: all default parameters except that ITMS MS/MS match tolerance $0.6$ Da;
  - identification: PSM FDR and protein FDR $1$, with rest parameters set to default;
  - protein quantification: set "Use only unmodified peptides and..." and "Advanced ratio estimation" to false;
  - MS/MS FTMS: FTMS MS/MS match tolarance $20$ ppm and the rest parameters are set to default.

MaxQuant outputs a posterior error probability (PEP) for each target PSM and decoy PSM.

**MS-GF+** MS-GF+ was implemented with the following settings: $10$ ppm precursor tolerance; search decoy database: 1 (true); instrument ID: 1 (Orbitrap/FTICR/Lumos); Enzyme ID: 3 (Lys-C); protocal ID: 4(TMT); output additional features: 1 (true); maximum missed cleavages: 2; maximum number of variable modifications per peptide: 4; variable modifications oxidation (M); fixed modification carbamidomethyl (C), TMT6plex (K and N-term). To acquire the total lists of identified PSMs, peptides, and proteins, the internal FDR was set to $100\%$.

## 6. Existing aggregation methods

### 6.1. Description of methods

**Scaffold** Scaffold (Proteome Software, Portland, Oregan, USA) adopts a Bayesian approach to aggregate probabilities of the individual database search algorithm results into a single probability for each PSM. One of its key step is to generate for each database search algorithm a peptide probability model that estimates the probability of an individual spectrum being correctly assigned to a peptide based on that database search algorithm's score. To realize this, Scaffold designs a different statistical model for the internal scores from each database search algorithm [25], making it difficult to generalize its approach to other database search algorithms. Scaffold supports Byonic (Protein Metrics), Mascot (Matrix Science), Mascot Distiller (Matrix Science), MaxQuant/Andromeda (Max Planck Institute), Peaks (Bioinformatics Solutions), and Proteome Discoverer (Thermo Fisher Scientific) database search algorithms including Byonic, SEQUEST, and Mascot.

**MSblender** MSblender is an open-source software that uses a probability mixture model to model the scores of correct and incorrect PSMs. In particular, the correct PSM scores across database search algorithms are assumed to follow a two-component (by default) multivariate Gaussian [21]. Search algorithms that are compatible with MSblender include SEQUEST (Thermo Fisher Scientific), X!Tandem [93], OMSSA [94], InsPecT [95], MyriMatch [96], MSGFDB [97].

**ConsensusID** ConsensusID is part of the OpenMS Proteomics Pipeline [98]. It adopts a probabilistic approach to aggregate the top-scoring PSM results from several database search algorithms. A key feature of this tool is its sequence similarity scoring mechanism, which is a method to estimate the scores for PSMs in cases when the peptide is missing from the high-ranking results of a database search algorithm. It involves fitting the scores from each database search algorithm as a two-component mixture model. The two components are a Gumbel distribution for the incorrect PSMs and a normal distribution for the correct PSMs [19]. Although the paper Nahnsen et al. [19] claims that ConsensusID supports all database search algorithms, the OpenMS pipeline only supports the following database search algorithms: Comet [99], CompNovo [100], Crux [101], Mascot (Thermo Fisher Scientific), MS-GF+ [10], MyriMatch [96], OMSSA [94], PepNovo [102], and X!Tandem [93].

**PepArML** PepArML is an unsupervised, model-free, machine-learning-based method to aggregate search results [12]. It is compatible with Mascot [7], Tandem [93] with native, K-score, and s-score [103] scoring, OMSSA [94], MyriMatch [96], InSpecT [95], and MS-GF [104] spectral probability scores.

**iProphet** The iProphet software is an open-source software within the Trans Proteomic Pipeline (TPP) suite [20]. It is used between PeptideProphet [105] and ProteinProphet [106]. It calculates peptide level probabilities via mixture models [20]. The TPP suite is compatible with COMET [99], X!Tandem [93],

SEQUEST (Thermo Fisher Scientific), MS-GF+ [10], InSpecT [95], OMSSA [94], MyriMatch [96], ProbID [107], Mascot (Matrix Science), and Phenyx [108].

### 6.2. Implementation of Scaffold

We used Scaffold to combine the outputs of Byonic, Mascot, SEQUEST, MaxQuant, and MS-GF+ on the proteomics standard. For each combination of database search algorithms, the result files were inputted into Scaffold Q+ (version 4.10.0, Proteome Software Inc., Portland, OR) to generate peptide and protein identification probabilities. Peptide probabilities were assigned by the Scaffold Local FDR algorithm, protein groups were generated using standard experiment-wide protein grouping, and protein probabilities were assigned by the Protein Prophet algorithm [106]. To compare Scaffold with APIR which aims to control the FDR at the PSM level, we implemented Scaffold in two ways. In the first implementation, we set both the peptide threshold and the protein threshold to be $q$ FDR, the FDR threshold of APIR. In the second comparison, we set the peptide threshold to be $q$ FDR and varied the protein threshold among all default thresholds: $20\%$ ($1-$ PEP), $50\%$, $80\%$, $90\%$, $95\%$, $99\%$, $99.9\%$, $1\%$ FDR, $2\%$ FDR, $3\%$ FDR, $5\%$ FDR and $10\%$ FDR to maximize the number of identified peptides.

### 6.3. Implementation of ConsensusID

We used OPENMS (version 2.6.0) to combine the search results of Byonic, Mascot, SEQUEST, MaxQuant, and MS-GF+ on the proteomics standard dataset using the following procedure.

- We first converted xlsx output from search engines to IDXML files in Python.
- We used TOPPAS GUI: the openms proteomics pipeline assistant with the following nodes: `input file`, `merge`, `IDMerger`, `consensusID`, and `output file`.
- We used the default settings of IDMerger.
- The settings of consensusID: set "per spectrum" to TRUE, with the rest parameters set to the default.

## 7. DE peptides analysis of the phospho AML1 dataset

Here we describe how we performed DE analysis on the phospho AML1 dataset. This dataset contains six bone marrow samples: one LSC enriched sample and one LSC depleted sample from patient P5337, two LSC enriched samples and one LSC depleted sample from patient P5340, and one control.

Using all six samples from the phospho AML1 dataset, we first applied APIR to combine the search results by MaxQuant and MS-GF+. Then we applied APIR to adjust the search results of MaxQuant and MS-GF+ separately. Next, we selected fours sample: the two samples from P5337 and the LSC depleted sample from P5340 and one of the two LSC enriched samples from patient P5340, as shown in Fig. 5a. We treated the LSC enriched samples and the LSC depleted samples as from two conditions and applied DESeq2 with FDR threshold $5\%$ for DE analysis [85]. We use the R package `DESeq2` version 1.28.1.

# Software and code

The **APIR** `R` package is available at `https://github.com/yiling0210/APIR`. The code and processed data for reproducing the figures are available at XXX.

## Data deposition

We plan to submit the in-house generated raw data of proteomics standard and phospho-proteomics AML datasets to ProteomeXchange http://www.proteomexchange.org/.

## Funding

NCI K08 CA201591

Margaret Early Memorial Research Trust

Pediatric Cancer Research Foundation

And NCI P30CA033572, the NCI Cancer Center Support Grant This work was supported by the following grants: NIH-NCI T32LM012424 (to Y.E.C.); NCI K08 CA201591, Margaret Early Memorial Research Trust, and Pediatric Cancer Research Foundation (to L.D.W.); NCI P30CA033572, the NCI Cancer Center Support Grant (to the mass spectrometry facility at City of Hope); NIH/NIGMS R01GM120507 and R35GM140888, NSF DBI-1846216 and DMS-2113754, Johnson & Johnson WiSTEM2D Award, Sloan Research Fellowship, and UCLA David Geffen School of Medicine W.M. Keck Foundation Junior Faculty Award (to J.J.L.).

## Conflicts of interests

L.D.W. holds equity in Magenta Therapeutics.

## Acknowledgements

The authors appreciate the comments and feedback from all members of the Junction of Statistics and Biology at UCLA (http://jsb.ucla.edu).

## References

[1] Oscar Alzate. *Neuroproteomics*. CRC Press, 2009.

[2] John M Koomen et al. “Proteomic contributions to personalized cancer care”. In: *Molecular & Cellular Proteomics* 7.10 (2008), pp. 1780–1794.

[3] Mark A Eckert et al. “Proteomics reveals NNMT as a master metabolic regulator of cancer-associated fibroblasts”. In: *Nature* 569.7758 (2019), pp. 723–728.

[4] Gali Yanovich et al. “Clinical proteomics of breast cancer reveals a novel layer of breast cancer classification”. In: *Cancer research* 78.20 (2018), pp. 6001–6010.

[5] Marjorie L Fournier et al. “Multidimensional separations-based shotgun proteomics”. In: *Chemical reviews* 107.8 (2007), pp. 3654–3686.

[6] Michael P Washburn, Dirk Wolters, and John R Yates. “Large-scale analysis of the yeast proteome by multidimensional protein identification technology”. In: *Nature biotechnology* 19.3 (2001), pp. 242–247.

[7] David N Perkins et al. “Probability-based protein identification by searching sequence databases using mass spectrometry data”. In: *ELECTROPHORESIS: An International Journal* 20.18 (1999), pp. 3551–3567.

[8] Jürgen Cox and Matthias Mann. "MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification". In: *Nature biotechnology* 26.12 (2008), pp. 1367–1372.

[9] Marshall Bern, Yong J Kil, and Christopher Becker. "Byonic: advanced peptide and protein identification software". In: *Current protocols in bioinformatics* 40.1 (2012), pp. 13–20.

[10] Sangtae Kim and Pavel A Pevzner. "MS-GF+ makes progress towards a universal database search tool for proteomics". In: *Nature communications* 5 (2014), p. 5277.

[11] Joshua E Elias and Steven 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.

[12] Nathan Edwards, Xue Wu, and Chau-Wen Tseng. "An unsupervised, model-free, machine-learning combiner for peptide identifications from tandem mass spectra". In: *Clinical Proteomics* 5.1 (2009), pp. 23–36.

[13] Kyowon Jeong, Sangtae Kim, and Nuno Bandeira. "False discovery rates in spectral identification". In: *BMC bioinformatics* 13.16 (2012), pp. 1–15.

[14] Kristen Emery et al. "Multiple competition-based FDR control for peptide detection". In: *arXiv preprint arXiv:1907.01458* (2019).

[15] Lukas Käll et al. "Posterior error probabilities and false discovery rates: two sides of the same coin". In: *Journal of proteome research* 7.01 (2008), pp. 40–44.

[16] Lukas Käll et al. "Assigning significance to peptides identified by tandem mass spectrometry using decoy databases". In: *Journal of proteome research* 7.01 (2008), pp. 29–34.

[17] Oliver Serang and William Noble. "A review of statistical methods for protein identification using tandem mass spectrometry". In: *Statistics and its interface* 5.1 (2012), p. 3.

[18] Alexey 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.

[19] Sven Nahnsen et al. "Probabilistic consensus scoring improves tandem mass spectrometry peptide identification". In: *Journal of proteome research* 10.8 (2011), pp. 3332–3343.

