Integrative Data Analysis to Improve Protein Identification in Shotgun Proteomics Experiments

Supplementary Notes

Smriti R. Ramakrishnan*[1], Christine Vogel*[2], John T. Prince[2], Zhihua Li[2], Luiz Penalva[3], Margaret Myers[1], Edward M. Marcotte[2], Daniel P. Miranker[1]

*These authors contributed equally.

Correspondence and requests for materials should be addressed to DPM (miranker@cs.utexas.edu) or EMM (edward.marcotte@gmail.com)

[1] Department of Computer Sciences, 1 University Station C0500, The University of Texas at Austin, Austin, TX 78712

[2] Center for Systems and Synthetic Biology, Department of Chemistry and Biochemistry & Institute for Cellular and Molecular Biology, 2500 Speedway, The University of Texas at Austin, Austin, TX 78712

[3] Children's Cancer Research Institute; The University of Texas Health Science Center at San Antonio; San Antonio, TX 78229

Abbreviations:

2D – two-dimensional; FDR - False Discovery Rate; FP - false positive; FPR - False Positive Rate; GFP – green fluorescent protein; LC - liquid chromatography; MS - mass spectrometry; TP - true positive

Content:

Supplementary Website: **http://marcottelab.org/MSpresso/**
Raw MS data at: **http://marcottelab.org/MSdata/**

1. Experimental details and data sources

This section describes preparation of cells and mass spectrometry samples, MS setup and data analysis, mRNA collection and analysis (for human) and references of published data used.

Table S1. Experiment setup and datasets - Overview

In each experiment, we collected protein identifications via MS/MS to form a 'test set'. We generated MSpresso protein identification scores for each 'boostable' protein: a protein that has an observed primary protein identification score (S>0) and an observed mRNA concentration (M>0). We used the 'protein reference set' as a ground-truth set to estimate MSpresso probabilities, and to determine true and false identifications during evaluation. inj – injection, i.e. technical replicate during MS/MS experiment; LCQ – LCQ DecaXP+ MS/MS instrument; ORBI – LTQ-OrbiTrap MS/MS instrument; gte2 – greater than or equal 2, i.e. requirement for protein being observed in at least two independent datasets

Experiment	Test set	Protein reference set	'Boostable' proteins
Yeast-YPD-LCQ	Cell lysate, rich medium (YPD), LCQ (5inj)	4 published MS-based datasets (gte2)	867
Yeast-YPD-ORBI	Cell lysate, rich medium (YPD), ORBI (8inj)	4 published MS-based datasets (gte2)	2136
Yeast-YMD-LCQ	Cell lysate, minimal medium (YMD), LCQ (6inj)	3 published protein datasets	886
Yeast-Fraction-LCQ	Cell lysate, fractionated in polysomal gradient, rich medium (YPD), LCQ (3inj)	Known ribosomal, translation and ribosome biogenesis proteins	270
E. coli-ORBI	Cell lysate, minimal medium (MOPS9), ORBI (3inj)	2 published 2D-gel-based datasets	860
Human-DaoyWT-LCQ	Cell lysate from Daoy, LCQ (2inj)	10 injections of same sample (ORBI)	891
Human-DaoyWT-ORBI	Cell lysate from Daoy, ORBI (1inj)	9 injections of same sample (ORBI)	515

1.1. Sample preparation for MS/MS analysis

The following general protocol was used for yeast and human. The *E.coli* sample was prepared by a protocol described by Lu et al [1]. Cells were broken using glass beads or a homogenizer, and cellular lysate was extracted by 50 min centrifugation at 5,000g. Lysis buffer consisted of 25mM Tris HCL pH 7.5, 5mM DTT, 1.0mM EDTA, 1X CPICPS (Calbiochem protease inhibitor cocktail). Protein concentration was measured and lysate diluted to 2mg/ml with buffer (50mM Tris, pH 8.0). For a typical sample preparation (~2 injections on LCQ; ~4 injections on LTQ-OrbiTrap) 50ul of diluted cell lysate was mixed with 50 µl 100% trifluoroethanol and incubated at 55C for 45min (15mM DTT). The sample was cooled to room temperature and incubated with 55mM iodoacetamide in dark for 30 min. The sample was then diluted to 1ml with buffer (50mM Tris, pH 8.0) and 1:50 w/w Trypsin was added to digest for 4.5hrs. Tryptic digest was halted by adding 20 µl in 1ml formic acid (resulting in 2% v/v). The sample was lyophilized to 20ul, resuspended in buffer C (95% H2O, 5% acetonitrile, 0.01% formic acid) and washed using a C18 tip (ThermoFinnegan). The eluted sample was again lyophilized to 10 µl, resuspended in 120 µl buffer C and filtered through a Microcon-10 filter (for 50min at 12,000g). The sample was ready for MS/MS analysis.

1.2. Yeast grown in rich or minimal medium

The yeast experimental data for the LCQ analysis was prepared as described by Lu et al. [1]. The yeast protein extraction and trypsin digestion for the LTQ-OrbiTrap analysis were performed identically to that described above (section 1.1.). For mass spectrometry analysis, eight runs were performed on an LTQ-Orbitrap varying a range of parameters for optimization. SCX salt steps were performed by injecting 10 µl of Ammonium Chloride solutions of varying molarity, namely (0, 15, 60, 900) mM or (0, 20, 100, 900) mM in a 5% ACN, 0.1% Formic Acid background onto a strong cation exchange column (Thermo BioBasic-SCX 100mm X 0.180 mm ID) with a flow rate of 800 nl/min. Reverse phase chromatography was performed on a Thermo BioBasic-18 column 100mm X 0.10 mm ID running 38 nl/min for 90 or 120 minutes with varying ACN concentrations on a background of 0.1% Formic Acid. The column eluent was nano-electro-sprayed at 1.95kV from a 10 µm tip (New Objective). FTMS resolution was set at 60,000 or 100,000. Between 6 and 10 MS/MS spectra (using 1 or 3 microscans) were acquired per MS scan. Many other individual parameters were varied including exclusion list settings, charge

state rejection, mono-isotopic precursor selection, minimum signal required for MS2 scanning, MS2 isolation width, and use of a mass lock (445.120025).

Each of the eight runs was analyzed independently with Bioworks (ThermoFinnegan), searching a database of yeast sequences. The results were combined for analysis by PeptideProphet [2] and ProteinProphet [3].

Further information on the MS/MS analyses and the raw data can be found at http://marcottelab.org/MSdata/.

Table S2. Datasets for yeast grown in rich medium (YPD)

All data sets were derived from wild-type yeast, grown at 30°C in rich medium to mid-log phase. Experimental details are provided in the respective publications. **non-membrane proteins only (some proteins in this set may not have observed mRNA abundances); *non-membrane proteins that also have observed mRNA abundances; ^ including membrane proteins that also have observed mRNA abundance; ^^ including membrane protein.

Method	Reference
RNA	
Serial analysis of gene expression (SAGE)	[5]
Single channel microarrays	[6]
Dual channel microarray with genomic DNA as reference	[7]
mRNA set (YR3gte2)	**4148***(5174^)
PROTEIN	
Flow cytometry of GFP-labeled proteins	[8]
Western blot	[9]
2D-gel electrophoresis	[10]
Non-MS-based reference (YP3)	**3191*** (3443**, 3806^, 4087^^,4097)
1D-gel electrophoresis, LC-MS on linear ion-trap FT	[11]
Multi-dimensional LC MS/MS	[12]
LC/LC MS/MS	[13]
Electron transfer dissociation (ETD) mass spectrometry (to measure phosphorylation); nanoflow-HPLC/ESI MS/MS	[14]
MS-based reference set (YP4gte2)	**1433*** (1648**)
LC/LC MS/MS LTQ-OrbiTrap	*Own data*
Union of all 8 protein reference sets (YP8)	**3582*** (3895**)

Table S3. Training and reference datasets for yeast grown in minimal medium (YMD)

All datasets are derived from wild-type yeast grown in minimal medium (YMD) to log-phase. Experimental details are described in the respective publications. * non-transmembrane proteins with mRNA abundances only; ** non-transmembrane proteins only (some of these proteins might not have observed mRNA abundances)

Method	Reference
RNA	
Single channel microarray (AFFYMETRIX)	[15]
mRNA set	**4716* (6014)**
PROTEIN	
Flow-cytometry analysis of GFP labeled proteins	[8]
Non-MS-based reference set	**1791* (1831**, 2214)**
MS/MS	[16]
MS/MS	[17]
MS-based reference set	**768* (799**, 1025)**
Union of all 3 protein sets	**1948* (2063**, 2529)**

1.3. *E. coli* grown in minimal medium

Table S4. E. coli training and reference datasets.

