Assessment of the Yeast Proteome Repertoire

Using the Mascot Algorithm

By
Samuel J. Roth

Faculty Advisors: Michael P. Weir and Daniel Krizanc

A Dissertation submitted to the Faculty of Wesleyan University in partial fulfillment of the requirements for the degree Master of Arts

Middletown, Connecticut

May 2014
Acknowledgements

Given that I have limited space, I cannot help leaving out people who contributed to this thesis. If you are among these people, I can assure you that your contribution to my thesis and my life are greatly appreciated. Isaac Newton once said, “If I have seen further, it is by standing on the shoulders of giants.” If you are in my life, I have stood on your shoulders and, for that, I am eternally grateful.

I would like to begin by thanking Dr. Michael Weir, my thesis advisor and mentor. The past 4 years in the lab have flown by and I still remember when I was a clueless sophomore in the lab. Dr. Weir, you have always offered me help and support in every endeavor. I would not be half the thinker and scientist I am today without your continual guidance and infinite kindness. This thesis would not have been possible without you.

I would also like to thank the rest of my thesis committee. Dr. Daniel Krizanc was my computer science advisor as an undergraduate and one of my lab advisors. Danny, I am proud to have you on my committee and am grateful for your help in all of my projects. You provide a wonderful blend of insight and kindness and I would not be here without you. Dr. Fred Cohan is the chair of my thesis committee. Fred, you have always offered inspiration at every turn. Your enthusiasm and creativity shows that science can be adventurous and fun. You will always inspire me.
Thanks are due to all of my lab mates, past and present, for all of your contributions to my projects. Without your help, I certainly wouldn’t be in the position I am right now. Science is a cooperative process and it has been a pleasure and a privilege to cooperate with all of you. In particular, I would like to thank Adam Robbins-Pianka for guiding me on my initial journey towards becoming a bioinformatics scientist, Justin Cherny for being my mentor and teaching me how to think like a scientist, and Claire Fournier and Miin Lin for their help in the initial implementation of the Mascot Algorithm.

I wouldn’t be who I am today without my countless friends. Thanks to my fellow founding fathers of the Alpha Psi Chapter of the Alpha Epsilon Pi Fraternity. We did it! Thanks to all current and past brothers. You showed me that brotherhood isn’t defined by blood but rather by support and friendship. Thanks to all my teammates on the Cross Country and Track teams. You showed me that being a good person and being a good athlete are not mutually exclusive. A special thanks to Mike Bali, Michael Creager, Sam Ebb, Agbon Edomwonyi, Max Kaplan, Bryan Marsh, Richie Park, Zander Parkinson, Ariel Rejae, Aaron Rosen, and Adam Saul. You guys are my best friends in the world and have always been there when I needed you.

Thanks to the Biology Department and my fellow graduate students. You are what make the BA/MA program so enriching and fulfilling.

Finally, thank you to my family for supporting me. A special thanks are necessary for my wonderful parents. Their support for me has been unwavering,
even when we disagree. They taught me to believe in myself and that I could achieve anything through discipline and perseverance. They are the best parents an aspiring scientist could ever wish for.
# Table of Contents

Acknowledgments .................................................................................................................. I-III
Abstract ................................................................................................................................... V-VI
Introduction to Thesis ............................................................................................................... 1-6

Chapter 1: Assessment of the Mascot Algorithm with Parent-Protein Profiling
  Introduction .............................................................................................................................. 7-10
  Methods ................................................................................................................................. 10-11
  Results .................................................................................................................................... 11-19
  Discussion .............................................................................................................................. 19-22
  Figures and Tables .................................................................................................................. 22-29
  References ............................................................................................................................... 30

Chapter 2: Investigation of dnAUG Translation Initiation Using the Mascot Algorithm
  Introduction .............................................................................................................................. 31-32
  Methods .................................................................................................................................. 32-35
  Results ..................................................................................................................................... 35-40
  Discussion .............................................................................................................................. 40-42
  Future Directions .................................................................................................................... 42-44
  Figures and Tables .................................................................................................................. 45-53
  References ............................................................................................................................... 54

Chapter 3: Investigation of Translated long non-coding RNA ORFs
  Introduction .............................................................................................................................. 55-58
  Methods .................................................................................................................................. 58-61
  Results ..................................................................................................................................... 62-67
  Discussion .............................................................................................................................. 67-68
  Future Directions .................................................................................................................... 68-70
  Figures and Tables .................................................................................................................. 71-83
  References ............................................................................................................................... 84-85
Abstract

Recent studies have indicated that transcription and translation are more pervasive on a genome-wide level than previously thought (Xu et al., 2009, Nature 457: 1033; Ingolia et al., 2009 Science 324:218). Specifically, studies using ribosome profiling have shown evidence of translation at previously unannotated open reading frames (ORFs). These findings challenge current approaches to genome annotation. As a result, new methods are being developed in order to more precisely visualize the proteome. One methodology is peptide mass spectrometry wherein we perform tandem MS/MS on whole cell lysate and use a peptide search algorithm to match theoretical peptides provided by the user to observed mass spectra. To test the accuracy of these algorithms, we developed a gel slice method for parent-protein profiling in order to assess the accuracy of these algorithms (Lin et al., 2014, J. Prot. Res. 13: 1823). In Chapter 1, we used this methodology to assess the performance of the Mascot search algorithm. In Chapter 2, we applied the findings of Chapter 1 in order to assess a previously examined set of novel peptides resulting from the translation of open reading frames (ORFs) downstream of the annotated ORF (dnORFs) (Fournier et al., 2012, J. Prot. Res 11: 5712). In Chapter 3, we applied the methodologies described in Chapters 1 and 2 to detect peptides resulting from the translation of long non-coding RNAs (lncRNAs). In summary, the use of peptide mass spectrometry with peptide search algorithms can provide a high-confidence
assessment of the proteome. This enabled us to detect previously un-annotated proteins and begin to characterize them.
Introduction to Thesis

One of the central concerns of biology in the post-genomic era has been characterizing the repertoire of the proteome. The standard model of genetics dates back to Francis Crick’s “Central Dogma of Biology”. This model states that DNA is transcribed into RNA which is then translated into protein. This picture has largely been verified with a few caveats such as the existence of functional non-coding RNAs and is the current view of how genetic information is typically expressed in the cell. Newer methodologies such as transcriptome-wide RNA detection (e.g., RNA-seq and genome-wide tiled cDNA microarrays) and ribosome profiling have allowed for precise and large-scale visualization of this process.

Transcriptome-wide RNA detection provides a characterization of the transcriptome. This allows researchers to observe the products of transcription. Ribosome profiling provides a snapshot of in vivo translation. However, neither method provides a characterization of the proteome. As a result, new methods are being developed in order to more precisely visualize the proteome. This is extremely important for understanding the cellular phenotype as proteins are responsible for the majority of cellular functions. One methodology for understanding the proteome is peptide mass spectrometry. In this methodology, we perform tandem MS/MS on whole cell lysate and use a peptide search algorithm to match theoretical peptides provided by the user to observed mass
spectra. This can provide a proteome-wide assessment of expressed proteins (Aebersold and Mann, 2003, Elias et al., 2005).

The application of this methodology raises a number of issues. Firstly, how accurate are peptide search algorithms? These algorithms are critical because they perform the peptide-spectrum matching that assigns theoretical peptides to observed spectra. Secondly, how wide of a sampling of the proteome does this method provide? We have discussed methods for detecting RNA and ribosomes on a genome-wide level. One of the goals is to have a methodology that can detect proteins on a genome-wide level. Another question is whether different algorithms sample the proteome differently? Although all peptide search algorithms perform the same task, they do not perform it in the same way. Finally, are low expression proteins detected successfully? This is important because if we are to search for un-annotated proteins that are expressed at low levels, we want to ensure their detection. This thesis attempts to address these questions in the upcoming chapters.

Recently, studies have indicated that both transcription and translation are more pervasive on a genome-wide level than previously thought (Xu et al., 2009, Ingolia et al., 2009, 2012). This provides a challenge for genome annotation as studies that utilize ribosome profiling have shown evidence of translation at previously un-annotated ORFs (Ingolia et al., 2009, 2012). Current annotations underestimate the complexity of the eukaryotic proteome. For example, in *S. cerevisiae*, the standard annotation methodology is to search for
ORFs that are longer than 300 nucleotides (nt) and do not overlap with any other large ORFs (Kellis et al., 2003). One problem with this methodology is that smaller ORFs can be functional (Kondo et al., 2007). Another problem is that ORFs are identified computationally and then subsequently verified. However, there is an alternative to this approach.

In order to address this concern, we decided to use peptide mass spectrometry in order to characterize the proteome. This approach has several advantages. Firstly, it provides in vivo evidence of the products of translation. While techniques like ribosome profiling can visualize the process of translation on mRNA, it doesn’t identify the protein products of translation. Secondly, as mentioned above, it can sample the entire proteome in a high-throughput manner. For these reasons, we take this approach in characterizing the proteome.

One of the concerns in using this methodology is its accuracy. For this reason, we developed the parent-protein profiling approach to assess peptide search algorithm performance (Lin et al., 2014). In this approach, we ran cell lysate on on a gel and extracted specific size ranges (25-37, 37-50, 50-75 kDa) in the form of gel slices (Lin et al., 2014). We then performed MS/MS on those gel slices and used peptide search algorithms to perform peptide spectrum matching. In Chapter 1, we used this methodology to assess the performance of the Mascot search algorithm and also compare its performance to the SEQUEST algorithm and the Open Mass Spectrometry Search Algorithm (OMSSA).
allowed us to develop high-stringency filters such that we have high confidence in detected peptides.

In Chapter 2, we expand upon the work of Fournier et al. (2012). In that study, they used the SEQUEST algorithm in order to detect a novel set of N-terminal peptides (Fournier et al., 2012). These peptides were the result of translation initiation at start sites that are downstream of the annotated start site (called dnAUGs). We perform a similar assessment of the proteome while applying the filters we recommend in Lin et al. (2014) and Chapter 1. We found that we were able to detect a new set of peptides resulting from dnAUG translation initiation and begin to characterize these peptides.

In Chapter 3, we further expand the proteome using the methodologies from the first two chapters. We searched for ORFs that reside on long non-coding RNAs (IncRNAs). IncRNAs were previously thought to not code for any proteins. We were able to identify 436 IncRNAs that coded for short peptides. Interestingly many of these RNAs were polycistronic. We performed a preliminary characterization of these peptides in order to investigate their properties.

