As a test case for optimizing how to perform proteomics experiments, we chose a yeast model system in which the UPF1 gene, a protein involved in nonsense-mediated mRNA decay, was knocked out by homologous recombination. The results from five complete isotope-coded affinity tag (ICAT) experiments were combined, two using matrix-assisted laser desorption/ionization (MALDI) tandem mass spectrometry (MS/MS) and three using electrospray MS/MS. We sought to assess the reproducibility of peptide identification and to develop an informatics structure that characterizes the identification process as well as possible, especially with regard to tenuous identifications. The cleavable form of the ICAT reagent system (Gygi et al. (1999) Nat. Biotechnol. 17, 994–999) was used for quantification. Most proteins did not change significantly in expression as a consequence of the upf1 knock-out. As expected, the Upf1 protein itself was down-regulated, and there were reproducible increases in expression of proteins involved in arginine biosynthesis. Initially, it seemed that about 10% of the proteins had changed in expression level, but after more thorough examination of the data it turned out that most of these apparent changes could be explained by artifacts of quantification caused by overlapping heavy/light pairs. About 700 proteins altogether were identified with high confidence and quantified. Many peptides with chemical modifications were identified, as well as peptides with noncanonical tryptic termini. Nearly all of these modified peptides corresponded to the most abundant yeast proteins, and some would otherwise have been attributed to "single hit" proteins at low confidence. To improve our confidence in the identifications, in MALDI experiments, the parent masses for the peptides were calibrated against nearby components. In addition, five novel parameters reflecting different aspects of identification were collected for each spectrum in addition to the Mascot database sequence corresponding to an identified peptide; EID, experiment ID (HPLC run index); EM, electrospray versus MALDI; Ex, experiment number; Fr, ion exchange fraction number; FrT, Peptide Fragment TriScore; HL, heavy-to-light ratio; HLI_1, a key that indicates whether an identified HL pair is free of overlapping isotope clusters; HL_1, a key that indicates whether an identified HL pair is consistent with the number of Cys residues and HL_Type of the corresponding peptide sequence; HL_S, a string that indicates whether there are confounding isoform clusters at the position of each possible HL pair; HLI_1, a key that indicates whether an identified HL pair is free of overlapping isoform clusters; HL_1, a key that indicates whether an identified HL pair is consistent with the number of Cys residues and HL_Type of the corresponding peptide sequence; HL_S, a string that indicates whether there are potential overlapping HL pairs; HL_Type, a key that indicates whether a peptide is in the c0 form, the c9 form, or mixed; HPLC, high-performance liquid chromatography; IEX, ion exchange chromatography; Int, intensity; IntD, sum of the intensity of the fragment masses in MFU; IntP, parent ion intensity; Ippm, ppm difference between the masses in a HL pair; ISD, indicates whether the parent ion may have been generated by in-source decay from a different parent ion; IT, the number of missed trypsin cleavage sites within a sequence; KICAT,
quantify the relative expression levels of each protein based on the relative intensity of the heavy and light forms of Cys-containing tryptic peptides as measured by mass spectrometry. More than 700 yeast mRNAs are regulated by UPF1; that is, their expression levels increase 2-fold or more when Upf1p is inactivated (7), and their respective extents of translation may thus be comparably enhanced. However, genes encoding the most abundant proteins in yeast have evolved such that their mRNA levels are not affected by this pathway (7), making this system a good test for detection of small changes in expression levels.

One simple statistic for determining the relative success of a proteomics experiment is to count the number of correctly identified peptides and proteins. Because identification of peptides by electro spray is dependent on appropriate automated selection of precursor ions, repetition of experiments has frequently been observed to result in a higher number of identifications. In the matrix-assisted laser desorption/ionization (MALDI) approach, the output of the reversed-phase columns is deposited on the MALDI plate, so that time is much less of a limitation in selecting the most informative set of precursors for fragmentation. However, even in this approach, one can expect a larger number of identifications upon repetition because of slight changes in elution times, differences in matrix crystalization, experiment-specific loss of peptides by electrospray is dependent on appropriate automation, or limitations on sample consumption by the acquisition process itself. For both approaches, another desired output is the dependable selection of precursor ions, repetition of experiments may thus be comparably enhanced. However, genes encoding the most abundant proteins in yeast have evolved such that their mRNA levels are not affected by this pathway (7), making this system a good test for detection of small changes in expression levels.

In these experiments, we combined together identifications from two different instrument types that initially used two different search engines. After combining the data into a common relational database, we submitted all of the spectra corresponding to identified peptides to Mascot. In addition, we developed new measures of reliability of identification and quantification that proved useful in resolving discrepancies in identifications. These parameters are adapted from the parameters we recently described for peptide mass fingerprinting experiments (8). In addition to the percent intensity matched parameter, we describe a percent ChemScore matched parameter that semiquantitatively assesses what percentage of the critically important ion fragments were detected. A third parameter was the internal mass consistency of the fragment ions. We also describe a fourth parameter, called Fragment TriScore, that gives higher credit to spectra in which the fragments with the highest ChemScore are also the most intense ions. These parameters were combined in an overall MatchScore parameter that is useful in documenting the credibility of an identification, especially in marginal cases. The parent mass accuracy parameter was left separate as an independent measure of credibility of identification. Upon calibration, based on added calibrants or masses that can be used as calibrants because they can be confidently assigned, the mass of a precursor ion should not deviate from the theoretical peptide mass by more than 20 ppm.

In order to determine which factors were most significant in preventing additional protein identifications, a lot of attention was focused on spectra that were not readily identifiable. In many cases, these spectra could be attributed to chemically modified forms of the peptides that had already been identified many times and were therefore surely very abundant. A second class of these peptides corresponded to peptides with one noncanonical tryptic terminus from these same abundant proteins. An additional problem is spectra whose measured masses were several mass units different from the mass of the peptide to which they were matched. To quantify these problems, each of the spectra were classified in groups corresponding to chemical modification status, mass accuracy, and tryptic specificity.

**MATERIALS AND METHODS**

**Yeast**

Two strains of yeast were studied in these experiments. The strain we describe herein as “wild-type” has been designated HFY1200 (5); it has mutations in ade2, his3, leu2, trp1, and can1, which come in to play when the yeast is grown in restricted media. The upf1 knockout strain has been designated HFY871 (5). It has the same genetic background as HFY1200, but has the HIS3 gene inserted in place of...
the UPF1 gene. Wild-type yeast and the upf1 knockout strain were grown to mid-log phase (OD600 = 0.7) in 2 liters of yeast extract-peptone-dextrose (YPD) medium at 30°C in a fermentor. All subsequent procedures were performed at 4°C. Yeast cells were collected by centrifugation at 4,000 × g for 5 min and washed with 200 ml of water and then 200 ml of 50 mM Tris-CI, pH 7.5 (buffer A). The yeast extracts were prepared using the liquid nitrogen (LN₂) grinding method (9). The cell pellets were resuspended in 1/10 volume of buffer A and then carefully mixed into LN₂ to form beads. The beads were crushed and ground to fine powder in LN₂ using a prechilled mortar and pestle. The fine powder was stored at −70°C. The soluble fraction of the yeast extracts was prepared by thawing the fine powder on ice for 15 min and then collecting the supernatant by centrifugation at 14,000 rpm for 5 min using a microcentrifuge. The protein concentration of the soluble fraction was determined using a Bradford assay (10). Each 2-liter culture yields about 4 g of cell pellet, and the estimated protein yield for each soluble fraction is about 400 mg.

Peptide Chemistry

ICAT Reagent Labeling Procedure—Two 500-µg aliquots from each strain were resuspended in 6 µL guanidine-HCl, 1% Triton X-100, 50 mM Tris HCl, pH 8.5 (buffer B). The proteins were then reduced by the addition of 10 µL of 50 mM tricarboxyethylphosphine and boiled at 100°C for 10 min. After cooling for 5 min to room temperature, 1 mg of the acid-cleavable form of the ICAT light reagent, dissolved in acetonitrile (ACN), was added to the wild type, whereas 1 mg of the acid-cleavable form of the ICAT heavy reagent, dissolved in acetonitrile (ACN), was added to the knockout sample. After incubation for 2 h at 37°C, the two aliquots were combined and precipitated with acetone (6:1 volume of acetone: volume of sample). The precipitated proteins were centrifuged for 10 min at 13,000 × g, the acetone was decanted, and the pellet was resuspended in 100 µL of ACN. The sample was then diluted with 900 µL of 50 mM Tris, pH 8.5, 10 mM CaCl₂, 20% ACN. Then 12 µg of porcine trypsin (Promega, Madison, WI) was added, the sample was incubated for 2 h at 37°C, then another 12 µg of porcine trypsin was added, followed by overnight digestion.

Ion Exchange Chromatography (IEX)—The sample (1 mL) was diluted to 10 ml with 10 mM K₃PO₄, 25% ACN, pH ~2.5 (buffer C). In two batches, the sample was injected onto a 4.6 × 100 mm polyethylic A cation exchange column at a flow rate of 1 mL/min. The high salt buffer contained 350 mM KCl, 10 mM K₃PO₄, 25% ACN, pH ~2.5 (buffer D). Peptides were separated over four linear gradient segments using an Applied Biosystems Vision Work station (Applied Biosystems, Foster City, CA) in order to separate the peptides as efficiently as possible: 2 min to 10% buffer D, 15 min to 20% buffer D, 3 min to 45% buffer D, and 10 min to 100% buffer D. Seven to 23 fractions (see Table V, column No. IEX) consisting of 1.5 ml were collected beginning 4 min into the gradient. Prior to affinity chromatography, 250 µl of 100 mM Na₂PO₄, 1500 mM NaCl, pH 10 was added to each fraction, which was sufficient to bring the pH to ~7.2.

Avidin Affinity Chromatography—Each ion exchange fraction was separated using the monomeric avidin beads supplied with the ICAT reagent kit (Applied Biosystems), according to the instructions.

Cleavage of Biotin—Each eluate was dried completely using reduced pressure. A 200-µL aliquot of ICAT cleaving reagent from the ICAT reagent kit was added, followed by incubation at 37°C for 2 h. Once again the sample was dried under reduced pressure until time for reversed-phase separation. At that time, each sample was resuspended in 100 µL of 2% ACN, 0.1% trifluoroacetic acid (TFA).

Mass Spectrometry

Electrospray Analysis—Three or four dependent scans were collected per mass spectrometry (MS) scan using a QStar® Pulsar I (Applied Biosystems) equipped with a nanospray source using Analyst software. Typically, data was collected over 110 min at a flow rate of 0.3 µL/min, with 3-s parent scans from 300 to 1500 m/z, and with tandem MS (MS/MS) scans from 70 to 1500 m/z every 3 s. Information-dependent acquisition was used to select the most intense parent masses excluding singly charged ions and using dynamic exclusion to prevent the same parent mass from being chosen for fragmentation within a 45-s window. Samples were injected onto a capillary trap cartridge (Captrap, Michrom Biosources, CA) at a flow rate of 20 µL/min to concentrate and desalt the samples. After 10 min, the trap cartridge was automatically switched in-line with the analytical column. Peptides were separated using a 75-μm × 15-cm reversed-phase C18 column, 3-μm particle size (PepMap; LC Packings, Hercules, CA), by means of an Ultimate™ System ( Dionex Corporation, Sunnyvale, CA). The gradient was typically from 5 to 30% buffer B, where buffer B is 85% ACN/10% water/5% n-propanol/0.1% formic acid/0.1% TFA and buffer A is 98% water/2% ACN/0.1% formic acid/0.1% TFA.

MALDI Analysis—After cleavage, the peptides were separated using an Ultimate Chromatography system ( Dionex-LC Packings, Hercules, CA) equipped with a Probot MALDI spotting device. A total of 50 µL of each digested protein fraction was injected and captured on a 0.3 × 5-mm trap column (3 µm, Dionex-LC Packings) with a 0.1 × 150-mm resolving column (3 µm, Dionex-LC Packings) connected in series. Peptides were resolved by dual-solvent gradient elution at a flow rate of 800 nL/min, using a gradient of 5–45% buffer B over 35 min, followed by a gradient of 35–90% buffer B over 5 min, where buffer B is 98% water, 2% ACN, 0.1% TFA and buffer A is 85% ACN, 10% water, 5% isopropanol, 0.1% TFA. Column effluent was monitored using a 3-nl ultraviolet flow cell and spotted directly onto a MALDI target using the Probot. Column effluent was mixed 1:2 with MALDI matrix (7.5 mg/ml α-cyano-4-hydroxycinnamic acid dissolved in 75:25 ACN:water containing 0.15 mg/ml dibasic ammonium citrate) by means of a 25-nl mixing tee (Upchurch Scientific, Oak Harbor, WA) and was spotted onto the target in a 12 × 12 array at 20-s intervals. All MALDI spectra were acquired using a 4700 Proteomics Analyzer (Applied Biosystems) equipped with GPS Explorer version 1.0, and peaks were selected for MS/MS analysis using a strategy to collect as many useful spectra as possible without regard to heavy-to-light (HL) ratio (11). To accomplish this, the parent spectra were collected first, and masses were chosen for fragmentation from the spots in which they were most intense using the PeakPicker program (11).

Peptide Identification Methods—QStar files were submitted to the ProCat software package (Applied Biosystems) for the process of both identification and quantification. The database used was Swiss-Prot release 36. ProCat generates a set of tables housed in an Access relational database. From these tables, two new tables were generated using SQL queries: A table that contained one row for each mass spectrum (see Table III) and a table that contained the peak list information from each mass spectrum (see Table VI). Although the HL ratio and peptide sequence were originally in distinct tables, these values were incorporated into Table III.

The MALDI spectra were identified using Mascot version 1.9 (12). The database used was MSDB (Matrix Sciences Ltd., London, United Kingdom) from the June 1, 2003 release, containing 9722 Saccharomyces cerevisiae sequences. MALDI files were originally stored in an Oracle database generated by GPS 1.0. All of the mass and intensity information was extracted into tables in Access, along with the sequences of the peptides proposed to match that were obtained by the Mascot search engine. The isotope cluster intensities for the HL pair peaks of the parent spectrum that had been deposited into the Oracle database were used for quantification. Like the electrospray data, these data were housed in Tables III and VI. These steps correspond to Fig. 1, Workflow 1, steps 1–3.
A Visual Basic program (MatSc_Calc) was written that calculates MatSc (see below in “Calculation of Overall Score”) as well as most of the additional parameters in Table III starting from the proposed sequence, the experimental mass, and the peak/intensity list. In order to compare identifications from electrospray data to MALDI data, the peak lists for identified peptides from both electrospray and MALDI were combined together and resubmitted to Mascot (Fig. 1, Workflow 1, step 5). This generates a Mascot score for the electrospray data. In addition, the protein consolidation capability of Mascot ensures that the smallest possible list of proteins is deposited into Table III, with common accession numbers regardless of which databases were initially searched (Fig. 1, Workflow 1, step 6). Efforts were made to ensure that no identifications were lost in this process, which centered on determining which spectra were most crucial for the identifications of each peptide and protein (Fig. 1, Workflow 1, step 7). In many instances, spectra were examined manually (Fig. 1, Workflow 1, step 8). In some cases, MatSc_Calc was enabled to overwrite the initial sequence with alternative sequences (for example, sequences that Mascot or ProICAT had identified with lower confidence) when the alternative sequence had a higher MatSc (see Fig. 1, Deciding Between Alternative Sequences). Occasionally, queries were performed to determine how many identifications were changed to different sequences by this means, and MatSc_Calc or its input parameters was altered so that those sequences that appeared to be correct by manual examination of selected spectra resulted in the highest MatSc. The final set of input parameters used are listed in Table I. Many of the tracking parameters in Table III were manually appended using combinations of additional programs and update queries within Access.

To obtain the identifications in Table III, certain spectra were submitted to multiple additional rounds of Mascot searching (Fig. 1, Workflows 2–4). In most cases, whenever a peptide was identified with high confidence, all spectra with parent masses within 2 amu were searched (using MatSc_Calc) to find additional spectra that matched the same peptide (Fig. 1, Workflow 2). A modified form of MatSc_Calc was used to search for the oxidized form of some of the previously identified Cys-containing ICAT reagent modified (CICAT) peptides (Fig. 1, Workflow 2, step c). In other cases, MatSc_Calc was used in search mode to look for peptides of particular interest, for example, from Upf1p (Fig. 1, Workflow 2, step d). In these cases,
MatSc_Calc was routinely set to identify any spectrum that had the minimal criteria for identification; typically, >3 yb ions, a MatSc >5000, and parental mass accuracy of 300 ppm. In other cases, these criteria were lowered so as to annotate any spectrum that matched within 1 amu of the proposed sequence that could not be better explained by alternative sequences. This normally did not result in additional identifications with any confidence (except for e.g. human keratin peptides that were missed at the database level). Instead, this process enabled us to annotate parent masses from which additional spectra could in principle be acquired to substantiate or to rule out such an identification.

A large number of spectra were examined manually (Fig. 1, Workflow 3). Special attention was paid to spectra that were derived from unmatched, intense MS/MS spectra (Fig. 1, Workflow 3, step 1a), or intense precursors that were not automatically matched (Fig. 1, Workflow 3, step 1b). We were also particularly interested in spectra that had precursor masses that appeared to belong to HL pairs that indicated differential expression (Fig. 1, Workflow 3, step 1c). Other spectra were selected for special attention because they matched to proteins that we did not expect to be abundant, or that contained unusual modifications, or did not conform to trypsin cleavage rules (Fig. 1, Workflow 3, step 1d). Some spectra were examined randomly as a spot check of the annotation process (Fig. 1, Workflow 3, step 1e). Some of the methods used to identify these spectra are listed in Fig. 1, Workflow 3, section 2. For some spectra, the precursor mass or charge state was adjusted (Fig. 1, Workflow 3, step 2b). In other cases, the peak list was refined manually because of inappropriate de-isotoping, or because some fragment ions appeared to be derived from substances other than peptides (Fig. 1, Workflow 3, step 2c). In this case, the peak list was usually not permanently altered, and MatSc was calculated using the peptide sequence that was determined manually.

After we had uncovered evidence for peptides with new modifications, selected high-performance liquid chromatography (HPLC) runs were resubmitted to Mascot, for example, with lysine-modified ICAT reagent (KICAT), oxidized CICAT, or N-terminal ICAT reagent modified as alternative variable modifications (Fig. 1, Workflow 4). Once again, MatSc_Calc was used as the criterion to resolve conflicts. Regardless of how many searches were performed, in the end, each spectrum was allowed to match to no more than one peptide sequence. Some spectra were manually placed into Y1 class 99 (see Table IVB), because manual examination of the spectrum indicated that a proposed sequence was implausible, yet MatSc was significant.