[20] David Shteynberg et al. "iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates". In: *Molecular & cellular proteomics* 10.12 (2011).

[21] Taejoon Kwon et al. "MSblender: A probabilistic approach for integrating peptide identifications from multiple database search engines". In: *Journal of proteome research* 10.7 (2011), pp. 2949–2958.

[22] David Shteynberg et al. "Combining results of multiple search engines in proteomics". In: *Molecular & Cellular Proteomics* 12.9 (2013), pp. 2383–2393.

[23] Tommi Välikangas, Tomi Suomi, and Laura L Elo. "A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation". In: *Briefings in bioinformatics* 19.6 (2018), pp. 1344–1355.

[24] Ruben K Dagda, Tamanna Sultana, and James Lyons-Weiler. "Evaluation of the consensus of four peptide identification algorithms for tandem mass spectrometry based proteomics". In: *Journal of proteomics & bioinformatics* 3 (2010), p. 39.

[25] Brian C Searle, Mark Turner, and Alexey I Nesvizhskii. “Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies”. In: *The Journal of Proteome Research* 7.1 (2008), pp. 245–253.

[26] Dominique Tessier et al. “Origin of disagreements in tandem mass spectra interpretation by search engines”. In: *Journal of proteome research* 15.10 (2016), pp. 3481–3488.

[27] David C Wedge et al. “FDRAnalysis: a tool for the integrated analysis of tandem mass spectrometry identification results from multiple search engines”. In: *Journal of proteome research* 10.4 (2011), pp. 2088–2094.

[28] Xinzhou Ge et al. “Clipper: p-value-free FDR control on high-throughput data from two conditions”. In: *bioRxiv* (2020).

[29] “UniProt: the universal protein knowledgebase in 2021”. In: *Nucleic Acids Research* 49.D1 (2021), pp. D480–D489.

[30] Dattatreya Mellacheruvu et al. “The CRAPome: a contaminant repository for affinity purification–mass spectrometry data”. In: *Nature methods* 10.8 (2013), pp. 730–736.

[31] Joao A Paulo. “Practical and efficient searching in proteomics: a cross engine comparison”. In: *Webmedcentral* 4.10 (2013).

[32] Simon Raffel et al. “BCAT1 restricts $\alpha$KG levels in AML stem cells leading to IDH mut-like DNA hypermethylation”. In: *Nature* 551.7680 (2017), pp. 384–388.

[33] Yi Fang et al. “Quantitative phosphoproteomics reveals genistein as a modulator of cell cycle and DNA damage response pathways in triple-negative breast cancer cells”. In: *International journal of oncology* 48.3 (2016), pp. 1016–1028.

[34] Sean J Humphrey, David E James, and Matthias Mann. “Protein phosphorylation: a major switch mechanism for metabolic regulation”. In: *Trends in Endocrinology & Metabolism* 26.12 (2015), pp. 676–687.

[35] Wen-Wei Tsai et al. “TRIM24 links a non-canonical histone signature to breast cancer”. In: *Nature* 468.7326 (2010), pp. 927–932.

[36] Zhibin Cui et al. “TRIM24 overexpression is common in locally advanced head and neck squamous cell carcinoma and correlates with aggressive malignant phenotypes”. In: *PloS one* 8.5 (2013), e63887.

[37] Anna C Groner et al. “TRIM24 is an oncogenic transcriptional activator in prostate cancer”. In: *Cancer cell* 29.6 (2016), pp. 846–858.

[38] Haiying Li et al. “Overexpression of TRIM24 correlates with tumor progression in non-small cell lung cancer”. In: *PloS one* 7.5 (2012), e37657.

[39] Xiao Liu et al. “Overexpression of TRIM24 is associated with the onset and progress of human hepatocellular carcinoma”. In: *PloS one* 9.1 (2014), e85462.

[40] Jianwei Wang et al. “Knockdown of tripartite motif containing 24 by lentivirus suppresses cell growth and induces apoptosis in human colorectal cancer cells”. In: *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics* 22.1 (2014), pp. 39–45.

[41] C Li et al. “Knockdown of TRIM24 suppresses growth and induces apoptosis in acute myeloid leukemia through downregulation of Wnt/GSK-3$\beta$/$\beta$-catenin signaling”. In: *Human & Experimental Toxicology* 39.12 (2020), pp. 1725–1736.

[42] Yan Ye et al. “PI (4, 5) P2 5-phosphatase A regulates PI3K/Akt signalling and has a tumour suppressive role in human melanoma”. In: *Nature communications* 4.1 (2013), pp. 1–15.

[43] Laura J Van't Veer et al. "Gene expression profiling predicts clinical outcome of breast cancer". In: *nature* 415.6871 (2002), pp. 530–536.

[44] Sang-Uk Han et al. "Loss of the Smad3 expression increases susceptibility to tumorigenicity in human gastric cancer". In: *Oncogene* 23.7 (2004), pp. 1333–1341.

[45] Patrick Ming-Kuen Tang et al. "Smad3 promotes cancer progression by inhibiting E4BP4-mediated NK cell development". In: *Nature communications* 8.1 (2017), pp. 1–15.

[46] C Liu et al. "MicroRNA-34b inhibits pancreatic cancer metastasis through repressing Smad3". In: *Current molecular medicine* 13.4 (2013), pp. 467–478.

[47] Maj Petersen et al. "Smad2 and Smad3 have opposing roles in breast cancer bone metastasis by differentially affecting tumor angiogenesis". In: *Oncogene* 29.9 (2010), pp. 1351–1361.

[48] Nicholas I Fleming et al. "SMAD2, SMAD3 and SMAD4 mutations in colorectal cancer". In: *Cancer research* 73.2 (2013), pp. 725–735.

[49] Jianfei Xue et al. "Sustained activation of SMAD3/SMAD4 by FOXM1 promotes TGF-$\beta$–dependent cancer metastasis". In: *The Journal of clinical investigation* 124.2 (2014), pp. 564–579.

[50] Konstanze Döhner and Hartmut Döhner. "Molecular characterization of acute myeloid leukemia". In: *Haematologica* 93.7 (2008), pp. 976–982.

[51] Raed A Alharbi et al. "The role of HOX genes in normal hematopoiesis and acute leukemia". In: *Leukemia* 27.5 (2013), pp. 1000–1008.

[52] A Renneville et al. "Cooperating gene mutations in acute myeloid leukemia: a review of the literature". In: *leukemia* 22.5 (2008), pp. 915–931.

[53] Antonella Di Costanzo et al. "The HDAC inhibitor SAHA regulates CBX2 stability via a SUMO-triggered ubiquitin-mediated pathway in leukemia". In: *Oncogene* 37.19 (2018), pp. 2559–2572.

[54] M Terol et al. "HBZ-mediated shift of JunD from growth suppressor to tumor promoter in leukemic cells by inhibition of ribosomal protein S25 expression". In: *Leukemia* 31.10 (2017), pp. 2235–2243.

[55] Kristopher R Bosse et al. "Identification of GPC2 as an oncoprotein and candidate immunotherapeutic target in high-risk neuroblastoma". In: *Cancer cell* 32.3 (2017), pp. 295–309.

[56] Hemanth Tummala et al. "DNAJC21 mutations link a cancer-prone bone marrow failure syndrome to corruption in 60S ribosome subunit maturation". In: *The American Journal of Human Genetics* 99.1 (2016), pp. 115–124.

[57] Jia Liu et al. "ZFP36L2, a novel AML1 target gene, induces AML cells apoptosis and inhibits cell proliferation". In: *Leukemia research* 68 (2018), pp. 15–21.

[58] Eisaku Iwanaga et al. "Mutation in the RNA binding protein TIS11D/ZFP36L2 is associated with the pathogenesis of acute leukemia". In: *International journal of oncology* 38.1 (2011), pp. 25–31.

[59] Lijuan Chen et al. "LncRNA MAGI2-AS3 inhibits the self-renewal of leukaemic stem cells by promoting TET2-dependent DNA demethylation of the LRIG1 promoter in acute myeloid leukaemia". In: *RNA biology* 17.6 (2020), pp. 784–793.

[60] Luke W Thomas et al. "CHCHD4 regulates tumour proliferation and EMT-related phenotypes, through respiratory chain-mediated metabolism". In: *Cancer & metabolism* 7.1 (2019), pp. 1–17.

[61] David Ross et al. "Cell-specific activation and detoxification of benzene metabolites in mouse and human bone marrow: identification of target cells and a potential role for modulation of apoptosis in benzene toxicity." In: *Environmental health perspectives* 104.suppl 6 (1996), pp. 1177–1182.

[62] William B Slayton et al. “The first-appearance of neutrophils in the human fetal bone marrow cavity”. In: *Early human development* 53.2 (1998), pp. 129–144.

[63] Diane G Schattenberg et al. “Peroxidase activity in murine and human hematopoietic progenitor cells: potential relevance to benzene-induced toxicity.” In: *Molecular pharmacology* 46.2 (1994), pp. 346–351.

[64] Michelle W Wong-Brown et al. “Prevalence of BRCA1 and BRCA2 germline mutations in patients with triple-negative breast cancer”. In: *Breast cancer research and treatment* 150.1 (2015), pp. 71–80.

[65] DG Evans et al. “Prevalence of BRCA1 and BRCA2 mutations in triple negative breast cancer”. In: *Journal of medical genetics* 48.8 (2011), pp. 520–522.

[66] E Comen et al. “Relative contributions of BRCA1 and BRCA2 mutations to “triple-negative” breast cancer in Ashkenazi Women”. In: *Breast cancer research and treatment* 129.1 (2011), pp. 185–190.

[67] C Villarreal-Garza et al. “The prevalence of BRCA1 and BRCA2 mutations among young Mexican women with triple-negative breast cancer”. In: *Breast cancer research and treatment* 150.2 (2015), pp. 389–394.

[68] Rachel Greenup et al. “Prevalence of BRCA mutations among women with triple-negative breast cancer (TNBC) in a genetic counseling cohort”. In: *Annals of surgical oncology* 20.10 (2013), pp. 3254–3258.

[69] Johanna Tommiska et al. “The DNA damage signalling kinase ATM is aberrantly reduced or lost in BRCA1/BRCA2-deficient and ER/PR/ERBB2-triple-negative breast cancer”. In: *Oncogene* 27.17 (2008), pp. 2501–2506.

[70] Toshiyasu Taniguchi et al. “Disruption of the Fanconi anemia–BRCA pathway in cisplatin-sensitive ovarian tumors”. In: *Nature medicine* 9.5 (2003), pp. 568–574.

[71] Carmen J Marsit et al. “Inactivation of the Fanconi anemia/BRCA pathway in lung and oral cancers: implications for treatment and survival”. In: *Oncogene* 23.4 (2004), pp. 1000–1004.

[72] Helong Zhao et al. “Endothelial Robo4 suppresses breast cancer growth and metastasis through regulation of tumor angiogenesis”. In: *Molecular oncology* 10.2 (2016), pp. 272–281.