All datasets are derived from wild-type *E. coli* grown in minimal medium (MOPS9) to log-phase. For experimental details, please refer to the individual publications. The proteomics analysis of the *E. coli* cell lysate was conducted in a manner identical to that described above and in reference [1]. * non-transmembrane proteins that also have observed mRNA abundances

Method	Reference
RNA	
Single channel microarray	[18]
Single channel microarray	[19]
Single channel microarray	[20]
mRNA set (ERgte2)	**1769* (2470)**
PROTEIN	
2D-gel electrophoresis	[21]
2D-gel electrophoresis	[22]
ECOLI-2: Non-MS-based reference set	**370* (394)**

1.4. Human Daoy medulloblastoma cells

The medulloblastoma Daoy cell line was obtained from American Type Culture Collection (ATCC). Cells were cultured in improved minimum essential medium (IMEM) (Invitrogen, Carlsbad, CA) supplemented with 10% fetal bovine serum (Atlanta Biologicals, Inc., Lawrenceville, GA). Cells were grown until 90% confluence and then harvested to prepare RNA and protein extracts for further analyses.

RNA was extracted with Trizol (Invitrogen) and subsequently purified with an RNAeasy micro kit (Qiagen, Germany). Samples were labeled and hybridized to Nimblegen human HG18 microarrays (Madison, WI) according to their protocols. Microarrays were scanned using an Agilent Microarray Scanner G2565AA and quantified using Agilent feature Extraction Software version 9.1. Analysis of microarray data was done using the Bioconductor (www.bioconductor.org) packages marray, arrayQuality, limma and arrayMagic. Array quality was assessed using MA and other diagnostic plots. Arrays were background corrected and normalized between arrays using the quantile normalization method. Spot fluorescence intensities were then used as the measure of mRNA concentration (arbitrary units).

To prepare protein extracts, Daoy cells were re-suspended in an equal volume of lysis buffer (50mM Tris pH 8.0, 50mM NaCl, 1mM EDTA and 1 tablet of Complete Proteinase Inhibitor (Roche)/10ml) and incubated on ice for 30 minutes. Cells were then lysed with a dounce homogenizer and the soluble protein extracts obtained after centrifugation. Protein samples were prepared for mass spectrometry as described above.

Table S5. Human test and reference datasets

The proteomics and transcriptomics analysis was performed as described in section 1.1. and 1.4. * non-transmembrane proteins that also have observed mRNA abundances; * non-transmembrane proteins that also have observed mRNA abundances; **non-membrane proteins only (some proteins in this set may not have observed mRNA abundances)

Method	Reference
RNA	
Single channel microarray	*Unpublished data (L. Penalva)*
mRNA set	**9784* (13340)**
Replicate A – MS/MS analysis LTQ-OrbiTrap	http://marcottelab.org/MSdata/
HUMAN-9: PROTEIN REFERENCE	**786* (1170**, 1477)**
HUMAN-10: PROTEIN REFERENCE	**844* (1266**, 1586)**

2. Discussion of computation and statistics

2.1. Derivation of the MSpresso score

The MSpresso score is defined as the posterior probability *P(K=1|S=s, M=m)* that the protein is present in the sample having observed it in an MS/MS experiment with an identification score (*S=s*), and having observed its mRNA at particular concentration (*M=m*) under similar experimental conditions. Notation, and a partial derivation were detailed in the Methods section of the main text. The full derivation is given below.

Equation SE1

$$
\begin{aligned}
& P(K=1 \mid M=m, S=s) \\
&= P(K \mid M, S) \text{ (equivalent short notation)} \\
&= \frac{P(K, M, S)}{P(M, S)} \text{ (Bayes rule)} \\
&= \frac{P(K, M, S)}{\sum_{K=0,1} P(K, M, S)} \\
&= \frac{P(S) P(K \mid S) P(M \mid K, S)}{\sum_{K=0,1} P(S) P(K \mid S) P(M \mid K, S)} \\
&= \frac{P(S) P(K \mid S) P(M \mid K)}{\sum_{K=0,1} P(S) P(K \mid S) P(M \mid K)} \text{ (conditional independence: } (M \perp S) \mid K) \\
&= \frac{P(S) P(K \mid S) \left(P(K \mid M) P(M) \big/ P(K) \right)}{\sum_{K=0,1} P(S) P(K \mid S) \left(P(K \mid M) P(M) \big/ P(K) \right)} \\
&= \frac{P(K \mid S) P(K \mid M) \big/ P(K)}{\sum_{K=0,1} P(K \mid S) P(K \mid M) \big/ P(K)} \text{ (P(S), P(M) can be moved out of the summation, and cancelled)}
\end{aligned}
$$

2.2. Estimating *P(K)*

We use a uniform prior distribution for *P(K)*, and estimate *P(K=1)=2/3* based on observed proteins in various yeast rich-medium datasets. These datasets and their intersections (P-value<0.001, hypergeometric distribution) are illustrated in **Figure S1**. *P(K)* acts as a proportionality constant (main text, Methods, Equation 3), and changing *P(K)* does not change the relative ordering of the scores. This implies that ROC plots do not change with varying *P(K)*. However, varying *P(K)* affects the actual *values* of the MSpresso scores, and the more realistic values of *P(K)* are, the better do the MSpresso scores estimate 'true probabilities'.

2.3. Generalizing *P(K|M)*

When high-quality protein reference data is unavailable, we derive two models to generalize *P(K|M)*: SCALE_UP and SCALE_DOWN, as described in the main text. These are dubbed 'reuse' models – the *P(K|M)* distribution is learnt from a reliable dataset, generalized and reused on other datasets where protein reference data is limited, but mRNA data is available. **Figure S2(A-C)** shows the *P(K|M)* models learned from reference data for the different datasets. The general trend is towards a step-function with linear interpolation between steps as shown in **Figure S2D**, scaling the x- and y- axis results in different 'reuse' models. **Table S10** includes results on using the 'reuse' model on other datasets and organisms.

2.4. Conditional independence of *S* and *M*

The MSpresso score assumes conditional independence between *M* and *S* given *K*. In other words, the value of mRNA abundance should be independent of the value of the protein identification score, when we know the protein is (is not) in the sample. **Figures S3A** and **B** show scatter-plots for *S* and *M* for proteins that are present (or absent) from a reference set, i.e. the value of *K* is known (1 or 0, respectively). **Figure S3A** plots those proteins for which *K=1*, the proteins that are present in at least two of the four mass spectrometry based reference datasets (**Table S1**, YP4GTE2). **Figure S3B** is the corresponding figure for *K=0*, only those proteins that are absent from all four datasets.

There is little correlation between *S* and *M* for *K=0* (**Figure S3B**), though *K=1* shows a slightly stronger correlation, as revealed by bin-averaging (**Figure S3A**). Protein identification scores *S* are also only weakly correlated with abundance measurements other than mRNA concentration (*M*), i.e. protein concentration determined by Western blotting (**Figure S3C).** For these reasons, we think that it is justified to employ the conditional independence assumption, as a first simple model of the relationship between *S* and *M* on a global proteome-wide scale.

2.5. Discussion of error estimates

Protein identification methods (e.g. ProteinProphet [3], MSpresso) generate protein scores that indicate the probability that a protein is present in the sample. For practical purposes, current methods also present a list of high-confidence proteins, based on some error estimate e.g. 5% False Positive Rate. A review of different error estimates for mass spectrometry proteomics is in Choi et al. [23]

To compute such a list of high-confidence, above-threshold identifications we need a list of proteins with scores and a classical hypothesis testing framework. Our goal is to compute a significance threshold such that proteins with scores better than the threshold are marked 'significant'. To do so we must a) define the notion of a null hypothesis, and b) define a ground truth set that identifies which proteins satisfy the null hypothesis ('null' or 'decoy' proteins) and thereby also tell which proteins do not satisfy it. The null hypothesis is that the tested protein is not present in the sample. There is no such 'ground-truth' available today that describes the complete expressed proteome in yeast, and to employ a hypothesis testing framework we must first estimate this set. As described in the main text, we compile a reference set to estimate this ground truth, and assume all proteins absent from this set are 'null' or decoy proteins. In section 2.5 below we discuss alternate definitions of ground truth or 'decoy' sets and how to choose good decoy datasets.

Figure S4 illustrates the concept of true and false positive identifications**.** True and false identifications are dependent on the ground truth set and hence all error estimates should be considered *in the context of* the ground truth (reference) set. Ideally, all ground truth sets will converge to a single truth: the 'true' expressed proteome.

Once we have some notion of a ground truth to define null proteins, we can define different error estimates. A commonly used error estimate is the False Discovery Rate (FDR), which is defined as the percentage of false positive identifications amongst all identified proteins. FDR may be estimated by the quantity *FP/(FP+TP)* (**Figure S4**). The FDR is the *global* or cumulative version of what is referred to as the *local FDR*, or in Bayesian terms, the posterior error probability P(false-hit|data)[24]. Another error estimate is the False Positive Rate (used in the main paper): defined as the percentage of all false hits present among reported proteins.

