

Improving Gene Model Accuracy for Nitrogen-Use Efficiency Genes in *Zea mays*

by

Russell Walden

A thesis presented to the Honors College of Middle Tennessee State University in partial fulfillment of the requirements for graduation from the University Honors College

Fall 2020

Improving Gene Model Accuracy for Nitrogen-Use Efficiency Genes in *Zea mays*

by Russell Walden

APPROVED:

Dr. Rebecca Seipelt-Thiemann, Thesis Director
Professor, Biology Department

Dr. John Dubois, Second Reader
Professor, Biology Department

Dr. Dennis Mullen, Thesis Committee Chair
Chair, Biology Department

ACKNOWLEDGEMENTS

I'd like to thank Dr. Rebecca Seipelt-Thiemann for her guidance through the research for and the writing of this thesis. Dr. Seipelt was always incredibly patient, kind, and helpful whenever I needed guidance, and I would like to thank her for that. I'd also like to thank my wonderful parents, my brother and sister Gabriel and Katerina, my grandmother Lynn Russell (Bom Bom), my girlfriend Jerrica and her parents James and Kelly Voyles and her brother and sisters Wyatt Voyles, Katie and Michael Hinson, and Alex and Mark Vibbert, my Aunt Dana, Uncle Frank, Aunt Candy, my cousins Brandon and Jodi Russell and their sons Kohan, Walker, and Ashton, my friend Hank Clement, and my roommate and friend Jack Shreeve for all of their support as well.

ABSTRACT

Zea mays is one of the most highly produced crops in the world. It plays a big role in food security, fuel production, and economic stability. A factor that contributes to the production cost and yield of *Zea mays* is its nitrogen-use efficiency. At least twenty genes in *Zea mays* are related to nitrogen-use efficiency. The purpose of this study was to improve the accuracy of existing models of nine genes relating to nitrogen-use efficiency in *Zea mays* by using data present in the Apollo gene annotation platform to inform changes required for updated gene models. These changes were made and used to produce a supertranscript for each gene. Alignment of proteins encoded by all possible transcripts was performed to identify differences in protein structure and domain presence where applicable. These models can be used to provide insight into the gene regulation and protein isoform function

TABLE OF CONTENTS

AKNOWLEDGMENTS.....	iii
ABSTRACT	iv
LIST OF FIGURES	vi
LIST OF TABLES	vii
INTRODUCTION	1
MATERIALS AND METHODS	4
RESULTS.....	9
CONCLUSIONS	53
LITERATURE CITED.....	55

LIST OF FIGURES

FIGURE 1: Example of gene structure models and RNA sequence data in the Apollo genome annotation platform.....	6
FIGURE 2: Example of workflow used to create improved gene models	8
FIGURE 3: Evidence for reannotation for gene Zm00001d031769	10
FIGURE 4: Alignment of proteins encoded by Zm00001d031769	13
FIGURE 5: Evidence for reannotation for gene Zm00001d049995	15
FIGURE 6: Alignment of proteins encoded by Zm00001d049995	17-19
FIGURE 7: Evidence for reannotation for gene Zm00001d018206	21
FIGURE 8: Alignment of proteins encoded by Zm00001d018206	23-24
FIGURE 9: Evidence for reannotation for gene Zm00001d017958	27
FIGURE 10: Alignment of proteins encoded by Zm00001d017958	29
FIGURE 11: Evidence for reannotation for gene Zm00001d022388	31
FIGURE 12: Alignment of proteins encoded by Zm00001d022388	33-35
FIGURE 13: Evidence for reannotation for gene Zm00001d052165	37
FIGURE 14: Alignment of proteins encoded by Zm00001d052165	39
FIGURE 15: Evidence for reannotation for gene Zm00001d018161	41
FIGURE 16: Alignment of proteins encoded by Zm00001d018161	43
FIGURE 17: Evidence for reannotation for gene Zm00001d025984	45
FIGURE 18: Alignment of proteins encoded by Zm00001d025984	47
FIGURE 19: Evidence for reannotation for gene Zm00001d028750	49
FIGURE 20: Alignment of proteins encoded by Zm00001d028750	51-52

LIST OF TABLES

TABLE 1: Summary of twenty genes involved in nitrogen-use efficiency in *Zea mays* 5

TABLE 2: Changed features of genes 11

Introduction

Grains such as corn, rice, and wheat play important roles in food security, fuel production, and economic stability. Corn is the world's largest grain crop with world production of maize in 2013-14 at 967 million metric tons (Shah *et al.* 2016). It is considered a staple food in numerous parts of the world and is the third leading crop in the world after rice and wheat (Shah *et al.* 2016). Approximately 25% of U.S. corn croplands are used for ethanol production (Mumm *et al.* 2014). Corn distillers' oil is a by-product of starch and ethanol production in corn and this oil can be used to synthesize biodiesel (Veljković *et al.* 2018). In relation to the economy, the B73 maize reference sequence promises to advance basic research and to facilitate efforts in an era of global climate change (Schnable *et al.* 2009).

One nutrient that is critical for development and growth of many agriculturally important plants, including *Zea mays*, is nitrogen (Sharma and Bali 2018). Corn's ability to deplete the soil of nitrogen requires addition of fertilizer to the soil, as well as crop rotation to maintain agriculture sustainability. Both of these solutions have accompanying problems. Crop rotation reduces annual yield and addition of fertilizers has been shown to have negative environmental impacts, such as ecosystem disruption and soil acidification (Singh 2018). A new approach to improve corn growth and yield without the disadvantages is to utilize new genome-level data to investigate the metabolic and physiological aspects of corn's nitrogen use with the aim of engineering new corn varieties to use less nitrogen (Simons *et al.* 2014).

To understand how corn uses nitrogen, metabolic and physiological pathways, including the genes that encode the enzymes within these pathways, must be accurately

understood at the genetic, transcript, protein, and regulatory levels. In 2009, the corn genome was sequenced, which contains over 32,000 predicted genes, and resulted in the B73 reference genome version 1 (B73 reference genome version 5 was used as reference data in this study) (Schnable *et. al.* 2009). 32,540 protein-encoding genes were predicted from assembled or improved bacterial artificial chromosome (BAC) contigs by a combination of evidence based and ab initio approaches (Schnable *et. al.* 2009). Maize was also found to exhibit extremely high levels of both phenotypic and genetic diversity (Schnable *et. al.* 2009). Extensive structural variation, including hundreds of copy number variants (CNVs) and thousands of present-absent variants (PAVs) were revealed by resequencing and array-based comparative genomic hybridization between the B73 and Mo17 inbred lines (Schnable *et. al.* 2009).

While the corn genome has been sequenced, and the genes computationally predicted and revised, transcript evidence showing how and when gene products are produced in different corn tissues is just now becoming available (Ware, personal communication) (Xu *et al.* 2009). Recent advances in RNA analyses called RNA sequencing have enable scientists to assay the entire pool of RNAs produced within a particular tissue or under a distinct condition, the transcriptome. By comparing the RNAs produced across tissues, developmental stages, and conditions, a full picture of how each gene's structure could differ at the RNA level though alternative promoter use, alternative splicing, alternative transcription termination, and alternative polyadenylation can be identified. Additionally, this information would then form the basis for regulatory studies to determine how these transcripts are generated, the protein isoforms encoded, and the functional role each plays in corn life. This, in turn, could then be used to

identify points in metabolic pathways that could be better engineered for more efficient or different use of the plant's resources.

The goal of this study was to use existing genome, gene, and tissue-specific RNA evidence to develop supertranscript gene models for nine genes that are known to be involved in nitrogen use efficiency, nitrogen assimilation, amino acid metabolism, and auxin signaling which would then be tested experimentally and corrected within the community-based Apollo genome annotation platform (Tello-Ruiz *et al.* 2017). This study contributes to research to determine the genomic structure, expression, and regulation of all corn genes, but especially the nitrogen-use efficiency (NUE) genes which may then be targeted for genetic based improvements to corn agriculture sustainability.

Materials and Methods

NUE Genes

The names and genome coordinates of a selection of 20 genes relating to nitrogen use efficiency, nitrogen assimilation, amino acid metabolism, and auxin signaling in *Zea mays* genome version 5 (Table 1) were provided by Dr. Doreen Ware of Cold Spring Harbor Laboratories (personal communication).

Data Viewer

The *Zea mays* version 5 (v5) gene models (Figure 1), as well as multiple short and long-read RNA evidence tracks including: RNA sequencing data for six tissues, each with six replicates, IsoSeq RNA sequencing, and full-length RNA sequencing were visualized within the Apollo genome viewer and annotation platform. Access was provided by Cold Spring Harbor Laboratory (CSHL) (Ware, personal communication). Briefly, different gene features are represented visually in this user space (Figure 1A). Within the gene models, exons are denoted by rectangles, while introns and intergenic spaces are denoted by horizontal lines. The direction of transcription is noted with an arrowhead at the 3' end of the terminal exon. Within transcriptome data, a single read aligned to its corresponding genome sequence is positioned below the gene model in exactly the position identified by genome sequence. Multiple reads aligning to the same region are stacked vertically to show “depth of coverage” for a particular genomic region. Because alternative splicing joins non-contiguous sequences, for an individual read, the sequence that is actually present in the read is noted as a rectangle, and the implied sequence is noted as a horizontal line.

Table 1. Summary of twenty genes involved in nitrogen-use efficiency in *Zea mays*.
Gene coordinates and functions are from Gramene (Tello-Ruiz *et al.* 2018).

	Gene ID#	Gene Coordinates	Function/Description
1	Zm00001d031769	chr1:203089608..203094203	Nitrate reductase [NADH] 2
2	Zm00001d049995	chr4:58905791..58910442	Nitrate reductase
3	Zm00001d018206	chr5:219126820..219129945	Nitrate reductase [NADH] 2
4	Zm00001d048050	chr9:152280180..152287969	Glutamine synthetase 3 isoform 1%3B Glutamine synthetase 3 isoform 2
5	Zm00001d028260	chr1:27922075..27924590	Glutamine synthetase 6
6	Zm00001d017958	chr5:213469469..213473069	Glutamine synthetase root isozyme 3
7	Zm00001d022388	chr7:180078625..180096312	Ferredoxin-dependent glutamate synthase%2C chloroplastic
8	Zm00001d011610	chr8:155140627..155152027	Glutamate synthase 1 [NADH] chloroplastic
9	Zm00001d043845	chr3:213907196..213918549	Glutamate synthase 1 [NADH] chloroplastic
10	Zm00001d038948	chr6:174758495..174770022	Glutamate synthase 1 [NADH] chloroplastic
11	Zm00001d038948	chr4:183572732..183576346	Nitrite reductase 2
12	Zm00001d018161	chr5:218273575..218276598	Ferredoxin--nitrite reductase chloroplastic
13	Zm00001d034420	chr1:294101398..294107663	Glutamate dehydrogenase
14	Zm00001d025984	chr10:136652035..136656257	Glutamic dehydrogenase 2
15	Zm00001d002052	chr2:5280850..5283996	Probable isoaspartyl peptidase/L-asparaginase 2
16	Zm00001d028750	chr1:44908940..44912682	Asparagine synthetase 3
17	Zm00001d045675	chr9:34625498..34634881	Asparagine synthetase 1
18	Zm00001d022152	chr7:174854238..174860476	Alanine aminotransferase 9
19	Zm00001d014258	chr5:38654080..38663968	Alanine aminotransferase 5
20	Zm00001d007357	chr2:229098421..229111535	Protein Auxin Signaling F-Box 3

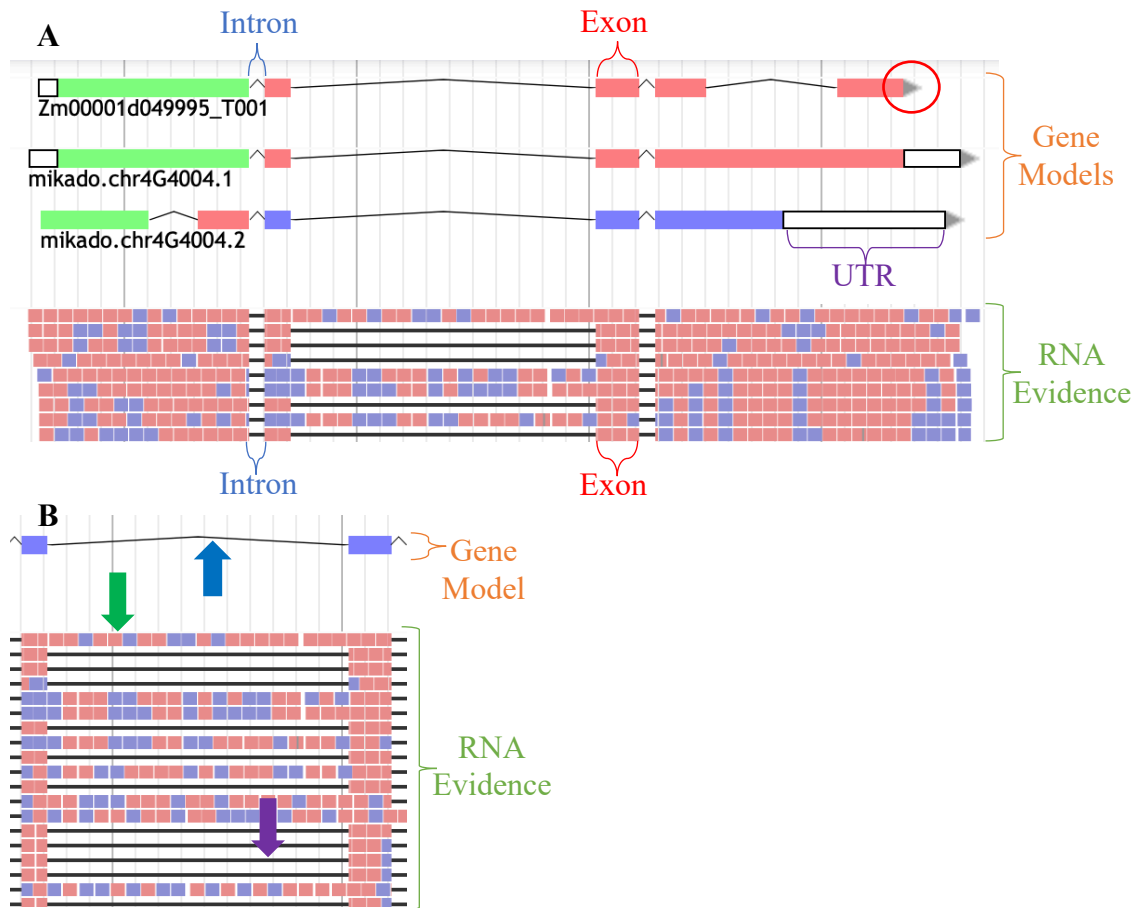


Figure 1. Example of gene structure models and RNA sequence data in the Apollo genome annotation platform. Panel A shows an example of gene structure models and RNA sequence data. Introns are denoted in both the gene models and in the RNA sequence data by black lines (Blue bracket). Exons are denoted by colored boxes (Red bracket). Gene models are at the top (Orange bracket) and have been produced previously using RNA evidence data (Green bracket). Boxes filled with white denote untranslated regions (UTR, purple bracket). Arrows show the direction of transcription for the model (Red circle) which means the arrow is at the end of the transcript and is at the 3' end of the model. In panel B, the model shows presence of an intron between two exons (blue arrow). However, some RNA evidence suggests the presence an exon (green arrow) and some evidence suggests the presence of an intron (purple arrow).

Data Analysis and Consolidation

A collapsed gene model (or SuperTranscript) (Davidson *et al.* 2017) was created for each gene (Figure 2). First, the PASA-informed v5 gene model was compared to the RNA evidence in the consolidated Mikado RNA evidence, which represents RNA evidence from six replicates of six different corn tissues (Figure 2A). First, existing genome data in the Apollo genome annotation platform was compared to models to first determine the model accuracy and then to look at evidence that changes were necessary. Next, aligned RNA sequencing reads were visualized one set at a time to determine gather evidence for the gene features represented in the v5 model and any features not represented in the v5 model (Figure 2B). Evidence was in the form of aligned transcript reads from RNA sequencing experiments and included large reads (IsoSeq) and short reads from six tissues, as well as full-length RNAs. Possible changes include: retained intron, alternative 5' splice site, alternative 3' splice site, alternative exon, and alternative intron. Next, the gene features with adequate evidence in any of the RNA data was collapsed visually to form the SuperTranscript with exonic features noted as rectangles, alternative exonic regions noted as hatched rectangle areas, introns noted as horizontal lines, and exons numbered beginning with 1 at the first transcribed nucleotide (Figure 2C). This SuperTranscript was a diagrammatic representation of all transcript evidence combined with the new PASA-informed v5 gene model. It essentially allowed for visualization of all genomic regions that were retained in the mature mRNA for all transcripts in all tissues where data were available. All possible transcripts were then constructed with the Apollo user space to be available to the research community (Figure 2D).

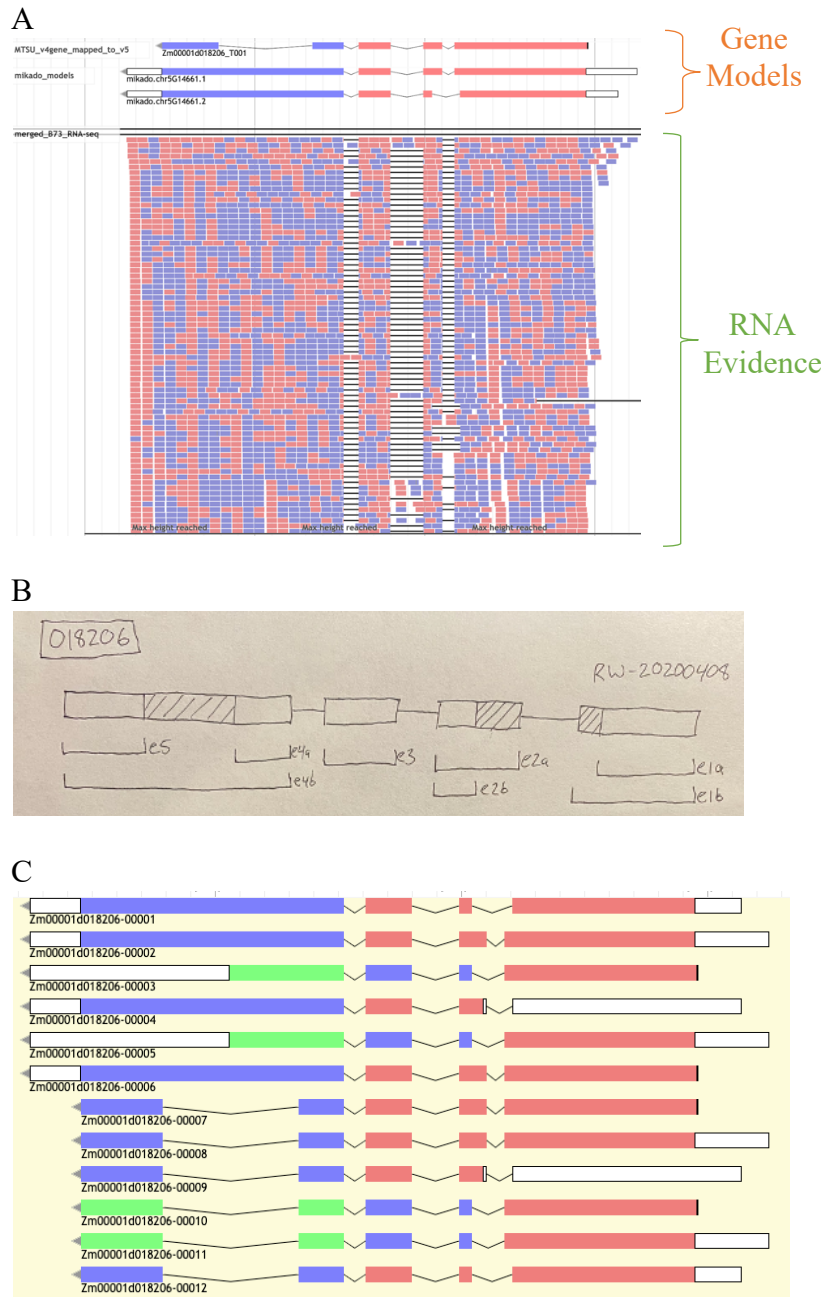


Figure 2 – Example of workflow used to create improved gene models for gene Zm000018206 – Panels A shows some of, but not all of the data used as evidence for reannotation of gene models. Previously compiled gene models are shown (orange bracket) with RNA sequencing data underneath (green bracket). Panel B shows an example of a hand-drawn collapsed gene model, or SuperTranscript, A SuperTranscript is a visual representation of the similarities and differences between all possible transcripts. Hashed boxes denote features that differ between possible transcripts. Panel C shows an example of transcripts constructed in the Apollo user-created annotations panel. This shows all the ways RNA could be constructed by the cell and was constructed to produce the amino acid sequence of each transcript to generate protein alignments.