[73] Rebecca Marlow et al. “Vascular Robo4 restricts proangiogenic VEGF signaling in breast”. In: *Proceedings of the National Academy of Sciences* 107.23 (2010), pp. 10520–10525.

[74] Steven Suchting et al. “Soluble Robo4 receptor inhibits in vivo angiogenesis and endothelial cell migration”. In: *The FASEB Journal* 19.1 (2005), pp. 121–123.

[75] Xiaodong Zhuang et al. “Robo4 vaccines induce antibodies that retard tumor growth”. In: *Angiogenesis* 18.1 (2015), pp. 83–95.

[76] Minoru Kanehisa and Susumu Goto. “KEGG: kyoto encyclopedia of genes and genomes”. In: *Nucleic acids research* 28.1 (2000), pp. 27–30.

[77] Minoru Kanehisa. “Toward understanding the origin and evolution of cellular organisms”. In: *Protein Science* 28.11 (2019), pp. 1947–1951.

[78] Minoru Kanehisa et al. “KEGG: integrating viruses and cellular organisms”. In: *Nucleic Acids Research* 49.D1 (2021), pp. D545–D551.

[79] JA Costoya, RM Hobbs, and PP Pandolfi. “Cyclin-dependent kinase antagonizes promyelocytic leukemia zinc-finger through phosphorylation”. In: *Oncogene* 27.27 (2008), pp. 3789–3796.

[80] CH Yam, TK Fung, and RYC Poon. “Cyclin A in cell cycle control and cancer”. In: *Cellular and Molecular Life Sciences CMLS* 59.8 (2002), pp. 1317–1326.

[81] Ida RK Bukholm, Geir Bukholm, and Jahn M Nesland. “Over-expression of cyclin A is highly associated with early relapse and reduced survival in patients with primary breast carcinomas”. In: *International journal of cancer* 93.2 (2001), pp. 283–287.

[82] Marcos Malumbres and Mariano Barbacid. “Cell cycle, CDKs and cancer: a changing paradigm”. In: *Nature reviews cancer* 9.3 (2009), pp. 153–166.

[83] Erica K Cassimere, Claire Mauvais, and Catherine Denicourt. “p27Kip1 is required to mediate a G1 cell cycle arrest downstream of ATM following genotoxic stress”. In: *PLoS One* 11.9 (2016), e0162806.

[84] Byung-Kwon Choi et al. “WIP1 dephosphorylation of p27Kip1 Serine 140 destabilizes p27Kip1 and reverses anti-proliferative effects of ATM phosphorylation”. In: *Cell Cycle* 19.4 (2020), pp. 479–491.

[85] Michael I Love, Wolfgang Huber, and Simon Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome biology* 15.12 (2014), pp. 1–21.

[86] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. In: *Bioinformatics* 26.1 (2010), pp. 139–140.

[87] Matthew E Ritchie et al. “limma powers differential expression analyses for RNA-sequencing and microarray studies”. In: *Nucleic acids research* 43.7 (2015), e47–e47.

[88] Yohann Couté, Christophe Bruley, and Thomas Burger. “Beyond Target–Decoy Competition: Stable Validation of Peptide and Protein Identifications in Mass Spectrometry-Based Discovery Proteomics”. In: *Analytical Chemistry* 92.22 (2020), pp. 14898–14906.

[89] Yoav Benjamini and Yosef 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.

[90] Anton A Goloborodko et al. “Pyteomics—a Python framework for exploratory data analysis and rapid software prototyping in proteomics”. In: *Journal of The American Society for Mass Spectrometry* 24.2 (2013), pp. 301–304.

[91] Lev I Levitsky et al. “Pyteomics 4.0: five years of development of a Python proteomics framework”. In: *Journal of proteome research* 18.2 (2018), pp. 709–714.

[92] Yasset Perez-Riverol 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.

[93] Robertson Craig and Ronald C Beavis. “TANDEM: matching proteins with tandem mass spectra”. In: *Bioinformatics* 20.9 (2004), pp. 1466–1467.

[94] Lewis Y Geer et al. “Open mass spectrometry search algorithm”. In: *Journal of proteome research* 3.5 (2004), pp. 958–964.

[95] Stephen Tanner et al. “InsPecT: identification of posttranslationally modified peptides from tandem mass spectra”. In: *Analytical chemistry* 77.14 (2005), pp. 4626–4639.

[96] David L Tabb, Christopher G Fernando, and Matthew C Chambers. “MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis”. In: *Journal of proteome research* 6.2 (2007), pp. 654–661.

[97] Sangtae Kim et al. “The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search”. In: *Molecular & Cellular Proteomics* 9.12 (2010), pp. 2840–2852.

[98] Hannes L Röst et al. “OpenMS: a flexible open-source software platform for mass spectrometry data analysis”. In: *Nature methods* 13.9 (2016), pp. 741–748.

[99] Jimmy K Eng, Tahmina A Jahan, and Michael R Hoopmann. “Comet: an open-source MS/MS sequence database search tool”. In: *Proteomics* 13.1 (2013), pp. 22–24.

[100] Andreas Bertsch et al. “De novo peptide sequencing by tandem MS using complementary CID and electron transfer dissociation”. In: *Electrophoresis* 30.21 (2009), pp. 3736–3747.

[101] Christopher Y Park et al. “Rapid and accurate peptide identification from tandem mass spectra”. In: *Journal of proteome research* 7.7 (2008), pp. 3022–3027.

[102] Ari Frank and Pavel Pevzner. “PepNovo: de novo peptide sequencing via probabilistic network modeling”. In: *Analytical chemistry* 77.4 (2005), pp. 964–973.

[103] Brendan MacLean et al. “General framework for developing and evaluating database scoring algorithms using the TANDEM search engine”. In: *Bioinformatics* 22.22 (2006), pp. 2830–2832.

[104] Sangtae Kim, Nitin Gupta, and Pavel A Pevzner. “Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases”. In: *Journal of proteome research* 7.8 (2008), pp. 3354–3363.

[105] Andrew Keller et al. “Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search”. In: *Analytical chemistry* 74.20 (2002), pp. 5383–5392.

[106] Alexey I Nesvizhskii et al. “A statistical model for identifying proteins by tandem mass spectrometry”. In: *Analytical chemistry* 75.17 (2003), pp. 4646–4658.

[107] Ning Zhang, Ruedi Aebersold, and Benno Schwikowski. “ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data”. In: *Proteomics* 2.10 (2002), pp. 1406–1412.

[108] Jacques Colinge et al. “OLAV: towards high-throughput tandem mass spectrometry data identification”. In: *PROTEOMICS: International Edition* 3.8 (2003), pp. 1454–1463.

# Figures

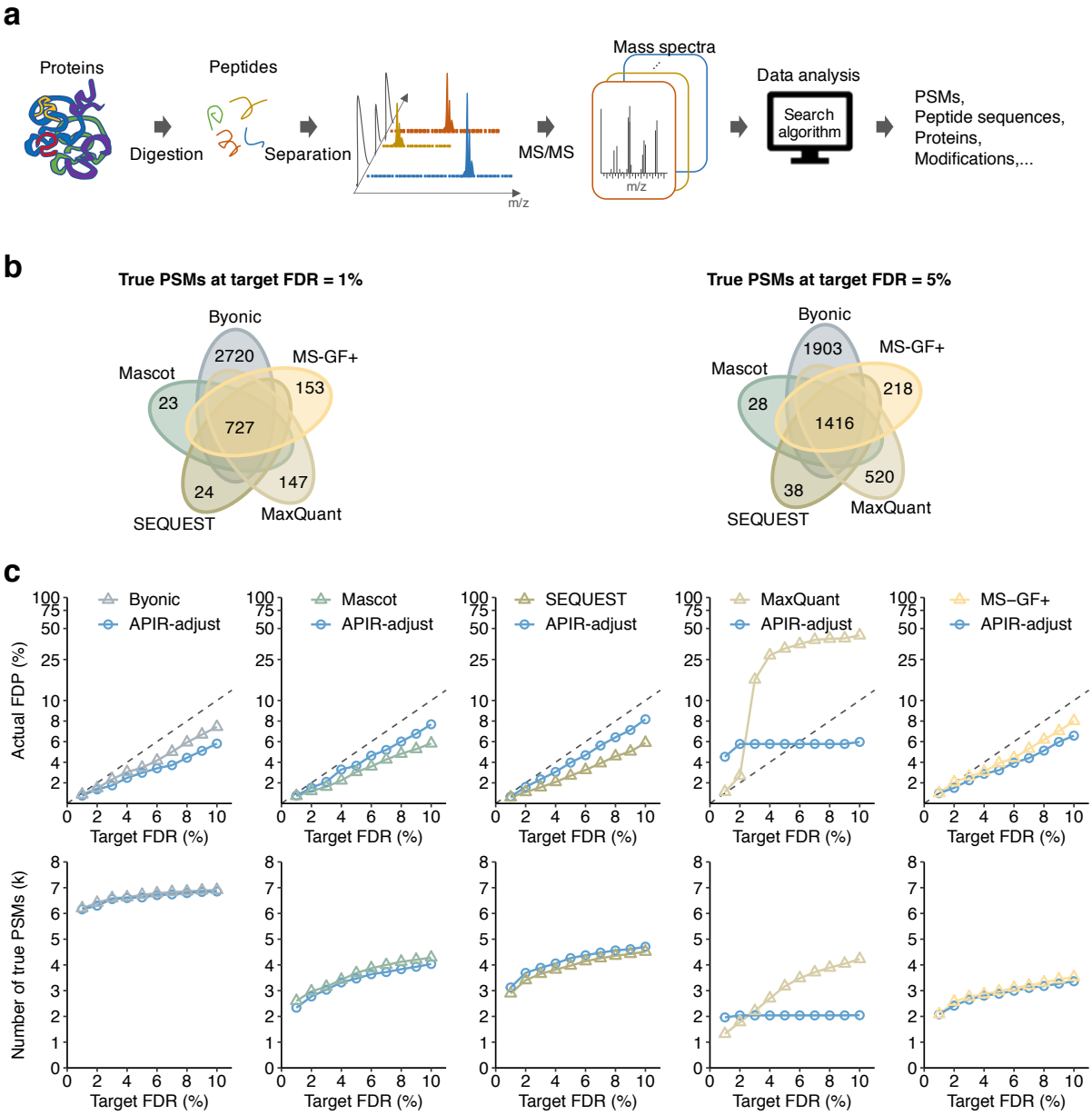



**Figure 1:** **(a)** The workflow of a typical shotgun proteomics experiment. The protein mixture is first enzymatically digested into peptides, i.e., short amino acid chains up to approximately $40$-residue long; the resulting peptide mixture is then separated and measured by tandem MS into tens of thousands of mass spectra. Each mass spectrum encodes the chemical composition of a peptide; thus, the spectrum can be used to identify the peptide's amino acid sequence and post-translational modifications, as well as to quantify the peptide's abundance with additional weight information. **(b)** Venn diagrams showing the overlap of true PSMs identified by the five database search algorithms from the proteomics standard dataset under the FDR threshold $q = 1\%$ (left) or $q = 5\%$ (right). Byonic, MaxQuant and MS-GF+ identify many unique true PSMs. **(c)** The FDP and power of each database search algorithm on the proteomics standard dataset at the FDR threshold $q \in \{1\%, 2\%, \cdots, 10\%\}$. MaxQuant fails to control the FDR, while the other four successfully control the FDR. APIR-adjust alleviates the FDR control issue of MaxQuat; for the other four database search algorithms, APIR-adjust controls the FDR and achieves similar power.