FDR and FPR are global significance estimates i.e. they determine a score threshold over all proteins, and answer the question 'how many proteins are reported at x% error rate, or at score threshold=s_x (**Figure S4**). Sometimes, we are interested in per-protein measures: asking the question 'what is the smallest error rate at which this protein can still be called a significant discovery'. The q-value, introduced by Storey & Tibshirani [25] is such a measure. It is defined as the minimum FDR at which a protein with score *s* will be called significant. Note the analogy to p-values: a p-value is the minimum FPR at which a protein will be called significant. A detailed discussion can be found in Storey & Tibshirani [25].

The FDR and FPR, as defined above (see below for another definition of FDR without a ground-truth set), do not correct for multiple hypothesis testing. To correct for multiple hypothesis testing, two methods can be applied: the Bonferroni correction [26] and a less conservative method proposed by Benjamini and Hochberg [27]. We have computed the Bonferroni correction, the FDR by Benjamini and Hochberg, and q-values (software from [25]) for MSpresso and primary identification scores, using a shuffled decoy database (Section 2.6, **Table S6**). The Bonferroni correction is usually very conservative, especially for large-scale datasets in genomics and proteomics [25] and does not result in any protein identifications using our reference datasets (*not shown*). We were able to compute lists of significant proteins using the latter two measures (see **Table S6**). Analysis and comparison of these measures is part of ongoing work.

A final note on the Bayesian viewpoint: by definition, the MSpresso score is the posterior probability of protein presence given evidence of protein observation in different systems biology experiments (MS/MS and mRNA). As suggested by Kall et al. [24], our score is mathematically equivalent to (*1-posterior error probability*), reported for *each* protein.

Note that ProteinProphet [3] uses a False Discovery Rate that is computed independent of a ground truth set. Computation of their FDR estimate requires that reported protein identification score is a 'true probability', i.e. a good estimate of the 'real' posterior probability of protein presence given observed data. It is hard to prove that any score is a 'true probability' since the truth is unknown, especially when using empirical methods of probability estimation.

Since by construction the ProteinProphet score is *not* a posterior probability, the authors demonstrate that their score satisfies this requirement by showing good correlation of the score with the True Discovery Rate (TDR = TP/(TP+FP)) [3]. This is shown on one benchmark dataset, where the TDR (an unknown quantity) is estimated from a decoy database constructed by appending human proteins to the sample database with *Halobacterium* proteins: identification of a human protein is considered a false hit. With this dataset, the ProteinProphet authors demonstrate that the identification score is a true probability, i.e. the score and the TDR correlate well and the plot lies on the diagonal [3].

However, this was not the case for our reference datasets when used to construct the decoy database, and we found that the ProteinProphet score deviates from the diagonal (**Figure S5A**). Similar behavior was observed when using a shuffled decoy database (**Figures S5B, C**). We conclude that the discrepancy is due to the use of a different reference/decoy dataset than in the original ProteinProphet work [3]. For this reason, to enable fair comparison, we do not report FDR, although **Figure S5B** suggests that MSpresso scores could be used to compute accurate FDRs.

2.6. Decoy databases

The definition of an appropriate decoy database in MS-based proteomics is matter of ongoing research [23,24,28], and it is, to the best of our knowledge, still not completely resolved, especially at the level of computing significance of *protein* scores (in contrast to *peptide* scores). In fact, most research focuses on decoy databases at the peptide level. Below we discuss the methods we used to determine the null distribution: shuffled protein decoy databases [3] and control databases. We tested the yeast data using these two decoys, also trying different sizes of decoy databases (**Figure S6**). We summarize our results in **Table S6**.

2.6.1 Assessing the performance of a protein identification scheme on a decoy database

We assessed how well a particular scoring scheme (here ProteinProphet [3]) performs on a given decoy database. We perform this analysis to be able to choose the best decoy database for MSpresso analysis. We assess performance of a scoring scheme on a decoy database by examining how well the scoring scheme is able to distinguish between true ('target') and null ('decoy') hits, i.e. how well-separated the respective score distributions are (**Figure S6**). Using this criterion, a database of yeast sequences and 20-times (20x) its size of shuffled sequences performed the best. However, for practical reasons (database size, computation time) we used the decoy database with 5-times (5x) as many shuffled as 'target' sequences. Note that some hits to the target proteins may be random hits [29]; we partially correct for this phenomenon by labeling all target proteins with only a single peptide hit (single spectral count) as decoy proteins. Proteins with multiple detected peptides are more likely to be correct identifications [3].

2.6.2 Generating MSpresso scores for decoy proteins

We now describe how to generate MSpresso scores for shuffled decoy proteins. Since both decoy and target proteins have MS/MS identification scores (S), $P(K|S)$ can be estimated as usual using a logistic regression classifier, setting $K=0$ for a decoy protein and $K=1$ for a target protein. $P(K|M)$ for target proteins can be estimated as described in the main paper. However, since decoy proteins do not have mRNA abundances, we must define $P(K|M)$ distributions differently for decoys. We tried different distributions: random uniform (rand), same as the target $P(K|M)$ distribution (target), randomly sampled from the target $P(K|M)$ distribution (rand-target), constant at the minimum of the target $P(K|M)$ distribution (min-target), and (rand-target-neg): randomly sampled from the $P(K|M)$ distribution of the proteins absent from the MS-based yeast reference set (YP4GTE2, **Table S2**). We obtain moderate boosting at 5%FPR of protein identifications (7-14%) by using the '*min-target*' and '*rand-target*' methods, shown in **Table S6**. Much larger gains are observed when using different error estimates like Benjamini-Hochberg 5%FDR (FDR_BH) and q-value (17-37%) instead of FPR (**Table S6**).

Table S6. MSpresso predictions using shuffled decoy sequences.

A. 5x shuffled sequences

	P(K|M)	rand-target	min-target
AUC	ProteinProphet	0.93	0.93
	MSpresso	0.96	0.98
	Percent increase	**3**	**5**
5% FDR (Benjamini, Hochberg)	ProteinProphet	148	148
	MSpresso	244	290
	Percent increase	**65**	**96**
5% FPR	ProteinProphet	281	281
	MSpresso	300	320
	Percent increase	**7**	**14**
5% q-value	ProteinProphet	212	212
	MSpresso	249	290
	Percent increase	**17**	**37**

B. 20x shuffled sequences

	P(K|M)	rand-target	target-min
AUC	ProteinProphet	0.97	0.97
	MSpresso	0.99	1
	Percent increase	**2**	**3**
5% FDR (Benjamini,Hochberg)	ProteinProphet	187	187
	MSpresso	233	240
	Percent increase	**25**	**28**
5% FPR	ProteinProphet	250	250
	MSpresso	264	269
	Percent increase	**6**	**8**
5% q-value	ProteinProphet	203	203
	MSpresso	223	240
	Percent increase	**10**	**18**

3. Additional results for yeast grown in rich medium (YPD, LCQ)

An enlarged version of Figure 2A (main text) is shown in **Figure S7A** to illustrate that MSpresso outperforms primary identifications especially at low false positive rates. **Figure S7B** illustrates the overlap of MSpresso identifications with primary protein identifications and the two reference datasets. The FPR (false positive rate) is determined based on the protein reference dataset and set on 5% for both datasets. There is significant overlap between the three different sets (P-value<0.001, hypergeometric distribution).

3.1. Analysis *including* membrane proteins

The results in the main paper were produced excluding membrane proteins since the yeast cellular lysate was biased against membrane proteins. We observe a similar trend in results when membrane proteins are included (**Table S7**), increasing the number of identified proteins at 5%FPR by 44%. ROC plots are shown in **Figure S7C**.

Table S7. Results of MSpresso experiments <u>including</u> membrane proteins

	No. proteins (5% FPR)	AUC
ProteinProphet	241	0.74
MSpresso	347	0.88
% Increase	**44**	**19**

3.2. Analysis using a non-MS based reference set

The results in the main text use an MS-based reference set (YP4GTE2, **Table S2**). Here we present results of training MSpresso ('self' model) on the non-MS based reference set (YP3) and evaluating on the union of MS- and non-MS based sets (YP8, Table S2). There is a 66% increase in the number of reported proteins at 5%FPR, and an 8% AUC increase. ROC plot: **Figure S7D**.

3.3. Analysis on OrbiTrap

Figure S7D is a ROC plot for MSpresso analysis ('self' model) on yeast grown in rich medium, analyzed on an OrbiTrap. The number of proteins identified at 5%FPR and the AUC were given in Table 1, main text (Yeast-YPD-ORBI)

3.4. Proteins predicted by MSpresso but not by ProteinProphet ("new" predictions)

Table S8. 'New' proteins predicted by MSpresso but not by ProteinProphet (5% FPR).

