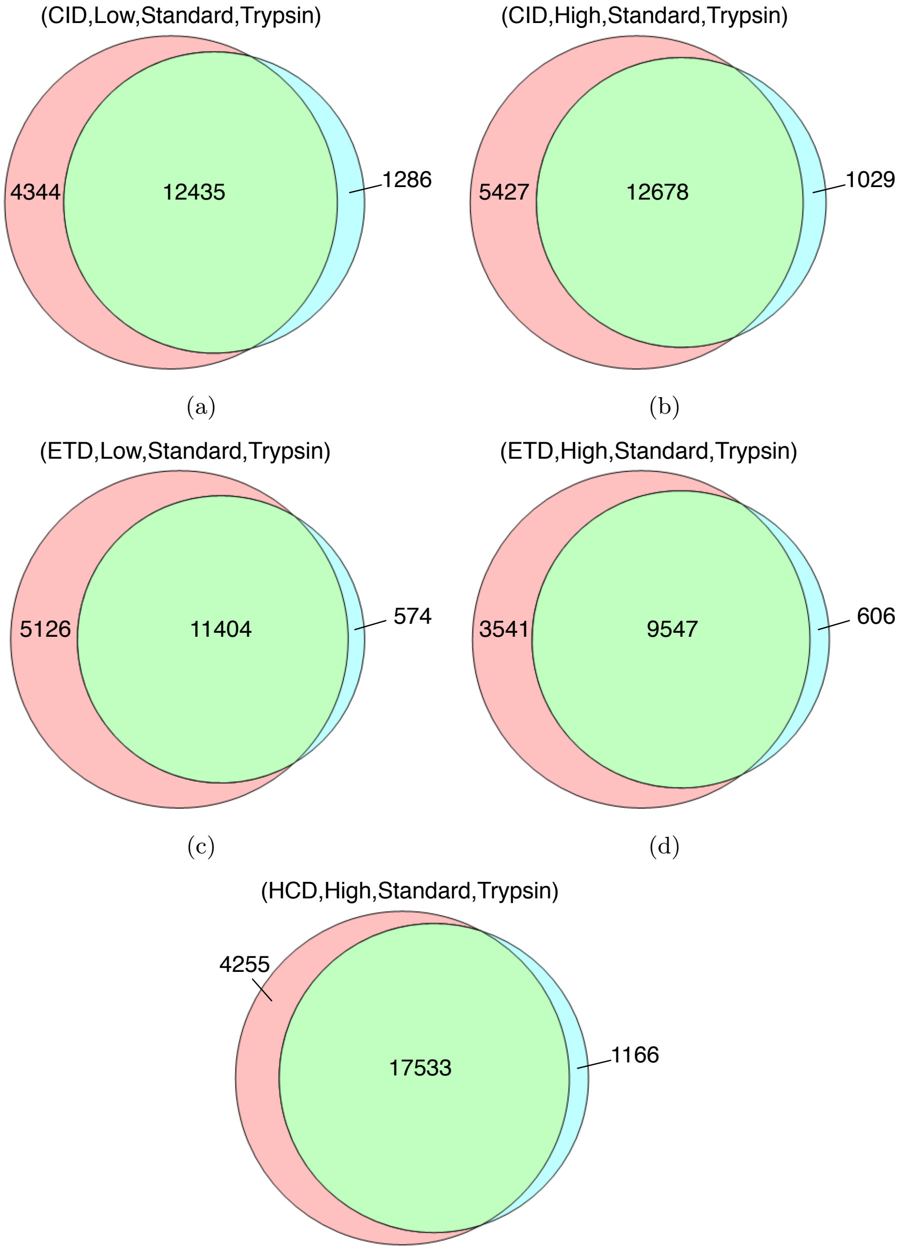
Supplementary Figure 1 - Venn Diagrams of MS-GF+ and Mascot+Percolator Identifications



(a)

(b)

