An Automated Proteogenomic Method Uses Mass Spectrometry to Reveal Novel Genes in Zea mays*†

Natalie E. Castellana‡, Zhouxin Shen§, Yupeng He¶, Justin W. Walley§, California Jack Cassidy‖, Steven P. Briggs§, and Vineet Bafna‡**

New technologies in genomics and proteomics have influenced the emergence of proteogenomics, a field at the confluence of genomics, transcriptomics, and proteomics. First generation proteogenomic toolkits employ peptide mass spectrometry to identify novel protein coding regions. We extend first generation proteogenomic tools to achieve greater accuracy and enable the analysis of large, complex genomes. We apply our pipeline to Zea mays, which has a genome comparable in size to human. Our pipeline begins with the comparison of mass spectra to a putative translation of the genome. We select novel peptides, those that match a region of the genome that was not previously known to be protein coding, for grouping into refinement events. We present a novel, probabilistic framework for evaluating the accuracy of each event. Our calculated event probability, or eventProb, considers the number of supporting peptides and spectra, and the quality of each supporting peptide-spectrum match. Our pipeline predicts 165 novel protein-coding genes and proposes updated models for 741 additional genes. Molecular & Cellular Proteomics 13: 10.1074/mcp.M113.031260, 157–167, 2014.

Accurate genome annotation, wherein the location and structure of all protein coding genes are identified, is critically important and yet it remains elusive for even the most extensively studied organisms. The wide availability of inexpensive next-generation sequencing technologies ensures that model organisms from all branches of the tree of life will continue to be sequenced at an ever increasing pace. However, the annotation pipelines are not able to keep up.

Much recent focus on computational gene finding is on incorporating transcript evidence. As with genomic sequencing such as RNA-Seq (1) has dramatically changed the genome annotation landscape. Although RNA-Seq provides valuable evidence for genome annotation (2–5) it does not provide a comprehensive solution either. Increasing evidence suggests that a discrepancy exists between protein isoforms that are transcribed versus translated (6). Indeed in our own observation, we find evidence for genes in sampling proteins that are not visible at the transcript level. Moreover, the transcript evidence is confounded by prespliced messages, nontargeted expression noise, ncRNA, and lack of strand and frame information. All of these pose challenges for gene finding.

Tandem mass spectrometry is a key technology for assaying the expressed proteome. In typical bottom-up workflows, enzymatically digested peptides are isolated via chromatography and then fragmented in the mass spectrometer. The collection of masses of peptide fragments (tandem mass spectrum) is used as a fingerprint for identification of expressed peptides.

Historically, the genomics community has provided the annotations (aa sequences) and the proteomics community has focused on identifying peptides and proteins from this annotated list to assay for expression of proteins in specific contexts. However, rapidly improving MS instrumentation and advances in sample preparation have enabled the field of proteogenomics, relying on direct interpretation from the genome (7). In this context, the evidence of peptide expression is used to annotate the genome and to reconstruct gene structures in model organisms (8–11), multiple organisms in parallel (12–14), and difficult to annotate genomes such as those with high GC content (15).

However, significant concerns remain in the development of this technology. The tandem mass spectra are noisy, and large-scale MS-based proteomic studies are confounded by false positives resulting from the intrinsic testing of millions of hypothesis. Further, many of the peptides cross splice-junctions, and would not be identified by searching against a six-frame translation. At the same time identification of spliced-peptides is key to reconstructing gene structures. Finally, the annotated peptides must be reconciled into
complete gene models, possibly with alternatively spliced isoforms.

Here we present a semi-automated proteogenomic method and apply it to the annotation of the maize (Zea mays) genome. Our method extends our previous proteogenomics efforts on human and Arabidopsis (16, 8). Like previous efforts, we use a splice-graph database for identifying splice peptides. We automate the refinement process to predict complete gene models, and automatically update existing structures. We developed a framework for evaluating the quality of mass spectrometry-based discovery of gene refinement ‘events’ to control the false discovery rate. The maize genome is particularly challenging because of its large size (2 billion nucleotides), and abundance of mobile elements that re-insert themselves into the genome creating many repetitive regions (17, 18). Instead of discarding peptides that appear in multiple locations, we use them in addition to uniquely mapping peptides for scoring novel discoveries.

We analyzed over 109 million tandem mass spectra generated from Zea mays seeds at multiple stages of development. Comparison of the spectra against our putative protein database containing nearly 2 billion amino acids required extensive computing power (~900 thousand CPU-hours). Our analysis revealed a revised genome annotation with updated gene models for 741 genes and the addition of 165 novel protein coding genes. This study represents one of the largest proteogenomic efforts undertaken on a single organism.

**EXPERIMENTAL PROCEDURES**

The core of our method is the identification of peptides that are discordant with the annotated proteome (See Fig. 1). A total of 109 million tandem mass spectra were acquired as described previously (19). Samples were acquired from maize embryo, endosperm, germinating kernel, and pericarp together with aleurone. The endosperm samples were acquired at 8, 10, and 12 days after pollination. The experiment was designed to capture the expression of every protein involved in seed development. Although the samples contain significant diversity, cross-tissue comparisons are outside the scope of this paper.