These sets of experiments focus on improving the detection of peptides in the proteome and applying those improvements towards identifying high confidence novel peptides. This will expand the set of detected proteins and ensure confidence in these matches. This will contribute importantly to understanding the repertoire of the proteome and the known functions of
proteins. One example of this is the discovery of novel small proteins (discussed in Ch. 3). As explained above, ORFs may be too short to be annotated by current algorithms. However, Kondo et al. (2007) discovered three small proteins (two 11 amino acid proteins and one 32 amino acid protein) that were important regulators of actin-based cell morphogenesis during development in *Drosophila*. Another example is human insulin, which is 51 amino acids long. It could be the case that many small proteins are similarly functionally significant and in some cases could begin to explain disease phenotypes if mutations occur at those loci. Thus, investigating novel peptides could provide an important first step in helping shed light on current phenotypic mysteries.
References


Chapter 1: Assessment of the Mascot Algorithm with Parent-Protein Profiling

Introduction

Peptide mass spectrometry is a useful methodology for detecting protein expression in whole cell lysate on a genome-wide scale. This approach is able to match multiple peptides to a collection of mass spectra. The use of peptide search algorithms is critical to this approach. These algorithms match a theoretical sequence “database” of proteins to observed spectra. Therefore, it is critical that the peptide matching performed by these algorithms is accurate in order to use mass spectrometry to characterize the proteome. In this set of experiments, we present a methodology that we used to assess the Mascot, SEQUEST, and OMSSA search algorithms and increase the confidence in their peptide matching. I describe first our analysis of Mascot, although similar observations were made with OMSSA and SEQUEST as described later.

The typical methodology for an experiment is to first obtain the whole cell lysate of the desired cells. Then, the proteins in that lysate are digested using an endonuclease such as trypsin. This allows the peptide fragments to be small enough to be detected by the peptide search algorithm. After this, the lysate is submitted to tandem mass spectrometry (MS/MS). During the first MS, the trypsin-digested peptide fragments are detected and then, in the second MS, they undergo collision induced dissociation (CID). CID causes further fragmentation of these peptides (usually at amide bonds resulting in N-terminal b ion and C-
terminal y ions, although a ions can occur if the cleavage is shifted by one carbon). This fragmentation yields a set of detected mass to charge (m/z) peaks in a spectrum. Peptide search algorithms compare these observed peaks to theoretical peaks based upon a set of sequences provided by the user (called a sequence “database”; usually in the form of a FASTA file) and assign matches based upon which theoretical peptides are closest to the observed spectra.

One commonly-used peptide search algorithm is Mascot. Mascot utilizes a model-based approach using statistical significance to assign matches based upon which peptide has the highest probability of matching to a given spectrum (Perkins, et al., 1999). Each match is ranked against other matches to the same spectrum and given a probability score (the rank corresponds to how good the match was in comparison to other matches to the same spectrum; i.e., a peptide with rank 1 is the best match to a given spectrum). The algorithm performs matching based upon a set of parameters determined by the user. These parameters include mass tolerances between theoretical and observed trypsin fragments (precursor mass tolerance) as well as between theoretical and CID fragments (fragment mass tolerance). However, it is also important to establish the confidence of peptide spectrum matches (PSMs) in addition to the scoring methods of the algorithm.

A standard method for assessing the accuracy of a PSM is decoy analysis (Elias and Gygi, 2007, Fitzgibbon et al., 2008). Decoy analysis measures false detection rates (FDRs) by including reverse sequences in the sequence
“database” as decoys. The scoring threshold (the minimum score at which a match is considered acceptable) is assigned based upon whether the ratio of decoy matches to non-decoy matches is below the desired FDR (e.g., 1% or 5%; calculated by # of decoys/# of all peptides matched X 100; Fitzgibbon et al., 2008). PSMs with probability scores better than this threshold are considered matches and all other PSMs are discarded.

We wanted to develop another method to assess the accuracy of the Mascot algorithm. To this end, we separated the proteins in whole cell lysate based upon their size using gel electrophoresis prior to trypsin digestion. This was done by running the lysate on a gel and extracting specific size ranges (25-37, 37-50, 50-75 kDa) in the form of gel slices (Lin, et al., 2014). Then, we performed MS/MS on those gel slices and used Mascot to perform peptide spectrum matching (this approach is outlined in Figure 1.1). We assessed those PSMs based upon whether the proteins to which the detected peptides mapped (called parent-proteins) were found in the proper size ranges. We called this approach parent-protein profiling. It allowed us to test the accuracy of matches made by the Mascot algorithm independently of FDRs and compare the performance of different algorithms and parameter settings.

We found that, when using Mascot, confidence in peptide spectrum matching is increased when a peptide is detected by multiple sets of parameter settings and when the algorithm is configured to search for a, b, and y ions (called the a/b/y ion screen) rather than the standard b/y ion screen. We also
found that some differences in performance between algorithms can be explained by the PSM ranking system (as discussed below).

**Methods**

**Implementation of Mascot**

Mascot version 2.4.0 was run on a Dell XPS 8300. Four different sets of parameter settings were used that varied in the mass tolerance, fragment tolerance, precursor ion charge, and fragment ion type (Table 1.1). As detailed in Lin et al. (2014), after gel slices were obtained they were subjected to reduction and alkylation followed by a trypsin digest. To account for this, all parameter sets included a trypsin digest and a static modification (a mass addition to the peptide) of +57 kDa for carbamidomethyl modification of cysteine residues. Optional modifications of +42 kDa for acetylation of any N-terminal amino acid residue and +16 kDa for oxidation of methionine residues were also included. We used a sequence “database” that consisted of the translated annotated open reading frames (ORFs) as well as ORFs that mapped to start sites downstream of the annotated start codon with their reverse sequences as the decoys as previously described (Fournier et al., 2012; see chapter 2).

Output from Mascot was stored in the Wesleyan IGS Database server and analyzed using a series of SQL scripts to compute scoring thresholds to yield a 5% FDR using decoy analysis (as described above and in Fournier et. al. (2012)). PSM matches were included if they had a peptide rank of 1 and the peptide was matched to only one FASTA entry (peptides of rank > 1 and peptides matching to
multiple FASTA headers were excluded). Additionally, any peptide with an internal predicted trypsin site was excluded.

**Computing Parent-Protein Conformance**

Given this parent-profiling approach, we wanted a method that could measure how well the algorithms matched peptides with parent proteins of the appropriate size. We developed the parent-protein conformance score. This score gives us the percentage of peptides found in the appropriate size range. It is computed as follows for a given MS/MS run:

\[
\text{conformation score} = \frac{\# \text{ of conforming peptides}}{\# \text{ of conforming peptides} + \# \text{ of nonconforming peptides}}
\]

In this set of experiments, we utilized 22 different MS/MS trials (detailed in Lin et al. (2014)) so we computed an overall conformance score in the following way:

\[
\text{overall conformance score} = \frac{\sum \# \text{ of conforming peptides}}{\sum \# \text{ of conforming peptides} + \# \text{ of nonconforming peptides}}
\]

A peptide was considered to be conforming to the given size range of the gel slice (size ranges listed above) if the mass of its parent-protein was within 10% of either end of the size range (e.g., for the size range 25-37 kDa, acceptable peptides had parent-proteins between 22.5 kDa and 40.7 kDa).

Overall conformance scores were computed for each set of parameter settings as well as the various assessments conducted.

**Results**
**Parameter Sets Show Similar Parent-Protein Conformance**

Mascot, like other peptide search algorithms, matches theoretical peptides to observed spectra in the form of m/z peaks obtained via CID of trypsin-digested peptide fragments. The algorithm does not take the parent-protein size into consideration when performing its matching. Using the gel slice approach outlined in Lin et al. (2014), we can test the effectiveness of this matching because we can calculate the parent-protein size of matched peptides and see whether it conforms to the appropriate size range. We utilized the conformance scoring method to evaluate the accuracy of Mascot under four different sets of parameter settings (Table 1.2).

In general, all parameter settings performed similarly (Table 1.2). In the standard b/y ion screen (where the algorithm searches for b and y ions that result from CID), the different parameter sets resulted in matches to between 4,086 and 4,952 peptides with conformance scores between 83.3 and 83.4%.

In addition to searching for the b and y ions that result from CID, Mascot can also search for N-terminal a ions (thus, performing an a/b/y ion screen). We performed an a/b/y ion screen in order to compare its accuracy to the b/y ion screen. In that screen, we found that 3,902-4,630 peptides were matched with conformance scores of 84.4-84.9% (Table 1.2). So, the performance of the a/b/y ion screen performed slightly better than the b/y ion screen (see discussion below).
These conformance scores provide an assessment of the accuracy of the algorithm when using different sets of parameter settings. However, these scores should not be considered absolute measures of accuracy for several reasons:

1. Parent-proteins may run aberrantly in the gel (Iakoucheva et al., 2001).
2. Post-translational modifications to the protein can change the molecular weights of proteins (e.g., glycosylation; Kung et al., 2009).
3. Randomly matched peptides may have parent proteins that fall within the specified size range. For example, in parameter set 0 for the b/y ion screen, decoy matches have a conformance score of 24.8%.

These concerns with the methodology likely affect each parameter set equally as evidenced by the similarity in parent-protein conformance scores (Table 1.2). However, the a/b/y ion screen had slightly higher conformance scores than the b/y ion screen (Table 1.2). For this reason, we decided to further investigate the differences in conformance of the peptides detected in each screen.

**a/b/y ion screen PSMs are a high-confidence subset of those matched in the b/y ion screen**

Given the observation that peptides matched in the a/b/y ion screen had a higher conformance than those in the b/y ion screen and there were fewer peptides matched in that screen across all parameter settings, we wanted to examine the differences in peptides matched in those screens. One explanation
for these differences could be that the peptides detected in the a/b/y ion screen represents a high-confidence subset of those detected in the b/y ion screen. To test this hypothesis, we looked at the overlap in peptides matched in both screens as well as peptides matched in a single screen.

Consistent with our prediction, peptides in the a/b/y ion screen represent a high-confidence subset of the peptides detected in b/y ion screen. All peptides detected in the a/b/y ion screen were detected in the b/y ion screen (Table 1.3). Peptides detected only in the b/y ion screen represented a small, low confidence subset. Only 184-322 peptides were matched only in the b/y ion screen and these peptides had conformance scores of 60.9%-63.0% (Table 1.3). Because this is a relatively small set of peptides (e.g., for parameter set 0, 322 out of 4,952 peptides were detected in only the b/y ion screen), it explains why the difference in conformance scores between the a/b/y ion screen and the b/y ion screen were relatively small (Table 1.2). This is consistent with our hypothesis that the peptides detected by the a/b/y ion screen are a high confidence subset of the b/y ion screen.