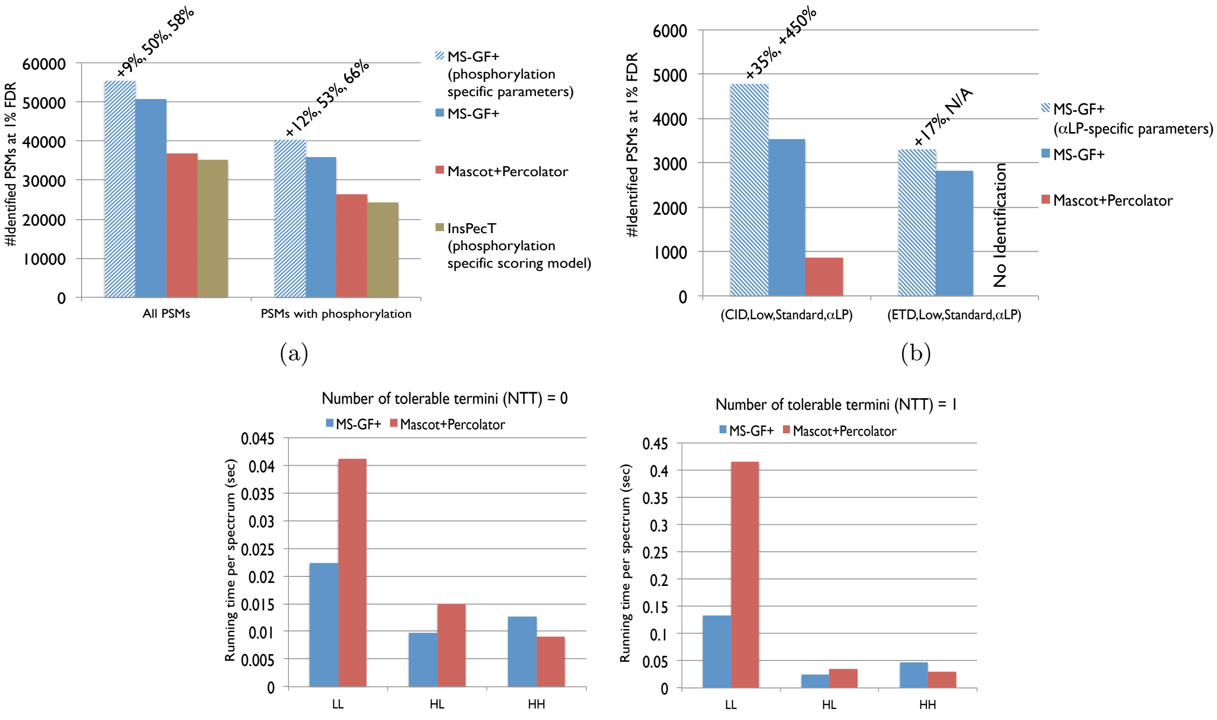
(c)

(d)

(e)

Venn diagrams of MS-GF+ and Mascot+Percolator identifications for the human datasets corresponding to (a) (CID,Low,Standard,Trypsin), (b) (CID,High,Standard,Trypsin), (c) (ETD,Low,Standard,Trypsin), (d) (ETD,High,Standard,Trypsin), and (e) (HCD,High,Standard,Trypsin). The red, blue, and green areas represent the number of PSMs at 1% FDR identified only by MS-GF+ (red), only by Mascot+Percolator (blue), and by both (green).

Supplementary Figure 2 - Comparison of MS-GF+ with Others



(a)

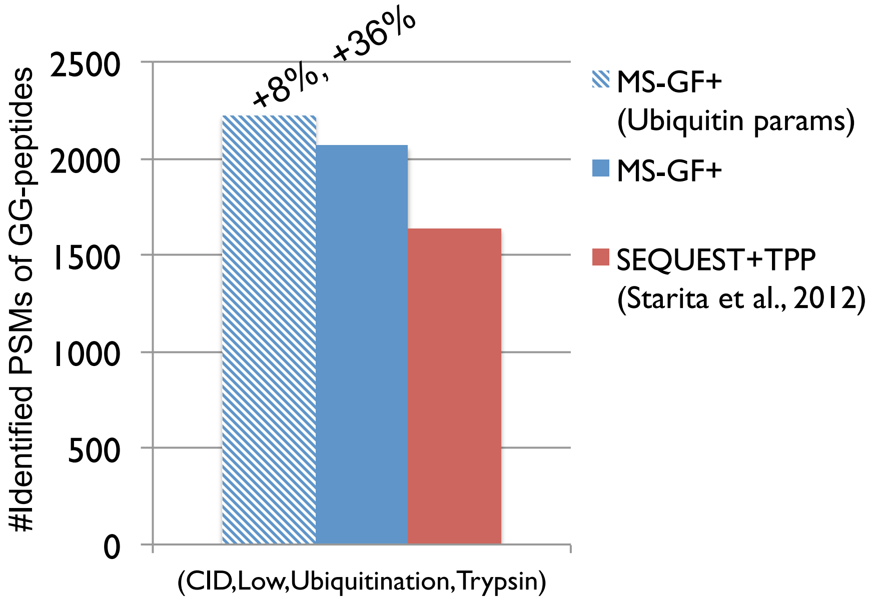
(b)

(c)

(d)

(a) Comparison of MS-GF+, Mascot+Percolator, and InsPecT for the mouse dataset corresponding to (CID,Low,Phosphorylation,Trypsin). The numbers of identified PSMs of all peptides (left) and phosphorylated peptides (right) are shown. MS-GF+ was run twice, first, with the parameter set for (CID,Low,Standard,Trypsin) (denoted by MS-GF+), and second, for (CID,Low,Phosphorylation,Trypsin) (denoted by MS-GF+ (phosphorylation specific parameters)). (b) Comparison of MS-GF+, Mascot+Percolator for the *S. Pombe* datasets corresponding to (CID,Low,Standard,αLP) and (ETD,Low,Standard,αLP). MS-GF+ was executed twice with the parameter set for (CID,Low ,Standard,Trypsin) (denoted by MS-GF+) and for (CID,Low,Standard,αLP) (denoted by MS-GF+ (αLP-specific parameters)). For (a), (b), and (e), shown on the top of the bars is the percentages of increase in the number of identifications for the leftmost tool compared to others. (c,d) Running times of MS-GF+ and Mascot+Percolator when the number of tolerable termini (NTT) is 0 (d) and NTT=1 (e). Average running time per spectrum is shown in second. LL, HL, HH represent LL, HL, and HH spectra, respectively. When NTT=0, MS-GF+ was 80% and 50% faster for LL and HL spectra, respectively, but 30% slower for HH spectra as compared to Mascot+Percolator. When NTT=1, MS-GF+ was 210% and 40% faster for LL and HL spectra, respectively, but 40% slower for HH spectra as compared to Mascot+Percolator.