Spectrum Classification

Tables I and II list the input parameters used in the calculation of MatSc.

Peak List Filtering—For intensity truncation, the initial peak lists were extracted by GPS 1.0 for MALDI analysis, and by ProICAT for electrospray analysis. To the first approximation, the higher the percentage of the intensity that can be accounted for a proposed match (the Percent Intensity Matched [PIM]), the more confident the identification. However, we found that if the spectrum contained from one to three very intense masses, these intense masses often eliminated the discriminating power of the PIM term because they overwhelmed the contributions by less-intense matched ions. To avoid this problem, the intensities of the most intense three daughter ions were “truncated” to the intensity of the fourth most intense daughter ion, thereby allowing the weaker ions to contribute to the calculations (Table II, rule 8). In the case of a “rich” spectrum with many nearly equally strong peaks, truncation is not necessary and has little effect.

For peak density filtration, in many instances, the most important ions for determining the correct sequence are relatively high in molecular mass but weak in overall intensity. To ensure that the peak list contains the most significant ions from all regions of the mass spectrum, the peak list was filtered to eliminate all but the six most intense masses every 100 amu. Later, these same processed peak lists were resubmitted to Mascot so that the Mascot scores listed were derived from these altered peak lists.

Percent Intensity Matched—PIM is calculated beginning with the filtered peak list. Only masses that are greater than 200 amu and smaller than 50 amu below the precursor mass are allowed to contribute to PIM because most masses outside this range are not sequence specific (Table II, rules 6 and 7).

Calculation of Fragment Ion ChemScore—The Fragment Ion ChemScore is designed to approximate the theoretical expected intensity for each ion type (see Refs. 13 and 14 for alternative methods to do this). In this article, the ChemScore for each fragment ion is calculated based on the rules listed in Table I. If it was desired, these assumptions could be tuned to the instrument type (electrospray versus MALDI), but in this article we have used the same settings for all spectra.

In this scheme, it is possible to count any desired ion type. In this article, we have counted only y ions, b ions, y-17 ions, b-17 ions, certain immonium ions, and y and b ions generated by neutral loss of 64 amu from oxidized Met.

The first principle is that y ions and b ions are the most important. y ions have been assigned an arbitrary score of 6000, versus 5000 for b ions, etc. (Table I, rules 9 and 10). The ChemScore for any ion that contains a Lys or Arg is further multiplied by 1.5-fold (Table I, rules 3 and 4). Because we and others (15, 16) have observed preferential fragmentation before Pro (in all ion series), and after Asp, and to a lesser degree Glu, the ChemScores of all such ions have been multiplied the factors described by Table I, rules 5, 7, and 8. In addition, the ChemScores of all ions C-terminal to Pro have been decreased by 1.5-fold (Table I, rule 6), because these ions are less frequently detected (13).

We have observed that y-17 and b-17 ions are more intense when the N-terminal amino acid (aa) is Gln, possibly because of facile elimination of ammonia by cyclization; thus such ions are given the same ChemScore as the corresponding y or b ionic (Table I, rule 14). The remaining (y-17) and (b-17) ions are given 20-fold more credit than other -17 ions if they contained R, H, or Q (Table I, rule 15). We have also observed unusually intense y-17 and b-17 ions after ICAT reagent labeling and acid cleavage; thus ICAT reagent-labeled Cys-containing (CICAT) fragments are dealt with the same as R, H, and Q (Table I, rules 16 and 17).

Most neutral losses of water have not been considered here, except for the case of N-terminal Glu, which are often unusually intense (Table I, rule 18).

Because the b2 ion and a2 ion often seem to be more intense than other a and b ions, the ChemScore for the b2 ion has been multiplied by 1.5-fold, whereas the ChemScore for the a2 ion is multiplied by 6.6-fold (Table I, rules 19 and 20).

Finally, it has been observed that an ion having the molecular mass of the parent ion minus the C-terminal aa is often present, especially when the sequence contains an internal arginine (which in general makes fragmentation of any kind more difficult). In this report, the ChemScore of the b(n−1)+18 ion is counted the same as an ordinary b ion (Table I, rule 21). In order to promote detection of ions, the timed ion selector in MALDI mode is typically relaxed so that small amounts of ions from the unselected HL pair are often detected. We decided to add these ions to the fragment list and assign them ChemScores 10-fold below the value for the ICAT reagent-labeled peptide that was selected (Table I, rule 22).

Differential ChemScore values have also been applied to the most common immonium ions. The immonium ion for His at 110 is usually the most reliable immonium ion, thus it was assigned a ChemScore of 100. This value is too low to have much impact on scoring, but these
TABLE I
MS/MS ChemScore rules

There is one rule per row, numbered by the No. column. Rules 1–8 are initial value rules. The ChemScore of the fragment in question is first defined by rules 1–4, which are mutually exclusive. Thus, His is not treated differently than other aa. Rules 5–8 are applied whenever the sequence of the fragment contains the condition described. Thus, if the fragment’s sequence ends in Pro, rule 6 applies. If the fragment begins with Pro, but the preceding residue is Asp, then both rules 5 and 7 apply, and the ChemScore gets multiplied by 4.5. Rules 9–13 adjust the ChemScore according to the ion type, by multiplying the ChemScore calculated by rules 1–8 by the factor listed in column Value. The SeqString, which is calculated separately, is based on a separate score for each peptide bond in the peptide. It is calculated by adding the score in column SeqString for each ion type, with a maximum value of 9. The SeqString is based on rules 9–13 only. Rules 14–21 are rules that allow the ChemScores of ion types to be adjusted in a sequence-specific fashion. Rule 14 applies only to y-17 or b-17 ions in which the first residue is Glu. These ions are often very prominent, and rule 14 sets the ChemScore for such ions equal to the corresponding y or b ion. Rule 15 applies to any y-17 or b-17 that contains a Lys, His, or Glu except when rule 14 applies. Rules 16 and 17 are analogous to rules 14 and 15, but apply to CICAT in peptides. Rule 18 causes a y-18 or b-18 ion to be calculated only if the N-terminal residue of the fragment is Glu. Rules 19 and 20 adjust the ChemScores for the a2 and b2 ions only. Rule 21 applies to a special +18 amu form of the penultimate b ion, which is especially prominent if there is an internal Arg in the fragment. Rule 22 should apply only to MALDI data (but has little effect on QStar data, where it need not apply). Because in MALDI MS/MS spectra there is often detectable transmission of the unselected member of an HL pair, both heavy and light forms of all y and b ions containing CICAT or KICAT are calculated. The unselected HL form gets penalized 10-fold. In rules 23 and 24, peptides containing MetOx or oxidized CICAT often have prominent neutral losses. Therefore, all y and b ions are duplicated for each fragment that contains m, 7 or 8 (see Table IV A). These ions get assigned one-half the value of the corresponding y or b ions. Because Mascot does not consider these ions, it has difficulty distinguishing between sequences, which is often desirable when there are overlapping precursors. The His immonium ion at 110 amu is by far the most reliable. Rule 41 sets PpmMin for calculating FrT and MatSc. Rule 42 sets the highest allowed ppm error for matching a fragment ion. Rule 43 sets the lower mass limit for calculation of ppw. Rule 44 sets the highest allowed ppm error for matching a fragment ion with mass < 200 amu (set by rule 49).

| No. | Rule | Value | SeqString |
|-----|------|-------|-----------|
| 1   | Default | 1     |           |
| 2   | H     | 1     |           |
| 3   | K     | 1.5   |           |
| 4   | R     | 1.5   |           |
| 5   | Before P | 1.5 |           |
| 6   | After P | 0.667 |           |
| 7   | After D | 3     |           |
| 8   | After E | 1.5   |           |
| 9   | y     | 6000  | 5         |
| 10  | b     | 5000  | 4         |
| 11  | a     | 500   | 2         |
| 12  | y-17  | 60    | 1         |
| 13  | b-17  | 50    | 1         |
| 14  | N q-17 | Set chemscore = b or y for N-terminal Gln in b-17 or y-17 |
| 15  | rqh-17 | Multiply by 20 if ROH in seq (except N-terminal Q) |
| 16  | N c-17 | Set chemscore = b/2 or y/2 for N-terminal Cys in b-17 or y-17 |
| 17  | c-17  | Multiply by 20 if ICAT C in seq (except N-terminal Q) |
| 18  | N e-18 | Multiply by 10 if N-terminal Glu in –18 ion |
| 19  | b2    | Multiply by 1.5 |
| 20  | a2    | Multiply by 6.6 |
| 21  | b(N-1)+water | 500 unless internal Arg, then 5000 |
| 22  | Cross c0c9 | Divide by 10 |
| 23  | MetOx | Set chemscore = b/2 or y/2 (–64 amu) |
| 24  | CICAT_Ox | Set chemscore = b/2 or y/2 (69 amu residual mass) |
| 25  | S60   | 1     |           |
| 26  | R70   | 20    |           |
| 27  | P70   | 20    |           |
| 28  | V72   | 10    |           |
| 29  | T74   | 1     |           |
| 30  | L186  | 60    |           |
| 31  | D88   | 1     |           |
| 32  | Q101  | 1     |           |
| 33  | E102  | 1     |           |
small values would break ties between peptides that otherwise seemed to be equally valuable. If the starting peptide samples were less complex, then the reliability of immonium ions for identification could be increased, but even small amounts of His-containing peptides appear to render the His 110 ion detectable (Table I, rules 25–40).

It has also been observed that when methionine sulfoxide is present, a second series of ions is observed that is 64 amu below the canonical y and b ions (Table I, rule 23). A similar neutral loss from the oxidized ICAT side-chain has also been observed (Table I, rule 24).

**Calculation of Percent ChemScore Matched (PCM)**—All of the masses in the filtered peak list that match to predicted ions contribute to PCM. The total ChemScore for the sequence in question is the sum of the Fragment Ion ChemScores of all of the considered ions. The ChemScore Matched is calculated by summing the Fragment Ion ChemScores for those ions that were matched to masses in the filtered peak list within tolerance (Table I, rule 42). Thus, PCM is calculated by:

$$\text{PCM} = 100 \times \frac{\sum_{i=m}^{MFM} \text{ChemScore}(n)}{\sum_{i=n}^{MFM} \text{ChemScore}(n)}$$

where \(t_i\) is the total number of ions considered, MFM is the number of ions matched, \(m\) is the set of matched ions, and \(n\) is the set of all ions. If two theoretical ions matched within tolerance to the same mass, then the Fragment Ion ChemScore for both ions was used. Because of the quantitative nature of the Fragment ChemScore index, PCM is not significantly affected by consideration of additional ions, so long as these additional ions get assigned low ChemScores. In contrast, consideration of large numbers of additional ions can by itself destroy the usefulness of the PIM term, because almost any mass can be explained by some ion.

**Calculation of Peptide Fragment TriScore (FrT)**—It is to be expected that the ions with the highest ChemScore (ChS) should correspond to the most intense ions. In addition, the observed masses are more credibly identified if they match the calculated fragment ion mass within experimental accuracy. To make this quantitative, we define PpmMin as a lower limit of mass error in parts per million (ppm) (Table I rule 41). Ions that match to a tolerance below PpmMin get increasingly less additional credit for doing so. PpmMin is in principle an instrument-specific factor, although in this study a value of 400 ppm was used for both instruments. The upper limits for matching were set to 500 ppm for ions with >200 amu (Table I, rule 42 and 43) and to 0.5 amu for ions <200 amu (Table I, rule 44). To prevent a small number of ions from dominating FrT, it was arbitrarily decided to truncate ChS and intensity (Int) parameters to the fourth highest value (Table II, rules 8 and 9). To calculate FrT, the theoretical ion list is sorted by decreasing ChS, and the peak list is sorted by decreasing Int. To normalize the value of FrT to the intensity distribution observed, the MaximumFrT was calculated as follows:

$$\text{MaximumFrT} = 4 \times \text{ChS}(4) \times \text{Int}(4) + \sum_{i=5}^{n} \text{ChS}(n) \times \text{Int}(n)$$

where \(i\) is the number of elements in the shorter of the two lists. If \(i < 5\), then MaximumFrT is calculated using:

$$\text{MaximumFrT} = i \times \text{ChS}(i) \times \text{Int}(i)$$

The MatchedFrT was then calculated for each matched fragment according to:

$$\text{MatchedFrT} = 4 \times \text{ChS}(4) \times \text{Int}(4) + \sum_{i=5}^{\text{ThIM}} \text{ChS}(n) \times \text{Int}(n) \times \text{PpmMin} \times \frac{\text{ThIM}}{(\text{PpmMin} + \text{MeasuredPpm})}$$

where ThIM is the number of matched ions, and the list is sorted by decreasing ChS × Int. As with MaximumFrT, the equation for MatchedFrT may need to be reduced to suit the number of elements. Finally, FrT was calculated according to:

$$\text{FrT} = 100 \times \frac{\text{MatchedFrT}}{\text{MaximumFrT}}$$

so that the peptide would receive a value of 100 if the intensity distribution of ions exactly matched the theoretical distribution postulated above, and with a mass accuracy of \(<\approx\text{PpmMin}\). Significantly lower scores result whenever the intensity distribution of ions does not conform to expectations. Note that FrT can still have a high value (near 100) if a small number of ions are detected, so long as these ions are predicted to be the most intense. This is one reason why no match is considered plausible unless the sum of b and y ions matched exceeds 3.

The FrT term acts to buffer against devaluation of the PIM term by random matching of intense masses to minor ion types, because when this takes place, although PIM increases, FrT decreases significantly.

**Internal Calibration of MSMS Spectra**—In order to improve the internal accuracy of the fragment ions for the MALDI spectra, for each tentative identification, masses were selected to be used to calibrate the remaining fragments. To accomplish this, two masses corresponding to y or b ions were sought that were as intense as possible.
and also well separated so that a slope measurement derived from them would be as accurate as possible. The rules to find appropriate masses are listed in Table II (rules 1–5). First, a low-molecular-mass fragment was sought that had a mass greater than 0.2 times the mass of the parent ion, and no greater than 0.4 times the mass of the parent ion. The second mass was then selected from the fragments under 0.6 times the parent mass using the rules in Table II. A two-point calibration was then performed. If no appropriate masses could be found, then no calibration was performed.

**Intensity Weighted Mass Error (ppw)**—So that matches to low-intensity ions do not distort the mass error term, intensity-weighting was performed. Because the immonium ion region was often poorly calibrated after this procedure, no fragments less than a value of 200 amu were allowed to contribute to ppw.

**Calculation of Overall Score (MatSc)**—The overall score (MatSc) is a compound index that includes contributions from many of the parameters that can be used to judge the quality of an identification. It is not yet clear how these parameters should be optimally combined, as each of the individual parameters have limitations under certain conditions (like if the peak list is too large). Sophisticated mathematical techniques have been used by others to optimize the weighting of such parameters (17).

In this article, MatSc is calculated according to:

\[
\text{MatSc} = \frac{(\text{PCM} \times \text{PIM} \times \text{FrT} \times \text{PpmMin})}{(\text{Ppw} + \text{PpmMin})}
\]

where PpmMin is the minimum ppm value below which matches are of no greater significance. In these calculations, a PpmMin of 400 ppm was used. This seems rather high, but there were a few identifications that by all other criteria appeared to be correct for which this value was appropriate. In most instances, MatSc would have higher discriminating power if PpmMin were around 50 ppm. Note that the PpmMin term can break ties between peptides that are identical except for Lys versus Gln (0.036 amu difference) even when PpmMin is 400 ppm.

**Calculation of SeqString**—The SeqString describes how the ions that match the proposed peptide are dispersed along the length of the sequence of the peptide. It is not used for spectrum classification, but is a useful visual guide. A similar scheme has been used previously by others (David Fenyo, personal communication). Each ion type is assigned a score, as listed in Table I, rules 9–13, column SeqString. Each peptide bond in the sequence is assigned the sum of those scores, with the exception that the sum must not exceed a value of 9. The most meaningful way to interpret the SeqString is to place it in TrueType font directly over the sequence to which it corresponds, shifted by half a space.

For example, the SeqString 9004555929 might correspond to the sequence ACDEFGHILMK. It could be displayed as:

```
9004555929
ACDEFGHILMK
```

This would indicate that both b and y ions were found that support the AC peptide bond, that no ions were found to corroborate the CD and DE bonds, a b ion corroborates the EF bond, etc.

**Counting y and b Ions**—Probably the simplest way of assessing the quality of a database identification is to count the number of y and b ions matched (see Ref. 18 for an alternative way to do this). If this number is small (e.g., ≤3), then the identification is uncertain. If three or four masses, then the identification might be correct. If the peak list is

---

**Table II**

| No. | Rule | Value        |
|-----|------|--------------|
| 1   | Low mass calibrant | Rank < 21 and > parent*.2 and < parent*.4 and (b or y) |
| 2   | High mass calibrant | Rank < 31 and > parent*.8 and (b or y) |
| 3   | Mass limits        | Rank < 51 and > parent*.6 and (b or y) |
| 4   | Truncations        | Rank < 101 and > parent*.6 and (b or y) |
| 5   | Counting y and b ions | Rank < 201 and > parent*.6 and (b or y) |
| 6   | % Intensity and ppw LMW | 200 |
| 7   | % Intensity and ppw HMW | Parent-50 |
| 8   | Truncate Intensity to 4th highest |
| 9   | Truncate FrT to the 4th highest FrT |
| 10  | Don’t count y max |
| 11  | Don’t count neutral losses or q-17 |
| 12  | If CICAT, count only correct HL form |

**Probable Proteomics 3.7**
In this report, four or more y ions were matched (if any). This happens if the peptide is deamidated, or if CalMW was incorrectly assigned to the second isotope of the isotope cluster, rather than to the monoisotopic mass. This is more likely to happen when MBC/MB corresponds to the integer itself. MB is especially useful for comparing the mass bin MB of the parent ion, which is obtained by dividing the parent ion mass by 1.0005, and then rounding to the nearest integer. MB is especially useful for comparing the results of multiple experiments, or adjacent ion exchange fractions, and is particularly useful as an index. Because a large number of identifications were made to peptides where CalMW-PepMW was nearly equal to small integers, another pair of fields were calculated to classify these spectra. The mass bin class field MBC corresponds to the integer itself. MBC designates electrospray and M designates MALDI; experiment number Ex; elution time (electrospray) or well number (MALDI), designated TiW; the ion exchange fraction number Fr; and a plate number (MALDI) or replicate number (electrospray) Pla. An experiment consists of all data from the same initial ICAT reagent labeling reaction and is instrument-specific. In some cases, replicate electrospray HPLC runs were performed on the same IEX fraction, resulting in the need for Pla. Because in some cases one HPLC run was collected on more than one plate, there is also an HPLC run index EID. Whenever the same spectrum was submitted to more than one database search, MatSc_Calc was used to resolve sequence discrepancies, using SpID as a key field.

