Metabolome Identification by Systematic Stable Isotope Labeling Experiments and False Discovery Analysis with a Target-Decoy Strategy

Drew R. Jones[1#], Xusheng Wang[2#], Tim Shaw[2,3], Ji-Hoon Cho[2], Ping-Chung Chen[1,4], Suiping Zhou[2], Yuxin Li, Nam Chul Kim[5], J. Paul Taylor[5,6], Udhghatri Kolli[7], Jiaxu Li[7], and Junmin Peng[1,2,4]*

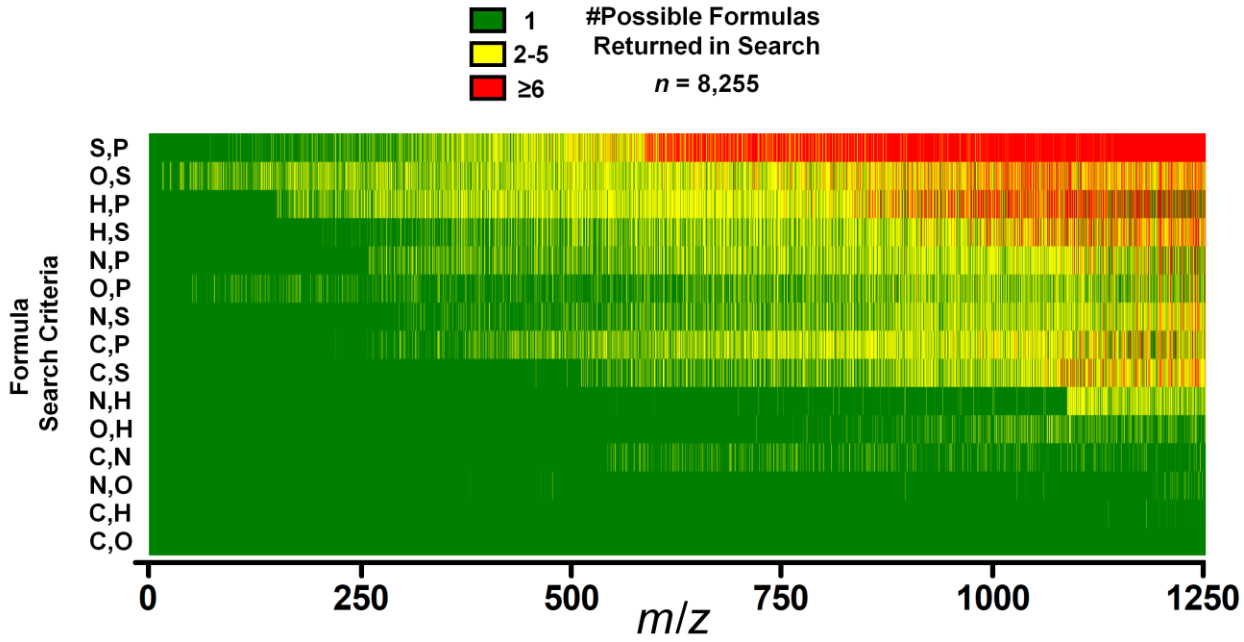
[1]Department of Structural Biology, [2]St. Jude Proteomics Facility, [3]Department of Computational Biology, [4]Department of Developmental Neurobiology, [5]Department of Cell and Molecular Biology, [6]Howard Hughes Medical Institute. [1,2,3,4,5]St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105, USA

[7]Department of Biochemistry, Molecular Biology, Entomology, and Plant Pathology, Mississippi State University, 32 Creelman Street Mississippi State, MS 39762, USA

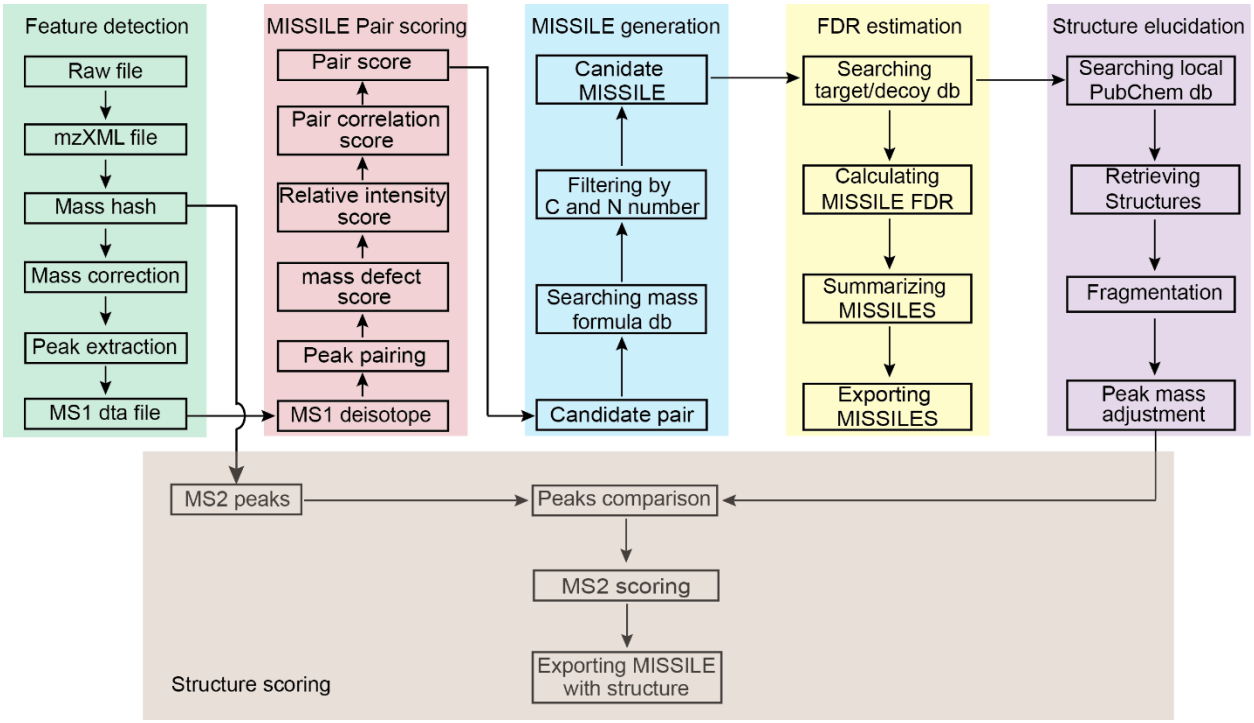
[#]These authors contributed equally to the work

*Correspondence: Junmin Peng (junmin.peng@stjude.org)