Supplementary Figure 3 - MS-GF+ for Ubiquitinated Peptides

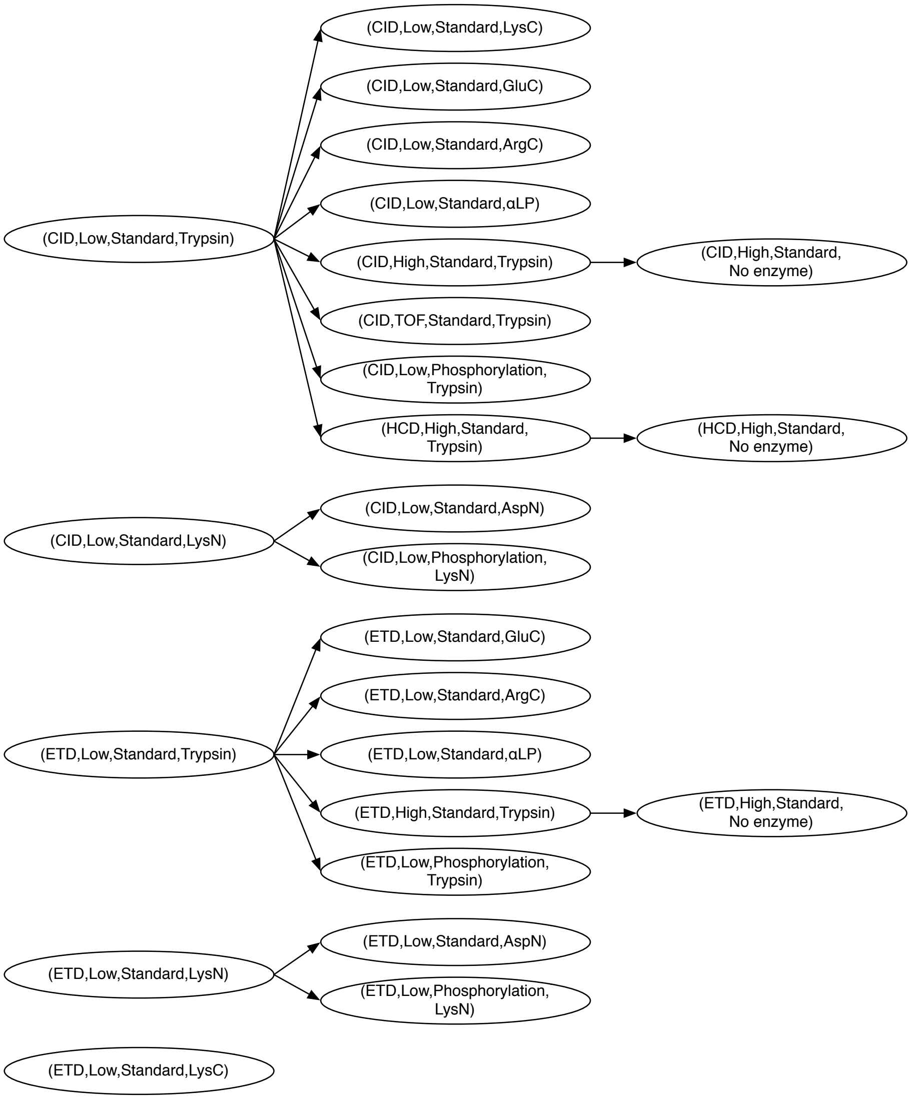


Comparison of MS-GF+, and results in [1] for the yeast dataset corresponding to (CID,Low, Ubiquitination,Trypsin). The numbers of identified PSMs of GG-peptides are shown. MS-GF+ was executed twice with the parameter set for (CID,Low,Ubiquitination,Trypsin) (denoted by MS-GF+ (Ubiquitin params)) and for (CID,Low,Standard,Trypsin) (denoted by MS-GF+). MS-GF+ identified 36% more PSMs of GG-peptides as compared to [1]. With scoring parameters for (CID,Low,Ubiquitination,Trypsin), MS-GF+ further identified 8% more PSMs of GG-peptides as compared to MS-GF+ for (CID,Low,Standard,Trypsin).

We used a yeast dataset corresponding to the spectral type (CID,Low,**Ubiquitination**,Trypsin) containing 452,812 spectra. This dataset was generated from the Fields laboratory (University of Washington). Yeast proteins are mock treated. Ubiquitinated proteins are purified by metal affinity resin and further digested with trypsin. Peptides are separated by SCX. The detailed experimental procedures are described in [1].