**Peptides detected by multiple sets of parameter settings have higher conformance scores**

One of the central goals in this assessment of peptide search algorithms was to increase both the number of peptides detected and the confidence in that detection. For this reason, we chose to run the MS/MS data using multiple sets of parameter settings. Because they each matched a different number of peptides,
we concluded that the different sets of parameter settings performed different samplings of the proteome (Table 1.2). However, while these were different samplings of peptides, many peptides were found in more than one set of parameter settings. This suggested that peptides matched in multiple sets represented higher confidence peptides and would have higher conformance scores.

We found that peptides that were matched using more than one set of parameter settings had higher conformance scores (Figure 1.2). For example, in the b/y ion screen, peptides that were matched using only one set had a conformance score of 52.0% whereas the union of all peptides detected had a conformance score of 78.6% (Figure 1.2). Thus, peptides detected in only one set of parameter settings represent a low confidence subset of all peptides detected.

Comparing the a/b/y ion screen with the b/y ion screen using multiple sets of parameter settings

As discussed above, peptides detected in the b/y ion screen alone showed relatively low parent-protein conformance (Table 1.3). Because peptides detected in only one set of parameter settings are low confidence, we wanted to see whether limiting our screens to matches found using two or more sets of parameter settings reduced the numbers of and increased the confidence of b/y only matches. To address this, we repeated the previous ion screen analysis with the added constraint that peptides had to be detected by more than one set of parameter settings.
We found that peptides detected in the b/y ion screen had higher conformance than before, although it was still much lower than those detected in both ion screens (Table 1.4). For this reason, we conclude that peptides detected in the b/y ion screen alone represent a set of low confidence peptides, even when they are detected in multiple parameter settings. These data suggest that peptides detected in the b/y ion screen alone as well as only in one set of parameter settings are low confidence matches. Therefore, searching for a/b/y ions and requiring that peptides be matched using multiple sets of parameter settings are preferred filters for using Mascot.

**Proteins with higher expression levels show higher conformance scores**

We next wanted to see what effect protein expression levels had on conformance. To address this, we looked at the parent-protein expression for all detected peptides using data obtained from genomic-scale Western blot analysis (i.e., they created epitope tags for all annotated ORFs, purified for the target proteins, and then quantified the protein expression; Ghaemmaghami et al., 2003). We found that conformance scores increased as proteins had higher expression levels. However, we noticed that the peptides that were undetected in the Ghaemmaghami et al. (2003) screen had a surprisingly high conformance score (Table 1.5). This suggests that many undetected proteins are likely expressed. This is evidenced by the observation that the conformance score of the undetected peptides (81.2%) was much higher than that of low expressing proteins (log(Protein Molecules per Cell) ≤ 3; conformance score = 53.9%).
These proteins are probably not detected due to experimental issues rather than a lack of expression (e.g., an inaccessible epitope tags when conducting the Western blot). Taken together these results indicate that peptides resulting from higher expression parent proteins are matched with greater confidence and that many proteins with undetected protein expression are likely expressed.

**Evaluating the Scoring of Mascot**

We next wanted to investigate the performance of Mascot’s scoring system using parent-protein profiling in order to independently assess the effectiveness of decoy analysis. Mascot assigns a probability score to each PSM based upon how well the theoretical peaks of a given peptide match to the experimental spectrum (Perkins, et al., 1999). We assessed the performance of this scoring system by developing a confidence score. This score was calculated as follows:

\[ d = -\log_{10}(PSM\_Prob\_Score / FDR\_threshold) \]

where \( d \) is the scoring confidence, \( PSM\_Prob\_Score \) is the probability score of the given PSM, and \( FDR\_threshold \) is the scoring threshold used to obtain the desired FDR (5% in this case). Because Mascot employs a log-based scoring system in its output, \( PSM\_Prob\_Score \) is computed using this equation:

\[ PSM\_Prob\_Score = -10 \times \log_{10}(PSM\_Prob) \]

where \( PSM\_Prob \) is the output score given by Mascot.

The confidence score provides a measure of how well a given PSM match was compared to the worst allowable match (i.e., the FDR scoring threshold). For
example, a confidence score of 3 means that the PSM is a $10^3$-fold better match than the worst match allowed. If decoy analysis effectively identified a proper scoring threshold, we should expect that PSMs with higher confidence scores should show higher conformance.

We computed the confidence score for every PSM match and grouped matches based upon their scores. We found that as the confidence score increases the conformance score also increases (Table 1.6). Most notably, if the confidence score is above 0.5 (PSM score is $\sim3$ fold better than the FDR scoring threshold), conformance is greatly increased (confidence $\leq 0.5$ had a conformance score of 58.6% whereas $0.5 < \text{confidence} \leq 1$ had a conformance score of 76.2%, Table 1.6). This suggests that we can have higher confidence in matches that have a confidence score greater than 0.5 and this could serve as a potential filter.

**Decoy detection differs among algorithms**

A similar analysis was performed on the other two peptide search algorithms (OMSSA and SEQUEST). One observation that we noted was that conformance scores differed between peptide search algorithms. SEQUEST, one algorithm, had similar conformance scores to Mascot (84.4-84.5%) while OMSSA, another algorithm, had higher scores ($\sim4\%$ higher across parameter settings) (Lin, et al., 2014). We first wanted to assess whether these differences between these algorithms (SEQUEST and OMSSA) were significant.
We performed a bootstrap analysis and the resulting score distributions indicated that these differences were indeed significant (Figure 1.3). Due to this difference in algorithm performance, we hypothesized that there must be a difference in the detection of decoys and the FDR. We found that prior to calculating a scoring threshold for a 5% FDR OMSSA had FDRs between 1.6% and 3.9% for the various parameter sets. This indicated that the standard implementation of OMSSA is more stringent in its PSM matching than SEQUEST and Mascot.

Additionally, OMSSA does not output a ranking of PSMs (as described above) while SEQUEST and Mascot do. If we only consider the best-scoring peptide for each spectrum, OMSSA yields FDRs between 0.6% and 1.7%. Because we only considered SEQUEST matches having rank 1 (i.e., the best-scoring peptide for each spectrum), we decided to perform the bootstrap analysis while giving SEQUEST a 0.6% FDR (Figure 1.3A). We also performed an analogous bootstrap analysis while giving Mascot parameter set 0 the same FDR (Figure 1.3B). We found that when a 0.6% FDR is applied to SEQUEST and Mascot the difference in conformance score when compared to OMSSA is not significant (Figure 1.3). This indicates that the standard implementation of OMSSA has more stringent filtering than the standard implementations of SEQUEST and Mascot, but differences in algorithm performance disappear when the filtering is made comparable.

Discussion
**Parent-Protein Profiling as a tool for evaluating algorithms**

As peptide search algorithms continue to improve, methodologies like the one presented in these experiments will become increasingly important. In this study, we outlined how parent-protein profiling was used to analyze one such algorithm, Mascot. We were able to evaluate the algorithm under various parameter settings in an unbiased manner that was independent of individual FDR scoring thresholds.

In particular, we found that the four sets of tested parameter settings performed similarly and yet gave different samplings of detected peptides. We also observed that peptides matched in multiple sets of parameter settings had higher confidence than those matched in a single set. Additionally, we found that the a/b/y ion screen represents a high confidence subset of the b/y ion screen and that this pattern is maintained when peptides are matched in multiple sets of parameter settings.

This is particularly interesting as it distinguishes the matching behavior of Mascot as compared to both OMSSA and SEQUEST. In contrast to Mascot, matches to only a/b/y ion screens and b/y ion screens are still fairly high confidence in OMSSA (Lin et al., 2014). With SEQUEST, matches found in only the a/b/y ion screen or the b/y ion screen represent low confidence subsets of matches (Lin et al., 2014). It is only when using Mascot that there are no matches detected in the a/b/y ion screen alone. We therefore recommend that the Mascot protocol be changed so that the algorithm searches for a/b/y ions rather than
b/y ions in order to have the highest confidence matches. To further increase confidence in matching, only peptides matched in multiple sets of parameter settings should be considered.

We also demonstrated that matches to higher expressing proteins represent higher confidence matches. This made sense because proteins that are more highly expressed will have more abundant spectra for the peptide search algorithms to match. Finally, we observed that matches that are more than ~3 fold better than the FDR scoring threshold have much higher confidence than those that are closer to the FDR scoring threshold.

In summary, we suggest that this type of analysis can assess peptide search algorithm performance. For example, we were able to better understand differences in decoy detection and filtering between algorithms by examining differences in parent-protein conformance between algorithms. Further, we were also able to outline a set of filters when using Mascot (outlined above) based upon which filters could optimize performance. However, one of the concerns of more stringent filtering is ruling out potential matches that deemed false negatives by our filters. Given our observation that lower expression proteins show lower conformance, by ruling out sets of peptides that have low conformance (i.e., for Mascot, peptides matched in only one set of parameter settings and in the b/y ion screen only), we could be potentially creating false negatives for low expressing proteins. While this is a concern for accurate peptide detection, we believe that more stringent filters are prudent when
detecting previously un-annotated proteins (see Ch. 2 and Ch. 3 as examples). In this situation, we want to be sure that we have confidence in these novel peptides and added filtering ensures this confidence. In all, these assessments can help optimize algorithm performance such that confidence in peptide matching can be substantially increased.
Figures and Tables

Figure 1.1. A general outline of the gel slice approach.
Figure 1.2. Peptides detected in multiple parameter settings have better conformance scores than peptides detected in a single parameter setting. (A) b/y ion screen. (B) a/b/y ion screen.
Figure 1.3. Bootstrap analysis of parent-protein conformance. 3,060 peptides were randomly selected with replacement from OMSSA run 0 using a 5% FDR. (A) At a 5% FDR, the conformance score for SEQUEST is below the 1%
distribution threshold, however, at a 0.6% FDR, the conformance score for SEQUEST falls within the 1% to 99%. (B) Similarly, for Mascot run 0, at a 5% FDR, the conformance score is below the 1% of the distribution threshold, but, at a 0.6% FDR, the conformance score falls within the distribution.