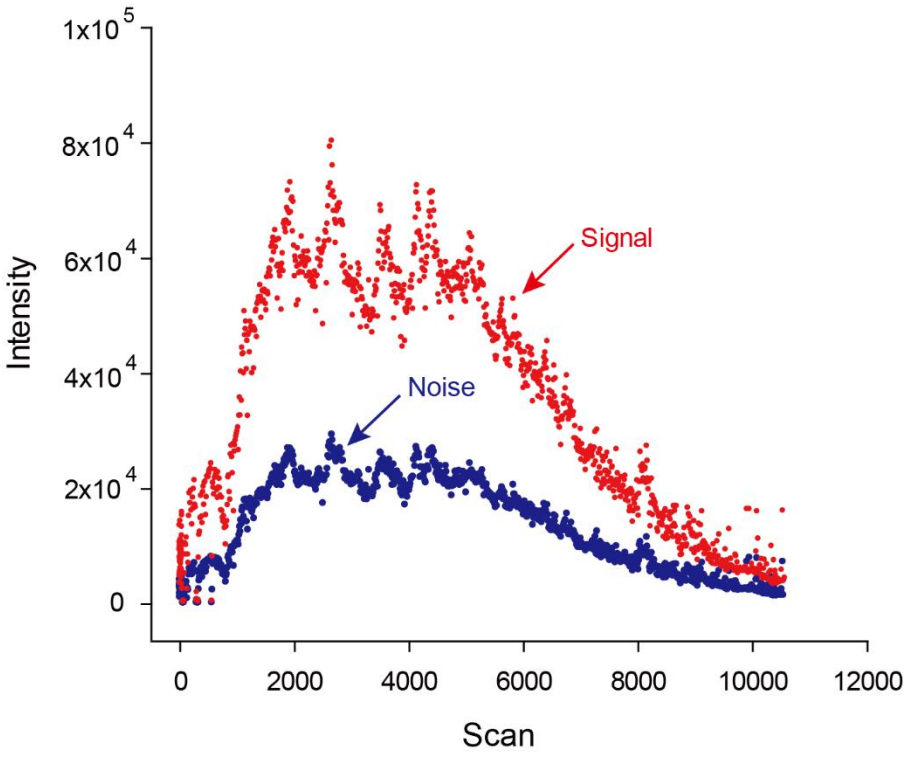
Table of Contents



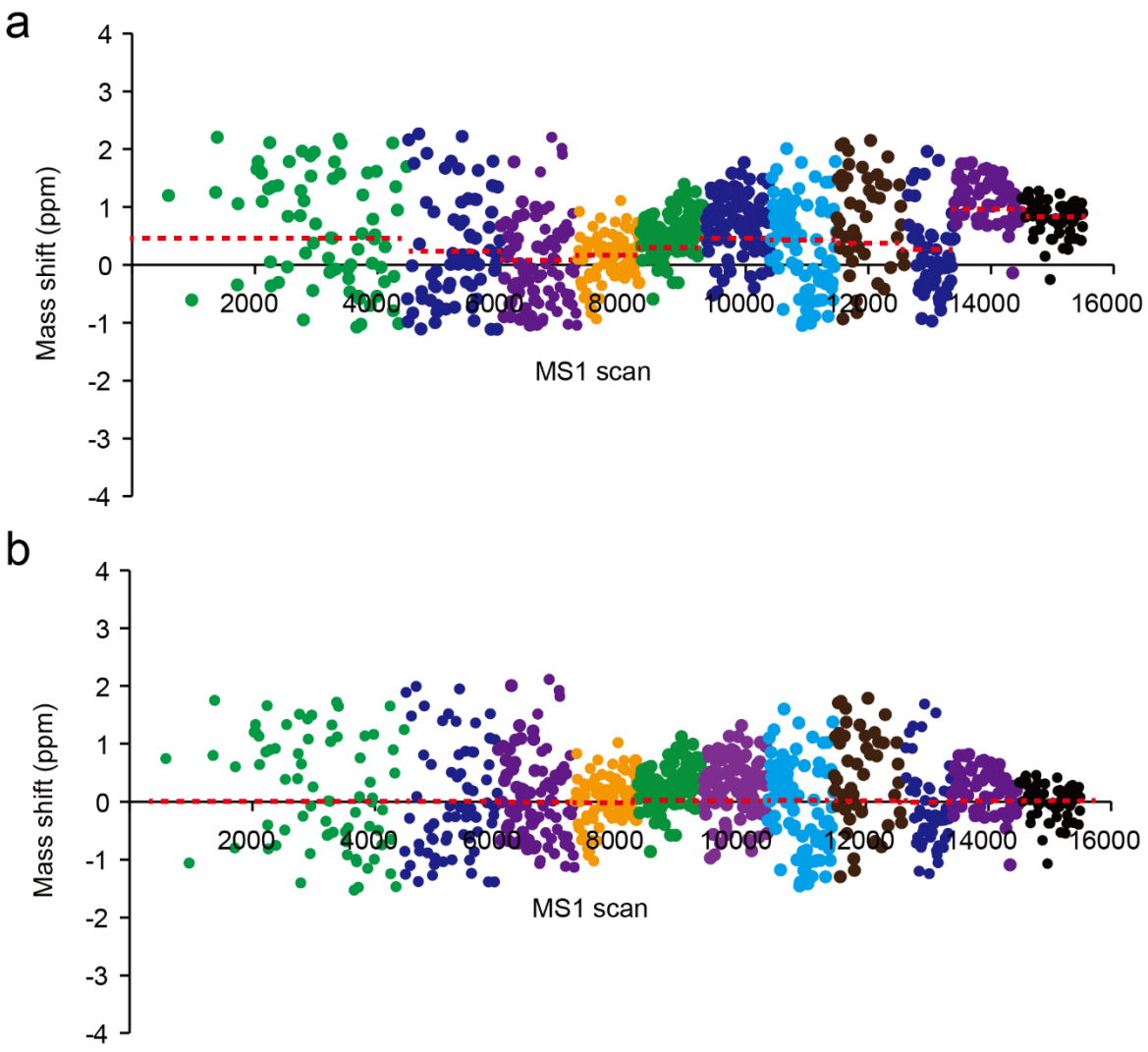
Supplementary Fig. 1. Heat map visualization of chemical formula searches restricted by known formula information (two elements) as indicated. Each search was performed with a mass tolerance of 4 ppm. E.g. for row "C,N" the available information was the accurate theoretical mass (+/- 4ppm) of the HMDB entry and the number of carbon and nitrogen atoms in that entry's chemical formula. Total number of searches = 8,255 for each row.



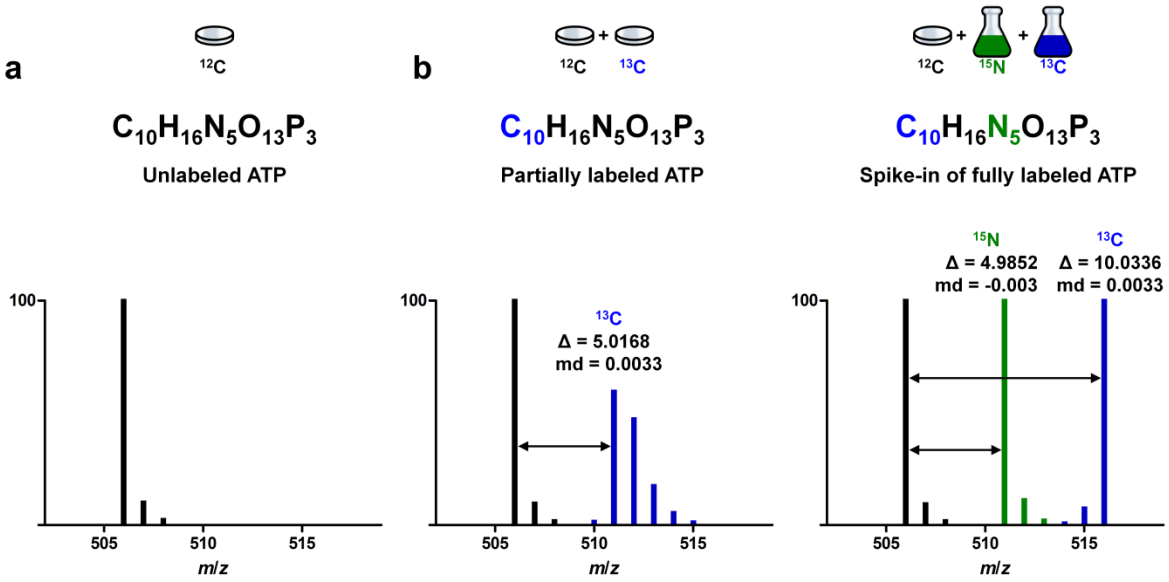
Supplementary Fig. 2. Detailed workflow for JUMPm. JUMPm has six major components, including feature detection, isotope peak pairing, formula identification, FDR estimation, structural elucidation, and spectrum match scoring. JUMPm accepts .raw or .mzXML files (including MS1 and MS2) and outputs tables of identified metabolites and formulas. db = database, FDR = false discovery rate, MISSILE = chemical formula determined from accurate mass and labeled mass shift information.