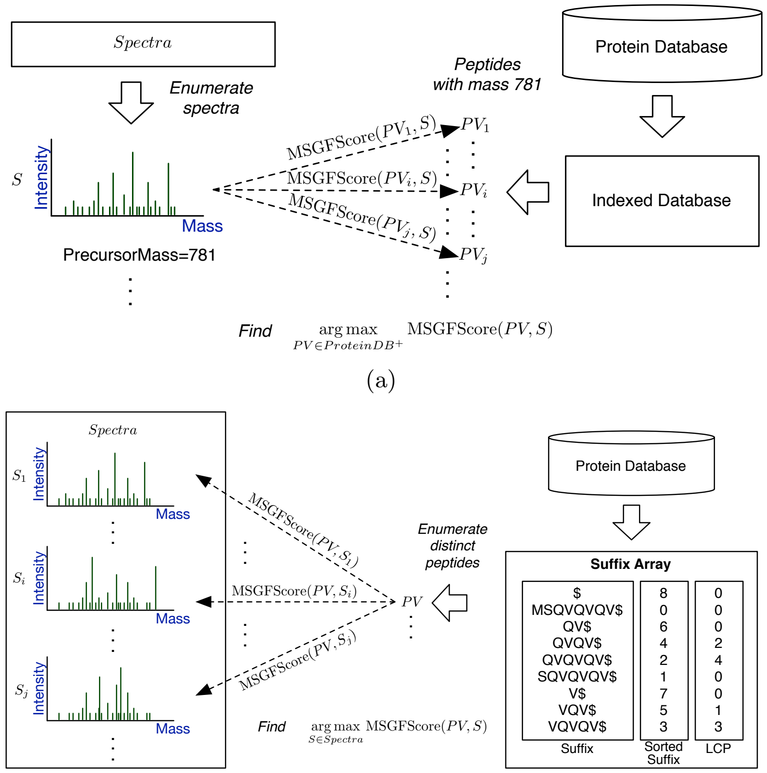
Using this dataset, we generated a scoring parameter set for (CID,Low,Ubiquitination,Trypsin) and compared the numbers of identified PSMs of GG-peptides (peptides with two glycines attached to a lysine) for MS-GF+ and SEQUEST+Trans-Proteomics Pipeline [2] (denoted by SEQUEST+TPP) reported in [1] (Figure). In this comparison, FDR 0.3% was used as in [1]. Without ubiquitination-specific scoring parameters, MS-GF+ outperformed SEQUEST+TPP, identifying 26% more PSMs of GG-peptides than SEQUEST+TPP. With ubiquitination-specific parameters, MS-GF+ further identified 8% more PSMs of GG-peptides.

Supplementary Figure 4 - Creating Scoring Parameter Sets


(CID,Low,Standard,LysC)
(CID,Low,Standard,GluC)
(CID,Low,Standard,ArgC)
(CID,Low,Standard,αLP)
(CID,Low,Standard,Trypsin)
(CID,High,Standard,Trypsin)
(CID,High,Standard, No enzyme)
(CID,TOF,Standard,Trypsin)
(CID,Low,Phosphorylation, Trypsin)
(HCD,High,Standard, Trypsin)
(HCD,High,Standard, No enzyme)
(CID,Low,Standard,AspN)
(CID,Low,Standard,LysN)
(CID,Low,Phosphorylation, LysN)
(ETD,Low,Standard,GluC)
(ETD,Low,Standard,ArgC)
(ETD,Low,Standard,Trypsin)
(ETD,Low,Standard,αLP)
(ETD,High,Standard,Trypsin)
(ETD,High,Standard, No enzyme)
(ETD,Low,Phosphorylation, Trypsin)
(ETD,Low,Standard,AspN)
(ETD,Low,Standard,LysN)
(ETD,Low,Phosphorylation, LysN)
(ETD,Low,Standard,LysC)

Various scoring parameter sets used by MS-GF+. Each vertex represents a scoring parameter set for the specified spectral type. Each edge represents what scoring parameter set (parent) is used to construct a new scoring parameter set (child). For example, a parameter set for (HCD,High,Standard,Trypsin) was constructed using the parameter set for (CID,Low,Standard,Trypsin), which in turn is used to construct a parameter set for (HCD,High,Standard,No enzyme). Starting with 5 parameter sets (vertices with no incoming edge), 20 more parameter sets were constructed. Note that users can also easily generate new parameter sets using existing parameter sets.

Supplementary Figure 5 - Approaches for Database Searching


Spectra
Enumerate spectra
S
Intensity
Mass
PrecursorMass=781
Peptides with mass 781
Protein Database
Indexed Database
$\mathrm{MSGFScore}(PV_1, S)$
$\mathrm{MSGFScore}(PV_i, S)$
$\mathrm{MSGFScore}(PV_j, S)$
PV_1
PV_i
PV_j
Find $\underset{PV \in ProteinDB^+}{\arg\max} \ \mathrm{MSGFScore}(PV, S)$
Spectra
S_1
S_i
S_j
Intensity
Mass
$\mathrm{MSGFScore}(PV, S_1)$
$\mathrm{MSGFScore}(PV, S_i)$
$\mathrm{MSGFScore}(PV, S_j)$
PV
Enumerate distinct peptides
Protein Database
Suffix Array
Find $\underset{S \in Spectra}{\arg\max} \ \mathrm{MSGFScore}(PV, S)$