**Parent Ions** — The parent ion fields include the recalibrated parent ion mass CalMW; its intensity IntP; the mass of the corresponding peptide sequence PepMW; and the difference between these two masses in ppm Df. A fifth field defines the mass bin MB of the parent ion, which is obtained by dividing the parent ion mass by 1.0005, and then rounding to the nearest integer. MB is especially useful for comparing the results of multiple experiments, or adjacent ion exchange fractions, and is particularly useful as an index. Because a large number of identifications were made to peptides where CalMW-PepMW was nearly equal to small integers, another pair of fields were calculated to classify these spectra. The mass bin class field MBC corresponds to the integer itself. When MBC = 1, CalMW is about 1 amu larger than PepMW. This happens if the peptide is deamidated, or if CalMW was incorrectly assigned to the second isotope of the isotope cluster, rather than to the monoisotopic mass. This is more useful, and it must be balanced with a term like % Intensity Matched. In this report, four or more y ions were required for all classifications of identified spectra. Table II, rules 10–12 list several additional considerations that apply to counting ions.

**CRUDE RESULTS**

**List of Spectra with Identifications** — All 73,009 spectra from experiments 2–6 are listed in Supplemental Table S3, with fields that define in what fraction they were obtained, the intensity, the HL ratio, and various parameters that define the confidence of the identification. Because Table S3 (as well as Tables S6–10, S12, and S13) is so large, a subset of this table is shown to exemplify the information it contains. Table III shows a subset of Table S3, which corresponds to seven spectra. The first three of these spectra correspond to Upf1p, the next two are from His3p, and the final two are from pyruvate decarboxylase. Table IIIID lists the fields in Table III, many of which are useful for filtering and processing these data. Because in many cases multiple rounds of database searches were performed, the Mascot scores are not always exactly comparable to one another, even though this score is based mainly on the number of ions that match. One reason for this is that the same mass tolerances were not always used for each search. In the discussion below, the significance of the Mascot score will be addressed. Regardless of how the tentative sequence was identified, it is possible to tabulate how well the spectrum corresponds to the proposed sequence, for example, by using MatSc. In Table IIIID, the column definitions for Table III are divided into five categories: those involved in tracking which HPLC run the spectrum belongs to, those that characterize the parent ion that was fragmented, those (labeled MS/MS) that describe how well the proposed peptide sequence corresponds to the spectrum, those that characterize the HL ratio, and those that characterize the peptide and protein sequence that was matched (if any).

Table S6 contains each of the peak lists that were automatically extracted for each of the 73,009 spectra in Table S3. This is an enormous table, and three examples of peak lists corresponding to two distinct spectra (one of which was obtained manually from the peak list in Table S6) are shown in Table VI.

**Tracking** — The tracking fields include such parameters as spectrum number SpID; instrument type EM, where E designates electrospray and M designates MALDI; experiment number Ex; elution time (electrospray) or well number (MALDI), designated TiW; the ion exchange fraction number Fr; and a plate number (MALDI) or replicate number (electrospray) Pla. A further set of fields are included in the tracking observations and manual modifications. The comment field is not displayed, but is used for tracking observations and manual modifications.

**Parent ions**

### Table IIIA

**Spectra**

The meanings of the column headings are described in “Crude Results.” SpID was derived from the MS/MS ID number from ProICAT, with values of 100,000, 600,000, or 700,000 added to prevent overlaps. SpID was derived from the Peak List ID from GPS. Pla was derived from the Plate ID from GPS, and from the run id from ProICAT. Many of the QStar spectra were not processed so as to extract the parent ion intensity, and therefore it is blank. HL is listed only when measured. Fields HL_S, HLI_S, HL_1, and HLI_1 are defined for MALDI spectra only. AutoSeq was calculated only in rare instances. ISD has meaning only when ChM = 10 or 12, or NT = 0 or 1, because only in these circumstances is it likely that the parent mass may have arisen from in-source decay. The comment field is not displayed, but is used for tracking observations and manual modifications.

| No. | SpID | EM | Ex | Fr | Pla | TiW | EID | CalMW | PepMW | Df | MAc | IntP | MB | MBC | DfC | Cal |
|-----|------|----|----|----|-----|-----|-----|-------|-------|----|-----|-----|----|-----|-----|-----|
| 1   | 602450 | 2  | 2  | 12 | 3   | 43.23 | 11  | 1948.08 | 1948.07 | 4   | 1   | 1337 | 1947 | 0   | 4   |     |
| 2   | 602448 | 2  | 2  | 12 | 3   | 43.14 | 11  | 1949.10 | 1948.07 | 529 | 1   | 1948 | 1715 | 0   | 12  |     |
| 3   | 140052 | 2  | 4  | 8  | 45  | 32.39 | 51  | 1948.09 | 1948.07 | 9   | 1   | 1947 | 0  | 9   |     |     |
| 4   | 52964  | 1  | 5  | 15 | 799 | 77.00 | 91  | 1305.73 | 1305.74 | -4  | 1   | 2865 | 1305 | 0   | 4   | 0    |
| 5   | 604059 | 2  | 2  | 15 | 6   | 34.67 | 14  | 1305.74 | 1305.74 | 2   | 1   | 1000 | 1305 | 0   | 2   |     |
| 6   | 48200  | 1  | 6  | 15 | 859 | 62.00 | 176 | 1948.08 | 1948.07 | 1733 | 91  | 1733 | 0  | 1733 | 0   | 7    |
| 7   | 612420 | 2  | 2  | 9  | 23  | 36.42 | 8   | 1948.08 | 1948.07 | 4   | 1   | 1337 | 1947 | 0   | 4   |     |

The tracking fields include such parameters as spectrum number SpID; instrument type EM, where E designates electrospray and M designates MALDI; experiment number Ex; elution time (electrospray) or well number (MALDI), designated TiW; the ion exchange fraction number Fr; and a plate number (MALDI) or replicate number (electrospray) Pla. The tracking fields include such parameters as spectrum number SpID; instrument type EM, where E designates electrospray and M designates MALDI; experiment number Ex; elution time (electrospray) or well number (MALDI), designated TiW; the ion exchange fraction number Fr; and a plate number (MALDI) or replicate number (electrospray) Pla. Whenever the same spectrum was submitted to more than one database search, MatSc_Calc was used to resolve sequence discrepancies, using SpID as a key field.

**Parent Ions** — The parent ion fields include the recalibrated parent ion mass CalMW; its intensity IntP; the mass of the corresponding peptide sequence PepMW; and the difference between these two masses in ppm Df. A fifth field defines the mass bin MB of the parent ion, which is obtained by dividing the parent ion mass by 1.0005, and then rounding to the nearest integer. MB is especially useful for comparing the results of multiple experiments, or adjacent ion exchange fractions, and is particularly useful as an index. Because a large number of identifications were made to peptides where CalMW-PepMW was nearly equal to small integers, another pair of fields were calculated to classify these spectra. The mass bin class field MBC corresponds to the integer itself. When MBC = 1, CalMW is about 1 amu larger than PepMW. This happens if the peptide is deamidated, or if CalMW was incorrectly assigned to the second isotope of the isotope cluster, rather than to the monoisotopic mass. This is more useful, and it must be balanced with a term like % Intensity Matched. In this report, four or more y ions were required for all classifications of identified spectra. Table II, rules 10–12 list several additional considerations that apply to counting ions.
AutoSeq was calculated only in rare instances. ISD has meaning only when ChM intensity, and therefore it is blank. HL is listed only when measured. Fields HL_S, HLI_S, HL_1, and HLI_1 are defined for MALDI spectra only.

the Plate ID from GPS, and from the run id from ProICAT. Many of the QStar spectra were not processed so as to extract the parent ion values of 100,000, 600,000, or 700,000 added to prevent overlaps. SpID was derived from the Peak List ID from GPS. Pla was derived from

likely to happen at higher peptide molecular masses where the monoisotopic mass is harder to distinguish. Because de-amidation is more interesting than a mistake in de-isotoping, the mass difference in ppm DfC was calculated, assuming an ideal mass difference of 0.984 amu, which is the mass difference between an acid and an amide. If the explanation is amidation is more interesting than a mistake in de-isotoping, the monoisotopic mass is harder to distinguish. Because de-amidation probably used as a calibrant. In a few cases, the ppm difference rounded to 0 to four significant figures even when the mass in question was not used as a calibrant.

MS/MS Ions—The MS/MS fields include the Mascot score Sc; the overall match score MatSc, defined above; the number of y and b ions matched yb; SeqString defined above; and the four major components of MatSc: namely, Percent ChemScore Matched PCM, Percent Intensity Matched PIM, intensity-weighted average ppm deviation ppw, and Fragment TriScore FrT. The number of masses detected prior to peak filtering MFD, after peak filtering MFU, the number of masses matched MFM, and the number of theoretical ions matched ThIM are also listed. Note that it is possible for more than one theoretical ion to match the same mass, and vice versa. IntD lists the sum of the intensity of the ions counted in MFU. Note that all of these scores are dependent on peak detection. In an ideal world, it would be possible to detect peaks reliably and

THE MEANINGS OF THE COLUMN HEADINGS ARE DESCRIBED IN “CRUDE RESULTS.” SPECID WAS DERIVED FROM THE MS/MS ID NUMBER FROM PROICAT, WITH VALUES OF 100,000, 600,000, OR 700,000 ADDED TO PREVENT OVERLAPS. SPECID WAS DERIVED FROM THE PEAK LIST ID FROM GPS. PLA WAS DERIVED FROM THE PLATE ID FROM GPS, AND FROM THE RUN ID FROM PROICAT. MANY OF THE QSTAR SPECTRA WERE NOT PROCESSED SO AS TO EXTRACT THE PARENT ION VALUES OF 100,000, 600,000, OR 700,000 ADDED TO PREVENT OVERLAPS. SpID WAS DERIVED FROM THE PEAK LIST ID FROM GPS. PLA WAS DERIVED FROM THE PLATE ID FROM GPS, AND FROM THE RUN ID FROM PROICAT. MANY OF THE QSTAR SPECTRA WERE NOT PROCESSED SO AS TO EXTRACT THE PARENT ION INTENSITY, AND THEREFORE IT IS BLANK. HL IS LISTED ONLY WHEN MEASURED. Fields HL_S, HLI_S, HL_1, AND HLI_1 ARE DEFINED FOR MALDI SPECTRA ONLY.
TABLE III
Fields for spectrum table

The meanings of these columns are described in “Crude Results.”

| No. | Abbr. | Description |
|-----|-------|-------------|
| 1   | SpID | Spectrum ID number |
| 2   | EM   | 1 for MALDI, 2 for electrospray |
| 3   | Ex   | Experiment number |
| 4   | Fr   | Ion exchange fraction number |
| 5   | Pla  | Plate number (MALDI) or replicate number (electrospray) |
| 6   | TW   | Time in minutes (QStar), or well number (MALDI) |
| 7   | EID  | A unique number for each HPLC run |
| 8   | CalMW| Recalibrated parental MW |
| 9   | PepMW| MW deduced from matched peptide sequence |
| 10  | Df   | ppm difference between CalMW and PepMW |
| 11  | MAC  | Mass accuracy class; depends on MBC and DIC |
| 12  | IntP | Intensity of parent mass |
| 13  | MB   | Mass bin; =round (CalMW/1.0005) |
| 14  | MBC  | Mass bin class |
| 15  | DIC  | ppm difference between CalMW and (PepMW-MBC) |
| 16  | Cal  | A binary field; 1 indicates the mass or its HL partner was used as a calibrant |
| 17  | Y1   | Specific categories of peptides (combination of ChM, MAC, HL; see Table IVB) |
| 18  | Y2   | General categories of peptides, mostly based on MBC and ChM, ranges from 0–5 |
| 19  | Sc   | Mascot Score |
| 20  | MatSc| MatchScore; based on pim, pcm, ppw, ft |
| 21  | yb   | Number of matching y and b ions |
| 22  | SeqString | String of length len(seq)−1, 5 for y, 4 for b |
| 23  | PIM  | Percent Intensity Matched |
| 24  | PCM  | Percent ChemScore Matched |
| 25  | FrT  | Fragment TriScore Calculated |
| 26  | Ppw  | Intensity weighted average ppm of daughter ions |
| 27  | MFD  | Number of daughter ions detected |
| 28  | MFU  | Number of daughter ions used in database searching and in calculations |
| 29  | MFM  | Number daughter ions matched |
| 30  | ThM  | Number theoretical ions matched |
| 31  | IntD | Sum of daughter ion intensities |
| 32  | HL   | ICAT ratio, heavy to light |
| 33  | Ippm | ppm difference between expected ICAT pair and observed masses |
| 34  | HL_S | 6-digit binary string, P1 = 9, 18, 27, −9, −18, −27 |
| 35  | HLI_S| 6-digit binary string, P1 = 9, 18, 27, −9, −18, −27 interferent within 2 amu |
| 36  | HL_1 | Binary field; 1 indicates no interferent from HL_S |
| 37  | HLI_1| Binary field; 1 indicates no interferent from HLI_S |
| 38  | Seq  | Sequence matched. See single aa dictionary in Table IVA |
| 39  | ChM  | Chemical Modification Class (see Table IVA) |
| 40  | <    | aa prior to DSeq |
| 41  | DSeq | Database sequence that corresponds to Seq |
| 42  | >    | aa after DSeq |
| 43  | Acc  | Accession number deduced from Seq |
| 44  | HL_Type | HL type (c0 Vs c9 Vs c0c0 etc.) |
| 45  | CysN | Number of cysteines |
| 46  | AutoSeq | Automatic sequence (as determined by de novo sequencing, or manual) |
| 47  | NT   | Number of tryptic cleavage sites; 2 = fully tryptic, 0 = nontryptic |
| 48  | IT   | Internal trypsin sites; the number of missed cleavage sites |
| 49  | ISD  | In-source decay; 1 indicates fully tryptic donor peptide co-migrates by RP-HPLC |

In these experiments, the peak lists for the MALDI data are usually reliable; that is, upon manual inspection, the peaks deposited in the database appear to reflect faithfully what one would expect from careful inspection of the raw spectra. For the electrospray data, however, it is difficult to define peak detection, de-isotoping, and de-charging pa-
rameters that result in a good peak list for all spectra. For this reason, we spent a lot of time validating electrospray identifications, as is regular practice. We hope to develop more robust peak extraction methods in the future.

**HL Fields**—The HL fields include the measured heavy-to-light ratio HL. In some cases, the ratio was not measurable because no HL partner was detected. HL_S summarizes how many possible HL partners were detected. HL_S has six digits, each of which can be either 0 or 1. If the first position = 1, then a potential HL partner was detected about 9.03 amu above the parent mass. The remaining five positions of HL_S correspond to the presence or absence of a peak 18, 27, -9, -18, and -27 amu above or below the parent mass in question. Under ideal circumstances, the HL pair would be nonambiguous, meaning there would be only one non-zero digit in HL_S, which would correspond to the number of CICAT residues in the peptide. In this case, the binary field HL_1 is set to 1. We have included as “identified” (Y1 class 1, see below) only those peptides in which HL_S is consistent with the number of Cys residues in the corresponding sequence. A second parameter, Ipptom, is the ppm difference between the observed masses of the HL pair and the theoretical mass difference of the HL pair (not shown in Table III). When the HL pair was ambiguous, the value for Ipptom listed corresponds to the sequence. A final HL parameter is the HLI_1. It is similar to HL_S, but lists instead whether there are any interfering masses within 2 amu of an HL pair at +9, +18, +27, -9, -18, or -18. When HL_S is nonambiguous, and HLI_1 is all zeroes, then the ratio of the HL pair is more reliable (depending on the intensity of the peaks), and the binary field HLI_1 is set to 1. When HL_S is ambiguous, the exact value of the HL ratio may be unknowable, because more than one peptide may contribute to the intensity of either the c0 form of the ICAT reagent or the c9 form of the ICAT reagent. This problem is lessened if it turns out that the ambiguous ICAT reagent-labeled mass belongs instead to a nonoverlapping HL pair. For example, if HL_S was 111,000 and the corresponding peptide had one Cys, the potentially confounding masses could correspond to a second unrelated HL pair whose masses were 18 and 27 amu above the first peptide’s mass. Because there are slight differences in HL ratio between experiments due to mixing inaccuracies, the HL values must be normalized. In the five experiments listed here, these normalization constants were small (between 0.855 and 1.11, see Table V, column Nor).

**Peptide and Protein Fields**—These fields indicate which peptide sequence Seq was identified, and which protein was matched to that sequence, which is designated by (usually) a Swiss-Prot accession number Acc. The amino acid preceding the peptide is listed in < (SeqA in Access), whereas the amino acid following the peptide is listed in > (SeqZ in Access). If field “<” is empty, then the peptide was N terminal; similarly for field “>” and C-terminal peptides. When modified amino acids were detected, the normal capital letter abbreviation for the amino acid is altered in field Seq, whereas DSeq corresponds exactly to the sequence in the database (not shown in Table III). Table IVA lists all of the single-letter amino acid codes for altered amino acids in column Sym, where column RMW lists the molecular mass difference between the modified aa and the natural aa, or in the case of modified CICAT peptides the molecular mass difference between the modified and unmodified form of the CICAT residue. For example, “C” refers to the c0 form of the ICAT reagent, whereas “c” refers to the c9 form. “Z” refers to unmodified Cys. If an Asn or Gln has been deamidated, it is labeled as the corresponding acid residue in Seq, but not in DSeq. In some cases, the accession numbers are PIR accession numbers from csc-fserve.hh.med.ic.ac.uk/delphos.html (referred to hereafter as DelPhos) because there was no corresponding protein in Swiss-Prot. Table IVA also lists how many spectra are classified into each grouping, either at high confidence (field HiMac), at high confidence but also considering different MBC bins (field High), or at any confidence (field Low).

The HL_Type field lists whether the peptide was modified with the c0 or the c9 form of the ICAT reagent, and how many such modifications there are. The CysN field lists the number of modified Cys residues in the peptide. The AutoSeq field lists the best sequence obtained by either manual de novo sequencing or automatic de novo sequencing, if that was determined (not shown in Table III). NT lists the number of tryptic cleavage sites; a value of 2 indicates that both ends correspond to trypsin cleavage rules. In these tables, all terminal Arg and Lys residues are defined to be acceptable tryptic termini, including sequences in which the next aa is Pro. In addition, peptides that include the protein’s N or C terminus are counted as tryptic. Finally, peptides whose N terminus is consistent with annotated biological cleavages are also counted as tryptic. This includes removal of an N-terminal Met or cleavage by signal peptidase. We did not happen upon any instances of known biologically relevant C-terminal cleavages. Some peptides contain missed trypsin cleavage sites; these are counted in field IT. In field IT, KP or RP sequences are not counted as internal missed cleavage sites. In some cases, a peptide may be generated from a longer peptide by fragmentation in the source region of the mass spectrometer. This is especially likely for certain chemically modified peptides, especially peptides containing “Z” for unmodified Cys, which commonly elute together with a corresponding Cys-alkylated peptide. If a suitable precursor for in-source decay was detected in the same MALDI spot or within 1 min of elution time, then this is noted in field ISD. Ideally, in-source decay in electrospray requirements should also require exact co-migration of the ISD ion and its precursor.