The proteins correspond to those listed in Figure 2C (main text).

YKL127W	YLR167W	YEL051W	YML063W	YER003C	YDR399W	YFL045C
YCL011C	YGR034W	YPL225W	YML022W	YDR188W	YLR276C	YFL022C
YBR177C	YLR448W	YER025W	YBR286W	YPR035W	YEL009C	YJL177W
YLR179C	YGR285C	YCL009C	YPL145C	YMR315W	YNL287W	YHR068W
YGR008C	YPL037C	YHR027C	YDR071C	YCL043C	YOR157C	YKL054C
YDR394W	YGR209C	YMR303C	YGR159C	YDR258C	YGL173C	YJR121W
YLR325C	YJL130C	YDR091C	YGR037C	YDL171C	YMR120C	YJL014W
YBL085W	YKL117W	YER155C	YER136W	YEL032W	YMR318C	YCL035C
YDL058W	YMR083W	YNL045W	YJR148W	YDR212W	YMR165C	YDL147W
YNL241C	YPR033C	YDL144C	YDR129C	YDR047W	YML061C	YDR101C
YHR165C	YKR057W	YNR050C	YGL030W	YBL024W	YMR074C	YDR190C
YBR115C	YAR015W	YMR251W	YNL244C	YOR234C	YOL061W	YOR187W
YPR036W	YNL007C	YPL093W	YBL039C	YJL026W	YPL091W	YBR158W
YCL030C	YOR375C	YJL140W	YOR323C	YPR004C	YGR257C	YOR007C
YPL028W	YOR261C	YOR020C	YOR168W	YDR172W		
YBR088C	YBL002W	YNL135C	YGL148W	YDL040C		

3.5. Proteins predicted by ProteinProphet but not by MSpresso ("negative boosting")

Some proteins were identified by the original identification software at probability cutoff corresponding to 5% FPR, but fell below the threshold of MSpresso. We call this phenomenon 'negative boosting' since the protein scores are reduced below their original values, based on low mRNA abundance (see main text).

Table S9. Proteins predicted by ProteinProphet but not by MSpresso (5% FPR)

The proteins correspond to those listed in Figure 2C (main text).

YBR208C	YPL070W	YIL125W	YPL042C	YDL223C
YLR182W	YGL062W	YEL021W	YFR016C	YJL187C
YGR255C	YJL209W	YLR430W	YIL159W	YPL151C

4. Additional results for other experiments

We applied MSpresso to a number of experimental datasets (described in Table S1). **Table 1** in the main text presented MSpresso results on these datasets using the 'self' model. First, we present the corresponding ROC plots in **Figure S8, A-D**.

Next, we present results on all datasets using the 'reuse' model in **Table S10.** The reuse model is applicable when good quality training reference data may not be available to train the MSpresso model. In the reuse models, we apply the SCALE-UP model (described in the main text) to determine the $P(K|M)$ distribution, where M is

now the test-set mRNA. Unless otherwise specified, the *P(K|S)* term is estimated by applying a logistic regression classifier learnt on good quality reference data to the test-set primary protein identification score *S*. The protein reference set for each experiment (**Tables S2-S5**) is used to define true and false hits when computing ROC curves ('reuse' model ROC plots not shown) and when computing the 5%FPR cutoff (**Table S10**).

The reuse models are able to achieve an increase in reported proteins by 16 to 44%, as shown in **Table S10**. This is a considerable improvement, although not as much as reported for the 'self' models (Table 1, main text), which are trained on high-quality, experiment specific data. As expected, the best 'boosting' results were achieved with 'self' models, i.e. when good experiment-specific training data was available. Lastly, if no training data was available, re-use of the yeast model still leads to decent increases in protein identifications (**Table S10**, 16% in *E. coli*).

4.1. Yeast grown in minimal medium, analyzed on LCQ (Yeast-YMD_LCQ)

The *P(K|M)* SCALE-UP model was trained on mRNA and protein reference data from yeast grown in rich medium. Reference set used for evaluation: YMD3, **Table S3**.

4.2. *E.coli* grown in minimal medium, analyzed on OrbiTrap (Ecoli_ORBI)

The *P(K|M)* SCALE-UP model was trained on mRNA and protein reference data from yeast grown in rich medium. Reference set used for evaluation: ECOLI-3, **Table S4**.

4.3. Human data, analyzed on LCQ (HUMAN_DaoyWT_LCQ)

The *P(K|M)* SCALE-UP model was derived from human mRNA data, using a reference set derived from codon bias index values for human protein sequences – all proteins with codon bias indices (CBI) in the top two-thirds of CBI values were considered 'present' (*K=1*). *P(K|S)* was estimated differently from other datasets, it was set to the ProteinProphet probability S. Reference set used for evaluation: HUMAN-10, **Table S5**.

4.4. Human data, analyzed on OrbiTrap (HUMAN_DaoyWY_ORBI)

P(K|M) and *P(K|S)* determined as for HUMAN_LCQ. Reference set used for evaluation: HUMAN-9, **Table S5**.

Table S10. MSpresso results for 'reuse' models

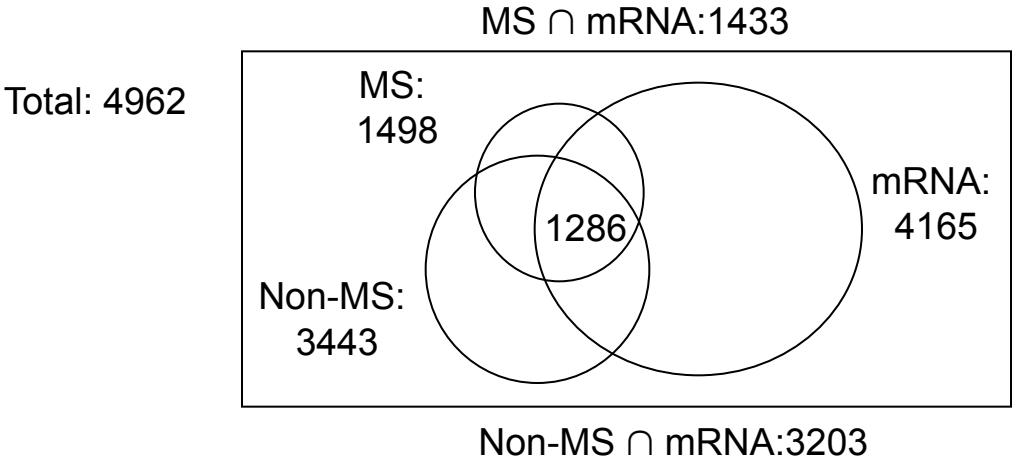
In all models the model for *P(K|M)* was reused unless denoted as 'self'. LCQ – collected on DecaXP+ LCQ; ORBI – collected on LTQ-OrbiTrap; Test set - LC-MS/MS dataset which was MSpresso-processed; Reference set – Protein reference dataset used for evaluation; inj – injection (technical replicate); AUC – area under the curve; PP – ProteinProphet (primary identification); MSp – MSpresso; * no proteins at 5%FPR, so numbers are derived from linear interpolation on ROC curve

Experiment	Test set	Reference set	AUC			Num. proteins at 5% FPR		
			PP	MSp	% increase	PP	MSp	% increase
Yeast-YPD-ORBI	ORBI-8inj	Yp4gte2	0.84	0.89	6	428*	618	44
Yeast-YMD-LCQ	LCQ-6inj	YMD-3	0.73	0.83	14	229	277	21
E.coli-ORBI	ORBI-3inj	ECOLI-3	0.69	0.8	16	63*	75*	20
Human-DaoyWT-LCQ	LCQ-6inj	HUMAN-10	0.71	0.74	4	96	111	16
Human-DaoyWT-ORBI	ORBI-1inj	HUMAN-9	0.79	0.79	0	105	104	0