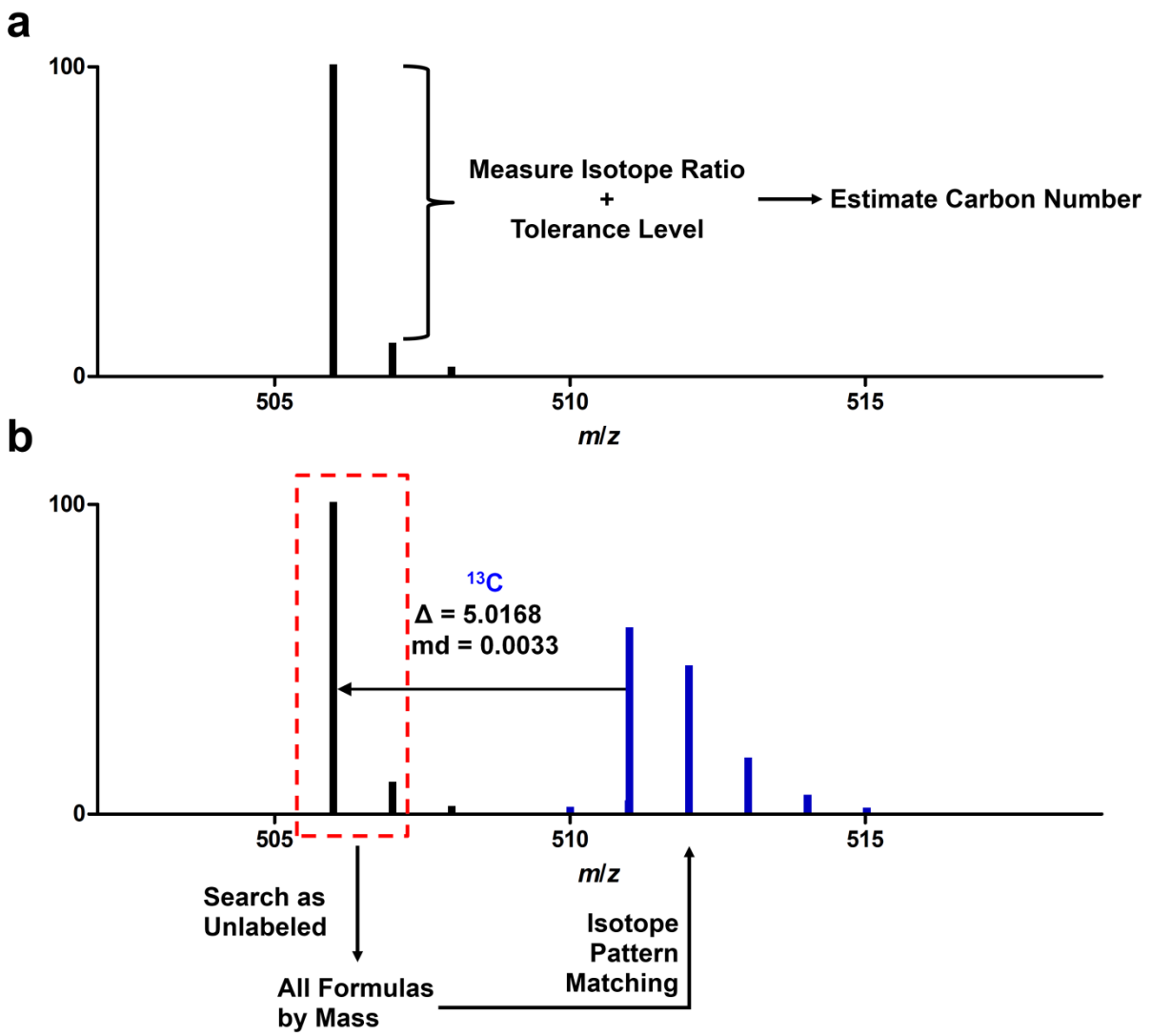
Supplementary Fig. 3. Distribution of signal and noise levels in the testing dataset. JUMPm first defines noise peaks as those which cannot be repeatedly detected in adjacent MS1 scans (i.e. peaks detected in only a single scan, **Online Methods**). For each scan, the program collects these noise peaks, removes outliers and obtains the average intensity of the remaining peaks. The average is used as scan-specific noise level to enable the calculation of the signal-to-noise (S/N) ratio of all peaks. An S/N ratio of 3 is set as the default cutoff for the peaks but is user-adjustable.



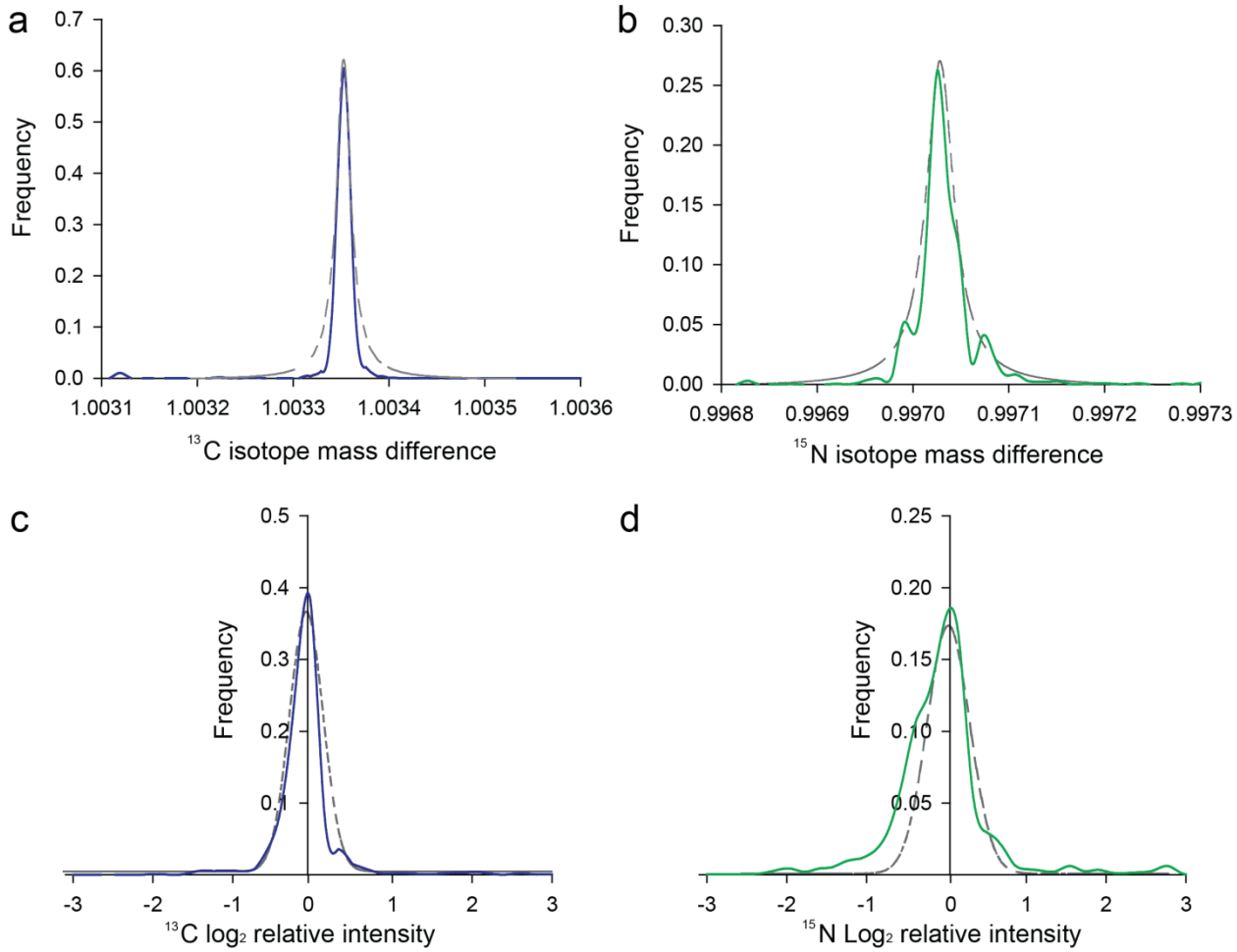
Supplementary Fig. 4. Mass correction by the moving average method. (**a**) Mass accuracy can drift during the course of a single run, and therefore the use of moving windows enables accurate mass correction. The mass error is determined by the mass difference between the detected and theoretical mass of the polysiloxane ion (445.120025, **Online Methods**). The entire chromatogram is typically divided into 10 windows, each containing the same number of scans. The mean of the mass errors in each window (dashed red line) is used to perform mass correction (**b**) on the ions within that window.