First, we clustered and quality filtered the tandem mass spectra to 18 million spectral clusters (see Supplemental Methods). On average 6 tandem mass spectra were combined into a single spectral cluster, increasing the signal to noise ratio. To identify the novel peptides, we constructed a database of putative translated sequences directly from the genome. The database contained two translations of the genomes; a six-frame translation and a splice graph (16). The splice graph compactly encodes potentially thousands of putative transcripts for each gene locus. The putative transcripts are informed by over 2 million mRNA sequences combined with ab initio predictions. Predictably, many putative transcripts share coding sequence, which are collapsed in the graph and can be searched efficiently. In total, the six-frame translation contained 1.98 billion amino acids whereas the splice graph contained 218 million amino acids. To complete the search of this enormous database, we divided the database sequences into 45 individual files to be distributed on a computer cluster.

**Database Construction and Peptide Identification**—The six-frame translation of the genome has previously been used with great success to capture novel coding regions for proteogenomics (8). In eukaryotic organisms, spliced peptides provide a wealth of information about the structure of a gene. Whereas the six-frame translation does not include peptides that span splice boundaries, an exon splice graph (16) can compactly represent many putative splice junctions. The exon splice graph construction method relies on putative gene predictions, including alternate transcripts for the same gene that will include significant amounts of sequence redundancy. The graph construction routine then merges transcripts that share sequence, so that every exon prediction appears only once in the graph.

We searched all tandem mass spectral clusters against the putative protein sequence databases and the annotated protein database (B73 RefGen_v2 5a.59 (16)) using InsPect (20) and rescoring using MS-Generating Function (21). We filtered peptide-spectrum matches to a 1% false discovery rate (FDR) (See Supplemental Methods). Peptides were then labeled as either “known,” meaning they can be derived from an annotated protein sequence, or “novel.” Novel sequences were additionally filtered to remove peptides that could easily be explained by known peptides with mutations or modifications undetectable by mass spectrometry (See Supplemental Methods). By our definition, we found 24,782 novel peptides. Known peptides, those that match to annotated proteins, confirm the translation of the protein. Using the known peptides that match to only one protein in the 5a.59 protein set, we identified expressed proteins. In the case of genes with several protein isoforms that share significant amounts of sequences, it is impossible to localize the peptide to a single isoform. Known peptides that only matched to isoforms from the same locus were used to identify expressed loci.

We mapped novel peptide sequences to all locations in our putative protein databases. On average, a novel peptide matched to 3.67 locations in the genome. In subsequent analysis, we treat peptides that match to multiple locations (shared peptides) differently than peptides that match to a single location (unique peptides). For gene annotation, the most important step is interpreting the peptide locations to identify refinements to the current annotation. We created an automated method for interpreting and scoring the suggested refinements. We call these refinements, “novel events.” Each novel event indicates an update to a specific annotated protein. The exception to this is novel genes, which are refining the annotation as a whole, not a specific gene. In our method, we define seven types of novel events besides novel genes. The event types accompanied by a brief description are provided in Table I. The events are ordered by precedence so as to prevent a novel peptide cluster from being interpreted as two distinct types of events while not providing any new information. For example, a peptide may overlap an exon in a different frame and also extend beyond the boundary of the exon. Using the precedence rules in Table I, this peptide would contribute to a “Frame Change” event and not an “Exon Extension” event.

The databases, database search results, and peptide locations can be found at http://maizeproteome.ucsd.edu.

**Novel Event Identification and Scoring**—Classification of peptides into events can be viewed as an extension of the protein inference problem. In protein inference, a collection of proteins is inferred from the set of identified peptides matched to them. Enforcing a minimum number of peptides matching each inferred protein is a common method for controlling false discoveries in protein inference (the “N-peptide” rule).

The N-peptide rule has several limitations, however. First, applying a peptide-based rule disregards all information about the quality and quantity of spectral matches to the peptides. In addition, proteins expressed at a low level or short proteins may only ever produce one detectable peptide. The rule also creates ambiguity about how to treat

---

1 The abbreviations used are: FDR, false discovery rate; UTR, untranslated region; MS/MS, tandem MS.
uniquely mapping versus shared peptides. If we only consider unique peptides, then we throw out 30% of our peptide identifications. If we consider all peptides, then we are giving equal weight to a peptide mapping to two locations and a peptide mapping to 100 locations. We face the same challenge in proteogenomics when inferring events. As an alternative to an event finding rule based on the number of peptide locations, and we assume that the identification of distinct peptides is independent. Therefore,

$$\Pr(E) = 1 - \prod_{l \in \text{Loc}(E)} (1 - \Pr(l))$$

where \(\text{Loc}(E)\) is the set of peptide locations supporting \(E\), and \(\Pr(l)\) is the probability of the location \(l\) being translated. Intuitively, this equation explains that the probability of an event being correct is the probability that at least one of the locations in the event is expressed. \(\Pr(l)\) is dependent on the confidence in the peptide sequence identification, and the number of locations the peptide matches in the genome. Conversely, the probability of a peptide sequence being

---

**TABLE I**

| Event Type                  | Description                                                                 |
|-----------------------------|-----------------------------------------------------------------------------|
| Antisense overlapping       | The peptide cluster appears within 3kb of the annotated transcript, but is on the reverse strand |
| Frame change                | The peptide cluster overlaps an annotated exon, but in a different frame of translation |
| Exon extension              | The peptide cluster overlaps an annotated exon, but extends beyond the exon boundary |
| Novel exon                  | The peptide cluster falls with the intron region of an annotated transcript |
| Novel splice junction       | The peptide sequences map to exons, but the indicated splice junction is not in the annotated transcript |
| Translated UTR              | The peptide sequences fall within the untranslated region (UTR) of the transcript |
| Gene extension              | The peptide sequences fall within 3kb of the annotated transcript |