Results

The first gene that was evaluated was Zm00001d031769, which encodes nitrate reductase [NADH] 2 (Tello-Ruiz *et. al.* 2017). The original gene structure was compared with the data in the Apollo annotation platform and a model was produced as noted in the methods (Figure 3). Comparison of the v5 and Mikado RNA models showed the extension of the gene with addition of an intron and exon, as well as an alternative 5' splice site (Figure 3A). RNA sequencing evidence (Figure 3B) supported both the presence and absence of an intron at that location, so it was labeled as a retained intron in the revised model (*, Figure 3C, Table 2). Additionally, the data supported the presence of the additional 3' terminal exon (Figure 3A, 3B), so this was labeled as an alternative exon in the revised model (**, Figure 3C; Table 2). Finally, a new 5' splice site was identified in the Mikado RNA model (Figure 3A) and supported in the RNA sequencing evidence, (Figure 3B) and incorporated into the SuperTranscript (***, Figure 3C).

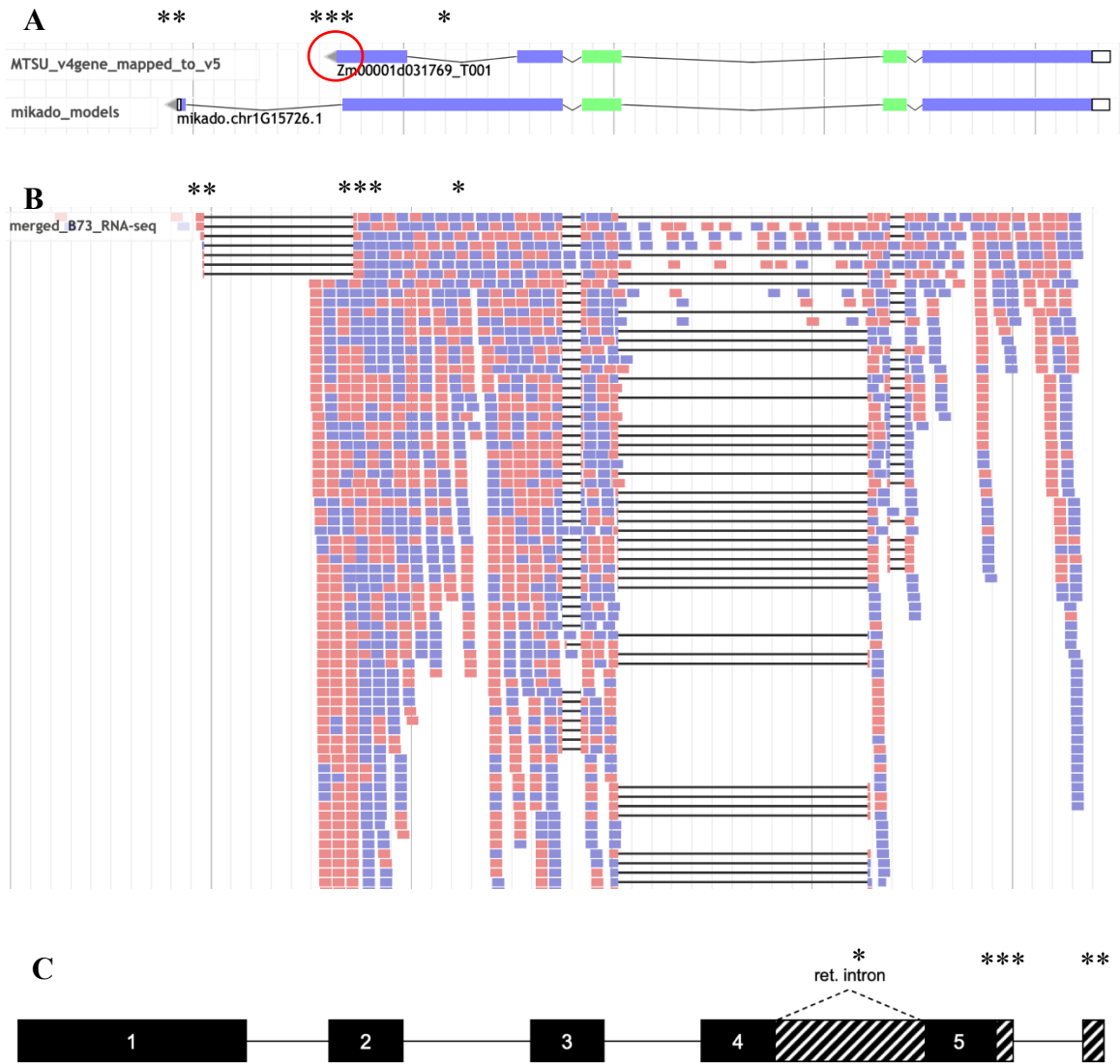


Figure 3. Evidence for reannotation for gene Zm00001d031769. Panel A shows the original structure of gene Zm00001d031769. Notice that the arrow on the original structure in panel A indicates the 3' end is on the left side of the gene (red circle) so the direction of the gene will be the reversed in the final model. The green and purple boxes show differentiation between the gene's reading frames. Panel B shows the data that was used to determine changes to be made to produce the final model. Panel C shows the final revised model. Note that the final model shows the RNA strand 5' to 3', which is the reverse of how the original model is displayed. The numbered black boxes indicate exons and their sequence from left to right. The lines between exons indicate introns.

Table 2. Changed features of genes.

Gene name	Features changed
Zm00001d031769	Retained intron between exons 4 and 5. Alternate 5' splice site after exon 5. Alternative exon at 3' end.
Zm00001d049995	Retained introns between exons 1 and 2, 3 and 4, and 5 and 6.
Zm00001d018206	Alternate 5' splice site after exon 1. Alternate 3' splice site before exon 2. Retained intron between exons 4 and 5.
Zm00001d017958	Alternate 3' splice sites before exons 2, 3, 4, 6, and 10. Two alternate 3' splice sites before exon 9. Alternate 5' splice sites after exons 5, 8, and 9.
Zm00001d022388	Alternative (or cassette) exons between exons 9 and 12. Alternative intron inside exon 28. Alternative 5' splice site after exon 30.
Zm00001d052165	Alternate start of transcription site. Retained intron between exons 1 and 2, and 2 and 3. Alternate end of transcription site after exon 5.
Zm00001d018161	Alternate 3' splice site before exon 3.
Zm00001d025984	Alternate 3' splice sites before exons 2 and 8.
Zm00001d028750	Retained introns between exons 1 and 2, 6 and 7, and 10 and 11. Two alternate end of transcription sites with possible retained intron.

Using the SuperTranscript as a guide, three RNAs were constructed in the Apollo space. The encoded amino acid sequences for each transcript were then downloaded from the Apollo workspace and aligned using Clustal Omega (Sievers *et. al.* 2011) and shaded using BOXSHADE to allow visualization of similarities and differences between the proteins encoded by the different transcripts. The three transcripts produce proteins that differ in the central region (Figure 4). Transcript 3 encodes a protein missing 216 amino acids compared to transcript 1. Transcript 2 encodes a protein missing 218 amino acids compared to transcript 1. Lastly, the sequence of amino acids encoded by each transcript was analyzed for the presence of known protein domains using the SMART domain sequence analysis program (Letunic and Bork 2017). The cytochrome b5-like Heme/Steroid binding domain was present in transcript 1, but not transcript 2 or 3, suggesting transcript 1, but not transcript 2 or 3, encodes a functional protein.

```

31769_3 1 MAAVEPRQFGRLEPARVGAYPPPPSHIPRRADSPARGCGFPPLVSPPRSTSDASSSDDE
31769_1 1 MAAVEPRQFGRLEPARVGAYPPPPSHIPRRADSPARGCGFPPLVSPPRSTSDASSSDDE
31769_2 1 MAAVEPRQFGRLEPARVGAYPPPPSHIPRRADSPARGCGFPPLVSPPRSTSDASSSDDE

31769_3 61 QDDWRELYGSQQLQLEVEPAAQDPDEGTADAWVERNPCLVRLTGKHLNCEPPLARLMQH
31769_1 61 QDDWRELYGSQQLQLEVEPAAQDPDEGTADAWVERNPCLVRLTGKHLNCEPPLARLMQH
31769_2 61 QDDWRELYGSQQLQLEVEPAAQDPDEGTADAWVERNPCLVRLTGKHLNCEPPLARLMQH

31769_3 121 GFITPAPLHYVRNHGAVPRGDWATWAVEVTGLVRRPARLTMDELARDFPAVEIPVTLACA
31769_1 121 GFITPAPLHYVRNHGAVPRGDWATWAVEVTGLVRRPARLTMDELARDFPAVEIPVTLACA
31769_2 121 GFITPAPLHYVRNHGAVPRGDWATWAVEVTGLVRRPARLTMDELARDFPAVEIPVTLACA

31769_3 181 GNRKEQNMVRQTAGFGWCAAGVSTSVWRGARLRDLVLRRCGVAPRHGGALNVCFEGAEDL
31769_1 181 GNRKEQNMVRQTAGFGWCAAGVSTSVWRGARLRDLVLRRCGVAPRHGGALNVCFEGAEDL
31769_2 181 GNRKEQNMVRQTAGFGWCAAGVSTSVWRGARLRDLVLRRCGVAPRHGGALNVCFEGAEDL

31769_3 241 PGGGGGSKYGTSPREWALDPSRDIMLAYMQNGEPLLPDHGFPVRVVIIPGCIIGRMVKW
31769_1 241 PGGGGGSKYGTSPREWALDPSRDIMLAYMQNGEPLLPDHGFPVRVVIIPGCIIGRMVKW
31769_2 241 PGGGGGSKYGTSPREWALDPSRDIMLAYMQNGEPLLPDHGFPVRVVIIPGCIIGRMVKW

31769_3 301 LKRIVVTPAESDNYHYKDNRLPSHVDALANAEAWWKPEYIINELNINSVITTPGHD
31769_1 301 LKRIVVTPAESDNYHYKDNRLPSHVDALANAEAWWKPEYIINELNINSVITTPGHD
31769_2 301 LKRIVVTPAESDNYHYKDNRLPSHVDALANAEAWWKPEYIINELNINSVITTPGHD

31769_3 361 EILPINGITTQRGYTMKGAYSGGKKVTRVEVTLDGGETWLVCDLAHEKPNKYGYWC
31769_1 361 EILPINGITTQRGYTMKGAYSGGKKVTRVEVTLDGGETWLVCDLAHEKPNKYGYWC
31769_2 361 EILPINGITTQRGYTMKGAYSGGKKVTRVEVTLDGGETWLVCDLAHEKPNKYGYWC

31769_3 421 WCFWSVEVEVLDLLGAKEIAVRAWDQSLNTQPERLIWNLMGMNNCWFKVKVNVCRPHRG
31769_1 421 WCFWSVEVEVLDLLGAKEIAVRAWDQSLNTQPERLIWNLMGMNNCWFKVKVNVCRPHRG
31769_2 421 WCFWSVEVEVLDLLGAKEIAVRAWDQSLNTQPERLIWNLMGMNNCWFKVKVNVCRPHRG

31769_3 481 EIGLVFEHPTQPGNQAGGWMARQKHLEKTAEEAAPGLKRSTSTPFMSTTDGGHQQLTMSE
31769_1 481 EIGLVFEHPTQPGNQAGGWMARQKHLEKTAEEAAPGLKRSTSTPFMSTTDGGHQQLTMSE
31769_2 481 EIGLVFEHPTQPGNQAGGWMARQKHLEKTAEEAAPGLKRSTSTPFMSTTDGGHQQLTMSE

31769_3 541 VSRHASRDS-----
31769_1 541 VSRHASRDSAWVVVHGHVYDCTRFLRDHPGGADSIILINAGTDCTEEFDAIHSKAKALLD
31769_2 541 VSRHASRDS-----

31769_3 550 -----
31769_1 601 AYRVGELIATGTSDSSVHGGSALPSHLLAPIREAAAPALSGPRDKVRCRLVGRTELS
31769_2 550 -----

31769_3 550 -----
31769_1 661 RDVRLRLRSLPSPDQALGLPIGKHSVCSIDGKLCMRAYTPTSVADEVGHFDDLKVYVF
31769_2 550 -----

31769_3 550 -----ADRYARRLAMVCG
31769_1 721 RDEHPKFPSSGGLMTQHLDLPLGSCIDVKGPLGHVEYTGGRGFVIDGRDRYARRLAMVCG
31769_2 550 -----ADRYARRLAMVCG

31769_3 563 GSGITPMYQVIQAVLRDQPEDRTEMHLVYANRTEDDILLRDELDRCAAEYPDLKVVYVV
31769_1 781 GSGITPMYQVIQAVLRDQPEDRTEMHLVYANRTEDDILLRDELDRCAAEYPDLKVVYVV
31769_2 563 GSGITPMYQVIQAVLRDQPEDRTEMHLVYANRTEDDILLRDELDRCAAEYPDLKVVYVV

31769_3 623 DVQKRPEEGWKYSVGFVTEDVLRHVPEGGDDTLALACGPPPMIQFAVSPNLEKMKYDMA
31769_1 841 DVQKRPEEGWKYSVGFVTEDVLRHVPEGGDDTLALACGPPPMIQFAVSPNLEKMKIFLE
31769_2 623 DVQKRPEEGWKYSVGFVTEDVLRHVPEGGDDTLALACGPPPMIQFAVSPNLEKMKIFLE

31769_3 683 NSFVVF
31769_1 901 APIG--
31769_2 683 APIG--

```

Figure 4. Alignment of proteins encoded by Zm00001d031769. Amino acid sequence of gene Zm00001d031769 showing three different versions of the protein that could be produced by the DNA sequence based on the revised model. The absence sequence is denoted by dashes, meaning the next amino acid in the chain would be the next letter after the dashes. Identities between different versions of protein are denoted by the shaded boxes. Note also that the boxes may be shaded even if the sequence is “skipped” in the comparison protein.

The next gene that was evaluated was Zm00001d049995, which encodes nitrate reductase (Tello-Ruiz *et. al.* 2017). The original gene structure was compared with the data in the Apollo annotation platform and a model was produced as noted in the methods (Figure 5). Comparison of the v5 and Mikado RNA models showed the shortening of the gene with addition of introns (Figure 5A). Mikado models supported both the presence and absence of an intron (*, Figure 5A), and was supported by RNA sequencing data showing the absence of an intron (*, Figure 5B), so it was labeled as a retained intron in the revised model (*, Figure 5C, Table 2). Additionally, v5 evidence supported the presence of an intron (***, Figure 5A), while RNA sequencing evidence (***, Figure 5B) supported the absence of an intron at that location, so it was labeled as a retained intron in the revised model (***, Figure 5C; Table 2). Finally, RNA sequencing data suggesting both the presence and absence of an intron was found (**, Figure 5B), and was then incorporated into the SuperTranscript and labeled as a retained intron in the revised model (**, Figure 5C, Table 2).

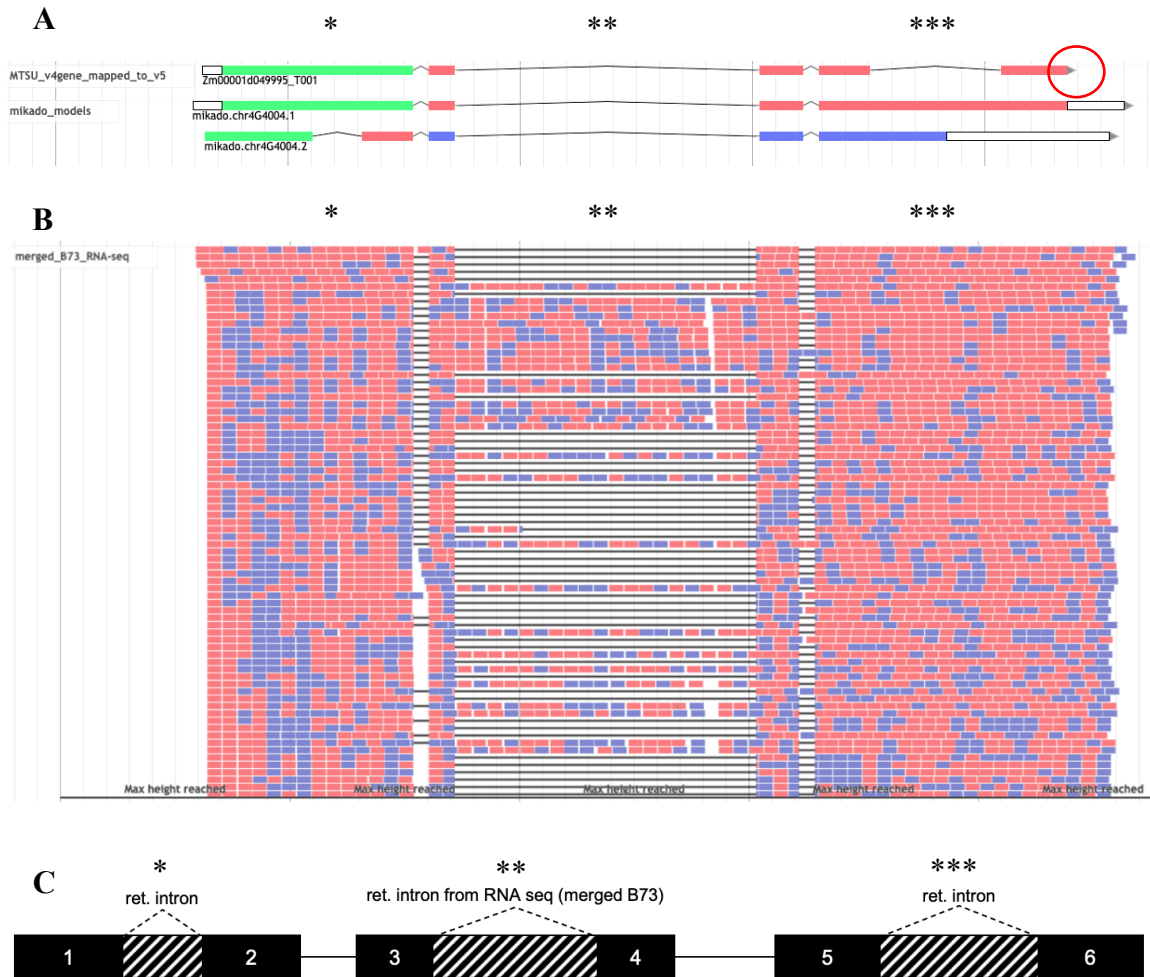


Figure 5. Evidence for reannotation for gene Zm00001d049995. Panel A shows the original structure of gene Zm00001d049995. Notice that the arrow on the original structure in panel A indicates the 3' end is on the right side of the gene (red circle) so the direction of the gene will be the same in the final model. The green, red, and purple boxes show differentiation between the gene's reading frames. Panel B shows the final revised model. Note that the final model shows the RNA strand 5' to 3', which is the same as how the original model is displayed. The numbered black boxes indicate exons and their sequence from left to right. The lines between exons indicate introns.