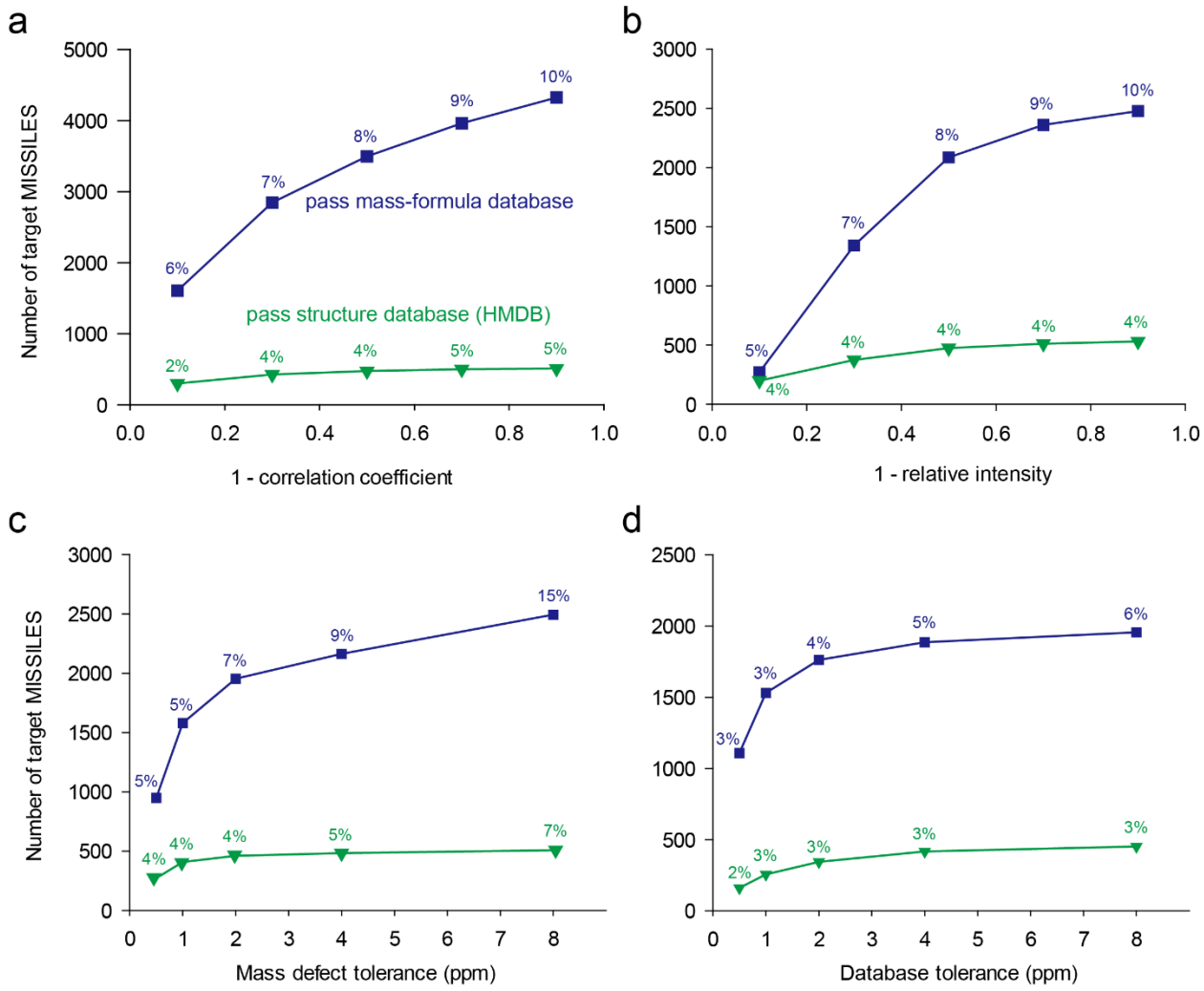
Supplementary Fig. 5. Other types of sample labeling compatible with JUMPm. (**a**) Example MS1 spectrum of unlabeled ($^{12}C^{14}N$) Adenosine triphosphate (ATP) from HEK293 cells (dish). Natural carbon atoms are ~98.9% pure, with ~1.1% of the atoms being the stable isotope carbon-13 (^{13}C). The tallest peak at 506 *m/z* is the monoisotopic peak, corresponding to ATP molecules composed of only ^{12}C carbon atoms. With 10 carbon atoms in the formula, there is a ~9% chance of any given ATP molecule containing 1 natural ^{13}C atom in its structure. Therefore, the peak at 507 *m/z* is ~9% as abundant as the monoisotopic peak. (**b**) an MS1 spectrum from a mixture of unlabeled ATP and partially labeled ATP (blue) from HEK293 cells. Unlabeled and partially labeled cells were grown separately then mixed (dishes). Although ATP has 10 carbons in its formula, we observed a mass shift of 5 *m/z* to the tallest labeled peak. The purity of the labeling determines the relative intensity and pattern of the labeled peaks. Δ = mass shift, due to isotope label, md = mass defect, unique to each stable isotope. (**c**) MS1 spectrum of ATP from a spike-in sample containing unlabeled ($^{12}C^{14}N$) HEK293 cells, $^{12}C^{15}N$ yeast, and $^{13}C^{14}N$ yeast (beakers). The $^{12}C^{15}N$ and $^{13C1}4N$ yeast samples are fully labeled, showing a shift of 10 carbons and 5 nitrogens, in agreement with the chemical formula of ATP. Compound identity was confirmed by authentic standards.