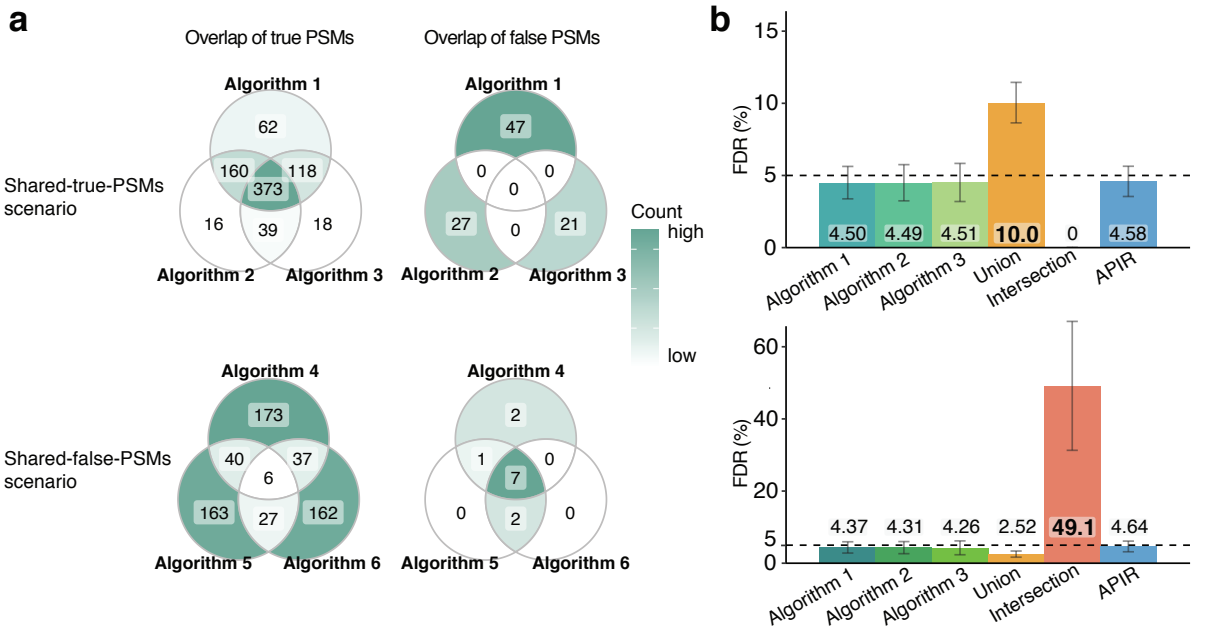
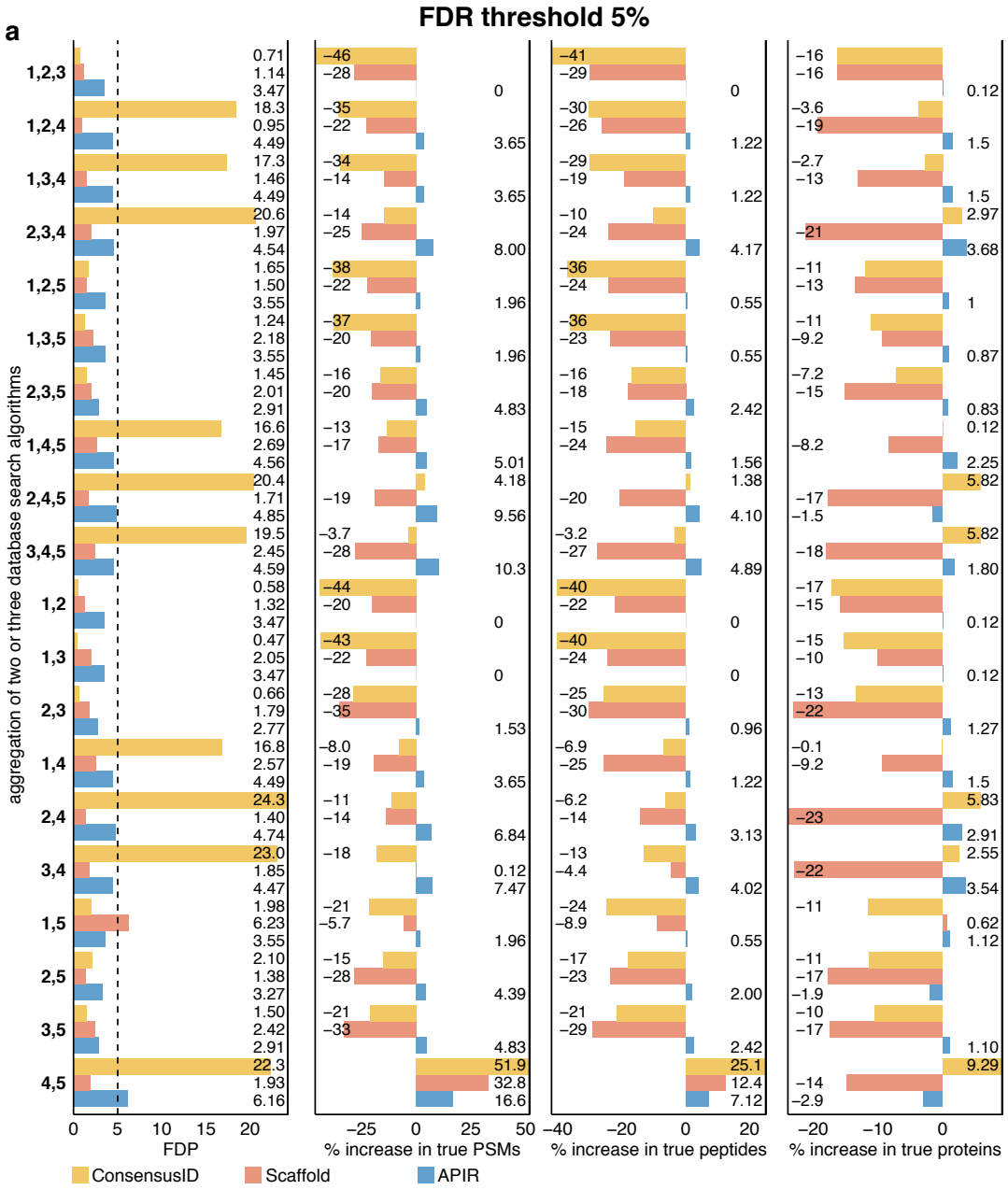



**Figure 2:** Simulation studies showing that neither intersection nor union of discovery sets (with controlled FDR) is guaranteed to maintain the FDR control. FDR control comparison of APIR, intersection, and union for aggregating three toy database search algorithms. Two scenarios are considered: the shared-true-PSMs scenario (top) and the shared-false-PSMs scenario (bottom). In the shared-true-PSMs scenario, the three database search algorithms tend to identify overlapping true PSMs but non-overlapping false PSMs. In the shared-false-PSMs scenario, the three database search algorithms tend to identify overlapping false PSMs but non-overlapping true PSMs. **(a)** Venn diagrams of true PSMs and false PSMs (identified at the FDR threshold $q = 5\%$) on one simulation dataset under each simulation scenario (top: shared-true-PSMs; bottom: shared-false-PSMs). **(b)** The FDRs of the three database search algorithms and three aggregation methods: union, intersection, and APIR. While union fails to control the FDR in the shared-true-PSMs scenario and intersection fails in the shared-false-PSMs scenario, APIR controls the FDR in both scenarios. Note that the FDR of each database search algorithm or each aggregation method is computed as the average of FDPs on $200$ simulated datasets under each scenario.




| Index | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Database search algorithm | Byonic | Mascot | SEQUEST | MaxQuant | MS-GF+ |
| Round 1 | q-thre | q-thre | q-thre | APIR-adjust | q-thre |

**Figure 3:** On the proteomics standard dataset, comparison of APIR, Scaffold, and ConsensusID at the FDR threshold $q = 5\%$ in terms of FDR control and power. We set both the peptide threshold and the protein threshold of Scaffold to be $5\%$ FDR. **(a)** FDPs (first column), the percentage increases in true PSMs (second column), the percentage increases in true peptides (third column), and the percentage increases in true proteins (fourth column) after aggregating two or three database search algorithms out of the five (Byonic, Mascot, SEQUEST, MaxQuant, and MS-GF+). The percentage increase in true PSMs/peptides/proteins is computed by treating as the baseline the maximal number of correctly identified PSMs/peptides/proteins by an individual database search algorithms in Round 1 of APIR. **(b)** The indices of database search algorithm in (a) and the implementation of APIR in Round 1. Based on the benchmarking results in Fig. 1c, in Round 1 of APIR, we applied q-value thresholding (q-thre) to Byonic, Mascot, SEQUEST, and MS-GF+, and we applied APIR-adjust to MaxQuant. In later rounds of APIR, we used APIR-adjust for FDR control.