---

**Fig. 1.** An overview of the proteogenomic method. Our method accepts tandem mass spectra and a genome. The genome is translated into two sequence databases: the six-frame translation and a splice graph. The spectra are compared with the genome to identify peptide sequences. The peptides either confirm protein sequences in the 5a.59 annotation or match a novel protein-coding region of the genome. The “novel” peptides are grouped into events and scored. The best events are then used to predict full gene models.
identified correctly is the sum of the probabilities of it being translated from all possible locations in the genome. In Fig. 2, the blue peptide is uniquely located, only appearing at the single location on the genome. However, the green peptide appears in two locations. We compute the probability of a location, \( l \), which matches peptide \( p \) being translated as

\[
\Pr(l) = \frac{1}{m_p} \Pr(p)
\]

where \( m_p \) is the number of genomic locations for peptide \( p \). We chose to consider each location of a peptide equally, but the distribution of probabilities could be extended to take into account neighboring peptides or genomic signals. The probability of a peptide sequence identification being correct, \( \Pr(p) \), is informed by the quality of the spectra identifying the peptide. Whereas we could treat the identification of distinct peptides as independent events, we cannot make that assumption of multiple spectra identifying the same peptide. Instead, we conservatively evaluate the probability of the peptide sequence being correct as the highest probability of a spectrum matching the peptide.

\[
\Pr(p) = \max_{s \in \text{Spec}(p)} \Pr(s, p)
\]

where \( \text{Spec}(p) \) is the set of spectra matching \( p \) and \( \Pr(s, p) \) is the probability of the spectrum \( s \) being generated from \( p \). We estimate the \( \Pr(s, p) \) using the local false discovery rate (LFDR) (23) of the match. The LFDR is a measure of the rate of false discoveries among peptide-spectrum matches of similar score. In our pipeline, we estimated match quality using the Spectral Probability produced by the MS-Generating Function (21), and computed the FDR in fixed-width bins with a minimum of 100 matches per bin. The LFDR of the match of \( p \) and \( s \) with MS-Generating Function Spectral Probability \( q \) is simply the LFDR in the bin containing the score \( q \). We then can compute the probability of a peptide-spectrum match being correct as

\[
\Pr(s, p) = 1 - \text{LFDR}
\]

We evaluated the ability of three filtering methods to identify events while maintaining an event-level FDR of no more than 5% (see Table II). We generated plausible decoy events (see Supplemental Methods) to evaluate the FDR of each method. We observed that the decoy events did not fall evenly across all types of events, with the majority of decoy events being novel gene events. This is intuitive because we expect decoy events to fall randomly on the genome, the majority of which is intergenic. For this reason, we divided the events into three distinct categories. Novel Gene events are in a one category, “Distal Events” are events that appear near an annotated gene (Gene Extension events, Translated Untranslated Region (UTR) events, and Antisense Overlapping events). “Proximal Events” are events that overlap the coding region of an annotated gene (Frame Change events, Exon Extension events, Novel Exon events, and Novel Splice events).

The first filtering technique we considered was the standard N-peptide rule, allowing shared peptides to count equally with unique peptides. For Proximal Events, requiring two peptides per event was sufficient to retain less than 5% false positives. However, for the other event types, we see that the number of required peptides is larger than two peptides (to maintain a FDR < 5%). This demonstrates that treating shared and unique peptides equally is not an effective method for separating true events from decoy events.

The second filtering technique we considered was the N-unique peptide rule, which considers only 16,659 of the 24,782 novel peptides. Compared with the general N-peptide rule, the N-unique peptide rule better identifies true events from false events except in the case of Proximal Events. Proximal Events require supporting peptides to be very close to one another. For example, for a frame change event to have two supporting peptides, both peptides must be within the same exon.

Finally, we evaluated the eventProb. For both Novel Genes and Distal Events, the eventProb was overly conservative compared with the N-unique peptide rule. However, for Proximal Events, for which it is difficult to have two or more supporting peptides, the eventProb performs the best of all methods. The eventProb is able to “rescue” single peptide events that have strong spectral support. Ultimately, we chose a hybrid approach to event filtering: using the two-unique peptide rule for Novel Genes and Distal Events, and the eventProb for Proximal Events.

**Gene Model Prediction**—We used the gene prediction tool Augustus (version 2.5.5) to predict gene models based on evidence from filtered events, expressed sequence tags (ESTs), RNA-Seq, and homology. Proteins and ESTs were aligned to the maize genome using BLAT (24) with default parameters. RNA-Seq reads were aligned using Bowtie version 0.12.7 with default parameters and TopHat version 1.4.1. We accepted only reads that mapped uniquely. For gene refinement events, we also include the currently annotated gene model from the 5a.59 protein set as a hint. Peptides were labeled as manual sources of evidence (M) whereas others were labeled as expression sources of evidence (E) in the augustus config file. Augustus was run with single stranded predictions on.
RESULTS

We identified 225,166 distinct peptide sequences by searching the MS/MS data against a database of both annotated and putative protein sequences. The peptide sequences either confirmed annotated protein sequences in the 5a.59 maize proteome release or mapped to a genomic location that was not previously believed to be protein-coding. Based on the genomic-mapped peptide sequences, we codified eight distinct refinement “events”: Novel Gene, Novel Exon, Frame Change, Exon Extension, Gene Extension, Translated UTR, Novel Splice Junction, and Overlapping Antisense Translation.