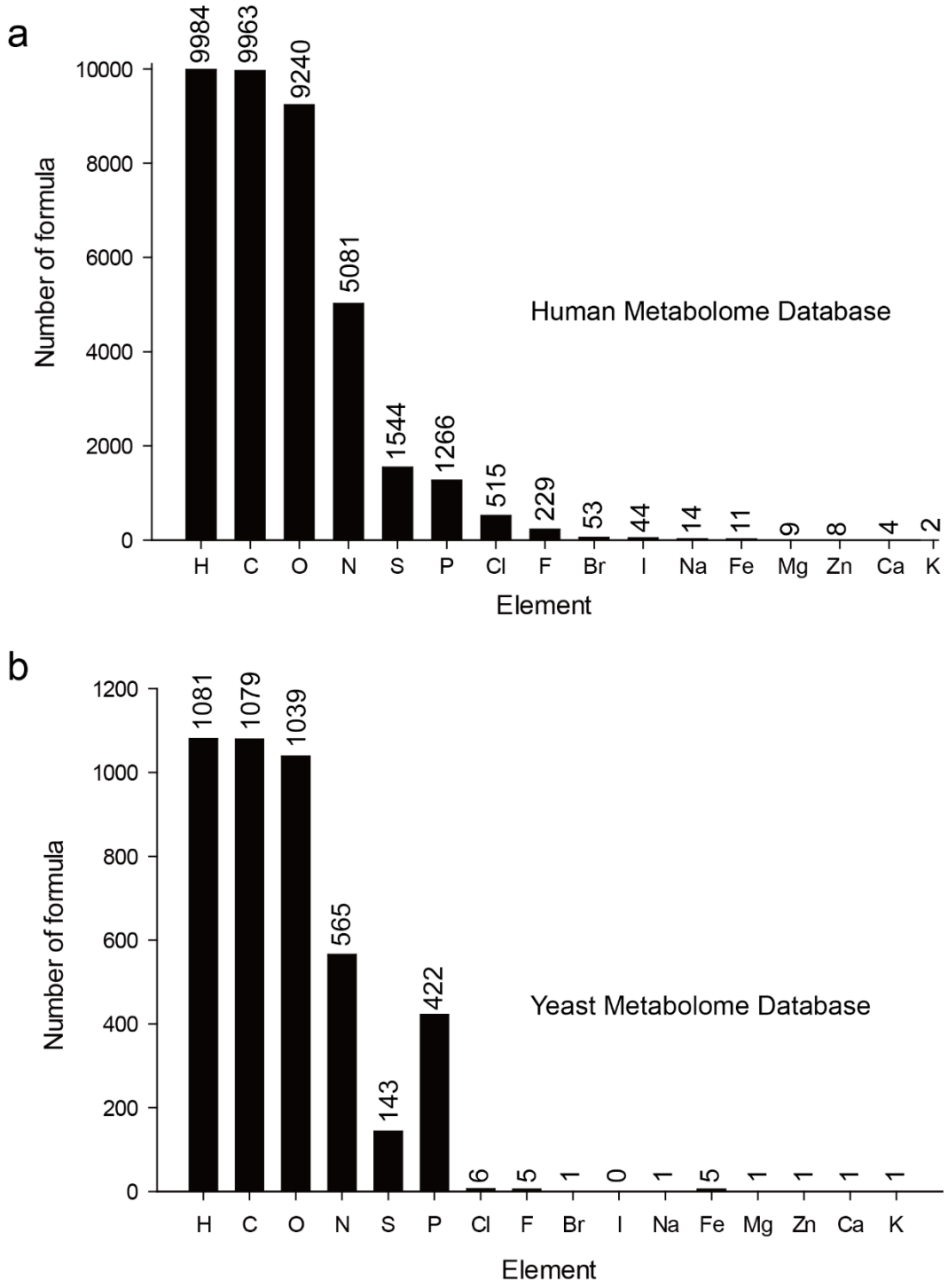
Supplementary Fig. 6. Conceptual workflow for JUMPm search of unlabeled and partially labeled metabolites. (**a**) MS1 spectrum of ATP from unlabeled ($^{12}C^{14}N$) HEK293 cells alone. Compound identity was confirmed by authentic standards. The relative intensity of the A+1 (nature abundance isotope peak) is used to estimate the number of carbon atoms in the formula. (**b**) MS1 spectrum of unlabeled and partially labeled ATP from a mixture of unlabeled and $^{13}C^{14}N$ labeled HEK293 cells. First, JUMPm detects ion pairs with a mass defect of 0.0033, specific to the ^{13}C label. Then, JUMPm searches the unlabeled peak, using the A+1 intensity to estimate the carbon number and identify potential formulas. From this list of formulas, JUMPm uses isotope pattern simulation on the partially labeled $^{13}C^{14}N$ (blue) cluster to determine the best matching formula.



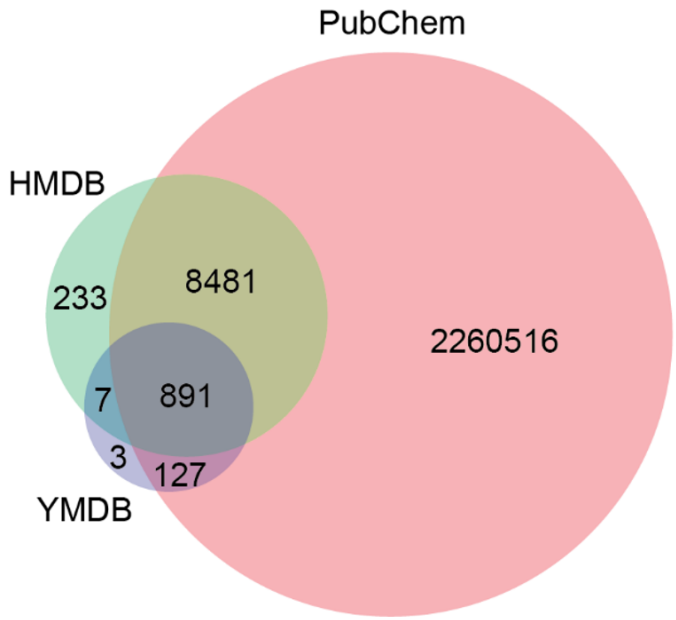
Supplementary Fig. 7. Determination of Pscore parameter tolerance. (**a-b**) Global frequency histograms for the isotope mass difference and relative intensity of the labeled peaks. The global tolerances for the ^{13}C and ^{15}N mass defects relative to the unlabeled monoisotopic peak were calculated by examining a population of the top 10% high confidence MISSILE formulas and their pairings. The mass defect is normalized about the monoisotopic mass, and is therefore primarily affected by mass resolution and mass accuracy, not the absolute mass value (small vs. large ions). The mass defect for both isotope labels was tightly distributed around the theoretical value. The nitrogen histogram is typically noisier because there are fewer nitrogen atoms in metabolites compared to carbon. This effectively reduces the sample size. (**c-d**) Relative ion intensity between labels also showed a normal distribution around ± 20%. Systematic shifts reflect unequal mixing of the samples while biological variation in metabolite levels contributes to the variation. For quantitative studies, the global ratio of labels can be used to fine-tune the mixture of samples for better quantitative sensitivity.



