FINDING AND FIXING ANNOTATION ERRORS THROUGH COMMUNITY CURATION

Marcela K. Tello-Ruiz, PhD

@marcelakarey

PB20 PLANT BIOLOGY
JULY 27 - 31 | WORLDWIDE SUMMIT

#PlantBio20

Cristina F. Marco

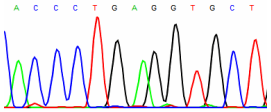
Outline

Marcela K. Tello-Ruiz

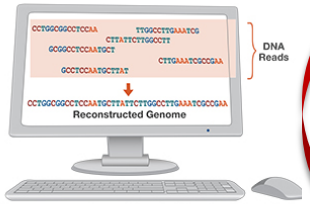
- Gene Structural Annotation & Community Curation – What?
- Maize Annotation Jamborees – How?
- Course-based Undergraduate Research Experience – Why?

GENE STRUCTURAL ANNOTATION

Sequence data



Assembly



AUTOMATED

MANUAL

Annotation

Structural

Create gene models

- Find ORFs and DNA features

Functional

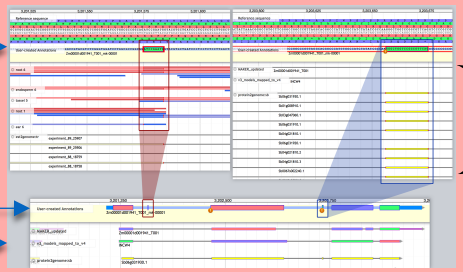
Annotate gene models

- Identify protein domains, gene families, expression patterns, inferred pathways

Assembly errors, limited coverage, conflicting evidence

MANUAL gene model

PREDICTED gene models



Experimental evidence (mRNA, protein)

GENE STRUCTURAL ANNOTATION

"The value of a genome is only as good as its annotation"

- Genome sequencing revolution → 1,000s new genomes

25 NAM

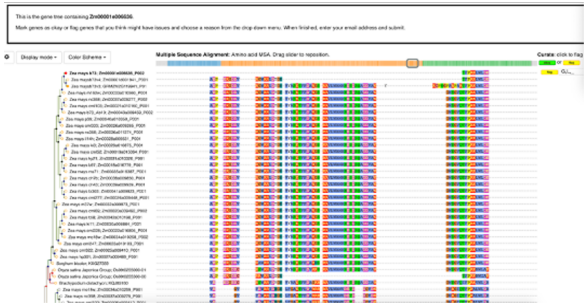
50 Sorghum

- Annotation errors are prevalent and very $$$
  - For example, CRISPR experiments
- Manual annotation is time-intensive ($$$)

Developing & Customizing Annotation Tools

- **Gramene Tree Visualizer**: Flag suspicious gene models

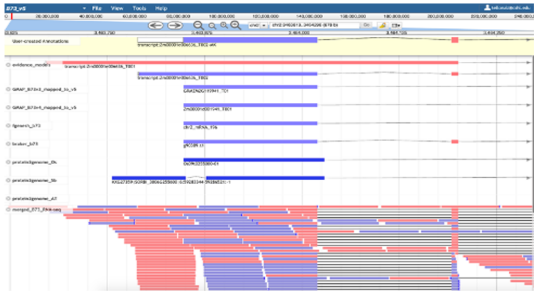
http://gramene.org



Andrew Olson

- **Apollo Browser**: Evaluate and curate flagged genes

http://genomearchitect.org



Kapeel Chougule

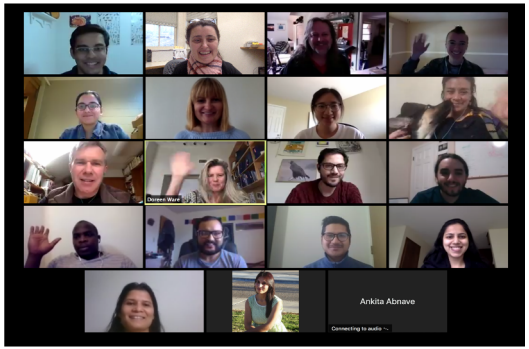
Scaling up to Community-Level

Cold Spring Harbor (Dec 2017)

St Louis (March 2019)

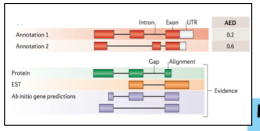
- 4 workshops (2017-2020)
- 51 participants; 38 institutions
- 12 PUI faculty, 8 researchers & 26 grad students
- ~250 genes manually curated