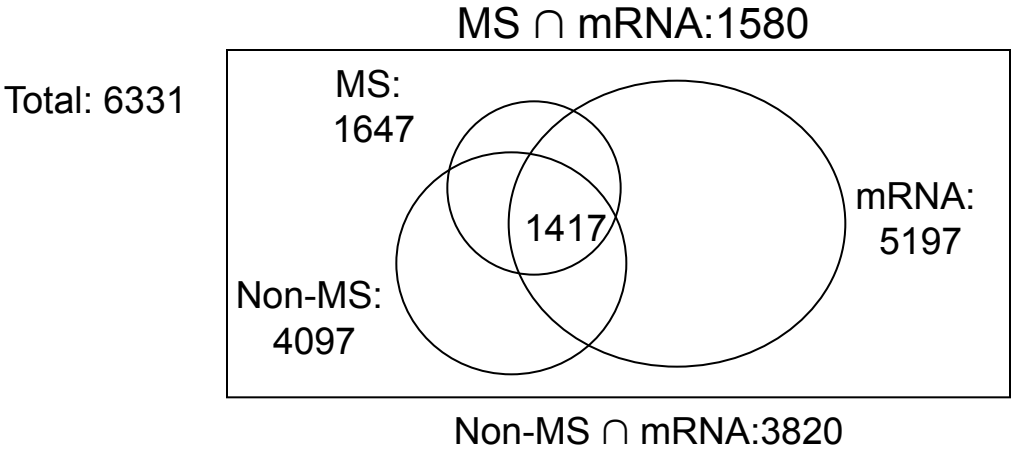
5. References

1. Lu P, Vogel C, Wang R, Yao X, Marcotte EM (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat Biotechnol 25: 117-124.
2. Keller A, Nesvizhskii AI, Kolker E, Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 74: 5383-5392.
3. Nesvizhskii AI, Keller A, Kolker E, Aebersold R (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem 75: 4646-4658.
4. Prince JT, Carlson MW, Wang R, Lu P, Marcotte EM (2004) The need for a public proteomics repository. Nat Biotechnol 22: 471-472.
5. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270: 484-487.
6. Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, et al. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. Cell 95: 717-728.
7. Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, et al. (2002) Precision and functional specificity in mRNA decay. Proc Natl Acad Sci U S A 99: 5860-5865.
8. Newman JR, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, et al. (2006) Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature.
9. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, et al. (2003) Global analysis of protein expression in yeast. Nature 425: 737-741.
10. Futcher B, Latter GI, Monardo P, McLaughlin CS, Garrels JI (1999) A sampling of the yeast proteome. Mol Cell Biol 19: 7357-7368.
11. de Godoy LM, Olsen JV, de Souza GA, Li G, Mortensen P, et al. (2006) Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. Genome Biol 7: R50.
12. Washburn MP, Wolters D, Yates JR, 3rd (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol 19: 242-247.
13. Peng J, Elias JE, Thoreen CC, Licklider LJ, Gygi SP (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. J Proteome Res 2: 43-50.
14. Chi A, Huttenhower C, Geer LY, Coon JJ, Syka JE, et al. (2007) Analysis of phosphorylation sites on proteins from Saccharomyces cerevisiae by electron transfer dissociation (ETD) mass spectrometry. Proc Natl Acad Sci U S A 104: 2193-2198.
15. Smirnova JB, Selley JN, Sanchez-Cabo F, Carroll K, Eddy AA, et al. (2005) Global gene expression profiling reveals widespread yet distinctive translational responses to different eukaryotic translation initiation factor 2B-targeting stress pathways. Mol Cell Biol 25: 9340-9349.
16. de Godoy LM, Olsen JV, de Souza GA, Li G, Mortensen P, et al. (2006) Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as model system. Genome Biol 7: R50.
17. Zybailov B, Coleman MK, Florens L, Washburn MP (2005) Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. Anal Chem 77: 6218-6224.
18. Allen TE, Herrgard MJ, Liu M, Qiu Y, Glasner JD, et al. (2003) Genome-scale analysis of the uses of the Escherichia coli genome: model-driven analysis of heterogeneous data sets. J Bacteriol 185: 6392-6399.
19. Corbin RW, Paliy O, Yang F, Shabanowitz J, Platt M, et al. (2003) Toward a protein profile of Escherichia coli: comparison to its transcription profile. Proc Natl Acad Sci U S A 100: 9232-9237.
20. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO (2004) Integrating high-throughput and computational data elucidates bacterial networks. Nature 429: 92-96.
21. Link AJ, Robison K, Church GM (1997) Comparing the predicted and observed properties of proteins encoded in the genome of Escherichia coli K-12. Electrophoresis 18: 1259-1313.
22. Lopez-Campistrous A, Semchuk P, Burke L, Palmer-Stone T, Brokx SJ, et al. (2005) Localization, annotation, and comparison of the Escherichia coli K-12 proteome under two states of growth. Mol Cell Proteomics 4: 1205-1209.
23. Choi H, Nesvizhskii AI (2008) False discovery rates and related statistical concepts in mass spectrometry-based proteomics. J Proteome Res 7: 47-50.
24. Kall L, Storey JD, MacCoss MJ, Noble WS (2008) Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. J Proteome Res 7: 29-34.
25. Storey J, Tibshirani R (2003 Aug 5) Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100: 9440 - 9445.
26. Bonferroni CE (1937) Teoria statistica delle classi e calcolo delle probabilita.: Universita di Firenza. 1-62 p.
27. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J Royal Stat Soc B 57: 289-300.
28. Choi H, Ghosh D, Nesvizhskii AI (2008) Statistical validation of peptide identifications in large-scale proteomics using the target-decoy database search strategy and flexible mixture modeling. J Proteome Res 7: 286-292.
29. Keller A, Purvine S, Nesvizhskii AI, Stolyar S, Goodlett DR, et al. (2002) Experimental protein mixture for validating tandem mass spectral analysis. Omics 6: 207-212.

A. Excluding membrane proteins



B. Including membrane proteins



Legend:	
MS	MS-based reference set
Non-MS	Non-MS-based reference set
mRNA	mRNA dataset

Figure S1. Estimating P(K)

(A) Of a total of 4962 yeast proteins without membrane helices, 3443 proteins (69%) are observed in the non-MS-based protein reference set, 1498 (30%) in the MS-based reference set. Both estimates are likely conservative given that the fraction of expressed mRNAs is even larger than 2/3 (4165 of 4962 genes; 83%). When computed over only proteins with detected mRNA abundances, the estimates are larger: *e.g.*, of 4165 total proteins without membrane helices that *also* have detected mRNA abundances, 77% are present in the non-MS based protein reference set and 34% are present in the MS-based reference set.

(B) Corresponding numbers including membrane proteins.