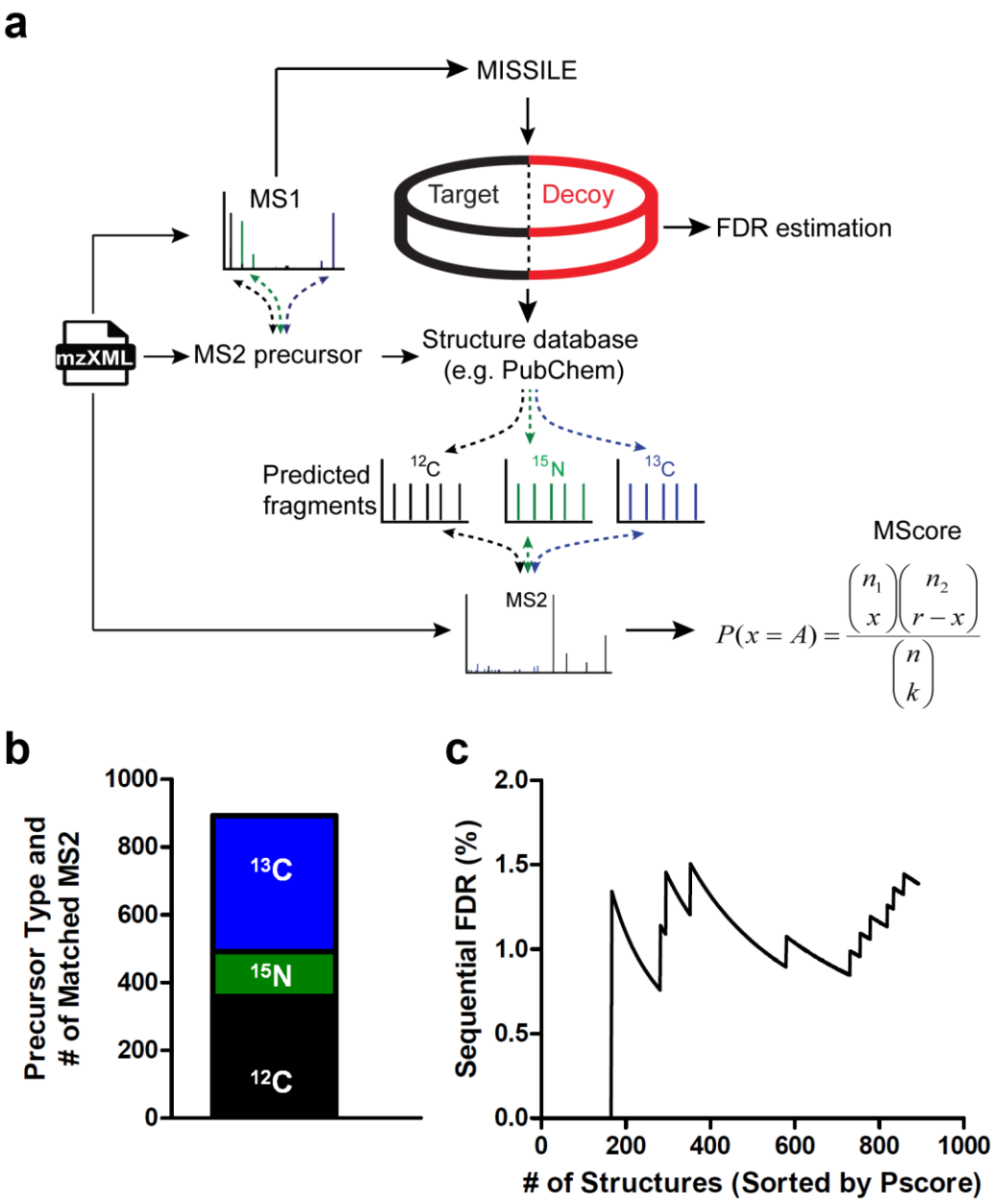
Supplementary Fig. 8. Systematic parameter optimization. (**a-b**) The number of target MISSILES (detected isotope labeled pairs) that pass both mass-formula and structure databases (e.g., HMDB) increases as the correlation coefficient and relative intensity decrease, respectively. (**c-d**) The number also increases as mass defect tolerance for pairing and database search tolerance increase. The structure database hits reach a plateau at 0.5 for both peak correlation coefficient and relative intensity, and at 4 ppm for both mass defect tolerance and database search tolerance, respectively. Percent numbers above the data points represent the apparent false discovery rate under those search parameters. These optimized parameters may be column and instrument dependent and should be reassessed for each user's platform.



Supplementary Fig. 9. Number of molecular formulas containing 16 selected elements in the HMDB and YMDB. We aimed to determine the relative representation of 16 common biological elements among the HMDB and YMDB. Outside of carbon, hydrogen, nitrogen, oxygen, phosphorus, and sulfur, the remaining elements were uncommon. The YMDB is relatively enriched for metabolites containing phosphorus compared to the HMDB. However, metabolites composed of C, H, N, and O account for the bulk of formula entries.

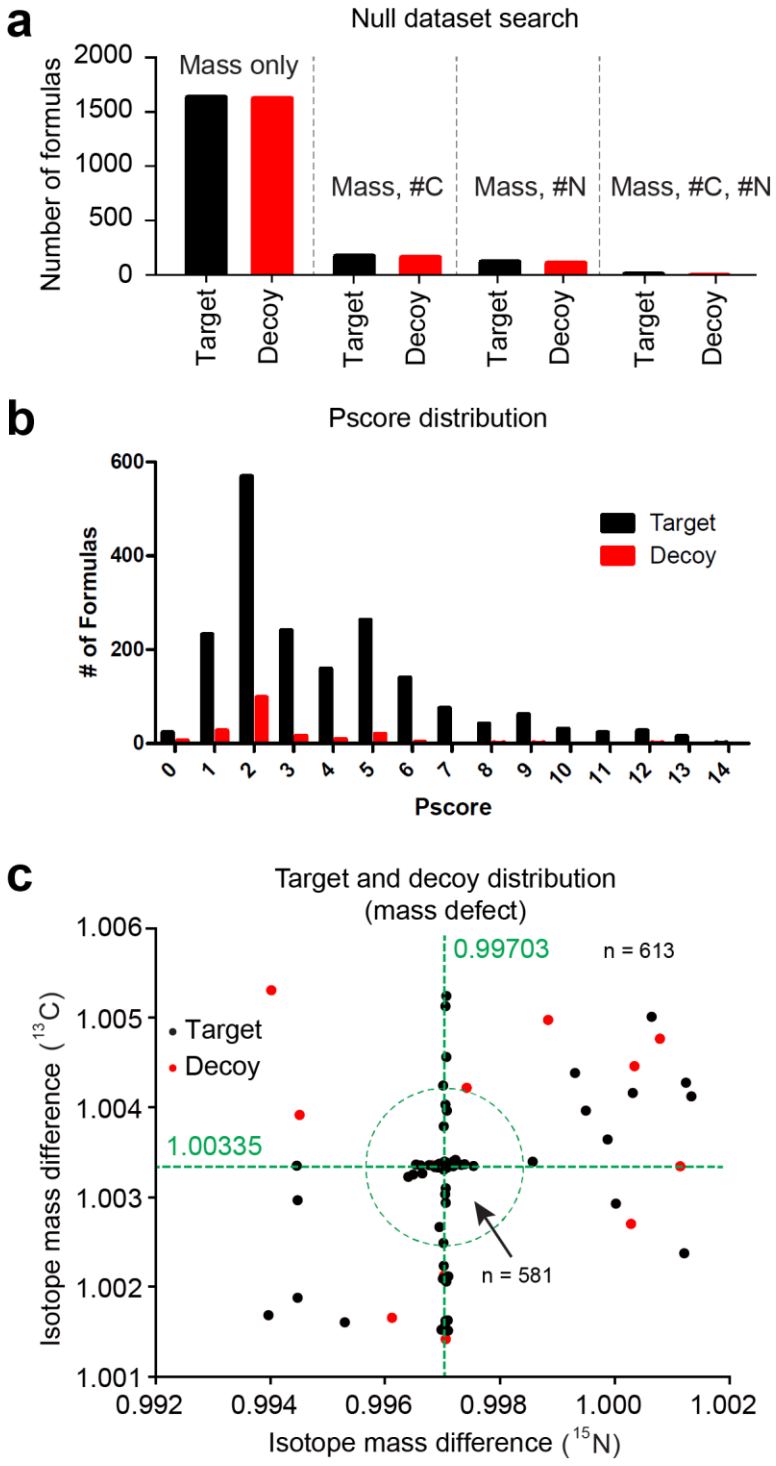


Supplementary Fig. 10. Comparison of the number of formulas between the PubChem, HMDB, and YMDB databases. We downloaded local copies of these publically available compound/metabolite databases to determine the degree of redundancy between data sources. Overall most formulas were contained within Pubchem.