Using the SuperTranscript as a guide, eight RNAs were constructed in the Apollo space. The encoded amino acid sequences for each transcript were then downloaded from the Apollo workspace and aligned using Clustal Omega (Sievers *et. al.* 2011) and shaded using BOXSHADE to allow visualization of similarities and differences between the proteins encoded by the different transcripts. The eight transcripts produce proteins that differ in various regions (Figure 6). Transcripts 3 and 7 both encode the same protein which is 400 amino acids shorter than transcript 5. Transcript 6 encodes a protein which is 235 amino acids shorter compared to transcript 5. Transcript 2 encodes a protein missing 322 amino acids compared transcript 5. Transcript 1 encodes a protein missing 303 amino acids compared to transcript 5. Transcript 4 encodes a protein missing 608 amino acids compared to transcript 5. Lastly, the sequence of amino acids encoded by each transcript was analyzed for the presence of known protein domains using the SMART domain sequence analysis program (Letunic and Bork 2017). The cytochrome b5-like heme/steroid binding domain was found present in the transcripts.

```

49995_2 1 RALSSIPAFPGQLQLLAHHQKPERTPATHLPEMAAVEPRQFGRLEPGSSPVRVATNGAKA
49995_1 1 RALSSIPAFPGQLQLLAHHQKPERTPATHLPEMAAVEPRQFGRLEPGSSPVRVATNGAKA
49995_4 1 RALSSIPAFPGQLQLLAHHQKPERTPATHLPEMAAVEPRQFGRLEPGSSPVRVATNGAKA
49995_8 1 RALSSIPAFPGQLQLLAHHQKPERTPATHLPEMAAVEPRQFGRLEPGSSPVRVATNGAKA
49995_3 1 -----
49995_7 1 -----
49995_5 1 RALSSIPAFPGQLQLLAHHQKPERTPATHLPEMAAVEPRQFGRLEPGSSPVRVATNGAKA
49995_6 1 RALSSIPAFPGQLQLLAHHQKPERTPATHLPEMAAVEPRQFGRLEPGSSPVRVATNGAKA

49995_2 61 YPPASHLPRRADSPVRGCSFPPLVSPRRLLDASDDEDEEQEDWRELYGSHLQLEVEPA
49995_1 61 YPPASHLPRRADSPVRGCSFPPLVSPRRLLDASDDEDEEQEDWRELYGSHLQLEVEPA
49995_4 61 YPPASHLPRRADSPVRGCSFPPLVSPRRLLDASDDEDEEQEDWRELYGSHLQLEVEPA
49995_8 61 YPPASHLPRRADSPVRGCSFPPLVSPRRLLDASDDEDEEQEDWRELYGSHLQLEVEPA
49995_3 1 -----
49995_7 1 -----
49995_5 61 YPPASHLPRRADSPVRGCSFPPLVSPRRLLDASDDEDEEQEDWRELYGSHLQLEVEPA
49995_6 61 YPPASHLPRRADSPVRGCSFPPLVSPRRLLDASDDEDEEQEDWRELYGSHLQLEVEPA

49995_2 121 VQDARDEGTADAWIERNPCLVRLTGKHLNCEPPLARLMHHGFI TPAPLHYVRNHGAVPR
49995_1 121 VQDARDEGTADAWIERNPCLVRLTGKHLNCEPPLARLMHHGFI TPAPLHYVRNHGAVPR
49995_4 121 VQDARDEGTADAWIERNPCLVRLTGKHLNCEPPLARLMHHGFI TPAPLHYVRNHGAVPR
49995_8 121 VQDARDEGTADAWIERNPCLVRLTGKHLNCEPPLARLMHHGFI TPAPLHYVRNHGAVPR
49995_3 1 -----
49995_7 1 -----
49995_5 121 VQDARDEGTADAWIERNPCLVRLTGKHLNCEPPLARLMHHGFI TPAPLHYVRNHGAVPR
49995_6 121 VQDARDEGTADAWIERNPCLVRLTGKHLNCEPPLARLMHHGFI TPAPLHYVRNHGAVPR

49995_2 181 GDWATWTVEVTGLVRRRGIQVRHQHARVGPFPVAGHHARLHAERRAAAGPRLPRAR--
49995_1 181 GDWATWTVEVTGLVRRRGIQVRHQHARVGPFPVAGHHARLHAERRAAAGPRLPRAR--
49995_4 181 GDWATWTVEVTGLVRRRGIQVRHQHARVGPFPVAGHHARLHAERRAAAGPRLPRAR--
49995_8 181 GDWATWTVEVTGLVRRRGIPARLTMEELA-RDFPA-----VEIPVTLACAGNRKRQNMV
49995_3 1 -----
49995_7 1 -----
49995_5 181 GDWATWTVEVTGLVRRRGIPARLTMEELA-RDFPA-----VEIPVTLACAGNRKRQNMV
49995_6 181 GDWATWTVEVTGLVRRRGIPARLTMEELA-RDFPA-----VEIPVTLACAGNRKRQNMV

49995_2 239 HHPLRHRWPHGQVAQAHHRHPRRVROLLPFQGGQPRPAVARRRRARQRRSVVVQAC---V
49995_1 239 HHPLRHRWPHGQVAQAHHRHPRRVROLLPFQGGQPRPAVARRRRARQRRSVVVQAC---V
49995_4 239 HHPLRHRWPHGQVAQAHHRHPRRVROLLPFQGGQPRPAVARRRRARQRRSVVVQAC---V
49995_8 232 QQTVGFNWGAAGVSTSVWRGA-RLRDVLRR-----CGTVPRKCGALNV
49995_3 1 -----
49995_7 1 -----
49995_5 232 QQTVGFNWGAAGVSTSVWRGA-RLRDVLRR-----CGTVPRKCGALNV
49995_6 232 QQTVGFNWGAAGVSTSVWRGA-RLRDVLRR-----CGTVPRKCGALNV

49995_2 295 HHQRAEHKLGDNDA GARRDPAHQQHHTARLHHERIRLLRRRQEGDAGGGDAGRRRDMAG
49995_1 295 HHQRAEHKLGDNDA GARRDPAHQQHHTARLHHERIRLLRRRQEGDAGGGDAGRRRDMAG
49995_4 295 HHQRAEHKLGDNDA GARRDPAHQQHHTARLHHERIRLLR-----
49995_8 274 CFEGAEDLPGGG--GSKYGTSVTREW---ALDPSRDIMLAYMONGEP-----
49995_3 1 -----
49995_7 1 -----
49995_5 274 CFEGAEDLPGGG--GSKYGTSVTREW---ALDPSRDIMLAYMONGEP-----
49995_6 274 CFEGAEDLPGGG--GSKYGTSVTREW---ALDPSRDIMLAYMONGEP-----

```

Figure 6. Alignment of proteins encoded by Zm00001d049995 – Part 1. Amino acid sequence of gene Zm00001d049995 showing eight different versions of the protein that could be produced by the DNA sequence based on the revised model. The absence sequence is denoted by dashes, meaning the next amino acid in the chain would be the next letter after the dashes. Identities between different versions of protein are denoted by the shaded boxes. Note also that the boxes may be shaded even if the sequence is “skipped” in the comparison protein.

```

49995_2 355 VPPRPPGEAQQVRQVLVLLVRRGGPRPARRQGDRRARMGPVAQHPAREAHME-----
49995_1 355 VPPRPPGEAQQVRQVLVLLVRRGGPRPARRQGDRRARMGPVAQHPAREAHME-----
49995_4 -----
49995_8 316 ---LLPDHCFPVRVII-----PGCIGGRM-----VKWLKRIVTPAESDNYHFKDN
49995_3 1 -----
49995_7 1 -----
49995_5 316 ---LLPDHCFPVRVII-----PGCIGGRM-----VKWLKRIVTPAESDNYHFKDN
49995_6 316 ---LLPDHCFPVRVII-----PGCIGGRM-----VKWLKRIVTPAESDNYHFKDN

49995_2 410 ---P-HGD-----DEQLVQGEGERVPSAQGRDRAGVRAPDAA
49995_1 410 ---P-HGD-----DEQLVQGEGERVPSAQGRDRAGVRAPDAA
49995_4 -----
49995_8 360 RVLPSHVDAELANAEAWYKPEYIINELNINSVTTPGHDEILPINSITTQRGYTMKGYA
49995_3 1 -----MRVERWL-----GEPYDFSRCFVV
49995_7 1 -----MRVERWL-----GEPYDFSRCFVV
49995_5 360 RVLPSHVDAELANAEAWYKPEYIINELNINSVTTPGHDEILPINSITTQRGYTMKGYA
49995_6 360 RVLPSHVDAELANAEAWYKPEYIINELNINSVTTPGHDEILPINSITTQRGYTMKGYA

49995_2 444 RQPARRVDGAAEAPGDGGGRAGPQAH-----
49995_1 444 RQPARRVDGAAEAPGDGGGRAGPQAH-----
49995_4 -----
49995_8 420 YSGKKNRHASITFPTFD---LVSS-----
49995_3 20 HVGGGKKVTRVETLDGGETWLVCHLDHEKPNKYGKYWCFSVEVEVLDLLGAKEIA
49995_7 20 HVGGGKKVTRVETLDGGETWLVCHLDHEKPNKYGKYWCFSVEVEVLDLLGAKEIA
49995_5 420 YSGGGKKVTRVETLDGGETWLVCHLDHEKPNKYGKYWCFSVEVEVLDLLGAKEIA
49995_6 420 YSGGGKKVTRVETLDGGETWLVCHLDHEKPNKYGKYWCFSVEVEVLDLLGAKEIA

49995_2 472 VHAVHEH---HRRRQAVHHVRGAQA---RVAGVGVDRHHPGAARAAGGPHGDAPRVRQPD
49995_1 472 VHAVHEH---HRRRQAVHHVRGAQA---RVAGVGVDRARPRLRL-HQVQGEPGRRR---
49995_4 -----
49995_8 -----
49995_3 80 VRAWDQSLNTQEKLIWNLMGMMNNCWFKVKVNVCRPHKGEIGLVFEHPTQPG---NQP
49995_7 80 VRAWDQSLNTQEKLIWNLMGMMNNCWFKVKVNVCRPHKGEIGLVFEHPTQPG---NQP
49995_5 480 VRAWDQSLNTQEKLIWNLMGMMNNCWFKVKVNVCRPHKGEIGLVFEHPTQPG---NQP
49995_6 480 VRAWDQSLNTQEKLIWNLMGMMNNCWFKVKVNVCRPHKGEIGLVFEHPTQPG---NQP

49995_2 527 GGRHPPPRRAR---PVGS---RV---PGA---QG-----VVRHRPGK-----APGGG
49995_1 523 ---QHPHQRRYR---LHRGVRRHP---LRQG---QG-----APRHLPHR---RAHHHGHR
49995_4 -----
49995_8 -----
49995_3 136 GGWMARQKHLETAEAAAPGLKRSTSTPFMNTTDVGKQFTMSEVRKHASQESAWIVVHGHV
49995_7 136 GGWMARQKHLETAEAAAPGLKRSTSTPFMNTTDVGKQFTMSEVRKHASQESAWIVVHGHV
49995_5 536 GGWMARQKHLETAEAAAPGLKRSTSTPFMNTTDVGKQFTMSEVRKHASQESAWIVVHGHV
49995_6 536 GGWMARQKHLETAEAAAPGLKRSTSTPFMNTTDVGKQFTMSEVRKHASQESAWIVI-----

49995_2 563 VEVQRWVRHGGRPAGARSGRWGRHAGPLRT---TADD---PVRHLAQLG-----
49995_1 563 LQLRQLRPRRLRPVAPRAHPRGRQCSRALQP---ARKD---PLPPRRQEG-----
49995_4 -----
49995_8 -----
49995_3 196 YDCTKFLKD---HPGGADSI-LINACTDCTEEFDAIHSDKAKALLDTYRIGELITTGTGYS
49995_7 196 YDCTKFLKD---HPGGADSI-LINACTDCTEEFDAIHSDKAKALLDTYRIGELITTGTGYS
49995_5 596 YDCTKFLKD---HPGGADSI-LINACTDCTEEFDAIHSDKAKALLDTYRIGELITTGTGYS
49995_6 592 -----

```

Figure 6. Alignment of proteins encoded by Zm00001d049995 – Part 2. Amino acid sequence of gene Zm00001d049995 showing eight different versions of the protein that could be produced by the DNA sequence based on the revised model. The absence sequence is denoted by dashes, meaning the next amino acid in the chain would be the next letter after the dashes. Identities between different versions of protein are denoted by the shaded boxes. Note also that the boxes may be shaded even if the sequence is “skipped” in the comparison protein.

```

49995_2 607 -----ED-----EVR-----HGQFFRRVL-----
49995_1 607 -----AVPRRPPLPLAAVAREPGARPPHRQAHLR-----LRQY-----
49995_4 -----
49995_8 -----
49995_3 253 SDNSVHGGSVLSHLAPIREAVRAPALSNPREKIHCR LVGKKELSRDVRLFRFSLSPDQV
49995_7 253 SDNSVHGGSVLSHLAPIREAVRAPALSNPREKIHCR LVGKKELSRDVRLFRFSLSPDQV
49995_5 653 SDNSVHGGSVLSHLAPIREAVRAPALSNPREKIHCR LVGKKELSRDVRLFRFSLSPDQV
49995_6 592 -----

49995_2 -----
49995_1 -----
49995_4 -----
49995_8 -----
49995_3 313 LGLPIGKHIFVCASIEGKLCMRAYTPTSMVDEIGHFDLLVKVYFKNEHPKFPNGGLMTQY
49995_7 313 LGLPIGKHIFVCASIEGKLCMRAYTPTSMVDEIGHFDLLVKVYFKNEHPKFPNGGLMTQY
49995_5 713 LGLPIGKHIFVCASIEGKLCMRAYTPTSMVDEIGHFDLLVKVYFKNEHPKFPNGGLMTQY
49995_6 592 -----

49995_2 -----
49995_1 -----
49995_4 -----
49995_8 -----
49995_3 373 LDSLPVGSYIDVGKPLGHVEYTRGGSFVINGKQRHASRLAMICGSGITPMYQIIQAVLR
49995_7 373 LDSLPVGSYIDVGKPLGHVEYTRGGSFVINGKQRHASRLAMICGSGITPMYQIIQAVLR
49995_5 773 LDSLPVGSYIDVGKPLGHVEYTRGGSFVINGKQRHASRLAMICGSGITPMYQIIQAVLR
49995_6 592 -----IQAVLR

49995_2 -----
49995_1 -----
49995_4 -----
49995_8 -----
49995_3 433 EQPEDHTEMHLVYANRTEDDILLRDELDRWAAEYPDR LKVWYVIDQVKRPEEGWKYSVGF
49995_7 433 EQPEDHTEMHLVYANRTEDDILLRDELDRWAAEYPDR LKVWYVIDQVKRPEEGWKYSVGF
49995_5 833 EQPEDHTEMHLVYANRTEDDILLRDELDRWAAEYPDR LKVWYVIDQVKRPEEGWKYSVGF
49995_6 598 EQPEDHTEMHLVYANRTEDDILLRDELDRWAAEYPDR LKVWYVIDQVKRPEEGWKYSVGF

49995_2 -----
49995_1 -----
49995_4 -----
49995_8 -----
49995_3 493 VTEAVLREHVPEGGDDTLALACGPPPMIQFAISP NLEKMKYDMANSFVVF
49995_7 493 VTEAVLREHVPEGGDDTLALACGPPPMIQFAISP NLEKMKYDMANSFVVF
49995_5 893 VTEAVLREHVPEGGDDTLALACGPPPMIQFAISP NLEKMKYDMANSFVVF
49995_6 658 VTEAVLREHVPEGGDDTLALACGPPPMIQFAISP NLEKMKYDMANSFVVF

```

Figure 6. Alignment of proteins encoded by Zm00001d049995 – Part 3. Amino acid sequence of gene Zm00001d049995 showing eight different versions of the protein that could be produced by the DNA sequence based on the revised model. The absence sequence is denoted by dashes, meaning the next amino acid in the chain would be the next letter after the dashes. Identities between different versions of protein are denoted by the shaded boxes. Note also that the boxes may be shaded even if the sequence is “skipped” in the comparison protein.

The next gene that was evaluated was Zm00001d018206, which encodes nitrate reductase [NADH] 2 (Tello-Ruiz *et. al.* 2017). The original gene structure was compared with the data in the Apollo annotation platform and a model was produced as noted in the methods (Figure 7). Comparison of the v5 and Mikado RNA models showed the shortening of the gene with addition of an intron, and alternate 5' and 3' splice sites (Figure 7A). The v5 model supported evidence of an intron while Mikado models supported the absence of an intron at that same location (***, Figure 7A). RNA sequencing evidence also supported the absence of an intron at that location (***, Figure 7B), so it was labeled as a retained intron in the revised model (***, Figure 7C, Table 2). Additionally, a new 5' splice site was identified in the Mikado RNA model (*, Figure 7A), supported in the RNA sequencing evidence, (*, Figure 7B) and incorporated into the SuperTranscript (*, Figure 7C; Table 2). Finally, a new 3' splice site was identified in the Mikado RNA model (**, Figure 7A) and supported in the RNA sequencing evidence, (**, Figure 7B) and incorporated into the SuperTranscript (**, Figure 7C, Table 2).

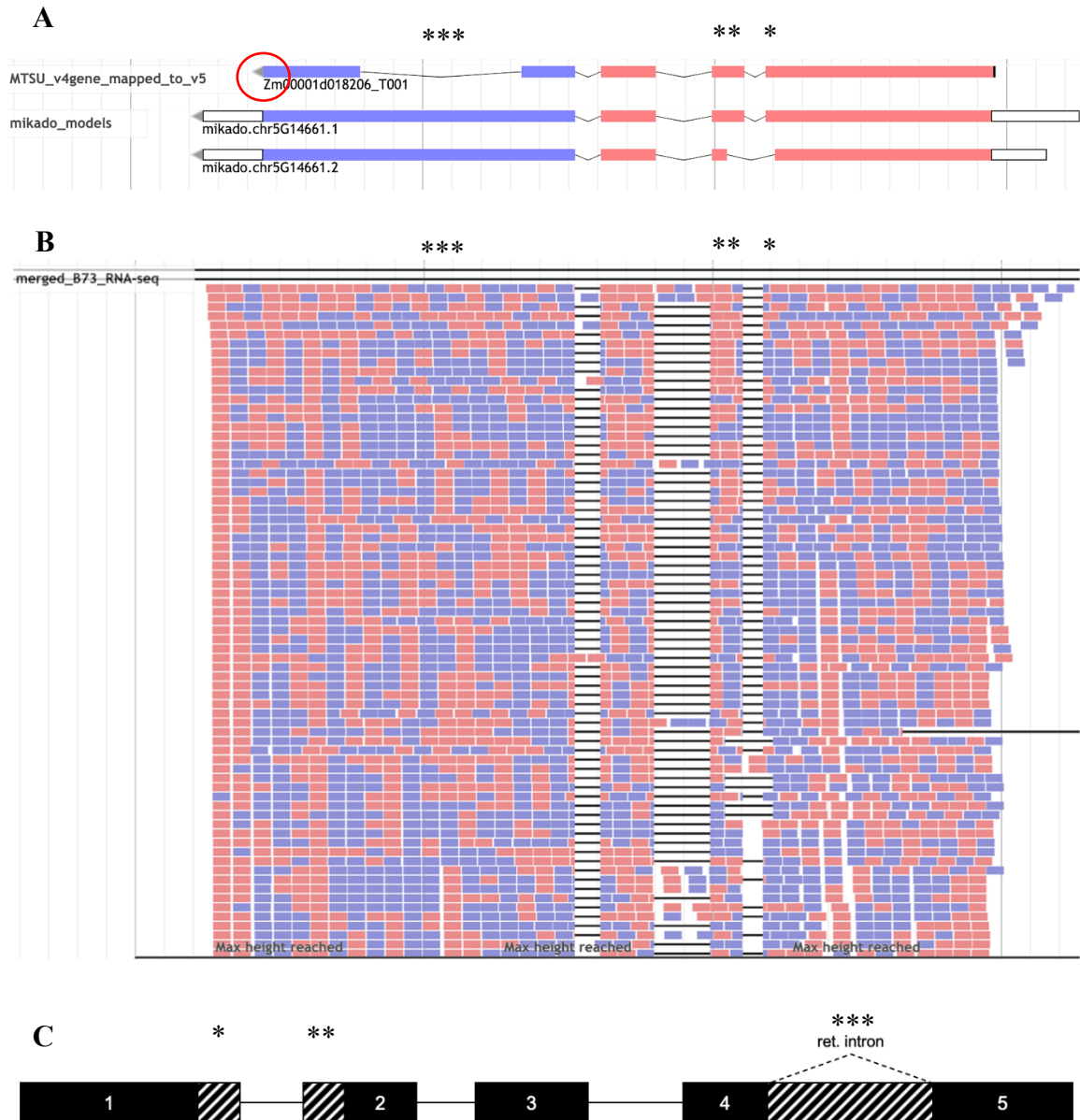


Figure 7. Evidence for reannotation for gene Zm00001d018206. Panel A shows the original structure of gene Zm00001d018206. Notice that the arrow on the original structure in panel A indicates the 3' end is on the left side of the gene (red circle) so the direction of the gene will be the reversed in the final model. The purple and red boxes show differentiation between the gene's reading frames. Panel B shows the final revised model. Note that the final model shows the RNA strand 5' to 3', which is the reverse of how the original model is displayed. The numbered black boxes indicate exons and their sequence from left to right. The lines between exons indicate introns.