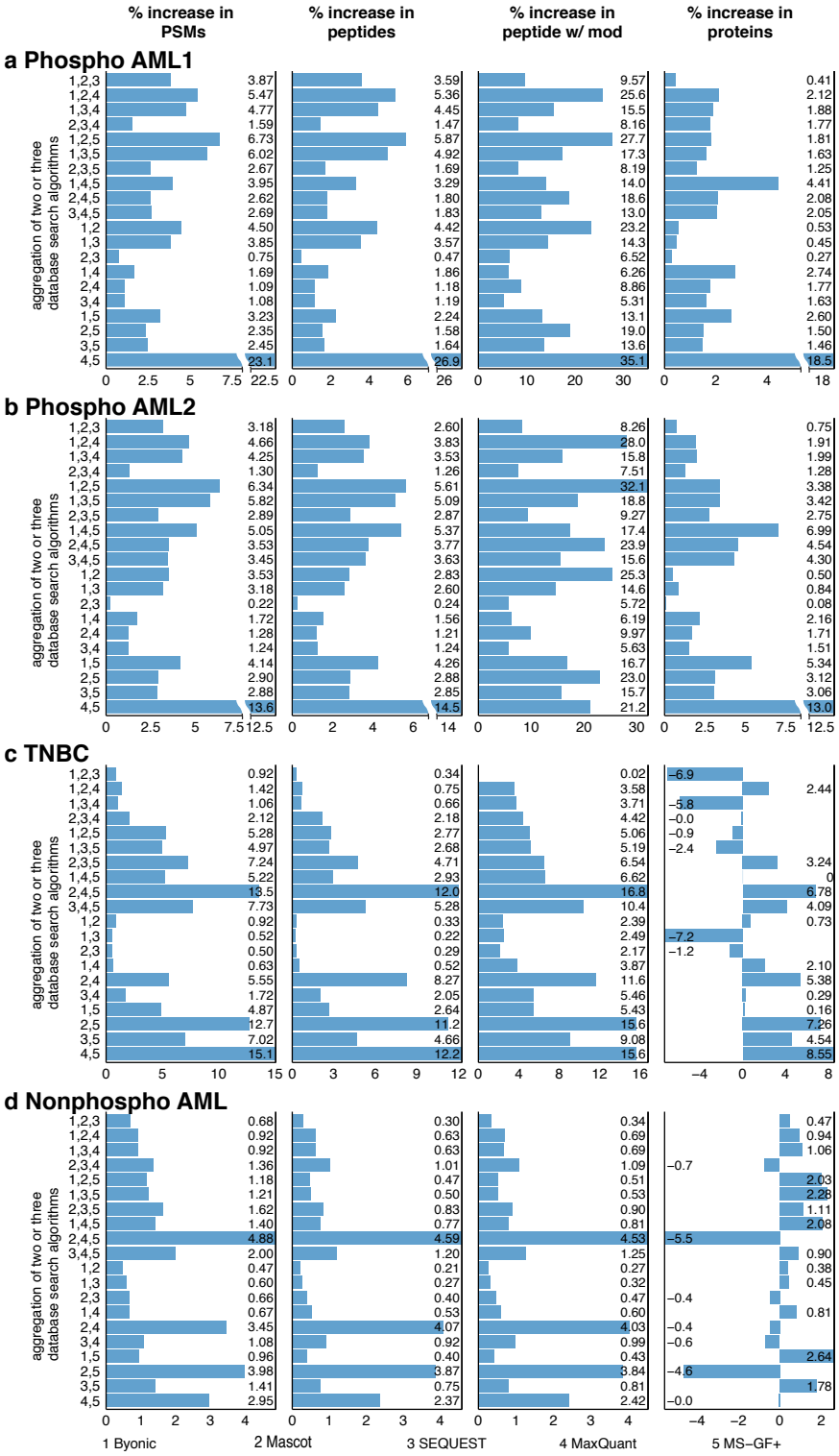



**Figure 4:** Power improvement of APIR over individual database search algorithms at the FDR threshold $q = 5\%$. The percentage increases in PSMs (first column), the percentage increases in peptides (second column), the percentage increases in peptides with modifications (third column), and the percentage increases in true proteins (fourth column) of APIR after aggregating two or three database search algorithms out of the five (Byonic, Mascot, SEQUEST, MaxQuant, and MS-GF+) at the FDR threshold $q = 5\%$ on **(a)–(b)** the phospho-proteomics AML datasets, **(c)** the nonphospho-proteomics AML dataset, and **(d)** the TNBC dataset. The percentage increase in PSMs/peptides/peptides with modifications/proteins is computed by treating as the baseline the maximal number of PSMs/peptides/peptides and modifications/proteins by an individual database search algorithm in the first round of APIR, where APIR-adjust was used for FDR control.

**a**

| Samples | Description |
|---|---|
| 1 | P5337 LSC depleted |
| 2 | P5337 LSC enriched |
| 3 | P5340 LSC depleted |
| 4 | P5340 LSC enriched |
| 5 | P5340 LSC enriched (not used for DE) |
| 6 | Control (not used for DE) |

**b**

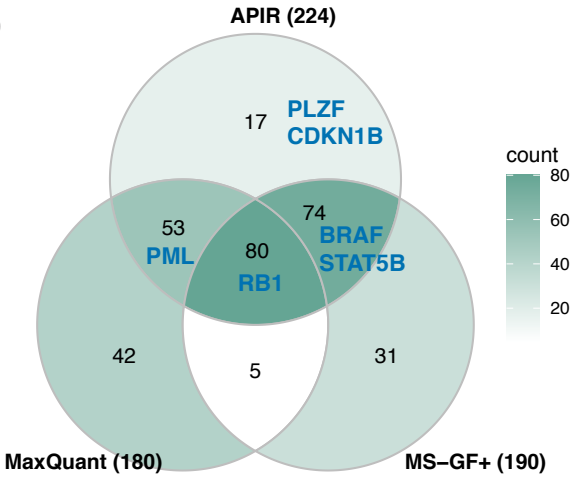



**Figure 5:** Comparison of APIR with MaxQuant and MS-GF+ by DE analysis on the phospho AML1 dataset. **(a)** Sample description of the phospho AML1 dataset. This dataset contains six bone marrow samples from two patients: P5337 and P5340. From P5337, one LSC enriched sample and one LSC depleted sample were taken. From P5340, two LSC enriched samples and one LSC depleted sample were taken. In our DE analysis, we compared samples 1 and 3 against samples 2 and 4. **(b)** Venn diagrams of DE proteins (proteins corresponding to DE peptides) based on the identified peptides by APIR aggregating MaxQuant and MS-GF+, APIR-adjusted MaxQuant, and APIR-adjusted MS-GF+. MaxQuant and MS-GF+ were chosen because their aggregation yielded the largest percentage increases in identified PSMs/peptides/peptides with modifications/proteins in Fig. 4. Six leukemia-related proteins were found as DE proteins based on APIR: PLZF, B-raf, STAT5B, PML, CDKN1B, and RB1, all of which belong to the AML KEGG pathway or the chronic myeloid leukemia KEGG pathway. In particular, PLZF and CDKN1B were uniquely identified as DE proteins based on APIR.

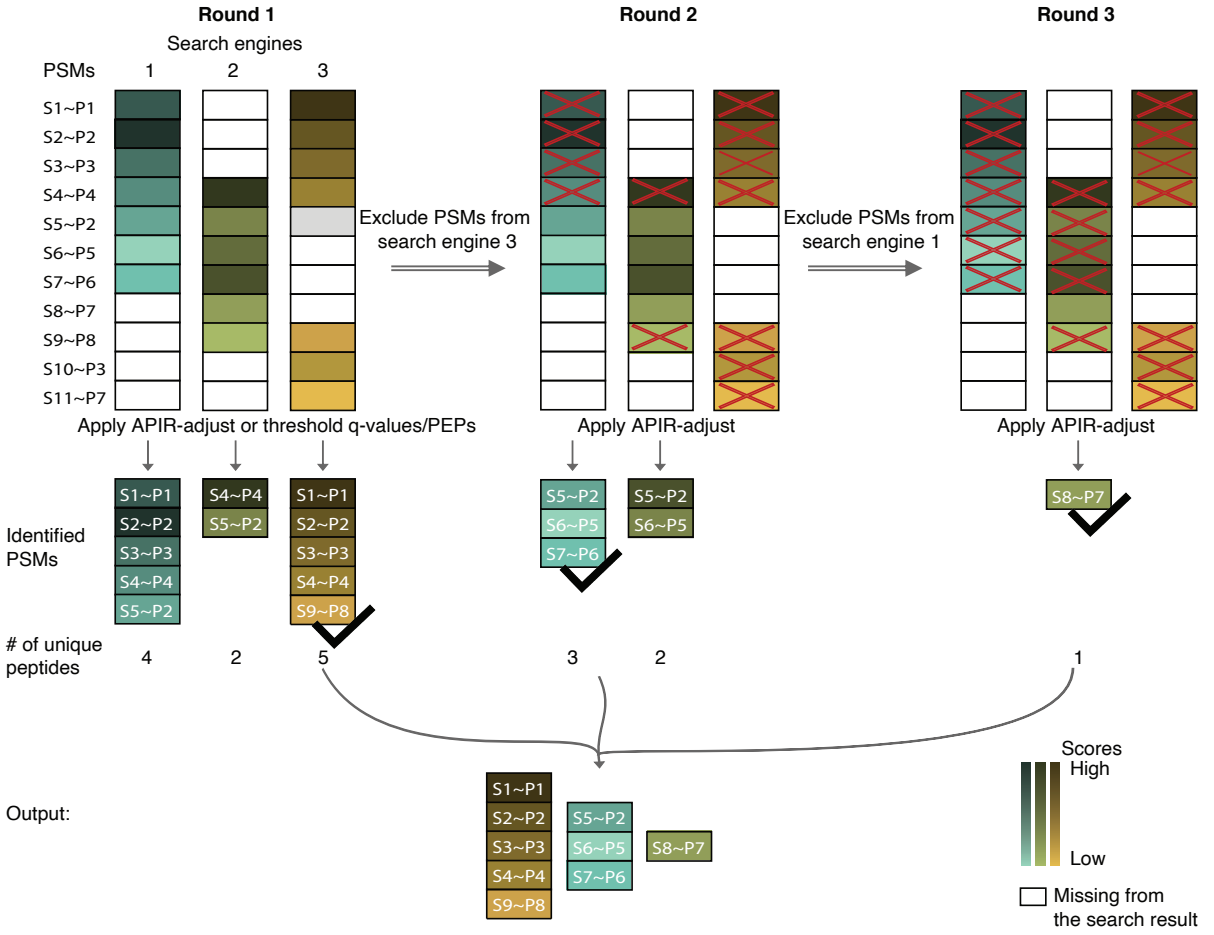



**Figure 6:** Illustration of APIR for aggregating three database search algorithms. We use S1∼P1 to denote a PSM of mass spectrum S1 matched to peptide sequence P1. In the output of a database search algorithm, a PSM with a higher score is marked by a darker color. White boxes indicate PSMs missing from the output. APIR adopts a sequential approach to aggregate the three database search algorithms. In Round 1, APIR applies APIR-adjust or q-value/PEP thresholding to identify a set of target PSMs from the output of each database search algorithm. APIR then selects the algorithm whose identified PSMs contain the largest number of unique peptides, and the identified PSMs are considered identified by APIR. In this example, APIR identified the same number of PSMs from algorithms 1 and 3 but more unique peptides from algorithm 3; hence, APIR selects algorithm 3. In Round 2, APIR excludes all PSMs, either identified or unidentified by the selected database search algorithm in Round 1 (algorithm 3 in this example), from the output of the remaining database search algorithms. Then APIR applies APIR-adjust to find the algorithm whose identified PSMs contain the largest number of unique peptides (algorithm 1 in this example). APIR repeats Round 2 in the subsequent rounds until all database search algorithms are selected and outputs the union of the PSMs identified in each round.