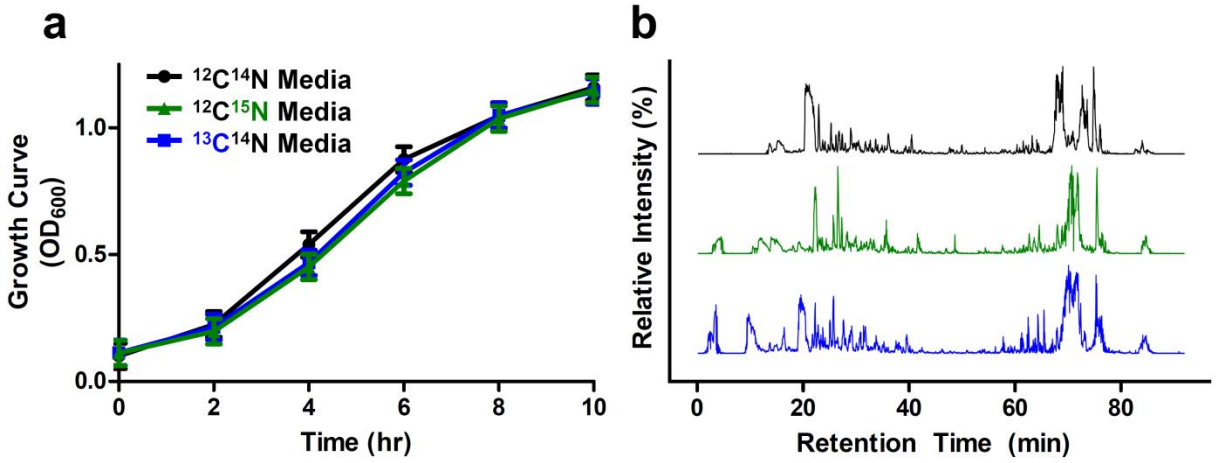


Supplementary Fig. 11. Scoring of metabolite-spectrum matches. (**a**) Putative MISSILE formulas are searched against the target and decoy databases to estimate the rate of false discovery. For each formula JUMPm identifies all known structures sharing the specified chemical formula. For each candidate structure, the tandem mass spectrometry fragments are predicted for the unlabeled and theoretical labeled versions using the CDK implementation of MetFrag. Empirical tandem scans are matched and scored against the predicted fragments using the formula shown (Mscore). (**b**) Proportion of MISSILEs with MS2 scans from each type of isotope label. The data-dependent Top N acquisition may target ^{12}C, ^{15}N, or ^{13}C parent ions. Therefore, only ~1/3rd of tandem scans come from carbon-12 metabolite precursors. There are fewer ^{15}N precursors in part because less than half of metabolites contain nitrogen. (**c**)

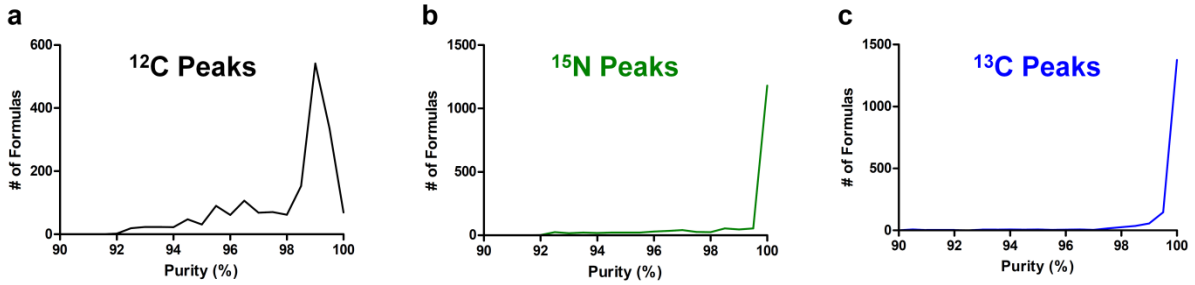
Cumulative FDR of structures identified by JUMPm. The identified structures were sorted by Pscore from high to low (left to right). As lower Pscore hits are included, the number of decoys accumulates, increasing the FDR for that subset of the results. Each "spike" in the plot is due to a decoy hit. JUMPm uses a Pscore cutoff to trim the data to achieve the user-specified FDR.



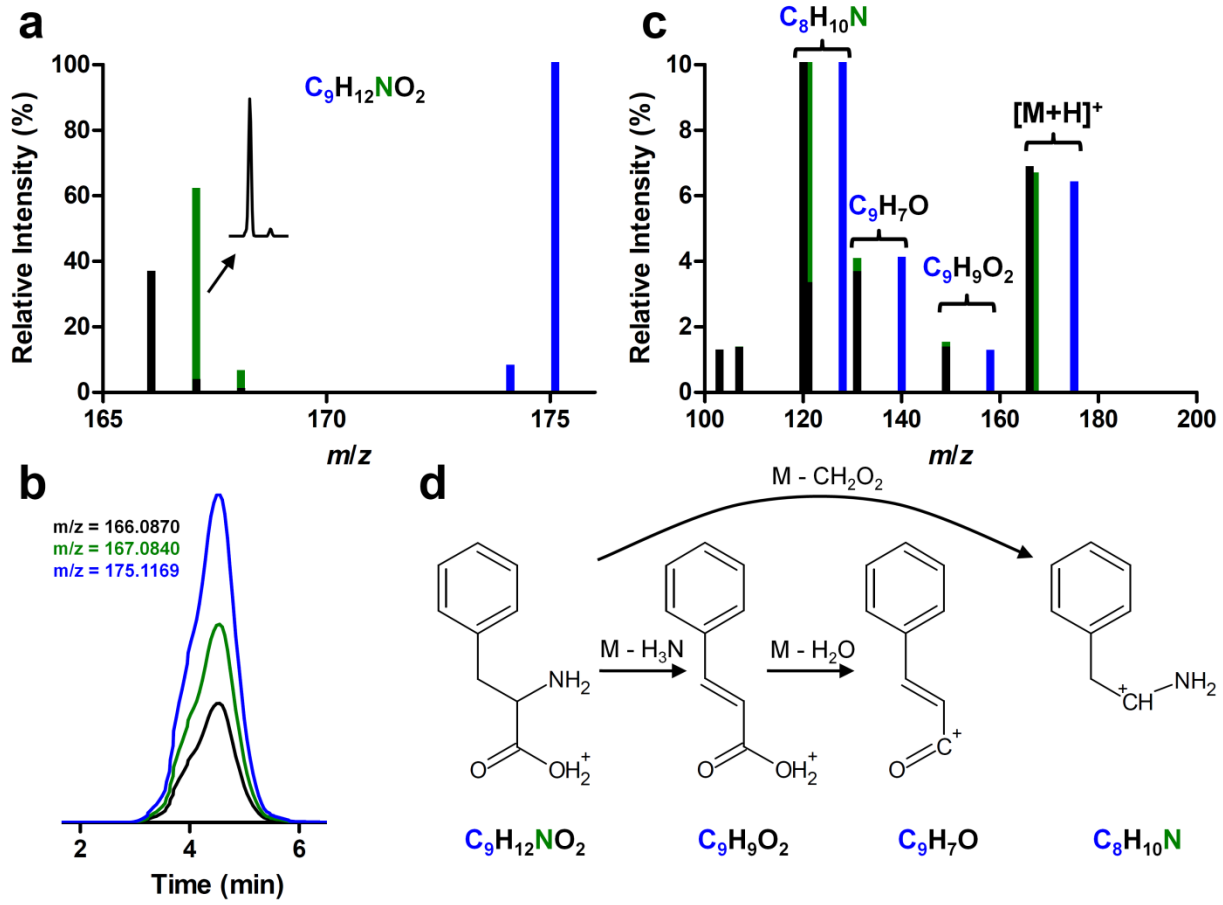
Supplementary Fig. 12. FDR characterization with labeled yeast. (**a**) Effect of formula search criteria on the number of hits and the FDR with a null (negative control) dataset. More stringent criteria (e.g., both mass and carbon number) reduced the number of known spurious hits (target and decoy) without affecting the apparent FDR. (**b**) Histogram of formulas as a function of Pscore. The number of decoys significantly decreases as Pscore increases. JUMPm can trim the data using the Pscore to achieve the user-specified FDR (e.g., 1%), real dataset used for analysis. (**c**) Two-dimensional distribution of target and decoy matches by isotope mass differences (581 out of 613 within the green circle), which are derived from the accurate delta mass between labeled and unlabeled ions divided by the integer delta mass (**Online Methods**). Decoy matches are randomly scattered in the plot while targets are tightly clustered around the expected mass differences of 1.00335 (carbon) and 0.99703 (nitrogen), real dataset used.