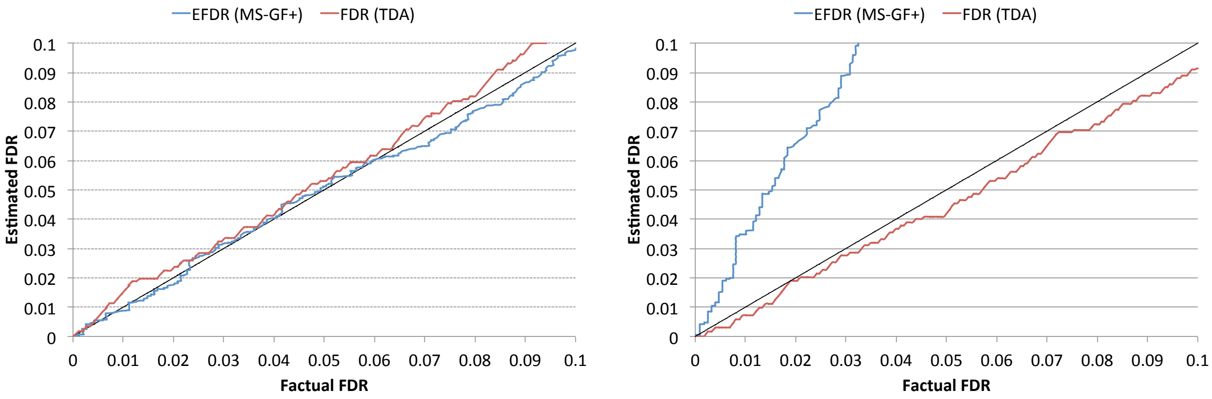
(a)

Suffix	Sorted Suffix	LCP
$	8	0
MSQVQVQV$	0	0
QV$	6	0
QVQV$	4	2
QVQVQV$	2	4
SQVQVQV$	1	0
V$	7	0
VQV$	5	1
VQVQV$	3	3

(b)

Two approaches for comparing spectra against a protein database. (a) Traditional spectrum-based approach (used by MS-GFDB [3]) compares each spectrum against all peptides. (b) MS-GF+ uses an alternative peptide-based approach that computes the suffix array to compare each peptide against all spectra with the same precursor mass. See Methods for details.

Supplementary Figure 6 - Accuracy of FDR estimation



(a) LL spectra

(b) HL spectra

Accuracy of the EFDR reported by MS-GF+ and the FDR via TDA for LL spectra (a) and HL spectra (b). The factual FDR was used as an estimator of true FDR. The EFDR is accurate for LL spectra but biased for HL spectra (and also HH spectra).

Supplementary Table 1 - Rescaling amino acid masses

Residue	NominalMass	RealMass	RescaledMass	ErrorRealMass	ErrorRescaledMass
G	57	57.021	56.993	0.021	-0.007
A	71	71.037	71.002	0.037	0.002
S	87	87.032	86.989	0.032	-0.011
P	97	97.053	97.004	0.053	0.004
V	99	99.068	99.019	0.068	0.019
T	101	101.048	100.997	0.048	-0.003
L	113	113.084	113.028	0.084	0.028
I	113	113.084	113.028	0.084	0.028
N	114	114.043	113.986	0.043	-0.014
D	115	115.027	114.969	0.027	-0.031
Q	128	128.059	127.995	0.059	-0.005
K	128	128.095	128.031	0.095	0.031
E	129	129.043	128.978	0.043	-0.022
M	131	131.040	130.975	0.040	-0.025
H	137	137.059	136.990	0.059	-0.010
F	147	147.068	146.995	0.068	-0.005
R	156	156.101	156.023	0.101	0.023
C+57	160	160.031	159.951	0.031	-0.049
Y	163	163.063	162.982	0.063	-0.018
W	186	186.079	185.986	0.079	-0.014
Average rounding error				0.057	-0.004
Average absolute rounding error				0.057	0.017

Rescaling amino acid masses. MS-GF+ is designed under the assumption that amino acid masses are integers and uses nominal masses as integer masses of amino acids. While this enables efficient computing of MS-GF+ scores, it causes rounding errors because peaks in the spectrum correspond to real rather than nominal masses of amino acid sequences. To minimize the rounding errors, MS-GF+ *rescales* every mass m into $0.9995m$ [4, 5, 6]. This dramatically reduces the rounding errors as shown above, so one can estimate the nominal mass of a peptide (or a peak) of mass m by simply taking $[0.9995m]$ where $[x]$ represents the closest integer to x. For example, for a peptide "RESCAILINGMASSES" of mass 1705.776, $[1705.776 \cdot 0.9995 = 1704.92] = 1705$ represents the correct nominal mass of the peptide. We investigated all 188 million unique peptides of length up to 20 in the human IPI database (version 3.87), and the estimation was inaccurate (i.e., $[0.9995m]$ is not equal to the nominal mass of the peptide of mass m) only for 874 peptides (4.6×10^{-6} %). The percentage only increases to 0.07% for peptides of lengths up to 40.