**Overall Fields**—Two final parameters, Y1 and Y2, group spectra into classes of overall reliability. The most important class is Y1 class 1, which corresponds to the highest confidence category of CICAT peptides. Table IVB lists each of these classes. Note that most of the classifications of spectra
according to fields Y1 and Y2 depend on the classifications in fields ChM, MAc, NT, HL_S, HL, yb, MatSc, or Sc. Y1 classes 8 and 9 are special exceptions that correspond to Cys-modified tryptic peptides that derive from trypsin or human keratin, respectively, and therefore do not contribute to any of the yeast peptide or protein statistics.

Column No. lists how many spectra are in each Y1 class. Field Y2 groups spectra according to how they are modified. Y2 class 1 refers to CICAT peptides, whereas Y2 class 3 correspond to Lys-modified (KICAT) peptides. Y2 class 2 refers to peptides that are not modified at all, whereas Y2 class 4 have an unalkylated Cys. In most cases, this appears to be a result of ISD. Y2 class 5 are spectra matched to peptides with chemical modifications that correspond to more than one of classes 1–4, whereas Y2 class 0 have fewer than three y or b ions matched and are therefore essentially unmatched. Note, however, that such a spectrum could be matched with high confidence due to co-migration with second spectrum with high Mascot score (Sc) and yb. In addition, in MALDI experiments, such spectra could become identifiable after subsequent MS/MS experimentation.

Data Processing

Low Stringency Identifications—The first task in extracting information from proteomics experiments is to determine which identifications are to be considered reliable. We chose to start with identifications using either ProICAT (for electrospray samples) or Mascot (for MALDI samples). Normally, threshold criteria are chosen that appear to exclude the bulk of the incorrect assignments while retaining the bulk of the correct assignments. In these experiments, we were particu-
larly interested in studying the borderline identifications, and therefore have collected statistics on many identifications that would normally be discarded. We decided upon four minimal requirements for tentative identification: 1) at least four y and b ions must match within 300 ppm; 2) the sequence must be at least 6 aa long; 3) MatSc must exceed 5000; and 4) the tentative sequence must match the spectrum in question better than any other considered sequence. This last requirement requires some judgment regarding which chemical modifications are at all plausible for consideration. Generally a chemical modification is considered reasonable only if the same modification has been found on a peptide that was matched with high confidence to a high-quality spectrum from within the same experiment, or alternatively if the unmodified form of the peptide in question is known to be so abundant that nearly any chemical modification seems plausible. There is a special category of spectra that we excluded from further consideration: they consist of spectra that by the criteria of 

**TABLE IVB**

**Overall classifications**

Y1, A collection of identifications, corresponding to line 17 in Table III. All Y1 classes require MAc < 4, yb > 3 and MatSc ≥ 10,000 except classes 0, 13, 14, 98, and 99. All Y1 classes except 0, 5, 8, and 9 are dependent on Y2 classifications. Y2, A collection of identifications with six general categories, based mostly on ChM without regard to confidence of identification. The Y2 category is listed for each Y1 when that is appropriate. In cases where there is more than one Y2 category for a single Y1 category, Y2 is left blank. Spectra classified as Y1 = 98 or 99 are defined as Y2 = 0. No., the number of identifications in each category. M, the number of MALDI identifications. E, the number of electrospray identifications.

| Y1 Y2 | Y1 categories | No. | M  | E  |
|-------|---------------|-----|----|----|
| 1 0   | Spectrum not classified: if y2 = 1 then yb < 3. Else, MAc > 3 or MatSc < 10,000 | 49736 | 8605 | 41131 |
| 2 1 1 | CICAT peptide; yb ≥ 3, MatSc ≥ 10,000, MAc = 1, MBC = 0, NT = 2; ChM = 2; H1 = 1; 0.2 < Hl < 5; len > 5 | 10888 | 3502 | 7386 |
| 3 2 2 | Unmodified peptide; yb > 3, MatSc ≥ 10,000, NT = 2; ChM < 9 | 12699 | 451 | 818 |
| 4 3 3 | KICAT peptide; yb > 3, MatSc ≥ 10000, NT = 2; ChM = 9 | 376 | 41 | 335 |
| 5 4 4 | Contains Z in Seq; yb > 3, MatSc ≥ 10,000, NT = 2; ChM = 10 | 202 | 132 | 70 |
| 6 5 | NT = 2; ChM = 99 | 323 | 23 | 300 |
| 7 6 1 | ChM = 3 – 7 | 203 | 116 | 87 |
| 8 7 1 | <6 aa long | 90 | 3 | 87 |
| 9 8 | From trypsin | 4 | 2 | 2 |
| 10 9 | From keratin | 9 | 9 | |
| 11 11 | Semitryptic CICAT | 404 | 82 | 322 |
| 12 12 | Semitryptic, not labeled | 875 | 33 | 842 |
| 13 13 | MatSc < 10000; yb > 3, MAc < 4 | 3437 | 1995 | 1442 |
| 14 14 | MatSc < 10000; yb > 3, MAc < 4 | 1011 | 680 | 331 |
| 15 21 | Nontrypic CICAT | 31 | 1 | 30 |
| 16 22 | Unmodified, nontrypic | 164 | | 164 |
| 17 51 | Contains CICAT Ox (7 or 8 in [seq]) | 1003 | 176 | 827 |
| 18 52 | Contains de-amidated CICAT (5 or 6 in [seq]) | 150 | | 150 |
| 19 53 | Contains N-terminal ICAT label (1 or 2 in [seq]) | 31 | 2 | 29 |
| 20 54 | Contains CICAT −17 amu (3 or 4 in [seq]) | 84 | 6 | 78 |
| 21 61 | Sodiated; may have [Df] corresponding to 22 amu for MALDI | 88 | 24 | 64 |
| 22 71 | Contains CICAT, [CalMW] − [PepMW] = −1 | 170 | 95 | 75 |
| 23 72 | Contains CICAT, [CalMW] = −2 | 131 | 68 | 63 |
| 24 73 | Contains CICAT, [CalMW] = −3 | 8 | 8 | |
| 25 75 | Contains CICAT, [Df] > 20 (MALDI), or [Df] > 80 (electrospray); [Df] < 300 | 515 | 292 | 223 |
| 26 76 | Contains CICAT, [CalMW] − [PepMW] < −3 or [CalMW] − [PepMW] > 3 | 0 | | |
| 27 81 | Contains CICAT, [CalMW] − [PepMW] = +1 | 1159 | 237 | 922 |
| 28 82 | Contains CICAT, [CalMW] − [PepMW] = +2 | 59 | 31 | 28 |
| 29 83 | Contains CICAT, [CalMW] − [PepMW] = +3 | 22 | 21 | 1 |
| 30 97 | By inspection clearly a viable explanation: protect against being overwritten | 0 | 0 | 0 |
| 31 98 | [Seq] not in yeast database, not trypsin or keratin or other plausible contaminant | 480 | 130 | 350 |
| 32 99 | By inspection clearly not a viable explanation | 87 | 87 | |
| 33 | Total | 73009 | 16765 | 56244 |

| Y1 Y2 | Y2 categories | No. | M  | E  |
|-------|---------------|-----|----|----|
| 34 0 | Unmatched | 48982 | 8092 | 40890 |
| 35 1 | Matched to a sequence containing CICAT, >3 yb, MatSc ≥ 1000, or manual | 18909 | 6827 | 12082 |
| 36 2 | Matched to a sequence with no ICAT modification, >3 yb, MatSc ≥ 1000 | 3522 | 1218 | 2304 |
| 37 3 | Matched to a sequence containing KICAT, >3 yb, MatSc ≥ 1000 | 523 | 158 | 365 |
| 38 4 | Matched to a sequence with at least one Cys that is not CICAT, >3 yb, MatSc ≥ 1000 | 517 | 355 | 162 |
| 39 5 | Contains mixed modifications (X or Z and another, unless confirmed), >3 yb, MatSc ≥ 1000 | 556 | 115 | 441 |
| 40 | Total | 73009 | 16765 | 56244 |
MatSc and yb seem to match a certain sequence, yet manual inspection of the spectrum indicates that the identification is not credible (Y1 class 99). In some cases, this results because the spectrum is so weak that the process of extracting a peak list failed. In other cases, examination of the spectrum indicates that the substance that was fragmented does not correspond to a peptide at all and is probably derived from contamination. In a few cases, the spectrum seems to correspond to a peptide, but has intense ions that cannot be explained easily by the proposed sequence. This process has been selectively applied to the data, with special attention paid to spectra that uniquely define proteins. Therefore, there are surely still examples of spectra that are matched to peptide sequences that can be shown by manual examination to correspond to other sequences, or substances other than peptides. Next, we classified spectra as 1) ICAT reagent modified on Cys (CICAT), 2) not ICAT reagent labeled, 3) ICAT reagent modified on Lys (KICAT), or 4) incompletely alkylated. These classifications correspond to Y2 classes 1–4.

Using the criteria for tentative identification described above, there are 12,249 identifications in Y2 class 1, corresponding to 2181 distinct CICAT peptides, or 1029 distinct proteins, some of which are surely misidentifications (Fig. 2A).

**Protein Identification and Consolidation**—The ICAT Dictionary—Whenever more than one protein encodes the same peptide, one needs to determine if possible which protein is the most likely source protein. This can sometimes be accomplished if there are additional MS/MS identifications pointing to one protein but not the alternatives, or if there is independent evidence that one protein is more likely to be encountered. In any case, the data need to be annotated with a consistent protein accession number for tracking purposes. All of these issues can be addressed by making a database of all possible CICAT peptides in the yeast proteome, annotated with protein accession numbers. We refer to this database as an ICAT dictionary. In yeast, another useful tracking field is the codon bias index (3), which correlates positively with protein abundance (3), and therefore is useful for tie-breaking purposes.

It is not possible to derive the optimized list of proteins until all of the relevant data are consolidated together (see Ref. 19 for a discussion of this issue). In this report, the relevant data include identifications from all five experiments. Another complication is that some peptides are virtually identical by mass spectrometry (I versus L, and combinations of amino acids that add up to the same mass without distinguishing backbones). It shows that 609 proteins were identified by both techniques. These identifications had yb > 3; MatSc ≥ 5000; len ≥ 8; Y1 <= 9; Y2 = 1; NT = 2; MBC = 0; Mac < 4; CHM 0 or 1. B, High-stringency identifications. The number of identifications decreases significantly. Many of the peptides and proteins that are unique to an instrument type in B were identified with low confidence in A. In addition to the requirements in A, MatSc ≥ 10,000; Sc ≥ 20; Y1 = 1; Mac = 1. C, Same as B, except all spectra also had HL ratios where 0.2 < HL < 5.

**Fig. 2. Degree of overlap of peptides and proteins across instrument types.** Identifications of CICAT peptides, segregated by instrument type. The number in parentheses indicates the number of distinct identifications. There were 7550 measurements of CICAT peptides that were chromatographically distinct. The Venn diagram in the middle of A shows that 990 peptides were identified by both techniques, out of 2181 distinct peptides. On the right, the diagram shows that 609 proteins were identified by both techniques. These identifications had yb > 3; MatSc ≥ 5000; len ≥ 8; Y1 <= 9; Y2 = 1; NT = 2; MBC = 0; Mac < 4; CHM 0 or 1. B, High-stringency identifications. The number of identifications decreases significantly. Many of the peptides and proteins that are unique to an instrument type in B were identified with low confidence in A. In addition to the requirements in A, MatSc ≥ 10,000; Sc ≥ 20; Y1 = 1; Mac = 1. C, Same as B, except all spectra also had HL ratios where 0.2 < HL < 5.

**Molecular & Cellular Proteomics 3.7** 639
bone fragment ions). Because of this, there is a danger that two distinct peptide sequences could get mapped to indistinguishable MS/MS spectra. To derive an optimized protein list, we submit the entire list of identified CICAT peptides to the Mascot search engine in the form of theoretical ion spectra. Mascot then automatically selects the smallest number of distinct protein sequences that can explain the spectra. Later, the sequences are mapped to the other tracking fields like codon bias using the ICAT dictionary.

To generate our theoretical ICAT dictionary, we started with the yeast proteome at genome-ftp.stanford.edu/pub/yeast, updated ~1/2003. In this database, there are 23,308 distinct Cys-containing peptides with no missed trypsin cleavage sites that contain at least 6 aa and have a MW of <4000. Of these, 481 (2%) are encoded by more than one distinct protein. The ICAT dictionary has a degeneracy field that enumerates how many genes encode each peptide. After the list of these sequences was generated, the sequences were grouped using an SQL query so that each unique sequence would be paired with the protein accession number that corresponded to the protein with the highest codon bias, as well as a number that listed how many distinct genes encoded the peptide. In some cases, it was necessary to add new sequences to the ICAT dictionary to accommodate additional sequences that were derived from other databases, for example, the Mascot nonredundant database. The final protein list contains Swiss-Prot accession numbers whenever possible, and PIR-derived “S” numbers from the Mascot nonredundant database when the proteins in question were not in Swiss-Prot. Extensive use was made of the web site cscfserve.hh.med.ic.ac.uk/delphos.html to resolve questions of protein identity. At this web site, one can input a query sequence and rapidly obtain a list of matching proteins from public databases. The codon adaptation indices were downloaded from genome-ftp.stanford.edu/pub/yeast/data_download/protein_info/protein_properties.tab. Efforts were made to match Swiss-Prot accession numbers to the “Orf” field in this table by sequence.

**Identifications for Quantification**—When the goal is to determine which proteins have changed in expression level, it is desirable to weed out all confounding data. To accomplish this, only measurements that corresponded to fully tryptic CICAT modifications were included in further quantification analyses. As before, at least four y and b ions were required for a match, but in addition MatSc >10000 and a Sc of at least 20 was required. In addition, internally consistent mass accuracy was required; for MALDI experiments, the mass accuracy requirement was 20 ppm; for electrospray experiments, the mass accuracy was 80 ppm (MACE = 1). Had we taken greater trouble to internally calibrate the QStar HPLC runs, the electrospray mass accuracy could have been additionally constrained, as there was typically <25 ppm systematic error in the electrospray measurements. Next, measurements were eliminated for which no HL data was extractable (using ProICAT data for QStar, using the parent spectra for MALDI data). At this point, 7441 spectra still qualify (Fig. 2B). Finally, MALDI measurements were excluded if there were two possible HL pairs that could contribute to the HL ratio that was measured that were 9, 18, or 27 amu above or below the HL pair that corresponded to the sequence in question (see HL_S in Table IIIID) or confounding isotope clusters (see HLI_S in Table IIID).

In addition, 24 peptide identifications were eliminated because they had ratios that were either <0.2 or >5. Manual examination of these measurements indicated that they were measured incorrectly.

**Reduction to Unique HL Measurements**—The list now contains 5668 predominantly credible measurements (Fig. 2C), but in many cases these measurements are not independent because in some cases the measurements derive from each member of an HL pair, or from multiple charge states of the same HL pair. The ProICAT software automatically reduces these measurements to unique measurements, but this feature was overridden so that statistics on individual spectra could be collected. To restore uniqueness, electrospray HL measurements were binned by time and peptide sequence by dividing the elution time in minutes by 3, rounding to the nearest integer, and multiplying by 3 to get a time bin integer. In most cases, the HL ratio for such binned measurements was the same as each individual measurement because the HL ratio that was deposited into the database had been generated by the ProICAT software using similar logic. The MALDI measurements were binned by spot and sequence. Each bin was considered a separate measurement.

Upon accomplishing this, the 5668 credible peptide identifications were reduced to 3726 measured HL ratios (Fig. 2C), which are attributable to 1361 distinct peptides and 702 different proteins. A total of 496 of these proteins had two different HL ratio measurements, but only 57 of them had standard deviations of >0.2. Manual examination of a large amount of discrepant data indicates that in most cases the reason for the large standard deviation has to do with overlapping isotope clusters (largely excluded previously) or low signal intensity. A second category of discrepancy may be based on low levels of reversible or isomeric chemical modifications to ICAT reagent-labeled peptides that cause MS/MS-identical peptides to separate by either IEX or reversed-phase chromatography, sometimes with distinguishable ratios. In this case, the most reliable measurement is from the major form of the peptide. A small minority of discrepancies may be due to false identifications or, more interestingly, biological complexity at the level of the proteome, based on multiple protein forms. The first two problems can be eliminated by assigning a single measurement per peptide per experiment based on an intensity-weighted average for that peptide. However, this removes much of the valuable redundancy in the data and should only be performed to test the seriousness of the third problem.

At this point, automated processes are not able to deal with
TABLE V

Spectra vs. experiment

| Ex | EM | No. IEX | No. HPLC | Date   | Spectra | Nor | Stringency ICAT | Stringency others |
|----|----|---------|----------|--------|---------|-----|----------------|------------------|
|    |    |         |          |        |         |     | L H             | N X Z Met Met_Ox % Ox CICAT_Ox |
| 2  | 2  | 20      | 24       | May-02 | 14,083  | 1.11| 3,284 1,611 | 831 506 503 1 2 297 26 8.05 11 |
| 3  | 2  | 7       | 7        | Aug-02 | 2,459   | 0.93| 1,060 480  | 42 6 5 0 0 4 77 95.06 92 |
| 4  | 2  | 23      | 44       | Oct-02 | 39,823  | 1.04| 6,369 2,349 | 1,256 402 195 204 3 57 408 87.74 240 |
| 5  | 1  | 20      | 22       | May-02 | 10,054  | 0.87| 2,060 735  | 471 205 178 0 27 104 19 15.45 8 |
| 6  | 1  | 14      | 14       | Aug-02 | 6,711   | 1.00| 2,460 1,162 | 131 25 5 13 7 25 51 67.11 25 |
| all| 5  | 84      | 111      |        | 73,130  | 1.00| 15,233 6,337  | 2,731 1,144 886 218 39 487 581 376 |

Overall Results—In all, six complete ICAT experiments were performed (Table V). One experiment was performed with a noncleavable form of the ICAT reagent and is not further described because all identified peptides were also identified in the other experiments. The remaining five experiments were performed with the cleavable form of the ICAT reagent. Three complete experiments were performed using electrospray, whereas two complete experiments were performed with
MALDI acquisition. For the electrospray experiments, in some cases, the ion exchange fraction was injected in several identical batches. In all, 84 IEX fractions were processed; because some fractions were analyzed several times, that works out to 111 separate HPLC runs. From these runs, 73,130 MS/MS spectra were collected (Table V, column Spectra).