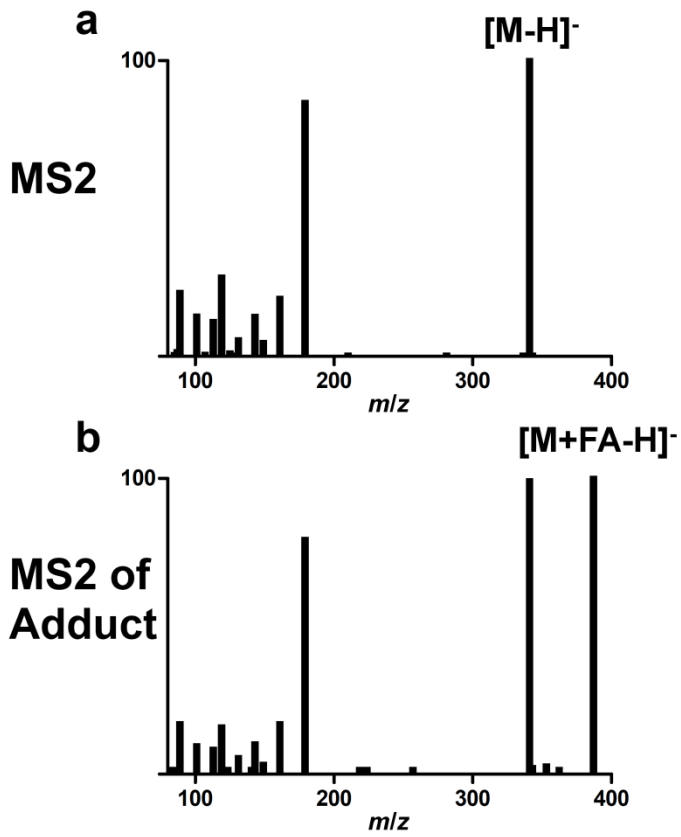
Supplementary Fig. 13. Stable isotope labeling in yeast. (**a**) Growth curves for each of the isotopic yeast cultures (unlabeled ^{12}C), ^{15}N, ^{13}C). (**b**) Base-peak LC-MS chromatograms for each of the yeast culture metabolite extracts, C18 column in positive mode (**Online Methods**).



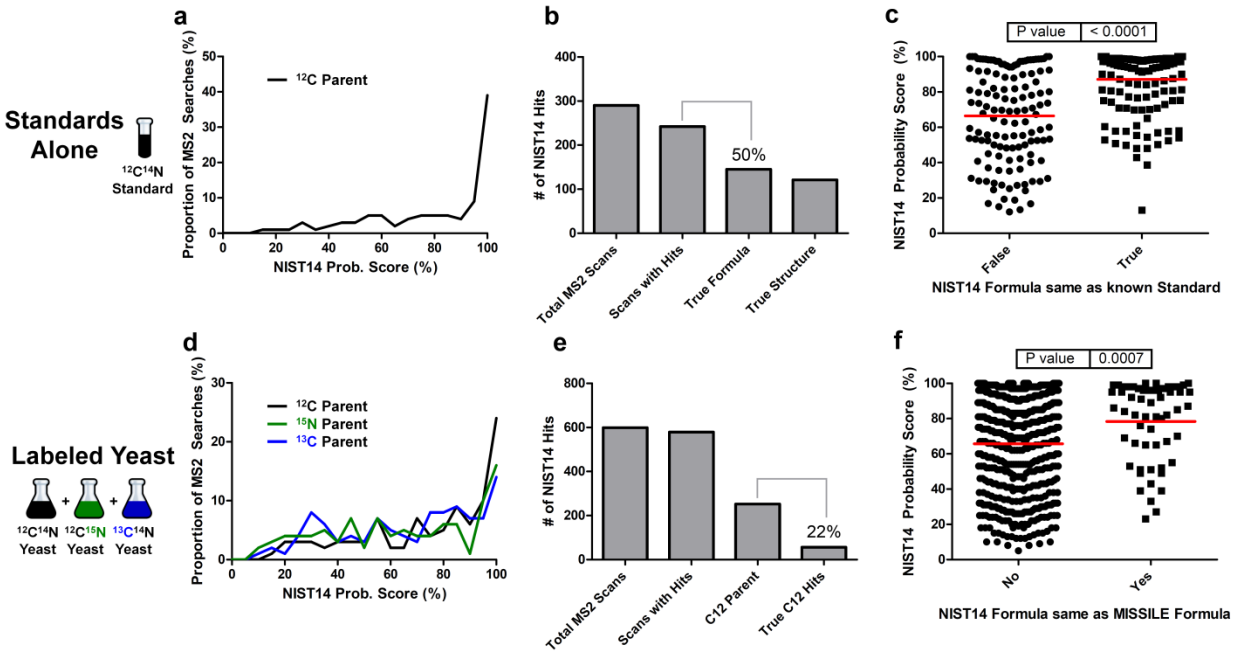
Supplementary Fig. 14. Purity of isotope incorporation in yeast metabolites. Formulas detected by JUMPm were subjected to isotopic pattern simulation analysis to determine the purity (% labeling) of each labeled ion (**Online Methods**). (**a**) Histogram of unlabeled (^{12}C) peaks (n=1722). The mode purity among unlabeled peaks was 99%, in close agreement with the natural isotopic purity of ^{12}C (98.9%). (**b**) Histogram of ^{15}N labeled peaks (n=1631). The mode purity among ^{15}N labeled peaks was 99%. (**c**) Histogram of ^{13}C labeled peaks (n=1722). The mode purity among 13C labeled peaks was 99%.



Supplementary Fig. 15. MS/MS annotation using stable isotope information. (**a**) MS1 spectrum of phenylalanine including unlabeled (black), $^{12}C^{15}N$ labeled (green), $^{13}C^{14}N$ labeled (blue) peaks. Arrow points to a profile trace of the raw MS1 data showing complete resolution of the natural isotope A+1 peak and the $^{12}C^{15}N$ labeled peak. (**b**) Co-eluting labeled ions of phenylalanine in the mixed yeast sample. (**c**) Overlay (n=3) of separate tandem MS scans for each of the labeled ions in (**a**). (**d**) Structural assignment of the phenylalanine fragmentation pathway.



Supplementary Fig. 16. Analysis of a formic acid adduct of trehalose from yeast. (**a**) MS2 spectrum of the unlabeled trehalose. The $[M-H]^-$ molecular ion still visible among the fragments. Characteristic fragments observed below 200 *m*/*z*. (**b**) MS2 spectrum of the unlabeled $[M+FA-H]^-$ adduct of trehalose. Parent ion and loss of formic acid molecular ion still visible among the fragments.



Supplementary Fig. 17. Spectral library search of standards and yeast MS2 scans. (**a-c**) Spectral library search results for the synthetic standard library. (**a**) Distribution of NIST14 probability scores for MS2 spectra from known standards (**Supplementary Table 5**). (**b**) NIST14 MS/MS database results by status. About 50% of spectra searches gave the same formula as the known standard. Overall, 42% were also the same structure as the standard compound. (**c**) NIST14 probability scores among True and False hits for the standard library. (**d-f**) Spectral library search results for selected yeast metabolites (n=599). (**d**) Distribution of NIST14 probability score for MS2 spectra from unlabeled (^{12}C), ^{15}N, and ^{13}C parent ions. (**e**) NIST14 MS/MS database results by status. ^{12}C parents are MS2 scans from unlabeled yeast precursor ions. "True" hits are those where NIST14 returned the best hit with the correct chemical formula while "False" hits are those where NIST14 returned the best hit with an incorrect chemical formula. (**f**) NIST14 probability scores among True and False hits for the yeast metabolites.