For each amino acid, shown are its nominal mass (NominalMass), real mass (RealMass), rescaled mass (RescaledMass), rounding error as the real mass minus nominal mass (ErrorRealMass), rounding error as the rescaled mass minus nominal mass (ErrorRescaledMass). Also the average of rounding errors and absolute rounding errors are shown. For every amino acid, rounding errors for rescaled masses are smaller (maximum 0.049 for Cys+57) as compared to real masses (maximum 0.101 for Arg). In addition, the average rounding error of 20 amino acids is only -0.004, meaning that even for long peptides the differences between their rescaled masses and nominal masses are usually small. Note that Carboxymethylated Cys (Cys+57) is considered instead of Cys here.

Supplementary Table 2 - Database Search Parameters

Parameter	MS-GF+	Mascot+Percolator
Precursor mass tolerance	7 ppm	50 ppm
Precursor mass tolerance (*S. pombe*)	2.5Da	2.5Da
Fragment mass tolerance	N/A	0.6 Da for Low, 0.05 Da for High
Number of missed cleavages	N/A	2
Fixed modification	Carbamidomethylation of Cys	
Variable modifications (Default)	Oxidation of Met, Acetylation of Protein N-term	
Variable modifications (Mouse)	Default + Phosphorylation of Ser, Thr, and Tyr	
Variable modifications (*S. pombe*)	Default + Pyro-glu of Gln and Glu	
Number of tolerable termini (NTT)	0 (trypsin), 2 (αLP), 1 (others)	
Number of ^{13}C	1	
Protein Database (Human)	IPI human (ver. 3.87) + contaminants	
Protein Database (Yeast)	UniProt yeast (release 2012_02)	
Protein Database (Mouse)	IPI mouse (ver. 3.87) + contaminants	
Protein Database (*S. pombe*)	Uniprot *S. pombe*	

Table 1: Database search parameters used for MS-GF+ and Mascot+Percolator searches. The number of tolerable termini (NTT) indicates the maximum allowed number of peptide termini (0, 1, or 2) that is not consistent with the specificity of the enzyme. For example, for trypsin, NTT=0 means that only fully-tryptic peptides are considered in the search. For αLP digests, NTT was set to 2 because the cleavage specificity of αLP was unknown. For other non-tryptic enzymes, NTT was set to 1 (instead of 0 for trypsin) because they often produce peptides that are not perfectly consistent with their cleavage specificities. The number of ^{13}C was set to 1 to correct the error of choosing ^{13}C rather than ^{12}C peak during the MS1 peak detection. MS-GF+ does not take the fragment mass tolerance as an input, and rather implicitly assumes 0.5 Da tolerance for all spectra. For HH spectra, instead of taking smaller fragment mass tolerance (e.g. 0.05Da), MS-GF+ benefits from accurate product ion peaks using the spectral DAG scoring model (Online Method). MS-GF+ does not take the number of missed cleavages as an input, and rather allows peptides with any number of missed cleavages. For all the tools, the decoy search option was enabled for all searches to estimate the FDR. Note that a large precursor mass tolerance (50 ppm) was used for Mascot to provide sufficient training data for the Percolator algorithm [7]. Since Percolator gives a penalty to peptide identifications whose parent masses deviate from the precursor ion masses of the spectra, using a wide precursor mass tolerance increases rather than decreases the number of identifications. For Mascot only searches, the precursor mass tolerance was set to 7 ppm because it produces more identifications than using 50 ppm. For the InsPecT search for the mouse dataset, InsPecT version 20100804 was used and the precursor and fragment mass tolerances were set to 2.5 Da and 0.5 Da, respectively. For InsPecT, using a narrower precursor mass tolerance decreased the number of identifications. To calculate the FDR, the spectral E-value, ion score, and F-score were used for MS-GF+, Mascot, InsPecT, respectively, according to the procedure described in the Supplementary Note 2. For Percolator, instead of calculating the FDR, we used the q-value reported by Percolator.

Supplementary Note 1 - The Peptide-based Approach to MS/MS Database Search

Solving the database search problem involves the following three steps: (1) for every spectrum $S \in \mathit{Spectra}$, computing its spectral vector $\mathbf{S}$, (2) for every variant $PV \in \mathit{ProteinDB}^{+}$, computing its peptide vector $\mathbf{PV}$, and (3) for every pair (PV, S) with $\text{Mass}(PV) = \text{PrecursorMass}(S)$, computing $\text{MSGFScore}(PV, S) = \mathbf{PV} \cdot \mathbf{S}$. To execute these steps efficiently, one may simply execute the step (1) and (2), store all $\mathbf{S}$ and $\mathbf{PV}$ in the main memory and execute the step (3). But this is often infeasible because the number of variants is usually too large to fit all peptide vectors $\mathbf{PV}$ in the main memory. Alternatively, one may consider executing the step (2) on the spot for each spectrum, but this is prohibitively slow. For example, for the IPI human database, using a Core i7 920 (quad core, 2.67Ghz) with 12GB main memory, executing the step (2) alone took 54 seconds, considering partially tryptic peptides of lengths between 6 and 40, and two variable modifications Oxidation of Met and protein N-term Acetylation.