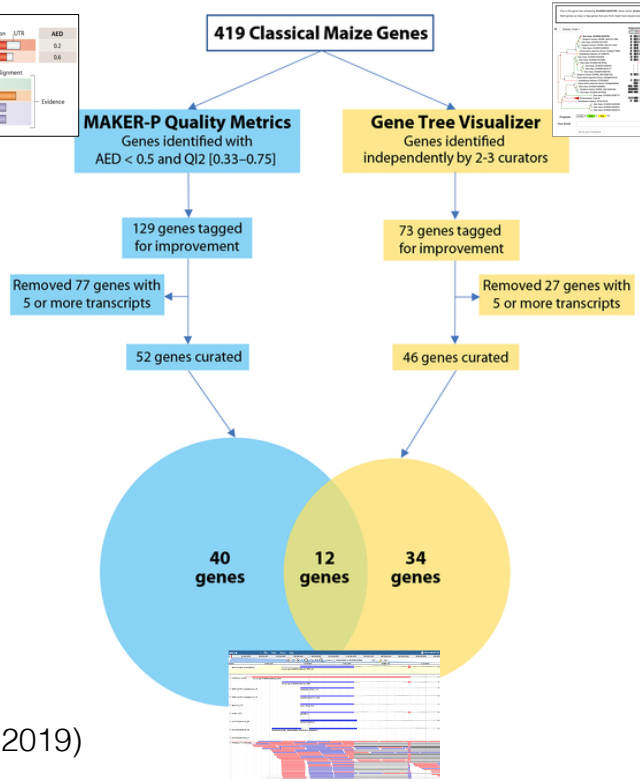
San Diego (Jan 2019)



Virtual (March 2020)

Community Annotation – Pilot Study





PLOS ONE | PUBLISH ABOUT BROWSE

Double triage to identify poorly annotated genes in maize: The missing link in community curation

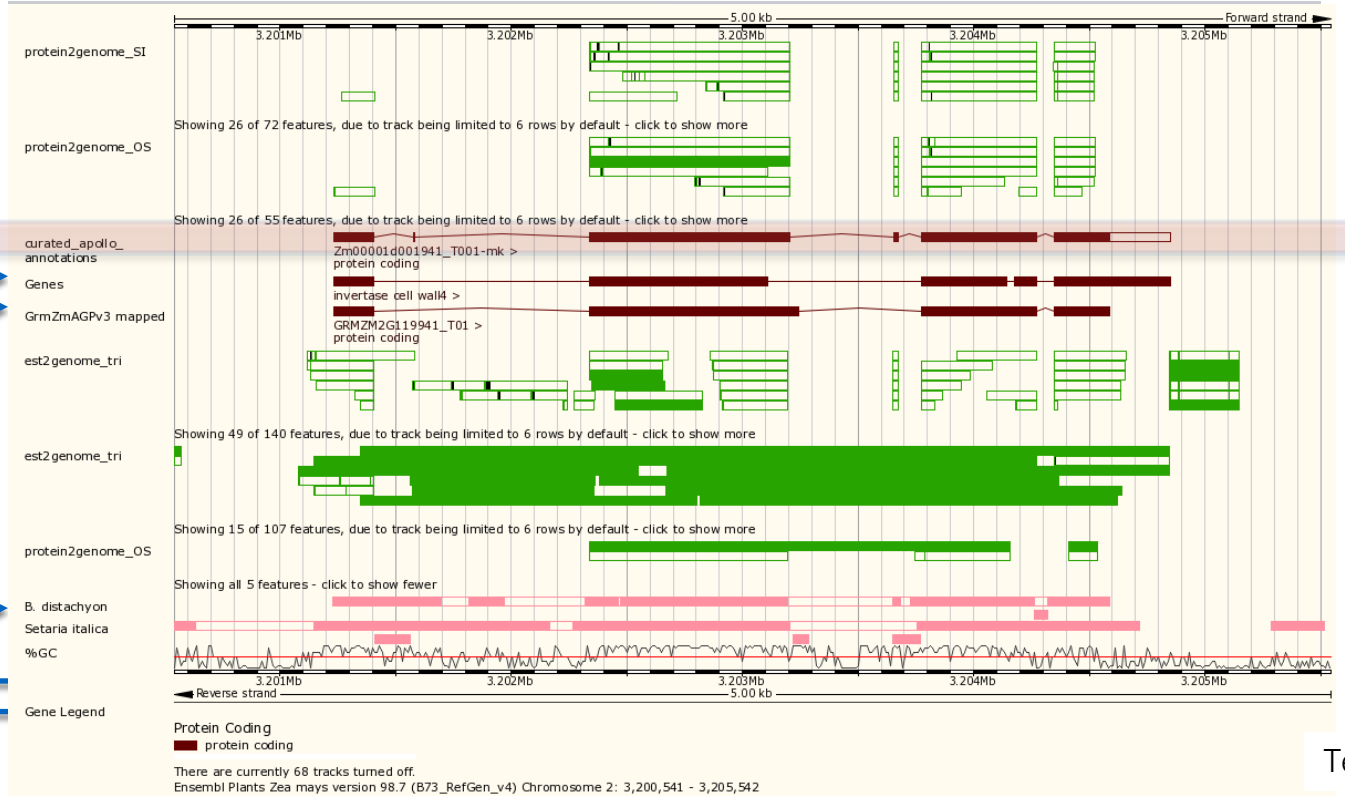
Marcela K. Tello-Ruiz, Cristina F. Marco, Fei-Man Hsu, Rajdeep S. Khangura, Pengfei Qiao, Sirjan Sapkota, Michelle C. Stitzer, Rachael Wasikowski, Hao Wu, Junpeng Zhan, Kapeel Chougule, Lindsay C. Barone, Cornel Ghiban, Demitri Muna, Andrew C. Olson, Liya Wang, Doreen Ware, David A. Micklos [view less]