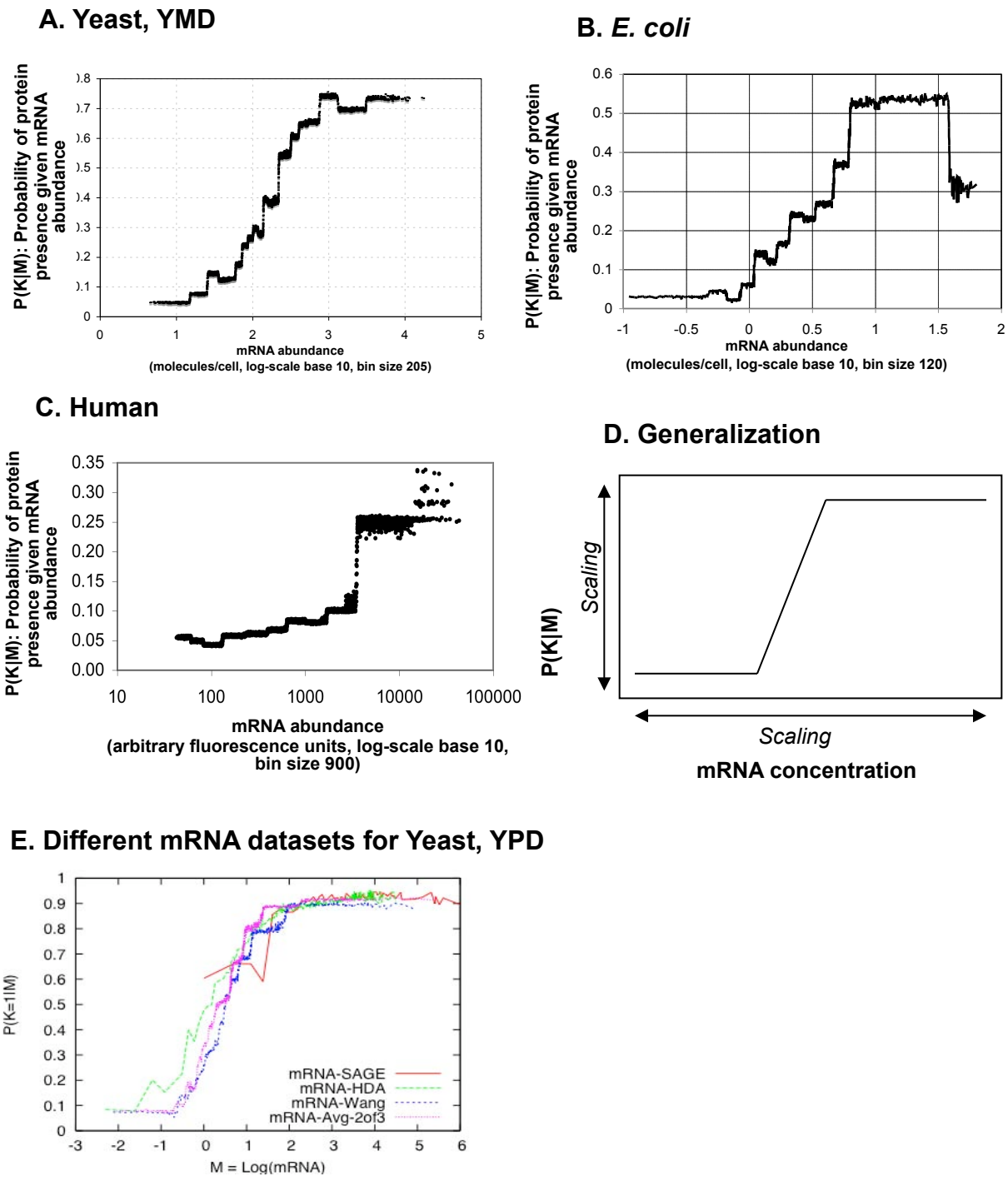


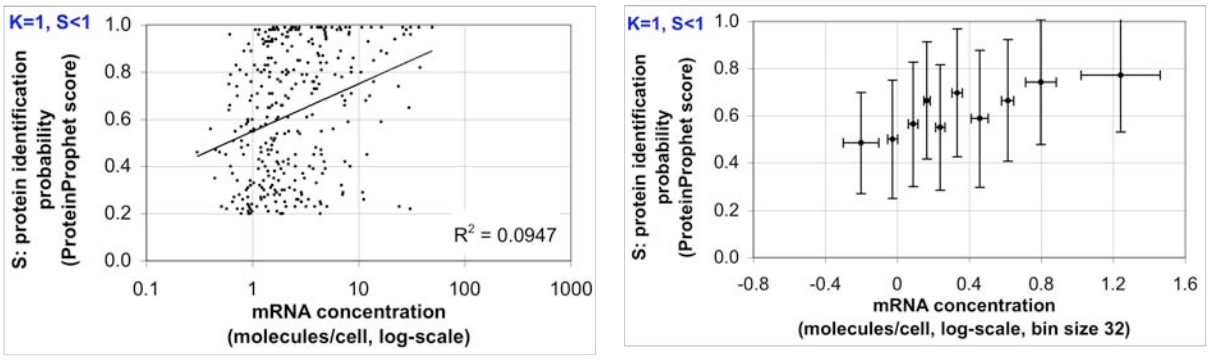
Figure S2. Estimating and generalizing P(K|M=m)

A, B, C. Experimental data describes the relationship between P(K|M) and M across different datasets: YMD, *E.coli* and Human. We plot P(K|M), the probability of protein presence given the corresponding measurement of mRNA certain abundance.

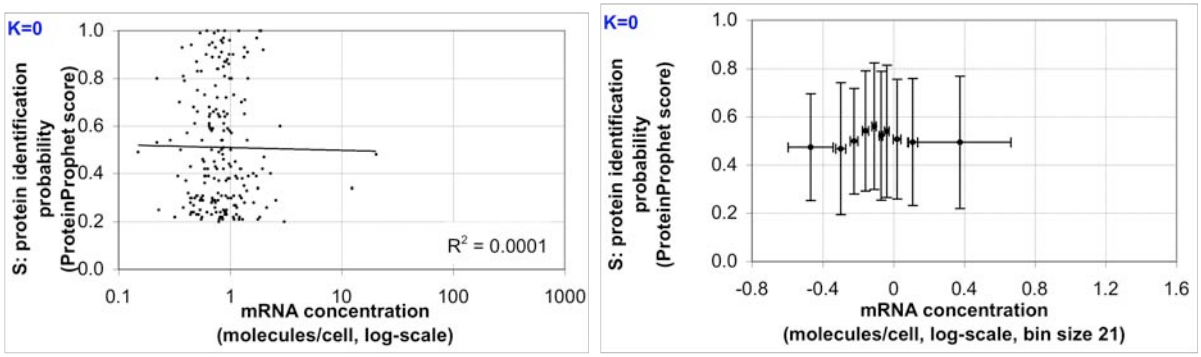
D. A generalized step-model for P(K|M) based on A, B, C and Figure 1B in the main text.

E. We use three mNRA datasets to estimate P(K|M) (Section 2.3.1, Yeast, YPD), setting M= average mRNA when at least two datasets have non-zero mRNA values, and zero otherwise. Using the average of 3 datasets allows us to overcome machine error in single datasets. The plot shows that the estimation is similar when using each individual dataset.

A. K=1 (TP according to MS-based protein reference set)



B. K=0 (TN according to MS-based protein reference set)



C. Protein abundance from Western blot data

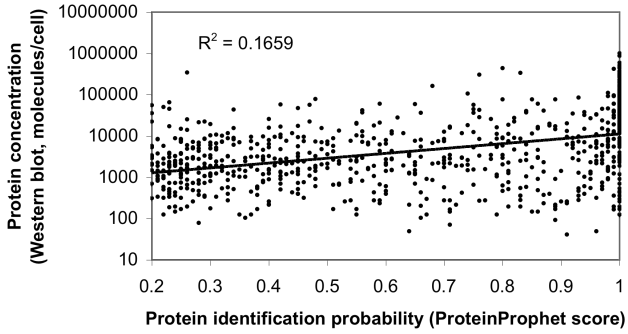


Figure S3. Independence of S and M given K

Using the MS-based protein dataset (YP4gte2) as a reference, we grouped proteins according to their absence (K=0, **A**) or presence (K=1, **B**) in the reference and plotted the primary identification score vs. mRNA concentration of the respective protein. The second plot in each panel shows the data binned to equal bin sizes. Panel **C.** shows the primary identification score vs. protein concentration measured by Western blots.

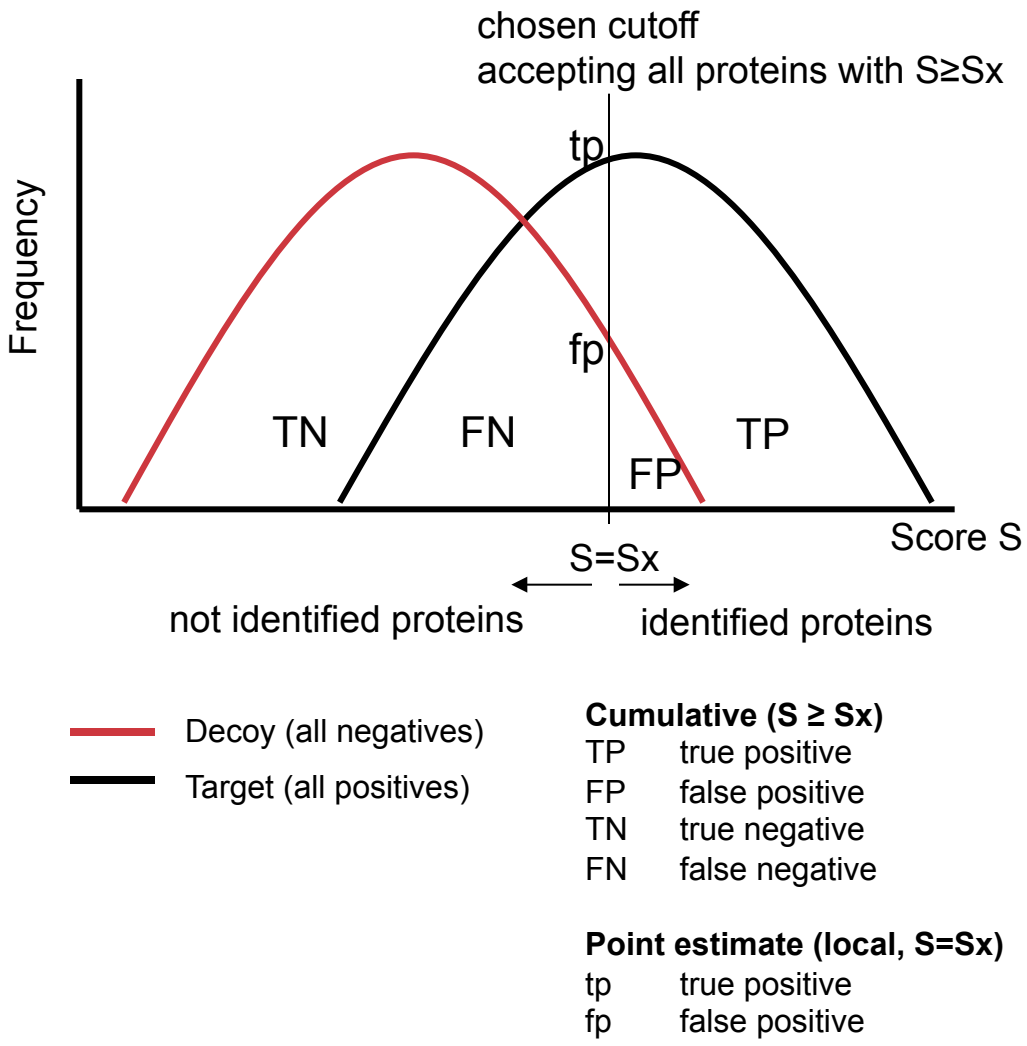
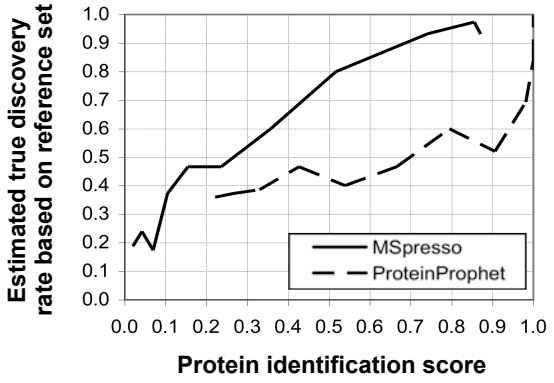
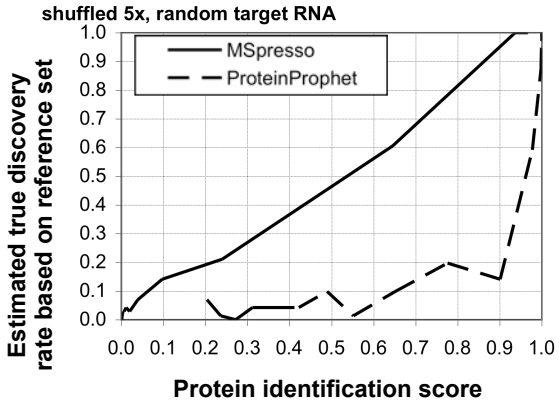