Table III and many of the other tables are so large, many tables have been prepared in two forms, the supplemental (S) table form, which is available electronically, and the published table form, which consists of selected entries from the larger table. Thus, Table III contains seven entries from three different proteins. Three identifications are shown from Upf1p (ILVCAPSNVAVDHLAAK), which was the most down-regulated protein, as well as two identifications from His3p (ITLHVDcLR), which was the most up-regulated protein. The remaining two peptides are solid identifications of NPVILADAccSR from pyruvate decarboxylase. These identifications were selected to illustrate how tenuous identifications can be annotated. In each of these seven cases, manual examination indicates that the identifications are likely to be correct.

Each spectrum was matched whenever possible to yeast peptides using Mascot or ProICAT, followed by Mascot as a search engine. A great deal of effort was spent in determining a plausible sequence that could explain high-quality spectra. In many cases, spectra matched chemically modified forms of abundant peptides (see Table IVB). In most cases, these chemical modifications were either oxidations of Met, Trp, or the ICAT reagent itself. Whenever such a modification was found, much of the original data was submitted to another round of database searching with Mascot, with altered settings so that additional instances of the same modification could be identified. This had the effect of increasing the total number of spectra that could be identified without increasing the number of peptides or proteins that could be matched. In some cases, spectra that had originally been matched to ICAT peptides with low confidence could be better explained by...
modified abundant peptides. In most cases, MatSc (described above) correctly determined which identification was most plausible. In a few cases, manual examination of the spectra indicated that these identifications were unlikely to be correct (due to strong peaks that could not be explained by known fragmentation patterns), and these spectra were manually excised from the list of identified spectra (Y1 class 99). By this means, 15,233 CICAT tryptic peptide identifications from yeast have been made that pass a minimum threshold of credibility (MatSc > 5000, at least four b or y ions matched, MAc = 1), and that could not be attributed to alternative sequences by automatic means (Table V).

When the criteria for matching was tightened (requiring in additional Sc = 20, NT = 2, MBC = 0, 0.2 < HL < 5), this list was reduced about 2-fold to 6337 identifications (Table V).

Example of a Borderline Identification—Table VI contains three peak lists corresponding to two different spectra. The first list corresponds to MALDI spectrum 52964 (Table III, spectrum 4), which was matched to peptide ITLHVDcLR from imidazoleglycerol-phosphate dehydrogenase, also known as His3p (see “Biological Significance” below). This is an example of a noisy, borderline MALDI spectrum (Fig. 3A) that has a low Mascot score of 9, but MatSc = 36,188, and yb = 6. Many spectra from Table S3 of similar low quality, MatSc, Sc, and # yb ions are probably incorrectly identified, based on selected manual examination of spectra. However, this spectrum derived from a spot with two other identified peptides. Immediately adjacent spots supplied an additional three peptides. Upon internal calibration, all five of these peptides matched each other to within 10 ppm. Peptide ITLHVDcLR matched to 4 ppm to this recalibrated spectrum, suggesting that ITLHVDcLR could indeed explain this spectrum.

Neither ITLHVDCLR nor ITLHVDcLR was matched automatically to any other spectra, but when all spectra with the appropriate parent mass were examined (34 spectra within 1 amu), an originally unmatched electrospray spectrum was identified that might correspond to ITLHVDcLR, based on only two yb ions (Table III, spectrum 5). Thus, this spectrum did not even meet the minimal criteria for tentative identification. When the peak list (Table IV, 2) was submitted to Mascot, peptide ITLHVDcLR was nonetheless identified as the top candidate from Saccharomyces with a score of 11, whereas MatSc was calculated to be 11,235 (Fig. 3B). When the original spectrum was examined (Fig. 3C), some strong peaks were evident that had not been accounted for. Manual examination revealed that these peaks correspond to doubly charged y ions, whose charge state had not been automatically deduced. A revised peak list (Table VI, 3) was manually prepared from the six strongest peaks, excluding the doubly charged parent ion. When peak list 3 was submitted to Mascot, a score of 45 was returned (a score of 26 is required for identity according to Mascot), based on matching five of the six peaks to five distinct y ions (Fig. 3D), strongly supporting the identification of ITLHVDcLR. Thus, by extrapolation, in all likelihood some of the spectra in Table S2 remain unidentified because of poor peak extraction, which is often a problem with borderline MS/MS spectra. The HL pattern for the electrospray parent spectrum is shown in Fig. 3E. The heavy form of this HL pair is readily detected, but the light form of this peptide is in the noise. Indeed, the isotope cluster for the light form would appear to be assignable to a parent mass that is 0.5 amu too low to be the light form of ITLHVDCLR. Thus, this peptide appears to be significantly up-regulated in the upf1 knockout strain, perhaps by even more than was calculated from the corresponding MALDI spectrum (4.61-fold).

Noncanonical Peptides—Database searching was also directed against three additional categories of peptides: peptides that have no Cys residues and were not modified that arise from incomplete avidin column washing (Y2 class 2), peptides that were modified by the ICAT reagent on Lys residues (KICAT; Y2 class 3), and CICAT peptides that were incompletely alkylated (Y2 class 4). A total of 2731 spectra fell into one of those three categories at low stringency (Table V), whereas 1144 spectra could be so classified at higher stringency (Mascot score of 20, two tryptic termini, and MAc = 1).

There were 886 spectra that corresponded to unlabeled tryptic peptides, many of which were observed repeatedly. Table VII lists each of the 24 proteins that were identified based on at least 10 different spectra. Most of these proteins are metabolic enzymes or ribosomal proteins with high codon adaptation indices (CAI); the lowest CAI in Table VII is 0.52.

Another 218 spectra matched with high confidence to KICAT peptides (Table VIII). The MALDI spectra in this category mostly had HL ratios near 1.0, but were not quantified automatically by the ProICAT software because quantification is coupled to identification, and they were not identified in the initial round of database searching. These spectra corresponded to 117 different peptides, from 70 different proteins. A total of 82 of the peptides, and 54 of the proteins, were encountered two times or more. As with the unlabeled peptides, these proteins were mostly common metabolic enzymes (Table VIII). Of the proteins in Table VIII, the lowest CAI observed was 0.52 (for a heat shock protein), but all other proteins had a codon bias of at least 0.7.

The final special category consists of 119 spectra (at low stringency) that could be matched to partially alkylated peptides. This corresponded to 82 distinct peptides, 16 of which were encountered more than once (Table IX). These 16 peptides correspond to 15 proteins, which are also rather abundant (Table S9). All but one of the peptides identified more than once were from MALDI spectra, and in every case the same MALDI spot also contained the fully alkylated form of the peptide. On this account, the ISD parameter was set to 1, indicating that in these cases the partial alkylation observed was probably an artifact of in-source decay. The one electrospray peptide that was identified more than once corresponded to glyceraldehyde phosphate dehydrogenase, the most commonly encountered ICAT reagent-labeled protein in yeast.

Depth of Proteome Issues in Yeast
A

B

**FIG. 3.** **Electrospray and MALDI spectra that support the identification of peptide ITLHVDcLR.** *A,* a noisy MALDI spectrum with a few y and b ions that support the identification. The peak list is shown in Table VI, *spectrum 1.* *B,* a weak electrospray spectrum drawn from the peak list that was automatically deposited into Table VI (*spectrum 2*). Two y ions match to peptide ITLHVDcLR. *C,* raw spectrum from which the peak list in *B* was derived. Many more peaks correspond to peptide ITLHVDcLR. *D,* same spectrum as *B* and *C,* but with manual peak de-isotoping and de-charging. The area of each member of the appropriate isotope cluster (up to four peaks) was added to obtain the plotted intensity. If more than one charge state was detected, these areas were combined. In some instances, there was some ambiguity as to where the monoisotopic mass was. The peptide sequence was used to ensure that the correct mass was used. The strongest peak that could not easily be attributed to peptide ITLHVDcLR is also shown below mass 400. *E,* the parent spectrum in the neighborhood of the parent ion selected for fragmentation. The isotope cluster at the nominal mass of the light form of the peptide is barely detectable.

In this case, the fully alkylated form of this peptide did not co-elute, and therefore the detection of this peptide may indicate a small amount of incompletely alkylated peptides in this preparation. A fourth category of noncanonical peptides consist of peptides that do not have two normal tryptic termini (Table X). For the purposes of this study, a nontryptic terminus that is known...
to be formed during normal protein biosynthesis is not classified as nontryptic; for example, if peptides are generated by removal of N-terminal Met residues or signal sequences. A great deal of effort was spent in manually examining these peptides, with the result that many of these spectra eventually were explained by chemical modifications of other peptides, or were reclassified as completely uninterpretable (Y1/H11005 99).

So far, no peptide with two nontryptic termini (nontryptic) has held up to manual scrutiny; many of these spectra can be positively identified and turn out to be chemically modified CICAT peptides. Only 23 distinct sequences have been identified more than once with even one nonbiological nontryptic terminus (semitryptic peptides; Table X). All but two of these correspond to abundant yeast proteins; the third lowest CAI is 0.709. (One protein, the major coat protein, was not listed in the CAI table but may nonetheless be abundant). No nontryptic peptides with a Mascot score of Y2/H11005 20 matched within 1 amu (MBC/H11005 0), compared with 36 semitryptic peptides. Many of these semitryptic peptides do hold up to scrutiny, but in all likelihood, based on previous experience, many will not. In Table X, these peptides are listed alphabetically starting with the amino acid preceding the sequence that was matched. By this means, it becomes clear that many of the semitryptic peptides belong to subcategories. For example, seven distinct sequences (out of 23 total) match to peptides in which there is an Asp residue immediately preceding the peptide in place of the canonical tryptic Lys or Arg residue, and six additional sequences have a preceding Asn residue. Surprisingly, there is no evidence of cleavage due to contaminating chymotrypsin-like enzymes. In fact, only two of the 36 peptides terminate with an aa other than Lys or Arg, and both of these sequences terminated with Asn. Perhaps the strong preference for N-terminal semitryptic sites is because Lys and Arg promote ionization and are therefore more easily detected.

Table X summarizes the number of fully tryptic, semitryptic, and nontryptic identifications as a function of stringency and Y2 category. As mentioned above, no peptides have been identified with a Mascot score of Y2/H11005 20 that are nontryptic and also match within 1 mass unit, and only 2.4% of the high-confidence CICAT peptides are semitryptic. The absence of KICAT peptides in the semitryptic and nontryptic categories at low stringency is because no searches were performed to look for them, but the other data would indicate they are unlikely to be very common. The low-stringency column for nontryptic peptides indicates that 204 spectra fall within this category; however, these are likely to be misidentifications. Extensive manual examination of spectra that corresponded to the highest scoring nontryptic peptides has not revealed any nontryptic peptides that appear to be credible identifications.

### Table VII

**Unlabeled Peptides**

All proteins from which unlabeled peptides were identified at least 10 times are listed, sorted by decreasing frequency of identification. Acc, the Swiss-Prot accession number of the protein. If there were two yeast proteins that could explain the same set of peptides, the protein with the higher codon bias was chosen. No. acc, the number of identifications of each peptide, with Sc/H11350 20; MBC/H11005 0, MAC/H11021 4; Y2/H11005 2. Name, the protein name. Codon, the codon bias for the protein.

| Acc     | No. acc | Name                                             | Codon |
|---------|---------|--------------------------------------------------|-------|
| P00925  | 47      | Enolase 2                                        | 0.89  |
| P00359  | 32      | Glyceraldehyde 3-phosphate dehydrogenase         | 0.92  |
| P00580  | 31      | Phosphoglycerate kinase                          | 0.82  |
| P00824  | 28      | Enolase 1                                        | 0.87  |
| P06169  | 27      | Pyruvate decarboxylase isozyme 1                 | 0.9   |
| P00330  | 26      | Alcohol dehydrogenase i                          | 0.81  |
| P05735  | 26      | 60s ribosomal protein 119                        | 0.71  |
| P00549  | 25      | Pyruvate kinase 1                                | 0.89  |
| P14126  | 25      | 60s ribosomal protein 13                         | 0.83  |
| P02994  | 24      | Elongation factor 1-alpha                        | 0.88  |
| P10592  | 22      | Heat shock protein ssa2                          | 0.8   |
| P11484  | 20      | Heat shock protein ssb1                          | 0.82  |
| P02829  | 15      | Heat shock protein hsp82                         | 0.52  |
| P46654  | 13      | 40s ribosomal protein s0-b                       | 0.73  |
| P14540  | 12      | Fructose-bisphosphate aldolase                   | 0.87  |
| P23301  | 12      | Initiation factor 5a-2 (elf-5a)                  | 0.8   |
| P32324  | 12      | Elongation factor 2                              | 0.8   |
| P38706  | 12      | 60s ribosomal protein 127                        | 0.74  |
| P02365  | 11      | 40s ribosomal protein s6                         | 0.85  |
| P02407  | 11      | 40s ribosomal protein s17-a                      | 0.81  |
| P35030  | 10      | 60s ribosomal protein 11                         | 0.83  |
| P02400  | 10      | 60s acidic ribosomal protein p2-beta             | 0.76  |
| P05030  | 10      | Plasma membrane atpase 1                        | 0.73  |
| P33442  | 10      | 40s ribosomal protein s1-a                       | 0.7   |
### TABLE VIII
**Peptides alkylated on lysine**

All identifications that were observed at least five times are listed; sorted by Acc, then No. descending, then Seq. This emphasizes that some abundant proteins contributed several of the repeatedly observed KICAT peptides. No., number of identifications where Y1/H110053, Y2/H110053, MatSc/H1135010,000; MBC/H110050, MAc/H110214, CalMW/H110221000. EM, 1 means MALDI, 2 means electrospray, 3 means both. NT, number of tryptic termini.

| No. | EM | NT | IT | ISD/H11021 | Seq              | Acc    | MatSc  | Sc    | Name                  | Codon  |
|-----|----|----|----|------------|------------------|--------|--------|------|-----------------------|--------|
| 1   | 7  | 2  | 0  | K          | SPIxVVGGLSTPLEIYEK | M      | P00330 | 72954 | Alcohol dehydrogenase i | 0.81   |
| 2   | 11 | 3  | 2  | N          | AVGxVLPELOGK      | L      | P00359 | 325494 | Glyceraldehyde 3-phosphate DH | 0.92    |
| 3   | 11 | 3  | 2  | R          | VWDLEHVxA         | P00359 | 482985 | 87    | Glyceraldehyde 3-phosphate DH | 0.92    |
| 4   | 10 | 3  | 2  | R          | TASNIP3GSTGAAxAVGK | V      | P00359 | 179265 | Glyceraldehyde 3-phosphate DH | 0.92    |
| 5   | 9  | 2  | 2  | N          | GDLGIPAEVLAVoXK   | L      | P00549 | 182774 | Pyruvate kinase 1       | 0.89    |
| 6   | 7  | 3  | 2  | N          | IVAALTxTVYLEHHPR  | Y      | P00560 | 152052 | Phosphoglycerate kinase   | 0.82    |
| 7   | 6  | 2  | 2  | K          | YSLAPxVAXELSLLGK  | D      | P00560 | 61468  | Phosphoglycerate kinase   | 0.82    |
| 8   | 9  | 2  | 2  | K          | AADALLxVNIQTLSEISK| A      | P00825 | 162525 | Enolase 2               | 0.89    |
| 9   | 8  | 3  | 2  | K          | HEEAGxXHGDXL       | P00825 | 37671  | 15    | Enolase 2               | 0.89    |
| 10  | 7  | 3  | 2  | 1         | RIEEGDxAVYGENFHFHDGKL | P00825 | 72254  | 54    | Enolase 2               | 0.89    |
| 11  | 5  | 2  | 2  | R          | AAAAExNVPLYQHLADLSK| S      | P00825 | 103207 | Enolase 2               | 0.89    |
| 12  | 5  | 3  | 2  | R          | AVDDFLLSLDGTANxSK | L      | P00925 | 233347 | Heat shock protein hsp82  | 0.52    |
| 13  | 6  | 2  | 2  | R          | VVVSYxLLDAPAAIR    | T      | P02829 | 144937 | Heat shock protein hsp82  | 0.52    |
| 14  | 5  | 2  | 2  | K          | FLxSGDAALVK        | F      | P02994 | 197195 | Elongation factor 1-alpha | 0.88    |
| 15  | 6  | 3  | 2  | R          | YxYVDPNVLPETESLALVIDR | P02994 | 90466  | 27    | Elongation factor 1-alpha | 0.88    |
| 16  | 8  | 2  | 2  | R          | YGGVYVGTLSxPEVK    | E      | P06169 | 286267 | Pyruvate decarboxylase    | 0.9     |
| 17  | 9  | 2  | 2  | K          | AADDxLVNIQTLSEISK  | A      | P06169 | 74849  | Pyruvate decarboxylase    | 0.9     |
| 18  | 6  | 3  | 2  | R          | IEEELGDxAVYAGENFHHGDKL | P06169 | 72254  | 54    | Enolase 2               | 0.89    |
| 19  | 5  | 2  | 2  | 1         | RADDxNVPLYQHLADLSK | P06169 | 74849  | 57    | Enolase 2               | 0.89    |
| 20  | 2  | 1  | 2  | 0         | 1 R               | P06169 | 74849  | 57    | Enolase 2               | 0.89    |

### TABLE IX
**Incompletely alkylated peptides**

As with Table VIII, all identifications that were observed twice were listed, using the same criteria as with Table VIII, except that Y1/H110054 and Y2/H110054. Field abbreviations are as in Table VIII. All peptides except 2, which was also identified by electrospray, were identified by MALDI in the same spot as the fully alkylated CICAT peptide. In electrospray, peptide 2 did not co-elute with the fully alkylated CICAT peptide.

| No. | EM | NT | IT | ISD/H11021 | Seq              | Acc    | MatSc  | Sc    | Name                  | Codon  |
|-----|----|----|----|------------|------------------|--------|--------|------|-----------------------|--------|
| 1   | 3  | 1  | 2  | 0         | K                | CZSDxFNOVQVK  | S      | P00330 | 238416 | Alcohol dehydrogenase i | 0.81   |
| 2   | 7  | 3  | 2  | 0         | K                | IVSNASZTNCLAPLAK | V    | P00359 | 80420  | Glyceraldehyde 3-phosphate DH | 0.92    |
| 3   | 2  | 1  | 2  | 0         | K                | YRPNZPILVTR   | C      | P00549 | 26367  | Pyruvate kinase 1       | 0.89    |
| 4   | 2  | 1  | 2  | 0         | K                | DVTFLNVDZVPEVEAAVK | A    | P00560 | 60248  | Phosphoglycerate kinase   | 0.82    |
| 5   | 13 | 1  | 2  | 0         | K                | NPVLZQPIIAPFLGxLPGLK | S    | P01625 | 112896 | Pyruvate decarboxylase isoyme 1 | 0.90    |
| 6   | 2  | 1  | 2  | 0         | K                | ELVSZTNCTDYQSR | N      | P07284 | 44327  | Seryl-trna synthetase cytoplasmic | 0.39    |
| 7   | 2  | 1  | 2  | 0         | K                | VNLTDZQAYLTLGIR | D      | P14540 | 98317  | Fructose-biphosphatase aldolase | 0.87    |
| 8   | 2  | 1  | 2  | 0         | K                | YZGIPGNCGCCG | S      | P16521 | 37239  | 40s ribosomal protein s11   | 0.73    |
| 9   | 2  | 1  | 2  | 0         | K                | TXGZPQGNGVG   | T      | P26781 | 27407  | 40s ribosomal protein s11   | 0.73    |
| 10  | 2  | 1  | 2  | 0         | K                | TPLGZPVQGNGVG | T      | P26781 | 27407  | 40s ribosomal protein s11   | 0.73    |
| 11  | 2  | 1  | 2  | 0         | K                | GNFZPQGNGVG   | T      | P26781 | 27407  | 40s ribosomal protein s11   | 0.73    |
| 12  | 2  | 1  | 2  | 0         | K                | TXGZPQGNGVG   | T      | P26781 | 27407  | 40s ribosomal protein s11   | 0.73    |
| 13  | 2  | 1  | 2  | 0         | K                | TXGZPQGNGVG   | T      | P26781 | 27407  | 40s ribosomal protein s11   | 0.73    |
| 14  | 2  | 1  | 2  | 0         | K                | TXGZPQGNGVG   | T      | P26781 | 27407  | 40s ribosomal protein s11   | 0.73    |
| 15  | 2  | 1  | 2  | 0         | K                | TXGZPQGNGVG   | T      | P26781 | 27407  | 40s ribosomal protein s11   | 0.73    |

**Depth of Proteome Issues in Yeast**

Molecular & Cellular Proteomics 3.7 647
tides that began at the protein bonds were classified as canonical tryptic cleavages, as were peptides eluted several minutes prior to the corresponding unoxidized Met (20). In most cases, these oxidized CICAT peptides modified CICAT peptides, which is also similar to what was found with peptides containing oxidized Met residues.