Spectral library (25) search is an emerging approach to peptide identification that relies on an archive of identified tandem mass spectra. Comparing spectra to previously identified spectra is faster, more sensitive, and more accurate than comparing spectra to a sequence database. One major challenge facing spectral library search tools is the construction of an archive of mass spectra representing all possible peptides in the organism of interest. Our identifications would contribute over 2.7 million peptide spectrum matches to the construction of the first spectral library available for peptide identification in maize.

To control false positive novel event predictions, we used two types of event filters depending on the type of event. One filter is based on the popular “N unique peptide rule” that scores events based on the number of uniquely-located supporting peptides. The second filter is based on the eventProb that scores events based on supporting spectral evidence (see Experimental Procedures). Using the peptides from the filtered novel events, combined with over 2 million Zea mays ESTs, 875 million RNA-Seq reads (26, 27) and the alignment of annotated proteins from maize, rice, and sorghum, we

TABLE III

| Event type                      | Events identified | Genes affected |
|---------------------------------|-------------------|---------------|
| Antisense overlapping           | 318               | 246           |
| Frame change                    | 226               | 163           |
| Exon extension                  | 291               | 216           |
| Novel exon                      | 132               | 93            |
| Novel splice junction            | 37                | 28            |
| Translated UTR                  | 330               | 234           |
| Gene extension                   | 570               | 390           |

Fig. 3. Genome browser views of three refined gene models. A, We observed two peptides (dark blue) matching to the annotated pseudogene, GRMZM5G883336 (green), as well as 25 novel peptide sequences (light blue) downstream of the gene. The predicted sequence extends the annotated gene by over 400 amino acids (shown in orange). Our peptides confirmed that this locus was translated and should be reclassified as protein coding. Support from RNA-seq data sets available on maizesequence.org is shown in purple. B, We observed 44 peptides uniquely matching to two annotated genes, GRMZM5G881353 and GRMZM5G831724. In addition, we observed 114 uniquely located novel peptides in the UTR of GRMZM5G881353 as well as downstream of both genes. We predict a single gene at this location that is consistent with all peptides. C, The genome view of a newly predicted gene overlapping the annotated maize gene, AC212100.3_FG001. The predicted gene was supported by three novel peptides, including two that confirm introns.
generated putative gene structures using the *ab initio* tool, *augustus* (3).

**Validation of Existing Maize Annotations**—The maize proteome has two sets of annotated proteins. The filtered gene set (version 5b.60) contains 39,656 trusted gene predictions encoding 63,540 proteins. The working gene set (version 5a.59) contains an expanded, hypothetical set of genes and is a superset of the filtered set, containing a total of 136,770 proteins in 110,028 genes.

We identified 200,384 distinct peptide sequences matching one or more proteins in the filtered set or working set. These peptides confirmed the expression of 14,615 genes. The majority of these genes are from the filtered set (13,811 genes, 94%), suggesting that the 5b.60 annotation contains most maize genes expressed in seed tissue. Using uniquely mapping peptides, we determined that 10,604 specific protein isoforms from 10,507 genes were expressed, each with at least one uniquely mapping peptide. Again, the vast majority of the proteins identified were from the filtered protein set (9874 proteins).

We identified unique peptides in 730 working set proteins, suggesting that these proteins should be promoted to the trusted protein-coding set for future proteome releases. The remaining 4,108 identified genes had no uniquely mapping peptides. Every peptide that matched to these genes mapped to two or more protein isoforms of the gene.

Sources of evidence commonly used in the annotation of the gene sets include cDNA, ESTs, mRNA, *ab initio* predictions, and protein mappings from both maize and other species. We found that 65% of filtered set proteins had two or more types of evidence, whereas only 23% of working set proteins had as much evidence. The lack of evidence for most working set proteins is likely a contributing factor in the classification. Of the working set proteins we believe should be promoted, 45% had at least two other types of evidence. It appears that peptide mass spectrometry provides an orthogonal source of information that improves the identification of a trusted proteome. The full list of maize proteins and genes we identified in this study can be found in supplemental Table S1.

**Deep versus Broad Proteomic Sampling**—In a previous study (8), we demonstrated that broad sampling of a diverse set of tissues improves coverage of the proteome by MS/MS data. Here we evaluate the benefit of deep sampling of a small collection of tissues. In this study, we use a more conservative scoring scheme compared with the previous study. To create a fair comparison, we re-scored the *Arabidopsis* peptides using an identical procedure (see Experimental Procedures).