| Dataset | Protein | Biological relevance | References |
|---|---|---|---|
| Phospho-AML 1 & 2 | TIF1$\alpha$ | High levels of TIF1$\alpha$ are associated with oncogenesis and disease progression in a variety of cancer lineages such as AML. | [35–41] |
| | PIB5PA | PIB5PA has a tumor-suppressive role in human melanoma; its high expression has been correlated with limited tumor progression and better prognosis in breast cancer patients. | [42, 43] |
| | SMAD3 | SMAD3 plays key roles in the development and progression of various types of tumor. | [44–49] |
| | HOXB5 | HOXB5 is among the most affected transcription factors by the genetic mutations that initiate AML. | [50–52] |
| | SUMO-2 | SUMO-2 plays a key role in regulating CBX2, which is overexpressed in several human tumors (e.g., leukemia) and whose expression is correlated with lower overall survival. | [53] |
| | JUND | JUND plays a central role in the oncogenic process leading to adult T-cell leukemia. | [54] |
| | GPC2 | GPC2 has been identified as an oncoprotein and a candidate immunotherapeutic target in high-risk neuroblastoma. | [55] |
| | DNAJC21 | DNAJC21 mutations have been linked to cancer-prone bone marrow failure syndrome. | [56] |
| | ZFP36L2 | ZFP36L2 induces AML cell apoptosis and inhibit cell proliferation; its mutation is associated with the pathogenesis of acute leukemia | [57, 58] |
| | CHCHD4 | CHCHD4 plays key roles in regulating tumor proliferation. | [60] |
| | MPO | MPO is expressed in hematopoietic progenitor cells in prenatal bone marrow, which are considered initial targets for the development of leukemia. | [61–63] |
| TNBC | BRCA2 | BRCA2 is an inherited genetic mutation inactivating the BRCA2 gene can be found in people with TNBC. | [64–69] |
| | FANCE | Inactivation of the FANC–BRCA pathway has been identified in ovarian cancer cell lines and sporadic primary tumor tissues. | [70, 71] |
| | ROBO4 | ROBO4 regulates tumor growth and metastasis in multiple types of cancer, including breast cancer. | [72–75] |

**Table 1:** A summary of biologically relevant proteins missed by individual database search algorithms but recovered by APIR from the phospho AML1 and AML2 and TNBC datasets.

| Protein | Phosphorylation site | Biological relevance | Reference |
|---|---|---|---|
| PLZF | Threonine 282 | Phosphorylation at Threonine 282 activates cyclin-A2, a core cell cycle regulator of which the deregulation seems to be closely related to chromosomal instability and tumor proliferation. | [79–82] |
| CDKN1B | Serine 140 | Phosphorylation of CDKN1B at Serine 140 is important for stabilization and enforcement of the CDKN1B-mediated G1 checkpoint in response to DNA damage; inability to phosphorylate CDKN1B at Serine 140 is associated with enhanced cellular proliferation and colony . | [83, 84] |

**Table 2:** A summary of biologically relevant phosphorylation sites in the DE peptides identified by DESeq2 from the aggregated peptides by APIR from the outputs of MaxQuant and MS-GF+ on the phospho AML1 dataset.

# Supplementary figures

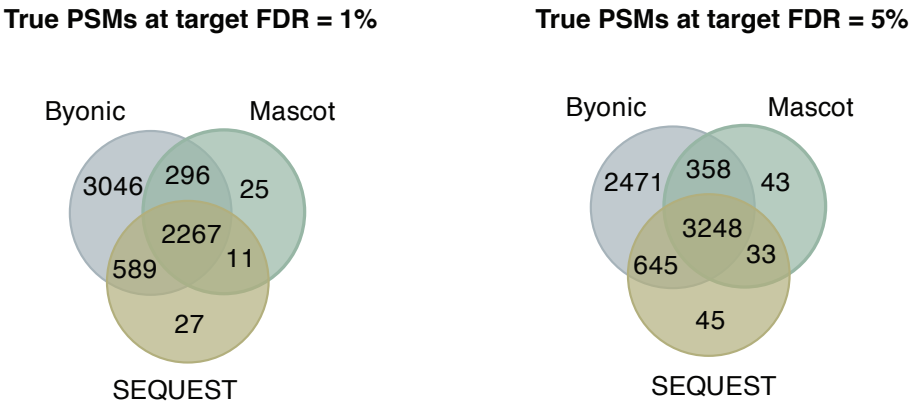



**Figure S1:** Venn diagrams of true PSMs identified by the three database search algorithms from Proteome Discoverer™ Software—Byonic, Mascot, and SEQUEST—under the FDR threshold $q = 1\%$ (left) or $q = 5\%$ (right) on the proteomics standard dataset. The true PSMs identified by Byonic nearly cover the true PSMs identified by Mascot or SEQUEST.

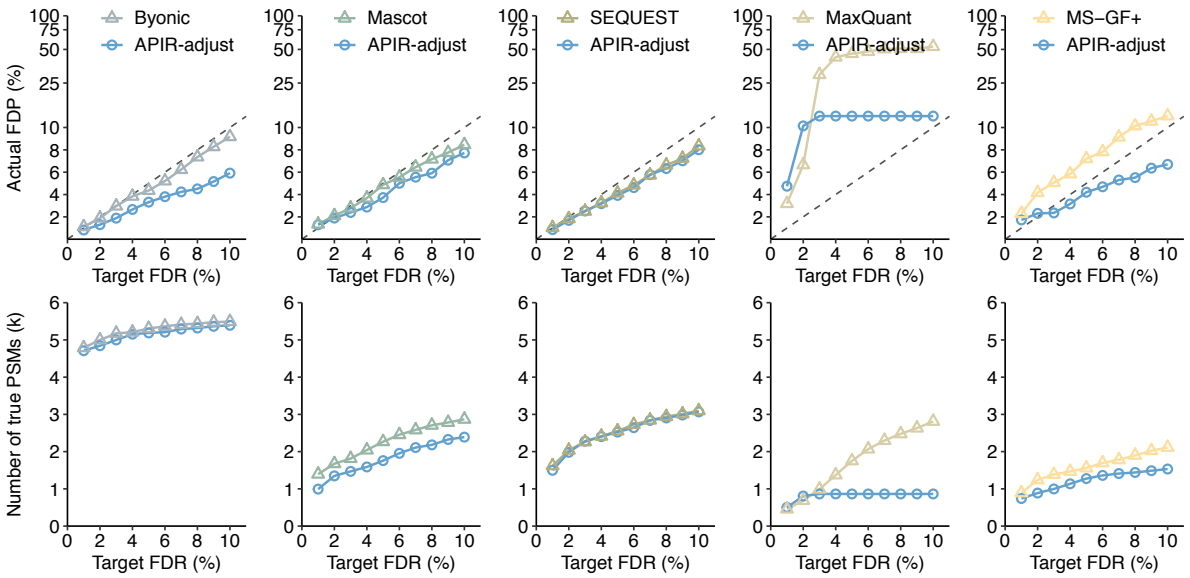



**Figure S2:** Verifying the FDR control of APIR-adjust and q-value thresholding on the incomplete output of the five popular database search algorithms—Byonic, Mascot, SEQUEST, MaxQuant, and MS-GF+—on the complex proteomics standard dataset in terms of FDR control and power. At the FDR threshold $q \in \{1\%, 2\%, \cdots, 10\%\}$, FDPs and power of each of the five database search algorithms when the $1416$ target PSMs identified by all five database search algorithms at the FDR threshold $q = 5\%$ are removed from the output of database search algorithms.

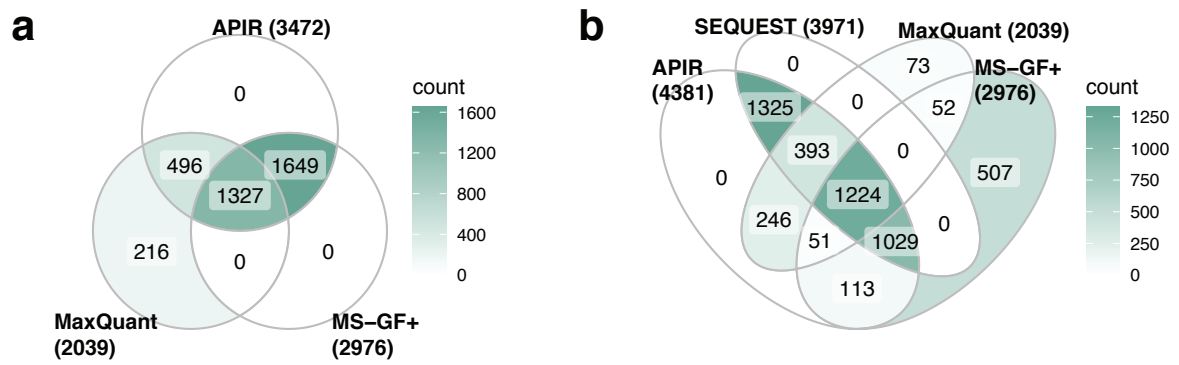



**Figure S3:** Venn diagrams of true PSMs by APIR and individual database search algorithms from two example combinations in Fig. 3a. Venn diagrams comparing APIR with **(a)** MS-GF+ and APIR-adjusted MaxQuant; with **(b)** SEQUEST, MS-GF+, and APIR-adjusted MaxQuant demonstrate that APIR identifies almost all true PSMs by individual database search algorithms at the same FDR threshold $q = 5\%$.