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max. missed cleavage sites</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Precursor mass tolerance (Da)</td>
<td>3.0</td>
<td>1.5</td>
<td>2.0</td>
<td>1.5</td>
</tr>
<tr>
<td>Fragment mass tolerance (Da)</td>
<td>1.0</td>
<td>0.5</td>
<td>0.8</td>
<td>0.5</td>
</tr>
<tr>
<td>Precursor ion search type</td>
<td>mono</td>
<td>mono</td>
<td>avg.</td>
<td>avg.</td>
</tr>
<tr>
<td>Fragment ion search type</td>
<td>mono</td>
<td>mono</td>
<td>avg.</td>
<td>avg.</td>
</tr>
<tr>
<td>Mass tolerance charge scaling</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Table 1.1. Mascot parameter settings that were evaluated. These parameter settings were similar to those evaluated for OMSSA in Lin et al. (2014).
Table 1.2. (A) Conformance scores for the b/y ion screen. (B) Conformance scores for the a/b/y ion screen.

<table>
<thead>
<tr>
<th>Mascot Parameter Set</th>
<th>Distinct Peptides</th>
<th>Conforming Peptides</th>
<th>Nonconforming Peptides</th>
<th>Overall Conformance Score</th>
<th>Overall Decoy Conformance Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>4,952</td>
<td>4,123</td>
<td>829</td>
<td>83.3</td>
<td>24.8</td>
</tr>
<tr>
<td>1</td>
<td>4,404</td>
<td>3,674</td>
<td>730</td>
<td>83.4</td>
<td>16.2</td>
</tr>
<tr>
<td>2</td>
<td>4,086</td>
<td>3,409</td>
<td>677</td>
<td>83.3</td>
<td>22.3</td>
</tr>
<tr>
<td>3</td>
<td>4,649</td>
<td>3,878</td>
<td>771</td>
<td>83.4</td>
<td>23.1</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>4,630</td>
<td>3,927</td>
<td>703</td>
<td>84.8</td>
<td>27.2</td>
</tr>
<tr>
<td>1</td>
<td>4,130</td>
<td>3,506</td>
<td>624</td>
<td>84.9</td>
<td>17.8</td>
</tr>
<tr>
<td>2</td>
<td>3,902</td>
<td>3,293</td>
<td>609</td>
<td>84.4</td>
<td>21.4</td>
</tr>
<tr>
<td>3</td>
<td>4,366</td>
<td>3,703</td>
<td>663</td>
<td>84.8</td>
<td>24.4</td>
</tr>
</tbody>
</table>

Table 1.3. Conformance of peptides detected by a/b/y ion screen only, both screens, and the b/y ion screen only. There were no peptides detected in only the a/b/y ion screen and those peptides detected in only the b/y ion screen showed relatively low parent-protein conformance.

<table>
<thead>
<tr>
<th>Mascot Parameter Set</th>
<th>a/b/y Ions</th>
<th>a/b/y Ions &amp; b/y Ions</th>
<th>b/y Ions</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>N/A (0)</td>
<td>84.8 (4,630)</td>
<td>60.9 (322)</td>
</tr>
<tr>
<td>1</td>
<td>N/A (0)</td>
<td>84.9 (4,130)</td>
<td>61.3 (274)</td>
</tr>
<tr>
<td>2</td>
<td>N/A (0)</td>
<td>84.3 (3,902)</td>
<td>63.0 (184)</td>
</tr>
<tr>
<td>3</td>
<td>N/A (0)</td>
<td>84.8 (4,366)</td>
<td>61.8 (283)</td>
</tr>
</tbody>
</table>
Table 1.4. Conformance of peptides detected in multiple sets of parameter settings in a/b/y ion screen only, both screens, and the b/y ion screen. Consistent with previous analysis in Table 3, peptides detected in only the b/y ion screen had relatively low parent-protein conformance. Unlike Table 3, all peptides analyzed in Table 4 were detected by more than one set of parameter settings.

<table>
<thead>
<tr>
<th>Mascot Parameter Set</th>
<th>a/b/y ions</th>
<th>a/b/y ions &amp; b/y ions</th>
<th>b/y ions</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>N/A (0)</td>
<td>86.3 (4,364)</td>
<td>68.5 (260)</td>
</tr>
<tr>
<td>1</td>
<td>N/A (0)</td>
<td>87.3 (3,757)</td>
<td>78.9 (199)</td>
</tr>
<tr>
<td>2</td>
<td>N/A (0)</td>
<td>85.8 (3,786)</td>
<td>63.5 (197)</td>
</tr>
<tr>
<td>3</td>
<td>N/A (0)</td>
<td>85.7 (4,237)</td>
<td>65.2 (244)</td>
</tr>
</tbody>
</table>

Table 1.5. Conformance varies with protein expression. For proteins with detected expression levels, proteins with higher expression levels showed higher parent-protein conformance. Undetected proteins showed relatively high conformance. This suggests that many of these proteins are expressed at relatively high levels even though they are not detected by current methods. Protein expression levels were obtained from genome-wide Western blot analysis conducted by Ghaemmaghami et al. (2003).
Table 1.6. Peptides with better PSM probability scores relative to the FDR scoring threshold showed better conformance scores. Confidence scores were calculated using the following equations:

\[ d = -\log_{10}(\text{PSM}_\text{Prob}\_\text{Score} / \text{FDR\_threshold}) \]

\[ \text{PSM}_\text{Prob}\_\text{Score} = -10 \times \log_{10}(\text{PSM}_\text{Prob}) \]
References


Chapter 2: Investigation of dnAUG Translation Initiation Using the Mascot Algorithm

Introduction

In the era of high-throughput genomics, characterizing the size and complexity of the proteome has become one of the main challenges to understanding the cell. Recent studies suggest that there are more translated open reading frames (ORFs) and more sites of translation initiation than previously annotated in the genome of budding yeast, *Saccharomyces cerevisiae* (Ingolia et al., 2009, Fournier et al., 2012). One example, outlined in Fournier et al. (2012), is initiation at AUG codons located downstream (dnAUGs) of the annotated AUG codon (annAUG). This translational event would certainly add to the complexity of the yeast proteome by producing truncated in-frame proteins as well as novel out-of-frame proteins.

However, the task of characterizing the proteome has become easier as high-throughput technologies are developed. For example, Ingolia et al. (2009) developed a technique called ribosome profiling that allows for precise visualization of translating ribosomes in vivo. Another method that is useful for characterizing the proteome is peptide mass spectrometry. Submitting trypsin-digested cell lysate to tandem mass spectrometry (MS/MS) provides a proteome-wide assessment of the repertoire of expressed proteins (Aebersold and Mann, 2003, Elias et al., 2005).
One of the most important steps in peptide mass spectrometry is matching theoretical protein sequences (usually in the form of a FASTA file; called a sequence “database”) to observed mass spectra. This function is performed by peptide search algorithms. Two such algorithms are the Open Mass Spectrometry Search Algorithm (OMSSA) and the Mascot algorithm. In combination with MS/MS, these algorithms allow for the detection of both annotated and previously un-annotated proteins. They represent a powerful tool for helping to better understand the proteome.

As mentioned above, Fournier et al. (2012) found that many dnAUGs are sites of translation initiation. They utilized peptide mass spectrometry as outlined above. However, they used the SEQUEST algorithm (another peptide search algorithm) to perform peptide-spectrum matching (Fournier et al., 2012). In this set of experiments, we expand their analysis by using both OMSSA and Mascot for peptide-spectrum matching. Further, we use methods developed in Lin et al. (2014) in order to increase confidence in our matching. Finally, we characterize the detected dnAUGs and compare them with detected annAUGs. In all, we add to and clarify the findings presented in Fournier et al. (2012).

**Methods**

**Protein Search Space**

A FASTA sequence “database” was created for our set of theoretical proteins that are used for peptide-spectrum matching. This sequence “database” included the entire annotated proteome as well as all potential ORFs that would
result from translation at a dnAUG that was within 100 nucleotides (nt) of the annAUG (created in the Wesleyan IGS database using SQL scripts). We also included proteins that would result from N-terminal cleavage due to aminopeptidases (this can occur if the N-terminal methionine is followed by an alanine, a cysteine, a serine, a threonine, a glycine, a valine, or a proline; Chen et al., 2002) as well as proteins that underwent cleavage due to a signal peptide (predicted by the SignalP algorithm version 3.0; Bendtsen et al., 2004).

Additionally, we included reverse sequences for all proteins to serve as decoys. We used these decoy sequences to perform decoy analysis as outlined in Elias and Gygi (2007) as well as Fitzgibbon et al. (2008). Decoy analysis enables the calculation of false detection thresholds (FDRs). We used an FDR of 5% (i.e., # of decoys/# of all peptides matched X 100 ≤ 5) to determine our scoring threshold (the minimum score at which a match is considered acceptable). Peptide-spectrum matches (PSMs) with probability scores worse than this threshold were discarded.

**Peptide Matching by MS/MS Search Algorithms**

OMSSA was run on the Wesleyan University Computing Cluster and Mascot was run on a local Dell XPS Server using the Mascot Daemon program. Five different sets of parameter settings were used for OMSSA and four different sets of parameter settings were used for Mascot (Table 2.1, Table 2.2). We used both publicly available (PeptideAtlas Repository) and locally curated (Fournier et al., 2012) mass spectrometry data. All mass spectrometry experiments
included chemical modifications to peptides in order to strengthen detection (e.g., Fournier et al. (2012) subjected peptides to glutaraldehydation, which added 68 kDa to the N-termus) and trypsin digestion so that peptides were small enough to be detected by the mass spectrometer. For this reason, we tuned each set of parameter settings with both optional and static peptide modification specific to the data set being run (listed in PeptideAtlas.org and Fournier et al. (2012)). Additionally, each set of parameter settings included a trypsin digestion. We ran the data sets on both OMSSA and Mascot using these sets of parameter settings and the resulting output was stored in the Wesleyan IGS database server.

After peptide-spectrum matching was performed, all matches were subject to filtering using decoy analysis to obtain a 5% FDR as outlined above. We also discarded all matches where a given peptide matched to multiple FASTA entries. When different peptides matched to the same spectrum in OMSSA, we discarded all matches. Also, we only considered the best peptide matches to a given spectrum in Mascot (i.e., entries had pep_rank = 1).

These matches were compared to the matches from SEQUEST that were obtained in Fournier et al. (2012) in order to compare algorithm detection (Figure 2.1). After this comparison, we subjected matches from OMSSA and Mascot to further filtering. Previous work in our lab demonstrated that peptides that were matched in multiple sets of parameter settings represented higher confidence matches (Lin et al., 2014). For this reason, we then considered
peptides that were matched in multiple sets of parameter settings. We called these our high confidence (HiConf) matches. We then used scripts in the Wesleyan IGS database server to conduct further analysis such as assessments of ribosome profiles, mRNA levels, ribosome densities, protein expression and individual information content at start sites.