---

**Fig. 4.** *Refined gene model for GRMZM2G121186 with a frame change.* The predicted sequence is aligned to the annotated sequence at gene GRMZM2G121186 and the homolog sequence found in nr. The N-terminal portion of the prediction is consistent with the annotation, and we observed 47 peptides matching this region. We also identified 41 distinct novel peptide sequences matching the C terminus of the predicted sequence, which is in a different frame from the annotation. In the figure, the peptides are collapsed to conserve space. The coordinates of the predicted sequence do not differ significantly from the annotation. In fact, the prediction contains a late splice donor site, extending exon 9 by 16 nucleotides. Exon 10 in the prediction is identical in coordinates to exon 10 in the annotation, however, the difference in frame results in a dramatically different C terminus.
resulting in the identification of 128,432 distinct *Arabidopsis* peptides. Compared with the *Arabidopsis* study, the maize study analyzed five times more spectra, but identified only 50% more peptides. In total amino acids the maize proteome (version 5a.59) is 43% larger than the *Arabidopsis* proteome (TAIR7), which suggests that our growth in peptides naturally follows the increase in the number of potential proteins expressed. Our sampling may have reached the limit of detection by mass spectrometry, simply acquiring more spectra for the same peptide. We observed on average 12 clustered spectra (24 raw spectra) supporting each peptide in our maize study, compared with 19 raw spectra per peptide in *Arabidopsis*. In part, the diminished return in maize may be because of the much larger maize genome, which resulted in a protein sequence database 10 times larger than the *Arabidopsis* database. Larger search databases are known to reduce sensitivity (21, 22).

For proteogenomics, this raises the question of whether increasing the search space dramatically is worth the loss of sensitivity. In Jeong *et al.* (22), the authors determined that while database size appears to have little impact on FDR calculation accuracy, they observed an 11% decrease in identifications. These estimates are likely to be a bit more conservative than most proteogenomic experiments. In Jeong *et al.* the increase in database size was achieved by adding sequences that were unlikely to be found in the spectral datasets (e.g., adding *Arabidopsis thaliana* proteins while searching with spectra of the ISB Standard Protein Mix). In proteogenomics, we are adding sequences that likely contain peptides that are true matches to the spectra. However, the result indicates that searching expanded databases, as is done for proteogenomics, should only be performed when the goal is to identify novel proteins and not achieve broad proteome coverage. Searching a large, putative protein database is of most use for organisms without well-annotated proteomes.

Although broad sampling helps increase the diversity expressed peptides, deep sampling allows for robust, label-free quantification of each peptide. A Maize Protein Atlas describing the comparative abundance of proteins across seed samples was presented in a recent study (19). In addition to the number of identifications as database size increases. When increasing the size of the database from 3 million amino acids to 13.5 million amino acids (approximately a 4.5-fold increase), they observed an 11% decrease in identifications.

![FIG. 5](image)

**Fig. 5.** The alignment of a novel protein to the closest *nr* protein, [SORBIDRAFT_03g013090](http://example.com). The four identified peptides, bolded and highlighted in blue, are unique to this genomic location. The amino acid differences between the two proteins support the assignment of these peptides to the novel protein.
deeper sampling of each peptide, we also observed greater coverage for the identified proteins. Among all maize proteins that have at least one identified peptide (unique or shared), we achieve 27% amino acid coverage on average (13 peptides per protein). In contrast, in Arabidopsis we only achieved 20% coverage on average (nine peptides per protein). Multiple peptides identified for a protein can dramatically improve the spectral-count based quantification of proteins, and help in improved gene structure annotation.

**Discovery of Novel Events**—We identified 24,782 novel peptides (mapping to 91,059 locations in the genome) that did not match any protein in the filtered or working sets. Many of the novel peptides mapped to a single location in our genomic databases (16,659, 67.2%). The peptides were clustered based on co-location, and one or more events were called for each cluster, for a total of 2113 novel events. The identified events by type and the number of affected genes are shown in Table III and discussed below. Broadly, the clusters fell into two categories: clusters that were colocated within or near existing gene annotation, and clusters that fell in the intergenic region. Clusters of novel peptides overlapping or proximal to annotated genes were used to revise the gene structure, whereas intergenic peptide clusters indicated a novel locus being translated. Each novel event was scored using the eventProb, and the collection of events was filtered to a 5% FDR (see Experimental Procedures).

**Revision of Annotated Gene Models**—Using the novel events that were filtered to a 5% event-level, we constructed revised gene models with peptide support for 206 working set genes and 535 filtered set genes. We used augustus to predict updated gene models, using novel peptides, ESTs, RNA-Seq, homology with rice and sorghum, and the current gene models as hints. The complete list of predicted gene refinements is in supplemental Table S2.

---

**Fig. 6. Peptide-spectrum matches supporting a novel gene discovery.** A novel gene on chromosome 3 was identified with 4 distinct novel peptide sequences. All but one of the peptides was identified with multiple spectra. For each peptide, the best spectrum (with the lowest Spectral Probability) is shown labeled with the matching peptide.
A key challenge in genome annotation is distinguishing pseudogenes from protein coding genes. Pseudogenes are protein-coding genes that have lost their capacity to produce proteins, and therefore are considered non-functional. In 98 of our refined gene models, we found annotated pseudogenes that were incorrectly labeled. It is likely that many of these loci were labeled as pseudogenes because they are short or show little similarity with other proteins. For example, the gene GRMZM5G883336 is annotated as a pseudogene in the 5a.59 gene annotation. We propose that this gene is in fact translated, as we identified two peptides that match the protein produced at this locus, one of which matches uniquely. We also observed 25 distinct, uniquely-located novel peptides downstream of the gene (Fig. 3A). In this example, we predict that GRMZM5G883336 should be extended in the 3' direction by 419 amino acids. The alignment of the annotated gene sequence and the predicted gene sequence is shown in supplemental Fig. S1.