Instead of storing both $\mathbf{S}$ and $\mathbf{PV}$, MS-GF+ stores only $\mathbf{S}$ for all spectra in the main memory, and indexes them by precursor masses. Since spectral vectors compactly represent experimental spectra, MS-GF+ can store over 200,000 spectral vectors in the main memory of 4GB. Rather than finding the best scoring peptide for each spectrum, MS-GF+ then finds the best scoring spectra for each variant. This can be done efficiently by enumerating variants $PV \in \mathit{ProteinDB}^{+}$ one by one, generating $\mathbf{PV}$ *on the spot*, and computing $\text{MSGFScore}(PV, S) = \mathbf{PV} \cdot \mathbf{S}$ for all *pre-computed* $\mathbf{S}$ where $S \in S_{\text{Mass}(PV)}$. In practice, for each spectrum S, MS-GF+ records the best scoring variant PV^{*} while enumerating PVs, and updates PV^{*} to PV whenever it finds PV, where $\text{MSGFScore}(PV, S) > \text{MSGFScore}(PV^{*}, S)$.

Similar to pFind [8], MS-GF+ uses a suffix array (a lexicographically sorted list of all the suffixes of $\mathit{ProteinDB}$ [9]) to further optimize the database search. Since protein databases contain many similar proteins, many peptides appear in multiple copies in a database (*repeated peptides*). For example, the IPI human database (version 3.87) contains about 130,000 fully-tryptic peptides of length 10, but the number decreases to about 50,000 if only unique peptides are considered. If a protein database is indexed as a suffix array, peptide occurrences from the same repeated peptide appear in the neighboring indices in the suffix array. So, instead of searching peptides according to their ordering in the database, MS-GF+ searches peptides according to their ordering in the suffix array, and uses the Longest Common Prefix (LCP) data structure [9] to score each unique peptide only once (Supplementary Fig. 5).

MS-GF+ achieved an order of magnitude speedup compared to the early versions of MS-GFDB. Furthermore, pre-processing of the database required to run MS-GF+ is much faster as compared to MS-GFDB which uses indices from peptide masses to the peptide locations (Figure 2 in the main text). The indexing in MS-GFDB had to be repeated even for the same protein database depending on the enzymes and modifications. In contrast, MS-GF+ uses a suffix array of a protein database that needs to be constructed only once, i.e., it does not depend on the chosen enzyme and allowed modifications. Also suffix arrays in MS-GF+ can be constructed much faster than indices in MS-GFDB while being more memory-efficient [8].

Note that MS-GFDB was implemented as a prototype of MS-GF+ and in the first year of MS-GF+ developments, we were distributing it under the name MS-GFDB.

Supplementary Note 2 - Estimating FDRs

Given a spectral E-value threshold t, MS-GF+ uses the following variation of the TDA to compute FDR:

1. Generate a decoy database by reversing the target database.
2. Concatenate the target and decoy database and run a database search tool against the concatenated database. For each spectrum, consider only the best scoring PSM.
3. Sort all PSMs in decreasing order of spectral E-values.
4. For a threshold t, report the FDR as N_{decoy}/N_{target} where N_{target} (N_{decoy}) is the number of target (decoy) PSMs with spectral E-values equal or smaller than t.
5. Report the set of target PSMs with spectral E-values equal or smaller than t.

One can estimate FDRs without using TDA (denoted EFDR) by applying the method presented in [10]. We evaluated the accuracy of the EFDR estimates using the *de facto* standard method using spectra from samples of known proteins (described in [11]). For this purpose, we obtained the LTQ-Orbitrap dataset in Mix 7 from the ISB Standard Protein Mix Database generated from 18 known proteins [12]. Out of 10 replicates, we selected the replicate #2 containing 4,966 spectra and ran MS-GF+ against the yeast database appended to the 18 protein sequences. In these searches, PSMs matched to the 18 proteins and yeast proteins are distinguished as *factual* and *non-factual* PSMs, respectively. We computed *factual FDRs* as the portion of factual PSMs over non-factual PSMs (at a certain E-value threshold) and used this factual FDR as an estimator of "true" FDR.

MS-GF+ was run twice with varying precursor mass tolerances: 2.5 Da to represent LL spectra and 30 ppm to represent HL spectra. We compared how the EFDR (denoted by EFDR (MS-GF+)) and the FDR computed via TDA (denoted by FDR (TDA)) match with the factual FDR for varying thresholds (Supplementary Fig. 6). For LL spectra, the EFDR matched very well with the factual FDR in the entire range. In particular, for thresholds below 0.05 (used by most MS experiments), the EFDR estimation was better than the estimation via TDA. However, for HL spectra (and for HH spectra), the EFDR was biased towards conservative estimations.

The EFDR is biased for HH and HL spectra because E-values reported by MS-GF+ are biased. This is due to a discrepancy between the search space of the database search and the E-value computation presented in the previous section; in the high-precision setting, peptides considered are those with masses matching the precursor mass of the input spectrum within a narrow tolerance (e.g. 10 ppm); but in the E-value computation, peptides considered are those with masses matching the nominal precursor mass of the input spectrum. For HL and HH spectra, E-values computed by MS-GF+ are larger than they should be (conservative estimation) when a strict precursor mass tolerance (e.g. 10ppm) is used.