**Results**

**Different Algorithms Provide Different Samplings of the Proteome**

We ran Mascot and OMSSA on the same data set sampled by SEQUEST in Fournier et al. (2012). This consisted of locally obtained data and publicly available data from the PeptideAtlas repository (Fournier et al., 2012). One of our initial aims was to examine differences in peptide matching between algorithms. Specifically, we wanted to investigate how distinct the matching was.

We found that matches to N-terminal peptides showed modest overlap (Figure 2.1A, Figure 2.1B). Most N-terminal matches to peptides resulting from dnAUG translation initiation (dnPeptides) were found in only one algorithm (90.7%). There was a similar pattern for N-terminal matches to peptides resulting from annAUG translation initiation (annPeptides) (80.5% were matched in only one algorithm).

This was surprising as all three peptide search algorithms were using the same sequence “database” and were searching the same mass spectra. Moreover, our lab had previously shown that there was far more overlap in peptide matches between SEQUEST and OMSSA (55.8% of peptides were detected by
both algorithms; Figure 5B of Lin et al. (2014)). However, in that screen, we were searching for all peptide fragments rather than only for N-terminal fragments.

With this in mind, we compared the matches made by the algorithm for peptides that were not N-terminal (henceforth referred to as internal peptides; Figure 2.1C, Figure 2.1D). Similar to N-terminal dnPeptides, internal dnPeptides were primarily found in only one algorithm (94.7%). One possible explanation for this finding is that proteins are able to have different N-termini (supported by the previous observation that some annORFs show dnORF expression; Fournier et al., 2012). Also, expression of dnORFs could also be a relatively rare translational event (i.e., they are not highly expressed). This would make detection of all dnPeptide fragments (N-terminal or otherwise) more difficult.

We also found that, consistent with Lin et al. (2014), most internal annPeptides were detected by multiple algorithms (75.6% were detected by two or more algorithms; Figure 2.1D). This a sharp contrast with N-terminal annPeptides, which had very little overlap (Figure 2.1B). A possible explanation for reduced overlap for N-terminal peptides is that the same reading frame is able to utilize different N-termini (as evidenced by the number of dnAUGs in the annotated ORF; discussed below).

Overall, these data indicate that Mascot, OMSSA, and SEQUEST are sampling the proteome differently.

Creating a High Confidence Set of Peptides using Mascot and OMSSA
Recent work in our lab (Lin et al. (2014) and Chapter 1) has provided methods for identifying high confidence peptide matches. We wanted to apply this methodology to expand and further characterize our set of dnAUGs using Mascot and OMSSA. We had previously found that peptides that are matched in multiple sets of parameter settings are higher confidence than those only detected in a single set (Lin et al., 2014). We used this as a filter for N-terminal matches. All matches that were filtered were considered high confidence (HiConf) matches.

Using this filter, we ran Mascot and OMSSA on the same set of mass spectra as described above as well as additional mass spectra that were publicly available on PeptideAtlas. We detected 240 HiConf N-terminal dnPeptides and 624 HiConf N-terminal peptides (Table 2.3). We detected fewer distinct N-terminal peptides than Fournier et al. (2012), however, this can be attributed to differences in matching between algorithms as well as the higher stringency of our filtering.

**Preliminary Characterization of the HiConf Set**

Fournier et al. (2012) presented preliminary characterizations of the detected dnAUGs in their screen. We were very interested in comparing our set of HiConf dnAUGs with their detected dnAUGs. In particular, we wanted to investigate any differences between the two data sets.

Similar to Fournier et al. (2012), we found dnPeptides detected in all three reading frames (Figure 2.2). However, our HiConf set had a higher
proportion of peptides in frame 1 (49% vs. 41%; Table 2.4). Additionally, we also found that frequencies of frame 2 and frame 3 dnAUGs were depressed close to the annAUG (Figure 2.2), which is consistent with the findings in Fournier et al. (2012). Detected frame 2 and frame 3 dnORFs were significantly longer than the lengths of all theoretical frame 2 and frame 3 dnORFs (Figure 2.2D). This is consistent with our previous hypothesis that there has been selection against stop codons in these ORFs (Fournier et al., 2012).

Previous studies utilizing ribosome profiling have indicated that there is pervasive translation in the yeast (Ingolia et al., 2009) and mouse (Ingolia et al., 2012) genomes at sites upstream and downstream of annotated start sites. Ingolia et al. (2009) also noted a characteristic ribosome profile for sites of translation initiation. Fournier et al. (2012) noted that dnAUGs had similar ribosome profiles to annAUGs, consistent with there being translation initiation at these sites. We wanted to see if we noted the same pattern with our new HiConf set of both annPeptides and dnPeptides (Figure 2.3A, Figure 2.3B). Similar to Fournier et al. (2012), we noticed a characteristic translation initiation profile for the annPeptides (Figure 2.3A). The ribosome profiles of dnPeptides were consistent with translation initiation, but they differed from those of annPeptides (Figure 2.3B).

Fournier et al. (2012) also examined ribosome densities and mRNA levels of the annORFs of annPeptides and dnPeptides (i.e., the annORF resulting from translation initiation at the annAUG upstream of the dnAUG; called the annORF
associated with the dnPeptide(s)). They found that annORFs associated with
dnPeptides had lower ribosome densities and lower mRNA levels than the
annORFs of annPeptides. Using data obtained from Ingolia et al. (2009), we
performed a similar assessment on our HiConf sets (Figure 2.3C, Figure 2.3D).
Our findings were consistent with those of Fournier et al. (2012). dnPeptide
genes had lower mRNA levels and lower ribosome densities compared to
annPeptides.

Protein expression of the annORFs associated with annPeptides and
dnPeptides were another feature that was examined. Fournier et al. (2012)
found that annORFs associated with dnPeptides had lower protein expression.
However, this was not the case when examining our HiConf sets (Figure 2.4). We
found that there was no statistically significant difference in protein expression.
This is surprising given the fact that annORFs associated with dnPeptides were
found to have lower mRNA and ribosome levels (Figure 2.3). Since mRNA is the
product of transcription and ribosomes perform translation, it follows that lower
mRNA and ribosome levels may correlate with lower protein expression.
However, this is not what we observe (Figure 2.3, Figure 2.4). This is consistent
with the view that translation is regulated on multiple levels rather than solely
by mRNA and ribosome availability.

The final comparison with the Fournier et al. (2012) data that we made
was examining the individual information content of the sequence surrounding
the start sites of the annORFs associated with both annPeptides and dnPeptides
using the Translation Relative Individual Information (TRII) score (a measure of individual information content of the sequences flanking the start site; developed in Weir and Rice (2010)). They found that the sequences surrounding annORFs associated with dnPeptides had lower TRII scores than those associated with annPeptides (Fournier et al., 2012). Again, in contrast to their findings, we observed that there was no statistically significant difference (Figure 2.5). This indicates that even if annORFs have good sequence context for their start sites, this does not prohibit the expression of dnORFs.

In summary, we provide an additional characterization of translation initiation at dnAUGs using high-stringency filtering with two new PSM algorithms. We detected a new set of dnPeptides, adding to those detected previously (Fournier et al., 2012). This represents an initial effort to understand these new dnPeptides given these constraints (i.e., peptides being matched in multiple sets of parameter settings).

Discussion

In this set of experiments, we utilized methodology outlined in a previous publication by our lab to increase stringency in the detection of novel peptides. As a result, we were able to take an initial step towards clarifying the findings that we published in Fournier et al. (2012). We provide data that both challenge and confirm our previous findings.

One interesting finding was the discovery that there is very little overlap in algorithm detection of N-terminal peptides in comparison with other
fragments (Figure 2.1). This could be due to a variety of reasons. Firstly, a variety of studies, including this one, support the notion that eukaryotic cells utilize multiple N-termini (some of which are in the same frame as the annotated N-terminus; Fournier et al., 2012, Ingolia et al., 2009, 2012). As a result, there would be greater overlap among internal trypsin fragments, which would have higher effective concentrations in the cell. Additionally, there may be N-terminal post-translational modifications that we haven’t accounted for in our screens. There are numerous post-translational modifications such as glycosylation that were not included (Kung et al., 2009). Further work needs to be done to test these explanations.

The other intriguing discovery was that some of the characteristics of dnAUGs in this screen were less pronounced or different than those of the dnAUGs in Fournier et al. (2012) (Figure 2.4, Figure 2.5). The first explanation for this was the fact that Fournier et al. (2012) used SEQUEST while we used OMSSA and Mascot. While these algorithms do provide valid samplings of the proteome, they are not the same sampling (Lin et al., 2014). The next possible explanation is that we are using more stringent filtering. This rules out many matches that may be more similar to the matches in Fournier et al. (2012). Additionally, we found that higher confidence matches tend to correlate with higher protein expression (Lin et al., 2014). Thus, by filtering, we could be limiting the detection of low expression peptides. Given the variety of possible N-terminal peptides, this could be a concern in future screens. Finally, Mascot
and OMSSA were run on the entire set of mass spectra from Fournier et al. (2012) in addition to other mass spectra publicly available from the PeptideAtlas repository (data from PeptideAtlas was used in Fournier et al. (2012) but some available data was not in a usable format for SEQUEST). Thus, Mascot and OMSSA were run on a much larger data set. As a result, much of the changes in sampling could be due to the addition of these data sets.

In conclusion, this set of experiments demonstrates that different peptide search algorithms sample the proteome differently and provides a template approach (utilized in Chapter 3) for high-stringency screens of the proteome.

**Future Directions**

This set of experiments represents an initial characterization of a new set of dnPeptides using a high-stringency filtering of potential candidates. Fournier et al. (2012) carried out further characterization of their set of dnPeptides. They examined evolutionary conservation of protein sequences, the codon usage of dnPeptide sequences, predicted protein disorder, gene status, and gene ontology. As a preliminary step, we would like to do the same. Given that there are some differences between our set of dnPeptides that may be due to properties of the different peptide search algorithms, it would be interesting to see what other differences and similarities exist between our set and theirs.

Another future direction is to run Mascot on the same mass spectra while using an a/b/y ion screen as opposed to the standard b/y ion screen, which was used in this set of experiments (an approach discussed in Chapter 1). We have
shown previously that peptides detected in the a/b/y ion screen represent a high-confidence subset of those detected in the b/y ion screen in Mascot (Chapter 1, Lin et al. (2014)). Because we would like to increase confidence in detected peptides, using the a/b/y ion screen to find higher confidence dnPeptides is sensible.