Determining whether a translated region is a single gene or two proximal genes is a difficult problem for gene annotation. We observed 81 instances where two or more genes were annotated, but given the evidence we predicted a single gene that contained the annotated sequences as substrings. For example, two peptides were identified in the 5' UTR of GRMZM5G881353 giving evidence for an unambiguous extension of the protein-coding region toward the adjacent annotated gene GRMZM5G831724. Given these peptides, we predicted an updated gene model that merges GRMZM5G831724 and GRMZM5G881353. In addition, we observed peptides downstream of the genes, suggesting a gene extension. Fig. 3B shows the two genes as well as the novel prediction. Both GRMZM5G831724 and GRMZM5G881353 are annotated as pseudogenes, however our peptide evidence suggests that these loci are translated. Analysis with blastp recovered a homolog in Sorghum bicolor, SORBIDRAFT_01g017695 (evalue 7E-47). The alignment between the two annotated proteins, GRMZM5G881353 and GRMZM5G831724, the predicted sequence, and the homolog is shown in supplemental Fig. S2.

Annotation pipelines that rely heavily on transcriptomics may fail to correctly determine the frame of translation. By contrast, proteomic evidence provides unambiguous frame identification. We identify novel peptides in 163 genes that suggest the annotated frame is incorrect. As an example consider GRMZM2G121186_P01, a nucleosome/chromatin assembly factor group A (nfa102), which is supported by 47 expressed peptides (12 peptides uniquely matching GRMZM2G121186_P01). However, we also identified 41 novel peptides overlapping the annotation but in a different frame. The proposed revision is consistent with the N terminus of the annotated protein, but the final two exons, consisting of 78 altered amino acids, are updated. A search of homologs in the NCBI nr database revealed a deposited maize sequence (NP_001105594.1) similar to the predicted sequence. Fig. 4 shows the alignment of the predicted sequence to the annotated sequence for GRMZM2G121186_P01. The nr sequence is likely to be an alternate annotation for the GRMZM2G121186 gene locus, but was not accepted for maize annotation version 5a.59.

Antisense Overlapping Genes—Overlapping natural antisense transcripts (NATs), once believed to predominantly exist in bacterial and viral genomes (28), have recently been observed in eukaryotes (29) including humans (30). Although most overlapping NATs are identified by transcriptional evidence (31) and homology (32), there are few documented cases of overlapping NATs in which both genes are translated. Many hypotheses attempt to explain how overlapping genes are regulated, most of which suggest that the simultaneous expression of both genes is unlikely (33, 34).

In the maize 5a.59 annotation, there are 10,724 overlapping antisense gene pairs in the working gene set. Some large genes overlap multiple genes on the opposite strand, resulting in a total of 20,159 genes overlapping another gene on the opposite strand. We observed 3470 pairs of maize genes in which one of the genes has peptide support. We also observed 230 overlapping pairs of genes where both genes are expressed.

On examining our predicted gene models we found 86 predicted genes that were proximal to an annotated gene on the opposite strand. Among these cases, we observed 22 genes with peptides also matching the gene on the annotated strand. We compared the predicted sequences to NCBI nr to identify potential homologs. We found that 18 of the predicted 110 transcripts had a strong sequence match (evalue < 1E-10).

As an example, on chromosome 2, we found three novel peptides on the reverse strand proximal to annotated gene AC212100.3_FG001. We found no peptides supporting the expression of AC212100.3_FG001 in our samples. We predicted a 407 amino acid protein on the reverse strand that is supported by the peptides, including two peptides that span splice junctions. A blastp search against NCBI nr revealed a homolog sequence in Zea mays (NP_001169025.1). The deposited sequence is derived from a prediction on a cDNA sequence and does not appear in the 5a.59 maize annotation. Fig. 3C shows the genomic region containing the newly predicted sequence. The full table of antisense overlapping transcript predictions can be viewed in supplemental Table S3.

Novel Genes—We identified 209 novel gene events. Occasionally, augustus chooses to split or merge some novel gene events based on the available evidence. In total, augustus predicted 218 novel protein sequences in 165 genes. Of these proteins, 133 were only supported by peptides and would not have been identified in an annotation pipeline based on the EST and RNA-Seq evidence alone. We used blastp to identify homologs of the novel protein sequences in NCBI non-redundant database (nr). We found that 106 of the 218 novel proteins have significant sequence similarity to a known protein.
Given the homolog sequences, we used augustus to further refine the novel gene predictions. The full list of predicted novel genes is available in supplemental Table S4.

One example is on chromosome 3, where we identified a novel gene encoding a protein with strong sequence similarity to GRMZM2G090086_P01 (evalue 4E-88). Although GRMZM2G090086_P01 is not functionally annotated at MaizeSequence.org, we believe this protein and the novel on chromosome 3, are translocase subunits SecA based on their similarity to the protein by that name (Gene ID: 100841935) in Brachypodium distachyon (evalue 0.0). Fig. 5 shows the alignment between our predicted novel protein and the closest homolog in NCBI nr, SORBIDRAFT_03g013090 (evalue = 0.0). The identified peptides are highlighted with blue. For each peptide, we selected the best spectrum match (the lowest Spectral eventProb) to showcase in Fig. 6.