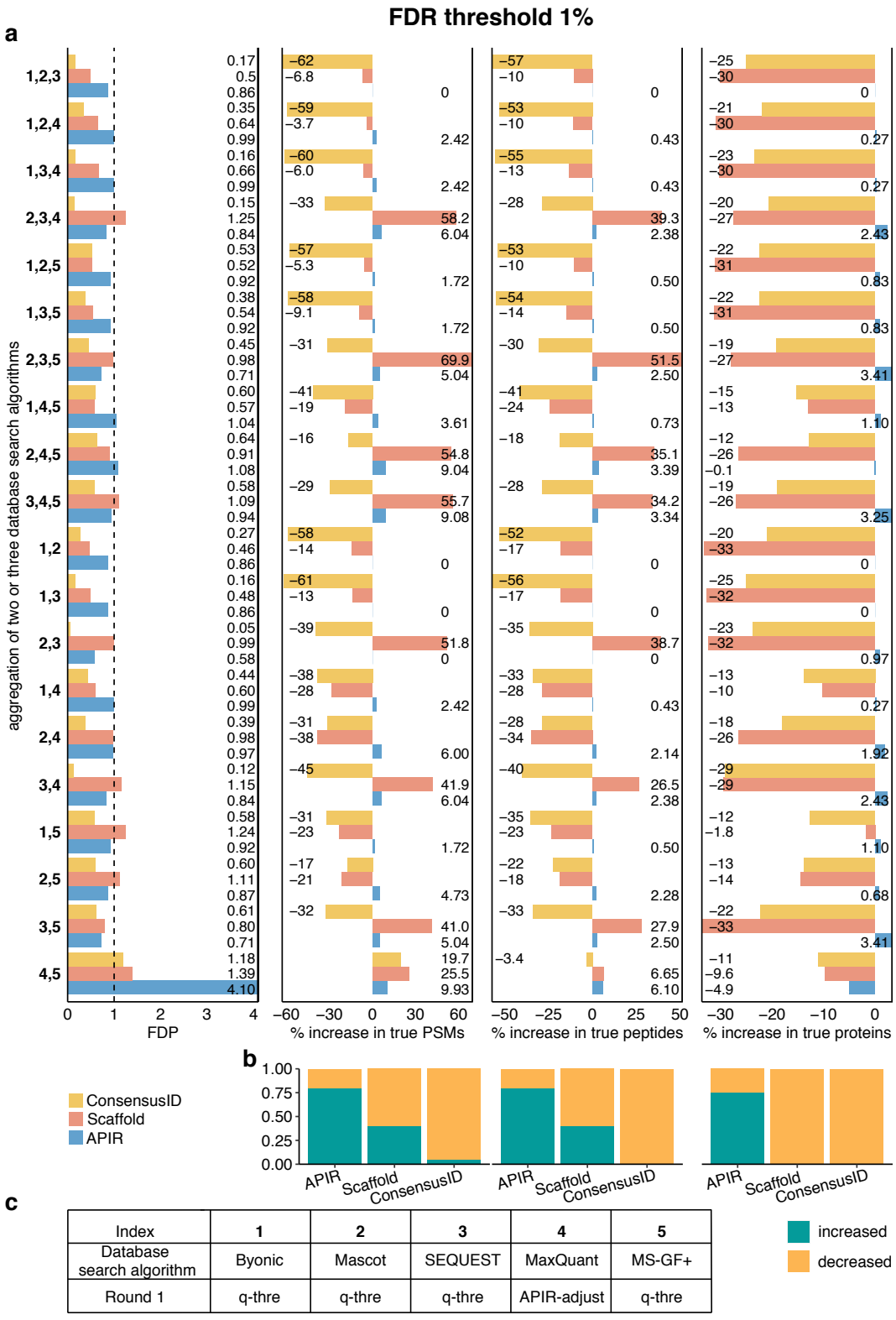
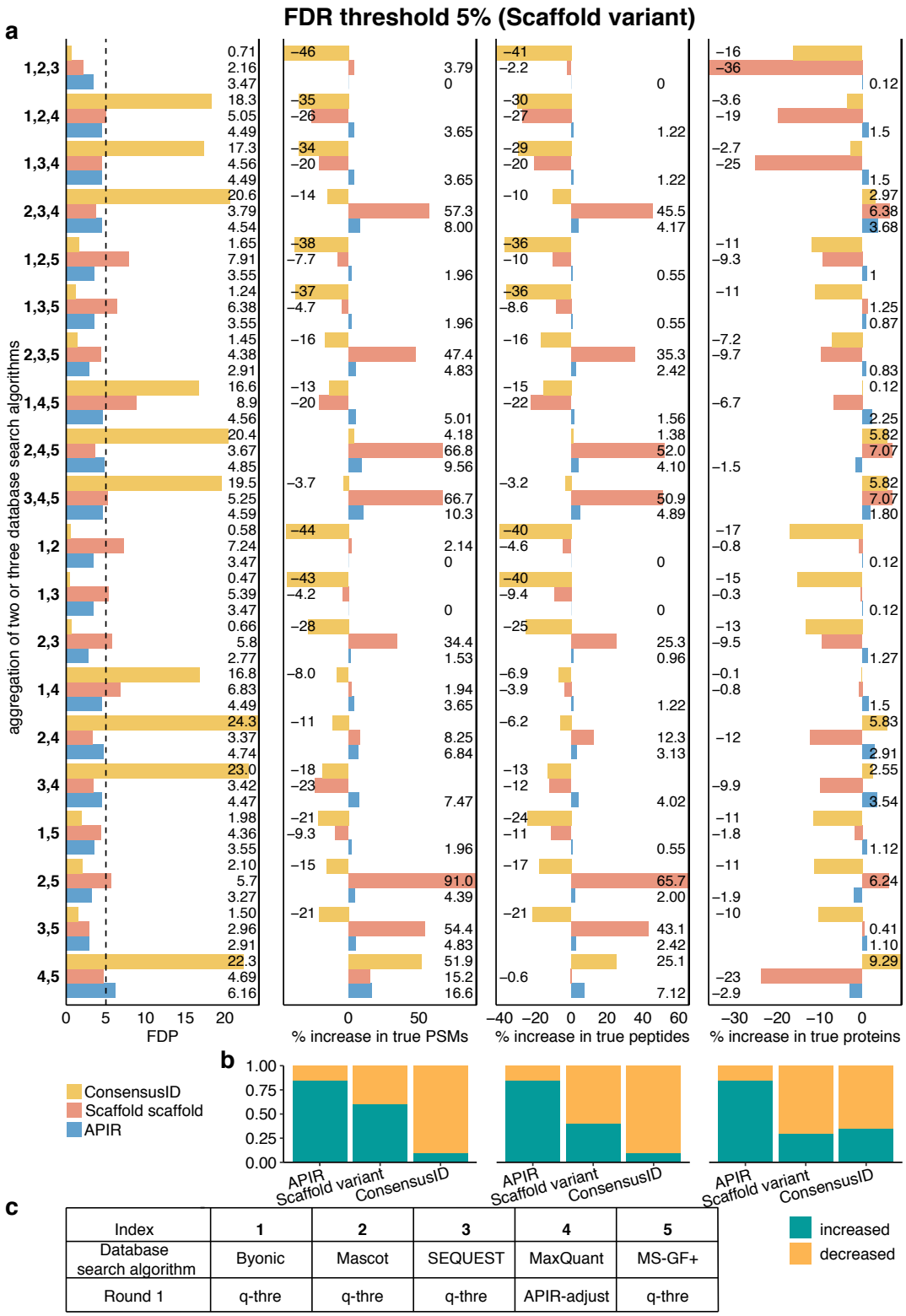



**Figure S4:** On the proteomics standard dataset, comparison of APIR, Scaffold, and ConsensusID at the FDR threshold $q = 1\%$ in terms of FDR control and power. We set both the peptide threshold and the protein threshold of Scaffold to be $1\%$ FDR. **(a)** FDPs (first column), the percentage increases in true PSMs (second column), the percentage increases in true peptides (third column), and the percentage increases in true proteins (fourth column) after aggregating two or three database search algorithms out of the five (Byonic, Mascot, SEQUEST, MaxQuant, and MS-GF+). The percentage increase in true PSMs/peptides/proteins is computed by treating as the baseline the maximal number of correctly identified PSMs/peptides/proteins by an individual database search algorithms in Round 1 of APIR. **(b)** Proportion of combinations that show a non-negative percentage increase (green bars) in true PSMs (first column), true peptides (second column), and true proteins (third column). **(c)** The indices of database search algorithm in (a) and the implementation of APIR in Round 1. Based on the benchmarking results in Fig. 1c, in Round 1 of APIR, we applied q-value thresholding (q-thre) to Byonic, Mascot, SEQUEST, and MS-GF+, and we applied APIR-adjust to MaxQuant. In later rounds of APIR, we used APIR-adjust for FDR control.




| Index | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Database search algorithm | Byonic | Mascot | SEQUEST | MaxQuant | MS-GF+ |
| Round 1 | q-thre | q-thre | q-thre | APIR-adjust | q-thre |

**Figure S5:** On the proteomics standard dataset, comparison of APIR, Scaffold variant, and ConsensusID at the FDR threshold $q = 5\%$ in terms of FDR control and power. We set Scaffold's peptide threshold to be $5\%$ FDR and varied its protein threshold to find the maximal number of identified peptides. **(a)** FDPs (first column), the percentage increases in true PSMs (second column), the percentage increases in true peptides (third column), and the percentage increases in true proteins (fourth column) after aggregating two or three database search algorithms out of the five (Byonic, Mascot, SEQUEST, MaxQuant, and MS-GF+). The percentage increase in true PSMs/peptides/proteins is computed by treating as the baseline the maximal number of correctly identified PSMs/peptides/proteins by an individual database search algorithms in Round 1 of APIR. **(b)** Proportion of combinations that show a non-negative percentage increase (green bars) in true PSMs (first column), true peptides (second column), and true proteins (third column). **(c)** The indices of database search algorithm in (a) and the implementation of APIR in Round 1. Based on the benchmarking results in Fig. 1c, in Round 1 of APIR, we applied q-value thresholding (q-thre) to Byonic, Mascot, SEQUEST, and MS-GF+, and we applied APIR-adjust to MaxQuant. In later rounds of APIR, we used APIR-adjust for FDR control.