**Peptides Identified and Quantified**—As it turned out, there were 5668 spectra that corresponded to peptides that are useful for quantification (Fig. 2C). This requires 1) that they be reliably identified (Sc ≥ 20, MAc = 1); 2) that they be CICAT peptides without chemical modifications other than oxidation of Met; 3) that they be intense enough to be automatically quantified; and 4) that they be chromatographically well resolved so that they do not overlap with any confounding HL pairs (HL_I = 1). Some of these spectra derive from both members of the HL pair, or from multiple charge states, or from the same chromatographic peaks, thus the 5668 identifications correspond to 3726 distinct HL ratio measurements. When consolidated by peptide, this results in 1361 distinct CICAT peptides (Table XII). When the lower-stringency data are included, there are 2181 peptides. Because this table is so large, only 39 peptides from 14 proteins of particular interest (see “Biological Significance” below) are displayed; the entire table may be viewed electronically as Table S12.

Table XII is sorted by protein by decreasing protein ratio, and then by peptide sequence, so one can get a sense of how reproducible the HL measurements are. For four different peptides, no HL ratio was automatically measured, so that column HL is blank. To demonstrate the consequences of including lower-quality data, the HL ratio, standard deviation, and number of measurements are calculated in two different ways. For peptides with multiple mature N terminus or C-terminal sequences, the MAc of the sequence with the highest MAc is used.

**Table X**

| No. | EM | NT | IT < | Seq | > | Acc | MatSc | Sc | Name | Codon |
|-----|----|----|------|-----|---|-----|-------|----|------|-------|
| 1   | 4  | 3  | 1    | 0 D | CASSEFFK | D | P00924 | 279660 | 32 | Enolase 1 | 0.871 |
| 2   | 9  | 2  | 1    | 0 D | COYALTGIR | D | P14540 | 179965 | 54 | Fructose-bisphosphate aldolase | 0.869 |
| 3   | 3  | 3  | 1    | 0 D | CAVPEVEAAK | A | P00560 | 100629 | 26 | Phosphoglycerate kinase | 0.815 |
| 4   | 3  | 2  | 1    | 0 D | CVMLSGETAK | G | P00549 | 202147 | 41 | Pyruvate kinase 1 | 0.893 |
| 5   | 4  | 2  | 1    | 0 D | GADGVMLGETAK | G | P00549 | 60255 | 32 | Pyruvate kinase 1 | 0.893 |
| 6   | 3  | 2  | 1    | 0 D | TDOYAYLGTIR | D | P14540 | 89640 | 56 | Fructose-bisphosphate aldolase | 0.869 |
| 7   | 2  | 2  | 1    | 0 D | VGNAIGDGVMLGETAK | G | P00549 | 34685 | 35 | Pyruvate kinase 1 | 0.893 |
| 8   | 13 | 3  | 1    | 0 H | MLDLSESEVENISTCVK | Y | P14540 | 255636 | 105 | Fructose-bisphosphate aldolase | 0.869 |
| 9   | 3  | 2  | 1    | 0 I | PVLHSDHCAC | K | P14540 | 173192 | 41 | Fructose-bisphosphate aldolase | 0.869 |
| 10  | 2  | 2  | 1    | 0 K | FYCVPRPN | A | P40075 | 196988 | 24 | scs2 protein | 0.194 |
| 11  | 4  | 2  | 1    | 0 K | LPNCPRPN | M | P53030 | 214725 | 41 | 60s ribosomal protein l1 | 0.832 |
| 12  | 4  | 2  | 1    | 0 L | PVGAIMNCADNSGAR | N | P04451 | 199756 | 55 | 60s ribosomal protein l23 | 0.648 |
| 13  | 58 | 3  | 1    | 0 M | CVEAFSEYPPLGR | F | P02994 | 321280 | 82 | Elongation factor 1-alpha | 0.876 |
| 14  | 4  | 3  | 1    | 0 N | AScTTCNLAPLAK | V | P00359 | 78052 | 39 | Glyceraldehyde 3-phosphate dehydrogenase | 0.924 |
| 15  | 9  | 3  | 1    | 0 N | cLAPLAK | V | P00359 | 329653 | 33 | Glyceraldehyde 3-phosphate dehydrogenase | 0.924 |
| 16  | 7  | 3  | 1    | 0 N | CPIILYTR | C | P00549 | 230077 | 49 | Pyruvate kinase 1 | 0.893 |
| 17  | 32 | 3  | 1    | 0 N | mSICIFGDAFDVDR | A | P53030 | 225208 | 87 | 60s ribosomal protein l9-b | 0.832 |
| 18  | 3  | 3  | 1    | 0 N | PVILADAccSR | H | P06169 | 200102 | 56 | Pyruvate decarboxylase isozyme i | 0.904 |
| 19  | 2  | 3  | 1    | 0 N | SVEDSONAALQLOICR | V | P51401 | 127115 | 70 | 60s ribosomal protein l9-b | 0.778 |
| 20  | 23 | 3  | 1    | 0 V | SSDLFSICSDR | G | P32503 | 158772 | 55 | Major coat protein | 0.778 |
| 21  | 6  | 2  | 1    | 1 W | LDLKLSGCYAPPRSPAGHK | L | P05753 | 116681 | 45 | 40s ribosomal protein s4 | 0.709 |
| 22  | 10 | 3  | 1    | 0 Y | IVOLVPACNTNAGNK | D | P16521 | 128602 | 44 | Elongation factor 3a | 0.778 |
| 23  | 2  | 3  | 1    | 0 Y | SORVVESSVEIANEQGNR | V | P11484 | 97480 | 100 | Heat shock protein ssb1 | 0.82 |

**Table XI**

| NT, number of noncanonical tryptic termini. Arg-Pro and Lys-Pro bonds were classified as canonical tryptic cleavages, as were peptides that began at the protein’s mature N terminus or C-terminal peptides. C, number of identifications of CICAT peptides. N, number of identifications of unlabelled peptides. X, number of identifications of KICAT peptides. Z, number of identifications of incompletely alkylated CICAT peptides. High confidence, same as Table X. Low confidence, MatSc ≥ 5000; yb > 3, MBC = 0, MAc < 4; CalMW > 900.

| NT | C | N | X | Z |
|----|---|---|---|---|
| 2  | 7884 | 840 | 217 | 35 |
| 1  | 188  | 39  | 0   | 1  |
| 0  | 0    | 0   | 0   | 0  |
| Low confidence |
| 2  | 13547 | 1233 | 371 | 129 |
| 1  | 351  | 500  | 0   | 31 |
| 0  | 31   | 150  | 0   | 23 |
Fig. 4. Spectrum corresponding to a peptide with an oxidized CICAT residue. Spectra that were matched to the oxidized light (E7ADLWPR; A) and heavy (E8ADLWPR; B) forms of peptide ECADLWPR, derived from ribosomal protein L23. In one aa code, 7 refers to oxidized light CICAT, whereas 8 refers to oxidized heavy CICAT. The most prominent y and b ions are labeled, as well as some y ions that appear to be derived by a neutral loss from the oxidized side chain, labeled yn7 and yn8. The peaks labeled u1 and u2 presumably are derived upon decay of the oxidized side chain itself, as they were detected 9 amu apart with similar intensities in each spectrum.
the standard deviation.

automatically measured. Note that in some instances, like 1, 7, and 9, HL measurements depend on identifications of lower stringency. Std, the number of spectra that supported the identification of the peptide. HL, the average HL ratio. When this is blank, the ratio was not

/3

High stringency here means Sc

/3

".

Depth of Proteome Issues in Yeast

The peptides are listed, sorted by Acc, and then by Seq, in order to emphasize how consistent the HL ratios were for each protein. The ratios much closer to 1.0 –

tides identified, respectively. All of the measurements for

–

3, Y2

/3

/3

/3

88 1 5 12 56.00 30649 0

8 1179.59 P40581 5 CGFTPQYK 1 5 2.03 1.69 1.01 2 3 4 6 23.98 144373 25

1875.90 P40706 10 ETHISGECVATAER 1 1 1.17 0.72 0.10 1 26 1 4 2 198.00 46356 23

2538.28 P40706 9 SADPSGCAFTK 10 16 1.61 1.68 1.03 31973 1 1 5 7 128.00 217205 40

2256.78 P40706 8 TVYFGDLGEGCK 7 7 0.68 0.68 0.07 0.07 7131

1282.34 P39635 7 DINIAESAFLCVDHELAAER 1 1

1626.98 P39635 7 HLSVIGECVQYALQPDGLYR 0 1

1952.06 P39635 6 IECLISLVK 0 1

1082.62 P39635 4 VIECNR 0 2 1.73 1.68 0.09 7 12319 11 1 6 5 49.00 26313 10

1233.76 P39635 4 CMNIVIYK 1 3 1.48 1.48 0.00 2601 43 2 2 8 42.25 218938 50

2842.34 P39635 3 CMVNYQ 1 3 1.17 0.72 0.10 1 26 0 1 4 2 198.00 46356 23

2747.21 P39635 3 YHIIDHEYDCVVIGAGGAGLR 1 2 0.80 0.80 0.00 16338 0 3 5 8 71.00 324702133

2696.14 P39635 2 GLPTTAACPSFQYK 1 5 2.03 1.69 1.01 2 3 4 6 23.98 144373 25

3009.38 P32528 10 QVDSAGQ 0 1 1.07 0.69 0.38 0.07 4 3353

2127.99 P32528 9 FPEDYGCQVPR 1 2 0.71 0.74 0.06 1674 0 1 5 8 66.00 61928 22

2878.73 P32528 8 OEGCFAVSHGCTGK 0 1 1.98 0.10 290 0 24.91 0.88 1 5 12 56.00 30649 0

2494.24 Q00711 9 THIHLDEDCEVSPTVTR 1 2 0.65 0.66 0.03 0.21 512 0 24.91 11 1 6 14 65.00 128847 49

1875.90 P40706 7 ETHISGECVATAER 1 2 1.41 1.48 0.25 1 3730 0 1 6 16 33.00 132276 61

2696.14 P32528 7 KIYFVTPTQ 0 1 1.07 0.69 0.38 0.07 4 3353

2494.24 Q00711 8 FASDPGCAFTK 10 16 1.61 1.68 0.22 0.18 2528 9 3 4 5 26.17 33778 50

3004.38 P39635 8 KFPADDYKFQYQAISQODADSECK 1 1 1.75 1.75 1 4114 0 1 6 15 100.00 81677 44

1875.90 P40706 6 LSNFQK 0 1 1.07 0.69 0.38 0.07 4 3353

2296.02 P49095 4 VAMCSTGQ 1 2 0.86 0.86 0.07 0.07 7131 5 3 3 5 51.27 188484 54

1978.98 P32528 4 KYSAVDCDSFEYKVR 1 1 1.18 1.18 1 913 0 6 1 6 16 80.00 18916 42

1233.76 P39635 3 CMVNYQ 1 3 1.17 0.72 0.10 1 26 0 1 4 2 198.00 46356 23

2842.34 P39635 3 CMVNYQ 1 3 1.17 0.72 0.10 1 26 0 1 4 2 198.00 46356 23
### TABLE XIII

| Acc | Pep No. | ID No. | HL | Std | Sc | MatSc | Seq | yb | Name |
|-----|---------|-------|----|-----|----|-------|-----|----|------|
| P06633 | 1 1 | 1 | 4.6 | 3.3 | 1.4 | 1.1 | 35 | 6 | 35510 ITHLVicLR | 6 Imidazoleglycerol-phosphate dehydratase |
| P07258 | 2 3 3 11 | 4.3 | 3.3 | 1.4 | 1.1 | 35 | 205500 ANVALIDcGVR | 12 Arg. specific carbamoylphosphate synthase, small |
| P48015 | 1 1 1 1 3 3 | 41 | 64090 ENDNDFEYLVTNAcGER | 14 Glycine decarboxylase, subunit T |
| P49095 | 1 2 1 2 2.5 2.4 | 0.4 | 41 | 85750 TFSIPHGcGAGPAGicVK | 20 Glycine decarboxylase subunit |
| P40581 | 1 1 1 5 1.1 2 1 | 25 | 144373 cGFTPOYK | 5 Glutathione peroxidase |
| P25282 | 1 3 1 1 3.2 1.8 | 0.4 | 9 | 40694 IEcLISLcVK | 6 Urea amidoase |
| P03965 | 2 7 3 12 1.9 1.8 0.6 0.3 | 50 | 218938 cMNIVNIcVK | 7 Arg. specific carbamoylphosphate synthase, large |
| P22768 | 3 7 | 1 | 0.1 | 45 | 249411 GoYEGAPLTvLR | 10 Arginosuccinate synthetase |
| P38013 | 3 3 27 34 | 1.6 | 1.6 | 0.2 | 0.2 | 35 | 337770 FASDPGcAFTK | 9 Sim. to C. boidinii peroxisomal membrane protein |
| P04076 | 1 1 | 1 1 2 1.7 | 0.2 | 61 | 132276 ETHHISEcGATAcER | 17 Arginosuccinate lyase |
| P06208 | 4 8 | 12 20 | 1.3 | 1.2 | 0.1 | 0.5 | 1 | 339379 VcIScTHicHNDR | 10 2-isopropylmalalate synthase |
| P13663 | 6 6 | 20 23 | 1.3 | 1.2 | 0.1 | 0.4 | 24 | 380370 EYcVicDAYK | 13 Aspartate-semialdehyde dehydrogenase |
| P53849 | 5 8 | 10 32 | 1.6 | 1.2 | 0.4 | 0.4 | 66 | 175947 cFNCcNQTHGicSR | 12 Similarity to nucleic acid-binding proteins |
| Q00095 | 6 9 | 13 29 | 1.1 | 1.2 | 0.2 | 0.2 | 31 | 189579 cDILcLR | 6 Acetyl-CoA carboxylase |
| P0931 | 7 9 16 25 | 1.1 | 1.1 | 0.2 | 0.2 | 110 | 259579 FGFDcQGYVPEALcVK | 19 Tripyrithion synthase |
| P14742 | 5 8 6 17 | 1.1 | 0.1 | 0.4 | 0.5 | 85 | 290930 DWTvSHGCcIAHTR | 17 Glucosamine—fructose-6-phosphate transaminase |
| P07264 | 9 13 25 | 46 | 1.1 | 0.4 | 0.3 | 33 | 292149 LCVdLPNQcK | 10 3-isopropylmalate dehydratase |
| P08566 | 7 11 13 | 23 | 1.2 | 0.2 | 0.6 | 61 | 201444 ALILcALGEGcQcK | 13 Arom pentafucntional enzyme |
| P38628 | 4 8 | 6 12 | 1.1 | 1.1 | 0.2 | 0.2 | 10 | 100746 VYAcEcK | 5 Phosphoacetylglyceromutase |
| P25491 | 7 8 16 | 26 | 1.1 | 0.1 | 0.1 | 0.1 | 68 | 322053 CALKcYHDPKcNPSEEAk | 20 Mitochondrial and ER import protein |
| P07149 | 6 9 | 24 41 | 1 | 0.1 | 1 | 0.1 | 61 | 273010 TclLHGPVAAcFTK | 18 Fatty-acyl-CoA synthase, β chain |
| P10964 | 6 8 | 7 15 | 1 | 1 | 0.1 | 0.4 | 43 | 76526 DGALLcGILcDK | 9 DNA-directed RNA polymerase I, 190 kDa subunit |
| P32505 | 3 8 | 5 22 | 1 | 0.1 | 0.2 | 0.2 | 16 | 187125 NcYcICLcFR | 7 Nuclear poly(A)-binding protein |
| P00958 | 7 9 | 16 24 | 0.1 | 0.1 | 0.1 | 0.1 | 52 | 297445 CGGLc RailroadPFCENPR | 14 Methionyl-tRNA synthetase |
| P18628 | 8 10 | 42 | 64 9 | 0.1 | 0.1 | 0.1 | 64 | 435078 SAVIDMNcK | 24 Translation elongation factor eEF3 |
| P00549 | 5 14 58 | 91 0.9 | 0.1 | 0.1 | 0.1 | 74 | 324802 ACCDKIMyVDYK | 21 Pyruvate kinase |
| P07259 | 10 | 12 | 27 | 42 0.9 | 0.1 | 0.1 | 31 | 189383 TDVLYcTR | 6 Multifunctional pyrimidine biosynthesis protein |
| P14540 | 5 | 12 | 44 78 | 1 | 0.9 | 0.1 | 0.1 | 100 | 25092 SIAPAPcVLPHLDSHCcAK | 26 Fruuctose-bisphosphate aldolase |
| Q12680 | 6 13 | 7 24 | 0.9 | 0.9 | 0.1 | 0.1 | 41 | 218882 YCGAILcDR | 7 Glutaminate synthase (NADPH) |
| P32324 | 6 9 28 | 36 | 0.9 | 0.1 | 0.1 | 0.1 | 88 | 204513 AEOLYEGPADDANcIAk | 27 Translation elongation factor eEF2 |
| P38205 | 7 11 | 10 29 | 0.8 | 0.9 | 0.2 | 0.2 | 22 | 149054 cGINLMcVK | 10 Similarity to nucleolar Nop2p |
| P02994 | 7 9 | 19 | 55 | 0.9 | 0.9 | 0.2 | 0.2 | 318119 cVEcAFSEYPPGL | 11 Translation elongation factor eEF1 α A chain |
| Q0071 | 2 3 | 2 5 | 0.7 | 0.8 | 0.2 | 0.2 | 49 | 128847 VDHTdLKEcPSVPPTVTR | 9 Succinate DH flavoprotein precursor |
| P33302 | 8 22 | 34 | 22 | 0.7 | 0.1 | 0.1 | 52 | 286686 TVYFGDcLGcECK | 12 Pleiotropic drug resistance protein |
| P21801 | 1 1 2 2 | 0.5 | 0.5 | 0.1 | 0.1 | 27 | 67577 NTLcACIC | 6 Succinate DH iron-sulfur protein subunit |
| P30771 | 1 | 1 2 | 0.5 | 19 | 54460 ILVCAPSNVADHLcAAcK | 7 Nonsense-mediated mRNA decay protein |