DISCUSSION

We presented a probabilistic framework, the eventProb, for scoring annotation events derived from mass spectrometry data. Proteogenomic studies, for which broad and deep sampling of the proteome is a key target, generate enormous data sets. False positive identifications in these experiments arise from the millions of hypotheses that are being tested, one for each peptide-spectrum match. Although controlling the false discovery rate at the level of the peptide-spectrum match addresses this problem, errors may be propagated and amplified by protein or event inference. We demonstrated that our eventProb could be used to limit the false discovery rate at the refinement event-level without sacrificing sensitivity.

Zea mays, with a human-scale genome and significant sequence redundancy, presented a challenging target for our proteogenomic methods. The parallelization of the database search was crucial to making the analysis tractable. Engineering alone, however, is insufficient to address the problem of repetitive genomic regions. Peptides that map to multiple genomic locations are not discarded in our eventProb, but instead are discarded by the number of locations from which they may arise.

By applying our semi-automated genome annotation method to Zea mays, we demonstrated that proteomics provides a much needed line of evidence for the identification of protein-coding genes. Over 70% of the novel protein sequences supported by the peptide data were lacking corroborating transcriptomic and homology evidence.

Acknowledgments—We thank Professor Laurie G. Smith for helpful discussions and constructive comments on the paper.

REFERENCES

1. Wang, Z., Gerstein, M., and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63
2. Lomadoszat, A., Ter-Hovanessian, Y., Chernoff, Y. O., and Borodovsky, M. (2000) Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 33, 6494–6506
3. Stanke, M., Schöeffmann, O., Morgenstern, B., and Waack, S. (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics, 7, 62
4. Blanco, E., Parra, G., and Guigó, R. (2007). Using geneid to identify genes. Curr Protoc Bioinformatics, Chapter 4, Unit 4.3
5. Holt, C., and Yandell, M. (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12, 491
6. Eddy, S. R. Non-coding RNA genes and the modern RNA world. Nat. Rev. Genet. 2, 919–929, 2001
7. Castellana, N., and Bafna, V. (2010) Proteogenomics to discover the full coding content of genomes: a computational perspective. J. Proteomics 73, 2124–2135
8. Castellana, N. E., Payne, S. H., Shen, Z., Stanke, M., Bafna, V., and Briggs, S. P. (2008) Discovery and revision of Arabidopsis genes by proteogenomics. Proc. Natl. Acad. Sci. U.S.A., 105, 21034–21038
9. Baenertailer, K., Grossmann, J., Gobrei, M. A., Hull, R., Hirsch-Hoffmann, M., Yalovsky, S., Zimmermann, P., Grossniklaus, U., Gruissen, W., and Batista, S. (2008) Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. Science 320, 938–941
10. Merrihew, G. E., Davis, C., Ewing, B., Williams, G., Käll, L., Flewren, B. E., Noble, W. S., Green, P., Thomas, J. H., MacCoss, M. J. (2008) Use of shotgun proteomics for the identification, confirmation, and correction of C. elegans gene annotations. Genome Res. 18, 1680–1689
11. Brunner, E., Ahrens, C. H., Mohanty, S., Baetschmann, H., Loewenich, S., Potthatt, F., Deutsch, E. W., Panse, C., de Lichtenberg, U., Rinner, O., Lee, H., Pedrioli, P. G., Malmstrom, J., Koehler, K., Schirmpf, S., Krijgsveld, J., Kregenow, F., Heck, A. J., Hafen, E., Slabachpab, R., and Abersold, R. (2007) A high-quality catalog of the Drosophila melanogaster proteome. Nat. Biotechnol. 25, 576–583
12. Gupta, N., Benhamida, J., Bharagava, V., Goodman, D., Kain, E., Kermani, I., Nguyen, N., Ollikainen, N., Rodriguez, J., Wang, J., Lipton, M. S., Ro- mine, M., Bafna, V., Smith, R. D., and Pevzner, P. A. (2008) Comparative proteogenomics: combining mass spectrometry and comparative genetics to analyze multiple genomes. Genome Res. 18, 1133–1142
13. Gallien, S., Perroud, E., Carapito, C., Deshayes, C., Reyrat, J. M., Van Dorselaer, A., Poch, O., Schaeffer, C., and Lecompte, O. (2009) Ortho-proteogenomics: multiple proteomes investigation through orthology and a new MS-based protocol. Genome Res. 19, 129–135
14. Venter, E., Smith, R. D., and Payne, S. H. (2011) Proteogenomic analysis of bacteria and archaea: a 46 organism case study. PLoS ONE 6, e27587
15. Kelkar, D. S., Kumar, D., Kumar, P., Balakrishnan, L., Muthusamy, B., Yadav, A. K., Shrivastava, P., Marimuthu, A., Anand, S., Sundaram, H., Kingsbury, R., Harsha, H. C., Nair, B., Prasad, T. S., Chauhan, D. S., Katoch, K., Katoch, V. M., Kumar, P., Chaerkady, R., Ramachandran, S., Dash, D., and Pandey, A. (2011). Proteogenomic analysis of Mycobacterium tuberculosis by high resolution mass spectrometry. Mol. Cell Proteomics 10, M111.01627
16. Tanner, S., Shen, Z., Ng, J., Florea, L., Guigó, R., Briggs, S. P., and Bafna, V. (2007) Improving gene annotation using peptide mass spectrometry. Genome Res. 17, 231–239
17. Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., Minx, P., Reily, A. D., Courtney, L., Kruchowski, S. S., Tomlinson, C., Strong, C., Deleaultna, K., Fronick, C., Courneye, B., Rock, S. M., Belter, E., Du, F., Kim, K., Abbott, R. M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S. M., Gilliam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Than, T., Scione, A., Than, N., Hinkle, J., Want, T., Ruppert, J., Shah, N., Ritter, C., Hodges, J., Ingenthen, E., Cordes, M., Kohlpeter, S., Goro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasek, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidottti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A,
Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Abroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M. J., McMahan, L., Van Buren, P., Vaugh, M. W., Ying, K., Yeh, C. T., Emrich, S. J., Jia, Y., Kalyanaraman, A., Haia, A. P., Barbazuk, W. B., Baucom, R. S., Brutnell, T. P., Carpta, N. C., Chaparro, C., Chia, J. M., Deragon, J. M., Estill, J. C., Fu, Y., Jeddeloh, J. A., Han, Y., Lee, H., Li, P., Lisch, D. R., Liu, S., Liu, Z., Nagel, D. H., McCann, M. C., San Miguel, P., Myers, A. M., Nettleton, D., Nguyen, P. T., Pennings, B. W., Ponnala, L., Schneider, K. L., Schwartz, D. C., Sharma, A., Soderlund, C., Springer, N. M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T. K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J. L., Dawe, R. K., Jiang, J., Jiang, N., Presting, G. G., Wessler, S. R., Aluru, S., Martienssen, R. A., Clifton, S. W., McCombie, W. R., Wing, R. A., and Wilson, R. K. (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326, 1112–1115.