Using the SuperTranscript as a guide, seven RNAs were constructed in the Apollo space. The encoded amino acid sequences for each transcript were then downloaded from the Apollo workspace and aligned using Clustal Omega (Sievers *et. al.* 2011) and shaded using BOXSHADE to allow visualization of similarities and differences between the proteins encoded by the different transcripts. The seven transcripts produce proteins that differ in various regions (Figure 8). Transcript 1 encodes a protein missing 44 amino acids compared to transcript 4. Transcript 7 encodes a protein missing 274 amino acids compared to transcript 4. Transcript 3 encodes a protein missing 333 amino acids compared to transcript 4. Transcript 5 encodes a protein missing 563 amino acids compared to transcript 4. Transcript 6 encodes a protein missing 255 amino acids compared to transcript 4. Transcript 2 encodes a protein missing 277 amino acids compared to transcript 4. Lastly, the sequence of amino acids encoded by each transcript was analyzed for the presence of known protein domains using the SMART domain sequence analysis program (Letunic and Bork 2017). The cytochrome b5-like heme/steroid binding domain was found present in the gene.

```

18206_5 1 -----
18206_7 1 -----MAASVERHLAPHWPANAPPKS FDMFRSGGGGKRRTPGDSDESEDSIPPDWRSI
18206_1 1 -----MAASVERHLAPHWPANAPPKS FDMFRSGGGGKRRTPGDSDESEDSIPPDWRSI
18206_3 1 -----
18206_4 1 LVPTMAASVERHLAPHWPANAPPKS FDMFRSGGGGKRRTPGDSDESEDSIPPDWRSI
18206_2 1 LVPTMAASVERHLAPHWPANAPPKS FDMFRSGGGGKRRTPGDSDESEDSIPPDWRSI
18206_6 1 LVPTMAASVERHLAPHWPANAPPKS FDMFRSGGGGKRRTPGDSDESEDSIPPDWRSI

18206_5 1 -----
18206_7 57 YSPRLVEPPAHDPRDEATSDAWVRRHPALVRLTGKHPFNSEPPVPRMAHGFIITPAPLH
18206_1 57 YSPRLVEPPAHDPRDEATSDAWVRRHPALVRLTGKHPFNSEPPVPRMAHGFIITPAPLH
18206_3 1 -----
18206_4 61 YSPRLVEPPAHDPRDEATSDAWVRRHPALVRLTGKHPFNSEPPVPRMAHGFIITPAPLH
18206_2 61 YSPRLVEPPAHDPRDEATSDAWVRRHPALVRLTGKHPFNSEPPVPRMAHGFIITPAPLH
18206_6 61 YSPRLVEPPAHDPRDEATSDAWVRRHPALVRLTGKHPFNSEPPVPRMAHGFIITPAPLH

18206_5 1 -----
18206_7 117 YVRNHGAVPRADWSTWTVEVAGLVRRPARLTMEQLVTEFEAVELPVTLCAGNRRKEQNM
18206_1 117 YVRNHGAVPRADWSTWTVEVAGLVRRPARLTMEQLVTEFEAVELPVTLCAGNRRKEQNM
18206_3 1 -----
18206_4 121 YVRNHGAVPRADWSTWTVEVAGLVRRPARLTMEQLVTEFEAVELPVTLCAGNRRKEQNM
18206_2 121 YVRNHGAVPRADWSTWTVEVAGLVRRPARLTMEQLVTEFEAVELPVTLCAGNRRKEQNM
18206_6 121 YVRNHGAVPRADWSTWTVEVAGLVRRPARLTMEQLVTEFEAVELPVTLCAGNRRKEQNM

18206_5 1 -----
18206_7 177 VRQTVGFNWGPAGISTSVWRGARLRDVLRRCGVMGAADGAANVCFEAGEDLPGGGGGGKY
18206_1 177 VRQTVGFNWGPAGISTSVWRGARLRDVLRRCGVMGAADGAANVCFEAGEDLPGGGGGGKY
18206_3 1 -----
18206_4 181 VRQTVGFNWGPAGISTSVWRGARLRDVLRRCGVMGAADGAANVCFEAGEDLPGGGGGGKY
18206_2 181 VRQTVGFNWGPAGISTSVWRGARLRDVLRRCGVMGAADGAANVCFEAGEDLPGGGGGGKY
18206_6 181 VRQTVGFNWGPAGISTSVWRGARLRDVLRRCGVMGAADGAANVCFEAGEDLPGGGGGGKY

18206_5 1 -----
18206_7 237 GTSLLRGGVAMPDARDVILAYMONGEPLAPDHGFFVPRVIVPGFIGGRMVKWLKRIIVASSE
18206_1 237 GTSLLRGGVAMPDARDVILAYMONGEPLAPDHGFFVPRVIVPGFIGGRMVKWLKRIIVASSE
18206_3 1 -----
18206_4 241 GTSLLRGGVAMPDARDVILAYMONGEPLAPDHGFFVPRVIVPGFIGGRMVKWLKRIIVASSE
18206_2 241 GTSLLRGGVAMPDARDVILAYMONGEPLAPDHGFFVPRVIVPGFIGGRMVKWLKRIIVASSE
18206_6 241 GTSLLRGGVAMPDARDVILAYMONGEPLAPDHGFFVPRVIVPGFIGGRMVKWLKRIIVASSE

18206_5 1 -----MINEININSVITTPGHDEVLPINALTT
18206_7 297 SESYYHYRDNRLV-----PINALTT
18206_1 297 SESYYHYRDNRLV-----PINALTT
18206_3 1 -----MINEININSVITTPGHDEVLPINALTT
18206_4 301 SESYYHYRDNRLVPSHVDADLANAEAWYKPECMINEININSVITTPGHDEVLPINALTT
18206_2 301 SESYYHYRDNRLVPSHVDADLANAEAAHQRPDAAAVY-----DQRIRI---LR
18206_6 301 SESYYHYRDNRLVPSHVDADLANAEAAHQRPDAAAVY-----DQRIRI---LR

18206_5 28 QRPYTIKGAYSGGGRKVTRVEVTLDGGETWHVCSLDHPERPTKYGYWCWCFWSVDVEV
18206_7 317 QRPYTIKGAYSGGGRKVTRVEVTLDGGETWHVCSLDHPERPTKYGYWCWCFWSVDVEV
18206_1 317 QRPYTIKGAYSGGGRKVTRVEVTLDGGETWHVCSLDHPERPTKYGYWCWCFWSVDVEV
18206_3 28 QRPYTIKGAYSGGGRKVTRVEVTLDGGETWHVCSLDHPERPTKYGYWCWCFWSVDVEV
18206_4 361 QRPYTIKGAYSGGGRKVTRVEVTLDGGETWHVCSLDHPERPTKYGYWCWCFWSVDVEV
18206_2 348 WRPE-----SNPGGGDAGRRL-----DVACVILARPPGASNQVR-----QVLV
18206_6 348 WRPE-----SNPGGGDAGRRL-----DVACVILARPPGASNQVR-----QVLV

18206_5 88 LDVLGAKEIAVRAWDEAMNTQPEKLVWNLGMNNNCWFRVKINACRPHKG-EIGMVF---
18206_7 377 LDVLGAKEIAVRAWDEAMNTQPEKLVWNLGMNNNCWFRVKINACRPHKG-EIGMVF---
18206_1 377 LDVLGAKEIAVRAWDEAMNTQPEKLVWNLGMNNNCWFRVKINACRPHKG-EIGMVF---
18206_3 88 LDVLGAKEIAVRAWDEAMNTQPEKLVWNLGMNNNCWFRVKINACRPHKG-EIGMVF---
18206_4 421 LDVLGAKEIAVRAWDEAMNTQPEKLVWNLGMNNNCWFRVKINACRPHKG-EIGMVF---
18206_2 385 LVLLVRRRGARRARGQGNRRPR-----LGR---GHEHPAGEACLEPHCHDEQLIVPGE
18206_6 385 LVLLVRRRGARRARGQGNRRPR-----LGR---GHEHPAGEACLEPHCHDEQLIVPGE

```

Figure 8. Alignment of proteins encoded by Zm00001d018206 – Part 1. Amino acid sequence of gene Zm00001d018206 showing seven different versions of the protein that could be produced by the DNA sequence based on the revised model. The absence sequence is denoted by dashes, meaning the next amino acid in the chain would be the next letter after the dashes. Identities between different versions of protein are denoted by the shaded boxes. Note also that the boxes may be shaded even if the sequence is “skipped” in the comparison protein.

```

018206_5 144 --EHPAQPG-----NPPGGWMARQKHLETSESAQSTLKKSTSTPFMNTATAQY
018206_7 433 --EHPAQPG-----NPPGGWMARQKHLETSESAQSTLKKSTSTPFMNTATAQY
018206_1 433 --EHPAQPG-----NPPGGWMARQKHLETSESAQSTLKKSTSTPFMNTATAQY
018206_3 144 --EHPAQPG-----NPPGGWMARQKHLETSESAQSTLKKSTSTPFMNTATAQY
018206_4 477 --EHPAQPG-----NPPGGWMARQKHLETSESAQSTLKKSTSTPFMNTATAQY
018206_2 436 DQRVPAACGRDRHGVRAPGAAGQFACRUDGAA-----EAPR
018206_6 436 DQRVPAACGRDRHGVRAPGAAGQFACRUDGAA-----EAPR

018206_5 190 TMSE-----
018206_7 479 TMSE-----
018206_1 479 TMSEVRRHTSPDSAW-IIVHGH-----IYDCTGFLKDHPGASII-I---NAGTD
018206_3 190 TMSEVRRHTSPDSAW-IIVHGH-----IYDCTGFLKDHPGASII-I---NAGTD
018206_4 523 TMSEVRRHTSPDSAW-IIVHGH-----IYDCTGFLKDHPGASII-I---NAGTD
018206_2 472 D I G R A E H A E E S H V H A I H E H G H R A V H H V R G A P P H V P G L R L D H R R P H L R I H G L P Q G P P G R
018206_6 472 D I G R A E H A E E S H V H A I H E H G H R A V H H V R A P A R A Q A --R H I R R R --I G H H A G V P -G D P G R

018206_5 194 -----
018206_7 483 -----
018206_1 525 C T E E F D A I H S D K A R G L L E M Y R V G E L V V T G S D Y S P N S H A D L K A I V E A P A A A A P L S V T S T V
018206_3 236 C T E E F D A I H S D K A R G L L E M Y R V G E L V V T G S D Y S P N S H A D L K A I V E A P A A A A P L S V T S T V
018206_4 569 C T E E F D A I H S D K A R G L L E M Y R V G E L V V T G S D Y S P N S H A D L K A I V E A P A A A A P L S V T S T V
018206_2 532 C R Q H -----P H C R R R L R L R G V R R H -----P L R Q G P R P
018206_6 527 A E G -----P A R R R H G D A P R V -----R E P N G

018206_5 194 -----
018206_7 483 -----
018206_1 585 A L S N P R E K V R C R L V D K K S L S Y N V R L F R F A L P S P D C -----K L G L P V G R H V Y V C A S I G K L
018206_3 296 A L S N P R E K V R C R L V D K K S L S Y N V R L F R F A L P S P D C -----K L G L P V G R H V Y V C A S I G K L
018206_4 629 A L S N P R E K V R C R L V D K K S L S Y N V R L F R F A L P S P D C -----K L G L P V G R H V Y V C A S I G K L
018206_2 559 P R D V P R C F A R C I R Q R -----I L P A E Q P R R P C G H R R G P
018206_6 547 C R H A F A G G -----D R -----P L G C R A P G A P G V V R G Q Q G A T C G S V G V R R G E S C R A

018206_5 194 -----
018206_7 483 -----
018206_1 640 C M R Y T P T S P V D E V G H V D L L I K Y F K D E D P K Y P N G G L M S Q Y L D S L P L G A T I D I K G P I G H I
018206_3 351 C M R Y T P T S P V D E V G H V D L L I K Y F K D E D P K Y P N G G L M S Q Y L D S L P L G A T I D I K G P I G H I
018206_4 684 C M R Y T P T S P V D E V G H V D L L I K Y F K D E D P K Y P N G G L M S Q Y L D S L P L G A T I D I K G P I G H I
018206_2 591 -----R C S R A V I G
018206_6 593 C H E T A P A S G R Q R -----D H C A R V R A A

018206_5 194 -----
018206_7 483 -----
018206_1 700 E Y A G R G G F V V N G E R R L A R R L A M I A G G T G I T P V Y Q V I Q A V L R D Q P D D D T E M H L V Y A N R T E D
018206_3 411 E Y A G R G G F V V N G E R R L A R R L A M I A G G T G I T P V Y Q V I Q A V L R D Q P D D D T E M H L V Y A N R T E D
018206_4 744 E Y A G R G G F V V N G E R R L A R R L A M I A G G T G I T P V Y Q V I Q A V L R D Q P D D D T E M H L V Y A N R T E D
018206_2 599 -----D V D R R A -----L Q -S A R E G Q V P A R R
018206_6 614 -----G D D R V H S --A P G P G E D G V R P R Q G S R -V L

018206_5 241 D M L L R E E I D R L A A A H P A R L K V W Y V V S K V A R P E D G W E Y G V G R V D E H V M R E H L P L G D S E T I A
018206_7 530 D M L L R E E I D R L A A A H P A R L K V W Y V V S K V A R P E D G W E Y G V G R V D E H V M R E H L P L G D S E T I A
018206_1 760 D M L L R E E I D R L A A A H P A R L K V W Y V V S K V A R P E D G W E Y G V G R V D E H V M R E H L P L G D S E T I A
018206_3 471 D M L L R E E I D R L A A A H P A R L K V W Y V V S K V A R P E D G W E Y G V G R V D E H V M R E H L P L G D S E T I A
018206_4 804 D M L L R E E I D R L A A A H P A R L K V W Y V V S K V A R P E D G W E Y G V G R V D E H V M R E H L P L G D S E T I A
018206_2 -----
018206_6 -----

018206_5 301 L V C G P P A M I E C T V R P G L E K M G Y D L D K A C L V F
018206_7 590 L V C G P P A M I E C T V R P G L E K M G Y D L D K A C L V F
018206_1 820 L V C G P P A M I E C T V R P G L E K M G Y D L D K A C L V F
018206_3 531 L V C G P P A M I E C T V R P G L E K M G Y D L D K A C L V F
018206_4 864 L V C G P P A M I E C T V R P G L E K M G Y D L D K A C L V F
018206_2 -----
018206_6 -----

```

Figure 8. Alignment of proteins encoded by Zm00001d018206 – Part 2. Amino acid sequence of gene Zm00001d018206 showing seven different versions of the protein that could be produced by the DNA sequence based on the revised model. The absence sequence is denoted by dashes, meaning the next amino acid in the chain would be the next letter after the dashes. Identities between different versions of protein are denoted by the shaded boxes. Note also that the boxes may be shaded even if the sequence is “skipped” in the comparison protein.

The next gene that was evaluated was Zm00001d017958, which encodes glutamine synthetase root isozyme 3 (Tello-Ruiz *et. al.* 2017). The original gene structure was compared with the data in the Apollo annotation platform and a model was produced as noted in the methods (Figure 9). Comparison of the v5, Mikado, IsoSeq, and full-length cDNA (flc) RNA models showed alternative 5' splice sites and alternative 3' splice sites (Figure 9A). Data from the IsoSeq model supported the presence of an alternative 3' splice site (*, Figure 9A), and its presence was supported in the RNA sequencing evidence (*, Figure 9B) so it was incorporated into the SuperTranscript (*, Figure 9C, Table 2). Data from an evidence model supported the presence of an alternative 3' splice site (**, Figure 9A), and its presence was supported in the RNA sequencing evidence (**, Figure 9B) so it was incorporated into the SuperTranscript (**, Figure 9C, Table 2). Data from the Mikado, IsoSeq, flc, and evidence models supported the presence of an alternative 3' splice site (*, Figure 9A), and its presence was supported in the RNA sequencing evidence (*, Figure 9B), so it was incorporated into the SuperTranscript (*, Figure 9C, Table 2). Data from an RNAseq Mikado model supported the presence of an alternative 5' splice site (***, Figure 9A), and its presence was supported in the RNA sequencing evidence (***, Figure 9B) so it was incorporated into the SuperTranscript (***, Figure 9C, Table 2). Data from an RNAseq Mikado model supported the presence of an alternative 3' splice site (**, Figure 9A), and its presence was supported in the RNA sequencing evidence (**, Figure 9B) so it was incorporated into the SuperTranscript (**, Figure 9C, Table 2). Data from the IsoSeq model supported the presence of an alternative 5' splice site (*, Figure 9A), and its presence was supported in the RNA sequencing evidence (*, Figure 9B) so it was incorporated into the

SuperTranscript (*, Figure 9C, Table 2). Data from the RNA sequencing evidence supported the presence of alternative 3' splice sites (***, Figure 9B), so they were incorporated into the SuperTranscript (***, Figure 9C, Table 2). Data from RNA sequencing data supported the presence of an alternative 5' splice site (**, Figure 9B), so it was incorporated into the SuperTranscript (**, Figure 9C, Table 2). Finally, a new 3' splice site was identified in the RNA sequencing data (***, Figure 9B) and incorporated into the SuperTranscript (***, Figure 9C).

Using the SuperTranscript as a guide, four RNAs were constructed in the Apollo space. The encoded amino acid sequences for each transcript were then downloaded from the Apollo workspace and aligned using Clustal Omega (Sievers *et. al.* 2011) and shaded using BOXSHADE to allow visualization of similarities and differences between the proteins encoded by the different transcripts. The four transcripts produce proteins that differ in a couple of regions (Figure 10). Transcripts 2 and 4 encode the same protein. Transcripts 1 and 3 also encode the same protein which is missing 17 amino acids compared to transcripts 2 and 4. Lastly, the sequence of amino acids encoded by each transcript was analyzed for the presence of known protein domains using the SMART domain sequence analysis program (Letunic and Bork 2017). The glutamine synthetase, catalytic domain was found present in the gene.

```

17958_1 1 -----MACLTDLVNLNLSDNTEKIIAEYIWIGGSGM
17958_3 1 -----MACLTDLVNLNLSDNTEKIIAEYIWIGGSGM
17958_2 1 PIPSSSISIPHHHHLLRSPTPVAPQPPAMACLTDLVNLNLSDNTEKIIAEYIWIGGSGM
17958_4 1 PIPSSSISIPHHHHLLRSPTPVAPQPPAMACLTDLVNLNLSDNTEKIIAEYIWIGGSGM

17958_1 32 DLRSKARTLSGPVTDPSKLPKWNVDGSSTGQAPGEDSEQTWERSLSLGHVAPCSPOAIFK
17958_3 32 DLRSKARTLSGPVTDPSKLPKWNVDGSSTGQAPGEDSEQTWERSLSLGHVAPCSPOAIFK
17958_2 61 DLRSKARTLSGPVTDPSKLPKWNVDGSSTGQAPGEDSEVI-----LYPQAIKF
17958_4 61 DLRSKARTLSGPVTDPSKLPKWNVDGSSTGQAPGEDSEVI-----LYPQAIKF

17958_1 92 DPFRRGNNILVMDCYTPAGEPIPTNKRYNAAKIFSSPEVAAEEPWYGIEQEYTLQKDT
17958_3 92 DPFRRGNNILVMDCYTPAGEPIPTNKRYNAAKIFSSPEVAAEEPWYGIEQEYTLQKDT
17958_2 109 DPFRRGNNILVMDCYTPAGEPIPTNKRYNAAKIFSSPEVAAEEPWYGIEQEYTLQKDT
17958_4 109 DPFRRGNNILVMDCYTPAGEPIPTNKRYNAAKIFSSPEVAAEEPWYGIEQEYTLQKDT

17958_1 152 NWPLGWPIGGFPGPGPYCGIGA EKSFGRDIVDAHYKACLYAGINISGINGEVMPGQWE
17958_3 152 NWPLGWPIGGFPGPGPYCGIGA EKSFGRDIVDAHYKACLYAGINISGINGEVMPGQWE
17958_2 169 NWPLGWPIGGFPGPGPYCGIGA EKSFGRDIVDAHYKACLYAGINISGINGEVMPGQWE
17958_4 169 NWPLGWPIGGFPGPGPYCGIGA EKSFGRDIVDAHYKACLYAGINISGINGEVMPGQWE

17958_1 212 FQVGPSVGISSGDQVWVARYILERITEIAGVVVTFDPKPIPGDWNGAGAHTNYSTESMRK
17958_3 212 FQVGPSVGISSGDQVWVARYILERITEIAGVVVTFDPKPIPGDWNGAGAHTNYSTESMRK
17958_2 229 FQVGPSVGISSGDQVWVARYILERITEIAGVVVTFDPKPIPGDWNGAGAHTNYSTESMRK
17958_4 229 FQVGPSVGISSGDQVWVARYILERITEIAGVVVTFDPKPIPGDWNGAGAHTNYSTESMRK

17958_1 272 EGGYEVIKAAIEKLKL RHREHIAAYGEGNERRLTGRHETADINTFSWGVANRGASVRVGR
17958_3 272 EGGYEVIKAAIEKLKL RHREHIAAYGEGNERRLTGRHETADINTFSWGVANRGASVRVGR
17958_2 289 EGGYEVIKAAIEKLKL RHREHIAAYGEGNERRLTGRHETADINTFSWGVANRGASVRVGR
17958_4 289 EGGYEVIKAAIEKLKL RHREHIAAYGEGNERRLTGRHETADINTFSWGVANRGASVRVGR

17958_1 332 ETEQNGKGYFEDRRPASNMDPYVVTSMIAETTTIWKP
17958_3 332 ETEQNGKGYFEDRRPASNMDPYVVTSMIAETTTIWKP
17958_2 349 ETEQNGKGYFEDRRPASNMDPYVVTSMIAETTTIWKP
17958_4 349 ETEQNGKGYFEDRRPASNMDPYVVTSMIAETTTIWKP

```

Figure 10. Alignment of proteins encoded by Zm00001d017958. Amino acid sequence of gene Zm00001d017958 showing four different versions of the protein that could be produced by the DNA sequence based on the revised model. The absence is denoted by dashes, meaning the next amino acid in the chain would be the next letter after the dashes. Identities between different versions of protein are denoted by the shaded boxes. Note also that the boxes may be shaded even if the sequence is “skipped” in the comparison protein.

The next gene that was evaluated was Zm00001d022388, which encodes ferredoxin-dependent glutamate synthase 2C chloroplastic (Tello-Ruiz *et. al.* 2017). The original gene structure was compared with the data in the Apollo annotation platform and a model was produced as noted in the methods (Figure 11). Comparison of the v5 and Mikado RNA models showed the presence of cassette exons, as well as an alternative 5' splice site (Figure 11A). The v5 model and Mikado model supported the presence of an exon, but in different locations (* and *, Figure 11A), and was supported in RNA sequencing data (* and *, Figure 11B), so they were incorporated as cassette exons in the SuperTranscript (* and *, Figure 11C, Table 2), meaning they could be present in that location or not depending on what transcript is produced. Additionally, RNA sequencing evidence supported both the presence and absence of an intron (**, Figure 11B), so it was labeled as an alternative intron in the final model (**, Figure 11C, Table 2). Finally, a new 5' splice site was identified in the Mikado RNA model (**, Figure 11A) and supported in the RNA sequencing evidence, (**, Figure 11B) and incorporated into the SuperTranscript (**, Figure 11C, Table 2).