**Proteins Identified and Quantified**—Table S13 lists all 1029 of the identified proteins, sorted from highest ratio to lowest ratio. Table XIII lists the 14 “interesting” proteins discussed below as well as all proteins that contributed at least eight distinct peptides. As with Table XII, the No. of Pep, No. of ID, HL, and Std fields are duplicated to show the consequences of including lower-stringency data. One representative peptide sequence, as well as the Mascot score, MatSc, and yb are also displayed to illustrate the confidence of the identification. As with the peptide data, including HL data from low-stringency identifications tends to support the data obtained from the high-stringency identifications. In all cases, we have used the ICAT dictionary to assign the identified peptides to the protein with the highest codon bias.

**Variation in HL Ratio**—Fig. 5 displays the variation in the HL ratio observed for the 702 proteins derived from 5668 high-stringency identifications. Each protein was categorized according to ln(HL ratio) into different bins, which were distributed so that each bin was 0.1 unit wide. It can be seen that the bulk of the proteins are accommodated in one of the five central bins, with HL ratios of between 0.8 and 1.2. Only 26 out of 702 proteins have ratios below 0.74, whereas 40 proteins have HL ratios higher than 1.35. The best measure of a meaningful up-regulated or down-regulated HL ratio is not the...
absolute value of the HL ratio, but the reproducibility of the measurements, which are expected to be different for different proteins, depending on how many individual measurements contribute to the measured HL ratio.

The measurements that held up to manual scrutiny range from 0.27 to 3.8. The protein at 0.27 is Upf1p itself, which if it were knocked out should be a singlet that was detected in the wild type only. Manual inspection of the one spectrum that matches Upf1p with high confidence (Table III, spectrum 1) reveals that there is a minor component at the \( m/z \) ratio expected for the c9 form of the peptide. This peak is presumably a background peak unrelated to Upf1p. Once this peptide was identified, the same peptide was sought in the corresponding IEX fractions from other experiments, using the HPLC comparison methodology described above. In some HPLC runs, an HL pair was detected that had a similar distorted ratio, but it was not identified automatically in any of the other experiments. As indicated in Table III, two additional spectra do correspond to the Upf1p peptide, though the evidence of this is weaker. In each case, the light form of the peptide is much more intense than the signals at the \( m/z \) ratio expected for the heavy form, which may belong to an unrelated HL pair.

**DISCUSSION**

*Reliability of Peptide Identification*—One of the issues that needs to be explored is the reliability of the peptide identification process. Our analyses confirm that both ProICAT and Mascot correctly determine which spectra can be classified as confidently identified. As is to be expected, as spectra of lower quality are examined, then some spectra are misidentified by each program. In most cases, we have found that incorrect identifications result because the correct identification was not considered by the search engine. This may result because the search was performed using an incorrect parent mass, an incomplete database, or because the correct identification bears a chemical modification or trypsin specificity requirement that was not considered. Currently, the Mascot search engine cannot search for all of the chemical modifications that we have identified in one round of searching. Moreover, when a large number of variable modifications are considered, or if trypsin specificity or mass tolerance is relaxed, many more incorrect identifications get assigned. To compensate for this, when these search restraints are relaxed, Mascot increases the minimal score for credible identification. Therefore, paradoxically, if purely automated methods are used, the more extensively one performs the search, the fewer correct results are returned. The solution to the paradox is that not all chemically modified peptides are equally likely to be found. Rare chemical modifications are likely to be encountered only on the most abundant peptides. In addition, manual examination of parent spectra can be performed so as to correct mistakes in the parent mass (for electrospray spectra) or to improve calibration in MALDI spectra, so that narrow parent mass ranges can be legitimately used in the searches. Nonetheless, in many cases, one has to decide between several possible peptide explanations, often a CICAT peptide and a noncanonical peptide.

This leads to the second way to assess the reliability of an identification: for an MS/MS expert to look carefully at how well the spectrum matches the peptide that is assigned to it. We have introduced many of the parameters in Table III to make this process more automatic; namely MatSc and Sc, which quantify overall reliability, as well as \( \text{yb} \), \( \text{PCM} \), \( \text{PIM} \), \( \text{FrT} \), and ppw. The higher the values for all but the last of these parameters, the more confident the identification. Similarly, SeqString enables one to discern which part of the peptide sequence is best supported by the evidence. Thus, one can gauge the reliability of each identification in Table III by examining these parameters. In most cases, one is primarily interested in the reliability of identification of a particular protein or a peptide from across the whole set of data. For that
reason, for each peptide and protein, the values of these parameters for the highest quality spectrum have been copied into the peptide and protein tables, respectively (Tables XII and XIII). Also, there are two additional parameters that separately can be used to gauge the reliability of identification; namely, the mass accuracy of the parent and the relative elution time of the peptide upon either ion exchange chromatography or reversed-phase chromatography, which are discussed below. Neither of these parameters are incorporated directly into Sc or MatSc (although mass accuracy is commonly used to limit the number of peptides to be considered in database searching). Finally, the more often a peptide is identified, the more likely the identification is correct. This is especially true when many alternative explanations have been considered. Alternative explanations can be tested by performing additional rounds of database searches directed to separate classes of chemical modifications, or by considering peptides without trypsin cleavage constraints. Only selected spectra have been subjected to these additional rounds of searches. However, these spectra routinely included the spectra that were most convincing in the identification of each protein.

**Mass Accuracy**—It is to be expected that the parent mass of every correctly identified peptide should be close to its theoretical mass. In MALDI experiments, one can substantially improve mass accuracy by performing internal calibration. In these experiments, unfortunately, no calibration standard was added directly to each MALDI spot; however, many spots contained multiple identified components, all of which ought to be consistent with each other. To improve the mass accuracy of the MALDI data, peptides that exceeded a MatSc of 10,000 were used to calibrate the spots (see above). There were 4668 identifications from MALDI that matched within 1 amu (MBC) and had MatSc > 5000 with Y2 = 1 (from any Y1, ChM, or NT category). Of these, 1952 were used as calibrants or were a member of an HL pair that was used as a calibrant. That leaves 2716 identifications whose mass accuracy can be assessed. To study the effect of mass accuracy on credibility of identification, the identifications were broken up into four classes: 1) those with Mascot scores (Sc) > 20, which are highly enriched in correct identifications due to the attention paid to questionable identifications; 2) those with MatSc > 5000 and Sc < 20; 3) those that matched at least four y or b ions and MatSc < 5000; and 4) those with four or more y or b ions. In Fig. 6, one can see that the highest confidence category also had high mass accuracy; that is, the most abundant categories were peptides that matched between 0 and 2.5 ppm and those between 2.5 ppm and 7.5 ppm. Only 22 peptides had mass accuracies that were higher than 100 ppm, which are probably due to badly calibrated spots or spots that were not calibrated at all. In the second category, an even larger number of spectra were matched to high mass accuracy, though there were also a few more peptides (53) that matched to >100 ppm. The third and fourth category of peptides had a few more identifications that

![Fig. 6. Identifications versus mass accuracy.](image-url)
matched to high mass accuracy, but a much larger number matched to >25 ppm, which indicates that many of these identifications are likely to be false. The spectra included in this analysis included all of the CICAT identifications regardless of chemical modifications or whether they adhere to trypsin cleavage rules, excluding only those matches that were >1 amu. The high percentage of high mass accuracy identifications in the second category indicates the power of the MatSc to distinguish borderline Mascot identifications from false-positive identifications. Nonetheless, the main power of MatSc is in discriminating between alternative sequences, and it probably would not be nearly as useful if it was used to identify the highest scoring sequences from all of sequence space.

Comparison of Elution Times—The data presented are derived from five independent ICAT experiments, starting from exactly the same preparations of yeast cells. To determine the reproducibility of the chromatography and to obtain more identifications, on some occasions individual ion exchange fractions were injected several times. The reproducibility of the chromatography can be estimated by selecting peptides that were identified in two different runs, and then plotting the elution times of the peptides against one another. To illustrate this, Fig. 7A shows the relationship between electrospray
The peptides that were identified both in MALDI experiment 1 fraction 15 and in electrospray experiment 2 fraction 15 are listed, sorted by spot number. These include very-low-stringency identifications (MatSc/H11022 1000; MAc/H11021 4, MBC/H11005 0, Y1/H11021 98). Spot, the MALDI spot number. Time, the elution time in minutes for electrospray data. Seq, the peptide sequence. HL, the HL ratio; M for MALDI, E for electrospray. rr, the relative ratio; HL for MALDI divided by HL for electrospray. MatSc, the Match Score for each peptide.

| No. | Spot | Time | Seq                                      | HL | rr  | MatSc |
|-----|------|------|-----------------------------------------|----|-----|-------|
| 1   | 31   | 10.76| DFGQDCDQCVVK                              | 1.022 | 0.99 | 1.04  |
| 2   | 32   | 12.58| VcSSHTGLIR                                | 0.761 | 0.79 | 0.96  |
| 3   | 33   | 11.54| DVPHDCPSLKL                                | 1.44 | 1.23 | 1.17  |
| 4   | 36   | 13.28| TYCQQHVEK                                  | 0.982 | 1.02 | 0.98  |
| 5   | 42   | 16.29| YGTcPHGcGYcGTER                            | 0.86 | 0.93 | 0.93  |
| 6   | 44   | 16.72| INEFCcSScGIR                               | 1.127 | 1.14 | 0.99  |
| 7   | 44   | 16.37| LYACISScRPGQGSR                            | 0.91 | 0.96 | 0.95  |
| 8   | 46   | 17.08| AGKLPGECcFK                                | 1.147 | 1.17 | 0.98  |
| 9   | 46   | 17.53| SCHYPNEIcVK                                | 1.091 |      |       |
| 10  | 49   | 18.73| ECINIKPQcDR                                | 1.063 | 1.07 | 0.98  |
| 11  | 49   | 20.69| HIVEcDcDMTVPGER                            | 0.8381 | 0.89 | 0.94  |
| 12  | 50   | 20.55| NLcGHcISAPYR                               | 0.716 |      |       |
| 13  | 52   | 21.15| SLVAcAGcLCcR                               | 1.278 | 1.54 | 0.83  |
| 14  | 53   | 20.93| KSAVIDNMcCK                                | 0.989 | 0.96 | 1.03  |
| 15  | 55   | 21.82| LcALDEcNEEScDKR                            | 1.256 |      |       |
| 16  | 56   | 21.87| GLcGSIcSHcQLK                              | 0.83 | 0.86 | 0.96  |
| 17  | 56   | 23.07| KNIcSISc8PR                                | 1.369 |      |       |
| 18  | 57   | 24.02| KcETIcEIQcNVK                              | 0.974 | 1.09 | 0.89  |
| 19  | 61   | 25.48| ALcHDCScLcSVR                              | 1.141 | 0.97 | 1.17  |
| 20  | 62   | 26.25| EVFcCHcSLyR                                | 0.985 | 1.05 | 0.94  |
| 21  | 63   | 27.59| EPHFDCNGcPTQcIyR                           | 1.18 | 1.52 | 0.78  |
| 22  | 63   | 26.34| VcHLcMENcEK                                | 0.969 | 1.03 | 0.94  |
| 23  | 66   | 30.15| YcNHcLcNSTLcAAcTQR                         | 0.974 | 0.94 | 1.11  |
| 24  | 67   | 27.28| NcFADcDLcHAK                               | 1.042 | 1.11 | 0.99  |
| 25  | 67   | 28.52| SEHQCcVELcScyR                             | 1.185 | 1.16 | 1.02  |
| 26  | 69   | 26.68| YNWc7DPcVHK                                | 0.768 | 0.89 | 0.89  |
| 27  | 69   | 30.04| YNWc8DPcVHK                                | 0.606 |      |       |
| 28  | 70   | 31.81| HcVDcELLAcACcANcVDK                        | 1.362 | 1.25 | 1.09  |
| 29  | 70   | 31.97| YcGYcGVcGTcLcSKEcVK                        | 1.334 | 1.34 | 0.995  |
| 30  | 71   | 32.71| ECcSNPcTCAGcGVcGFVLcANcVK                  | 0.767 | 0.73 | 1.03  |
| 31  | 77   | 34.67| ITLcHcDcLcDR                               | 4.617 | 4.47 | 1.03  |
| 32  | 77   | 36.42| YcHSALcGcTcSLcPCcKL                        | 1.007 |      |       |
| 33  | 78   | 35.7 | YNWcCDcPVHK                                | 0.834 | 0.85 | 0.98  |
| 34  | 78   | 38.53| SLcHDALcOVIR                               | 0.812 | 0.94 | 0.87  |
| 35  | 78   | 37.1 | YRPc7cPIcILcVTcR                            | 0.904 | 0.98 | 1.11  |
| 36  | 82   | 38.56| NKcSEVcAAcEACcAMcPDcLK                     | 0.917 | 0.95 | 0.96  |
| 37  | 82   | 38.77| RGcDNcIVLcLAcYcNDGcTPcNK                   | 1.094 | 1.08 | 1.01  |
| 38  | 84   | 37.39| GHEcHcFIcLr                                | 1.35 | 1.46 | 0.93  |
| 39  | 84   | 39.85| HcFGcONcDEcTcLEcQLr                       | 1.047 | 1.18 | 0.89  |
| 40  | 85   | 39.75| TcGcLcGcIcFK                                | 0.968 | 0.83 | 1.16  |
| 41  | 87   | 42.52| YRPc7cPIcILcVTcR                            | 0.968 | 1.02 | 0.95  |
| 42  | 73   | 44.09| KlcNESSEcVAcALcGcSEcISK                    | 0.899 |      |       |
| 43  | 93   | 45.43| LHcCIDcDYcDFK                              | 0.591 |      |       |
| 44  | 93   | 47.06| VcGYcTLcPSHcISTcSDcVt                    | 1.384 |      |       |
| 45  | 93   | 47.88| SWPSGcASTcGVcHEcALcEmr                    | 0.506 |      |       |
| 46  | 95   | 46.26| YACcIDLcOCFcK                              | 1.082 | 1.04 | 0.97  |
| 47  | 101  | 50.32| ATDcGAcGHcVINcVSVEcSAcAIEACcSTR            | 0.987 | 1.04 | 0.87  |
| 48  | 101  | 50.4 | VLYcQcDcLcHTc5AcASEcAR                    | 1.183 | 1.04 | 0.87  |
| 49  | 104  | 51.45| INEPTcAAcAIAYcGLcDKK                      | 4.572 |      |       |
| 50  | 105  | 51.42| KScAcVcLTcVcINEcQcQR                    | 1.144 | 1.03 | 1.11  |
| 51  | 107  | 54.14| TITPcMcGcGFcHYcGEcK                    | 4.521 |      |       |
| 52  | 108  | 52.53| KATVDEcCVLADcFDcPAcK                    | 0.903 |      |       |
| 53  | 108  | 53.08| YEFvLVcDPcOHK                              | 0.525 |      |       |
| 54  | 117  | 59.32| HAScFLcCSFcITcQcDK                     | 1.239 | 1.2 | 1.03  |
elution time and MALDI spot for fraction 15, which was chosen because it contains peptide ITLHVDcLR described above as an example of a borderline identification. There were a total of 57 peptides that were identified by both the electrospray and the MALDI runs, and they eluted in nearly the same order. This strongly suggests that each of these 57 pairs of spectra correspond to the same components. Table XIV illustrates that for four peptides, one of the identifications was tentative because the MatSc was below 5000. Note that the reliability of the each identification is no greater than the more credible of the two identifications, because although the evidence may be compelling that the same component has been detected, that component could be misidentified.

One can also estimate the reliability of the HL ratios by comparing them across the mapped peptides (Fig. 7B). As is typical for this study, most ratio measurements are close to 1.0, whereas peptide ITLHVDcLR is up-regulated about 5-fold according to both MALDI and electrospray. The ratios of the other 37 peptides were still well correlated; the correlation coefficient is 0.69 even though they are at most 2-fold different from one another. Nineteen peptides had ratios that were not automatically measurable by one technique or the other.

Chemical Modifications—Table IV lists how many spectra were matched to each class of chemical modifications. Even when only high-confidence identifications are considered, about 5% of the spectra correspond to peptides containing modifications other than oxidation of Met. It is likely that many more spectra could be assigned to each of these classes of chemical modifications with more extended searching, and there are undoubtedly additional classes of chemical modifications yet to be discovered. We conclude from these studies that in most cases these additional identifications do not add to the number of distinct peptides or proteins that can be identified, and may even eliminate borderline identifications by providing alternative explanations (21). We found that relatively few unmodified CICAT peptides were eliminated by this means. However, when Mascot was used to search for additional variable modifications or if the requirements for trypsin specificity were relaxed, then a much greater percentage of spectra were incorrectly matched to peptides that instead corresponded to chemically modified peptides.