18. Meyers, B. C., Tingey, S. V., and Morgante, M. (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res. 11, 1660–1676.

19. Walley, J. W., Shen, Z., Sartor, R. C., Wu, K. J., Osborn, J., Smith, L. G., Briggs, S. P. Reconstruction of protein networks from an atlas of maize seed proteotypes. Proc. Natl. Acad. Sci. U.S.A. 110, E4808–E4817.

20. Tanner, S., Shu, H., Frank, A., Wang, L. C., Zandi, E., Mumby, M., Pevzner, P. A., and Bafna, V. (2005) InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. Anal. Chem. 77, 4626–4639.

21. Kim, S., Gupta, N., and Pevzner, P. A. (2008) Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. J. Proteome Res. 7, 3354–3363.

22. Jeong, K., Kim, S., and Bandeira, N. (2012) False discovery rates in spectral identification. BMC Bioinformatics 13, S2.

23. Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001) Empirical Bayes analysis of a microarray experiment. J. Am. Stat. Assoc. 96, 1151–1160.

24. Kent, W. J. (2002) BLAT—the BLAST-like alignment tool. Genome Res. 12, 656–664.

25. Lam, H. (2011) Building and searching tandem mass spectral libraries for peptide identification. Mol. Cell Proteomics 10, 12:R111.008565.

26. Li, P., Ponnala, L., Gandotra, N., Wang, L., Si, Y., Tausta, S. L., Kebrum, T. H., Provat, N., Patel, R., Myers, C. R., Reidel, E. J., Turgeon, R., Liu, P., Sun, Q., Nelson, T., and Brutnell, T. P. (2010) The developmental dynamics of the maize leaf transcriptome. Nat. Genet. 42, 1060–1067, 2010.

27. Davidson, R. M., Hansey, C. N., Gowda, M., Childs, K. L., Lin, H., Vaillancourt, B., Sekhon, R. S., de Leon, N., Kaepller, S., Jiang, N., and Buell, C. R. (2011) Utility of RNA-seq for analysis of maize reproductive transcriptomes. Plant Genome 4, 191–203.

28. Wagner, E. G., and Simons, R. W (1994) Antisense RNA control in bacteria, phages, and plasmids. Annu. Rev. Microbiol. 48, 713–742.

29. Williams, B. A., Slamovits, C. H., Patron, N. J., Fast, N. M., and Keeling, P. J. (2005) A high frequency of overlapping gene expression in compacted eukaryotic genomes. Proc. Natl. Acad. Sci. U.S.A. 102, 10936–10941.

30. Yelin, R., Dahary, D., Sorek, R., Levanon, E. Y., Goldstein, O., Sooshan, A., Diber, A., Biton, S., Tamir, Y., Khoosravi, R., Nemzer, S., Pinner, E., Wileach, S., Bernstein, J., Savitsky, K., and Rotman, G. (2003) Widespread occurrence of antisense transcription in the human genome. Nat. Biotechnol. 21, 379–386.

31. Fahey, M. E., Moore, T. F., and Higgins, D. G. (2002) Overlapping antisense transcription in the human genome. Comp. Funct. Genomics 3, 244–253.

32. Sanna, C. R., Li, W. H., and Zhang, L. (2008) Overlapping genes in the human and mouse genomes. BMC Genomics 9, 169.

33. Borsani, O., Zhu, J., Verslues, P. E., Sunkar, R., and Zhu, J. K. (2005) Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis. Cell 123, 1279–1291.

34. Osato, N., Suzuki, Y., Ikeo, K., and Gojobori, T. (2007) Transcriptional interferences in cis natural antisense transcripts of humans and mice. Genetics 176, 1299–1306.