Figure S4. Illustration of true and false positives

The diagram explains the meaning of true positives, false positives, true negatives, and false negatives in the context of a decoy and target dataset. Different error estimates can be defined based on different relationships between the four terms: TP, FP, TN, FN. The area underneath the curves may be normalized to 1, resulting in a measurement of 'density' on the y-axis.

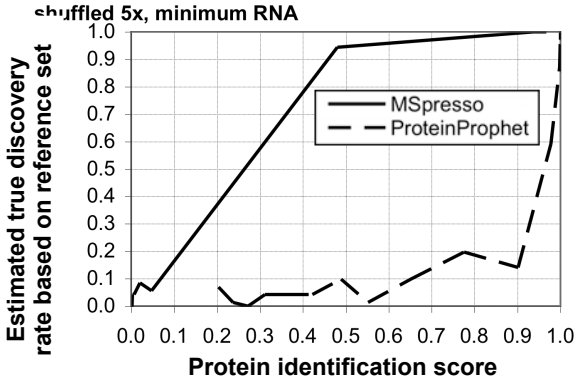


(A) Decoy= proteins absent from reference set (yp4gte2)



(B) Decoy = shuffled 5X

(B1) mRNA values randomly sampled from true (observed) values



(B2) mRNA values set to constant (minimum true (observed) value)

Figure S5. Protein identification score Vs. 'true probability' or True Discovery Rate

We ask whether the scores provided by ProteinProphet or by MSpresso are true probabilities, i.e. whether they are a good estimate of the True Discovery Rate (TDR). To test this, the score is plotted against the TDR. The closer the plot to the diagonal or the more parallel it is to the diagonal, the better the score in estimating the TDR. TDR at a score *c* is calculated as *TP/(FP+TP)*, where TP and FP are computed over proteins with scores >= *c.* In these plots, the ProteinProphet (and sometimes MSpresso) scores deviate from the main diagonal, implying they are not perfect estimates of the computed TDR I.e. they are not 'true probabilities', This observation rules out using FDR-PP (FDR as estimated in the ProteinProphet paper) as an evaluation measure for these scores, since FPR-PP requires that the scores be 'true probabilities'.

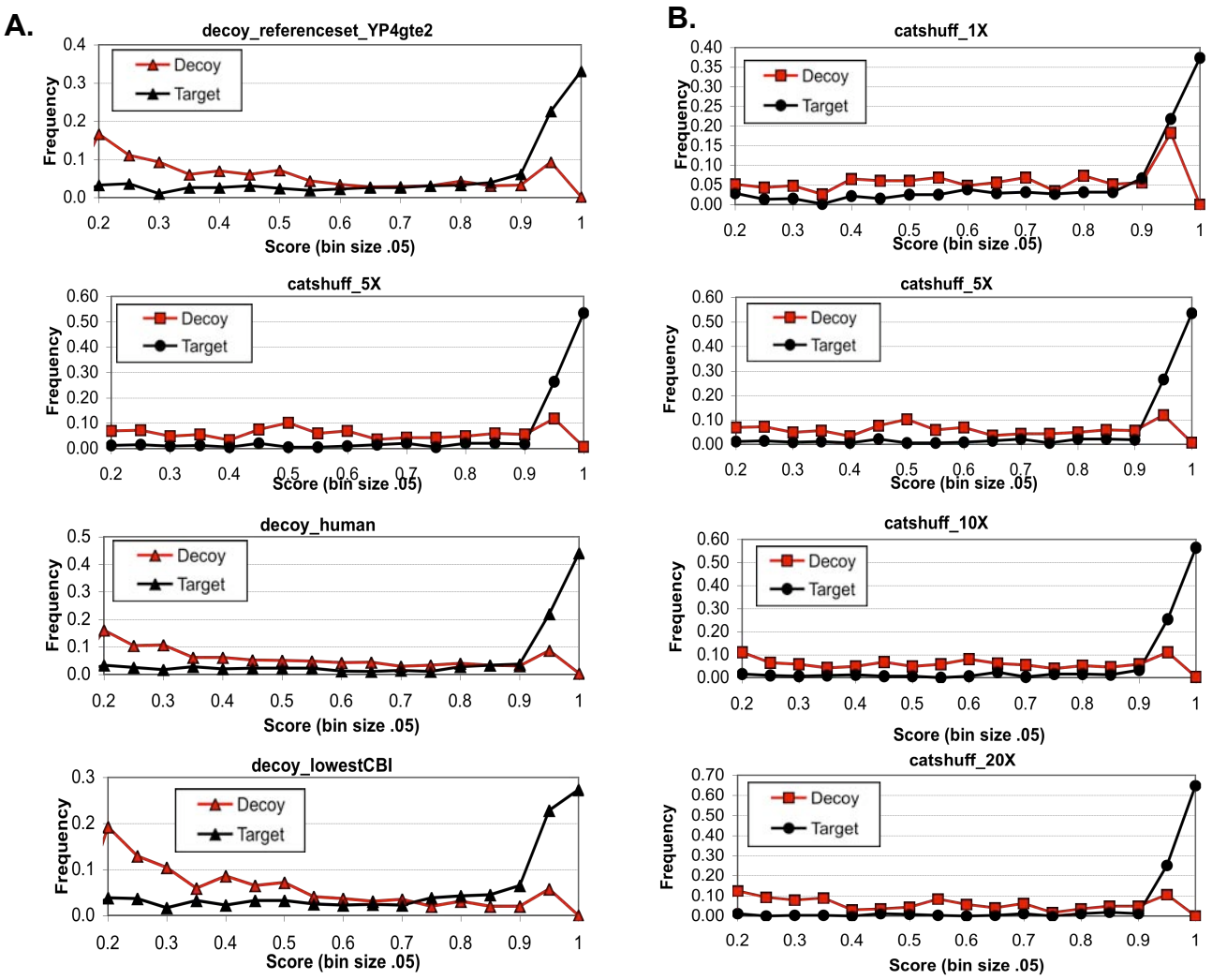


Figure S6. Assessing the performance of ProteinProphet's protein identification score on different decoy databases

Better separated 'target' and 'decoy' distributions imply better performance of the decoy database. Proteins from the yeast proteome are called 'targets' and assumed to be 'true' hits. We try four different notions of 'null' or decoy proteins **(A)** and different size of the decoy database relative to the target database **(B)**. The target yeast proteins are from yeast grown in YPD, analyzed on LCQ, and the protein identification score is from ProteinProphet.
A. Different types of decoy databases. From top to bottom the diagrams show the frequency distributions of the ProteinProphet scores obtained for decoy and target proteins defined i) the protein reference dataset YP4gte2 (**Table S1**); ii) shuffled yeast proteins (5x); iii) human proteins; and iv) the third of yeast proteins with the smallest codon bias index (CBI). The size of the human decoy database (~24,000) is approximately the size of 5 concatenated shuffled yeast proteomes – and has a similar score separation, implying similar performance when used as a decoy. Decoys defined by the experimental data (YP4gte2) have less well-separated distributions and the worst score separation is achieved by using CBI-based decoys implying it is not a very good decoy database.
B. Different sizes of shuffled protein databases: 1x, 5x, 10x, and 20x (0.25x not shown). As expected, score separation is best for large decoy databases (20x) and worst for small databases (1x). We used 5x for further analysis as it represented a good compromise between score separation (quality) and size (computation time).