Published: October 28, 2019 • https://doi.org/10.1371/journal.pone.0224086

Tello-Ruiz *et al* (2019)

109 Manually Curated Maize V4 genes

http://www.gramene.org/curated_maize_v4_gene_models



Manually curated

V4

V3

Manually curated

Predicted V4

Predicted V3

Evidence

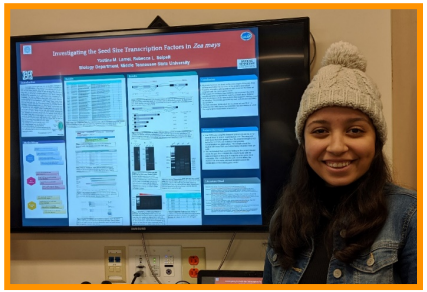
Tello-Ruiz *et al* (2019)

2019 Course-based Undergraduate Research Experiences

PUI Faculty
San Diego
Jan 2019

UCMO

MTSU



SUNY OW

MIDDLE TENNESSEE STATE UNIVERSITY

CUREs

Rebecca Seipelt

Regional University

- **Fall 2019**: Honors Genetics, Sophomores: 18 students (17 female, 1 male)
- **Spring 2020**: Undergrad Research, Soph-Junior (2 male, 1 female)

REVISING GENES IN THE NITROGEN USE EFFICIENCY PATHWAY OF ZEA MAYS FOR SUSTAINABLE AGRICULTURE

Jewel Larkins, Rija Asim, Rebecca L. Seipelt-Thiemann

Honors College, Biology Department, Middle Tennessee State University

MIDDLE TENNESSEE STATE UNIVERSITY

Introduction

•Maize (*Zea mays*) is one of the highest produced and most important grain crops worldwide.
•Nitrogen fertilizer application has been used to sustain high-yield agriculture as plant growth depends on the acquisition of nitrogen, an essential element deficient in most soils (Ladha et al. 2016).
•Of the total amount of nitrogen fertilizer applied to crops, less than 40% is taken up by the plant (Zuluaga et al. 2019).
•The remaining nitrogen is lost to the atmosphere as potent greenhouse gases or leached into the soil, polluting aquatic and terrestrial ecosystems (Hirel et al. 2011).
•Nitrogen Use Efficiency (NUE) comprises a plant's ability to take up nitrogen from the soil and convert it into grain yield (Kumar et al. 2018).
•Ultimately, NUE is a trait influenced by various genes. Knowing where in the plant these genes are expressed and when they are expressed, as well as in what forms, will be important in making genetic improvements aimed at expanding crop yield while being environmentally sustainable.

Hypothesis

Computationally predicted gene models for genes in the nutrient use efficiency pathway of maize will show both accuracies and inaccuracies that affect our understanding of maize's role in agricultural sustainability; inaccuracies will be revised to generate true, evidence-based gene models.

Materials & Methods

•Selection of the Nutrient Use Efficiency pathway in *Zea mays* and identification of the Reference Sequences (REFSEQ) for each gene using Gramene
•Identification of differences in conservation of amino acids and protein domains in *Zea mays* and related plant species
•Corn leaf RNA extraction and reverse transcription
•Resuspension of forward and reverse tubulin primers
•PCR and gel electrophoresis on corn leaf cDNA
•Determined splicing patterns observed for gene Zm00001d052233 using Apollo and designed primers for region of interest
•Found appropriate annealing temperature for primers through gradient PCR
•PCR and gel electrophoresis on leaf, root, and shoot cDNA for gene Zm00001d052233
•Analysis of all observed, expected and unexpected isoforms

Results

Table 1. Gene Summary Table

Table 2. Reference Sequence Domain Evidence

Domain source	Start	End	Description
PANTHER	5	331	-
Gene3D	68	302	Histidine phosphatase superfamily,
Superfamily	81	297	Histidine phosphatase superfamily,
SMART	80	266	Histidine phosphatase superfamily, clade-1
Pfam	81	156	Histidine phosphatase superfamily, clade-1
CDD	81	296	Histidine phosphatase superfamily, clade-1
Pfam	176	284	Histidine phosphatase superfamily, clade-1
PANTHER	1	331	Phosphoglycerate mutase 1
HAMAP	78	305	Phosphoglycerate mutase 1
PROSITE patterns	83	92	Phosphoglycerate/bisphosphoglycerate mutase, active site

Figure 1. Zm00001d052233 Reference Sequence and Domain Evidence. The region of interest includes exons 6 and 7 which are located near the end of Zm00001d052233 and include all domains.