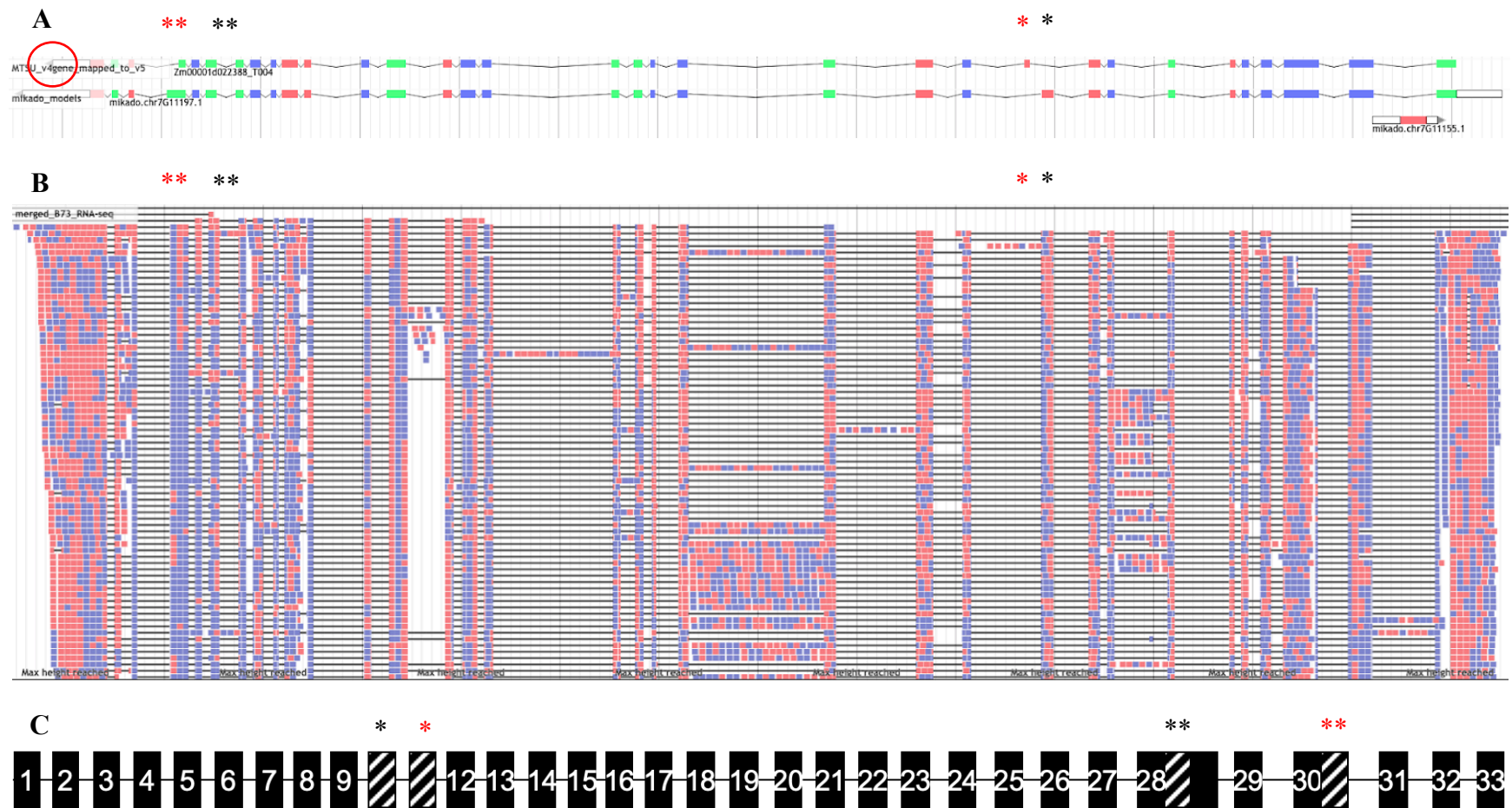


Figure 11. Evidence for reannotation for gene Zm00001d022388. Panel A shows the original structure of gene Zm00001d022388. Notice that the arrow on the original structure in panel A indicates the 3' end is on the left side of the gene (red circle) so the direction of the gene will be the reversed in the final model. The green and purple boxes show differentiation between the gene's reading frames. Panel B shows the final revised model. Note that the final model shows the RNA strand 5' to 3', which is the reverse of how the original model is displayed. The numbered black boxes indicate exons and their sequence from left to right. The lines between exons indicate introns.

Using the SuperTranscript as a guide, four RNAs were constructed in the Apollo space. The encoded amino acid sequences for each transcript were then downloaded from the Apollo workspace and aligned using Clustal Omega (Sievers *et. al.* 2011) and shaded using BOXSHADE to allow visualization of similarities and differences between the proteins encoded by the different transcripts. The four transcripts produce proteins that differ in a few regions (Figure 12). Transcript 4 encodes a protein missing 28 amino acids compared to transcript 2. Transcript 3 encodes a protein missing 47 amino acids compared to transcript 2. Transcript 1 encodes a protein missing 75 amino acids compared to transcript 2. Lastly, the sequence of amino acids encoded by each transcript was analyzed for the presence of known protein domains using the SMART domain sequence analysis program (Letunic and Bork 2017). No identifiable domains were found present in the transcripts.

```

22388_2 1 MATLPRAAPPTPAALLPLPRAAPPLLAGRAAAARRSRLRARGPSAAARRSWVVASAASS
22388_3 1 MATLPRAAPPTPAALLPLPRAAPPLLAGRAAAARRSRLRARGPSAAARRSWVVASAASS
22388_1 1 MATLPRAAPPTPAALLPLPRAAPPLLAGRAAAARRSRLRARGPSAAARRSWVVASAASS
22388_4 1 MATLPRAAPPTPAALLPLPRAAPPLLAGRAAAARRSRLRARGPSAAARRSWVVASAASS

22388_2 61 SSRVVGVARREAPPAPQKPTQQAADLNHILSERGACGVGFVANLKNMSSFDIVRDALM
22388_3 61 SSRVVGVARREAPPAPQKPTQQAADLNHILSERGACGVGFVANLKNMSSFDIVRDALM
22388_1 61 SSRVVGVARREAPPAPQKPTQQAADLNHILSERGACGVGFVANLKNMSSFDIVRDALM
22388_4 61 SSRVVGVARREAPPAPQKPTQQAADLNHILSERGACGVGFVANLKNMSSFDIVRDALM

22388_2 121 ALGCMHRGGCGADSDSGDGAGLMSAVPWLDFDDWASKQGLALFDRRNTGVGMVFLPQDE
22388_3 121 ALGCMHRGGCGADSDSGDGAGLMSAVPWLDFDDWASKQGLALFDRRNTGVGMVFLPQDE
22388_1 121 ALGCMHRGGCGADSDSGDGAGLMSAVPWLDFDDWASKQGLALFDRRNTGVGMVFLPQDE
22388_4 121 ALGCMHRGGCGADSDSGDGAGLMSAVPWLDFDDWASKQGLALFDRRNTGVGMVFLPQDE

22388_2 181 KSMEEAKAATEKVFVDEGLEVLGWRPVFPNVSVVGRNAKETMPNIQQIFVKVAKEDNADD
22388_3 181 KSMEEAKAATEKVFVDEGLEVLGWRPVFPNVSVVGRNAKETMPNIQQIFVKVAKEDNADD
22388_1 181 KSMEEAKAATEKVFVDEGLEVLGWRPVFPNVSVVGRNAKETMPNIQQIFVKVAKEDNADD
22388_4 181 KSMEEAKAATEKVFVDEGLEVLGWRPVFPNVSVVGRNAKETMPNIQQIFVKVAKEDNADD

22388_2 241 IERELYISRKLIERAAKSFWSADELYFCSLSSRTIVYKGLRSEVLGQFYLDLQNELYKS
22388_3 241 IERELYISRKLIERAAKSFWSADELYFCSLSSRTIVYKGLRSEVLGQFYLDLQNELYKS
22388_1 241 IERELYISRKLIERAAKSFWSADELYFCSLSSRTIVYKGLRSEVLGQFYLDLQNELYKS
22388_4 241 IERELYISRKLIERAAKSFWSADELYFCSLSSRTIVYKGLRSEVLGQFYLDLQNELYKS

22388_2 301 PFAIYHRRFSTNTSPRWPLAQPMRLLGHNGETINTIQGNLNMRSRETTLKS PVWRGREHE
22388_3 301 PFAIYHRRFSTNTSPRWPLAQPMRLLGHNGETINTIQGNLNMRSRETTLKS PVWRGREHE
22388_1 301 PFAIYHRRFSTNTSPRWPLAQPMRLLGHNGETINTIQGNLNMRSRETTLKS PVWRGREHE
22388_4 301 PFAIYHRRFSTNTSPRWPLAQPMRLLGHNGETINTIQGNLNMRSRETTLKS PVWRGREHE

22388_2 361 ICPFGDPKASDSANLDSTAELLRLSGRSPAEALMILVPEAYKNHPTLSIKYPEVTDFYDY
22388_3 361 ICPFGDPKASDSANLDSTAELLRLSGRSPAEALMILVPEAYKNHPTLSIKYPEVTDFYDY
22388_1 361 ICPFGDPKASDSANLDSTAELLRLSGRSPAEALMILVPEAYKNHPTLSIKYPEVTDFYDY
22388_4 361 ICPFGDPKASDSANLDSTAELLRLSGRSPAEALMILVPEAYKNHPTLSIKYPEVTDFYDY

22388_2 421 YKGQMEAWDGPALLLFSDGRTVGATLDRNGLRPARYWRTSDDFVYVASEVGVIPMDESKV
22388_3 421 YKGQMEAWDGPALLLFSDGRTVGATLDRNGLRPARYWRTSDDFVYVASEVGVIPMDESKV
22388_1 421 YKGQMEAWDGPALLLFSDGRTVGATLDRNGLRPARYWRTSDDFVYVASEVGVIPMDESKV
22388_4 421 YKGQMEAWDGPALLLFSDGRTVGATLDRNGLRPARYWRTSDDFVYVASEVGVIPMDESKV

22388_2 481 VMKGRLGPGMMITVDLQTGQVLENTEVKKTVASASPYGTWLQECTRLIKPVNLSSTIMD
22388_3 481 VMKGRLGPGMMITVDLQTGQVLENTEVKKTVASASPYGTWLQECTRLIKPVNLSSTIMD
22388_1 481 VMKGRLGPGMMITVDLQTGQVLENTEVKKTVASASPYGTWLQECTRLIKPVNLSSTIMD
22388_4 481 VMKGRLGPGMMITVDLQTGQVLENTEVKKTVASASPYGTWLQECTRLIKPVNLSSTIMD

22388_2 541 NETVLRHQAAGFYSSDQVVFVBSMASQGKEPTFCMGDDIPLAVLSORPHLIDYFKORF
22388_3 541 NETVLRHQAAGFYSSDQVVFVBSMASQGKEPTFCMGDDIPLAVLSORPHLIDYFKORF
22388_1 541 NETVLRHCHFVFLGKDSHYIRVG-----SR---NYIEF---S
22388_4 541 NETVLRHCHFVFLGKDSHYIRVG-----SR---NYIEF---S

22388_2 601 AOVTNPAIDPLREGLVMSLEVNIKGRGNILEVGPENADQVALSSPVLNEGELETLLNDSK
22388_3 601 AOVTNPAIDPLREGLVMSLEVNIKGRGNILEVGPENADQVALSSPVLNEGELETLLNDSK
22388_1 573 GVTNPAIDPLREGLVMSLEVNIKGRGNILEVGPENADQVALSSPVLNEGELETLLNDSK
22388_4 573 GVTNPAIDPLREGLVMSLEVNIKGRGNILEVGPENADQVALSSPVLNEGELETLLNDSK

22388_2 661 LKPKVLSTYFDIRKGLDGS LDKTIQALCEEADA AVRSQSLLVLSDRSEAPEPTRPAIPT
22388_3 661 LKPKVLSTYFDIRKGLDGS LDKTIQALCEEADA AVRSQSLLVLSDRSEAPEPTRPAIPT
22388_1 633 LKPKVLSTYFDIRKGLDGS LDKTIQALCEEADA AVRSQSLLVLSDRSEAPEPTRPAIPT
22388_4 633 LKPKVLSTYFDIRKGLDGS LDKTIQALCEEADA AVRSQSLLVLSDRSEAPEPTRPAIPT

```

Figure 12. Alignment of proteins encoded by Zm00001d022388 – Part 1. Amino acid sequence of gene Zm00001d022388 showing four different versions of the protein that could be produced by the DNA sequence based on the revised model. The absence sequence is denoted by dashes, meaning the next amino acid in the chain would be the next letter after the dashes. Identities between different versions of protein are denoted by the shaded boxes. Note also that the boxes may be shaded even if the sequence is “skipped” in the comparison protein.

```

22388_2 721 LLAVGAIHQHLIQNGLRMSASIVADTAQCFSTHHFACLIYGASAVCPYLAETCROWRI
22388_3 721 LLAVGAIHQHLIQNGLRMSASIVADTAQCFSTHHFACLIYGASAVCPYLAETCROWRI
22388_1 693 LLAVGAIHQHLIQNGLRMSASIVADTAQCFSTHHFACLIYGASAVCPYLAETCROWRI
22388_4 693 LLAVGAIHQHLIQNGLRMSASIVADTAQCFSTHHFACLIYGASAVCPYLAETCROWRI

22388_2 781 SNKTLNLMRNGKMPVTVTIEQAQRNFIKAVKSGLLKILSKMGISLLSSYCGAQIFEIYGLG
22388_3 781 SNKTLNLMRNGKMPVTVTIEQAQRNFIKAVKSGLLKILSKMGISLLSSYCGAQIFEIYGLG
22388_1 753 SNKTLNLMRNGKMPVTVTIEQAQRNFIKAVKSGLLKILSKMGISLLSSYCGAQIFEIYGLG
22388_4 753 SNKTLNLMRNGKMPVTVTIEQAQRNFIKAVKSGLLKILSKMGISLLSSYCGAQIFEIYGLG

22388_2 841 QEVVDLAFCGSVSKIGGLTLDDELGRETLSEFWVKAFSEDTAKRLENFGFIQSRPGGEYHAN
22388_3 841 QEVVDLAFCGSVSKIGGLTLDDELGRETLSEFWVKAFSEDTAKRLENFGFIQSRPGGEYHAN
22388_1 813 QEVVDLAFCGSVSKIGGLTLDDELGRETLSEFWVKAFSEDTAKRLENFGFIQSRPGGEYHAN
22388_4 813 QEVVDLAFCGSVSKIGGLTLDDELGRETLSEFWVKAFSEDTAKRLENFGFIQSRPGGEYHAN

22388_2 901 NPEMSKLLHKAIREKRDNAYTVYQQHLASRPVNVLRDLLELKS DRAPIPIGKVESATSIV
22388_3 901 NPEMSKLLHKAIREKRDNAYTVYQQHLASRPVNVLRDLLELKS DRAPIPIGKVESATSIV
22388_1 873 NPEMSKLLHKAIREKRDNAYTVYQQHLASRPVNVLRDLLELKS DRAPIPIGKVESATSIV
22388_4 873 NPEMSKLLHKAIREKRDNAYTVYQQHLASRPVNVLRDLLELKS DRAPIPIGKVESATSIV

22388_2 961 ERFCTGGMSLGAISRETHEAIAIAMNRIGGKSNSEGGGEDPIRWNP LTVVDGYSP TLPHE
22388_3 961 ERFCTGGMSLGAISRETHEAIAIAMNRIGGKSNSEGGGEDPIRWNP LTVVDGYSP TLPHE
22388_1 933 ERFCTGGMSLGAISRETHEAIAIAMNRIGGKSNSEGGGEDPIRWNP LTVVDGYSP TLPHE
22388_4 933 ERFCTGGMSLGAISRETHEAIAIAMNRIGGKSNSEGGGEDPIRWNP LTVVDGYSP TLPHE

22388_2 1021 LKGLQNGDTATSAIKQVASGRFGVTP TFLVNADQIEIKIAQGA KPGE GGLPGKKVSAYI
22388_3 1021 LKGLQNGDTATSAIKQVASGRFGVTP TFLVNADQIEIKIAQGA KPGE GGLPGKKVSAYI
22388_1 993 LKGLQNGDTATSAIKQVASGRFGVTP TFLVNADQIEIKIAQGA KPGE GGLPGKKVSAYI
22388_4 993 LKGLQNGDTATSAIKQVASGRFGVTP TFLVNADQIEIKIAQGA KPGE GGLPGKKVSAYI

22388_2 1081 ARLRNSKPGVPLISPPPHHDIYSIEDLAQLIYDLHQINPKAKVSVKLVSEAGIGTVASGV
22388_3 1081 ARLRNSKPGVPLISPPPHHDIYSIEDLAQLIYDLHQINPKAKVSVKLVSEAGIGTVASGV
22388_1 1053 ARLRNSKPGVPLISPPPHHDIYSIEDLAQLIYDLHQINPKAKVSVKLVSEAGIGTVASGV
22388_4 1053 ARLRNSKPGVPLISPPPHHDIYSIEDLAQLIYDLHQINPKAKVSVKLVSEAGIGTVASGV

22388_2 1141 SKANADI IQISGH DGGTGASPISSIKHAGGPWELGLTETNQ TLIQNGLRERVVLRVDGGF
22388_3 1141 SKANADI IQISGH DGGTGASPISSIKHAGGPWELGLTETNQ TLIQNGLRERVVLRVDGGF
22388_1 1113 SKANADI IQISGH DGGTGASPISSIKHAGGPWELGLTETNQ TLIQNGLRERVVLRVDGGF
22388_4 1113 SKANADI IQISGH DGGTGASPISSIKHAGGPWELGLTETNQ TLIQNGLRERVVLRVDGGF

22388_2 1201 RSGQDVLIAAAMGADEYFGFSVAMIATGCVMARICHTNNCPVGVASQREELRARFP GPVG
22388_3 1201 RSGQDVLIAAAMGADEYFGFSVAMIATGCVMARICHTNNCPVGVASQREELRARFP GPVG
22388_1 1173 RSGQDVLIAAAMGADEYFGFSVAMIATGCVMARICHTNNCPVGVASQREELRARFP GPVG
22388_4 1173 RSGQDVLIAAAMGADEYFGFSVAMIATGCVMARICHTNNCPVGVASQREELRARFP GPVG

22388_2 1261 DLVNYFLVAAEEVRAALAQLGYEKLDDIIGRTDLLKPKHISLVKTQHIDLGYLLSNAGLP
22388_3 1261 DLVNYFLVAAEEVRAALAQLGYEKLDDIIGRTDLLKPKHISLVKTQHIDLGYLLSNAGLP
22388_1 1233 DLVNYFLVAAEEVRAALAQLGYEKLDDIIGRTDLLKPKHISLVKTQHIDLGYLLSNAGLP
22388_4 1233 DLVNYFLVAAEEVRAALAQLGYEKLDDIIGRTDLLKPKHISLVKTQHIDLGYLLSNAGLP

22388_2 1321 EWSSSQIRSQDVHTNGPVLDETILADPEIADAIENEKEVSKAFQIYNVDRAVCGRVAGVI
22388_3 1321 EWSSSQIRSQDVHTNGPVLDETILADPEIADAIENEKEVSKAFQIYNVDRAVCGRVAGVI
22388_1 1293 EWSSSQIRSQDVHTNGPVLDETILADPEIADAIENEKEVSKAFQIYNVDRAVCGRVAGVI
22388_4 1293 EWSSSQIRSQDVHTNGPVLDETILADPEIADAIENEKEVSKAFQIYNVDRAVCGRVAGVI

22388_2 1381 AKKYGDTGFAGQLNITFNGSAGQSFSGCFLTPGMNIRLVGEANDYVVGK---GMAGGELVVV
22388_3 1381 AKKYGDTGFAGQLNITFNGSAGQSFSGCFLTPGMNIRLVGEANDYVVGKVLNGMAGGELVVV
22388_1 1353 AKKYGDTGFAGQLNITFNGSAGQSFSGCFLTPGMNIRLVGEANDYVVGKVLNGMAGGELVVV
22388_4 1353 AKKYGDTGFAGQLNITFNGSAGQSFSGCFLTPGMNIRLVGEANDYVVGK---GMAGGELVVV

```

Figure 12. Alignment of proteins encoded by Zm00001d022388 – Part 2. Amino acid sequence of gene Zm00001d022388 showing four different versions of the protein that could be produced by the DNA sequence based on the revised model. The absence sequence is denoted by dashes, meaning the next amino acid in the chain would be the next letter after the dashes. Identities between different versions of protein are denoted by the shaded boxes. Note also that the boxes may be shaded even if the sequence is “skipped” in an alternate RNA strand.

```

22388_2 1438 PVDKTFVVPEDATIVGNTCLY GATGGQVFVRGKAGERFAVRNSLCQAVVEGTGDHCCEYM
22388_3 1441 PVDKTFVVPEDATIVGNTCLA-----
22388_1 1413 PVDKTFVVPEDATIVGNTCLA-----
22388_4 1410 PVDKTFVVPEDATIVGNTCLY GATGGQVFVRGKAGERFAVRNSLCQAVVEGTGDHCCEYM

22388_2 1498 TGGCVVVLGKAGRNVAAAGMTGGLAYILDEDDTLVPKNKEIVKMQRVNAPAGQMQLKGLI
22388_3 1462 -----GRNVAAGMTGGLAYILDEDDTLVPKNKEIVKMQRVNAPAGQMQLKGLI
22388_1 1434 -----GRNVAAGMTGGLAYILDEDDTLVPKNKEIVKMQRVNAPAGQMQLKGLI
22388_4 1470 TGGCVVVLGKAGRNVAAAGMTGGLAYILDEDDTLVPKNKEIVKMQRVNAPAGQMQLKGLI

22388_2 1558 EAYVDKTGSEKGIAILREWEAYLPLFWQLVPPSEEDSPEACAEFERVLAKQATTQLSAK
22388_3 1511 EAYVDKTGSEKGIAILREWEAYLPLFWQLVPPSEEDSPEACAEFERVLAKQATTQLSAK
22388_1 1483 EAYVDKTGSEKGIAILREWEAYLPLFWQLVPPSEEDSPEACAEFERVLAKQATTQLSAK
22388_4 1530 EAYVDKTGSEKGIAILREWEAYLPLFWQLVPPSEEDSPEACAEFERVLAKQATTQLSAK

```

Figure 12. Alignment of proteins encoded by Zm00001d022388 – Part 3. Amino acid sequence of gene Zm00001d022388 showing four different versions of the protein that could be produced by the DNA sequence based on the revised model. The absence sequence is denoted by dashes, meaning the next amino acid in the chain would be the next letter after the dashes. Identities between different versions of protein are denoted by the shaded boxes. Note also that the boxes may be shaded even if the sequence is “skipped” in the comparison protein.