Chemical modifications can in principle be biological in origin, or may depend on the protocols used for peptide preparation and separation. Biological modifications will be consistently observed across experiments, unless a protocol variable partially masks the modification, like differential phosphatase activity. Several of the chemical modification classes we observed in these experiments represent oxidation of side chains, the most common of which is methionine sulfoxide formation (ChM class 1). Table V columns Met and MetOx list how many identifications were attributed to peptides that contained Met as a function of experiment number. It can be seen that the extent of Met oxidation is highly variable, across both instrument type and experiment, ranging from 8 to 95% oxidized. ICAT reagent oxidation is also variable; it was much more commonly observed with electrospray data, especially when the number of MetOx identifications was large. These data indicate that some subtle variation in protocol substantially affects the degree of oxidation, perhaps involving electrospray needle field strength (22). If oxidation was eliminated, the overall peptide complexity would be reduced, making it easier to identify more proteins. It is noteworthy that a great number of oxidized peptides are difficult to identify with Mascot even when they are explicitly considered (as a variable modification), because their spectra are often dominated by neutral losses that are not accounted for, especially by cleavage just before the sulfur atom of the side chain as has been previously observed for oxidized Met (20). Fig. 4 shows spectra for an HL pair containing oxidized CICAT in which the y ions containing the modified Cys residue appear as neutral loss species.

Unmodified Peptides—In addition to the modified peptides, another complication in ICAT reagent experiments is the presence of unlabeled peptides that should have been eliminated at the avidin affinity chromatography step (Table V, column N and Table VII). These peptides are a problem in two ways: they may suppress signals for CICAT peptides, and, more importantly, they commonly have HL ratios that on first glance appear to be biologically interesting. Of the 886 spectra that were identified that corresponded to unlabeled peptides, only 37 had apparent ratios between 0.2 and 5, presumably due to unrelated co-eluting peptides. Therefore, the majority of these spectra appeared to be “singlets” and thus were not quantifiable. These spectra corresponded to 473 different peptides, about 38% of which (183) were identified more than once. At the protein level, this corresponded to only 205 proteins, 120 of which were encountered more than once. These peptides arise from incomplete washing of the avidin affinity column; note that in experiments 3 and 6 only five high-stringency identifications were made to peptides in this category (Table V), indicating that many of these peptides can be eliminated if appropriate washing protocols are followed. Unfortunately, the degree to which this is a problem becomes evident only after the experiment has been completed. Many of these proteins were also identified from CICAT peptides, but many of these proteins do not have Cys. Despite the small number of these proteins, some proteins (like enolase) are separable into two distinct isoforms by their unmodified peptides, but encode identical Cys-containing peptides, and thus could not be distinguished in a normal ICAT experiment.

MatSc Versus Mascot Score—One of the purposes of these experiments was to derive new parameters that can be used to judge the validity of database identifications. One of the strengths and limitations of the Mascot score (12) is that the significance threshold of the score depends on the size of the database that is searched and how many chemical modifications are to be considered (or nonspecific cleavage sites). In general, this is reasonable, and in fact, it is easily demon-
strated that the larger the number of chemical modifications that are considered, the more likely a spectrum that had previously been correctly identified will become matched instead to an incorrect peptide, often to a peptide with multiple chemical modifications. The significance threshold that Mascot calculates does not take into account the likelihood of the modification; instead, it assumes that any modification that is being considered is just as likely as a completely unmodified peptide. One of the main conclusions from this investigation is that as a general rule, chemically modified peptides are only likely to be correctly identified if they derive from the most abundant peptides in the sample. This is reasonable to expect if the degree of overall chemical modification is small. For this reason, it is important to determine how common each modification is within the experiment. One way to proceed is to determine the likelihood of a particular modification, and use that probability to adjust the threshold for correctness of a score (14, 23). Another alternative is to apply a different threshold for each modification (or nonspecific cleavage), as is commonly done with Sequest scores (24). Another way is to take into account which protein is being considered for modification, as suggested here.

The MatSc is no different than the Mascot score with regard to significance thresholds, and in fact it is not wise to consider that peptides whose MatSc is greater than a threshold value are likely to be correct. However, it is more useful in distinguishing peptides from one another than the Mascot score so long as the predictions as to which fragment ions should be most intense are reasonably accurate. In its current manifestation, short peptides typically get assigned higher MatSc than longer peptides with similar Mascot scores. Thus MatSc is not useful for grouping together identifications of equal confidence based on peptides with significantly different lengths. Another spectral feature that contributes mightily to both MatSc and the Mascot score is spectral quality. In many ways, each of the terms that contribute to MatSc and also the Mascot Score are independent criteria for measuring both spectral quality and identification confidence.

Biological Significance—Generally speaking, few proteins changed in expression level upon knock-out of the UPF1 gene. Upr1p, previously shown to be present at ∼1600 copies per cell (25), was automatically fragmented and identified in wild-type yeast in one of the five experiments. It was definitely lower in the knockout, but there was an isotope cluster that was not selected for fragmentation close to the noise level at the nominal mass of the heavy form of the peptide. Therefore, from the data, one could conclude only that Upr1p was down-regulated by 5-fold or greater. Of course, in theory, Upf1p should not be expressed at all in the knockout strain.

A second protein that is notably lower in the UPF1 knockout strain is the pleiotropic drug-resistance protein (PDR5p), which was detected 27 different times based on nine different peptides, based on measurements that ranged between 0.52 and 0.99 with a median of 0.69. Only 132 of 5731 measurements were below 0.69; hence ∼10% of these measurements corresponded to PDR5p. At the mRNA level, this protein is only slightly repressed (0.88; see Ref. 7). Two transcription factors are thought to control the expression of PDR5, namely PDR1 and PDR3 (26). One hexapeptide from PDR1p was detected with a ratio of 0.98, which indicates that PDR1 is not responsible for the lower expression of PDR5p in the upf1 knockout strain. PDR3p, and two other proteins, SNQ2p and YOR1p, thought to be co-regulated with PDR5p (26), were not detected. Thus it is not obvious why PDR5p is down-regulated upon upf1 knockout.

Both subunits of succinate dehydrogenase (SDH1p and SDH2p), a tricarboxylic acid cycle (TCA) enzyme, are down-regulated (0.5 and 0.7), but the other TCA enzymes are nearly unchanged. As is the case for most of the up-regulated proteins, the remaining down-regulated proteins are not obviously related to one another, whether proteins are categorized by biological process, molecular function, or cellular compartment, using the Gene Ontology (GO) system (27).

When the up-regulated proteins are considered, the arginine biosynthesis pathway stands out, as four out of five proteins in this category are observed, and all four have increased expression in the UPF1 knockout strain (Fig. 8). Twenty-four HL measurements and 11 distinct peptides map to these four proteins, and the ratios range from 1.3 to 4.9. Of these, the top four measurements all map to Cpa2p, which is the large subunit of carbamoyl phosphate synthetase. A previous study indicated that mutations in UPF1 result in increased synthesis of the small subunit of carbamoyl phosphate synthetase, in agreement with our data (28), as the genes encoding the two subunits of carbamoyl phosphate synthetase.
synthetase (CPA1 and CPA2) are known to be co-regulated in *Saccharomyces* based on other experiments (29). In addition, urea amidolysase 
(DUR1) is related to arginine metabolism, and is also up-regulated (1.83) in the *UPF1* knockout strain. Of these proteins, only CPA1 and DUR1 appear to be up-regulated at the mRNA level (7), although DAL2, DAL3, DAL5, and DAL7, involved in allantoin/nitrogen metabolism, are all up-regulated at the mRNA level.

Imidazolglycerol-phosphate dehydratase (His3p) is the most up-regulated of all proteins (HL 4.6), but it is a special case because it was encoded by the construct used to eliminate the *UPF1* gene (5), and therefore presumably appears up-regulated on that account.

Other up-regulated proteins fit into smaller categories; there are three subunits of glycine decarboxylase, two were detected (Gcv1p and Gcv2p) with ratios of 3.0 (2.96 mRNA) and 2.4 (1.62 mRNA), the third subunit has no Cys. These proteins may also be related to nitrogen metabolism (30) and therefore arginine metabolism. Another pair of proteins of related function consists of peroxisomal alkyldihyroperoxide reductase (Ahp1p) and glutathione peroxidase (Hyr1p), with ratios of 1.6 (.93 mRNA) and 2.0 (.77 mRNA), which are related to the stress response according to the GO system. There are many other proteins that appear to be up-regulated, but that are not easily explainable in terms or function.

Surprisingly, proteins whose function are similar to *UPF1* itself appear to be mostly unchanged; Dcp2p has a ratio of 0.96; Dbp2p has a ratio of 1.0 even though it partially controls nonsense-mediated decay (31), whereas Nmd2p, Upf3p (5), and Hrp1p (32) were not detected. Hhf2p, also known to be up-regulated upon deletion of *UPF1* (33), has no Cys. In general, between two and three times more proteins are positively correlated with message expression than are negatively correlated, as can be seen from Fig. 9. Additional studies will be required to determine whether the observed changes are reproducible at the biological level, and to increase the number of proteins that can be identified and quantified to determine which of these biological processes are directly related to *UPF1* function.

**CONCLUSIONS**

Using two dimensions of peptide separations, about 700 proteins were quantifiable with high confidence from yeast. Some classes of proteins, like those involved in arginine biosynthesis, appear to be up-regulated as a consequence of knocking out the *UPF1* gene. Small numbers of other proteins appear to be up-regulated or down-regulated as well. Most of the proteins thought to be directly involved in the function of *UPF1* are either undetectable or are unchanged in HL ratio. The small number of proteins whose expression levels have changed is not surprising because RNA analysis indicates that the most salient characteristic of proteins whose expression is controlled by *UPF1* is that they are proteins of intrinsic low abundance (7). Additional experiments need to be performed to determine how best to detect these changes.

Many chemical modifications of peptides were found in these experiments, as well as small numbers of peptides in which one peptide terminus did not follow the trypsin cleavage rules. However, all reliable identifications of these peptides indicated that they derived from the most abundant proteins in yeast, or were formed by biological processing of the protein’s N terminus. The higher the percentage of spectra that are correctly attributed to these noncanonical tryptic peptides, the more reliable the remaining identifications, because noncanonical peptides from abundant proteins account for some incorrect borderline identifications of single-hit proteins. This problem is not too severe in yeast because of its small genome size, but it is likely to become a serious problem in mammalian studies. The spectrum scoring parameters described here allow many of the tenuous identifications to be classified and correctly distinguished from chemically modified peptides. These additional identifications are especially useful if they provide additional HL measurements that corroborate HL ratios of proteins that are significantly altered. The set of identifications made here should make it possible to perform additional biological experiments directed at the measurement of changes in HL ratio of those proteins that can most easily be detected. Additional levels of protein fractionation prior to digestion will be necessary to delve further into the proteome.
REFERENCES

1. Aebersold, R., and Goodlett, D. R. (2001) Mass spectrometry in proteomics. Chem. Rev. 101, 269–295
2. Griffin, T. J., Goodlett, D. R., and Aebersold, R. (2001) Advances in proteome analysis by mass spectrometry. Curr. Opin. Biotechnol. 12, 607–612
3. Futcher, B., Latter, G. I., Monardo, P., McLaughlin, C. S., and Garrels, J. I. (1999) A sampling of the yeast proteome. Mol. Cell. Biol. 19, 7357–7368
4. Jacobson, A., and Peltz, S. W. (2000). Destabilization of nonsense-contain-
ing transcripts in Saccharomyces cerevisiae. In Translational Control of Gene Expression (Sonenberg, N., Hershey, J. W. B., and Matthews, M. B., ed) pp. 827–847, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
5. He, F., and Jacobson, A. (2001) Upf1p, Nmd2p, and Upf3p regulate the decapping and exonucleolytic degradation of both nonsense-containing mRNAs and wild-type mRNAs. Mol. Cell. Biol. 21, 1515–1530
6. Gyg, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using iso-
to-coded affinity tags. Nat. Biotechnol. 17, 994–999
7. He, F. L., X. Spatrick, P., Casillo, R., Dong, S., and Jacobson, A. (2003) Genome-wide analysis of mRNAs regulated by the nonsense-mediated and 5’ to 3’ mRNA decay pathways in yeast. Mol. Cell. 12, 1439–1452
8. Parker, K. C. (2002) Scoring methods in MALDI peptide mass fingerprinting: ChenScore, and the ChemApplex program. J. Am. Soc. Mass Spectrom. 13, 22–39
9. Sachs, M. S., Wang, Z., Gaba, A., Fang, P., Belk, J., Ganesan, R., Amrani, N., and Jacobson, A. (2002) Toeprint analysis of the positioning of translation apparatus components at initiation and termination codons of fungal mRNAs. Methods 26, 105–114
10. Bradford, M. M. (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal. Biochem. 196, 76–85
11. Graber, A., Juhasz, P. S., Khainovski, N., Parker, K. C., Patterson, D. H., and Martin, S. A. (2004) Result driven protocols for strategy identification and quantitation—A way to optimize experimental design and derive reliable results. Proteomics 4, 474–489
12. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence data-
bases using mass spectrometry data. Electrophoresis 20, 3551–3567
13. Tabb, D. L., Smith, L. L., Breci, L. A., Wysocki, V. H., Lin, D., and Yates, J. R., 3rd (2003) Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. Anal. Chem. 75, 1155–1163
14. Zhang, N., Aebersold, R., and Schwikowski, B. (2002) ProBID: A probabil-
istic algorithm to identify peptides through sequence database search-
ing with tandem mass spectral data. Proteomics 2, 1406–1412
15. Loo, J. A., Edmonds, C. G., and Smith, R. D. (1993) Tandem mass spec-
 trometry of very large molecules. 2. Dissociation of multiple charged proline-containing proteins from electrospray ionization. Anal. Chem. 65, 425–438
16. Breci, L. A., Tabb, D. L., Yates, J. R., 3rd, and Wysocki, V. H. (2003) Cleavage N-terminal to proline: Analysis of a database of peptide tandem mass spectra. Anal. Chem. 75, 1963–1971
17. Gras, R., Muller, M., Gasteiger, E., Gay, S., Binz, P. A., Bienvenut, W., Hoogland, C., Sanchez, J. C., Balbroch, A., Hochstrasser, D. F., and Appel, R. D. (1999) Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. Electrophoresis 20, 3535–3550
18. Fenyo, D., and Beavis, R. C. (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. Anal. Chem. 75, 768–774
19. Von Haller, P. D., Yi, E., Donohoe, S., Vaughn, K., Keller, A., Nesvizhskii, A. I., Eng, J., Li, X. J., Goodlett, D. R., Aebersold, R., and Watts, J. D. (2003) The application of new software tools to quantitative protein profiling via isootope-coded affinity tag (iCAT) and tandem mass spec-
trometry: II. Evaluation of tandem mass spectrometry methodologies for large-scale protein analysis, and the application of statistical tools for data analysis and interpretation. Mol. Cell. Proteomics. 2, 428–442
20. Steen, H., and Mann, M. (2001) Similarity between condensed phase and gas phase chemistry: Fragmentation of peptides containing oxidized Cys residues and its implications for proteomics. J. Am. Soc. Mass Spectrom. 12, 228–232
21. Davis, M. T., Spahr, C. S., McGinley, M. D., Robinson, J. H., Bures, E. J., Beiere, J., Mort, J., Yu, W., Luehly, R., and Patterson, S. D. (2001) Towards defining the urinary proteome using liquid chromatography-
tandem mass spectrometry. II. Limitations of complex mixture analyses. Proteomics 1, 108–117
22. Morand, K., Talbo, G., and Mann, M. (1993) Oxidation of peptides during electrospray ionization. Rapid Commun. Mass Spectrom. 7, 738–743
23. Keller, A., Nesvizhskii, A. I., Kolker, E., and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal. Chem. 74, 5383–92
24. Ducret, A., Van Oostveen, I., Eng, J. K., Yates, J. R., 3rd, and Aebersold, R. (1998) High throughput protein characterization by automated reverse-phase chromatography/electrospray tandem mass spectrometry. Protein Sci. 7, 706–719
25. Maderazo, A. B., He, F., Mangus, D. A., and Jacobson, A. (2000) Upf1p control of nonsense mRNA translation is regulated by Nmd2p and Upf3p. Mol. Cell. Biol. 20, 4591–4603
26. Kolaczkowska, A., Kolaczkowski, M., Delahodde, A., and Goffeau, A. (2002) Functional dissection of Pdr1p, a regulator of multidrug resistance in Saccharomyces cerevisiae. Mol. Genet. Genomics. 267, 96–106
27. Dwight, S. S., Harris, M. A., Dolinski, K., Ball, C. A., Binkley, G., Christie, K. R., Fisk, D. G., Issel-Tarver, L., Schroeder, M., Sherlock, G., Sethur-
man, A., Weng, S., Botstein, D., and Cherry, J. M. (2002) Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO), Nucleic Acids Res. 30, 69–72
28. Messenguy, F., Vierendeels, F., Pierard, A., and Delbecq, P. (2002) Role of RNA surveillance proteins Upf1/CpaR, Upf2 and Upf3 in the translational control of nonsense mRNA translation is regulated by Nmd2p and Upf3p. Mol. Cell. Biol. 20, 4591–4603
29. Kinney, D. M., and Lusty, C. J. (1989) Arginine restriction induced by delta-N-(phosphonacetyl)-L-ornithine signals increased expression of HIS3, TRPS, CPA1, and CPA2 in Saccharomyces cerevisiae. Mol. Cell. Biol. 9, 4882–4888
30. Piper, M. D., Hong, S. P., Elissing, T., Sealey, P., and Dawes, W. (2002) Regulation of the yeast glycine cleavage genes is responsive to the availability of multiple nutrients. FEM Yeast Res. 2, 59–71
31. Bond, A. N., Mangus, D. A., He, F., and Jacobson, A. (2001) Absence of Dbp2p alters both nonsense-mediated mRNA decay and rRNA process-
ing. Mol. Cell. Biol. 21, 7366–7379
32. Gonzalez, C. I., Ruiz-Echevarria, M. J., Vasudevan, S., Henry, M. F., and Peltz, S. W. (2000) The yeast InRNP-like protein Hsr1/Naib4 marks a transcript for nonsense-mediated mRNA decay. Mol. Cell. 5, 489–499
33. Welch, E. M., and Jacobson, A. (1999) An internal open reading frame triggers nonsense-mediated decay of the yeast SPT10 mRNA. EMBO J. 18, 6134–6145