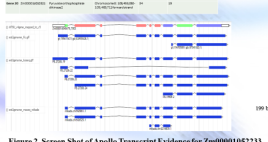


Figure 2. Screen Shot of Apollo Transcript Evidence for Zm00001052233

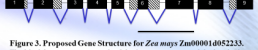
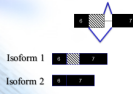


Figure 3. Proposed Gene Structure for *Zea mays* Zm00001d052233. The black boxes represent each of the 9 exons present and the hatched boxes represent alternative regions. Introns of Zm00001d052233 are represented by horizontal lines. Angled lines above and below the image represent splicing patterns. The horizontal line below the image indicates the chosen region of interest. Feature sizes are relative to their actual size in base pairs (bp). This structure was designed using RNA evidence provided via the Apollo instance for *Zea mays* genome, version 5 (unpublished data, Micklos and Ware 2019).



	Total base pairs (bp)	Expected PCR bp
Isoform 1	392 bp	237 bp
Isoform 2	362 bp	207 bp

Figure 4. Alternative Splices of Region of Interest Including Exon 6 and Exon 7. The splice noted by the angled line below the gene structure results in isoform 2 while the angled line above results in isoform 1. Expected RT-PCR base pair lengths in comparison to the total are noted at right. Primers to detect both transcripts were detected using Primer3plus (Rozen et al. 2000).

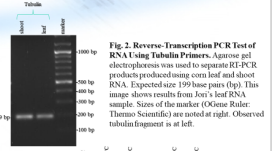


Fig. 2. Reverse-Transcription PCR Test of RNA Using Tubulin Primers. Agarose gel electrophoresis was used to separate RT-PCR products produced using corn leaf and shoot RNA. Expected size 199 base pairs (bp). This image shows results from Jori's leaf RNA sample. Sizes of the marker (OGene Ruler: Thermo Scientific) are noted at right. Observed tubulin fragment is at left.

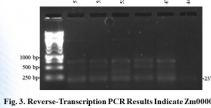


Fig. 3. Reverse-Transcription PCR Results Indicate Zm00001d052233 Primers are Visible at 237 base pairs (bp) for all Annealing Temperatures. Expected size 237 bp. Sizes of the marker (Ogene Ruler: Thermo Scientific) are noted at the left of the image. Observed expected PCR fragment is at right. Annealing temperatures for each lane labeled above the image.

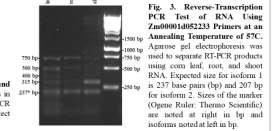


Fig. 3. Reverse-Transcription PCR Test of RNA Using Zm00001d052233 Primers at an Annealing Temperature of 57C. Agarose gel electrophoresis was used to separate RT-PCR products using corn leaf, root, and shoot RNA. Expected size for isoform 1 is 237 base pairs (bp) and 207 bp for isoform 2. Sizes of the marker (Ogene Ruler: Thermo Scientific) are noted at right in bp and isoforms noted at left in bp.

Results

Table 3. Isoform Comparison Table

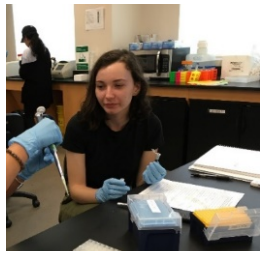
RT-PCR size (bp)	Transcript	Provides evidence for revision to gene model?	Expected and/or Observed?	Domains likely missing	Likely functional?
237	Isoform 1	no	E, O	no	yes
207	Isoform 2	no	E	yes	no
315	Isoform 3	yes	O	unknown	unknown
400	Isoform 4	yes	O	unknown	unknown
500	Isoform 5	yes	O	unknown	unknown
750	Isoform 6	yes	O	unknown	unknown