The next gene that was evaluated was Zm00001d052165, which encodes nitrate reductase 2 (Tello-Ruiz *et. al.* 2017). The original gene structure was compared with the data in the Apollo annotation platform and a model was produced as noted in the methods (Figure 13). Comparison of the v5, IsoSeq, full-length cDNA (flc), and Mikado RNA models showed the presence of retained introns, as well as alternative start and end of transcription sites (Figure 13A). The IsoSeq and flc models supported both the presence of an alternate start of transcription (*, Figure 13A), and was supported in RNA sequencing data (*, Figure 13B), so it was incorporated into the SuperTranscript (*, Figure 13C, Table 2). IsoSeq and flc models supported the presence and absence of an intron in one location (**, Figure 13A), and RNA sequencing data supported the absence of an exon in that location (**, Figure 13B), so it was labeled as retained intron in the final model (**, Figure 13C, Table 2). RNA sequencing data also supported the presence and absence of an intron (***, Figure 13B, so it was incorporated into the SuperTranscript (***, Figure 13C, Table 2). of the additional 3' terminal exon (Figure 13A, 13B), so this was labeled as an alternative exon in the revised model (**, Figure 13C; Table 2). Finally, a new end of transcription site was identified in the IsoSeq model (****, Figure 13A), and supported in the RNA sequencing evidence, (****, Figure 13B) and incorporated into the SuperTranscript (****, Figure 13C, Table 2).

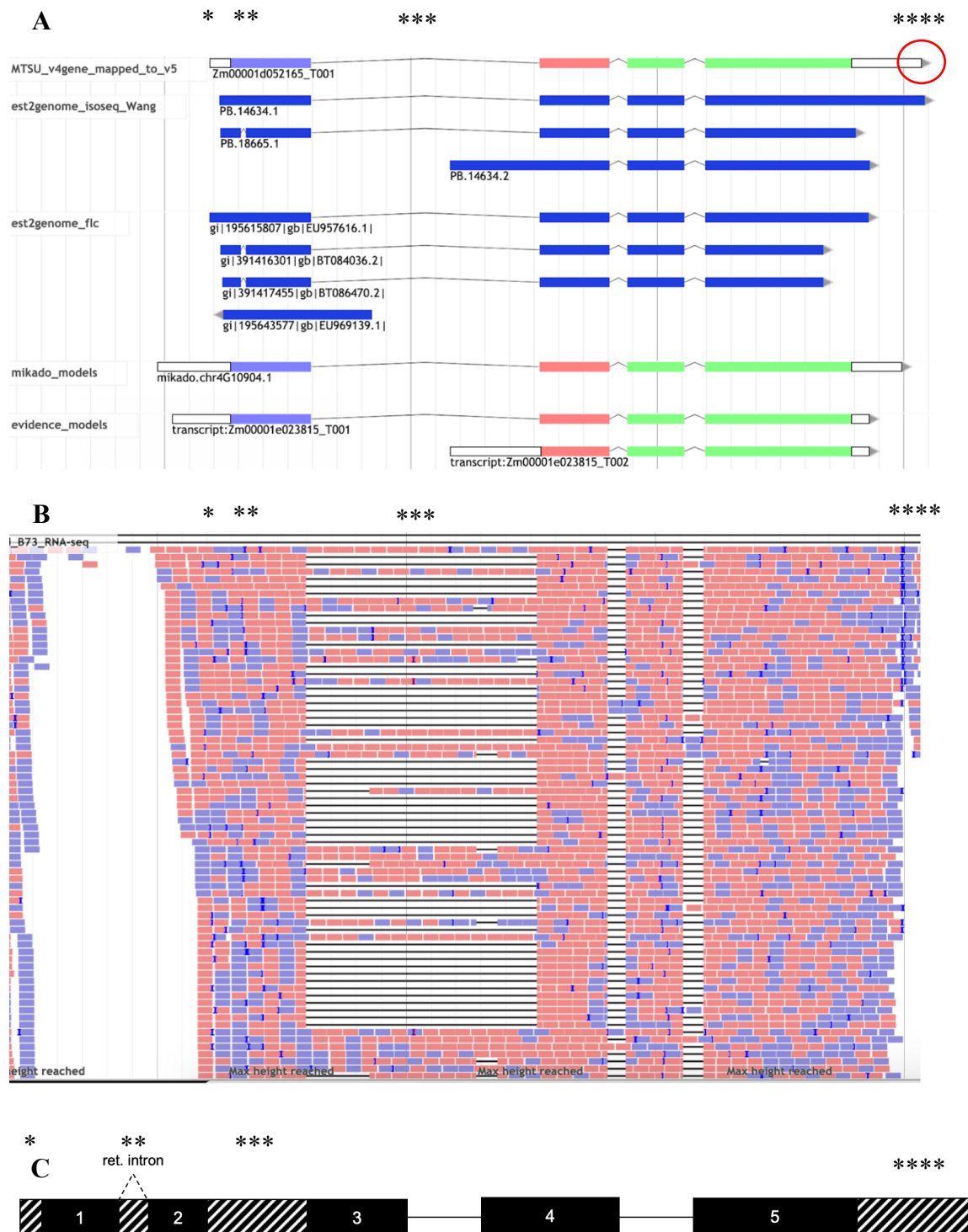


Figure 13. Evidence for reannotation for gene Zm00001d052165. Panel A shows the original structure of gene Zm00001d052165. Notice that the arrow on the original structure in panel A indicates the 3' end is on the left side of the gene (red circle) so the direction of the gene will be the reversed in the final model. The green and purple boxes show differentiation between the gene's reading frames. Panel B shows the final revised model. Note that the final model shows the RNA strand 5' to 3', which is the reverse of how the original model is displayed. The numbered black boxes indicate exons and their sequence from left to right. The lines between exons indicate introns.

Using the SuperTranscript as a guide, five RNAs were constructed in the Apollo space. The encoded amino acid sequences for each transcript were then downloaded from the Apollo workspace and aligned using Clustal Omega (Sievers *et. al.* 2011) and shaded using BOXSHADE to allow visualization of similarities and differences between the proteins encoded by the different transcripts. The five transcripts produce proteins that differ in various regions (Figure 14). Transcripts 1 and 3 encode the same protein. Transcript 5 encodes a protein missing 92 amino acids compared to transcripts 1 and 3. Transcript 4 encodes a protein missing 49 amino acids compared to transcripts 1 and 3. Transcript 2 encodes a protein missing 56 amino acids compared to transcripts 1 and 3. Lastly, the sequence of amino acids encoded by each transcript was analyzed for the presence of known protein domains using the SMART domain sequence analysis program (Letunic and Bork 2017). No identifiable domains were found present in the transcripts.

```

52165_1 1 -----MMRLKLPNGVTTSEQTRYLASVTE
52165_3 1 -----MMRLKLPNGVTTSEQTRYLASVTE
52165_5 1 -----
52165_2 1 SAPPPRPTIKQPHARHPWPPPPPPQHRPTATATANHGLLSVPAAVPP-----AIL
52165_4 1 SAPPPRPTIKQPHARHPWPPPPPPQHRPTATATANHGLLSVPAAVPP-----AIL

52165_1 25 AYCADGCADVTT-----RONWQIRGVTLDPVPAIID-----CLRAGVLTSTQSG
52165_3 25 AYCADGCADVTT-----RONWQIRGVTLDPVPAIID-----CLRAGVLTSTQSG
52165_5 1 MRADDDAVPF-E-----RCVQ-----DCGVPLGVVGVGEEKVGVGARVDDLVGVDA-
52165_2 52 AR-----VLFAAPHRARPRGRIRAAAGGGGGAGEDCAAGAEGRGACGRVLGPQGEVP-
52165_4 52 ARGGDGLAVLFAAPHRARPRGRIRAAAGGGGGAGEDCAAGAEGRGACGRVLGPQGEVP-

52165_1 69 MDNVRNPFVGNPLA-----CVDPHEIVDTRPYTNLLSSYVTNNSQGNPTI
52165_3 69 MDNVRNPFVGNPLA-----CVDPHEIVDTRPYTNLLSSYVTNNSQGNPTI
52165_5 47 --GER-----VADGVAVHVHAALQAGETDCAEAVQDGR-----DURE-----RHPADI
52165_2 104 --GGAEPNAGEGAGECAHGAVHGGCRHFGPCGPHGADRFRQAHCGR-----RRPPOV
52165_4 111 --GGAEPNAGEGAGECAHGAVHGGCRHFGPCGPHGADRFRQAHCGR-----RRPPOV

52165_1 113 -TNLRKWNVCVIGSHDLYEHPHINDLAYMPAVKDGEGFGNLLVGGFTSPKRWAEALPLD
52165_3 113 -TNLRKWNVCVIGSHDLYEHPHINDLAYMPAVKDGEGFGNLLVGGFTSPKRWAEALPLD
52165_5 88 -PVLPGGHVRAPVGA-----RIDDARQVPRILARRH-----AVG-QIQPHHE-----PPVN
52165_2 155 ARELPP--POAPVRAV-----HD--AAEA--AQRRD-----D-----
52165_4 162 ARELPP--POAPVRAV-----HD--AAEA--AQRRD-----D-----

52165_1 172 AWVAGDDVPVCKAILEYRDLGSRGNR-----OKTRMMWLI--DELCEVFRSEVE
52165_3 172 AWVAGDDVPVCKAILEYRDLGSRGNR-----OKTRMMWLI--DELCEVFRSEVE
52165_5 133 ---KTKMQIPEFAOKVNMSARKRSDCSVHS DASSKPAVSNDWQTLCLRENNCFVQSSAA
52165_2 181 ---ERADNVP--GERHRCVRRRRVRGR-DH-----PAE-----LATPRCDAGRPGHP
52165_4 188 ---ERADNVP--GERHRCVRRRRVRGR-DH-----PAE-----LATPRCDAGRPGHP

52165_1 222 ---KRMENGVLIERAA-PEDLVDKRERRDYLCVHFQ---KEGLSYVGLHVPVGRQLQAA
52165_3 222 ---KRMENGVLIERAA-PEDLVDKRERRDYLCVHFQ---KEGLSYVGLHVPVGRQLQAA
52165_5 190 -D---TARHSASRSEEQSRPRPNEWQ-----NARPHYELHASKLALLQSG--SCR---
52165_2 223 GRPPRRRPHQPAERHCQRAQPRRCPARRRRPPDRRHAEHLHQPSSLIRHQQL--PGEPHNH
52165_4 230 GRPPRRRPHQPAERHCQRAQPRRCPARRRRPPDRRHAEHLHQPSSLIRHQQL--PGEPHNH

52165_1 274 DMFEFA-RLADEYGHGETRLTVQNTVLPNVSNR-----LDALLAE-----PI
52165_3 274 DMFEFA-RLADEYGHGETRLTVQNTVLPNVSNR-----LDALLAE-----PI
52165_5 234 -----HLLVLAVEEAEPLEA--LVDVVLGELCGVDLLHGDGQVLDALHQRHGLLLOI
52165_2 282 QPVSLSLSLSHLTALSEPLTRCYQSDTLPLVITGR-----GNCTSASS
52165_4 289 QPVSLSLSLSHLTALSEPLTRCYQSDTLPLVITGR-----GNCTSASS

52165_1 317 LQEQRLSPRPSMLLRGLVACTGNQFCGQAI IETKARALQVAREVEKRVAVPRPVRMHWTC
52165_3 317 LQEQRLSPRPSMLLRGLVACTGNQFCGQAI IETKARALQVAREVEKRVAVPRPVRMHWTC
52165_5 287 HLLIRVCPRPVILLLEDVPFA-----RPLDRLQPLRR
52165_2 324 ARVCTCTSRITSTTSRTCRFS-----RTASSASTFWWAG
52165_4 331 ARVCTCTSRITSTTSRTCRFS-----RTASSASTFWWAG

52165_1 377 C-----PNSCGQVQVADIGFMGCLTSDSGKIVEAADIFVGGRVGSDSHLADVYRKSVP
52165_3 377 C-----PNSCGQVQVADIGFMGCLTSDSGKIVEAADIFVGGRVGSDSHLADVYRKSVP
52165_5 320 DLL-----PRRR--RR-----RHGDGRAGAPCAAQPGRRGRRRRVRVGG-----
52165_2 357 SSAPRGGPRRCRST-----PGSEGTTSPPCARPSSRRRTGTSAPGATER-----
52165_4 364 SSAPRGGPRRCRST-----PGSEGTTSPPCARPSSRRRTGTSAPGATER-----

52165_1 431 CKDLVPFVADLLVERFGAVPREREDEE
52165_3 431 CKDLVPFVADLLVERFGAVPREREDEE
52165_5 358 --QEEELQCR-----

52165_2 400 --RRA-----
52165_4 407 --RRA-----

```

Figure 14. Alignment of proteins encoded by Zm00001d052165. Amino acid sequence of gene Zm00001d052165 showing five different versions of the protein that could be produced by the DNA sequence based on the revised model. The absence sequence is denoted by dashes, meaning the next amino acid in the chain would be the next letter after the dashes. Identities between different versions of protein are denoted by the shaded boxes. Note also that the boxes may be shaded even if the sequence is “skipped” in the comparison protein.

The next gene that was evaluated was Zm00001d018161, which encodes ferredoxin-nitrite reductase chloroplastic (Tello-Ruiz *et. al.* 2017). The original gene structure was compared with the data in the Apollo annotation platform and a model was produced as noted in the methods (Figure 15). Comparison of the v5 and Mikado RNA models showed an alternative 3' splice site (*, Figure 15A), and RNA sequencing evidence supported presence of an alternative 3' splice site at that location (*, Figure 15B), so it was incorporated into the SuperTranscript (*, Figure 15C, Table 2).

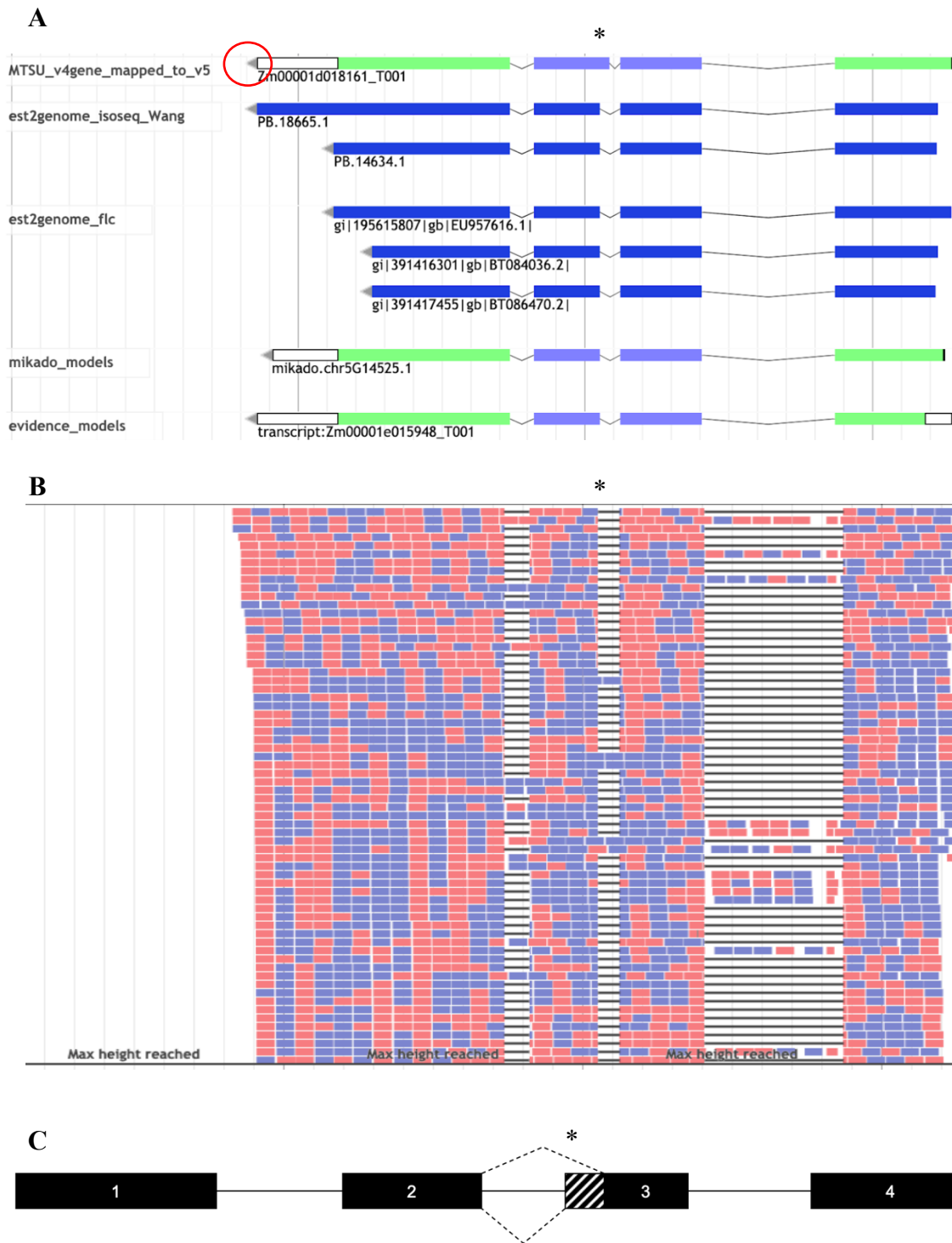


Figure 15. Evidence for reannotation for gene Zm00001d018161. Panel A shows the original structure of gene Zm00001d018161. Notice that the arrow on the original structure in panel A indicates the 3' end is on the left side of the gene (red circle) so the direction of the gene will be the reversed in the final model. The green and purple boxes show differentiation between the gene's reading frames. Panel B shows the final revised model. Note that the final model shows the RNA strand 5' to 3', which is the reverse of how the original model is displayed. The numbered black boxes indicate exons and their sequence from left to right. The lines between exons indicate introns.