Finally, we would like to verify the expression of the detected dnPeptides by running Western blots on epitope-tagged ORFs, providing independent testing of their expression. We would also like to run assays to assess the function of dnPeptides by knocking down expression of dnORF proteins. Our lab has recently begun doing this for a frame 2 dnORF of the Ace2 gene (Mary Vallo, MA thesis, 2014).

One potential concern (described at length in Ch. 1 discussion) is the potential for creating false negatives using our high-stringency filters. However, we do not believe that this is a major concern. In searching for dnPeptides, we are searching for novel protein products that are not annotated. Thus, our aim is to have the highest confidence in these peptides in order to ensure their expression pending further verification (described in the previous paragraph). In the future, we may loosen our stringency in order to capture all possible matches, but, as an initial first step, we believe that we should maintain high-stringency filters.

In summary, our aim is to further characterize and verify the expression and function of detected dnPeptides. These results further support our
hypothesis that yeast mRNAs can code for multiple protein products. We extend this idea in the next chapter by examining a class of RNAs that was previously thought to be non-coding.
Figures and Tables

A

OMSSA
168

SEQUEST
266

24
29
11
6

B

OMSSA
254

SEQUEST
303

135
17
27
54

27

Mascot
405

405
Figure 2.1. Venn diagram visualization of the numbers of peptides matched to one or more algorithms using the data used in Fournier et al. (2012). N-terminal peptides (A and B) show modest overlap in comparison with annotated internal
peptides (D). (A) N-terminal dnPeptides. (B) N-terminal annPeptides. (C) Internal dnPeptides. (D) Internal annPeptides.
Figure 2.2. (A-C) Distance from the annAUG to the dnAUGs. (A) All detected dnAUGs vs. all possible dnAUGs within 100 nt of the annAUG. (B) All possible dnAUGs within 100 nt of the annAUG by frame. (C) All detected dnAUGs by frame. (D) ORF length for dnPeptides (both detected and all possible dnAUGs).
Figure 2.3. (A) Ribosome profiles for detected HiConf annAUGs (ribosome profiling data available for 455 genes). Consistent with Ingolia et al. (2009), these genes have a ribosome profile indicative of translation initiation. (B) Ribosome profiles for detected HiConf dnAUGs (ribosome profiling data available for 149 genes). HiConf dnAUGs have a ribosome profile that is indicative of translation initiation (though, not as strongly as the HiConf annAUGs). Ribosome profiles for (A) and (B) were aligned with the putative start
site. (C) Normalized mRNA levels for HiConf annORFs and annORFs associated with HiConf dnAUGs. mRNA levels were depressed for annORFs associated with HiConf dnAUGs (p < 0.001). (D) Normalized ribosome densities for HiConf annORFs and annORFs associated with HiConf dnAUGs. Ribosome densities were depressed for annORFs associated with HiConf dnAUGs (p < 0.0002). All raw data for (A)-(D) were taken from Ingolia et al. (2009).

![Protein Expression](image)

Figure 2.4. Protein expression levels for HiConf annORFs and annORFs associated with HiConf dnAUGs with detected protein expression. There was no statistically significant difference between the protein expression levels of these two sets of genes. Data taken from genome-wide Western blot analysis conducted by Ghaemmaghami et al. (2003).
Figure 2.5. dnAUGs and annAUGs show similar sequence context as measured by TRII score. The TRII score is a measure of individual information content of the sequence surrounding the start site of an ORF. This is based upon a weight matrix derived from the start sites of high confidence annotated genes with no AUGs in the known 5’ UTR (methodology described in Weir and Rice (2010)).
### Table 2.1. Parameter Settings for Mascot

<table>
<thead>
<tr>
<th>Parameter</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max. missed cleavage sites</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Precursor mass tolerance (Da)</td>
<td>3.0</td>
<td>1.5</td>
<td>2.0</td>
<td>1.5</td>
</tr>
<tr>
<td>Fragment mass tolerance (Da)</td>
<td>1.0</td>
<td>0.5</td>
<td>0.8</td>
<td>0.5</td>
</tr>
<tr>
<td>Precursor ion search type</td>
<td>mono</td>
<td>mono</td>
<td>avg.</td>
<td>avg.</td>
</tr>
<tr>
<td>Fragment ion search type</td>
<td>mono</td>
<td>mono</td>
<td>avg.</td>
<td>avg.</td>
</tr>
<tr>
<td>Mass tolerance charge scaling</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

### Table 2.2. Parameter settings for OMSSA.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max. missed cleavage sites</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Precursor mass tolerance (Da)</td>
<td>3.0</td>
<td>1.5</td>
<td>2.0</td>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>Fragment mass tolerance (Da)</td>
<td>1.0</td>
<td>0.5</td>
<td>0.8</td>
<td>0.5</td>
<td>1.0</td>
</tr>
<tr>
<td>Precursor ion search type</td>
<td>avg.</td>
<td>avg.</td>
<td>avg.</td>
<td>mono</td>
<td>avg.</td>
</tr>
<tr>
<td>Fragment ion search type</td>
<td>mono</td>
<td>mono</td>
<td>avg.</td>
<td>mono</td>
<td>mono</td>
</tr>
<tr>
<td>Mass tolerance charge scaling</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>linear</td>
</tr>
</tbody>
</table>

### Table 2.3. Summary of matched N-Terminal peptides.

<table>
<thead>
<tr>
<th>Peptide Set</th>
<th>Number of Distinct N-Terminal Peptides Matched</th>
</tr>
</thead>
<tbody>
<tr>
<td>HiConf annPeptides</td>
<td>624</td>
</tr>
<tr>
<td>HiConf dnPeptides</td>
<td>240</td>
</tr>
</tbody>
</table>

Table 2.3. Summary of matched N-Terminal peptides.
References


Chapter 3: Investigation of Translated long non-coding RNA ORFs

Introduction

In the post-genomic era, the annotation of genes has become a unique challenge in understanding the full repertoire or the proteome. Genomes are now easily attainable by utilizing high-throughput sequencing methodologies. However, despite the wealth of genomic knowledge that we now possess, there are still difficulties in identifying the sequences that contribute to the proteome. Current annotation algorithms search for the longest, non-overlapping open reading frames (ORFs) that are more than 100 amino acids (aa) or 300 nucleotides (nt) in length (Kellis, et al., 2003). This implementation choice was made to ensure that annotated ORFs were too large to occur by chance alone (e.g., ORFs that are 60 aa or shorter will arise by chance ~5% of the time, Kellis, et al., 2003). This approach (outlined in Figure 3.1A) is problematic for several reasons.

First, many genes identified using this methodology remain uncharacterized and biologically unverified. According to the most recent annotation (available at Saccharomyces Genome Database), there are 6713 annotated genes in the S. cerevisiae genome, however, only 4673 of these genes have been “Verified” (Saccharomyces Genome Database). Moreover, 813 genes are categorized as “Dubious”, 4 as “Verified Silenced Gene”, 89 as “Transposable Element Gene”, 21 as “Pseudogene” and 1118 as “Uncharacterized”. According to genome-wide Western blot analysis, 572 of these unverified ORFs have
detectable levels of protein expression (Ghaemmaghami, et al., 2003). This means that 22% of the annotated genes have not been biologically verified.

Second, this approach does not utilize other high-throughput biological data sets (e.g., transcriptomic or proteomic data). Rather, gene annotation is performed computationally on whole genomes without high-throughput verification of protein expression. However, there are many techniques that allow for precise visualization of the transcriptome and proteome. Techniques such as RNA-seq (Wang, et al., 2009) and genome-wide tiled RNA microarrays (Bertone et al., 2004) allow for an unbiased characterization of the transcriptome (representing a possible pre-screening for potential protein products), while tandem MS/MS can similarly provide a proteome-wide picture of the products of translation (representing a possible method of high-throughput verification of protein expression) (Aebersold and Mann, 2003, Elias et al., 2005, Fournier et al., 2012). Moreover, studies utilizing ribosome profiling have indicated that translation is more widespread throughout the genome than previously thought, indicating that our current annotations are insufficient in characterizing the proteome (Ingolia et al., 2009, Ingolia et al., 2011). Current approaches to gene annotation do not fully incorporate these transcriptomic and proteomic data sets. As a result, current annotations do not fully address the complexity of the genomic and proteomic repertoire.

Finally, proteins that are shorter than 100 amino acids may have functional significance. For example, Kondo et al. (2007) found that a
polycistronic RNA encoded three short peptides (two were 11 aa long and the other was 32 aa long) that regulated actin-based cell morphogenesis in *Drosophila* embryos. More importantly, these three peptides were each individually capable of rescuing the phenotype in knockout strains (Kondo, et al., 2007). This is significant for several reasons. Firstly, the ORFs that encoded these peptides were too short to be annotated by current algorithms. Secondly, these short peptides were functionally significant despite their small size.

In all, this suggests that the proteome of *S. cerevisiae* is under-annotated with respect to small peptides given the constraints of current algorithms. There may be many small peptides (some of which could be encoded by polycistronic RNAs) that are functionally significant, yet not currently annotated. In order to address this concern, we took an alternative approach to identify protein-coding ORFs (Figure 1B). In this approach, we take unbiased transcriptomic data to determine the repertoire of potentially coding RNAs, then we search for putative ORFs, and test the expression of these ORFs using peptide MS/MS. This approach has the distinct advantage of being able to identify previously undetected ORFs in a high-throughput manner with high confidence.

In this set of experiments, we demonstrate that a previously un-annotated source of ORFs show evidence of protein expression. Moreover, these ORFs are generally shorter than 300 nt and reside on RNAs that were previously thought to be non-coding. Additionally, many of the ORFs are located on
polycistronic RNAs. Finally, we provide preliminary data to suggest that these ORFs are not highly expressed.

**Methods**

**Selecting a set of potential ORFs**

IncRNAs are believed to represent a class of functionally significant untranslated RNAs. They can be broadly classified as either Stable Un-annotated Transcripts (SUTs) or Cryptic Unstable Transcripts (CUTs). They are generally longer than 200 nt and are known to be implicated in functions such as regulation of chromatin modifications, RNA masking activities, recruitment of transcription factors, and regulation of development (Mercer and Mattick, 2013). Perhaps it is the case that some of these functions could be performed by small peptides similar to the earlier example in *Drosophila*.