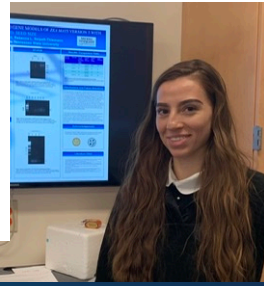
Conclusions and Future Directions

•Leaf RNA extraction and reverse transcription was unsuccessful as bands were not present on gel electrophoresis. Jori's leaf RNA was used for the duration of the research.
•Zm00001d052233 primers were visible at all annealing temperatures – 57C was used.
•Reverse transcription on corn leaf, root, and shoot tissues using Zm00001d052233 primers was successful.
•There were two expected isoforms with only one observed after RT-PCR. Additionally, four unexpected isoforms were observed. The expected and observed isoform would likely have no domains missing and therefore be functional. However, the isoform which was not observed would be nonfunctional due to missing domains and alternative splicing. The nonfunctional proteins include those in the histidine phosphatase superfamily and phosphoglycerate mutases. Proteins in the histidine phosphatase family are involved in metabolic regulations, intermediary metabolism and developmental processes. Moreover, phosphoglycerate mutase is an important enzyme in the pathway of glycolysis. No functionality in each of these proteins could influence nitrogen use efficiency in *Zea mays*. These results indicate that the original gene model is not entirely correct and should be revised.
•Further research is needed to determine whether the observed isoforms which were not expected would have missing domains or be functional.

References

•Hirel B, Tetu T, Lea PJ, Dubois F. 2011. Improving nitrogen use efficiency in crops for sustainable agriculture. Sustainability [Internet]. [cited 2019 August 30]; 3(9):1452-1485. Available from: https://doaj.org/article/eee7732a213a4e1c9b4bc03eaad953c4 doi: 10.3390/su3091452
•Kumar A, Pandeya A, Malik G, Sharma M, P HK, S AK, Gahlaut V, Gajula MNVP, Singh KP, Suravajhala P, et al. 2018. A web resource for nutrient use efficiency-related genes, quantitative trait loci and microRNAs in important cereals and model plants. F1000Res [Internet]. [cited 2019 August 30]; 7:1-12. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6073097/ doi: 10.12688/f1000research.14561.1
•Ladha JK, Tirol-Padre A, Reddy CK, Cassman KG, Verma S, Powlson DS, Kessel CV, Richter Zuluaga DL, Sonnate G. 2019. The use of nitrogen and its regulation in cereals: structural genes, transcription factors, and the role of miRNAs. Plants [Internet]. [cited 2019 September 11]; 8(8):1-15. Available from: https://www.mdpi.com/2223-7747/8/8/294/htm doi:10.3390/plants8080294
•Zuluaga DL, Sonnate G. 2019. The use of nitrogen and its regulation in cereals: structural genes, transcription factors, and the role of miRNAs. Plants [Internet]. [cited 2019 September 11]; 8(8):1-15. Available from: https://www.mdpi.com/2223-7747/8/8/294/htm doi:10.3390/plants8080294
•Rozen S, Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. Hum Press. 365-386.





MIDDLE TENNESSEE STATE UNIVERSITY

CUREs

Rebecca Seipelt

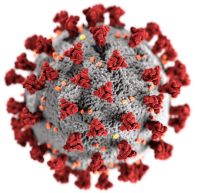
Outcomes: The Stats (F19)

- 207 genes examined
- 9+ genes evaluated in high detail
- 18 regions tested within 9 genes

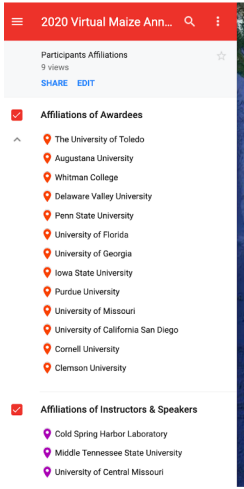
Medium term	Student Quote
Increased self-efficacy	"Personally, I learned more about **my strengths and weaknesses**, time management, flexibility, accommodation, trial and error and the **importance of learning from mistakes.**"
Increased motivation in science	"There is so much that goes on in the background of science that does not get appreciated and I found **a new passion for science** through this experience."
External validation from scientific community	"Although many sleepless nights were spent on this research, **the realness of it all was the most rewarding aspect**. I felt like **what I was doing in lab actually mattered beyond a simple grade**."
Increased tolerance for obstacles	**"sometimes results don't come in the first try. But continuing to try is what is important, you can't just give up** if something doesn't turn out the way you had hoped."