Using the SuperTranscript as a guide, four RNAs were constructed in the Apollo space. The encoded amino acid sequences for each transcript were then downloaded from the Apollo workspace and aligned using Clustal Omega (Sievers *et. al.* 2011) and shaded using BOXSHADE to allow visualization of similarities and differences between the proteins encoded by the different transcripts. The four transcripts produce proteins that differ in the beginning and central regions (Figure 16). Transcript 2 encodes a protein missing 12 amino acids compared to transcript 1. Transcript 3 encodes a protein missing 13 amino acids compared to transcript 1. Transcript 4 encodes a protein missing 25 amino acids compared to transcript 1. Lastly, the sequence of amino acids encoded by each transcript was analyzed for the presence of known protein domains using the SMART domain sequence analysis program (Letunic and Bork 2017). No identifiable domains were found present in the transcripts.

```

18161_1 1 AAPLAPIINNRRNARHPWLPYHHHSTGTAAVPPPPPTPTMASTASLQRFPLASPHASSRRR
18161_2 1 -----ARHPWLPYHHHSTGTAAVPPPPPTPTMASTASLQRFPLASPHASSRRR
18161_3 1 AAPLAPIINNRRNARHPWLPYHHHSTGTAAVPPPPPTPTMASTASLQRFPLASPHASSRRR
18161_4 1 -----ARHPWLPYHHHSTGTAAVPPPPPTPTMASTASLQRFPLASPHASSRRR

18161_1 61 AGRARAASIPSSSPPATRDNEVPAERLEPRVEAREGGYWSLKERYRTGLNPHEKVKLEK
18161_2 49 AGRARAASIPSSSPPATRDNEVPAERLEPRVEAREGGYWSLKERYRTGLNPHEKVKLEK
18161_3 61 AGRARAASIPSSSPPATRDNEVPAERLEPRVEAREGGYWSLKERYRTGLNPHEKVKLEK
18161_4 49 AGRARAASIPSSSPPATRDNEVPAERLEPRVEAREGGYWSLKERYRTGLNPHEKVKLEK

18161_1 121 EPMALFMDGGVRDLAKIPMEVIDAAKLTKDDVDVRLKWLGLFHRRKHQYGRFMMRLKLPN
18161_2 109 EPMALFMDGGVRDLAKIPMEVIDAAKLTKDDVDVRLKWLGLFHRRKHQYGRFMMRLKLPN
18161_3 121 EPMALFMDGGVRDLAKIPMEVIDAAKLTKDDVDVRLKWLGLFHRRKHQYGRFMMRLKLPN
18161_4 109 EPMALFMDGGVRDLAKIPMEVIDAAKLTKDDVDVRLKWLGLFHRRKHQYGRFMMRLKLPN

18161_1 181 GVTTSEQTRYLASVIEAYGADGCADVTTTRQNWQIRGVTLDPVPAILDGLRAVGLTSLQSG
18161_2 169 GVTTSEQTRYLASVIEAYGADGCADVTTTRQNWQIRGVTLDPVPAILDGLRAVGLTSLQSG
18161_3 181 GVTTSEQTRYLASVIEAYGADGCADVTTTRQNWQIRGVTLDPVPAILDGLRAVGLTSLQSG
18161_4 169 GVTTSEQTRYLASVIEAYGADGCADVTTTRQNWQIRGVTLDPVPAILDGLRAVGLTSLQSG

18161_1 241 MDNVRNPVGNPLAGVDPHEIVDTRPYTNLLSSYITSNSQGNPAITNLVRNLTVTRPGANR
18161_2 229 MDNVRNPVGNPLAGVDPHEIVDTRPYTNLLSSYITSNSQGNPAITNLVRNLTVTRPGANR
18161_3 241 MDNVRNPVGNPLAGVDPHEIVDTRPYTNLLSSYITSNSQGNPAITNI-----
18161_4 229 MDNVRNPVGNPLAGVDPHEIVDTRPYTNLLSSYITSNSQGNPAITNI-----

18161_1 301 PRKWNVCVIGSHDLYEHPHINDLAYMPAVKDGKFGFNLLVGGFISPKRWAEALPLDAWVA
18161_2 289 PRKWNVCVIGSHDLYEHPHINDLAYMPAVKDGKFGFNLLVGGFISPKRWAEALPLDAWVA
18161_3 288 PRKWNVCVIGSHDLYEHPHINDLAYMPAVKDGKFGFNLLVGGFISPKRWAEALPLDAWVA
18161_4 276 PRKWNVCVIGSHDLYEHPHINDLAYMPAVKDGKFGFNLLVGGFISPKRWAEALPLDAWVA

18161_1 361 GDDVVPACKAILEAYRDLGFRGNRQKTRMMWLIDELGMEVFRSEVEKRMPPNGVLERAAAE
18161_2 349 GDDVVPACKAILEAYRDLGFRGNRQKTRMMWLIDELGMEVFRSEVEKRMPPNGVLERAAAE
18161_3 348 GDDVVPACKAILEAYRDLGFRGNRQKTRMMWLIDELGMEVFRSEVEKRMPPNGVLERAAAE
18161_4 336 GDDVVPACKAILEAYRDLGFRGNRQKTRMMWLIDELGMEVFRSEVEKRMPPNGVLERAAAE

18161_1 421 DLVDKKWERRDYLGVHPQKQEGLSYVGLHVPVGRLQAADMFEALARLADEYGTGELRLTVE
18161_2 409 DLVDKKWERRDYLGVHPQKQEGLSYVGLHVPVGRLQAADMFEALARLADEYGTGELRLTVE
18161_3 408 DLVDKKWERRDYLGVHPQKQEGLSYVGLHVPVGRLQAADMFEALARLADEYGTGELRLTVE
18161_4 396 DLVDKKWERRDYLGVHPQKQEGLSYVGLHVPVGRLQAADMFEALARLADEYGTGELRLTVE

18161_1 481 QNVVLPNVSNRRLDALLAEPLLQRQLSPQPSLLLRLVACTGNQFCGQAI IETKARALQ
18161_2 469 QNVVLPNVSNRRLDALLAEPLLQRQLSPQPSLLLRLVACTGNQFCGQAI IETKARALQ
18161_3 468 QNVVLPNVSNRRLDALLAEPLLQRQLSPQPSLLLRLVACTGNQFCGQAI IETKARALQ
18161_4 456 QNVVLPNVSNRRLDALLAEPLLQRQLSPQPSLLLRLVACTGNQFCGQAI IETKARALQ

18161_1 541 VAREVEKRVAVPRPVRMHWTCGPNSCAQVQVADIGFMGCLTKDRDGKVVEAADIFVGGRV
18161_2 529 VAREVEKRVAVPRPVRMHWTCGPNSCAQVQVADIGFMGCLTKDRDGKVVEAADIFVGGRV
18161_3 528 VAREVEKRVAVPRPVRMHWTCGPNSCAQVQVADIGFMGCLTKDRDGKVVEAADIFVGGRV
18161_4 516 VAREVEKRVAVPRPVRMHWTCGPNSCAQVQVADIGFMGCLTKDRDGKVVEAADIFVGGRV

18161_1 601 GSDSHLADVYRKSVPCDLPVIVADLLVERFGAVPREREEDDEE
18161_2 589 GSDSHLADVYRKSVPCDLPVIVADLLVERFGAVPREREEDDEE
18161_3 588 GSDSHLADVYRKSVPCDLPVIVADLLVERFGAVPREREEDDEE
18161_4 576 GSDSHLADVYRKSVPCDLPVIVADLLVERFGAVPREREEDDEE

```

Figure 16. Alignment of proteins encoded by Zm00001d018161. Amino acid sequence of gene Zm00001d018161 showing four different versions of the protein that could be produced by the DNA sequence based on the revised model. The absence sequence is denoted by dashes, meaning the next amino acid in the chain would be the next letter after the dashes. Identities between different versions of protein are denoted by the shaded boxes. Note also that the boxes may be shaded even if the sequence is “skipped” in the comparison protein.

The next gene that was evaluated was Zm00001d025984, which encodes glutamic dehydrogenase 2 (Tello-Ruiz *et. al.* 2017). The original gene structure was compared with the data in the Apollo annotation platform and a model was produced as noted in the methods (Figure 17). Comparison of the v5 and Mikado RNA models showed the presence of an alternative 3' splice site (*, Figure 17A) and was supported by RNA sequencing evidence (*, Figure 17B), so it was incorporated into the SuperTranscript (* Figure 17C, Table 2) Additionally, RNA sequencing data supported the presence of another 3' alternative splice site (**, Figure 17B), so this was incorporated into the SuperTranscript (**, Figure 17C; Table 2).

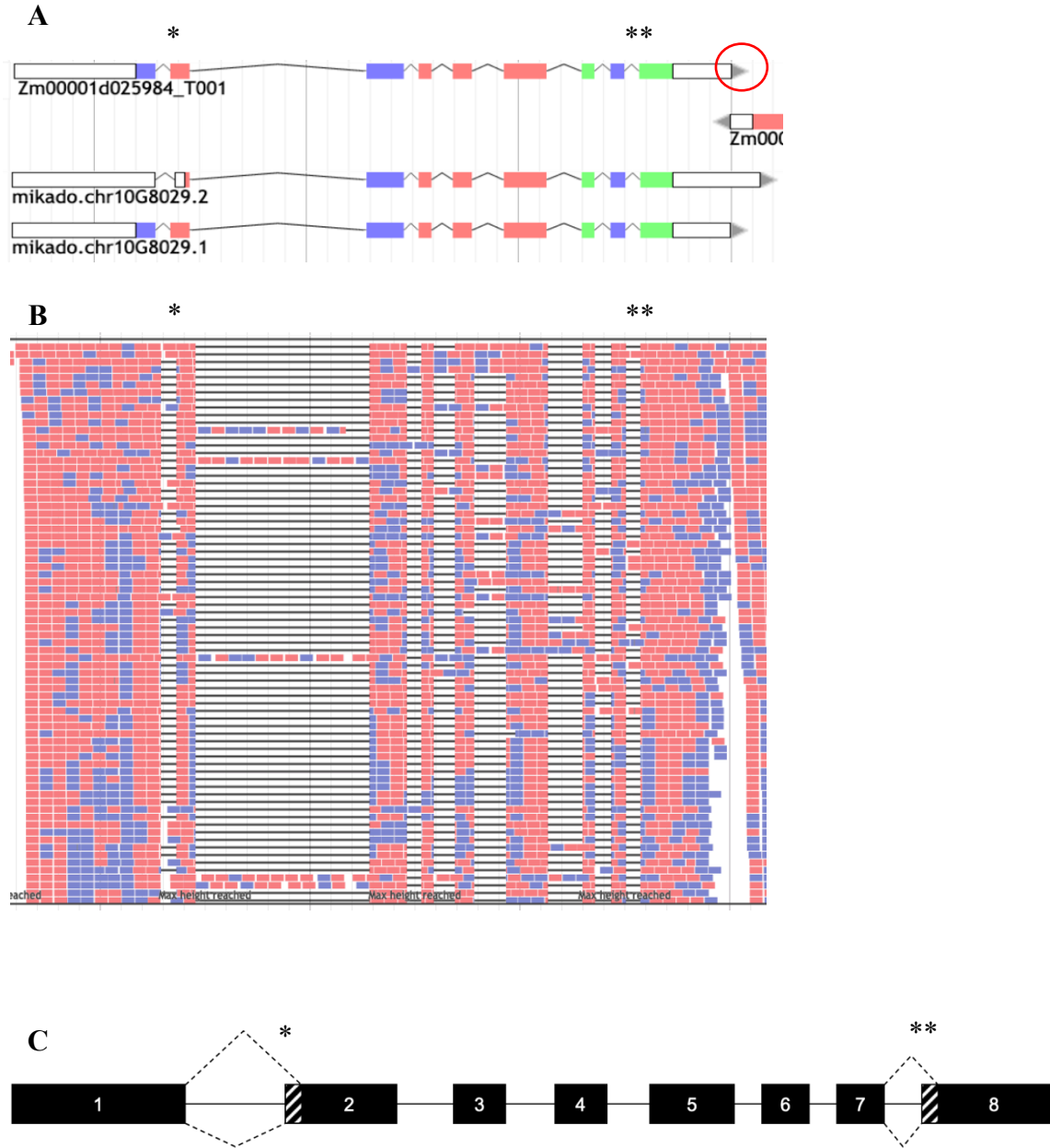


Figure 17. Evidence for reannotation for gene Zm00001d025984. Panel A shows the original structure of gene Zm00001d025984. Notice that the arrow on the original structure in panel A indicates the 3' end is on the left side of the gene (red circle) so the direction of the gene will be the reversed in the final model. The green and purple boxes show differentiation between the gene's reading frames. Panel B shows the final revised model. Note that the final model shows the RNA strand 5' to 3', which is the reverse of how the original model is displayed. The numbered black boxes indicate exons and their sequence from left to right. The lines between exons indicate introns.

Using the SuperTranscript as a guide, four RNAs were constructed in the Apollo space. The encoded amino acid sequences for each transcript were then downloaded from the Apollo workspace and aligned using Clustal Omega (Sievers *et. al.* 2011) and shaded using BOXSHADE to allow visualization of similarities and differences between the proteins encoded by the different transcripts. The four transcripts produce proteins that differ in a couple of regions (Figure 18). Transcript 2 encodes a protein missing 2 amino acids compared to transcript 1. Transcript 3 encodes a protein missing 62 amino acids compared to transcript 1. Transcript 4 encodes a protein missing 64 amino acids compared to transcript 1. Lastly, the sequence of amino acids encoded by each transcript was analyzed for the presence of known protein domains using the SMART domain sequence analysis program (Letunic and Bork 2017). The glutamate/leucine/phenylalanine/valine dehydrogenase domain was found present in the gene.

```

25984_3 1 -----
25984_2 1 MNALAATTRNFRRASKLLGLDSKLEQSLLIPFREIKVECTIPKDDGSLATFVGFRVQHDN
25984_1 1 MNALAATTRNFRRASKLLGLDSKLEQSLLIPFREIKVECTIPKDDGSLATFVGFRVQHDN
25984_4 1 -----

25984_3 1 ----MKGGIRYHNEVDPDEVNALAQLMTWKTAVAAVPYGGAKGGIGCSPGELSRSELERL
25984_2 61 ARGPMKGGIRYHNEVDPDEVNALAQLMTWKTAVAAVPYGGAKGGIGCSPGELSRSELERL
25984_1 61 ARGPMKGGIRYHNEVDPDEVNALAQLMTWKTAVAAVPYGGAKGGIGCSPGELSRSELERL
25984_4 1 ----MKGGIRYHNEVDPDEVNALAQLMTWKTAVAAVPYGGAKGGIGCSPGELSRSELERL

25984_3 57 TRVFTQKIHDLIGTHTDVPAPDMGTNAQTMAWMLDEYSKFHGHS PAVVTGKPIDLGGS LG
25984_2 121 TRVFTQKIHDLIGTHTDVPAPDMGTNAQTMAWMLDEYSKFHGHS PAVVTGKPIDLGGS LG
25984_1 121 TRVFTQKIHDLIGTHTDVPAPDMGTNAQTMAWMLDEYSKFHGHS PAVVTGKPIDLGGS LG
25984_4 57 TRVFTQKIHDLIGTHTDVPAPDMGTNAQTMAWMLDEYSKFHGHS PAVVTGKPIDLGGS LG

25984_3 117 RDAATGRGVMYATEALLAEYGKCSGSTFVIQGFNVGSWAARLIHEKGGKIIAIGDVTG
25984_2 181 RDAATGRGVMYATEALLAEYGKCSGSTFVIQGFNVGSWAARLIHEKGGKIIAIGDVTG
25984_1 181 RDAATGRGVMYATEALLAEYGKCSGSTFVIQGFNVGSWAARLIHEKGGKIIAIGDVTG
25984_4 117 RDAATGRGVMYATEALLAEYGKCSGSTFVIQGFNVGSWAARLIHEKGGKIIAIGDVTG

25984_3 177 SIRNTAGIDIPALVKHRNEGHAMKDFDGAEVLDSTELLVHDCDVLVPCALGGVLNKDNAP
25984_2 241 SIRNTAGIDIPALVKHRNEGHAMKDFDGAEVLDSTELLVHDCDVLVPCALGGVLNKDNAP
25984_1 241 SIRNTAGIDIPALVKHRNEGHAMKDFDGAEVLDSTELLVHDCDVLVPCALGGVLNKDNAP
25984_4 177 SIRNTAGIDIPALVKHRNEGHAMKDFDGAEVLDSTELLVHDCDVLVPCALGGVLNKDNAP

25984_3 237 DVKAKFVIEAANHPTDPEADEILAKKGVVVLVDIYANSGGVVVS YFEWVQVQNIQGMWD
25984_2 301 DVKAKFVIEAANHPTDPEADEILAKKGVVVLVDIYANSGGVVVS YFEW----NIQGMWD
25984_1 301 DVKAKFVIEAANHPTDPEADEILAKKGVVVLVDIYANSGGVVVS YFEWVQVQNIQGMWD
25984_4 237 DVKAKFVIEAANHPTDPEADEILAKKGVVVLVDIYANSGGVVVS YFEWVQVQNIQGMWD

25984_3 297 EEKVNDELEKYMSSAFQHKAMCKSLDCDLRMGAFTLGVNRVARATLLRGWEA
25984_2 357 EEKVNDELEKYMSSAFQHKAMCKSLDCDLRMGAFTLGVNRVARATLLRGWEA
25984_1 359 EEKVNDELEKYMSSAFQHKAMCKSLDCDLRMGAFTLGVNRVARATLLRGWEA
25984_4 295 EEKVNDELEKYMSSAFQHKAMCKSLDCDLRMGAFTLGVNRVARATLLRGWEA

```

Figure 18. Alignment of proteins encoded by Zm00001d025984. Amino acid sequence of gene Zm00001d025984 showing four different versions of the protein that could be produced by the DNA sequence based on the revised model. The absence sequence is denoted by dashes, meaning the next amino acid in the chain would be the next letter after the dashes. Identities between different versions of protein are denoted by the shaded boxes. Note also that the boxes may be shaded even if the sequence is “skipped” in the comparison protein.

The next gene that was evaluated was Zm00001d028750, which encodes asparagine synthetase 3 (Tello-Ruiz *et. al.* 2017). The original gene structure was compared with the data in the Apollo annotation platform and a model was produced as noted in the methods (Figure 19). RNAseq Mikado models supported both the presence and absence of an intron at one location (*, Figure 19A), and RNA sequencing data supported the absence of an intron in that location (*, Figure 19B), so it was labeled as a retained intron in the revised model (*, Figure 19C, Table 2). Comparison of v5 and IsoSeq models showed the presence and absence of an intron at one location (**, Figure 19A), and RNA sequencing data supported the absence of an intron in that location (**, Figure 19B), so it was labeled as a retained intron in the revised model (**, Figure 19C, Table 2). Comparison between v5 and an flc model shows both the absence and presence of an intron in the same location (***, Figure 19A), and RNA sequencing data shows the absence of that intron (***, Figure 19B), so it was incorporated into the SuperTranscript (***, Figure 19C, Table 2). Finally, two end of transcription sites were identified in the Mikado RNA models (****, Figure 19A), and supported in the RNA sequencing evidence, (****, Figure 19B) and incorporated into the SuperTranscript (****, Figure 19C).

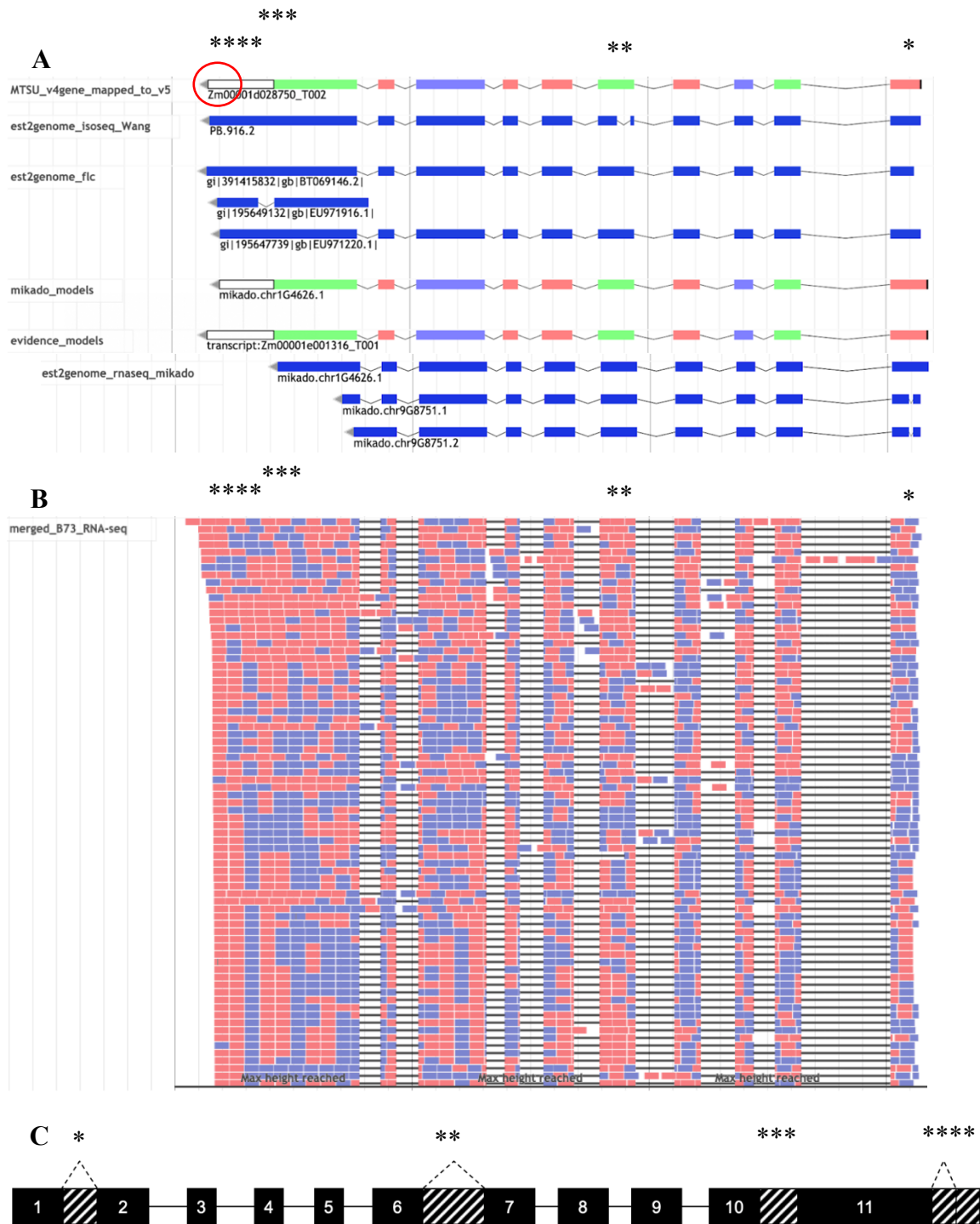


Figure 19. Evidence for reannotation for gene Zm00001d028750. Panel A shows the original structure of gene Zm00001d028750. Notice that the arrow on the original structure in panel A indicates the 3' end is on the left side of the gene (red circle) so the direction of the gene will be the reversed in the final model. The green and purple boxes show differentiation between the gene's reading frames. Panel B shows the final revised model. Note that the final model shows the RNA strand 5' to 3', which is the reverse of how the original model is displayed. The numbered black boxes indicate exons and their sequence from left to right. The lines between exons indicate introns.