Peptide identifications are usually reported for a fixed FDR either at the PSM-level (FDR for identified PSMs) or the peptide-level (FDR for identified peptides). Since a single peptide often generates multiple spectra, the same PSM- and peptide-level FDR may result in vastly different set of identified PSMs. MS-GF+ reports peptide-level FDRs along with PSM-level FDRs. To compute peptide level FDRs, for PSMs matched to the same peptide, MS-GF+ retains only the PSM with the lowest spectral E-value. The peptide-level FDR is calculated as $N_{decoyPep}/N_{targetPep}$ where $N_{targetPep}$ ($N_{decoyPep}$) is the number of *retained* target (decoy) peptides with spectral E-values equal or smaller than t.

Supplementary Note 3 - From Ion DAGs to a Single Spectral DAG

Given a set of ion DAGs of a spectrum S (for each $ion \in \mathcal{I}$), we generate a single *spectral DAG* (instead of a spectral vector) of S. A spectral DAG is a DAG with a vertex set $V = \{0, \ldots, M = \text{PrecursorMass}(S)\}$ and and edge set $E = \{(i, j) | j - i \in \text{Mass}(a) \text{ for } a \in \mathcal{A}^{+}\}$, where the score of a vertex i is the sum of VertexScore(i) over all ion DAGs of S, the score of an edge (i, j) is the sum of EdgeScore(i, j) over all ion DAGs of S, the probability of an edge is defined as $\frac{1}{20}$, the score of a path is defined as the sum of scores of its *vertices and edges*, and the probability of a path is defined as the product of probabilities of its edges. Note that a path of a spectral DAG represents a peptide (or a variant), and the score of a path represents the score of the peptide represented by the path.

Given a spectral DAG, one can compute spectral E-values of peptides (or variants) using the generating function approach [13]. The generating function approach works for any DAG as long as the score of a path (variant) in the DAG is represented as the sum of scores of the vertices along the path. While the spectral DAG has scores both on vertices and edges, it can be converted into a DAG having scores only on vertices by introducing a new vertex $v_{i,j}$ for every edge (i, j) and replacing the edge (i, j) by two edges $(i, v_{i,j})$ and $(v_{i,j}, j)$. Thus, the generating function approach can be applied to this "spectral DAG" scoring model. Note the spectral DAG scoring model is only applied to HH spectra (see Supplementary Note 3 for details).

In theory, one can apply this spectral DAG scoring model to all HH, HL, and LL spectra. However, for HL and LL spectra, using this spectral DAG scoring model only slightly increases the number of peptide identifications (by less than 5%) while doubling the running time. For HH spectra, we found that it is beneficial to convert multiply charged product ion peaks into singly charged ion peaks (charge deconvolution) prior to generating spectral DAGs. We use the following simple algorithm for the charge deconvolution: if two peaks are separated by (mass of ^{13}C$-$mass of ^{12}C)$/c$ within a small tolerance (e.g. 0.01 Da), we assume they are charge c and convert them into charge 1.

References

[1] Starita, L. M., Lo, R. S., Eng, J. K., von Haller, P. D. & Fields, S. Sites of ubiquitin attachment in saccharomyces cerevisiae. *Proteomics* **12**, 236–40 (2012).

[2] Deutsch, E. W. *et al.* A guided tour of the trans-proteomic pipeline. *Proteomics* **10**, 1150–9 (2010).

[3] Kim, S. *et al.* The generating function of cid, etd, and cid/etd pairs of tandem mass spectra: Applications to database search. *Mol. Cell. Proteomics* **9**, 2840–52 (2010).

[4] Kim, S., Gupta, N., Bandeira, N. & Pevzner, P. Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra. *Mol. Cell. Proteomics* **8**, 53–69 (2009).

[5] Taylor, J. A. & Johnson, R. S. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid. Commun. Mass Spectrom.* **11**, 1067–75 (1997).

[6] Bern, M., Cai, Y. & Goldberg, D. Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. *Anal. Chem.* **79**, 1393–400 (2007).

[7] Frese, C. K. *et al.* Improved peptide identification by targeted fragmentation using cid, hcd and etd on an ltq-orbitrap velos. *J. Proteome Res.* **10**, 2377–88 (2011).

[8] Zhou, C. *et al.* Speeding up tandem mass spectrometry-based database searching by longest common prefix. *BMC Bioinformatics* **11**, 577 (2010).

[9] Manber, U. & Myers, G. Suffix arrays: a new method for on-line string searches. *SIAM J. Computing* **22**, 935–48 (1990).

[10] Gupta, N., Bandeira, N., Keich, U. & Pevzner, P. A. Target-decoy approach and false discovery rate: when things may go wrong. *J. Am. Soc. Mass Spectrom.* **22**, 1111–20 (2011).

[11] Keller, A. *et al.* Experimental protein mixture for validating tandem mass spectral analysis. *OMICS* **6**, 207–12 (2002).

[12] Klimek, J. *et al.* The standard protein mix database: a diverse data set to assist in the production of improved peptide and protein identification software tools. *J. Proteome Res.* **7**, 96–103 (2008).

[13] Kim, S., Gupta, N. & Pevzner, P. Spectral probabilities and generating functions of tandem mass spectra: A strike against decoy databases. *J. Proteome Res.* **7**, 3354–63 (2008).