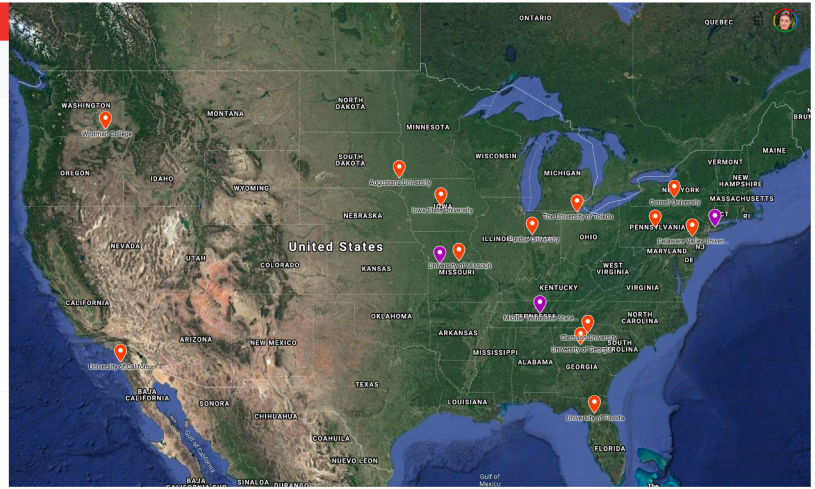
Long term	Student Quote
Enhanced science identity	"Not only did this research give me new skills and knowledge, but it helped me boost my confidence of being a science major. It showed me that **I have what it takes to be a female student in the STEM field, and that I can do much more**."
Career clarification	"This project personally has meant a great deal to me, because it was the first time I was actually able to experience real lab conditions. It **reassured me that I was in the right field of study**, and I whole-heartedly enjoyed every minute of the research (minus deadline stress)."
Persistence in science	"this lab experience because it has given me **the confidence to do other research** at MTSU, which I was unsure if I wanted to do previously."

2020 Virtual Jamboree



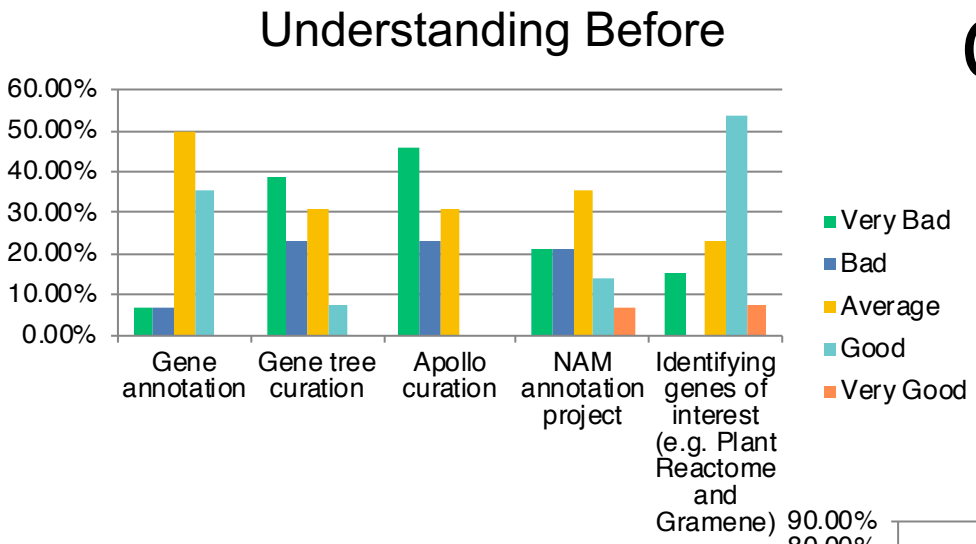
- 20 awardees (28 participants) from 17 institutions across US
- 5 teaching faculty (3 @ PUIs), 11 graduate students, 1 postdoc, 2 staff scientists
- 4 x 3-hour sessions (March 10-20)
- About 700 genes evaluated with the Gramene Gene Tree Visualizer and 42 genes curated in Apollo