Using the SuperTranscript as a guide, eight RNAs were constructed in the Apollo space. The encoded amino acid sequences for each transcript were then downloaded from the Apollo workspace and aligned using Clustal Omega (Sievers *et. al.* 2011) and shaded using BOXSHADE to allow visualization of similarities and differences between the proteins encoded by the different transcripts. The eight transcripts produce proteins that differ in various regions (Figure 20). Transcript 1 encodes a protein that is 12 amino acids shorter compared to transcript 2. Transcript 3 encodes a protein missing 19 amino acids compared to transcript 2. Transcript 6 encodes a protein missing 24 amino acids compared to transcript 2. Transcript 5 encodes a protein missing 36 amino acids compared to transcript 2. Transcript 4 encodes a protein missing 38 amino acids compared to transcript 2. Transcript 7 encodes a protein missing 43 amino acids compared to transcript 2. Transcript 8 encodes a protein missing 62 amino acids compared to transcript 2. Lastly, the sequence of amino acids encoded by each transcript was analyzed for the presence of known protein domains using the SMART domain sequence analysis program (Letunic and Bork 2017). No identifiable domains were found present in the transcripts.

```

28750_3 1 -----SQHVSSPSQKKK-----KNCSAGIMCGILAVLGCSDWSQAKRARI
28750_1 1 -----PSTSRLPKKKKKTARLLLLRRRAGIMCGILAVLGCSDWSQAKRARI
28750_2 1 YKKAAILAWFPH PSTSRLPKKKKKTARLLLLRRRAGIMCGILAVLGCSDWSQAKRARI
28750_4 1 -----MCGILAVLGCSDWSQAKRARI
28750_7 1 -----SQHVSSPSQKKK-----KNCSAGIMCGILAVLGCSDWSQAKRARI
28750_5 1 -----PSTSRLPKKKKKTARLLLLRRRAGIMCGILAVLGCSDWSQAKRARI
28750_6 1 YKKAAILAWFPH PSTSRLPKKKKKTARLLLLRRRAGIMCGILAVLGCSDWSQAKRARI
28750_8 1 -----MCGILAVLGCSDWSQAKRARI

28750_3 42 ACSRRRLKHRGPDWSGLYQHEGNFLAQORLAVVSPLSGDQPLFNEDRTVVVVANGEIYNHK
28750_1 49 ACSRRRLKHRGPDWSGLYQHEGNFLAQORLAVVSPLSGDQPLFNEDRTVVVVANGEIYNHK
28750_2 61 ACSRRRLKHRGPDWSGLYQHEGNFLAQORLAVVSPLSGDQPLFNEDRTVVVVANGEIYNHK
28750_4 23 ACSRRRLKHRGPDWSGLYQHEGNFLAQORLAVVSPLSGDQPLFNEDRTVVVVANGEIYNHK
28750_7 42 ACSRRRLKHRGPDWSGLYQHEGNFLAQORLAVVSPLSGDQPLFNEDRTVVVVANGEIYNHK
28750_5 49 ACSRRRLKHRGPDWSGLYQHEGNFLAQORLAVVSPLSGDQPLFNEDRTVVVVANGEIYNHK
28750_6 61 ACSRRRLKHRGPDWSGLYQHEGNFLAQORLAVVSPLSGDQPLFNEDRTVVVVANGEIYNHK
28750_8 23 ACSRRRLKHRGPDWSGLYQHEGNFLAQORLAVVSPLSGDQPLFNEDRTVVVVANGEIYNHK

28750_3 102 NVRKQFTGTHNFSTGSDCEVIIPLYEKYGENFVMDLGDVFAFVLYDTRDRTYVAARDAIG
28750_1 109 NVRKQFTGTHNFSTGSDCEVIIPLYEKYGENFVMDLGDVFAFVLYDTRDRTYVAARDAIG
28750_2 121 NVRKQFTGTHNFSTGSDCEVIIPLYEKYGENFVMDLGDVFAFVLYDTRDRTYVAARDAIG
28750_4 83 NVRKQFTGTHNFSTGSDCEVIIPLYEKYGENFVMDLGDVFAFVLYDTRDRTYVAARDAIG
28750_7 102 NVRKQFTGTHNFSTGSDCEVIIPLYEKYGENFVMDLGDVFAFVLYDTRDRTYVAARDAIG
28750_5 109 NVRKQFTGTHNFSTGSDCEVIIPLYEKYGENFVMDLGDVFAFVLYDTRDRTYVAARDAIG
28750_6 121 NVRKQFTGTHNFSTGSDCEVIIPLYEKYGENFVMDLGDVFAFVLYDTRDRTYVAARDAIG
28750_8 83 NVRKQFTGTHNFSTGSDCEVIIPLYEKYGENFVMDLGDVFAFVLYDTRDRTYVAARDAIG

28750_3 162 VNPLYIGWSDGSGVWIASEMKALEDVCFEIFFPPGHLYSSAGGGFRRWYTPHWFQEQVP
28750_1 169 VNPLYIGWSDGSGVWIASEMKALEDVCFEIFFPPGHLYSSAGGGFRRWYTPHWFQEQVP
28750_2 181 VNPLYIGWSDGSGVWIASEMKALEDVCFEIFFPPGHLYSSAGGGFRRWYTPHWFQEQVP
28750_4 143 VNPLYIGWSDGSGVWIASEMKALEDVCFEIFFPPGHLYSSAGGGFRRWYTPHWFQEQVP
28750_7 162 VNPLYIGWSDGSGVWIA-----SGGGFRRWYTPHWFQEQVP
28750_5 169 VNPLYIGWSDGSGVWIA-----SGGGFRRWYTPHWFQEQVP
28750_6 181 VNPLYIGWSDGSGVWIA-----SGGGFRRWYTPHWFQEQVP
28750_8 143 VNPLYIGWSDGSGVWIA-----SGGGFRRWYTPHWFQEQVP

28750_3 222 RMPYQPLVLRFAFEKAVIKRLMTDVPFGVLLSGGLDSSLVASVTKRHLVETEAAEKFGTE
28750_1 229 RMPYQPLVLRFAFEKAVIKRLMTDVPFGVLLSGGLDSSLVASVTKRHLVETEAAEKFGTE
28750_2 241 RMPYQPLVLRFAFEKAVIKRLMTDVPFGVLLSGGLDSSLVASVTKRHLVETEAAEKFGTE
28750_4 203 RMPYQPLVLRFAFEKAVIKRLMTDVPFGVLLSGGLDSSLVASVTKRHLVETEAAEKFGTE
28750_7 198 RMPYQPLVLRFAFEKAVIKRLMTDVPFGVLLSGGLDSSLVASVTKRHLVETEAAEKFGTE
28750_5 205 RMPYQPLVLRFAFEKAVIKRLMTDVPFGVLLSGGLDSSLVASVTKRHLVETEAAEKFGTE
28750_6 217 RMPYQPLVLRFAFEKAVIKRLMTDVPFGVLLSGGLDSSLVASVTKRHLVETEAAEKFGTE
28750_8 179 RMPYQPLVLRFAFEKAVIKRLMTDVPFGVLLSGGLDSSLVASVTKRHLVETEAAEKFGTE

28750_3 282 LHSFVVGLEGSPDLKAAREVADYLGTHHEFHFTVQDGDIAEEVIYHDETYDVTIRAS
28750_1 289 LHSFVVGLEGSPDLKAAREVADYLGTHHEFHFTVQDGDIAEEVIYHDETYDVTIRAS
28750_2 301 LHSFVVGLEGSPDLKAAREVADYLGTHHEFHFTVQDGDIAEEVIYHDETYDVTIRAS
28750_4 263 LHSFVVGLEGSPDLKAAREVADYLGTHHEFHFTVQDGDIAEEVIYHDETYDVTIRAS
28750_7 258 LHSFVVGLEGSPDLKAAREVADYLGTHHEFHFTVQDGDIAEEVIYHDETYDVTIRAS
28750_5 265 LHSFVVGLEGSPDLKAAREVADYLGTHHEFHFTVQDGDIAEEVIYHDETYDVTIRAS
28750_6 277 LHSFVVGLEGSPDLKAAREVADYLGTHHEFHFTVQDGDIAEEVIYHDETYDVTIRAS
28750_8 239 LHSFVVGLEGSPDLKAAREVADYLGTHHEFHFTVQDGDIAEEVIYHDETYDVTIRAS

28750_3 342 TPMFLMARKIKSLGVKMVLSGEGSDELLGGYLYFHFAPNKEEFHRETCKRVKALHQYDCI
28750_1 349 TPMFLMARKIKSLGVKMVLSGEGSDELLGGYLYFHFAPNKEEFHRETCKRVKALHQYDCI
28750_2 361 TPMFLMARKIKSLGVKMVLSGEGSDELLGGYLYFHFAPNKEEFHRETCKRVKALHQYDCI
28750_4 323 TPMFLMARKIKSLGVKMVLSGEGSDELLGGYLYFHFAPNKEEFHRETCKRVKALHQYDCI
28750_7 318 TPMFLMARKIKSLGVKMVLSGEGSDELLGGYLYFHFAPNKEEFHRETCKRVKALHQYDCI
28750_5 325 TPMFLMARKIKSLGVKMVLSGEGSDELLGGYLYFHFAPNKEEFHRETCKRVKALHQYDCI

28750_6 337 TPMFLMARKIKSLGVKMVLSGEGSDELLGGYLYFHFAPNKEEFHRETCKRVKALHQYDCI
28750_8 299 TPMFLMARKIKSLGVKMVLSGEGSDELLGGYLYFHFAPNKEEFHRETCKRVKALHQYDCI

```

Figure 20. Alignment of proteins encoded by Zm00001d028750 – Part 1. Amino acid sequence of gene Zm00001d028750 showing eight different versions of the protein that could be produced by the DNA sequence based on the revised model. The absence sequence is denoted by dashes, meaning the next amino acid in the chain would be the next letter after the dashes. Identities between different versions of protein are denoted by the shaded boxes. Note also that the boxes may be shaded even if the sequence is “skipped” in the comparison protein.

```

28750_3 402 RANKATSAWGLEVRVPFLDKEFINVAMGMDPEWKMYDKNLGRIEKWVMRKAFTDDDEHPYL
28750_1 409 RANKATSAWGLEVRVPFLDKEFINVAMGMDPEWKMYDKNLGRIEKWVMRKAFTDDDEHPYL
28750_2 421 RANKATSAWGLEVRVPFLDKEFINVAMGMDPEWKMYDKNLGRIEKWVMRKAFTDDDEHPYL
28750_4 383 RANKATSAWGLEVRVPFLDKEFINVAMGMDPEWKMYDKNLGRIEKWVMRKAFTDDDEHPYL
28750_7 378 RANKATSAWGLEVRVPFLDKEFINVAMGMDPEWKMYDKNLGRIEKWVMRKAFTDDDEHPYL
28750_5 385 RANKATSAWGLEVRVPFLDKEFINVAMGMDPEWKMYDKNLGRIEKWVMRKAFTDDDEHPYL
28750_6 397 RANKATSAWGLEVRVPFLDKEFINVAMGMDPEWKMYDKNLGRIEKWVMRKAFTDDDEHPYL
28750_8 359 RANKATSAWGLEVRVPFLDKEFINVAMGMDPEWKMYDKNLGRIEKWVMRKAFTDDDEHPYL

28750_3 462 PKHILYRQKEQFSDGVGYNWIDGLKSFTQQVTDMMNNAQMFYPNTPVNKEAYYYRMI
28750_1 469 PKHILYRQKEQFSDGVGYNWIDGLKSFTQQVTDMMNNAQMFYPNTPVNKEAYYYRMI
28750_2 481 PKHILYRQKEQFSDGVGYNWIDGLKSFTQQVTDMMNNAQMFYPNTPVNKEAYYYRMI
28750_4 443 PKHILYRQKEQFSDGVGYNWIDGLKSFTQQVTDMMNNAQMFYPNTPVNKEAYYYRMI
28750_7 438 PKHILYRQKEQFSDGVGYNWIDGLKSFTQQVTDMMNNAQMFYPNTPVNKEAYYYRMI
28750_5 445 PKHILYRQKEQFSDGVGYNWIDGLKSFTQQVTDMMNNAQMFYPNTPVNKEAYYYRMI
28750_6 457 PKHILYRQKEQFSDGVGYNWIDGLKSFTQQVTDMMNNAQMFYPNTPVNKEAYYYRMI
28750_8 419 PKHILYRQKEQFSDGVGYNWIDGLKSFTQQVTDMMNNAQMFYPNTPVNKEAYYYRMI

28750_3 522 FERLFPQDSARETVPWGPSIACSTPAAIEWVEQWKASNDPSGRFISSHDSAATDHTGGKF
28750_1 529 FERLFPQDSARETVPWGPSIACSTPAAIEWVEQWKASNDPSGRFISSHDSAATDHTGGKF
28750_2 541 FERLFPQDSARETVPWGPSIACSTPAAIEWVEQWKASNDPSGRFISSHDSAATDHTGGKF
28750_4 503 FERLFPQDSARETVPWGPSIACSTPAAIEWVEQWKASNDPSGRFISSHDSAATDHTGGKF
28750_7 498 FERLFPQDSARETVPWGPSIACSTPAAIEWVEQWKASNDPSGRFISSHDSAATDHTGGKF
28750_5 505 FERLFPQDSARETVPWGPSIACSTPAAIEWVEQWKASNDPSGRFISSHDSAATDHTGGKF
28750_6 517 FERLFPQDSARETVPWGPSIACSTPAAIEWVEQWKASNDPSGRFISSHDSAATDHTGGKF
28750_8 479 FERLFPQDSARETVPWGPSIACSTPAAIEWVEQWKASNDPSGRFISSHDSAATDHTGGKF

28750_3 582 AVANGGGHGAANGTVNGKDVAIAV
28750_1 589 AVANGGGHGAANGTVNGKDVAIAV
28750_2 601 AVANGGGHGAANGTVNGKDVAIAV
28750_4 563 AVANGGGHGAANGTVNGKDVAIAV
28750_7 558 AVANGGGHGAANGTVNGKDVAIAV
28750_5 565 AVANGGGHGAANGTVNGKDVAIAV
28750_6 577 AVANGGGHGAANGTVNGKDVAIAV
28750_8 539 AVANGGGHGAANGTVNGKDVAIAV

```

Figure 20. Alignment of proteins encoded by Zm00001d028750 – Part 2. Amino acid sequence of gene Zm00001d028750 showing eight different versions of the protein that could be produced by the DNA sequence based on the revised model. The absence sequence is denoted by dashes, meaning the next amino acid in the chain would be the next letter after the dashes. Identities between different versions of protein are denoted by the shaded boxes. Note also that the boxes may be shaded even if the sequence is “skipped” in the comparison protein.

Conclusions

Nitrogen-use efficiency plays an important role in *Zea mays*' growth and development. Nitrogen depletion by *Zea mays*, therefore, requires additional fertilizer and crop rotation in order to maintain both croppable soil and plant health. Understanding the biochemistry and genetics of nitrogen-use efficiency may allow for development of corn varieties less dependent on financially costly fertilizer addition and time costly crop rotation.

As a first step in understanding the genetic and biochemical aspects of genes involved in corn nitrogen use, a systematic categorization of all transcripts and encoded proteins for nine NUE genes was undertaken using newly available whole transcriptome data from multiple corn tissues. Using Apollo genome annotation collaboration software, (Ware) existing v5 corn gene models and transcriptome data were used to generate a SuperTranscript for each gene. Exon skipping, intron retention, alternate start and end of transcription, and alternate 5' and 3' splice sites were observed (Table 1). Thirty-five differences were found across the nine genes including: ten intron retentions, six alternate 5' splice sites, eleven alternate 3' splice sites, three alternative exons, three alternate end of transcription sites, one alternate start of transcription site, and one alternative intron (Table 2). Using these alternative features, old transcripts were re-annotated or constructed to accurately reflect the breadth of transcript versions (isoforms) produced by each gene. Furthermore, the encoded proteins for each gene were compared by multiple alignment and computational domain analysis. While most transcripts encoded slightly different proteins, only one protein's domain analysis showed a known protein domain. Therefore, until additional domain information becomes available, or these proteins are

studied biochemically, it is unclear how the proteins encoded by the different transcripts might function differently.

In addition to the nine genes of focus in this study, there are other genes pertaining to nitrogen use efficiency, including *Zm00001d048050*, *Zm00001d028260*, *Zm00001d011610*, *Zm00001d043845*, *Zm00001d038948*, *Zm00001d034420*, *Zm00001d002052*, *Zm00001d045675*, *Zm00001d022152*, *Zm00001d014258*, and *Zm00001d007357* (Table 1). Once all NUE genes have been systematically studied for transcript isoforms, researchers will have a more complete view of all proteins involved in the process and can use this knowledge to investigate mechanisms to alter regulation of these proteins to generate corn varieties with specific characteristics.

Literature Cited

- Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, de Castro E, Duvaud S, Flegel V, Fortier A, Gasteiger E, *et al.* 2012. ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res.* 40(W1):W597-W603. doi:10.1093/nar/gks400
- Davidson NM, Hawkins ADK, Oshlack A. 2017. SuperTranscripts: a data driven reference for analysis and visualisation of transcriptomes. *Genome Biology.* 18(148). doi:10.1186/s13059-017-1284-1
- Kennett DJ, Thakar HB, VanDerwarker AM, Webster DL, Culleton BJ, Harper TK, Kistler L, Scheffler TE, Hirth K. 2017. High-precision chronology for central American maize diversification from El Gigante rockshelter, Honduras. *PNAS.* 114(34):9026-9031. doi:10.1073/pnas.1705052114
- Letunic I, Bork P. 2017. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Research.* 46(D1):D493-D496. <https://doi.org/10.1093/nar/gkx922>
- Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, Lopez R. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research.* 47(W1):W636-W641. doi:10.1093/nar/gkz268
- Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, Brown SD, Chang HY, El-Gebali S, Fraser MI, *et al.* 2019. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research.* doi:10.1093/nar/gky1100
- Shah TR, Prasad K, Kumar P. 2016. Maize—A potential source of human nutrition and health: A review. *Cogent Food & Agriculture.* 2(1). doi:10.1080/23311932.2016.1166995

Sharma LK, Bali SK. 2018. A Review of Methods to Improve Nitrogen Use Efficiency in Agriculture. Sustainability. 10(1). <https://doi.org/10.3390/su10010051>

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7:539. doi: 10.1038/msb.2011.75

Simons M, Saha R, Guillard L, Clément G, Armengaud P, Cañas R, Maranas CD, Lea PJ, Hirel B. 2014. Nitrogen-use efficiency in maize (*Zea mays* L.): from ‘omics’ studies to metabolic modelling. Journal of Experimental Botany. 65(19):5657-5671. doi:10.1093/jxb/eru227

Singh B. 2018. Are nitrogen fertilizers deleterious to soil health?. Agronomy. 8(4). <https://doi.org/10.3390/agronomy8040048>

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, *et al.* 2009. The B73 maize genome: complexity, diversity, and dynamics. Science. 326(5956):1112-1115. doi:10.1126/science.1178534

Tello-Ruiz MK, Naithani S, Stein JC, Gupta P, Campbell M, Olson A, Wei S, Preece J, Geniza MJ, Jiao Y, *et al.* 2017. Gramene 2018: unifying comparative genomics and pathway resources for plant research. Nucleic Acids Research. 2018; 46(D1). doi:10.1093/nar/gkx1111

Veljković VB, Biberdžić MO, Banković-Ilić IB, Djalović IG, Tasić MB, Nježić ZB, Stamenković OS. 2018. Biodiesel production from corn oil: A review. Renewable and Sustainable Energy Reviews. 2018; 91:531-548. <https://doi.org/10.1016/j.rser.2018.04.024>

Xu Y, Skinner DJ, Wu H, Palacios-Rojas N, Araus JL, Yan J, Gao S, Warburton ML, Crouch JH. 2009. Advances in maize genomics and their value for enhancing genetic gains from breeding. *International Journal of Plant Genomics*. 2009:1-30.
<https://doi.org/10.1155/2009/957602>