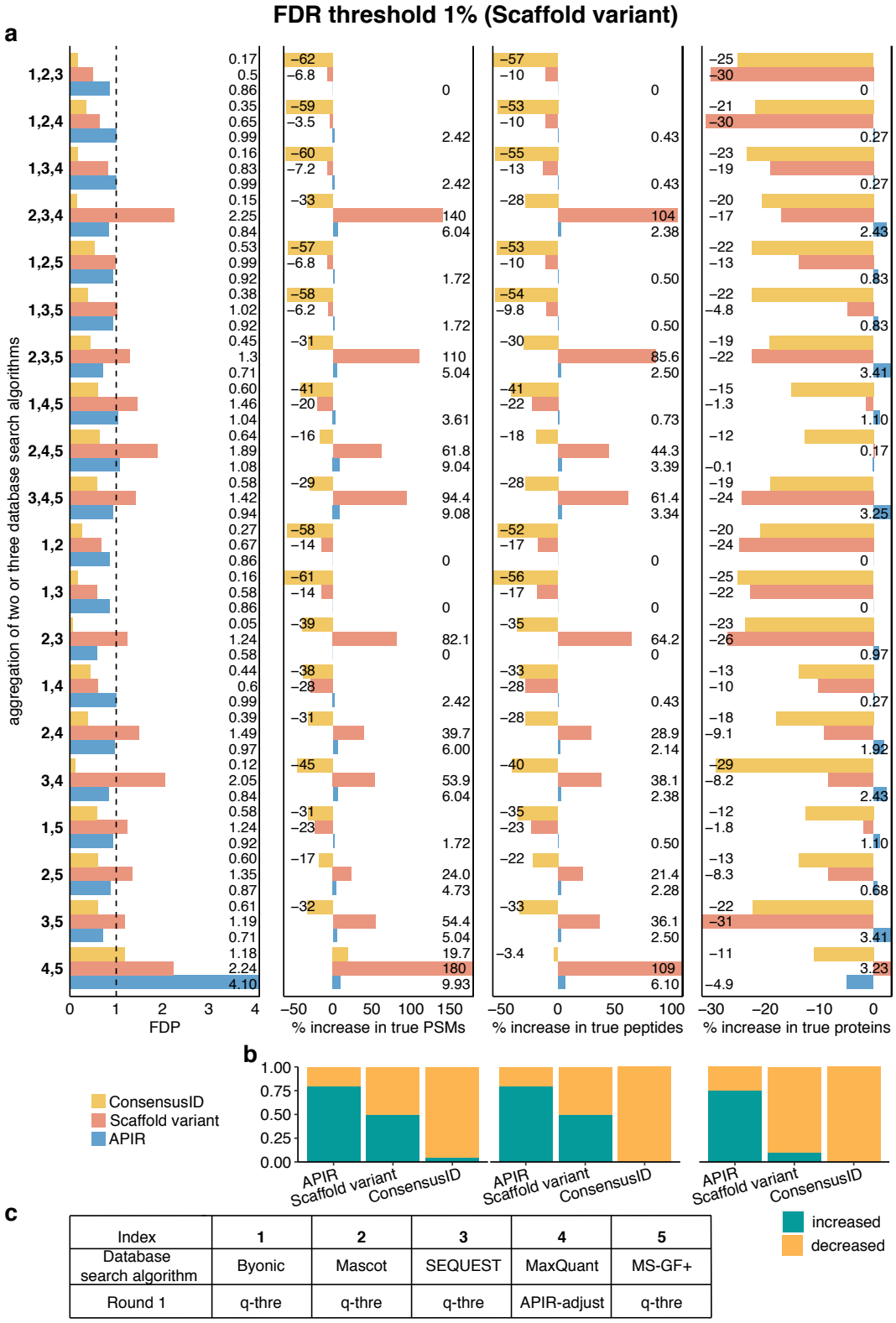



**Figure S6:** On the proteomics standard dataset, comparison of APIR, Scaffold variant, and ConsensusID at the FDR threshold $q = 1\%$ in terms of FDR control and power. We set Scaffold's peptide threshold to be $1\%$ FDR and varied its protein threshold to find the maximal number of identified peptides. **(a)** FDPs (first column), the percentage increases in true PSMs (second column), the percentage increases in true peptides (third column), and the percentage increases in true proteins (fourth column) after aggregating two or three database search algorithms out of the five (Byonic, Mascot, SEQUEST, MaxQuant, and MS-GF+). The percentage increase in true PSMs/peptides/proteins is computed by treating as the baseline the maximal number of correctly identified PSMs/peptides/proteins by an individual database search algorithms in Round 1 of APIR. **(b)** Proportion of combinations that show a non-negative percentage increase (green bars) in true PSMs (first column), true peptides (second column), and true proteins (third column). **(c)** The indices of database search algorithm in (a) and the implementation of APIR in Round 1. Based on the benchmarking results in Fig. 1c, in Round 1 of APIR, we applied q-value thresholding (q-thre) to Byonic, Mascot, SEQUEST, and MS-GF+, and we applied APIR-adjust to MaxQuant. In later rounds of APIR, we used APIR-adjust for FDR control.

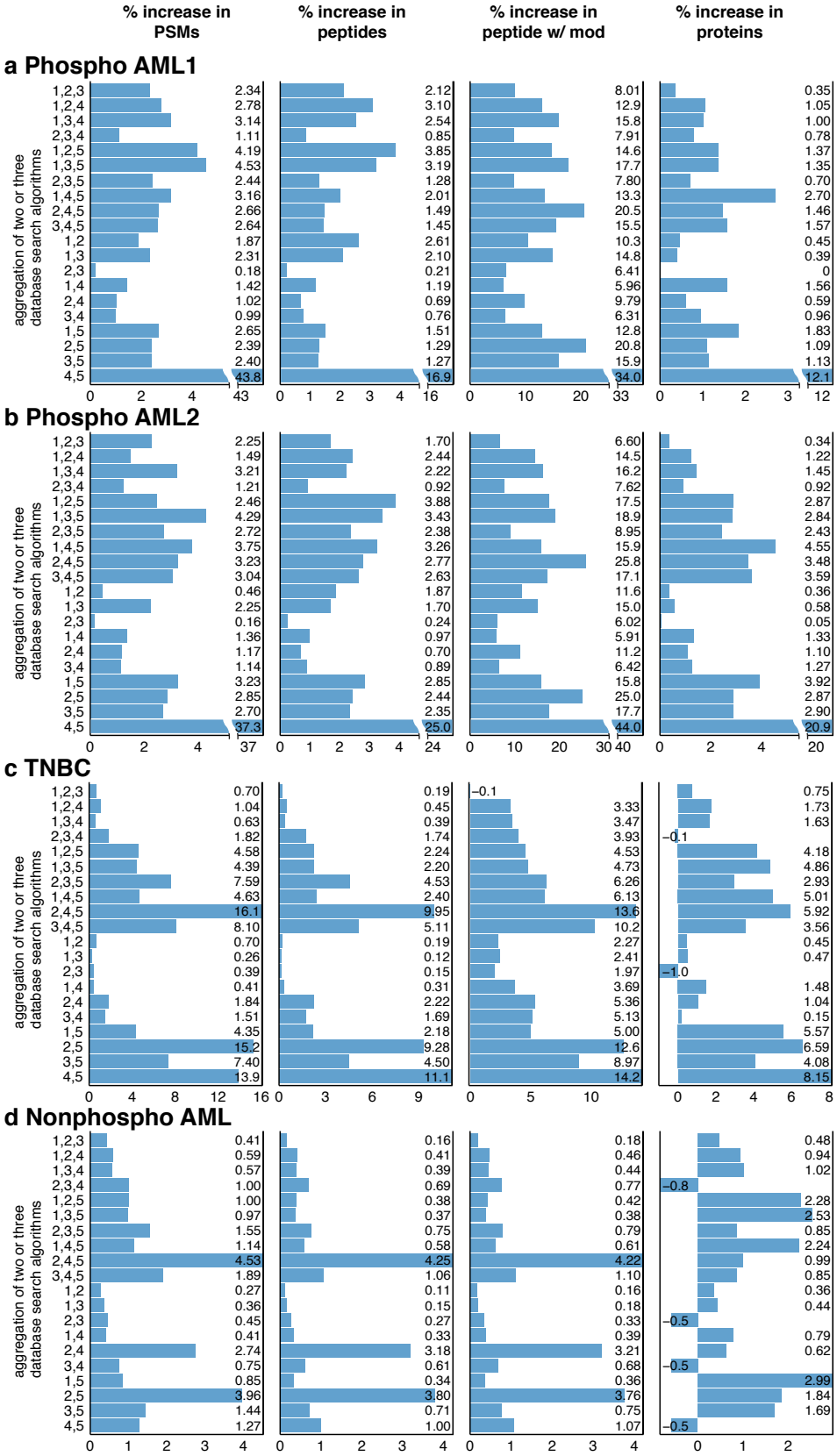



**Figure S7:** Power improvement of APIR over individual database search algorithms at the FDR threshold $q = 1\%$. The percentage increases in PSMs (first column), the percentage increases in peptides (second column), the percentage increases in peptides with modifications (third column), and the percentage increases in true proteins (fourth column) of APIR after aggregating two or three database search algorithms out of the five (Byonic, Mascot, SEQUEST, MaxQuant, and MS-GF+) at the FDR threshold $q = 5\%$ on **(a)–(b)** the phospho-proteomics AML datasets, **(c)** the nonphospho-proteomics AML dataset, and **(d)** the TNBC dataset. The percentage increase in PSMs/peptides/peptides with modifications/proteins is computed by treating as the baseline the maximal number of PSMs/peptides/peptides and modifications/proteins by an individual database search algorithm in the first round of APIR, where APIR-adjust was used for FDR control.

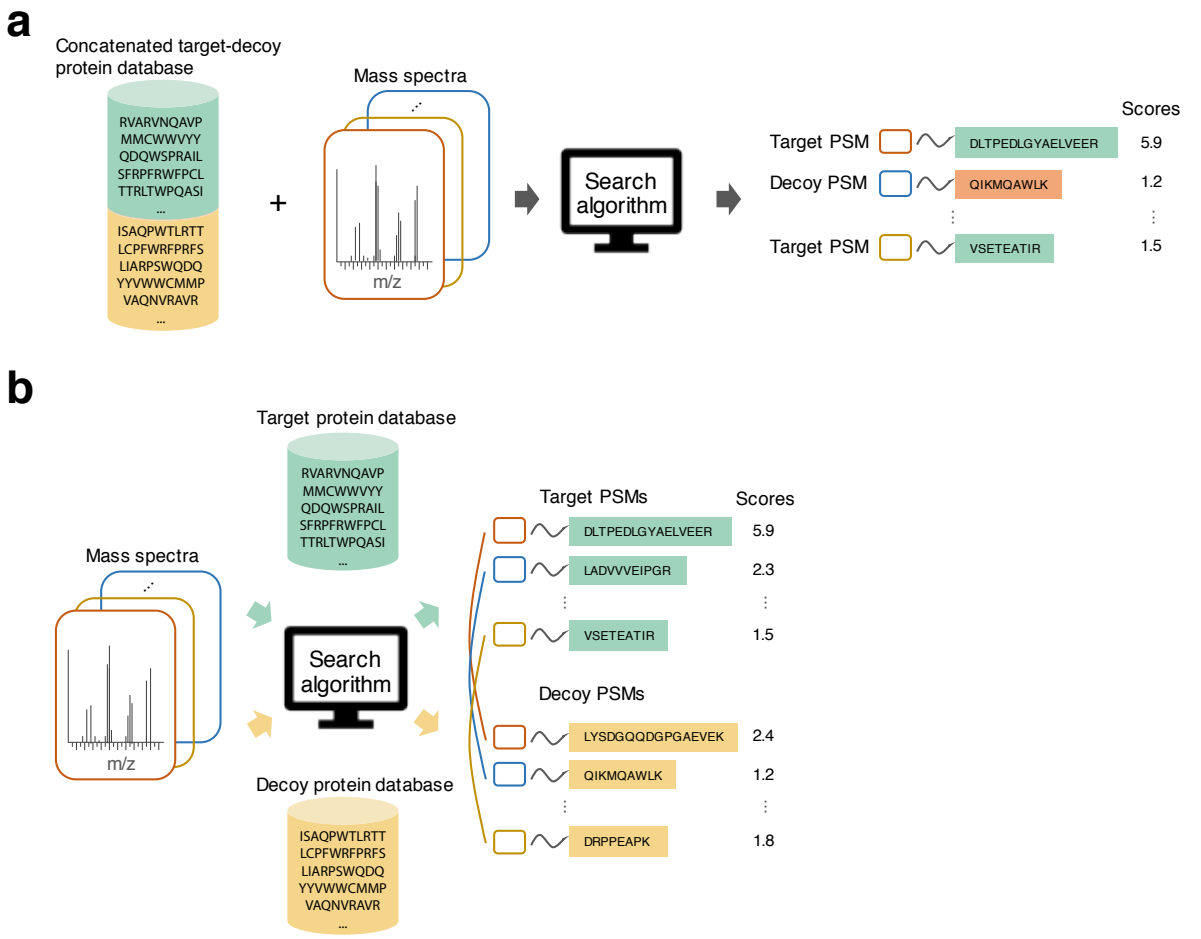



**Figure S8:** Two implementations of the target-decoy search strategy: concatenated (a) and parallel (b). In the concatenated search, a concatenated protein database is created by pooling original protein sequences, called "target" sequences, with the decoy sequences; then a database search algorithm uses the concatenated protein database to find PSMs; consequently, each mass spectra is mapped to either a target sequence or a decoy sequence with only one matching score. In the parallel search, a database search algorithm conducts two parallel searches: a target search where each mass spectrum is matched to target sequences and a decoy search where the mass spectrum is matched to decoy sequences; consequently, each mass spectrum receives two matching scores from the two searches. In both implementations, a PSM is called a target PSM or simply a PSM if it contains a target sequence; otherwise, it is called a decoy PSM.