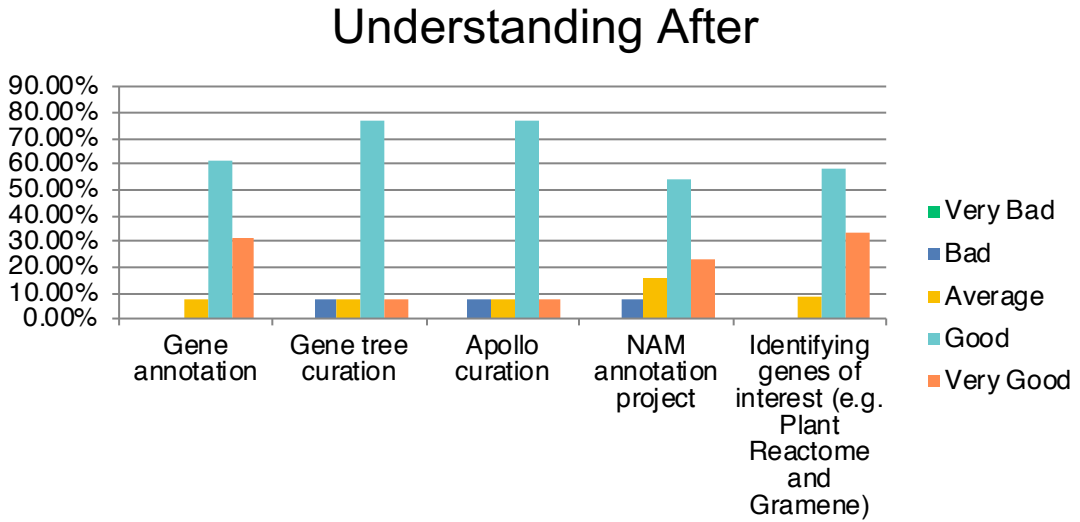




Cornell CURE
Spring 2020

Kevin Ahern





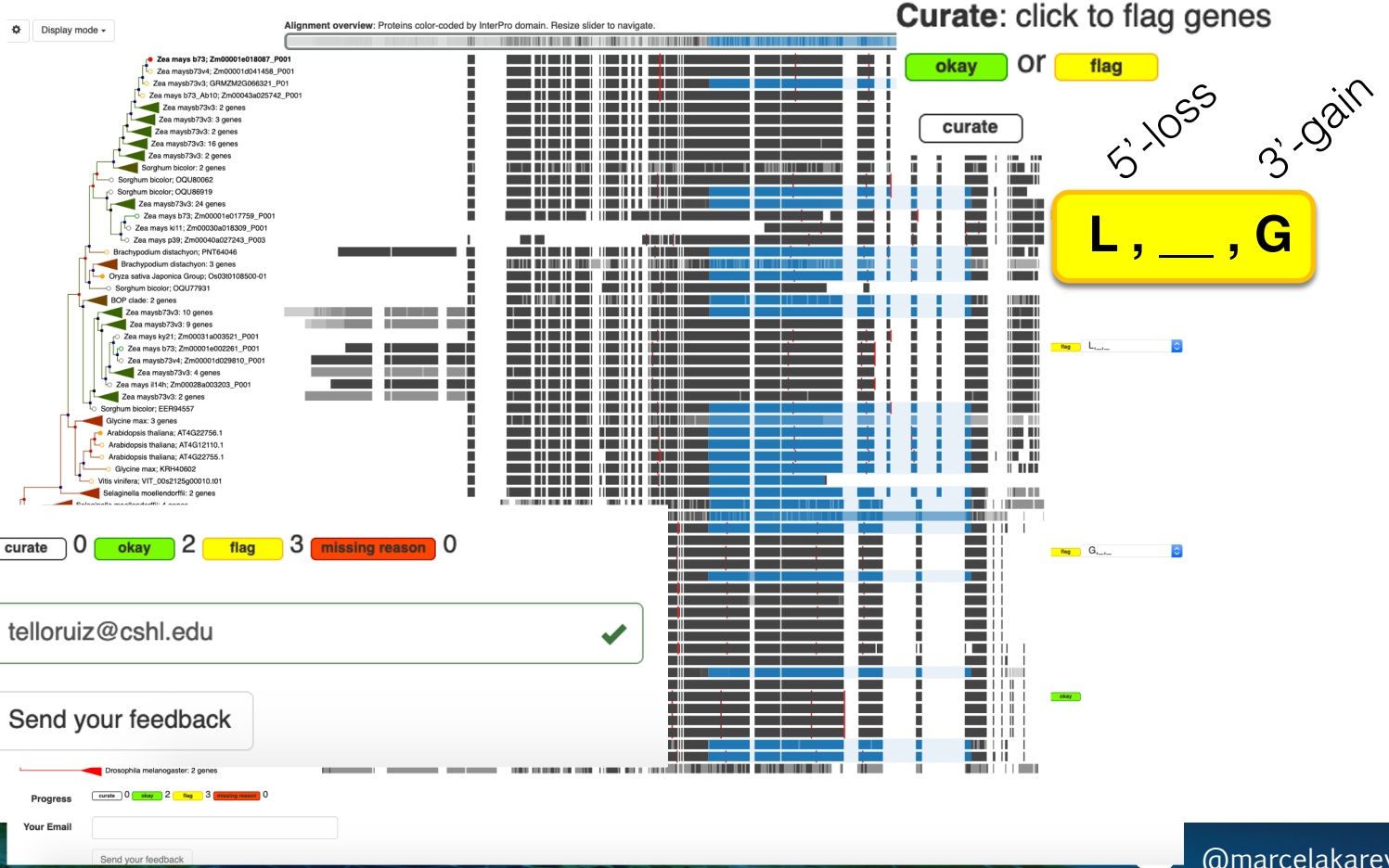
Plant Genetics
Annotation Lab (2-3 hours)
55 students