In 2013, Tuck and Tollervey conducted a transcriptome-wide screen of the binding of long non-coding RNAs (IncRNAs) and mRNAs to 13 proteins known to be a part of the mRNA maturation pathway using cross-linking and cDNA analysis (CRAC). The proteins and bound RNA form ribonucleoprotein particles (RNPs) during their transcription, packaging, processing, export, translation, and cytoplasmic decay. They analyzed the binding patterns of these RNPs and, based upon this, they were able to deduce how different classes of RNAs matured. One of their major findings was that many IncRNAs may mature similarly to mRNAs (Tuck and Tollervey, 2013). This suggested that IncRNAs could be potentially a source of un-annotated protein-coding ORFs.
Xu et al. (2009) performed a genome-wide RNA tiling microarray for the entire transcriptome. This method allowed for accurate visualization of the transcriptome with 8 base-pair resolution (Mancera et al., 2008). They identified 1757 lncRNAs (Xu, et al., 2009). The genomic sequences of these RNAs were loaded into the Wesleyan IGS server database and subjected to a search of potential ORFs. In all, 20,300 potential ORFs were found (Table 3.1). All ORFs coding for more than five amino acids (17,582 in total) were compiled into a FASTA search database along with the annotated genome. Reverse sequences to all entries were also included as decoys. This database was then submitted to the peptide MS/MS search algorithms Mascot and OMSSA.

**Peptide Matching by MS/MS Search Algorithms**

OMSSA was run on the Wesleyan University Computing Cluster and the Mascot Search Algorithm was run on a local Dell XPS Server using the Mascot Daemon program. Five different sets of parameter settings were used for OMSSA and four different sets of parameter settings were used for Mascot (Table 3.2, Table 3.3). MS/MS data was obtained from the PeptideAtlas repository as well as locally procured data (outlined in Fournier et al., 2012). All of the cell lysates were subject to chemical modifications (e.g., for one set of MS/MS experiments, proteins were treated with glutaraldehyde resulting in a +68 kDa mass addition at the N-terminus) and trypsin digest prior to undergoing MS/MS. To account for this, each parameter setting had both optional and static peptide modifications for the data set being run (listed in PeptideAtlas.org and Fournier et al. (2012)).
The algorithms were also tuned to incorporate a trypsin digest in all parameter settings. We ran the data sets on both OMSSA and Mascot using these sets of parameter settings and the resulting output was stored in the Wesleyan IGS database server.

Peptide matches were then subjected to filtering using a 5% false detection rate (FDR) by decoy analysis (decoys explained above). Subsequently, we discarded all instances where a given peptide matched to multiple FASTA entries. For OMSSA, we discarded all entries where different peptides matched to the same spectrum. For Mascot, we only considered the best peptide matches for a given spectrum (i.e., only entries where pep_rank = 1).

Some matches had ORFs that overlapped with other detected ORFs and both had the same stop site. In that case, only the longer ORF was considered as a match. Finally, matches to potentially coding lncRNA ORFs (IncORFs) were only considered if the peptide matched to spectra in the same MS/MS experiment in multiple parameter settings. This approach (outlined in Lin, et al. (2014)) provides for greater confidence in the quality of matches as opposed to peptides that were only matched in a single parameter setting. Most peptides matched to only one parameter setting; however, many were matched in multiple parameter settings (Figure 3.2). We considered these our high confidence (HiConf) IncORFs.
We wanted to ensure that we were picking the high confidence matches. To address this, we computed confidence scores for all matches. This score was computed using the following equation:

\[ d = -\log_{10}(PSM\_Prob\_Score / FDR\_threshold) \]

where \( d \) is the confidence score, \( PSM\_Prob\_Score \) is the probability score for a given peptide spectrum match (PSM), and \( FDR\_threshold \) is the FDR scoring threshold. For Mascot, \( PSM\_Prob\_Score \) was calculated in the following way:

\[ PSM\_Prob\_Score = -10 * \log_{10}(PSM\_Prob) \]

where \( PSM\_Prob \) is the output score given by Mascot. The confidence score provides a log-based measure of the quality of a given peptide spectrum matched as compared to the worst possible match allowed by the FDR scoring threshold. We compared the confidence scores of peptides matched to multiple parameter settings to those matched in only a single parameter setting for our detected lncORFs (Figure 3.3). Peptides that were detected in multiple parameter settings had higher confidence scores than peptides detected in a single parameter setting.

We considered the lncORF peptides matched in multiple sets of parameter settings to be our high confidence (HiConf) set. Because we were looking for previously un-annotated proteins, we were not concerned with potential false negatives due to the stringency of our filters (described in Ch. 1 Discussion). We then performed bioinformatic analyses on this HiConf lncORF set using SQL scripts written in the Wesleyan IGS database server.
Results

Translated IncORFs are primarily short proteins often found on polycistronic RNAs

Our HiConf IncORF set consisted of 568 detected proteins that were not previously annotated. We wanted to investigate why these proteins were not annotated. One approach we took was to examine the length of the IncORFs (Figure 3.4). Standard annotation algorithms search for potential ORFs that are longer than 300 nt or 100 aa (Kellis et al., 2003). We hypothesized that one reason that the IncORFs were not annotated was because they were shorter than that.

As we expected, a majority of detected proteins (~99%) were found to be smaller than 100 aa. Indeed, only 7 (~1%) of the 568 detected proteins had length longer than 100 aa. Thus, most of the detected IncORFs were not annotated due to being too short to be considered by standard gene annotation algorithms. This is also true of the majority of potential IncORFs (~0.5% were longer than 300 nt, Figure 3.4). It is not clear why IncORFs (both translated and un-translated) that were longer than 300 nt were not annotated. It is possible that these ORFs overlapped currently annotated ORFs.

Because translated IncORFs are relatively short, we wanted to examine whether they resulted from polycistronic RNAs. We found that many IncORFs reside on polycistronic RNAs (Figure 3.5). This suggests that, if these IncORFs are functionally significant, there could be a mechanism similar to what was
outlined by Kondo et al. (2007) where a polycistronic RNA encodes multiple short peptides that perform one or more cellular functions. More work needs to be done to see if this is the case.

The Position and Sequence Context of IncORFs suggests a distinct mechanism of Translation Initiation

Given that there are many IncORFs that are found on polycistronic RNAs, this suggested that there could be distinct mechanisms of translation initiation for these ORFs. To this end, we first investigated where the IncORF start site was found on the RNA in relation to the 5’ cap (Figure 3.6). Because of the number of polycistronic IncRNAs (93 out of 436 IncRNAs), we should expect that many of the IncORF start sites would be further away from the 5’ cap of the RNA. Consistent with this prediction, we found that the start sites of IncORFs were further downstream of the 5’ cap than those of annotated mRNAs (Figure 3.6).

This result suggested that perhaps there could be a different mechanism of translation initiation at work. To this end, we examined the sequence context of start sites. We used a scoring methodology that was previously developed in the lab called the Translation Relative Individual Information (TRII) score (Weir and Rice, 2010). This score measures the individual information of the sequence context of a given start site using reference sequences from high confidence mRNAs (Weir and Rice, 2010). If there is a different mechanism of translation initiation for the translated IncORFs, the start sites of IncORFs could have depressed TRII scores when compared to translated annotated ORFs (annORFs).
Consistent with our prediction, lncORFs show a depressed TRII score relative to annORFs (Figure 3.7). This TRII score depression could be due to lncORFs on polycistronic RNAs interfering with the sequence context of each other, a different mechanism of translation initiation being necessary for these short lncORFs, or some mixture of these factors along with many other possibilities. Because of the distance of lncORF start sites from the 5’ RNA cap (Figure 3.6), a mechanism comparable to translation initiation at internal ribosomal entry sites (IRESs) is also possible.

**Codon Usage and RNA levels of Translated lncORFs suggest that they are not highly expressed**

We next wanted to examine the codon usage of translated lncORFs. Codon usage refers to the specific codons found in a given sequence. If this sequence uses more preferential codons (i.e., those more conducive to efficient translation), it can be said to have better codon usage. One measure of codon usage is the Codon Adaptation Index (CAI). This index measures the bias in codon usage towards certain codons. It is calculated as follows for a given codon:

\[
W_{aa,i} = \frac{f_{aa,i}}{f_{aa,max}}
\]

where \(w_{aa,i}\) is the CAI score for the given codon \(i\), \(f_{aa,i}\) is the that codon \(i\) codes for its amino acid, \(aa\) (as compared to other codons coding for the same amino acid), and \(f_{aa,max}\) is the frequency of the codon that is most often used to encode the same amino acid (Jansen et al., 2003). It has been found that higher CAI scores in
the translated region of genes correlates with more protein expression (Jansen et al., 2003).

We computed the CAI score for the 7 codons both upstream and downstream of the start site for lncORFs, high confidence annORFs, and annORFs that are known to show low protein expression (LowExp) (expression data taken from genome-wide Western blots conducted by Ghaemmaghami et al., 2003). As expected, high confidence annORFs had higher CAI scores downstream (the translated region) of the start site than upstream of it (the untranslated region) (Figure 3.8). This is consistent with the finding that more highly translated regions have better CAI scores (Jansen et al., 2003). With this in mind, we took the difference between the regions upstream and downstream of the start site to be a relative measure of predicted protein expression.

As expected, LowExp annORFs showed a less pronounced difference between the CAI scores in the upstream and downstream regions of the start site (average difference for LowExp annORFs was 1.9 as compared to 7.7 for high confidence annORFs, Figure 3.8). HiConf lncORFs showed a similar pattern in CAI scores to LowExp annORFs (average difference in CAI score between the upstream and downstream regions of the start site 1.6, Figure 3.8) indicating that these ORFs may be low expressing based upon their codon usage.

We also wanted to investigate the RNA levels of the HiConf lncORFs as compared to high confidence annORFs and LowExp annORFs. We used data obtained by Ingolia et al. (2009) where they performed RNA-seq on poly(A)-
selected mRNAs (i.e., they purified for RNAs with poly(A) tails on a transcriptome-wide scale). This data allowed us to calculate the RNA levels of HiConf IncORFs, high confidence annORFs, and LowExp annORFs (Figure 3.9). Many HiConf IncORFs showed low RNA levels in comparison with high confidence annORFs and LowExp annORFs (Figure 3.9). This is consistent with the hypothesis that HiConf IncORFs have relatively low protein expression.

Taken together, these data indicate that proteins resulting from translated IncORFs are likely expressed at low levels, although further analysis is needed to confirm this prediction.