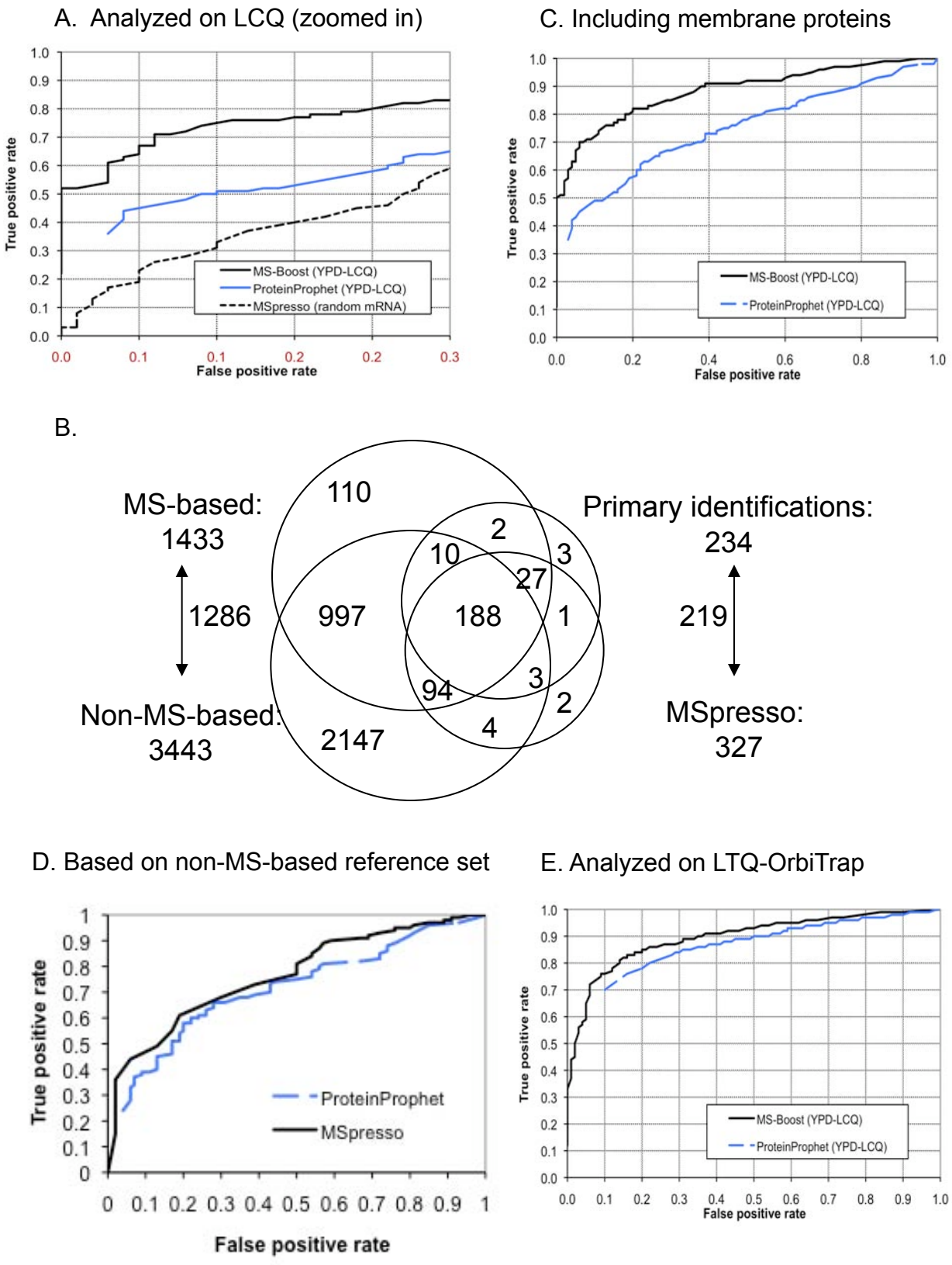
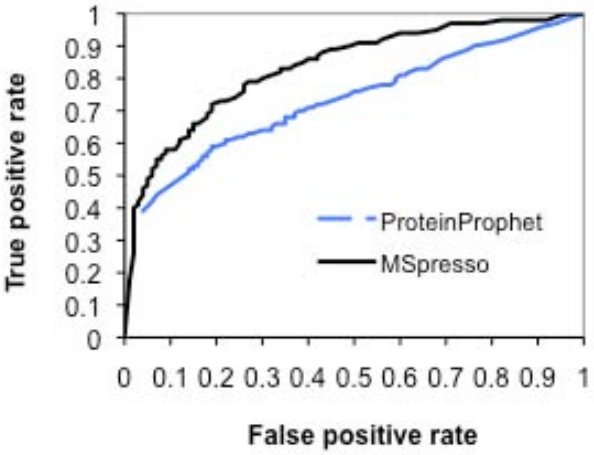


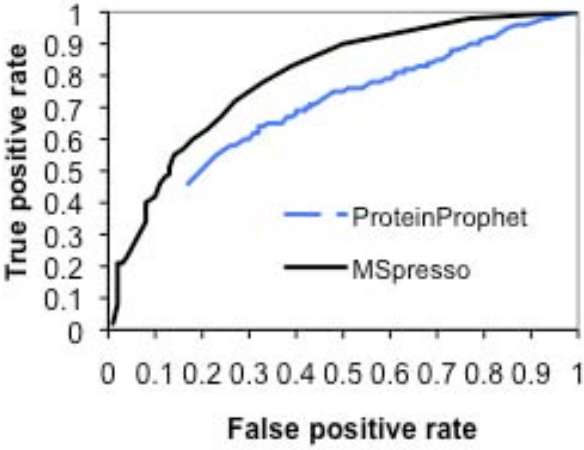
Figure S7. ROC for MSpresso analysis of yeast YPD data

We plot false vs.true positive rate as in Figure 2A (main text) for the region of low FPRs (A) and for different datasets from yeast grown in rich medium (YPD)(C-E). MSpresso outperforms primary protein identifications.

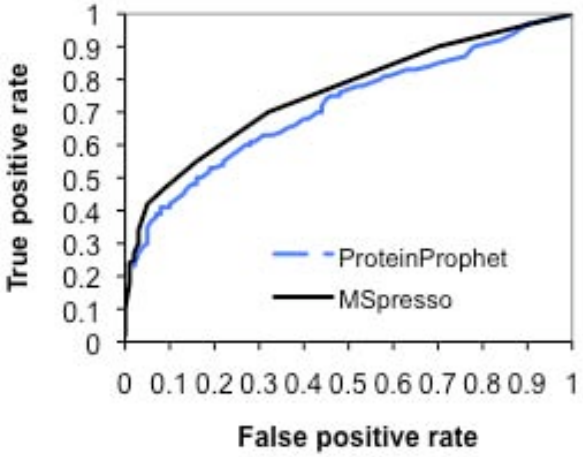
A. yeast, grown in minimal medium



B. *E. coli*



C. human, analyzed on LCQ



D. human, analyzed on ORBI

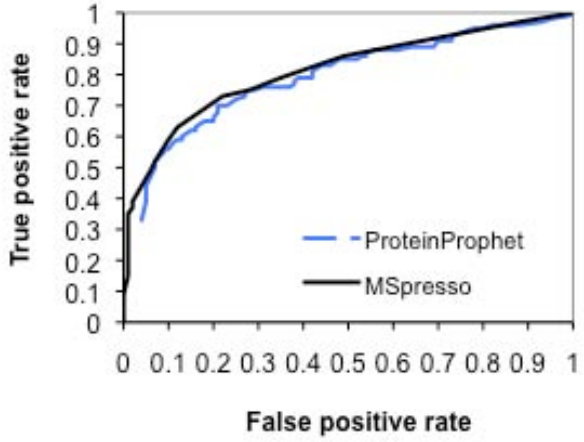


Figure S8. ROC plots for MSpresso analysis of yeast YMD, human (LCQ,ORBI) and *E.coli* data

We plot false vs.true positive rate as in Figure 2A (main text) for different datasets from yeast grown in minimal medium (A), *E.coli* (B) and human (C, D). In all four experiments, MSpresso outperforms primary protein identifications (see Table S9).