Teaching Opportunities

- Example tests:
  - Refined algorithm to select a canonical transcript
  - Previously curated V4 models improved in V5
  - Compared *ab initio* gene predictions (benchmark: 44 genes on chr1)
- Example errors:
  - Internal sequence gain/loss
  - Missing START Met
  - Multiple internal STOP codons

http://curate.gramene.org/nam?gene=Zm00001e018087

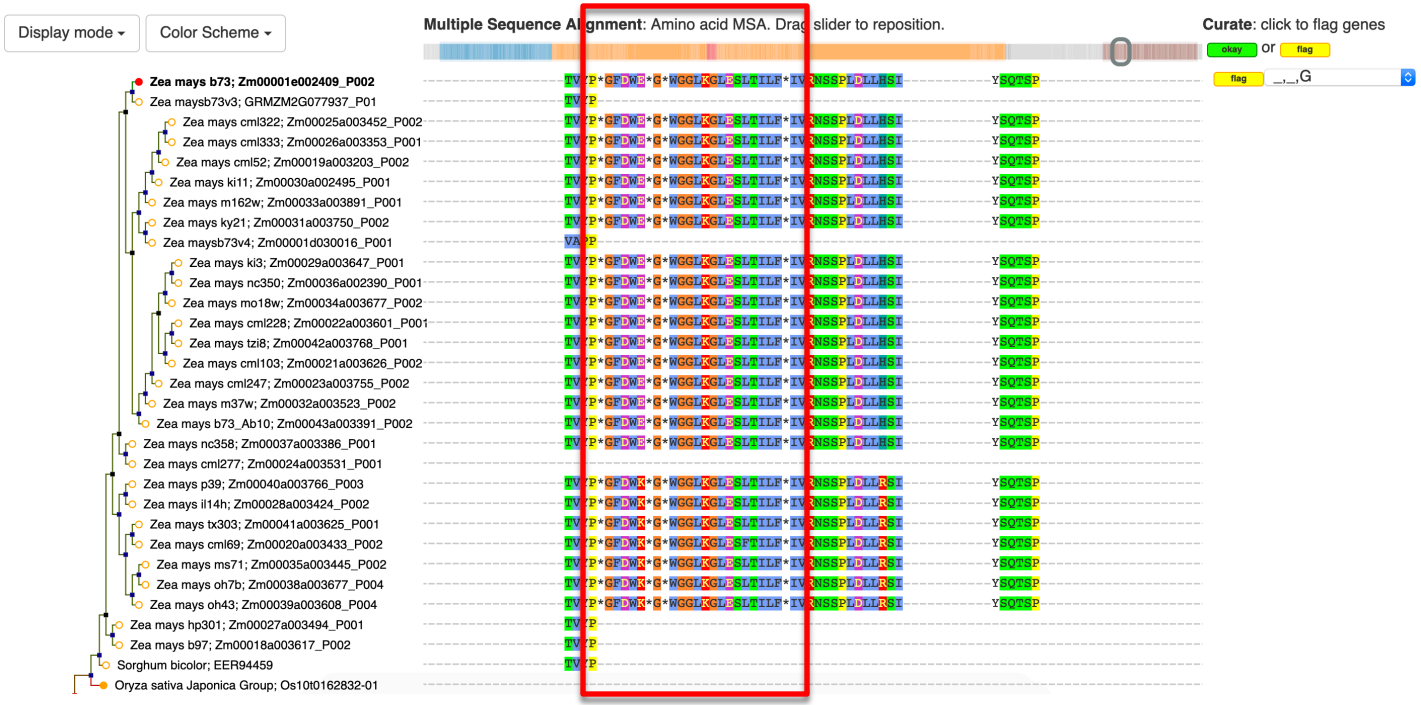
Gramene Tree Visualizer



Zm00001e017759 - Small internal deletion?

NOT a deletion → RNA evidence for alternative transcript in Apollo

Zm00001e002409 – Multiple STOP codons



Zm00001e002261 – Missing Starting Met

NAM genomes

@marcelakarey

CUREs derived from the 2020 VMAJ

- 2020 Spring CUREs:
  - Middle Tennessee State University – Rebecca Seipelt
  - University of Central Missouri – Selene Nikaido
  - Cornell University – Kevin Ahern
  - University of Toledo –Ankita Abnave & John Gray

- 2020 Fall CUREs under development at Whitman College and College of Southern Nevada

Interested in Joining?

- PUI faculty network
- Biweekly meetings, slack channel, office hours
- Educational resources: lesson plans, videotutorials, slides
- Resources to curate Maize and Sorghum genes
- Contact us: telloruiz@cshl.edu

THANKS!

Cristina Fernández-Marco
Kapeel Chougule
Andrew Olson
David Micklos
Doreen Ware

CUREs Faculty & Students

Rebecca Siepelt (MTSU)
Christos Noutsos (SUNY OW)
Selene Mikaido (UCM)
Kevin Ahern (Cornell U)
Ankita Abnave/John Gray (U Toledo)

The project is supported by NSF IOS-1127112, MCB-1744001, IOS-1445025, and USDA ARS 8062-21000-041

Outreach Funds