**Binding to Proteins in the mRNA Maturation Pathway indicates that many Translated IncRNAs are mRNA-like**

As discussed above, Tuck and Tollervey (2013) performed CRAC using 13 proteins that are known to be a part of the mRNA maturation pathway. The functions of these proteins include nuclear surveillance, cleavage and polyadenylation, translation, and subsequent cytoplasmic decay. Based upon the binding of RNAs to these proteins, Tuck and Tollervey (2013) were able to group RNAs into clusters. These clusters could be divided into three IncRNA-like clusters and seven mRNA-like clusters. One of their most startling findings was that many IncRNAs fell into mRNA-like clusters (Tuck and Tollervey, 2013). Given the fact that these IncRNAs seem to mature like mRNAs, we wanted to assess whether the IncRNAs we detected behaved similarly to mRNAs as well.
Using the clusters as assigned by Tuck and Tollervey (2013), we examined the set of detected lncRNAs. 281 of the lncRNAs detected in our screen were detected by Tuck and Tollervey (2013) (Table 3.4). While most of translated lncRNAs fell into lncRNA-like binding clusters, 65 (11 CUTs and 54 SUTs) were in mRNA-like clusters (Table 3.4). This suggests that some of the translated lncRNAs are actually mis-annotated mRNAs.

**Discussion**

**Under-annotation of the Yeast Proteome**

This study provides evidence that the proteome of *S. cerevisiae* is significantly under-annotated. We identify 568 ORFs that were previously thought to be non-coding. What is perhaps more interesting is that many of these proteins occur on polycistronic RNAs and are too short to be annotated by current annotation algorithms.

lncRNAs were previously thought to be a class of non-coding RNA. Indeed, analyses using both ribosome profiling and peptide mass spectrometry have suggested that lncRNAs do not contribute significantly to the proteomes of mice and humans (Guttman et al., 2013, Banfai et al., 2012). However, we provide evidence that lncRNAs contribute to the yeast proteome. This suggests that many lncRNAs have been mislabeled as non-coding and may in fact be mRNAs as suggested by our cluster analysis (Table 3.4). Perhaps we will be able to perform similar analyses in higher organisms to confirm that this is the case across species.
A New Approach to Genome Annotation

Figure 1 contrasts the standard approach to gene annotation with the approach taken in this study. The discovery of a new set of coding RNAs that contain previously un-annotated ORFs confirms the validity of our approach. The main strength of our approach is that we are integrating these data sets so that we can get a clearer picture of the proteome without relying upon assumptions made by annotation algorithms.

We have used *S. cerevisiae* as an example of our approach and we hope that we can begin to apply it to higher organisms. IncRNAs are found across many taxa and could be contributing to the proteomes of many organisms (Kung et al., 2013). The example of the *polished rice* gene from Kondo et al. (2007) illustrates that small peptides residing on polycistronic RNAs can contribute to the phenotype. Our data suggest that this is a possibility for IncRNAs. Our approach will enable future analyses that can test whether IncRNAs code for proteins in other organisms.

Future Directions

This study provides a preliminary characterization of translated IncORFs. We want to expand this analysis in the future. As a first step, we would like to examine the ribosome profiles of the translated IncORFs using data from Ingolia et al. (2009). This type of data provides a high-resolution, in vivo snapshot of translation as it is occurring. Previously, Ingolia et al. (2009) noted that there is a characteristic ribosome profile for the regions surrounding annotated start sites.
It would be interesting to see how the ribosome profiles of the regions surrounding start sites of translated IncORFs compare to this characteristic profile. If the profiles are markedly different, this could suggest that there is a different mechanism of translation initiation at work. Additionally, this data allows us to measure the number of ribosomes that are bound to a given ORF. This could aid in further predicting the protein expression of translated IncORFs.

Secondly, we would like to examine the protein to which the translated IncRNAs were bound using the data from Tuck and Tollervey (2013). Although Tuck and Tollervey (2013) have already clustered the IncRNAs based upon how they were bound to specific proteins, it would be interesting to investigate whether the individual binding patterns of the translated IncRNAs would give us any insight into the processing of these RNAs. For example, Mex67 is a protein involved in the nuclear export of mRNAs and there is data how it binds to IncRNAs (Tuck and Tollervey, 2013). If a given transcript showed elevated binding to this protein, it would be consistent with the idea that the transcript was exported from the nucleus to the cytoplasm. In principle, analyses like these could help us glean further insight into how translated IncRNAs are processed.

Thirdly, we would like to see how well the sequences of translated IncORFs are conserved. Kellis et al. (2003) observed that comparative genomics can aid in gene annotation. For example, they noticed that, across four species of yeast, ORFs tended to have fewer indels that altered the reading frame of the sequence than un-translated regions (Kellis et al., 2003). It would be interesting
to conduct a similar analysis to see how well these ORFs are conserved by comparing these sequences in *S. cerevisiae* to those in other yeast species. Additionally, it has been suggested by Kung et al. (2013) that IncRNAs could be loci for gene evolution and an evolutionary analysis could certainly help address this question. If IncRNAs are loci for the evolution of new genes, we should expect that their sequences are poorly conserved relative to other, more ancient genes.

Finally, we would like to verify the expression of these proteins and assess their function using standard molecular biological methods. We would like to verify the protein expression by conducting Western blot analysis using epitope-tagged IncORFs. We would also like to engineer deletion strains in order to examine the phenotypic effects of losing IncORF expression. One method would be to delete the IncORF and then examine the cellular phenotype. However, this approach would alter RNA-mediated effects that the IncRNAs may have so we could also mutate the start site of the IncORF such that translation initiation will not occur. This would allow us to examine the phenotypic consequences of losing IncORF protein expression without interfering with RNA-mediated IncRNA effects.
Figure 3.1. Two different approaches to predicting protein expression. A) The standard model of gene annotation. The genome is sequenced and all non-overlapping ORFs longer than 100 amino acids are considered genes. B) An alternative model of annotating coding regions. Take all detected cellular RNAs and then predict ORFs resulting from these RNAs. Verify the expression of these ORFs using peptide MS/MS.
Figure 3.2. Most peptides are matched in only one parameter setting. Peptides that were matched in multiple parameter settings were considered high confidence matches. A) Matches for OMSSA. B) Matches for Mascot.
A

Confidence for IncPeptides Matching 1 Parameter Set

B

Confidence for IncPeptides Matching >1 Parameter Sets
Figure 3.3. Peptides matched in multiple parameters show greater confidence than peptides matched by a single parameter setting. For OMSSA (A and B), confidence was computed by \(-\log(\text{PSM Score}/\text{FDR Threshold})\). For Mascot (C and D), confidence was computed by \((\text{PSM Score} - \text{FDR Threshold})/10\). Both...
scores are a log-based measure of the fold improvement of a given PSM score over the minimum score accepted by the false detection rate (FDR Threshold).

Figure 3.4. Many detected IncORFs are shorter than 300 nt. Most IncORFs fell between 30 nt and 90 nt (~56%). Very few IncORFs were longer than 300 nt (~1%). 300 nt is typically the minimum length that is considered for an ORF by standard annotation algorithms. Most detected IncORFs would not be annotated by these algorithms for this reason.
Figure 3.5. Multiple peptides were obtained from polycistronic RNAs. 343 (79%) translated IncRNAs had only one translated ORF. 59 (14%) translated IncRNAs had two non-overlapping ORFs. 11 (3%) translated IncRNAs had two overlapping ORFs. 23 (5%) translated IncRNAs had 3-6 ORFs.
Figure 3.6. Start sites of lncORFs are typically further downstream of the 5’ cap than annotated mRNAs. The median position for mRNAs is 70 nt. The median position for lncORFs is 356 nt.
Figure 3.7. lncRNA start sites show poorer sequence context than mRNA start sites as measured by TRII score. The TRII score measures the individual information content of the sequence surrounding the start site. This is based upon a weight matrix derived from the start sites of high confidence annotated genes with no AUGs in the known 5’ UTR (methodology described in Weir and Rice (2010)).
A

CAI for HiConf AnnPeptides

B

CAI for Low Expression Genes
Figure 3.8. IncORFs show similar codon usage to low expression genes. This was measured using the Codon Adaptation Index (CAI). The CAI of a given position was the average CAI value of all codons found in that position. (A) CAI of high confidence (HiConf) annotated genes. (B) CAI of low expression genes (data from Ghaemmaghami et al., 2003). (C) CAI of HiConf IncORFs.
Figure 3.9. IncORFs show lower RNA levels than both annORFs and low protein expression genes. RNA levels were from RNA-seq data for poly(A) selected RNAs from Ingolia et al. (2009).
### Table 3.1. Summary of potential IncORFs.

<table>
<thead>
<tr>
<th>Max. missed cleavage sites</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precursor mass tolerance (Da)</td>
<td>3.0</td>
<td>1.5</td>
<td>2.0</td>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>Fragment mass tolerance (Da)</td>
<td>1.0</td>
<td>0.5</td>
<td>0.8</td>
<td>0.5</td>
<td>1.0</td>
</tr>
<tr>
<td>Precursor ion search type</td>
<td>avg</td>
<td>avg</td>
<td>avg</td>
<td>mono</td>
<td>avg</td>
</tr>
<tr>
<td>Fragment ion search type</td>
<td>mono</td>
<td>mono</td>
<td>avg</td>
<td>mono</td>
<td>mono</td>
</tr>
<tr>
<td>Mass tolerance charge scaling</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>linear</td>
</tr>
</tbody>
</table>

### Table 3.2. Parameter Settings for OMSSA.

<table>
<thead>
<tr>
<th>Max. missed cleavage sites</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precursor mass tolerance (Da)</td>
<td>3.0</td>
<td>1.5</td>
<td>2.0</td>
<td>1.5</td>
</tr>
<tr>
<td>Fragment mass tolerance (Da)</td>
<td>1.0</td>
<td>0.5</td>
<td>0.8</td>
<td>0.5</td>
</tr>
<tr>
<td>Precursor ion search type</td>
<td>mono</td>
<td>mono</td>
<td>avg.</td>
<td>avg.</td>
</tr>
<tr>
<td>Fragment ion search type</td>
<td>mono</td>
<td>mono</td>
<td>avg.</td>
<td>avg.</td>
</tr>
<tr>
<td>Mass tolerance charge scaling</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

### Table 3.3. Parameter settings for Mascot.
Table 3.4. Many translated lncRNAs are found in mRNA-like clusters. Clusters obtained from Tuck and Tollervey (2013). Clusters 1-3 are considered lncRNA-like clusters and clusters 4-10 are considered mRNA-like clusters based upon binding profiles to proteins involved in the mRNA maturation pathway.

<table>
<thead>
<tr>
<th>Clusters</th>
<th>CUTs</th>
<th>SUTs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-3</td>
<td>121</td>
<td>95</td>
</tr>
<tr>
<td>4-10</td>
<td>11</td>
<td>54</td>
</tr>
</tbody>
</table>
References


