Finding and fixing annotation errors through community curation

Marcela K. Tello-Ruiz[1], Cristina F. Marco[2], Kapeel Chougule[1], Andrew C. Olson[1], Rebecca Seipelt-Thiemann[3], David A. Micklos[2], Doreen Ware[1,4]

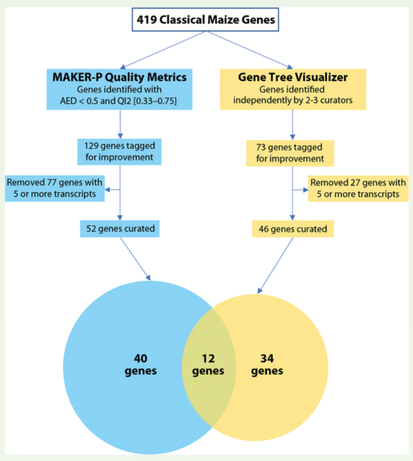
[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, [2]CSHL DNA Learning Center, Cold Spring Harbor, NY, [3]Middle Tennessee State University, Murfreesboro, TN, [4]USDA ARS, NEA Plant, Soil & Nutrition Laboratory Research Unit, Ithaca, NY

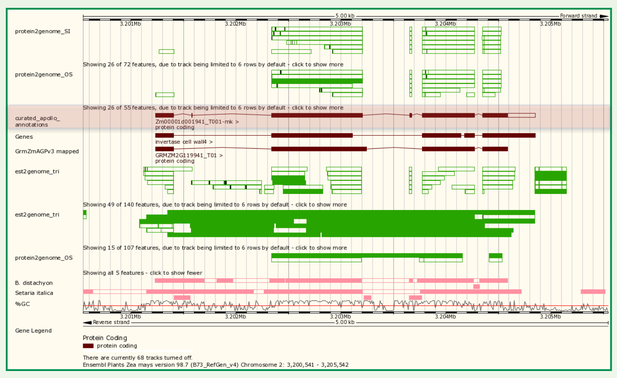
DOUBLE TRIAGE OF MAIZE GENES

In a pilot study (Tello-Ruiz et al, 2019), we evaluated the correctness of the structural annotation of B73 *Zea mays* RefGen_V4 classical gene models using two approaches: 1) MAKER-P quality metrics, and 2) Gramene gene tree visualizer.

MAKER-P generates quality metrics that assess how well a transcript model is supported by available biological evidence. We used two of these metrics: Annotation Edit Distance (AED) and Quality Index 2 (QI2), to identify low-quality gene models. To ensure the availability of evidence to use in manual curation, we flagged genes with AED scores less than 0.5 and QI2 values between 0.33 and 0.75.

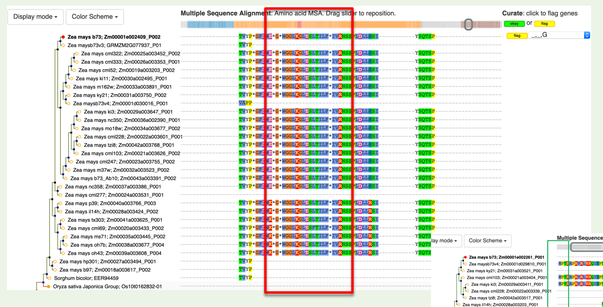
The Gramene tree visualizer provides an interactive interface to inspect protein sequence alignments for a given gene family and identify genes with potential annotation errors, further evaluated in the Apollo Genome Editor for improvement.





Left. **Flowchart of the double triage of classical maize genes and comparative number of genes flagged for curation by quality metrics and gene trees**. Parallel methods–MAKER-P quality metrics (blue) and the gene tree visualizer (yellow)–produced different but overlapping sets of genes for manual curation in the Apollo gene editor. Genes with five of more transcripts were excluded. *Right*. **The curated gene models are available in the Gramene genome browser**. Click on the genomic coordinates of the table at http://www.gramene.org/curated_maize_v4_gene_models to visualize a gene in its genomic context, as shown in the figure.

EXAMPLE ANNOTATION ERRORS

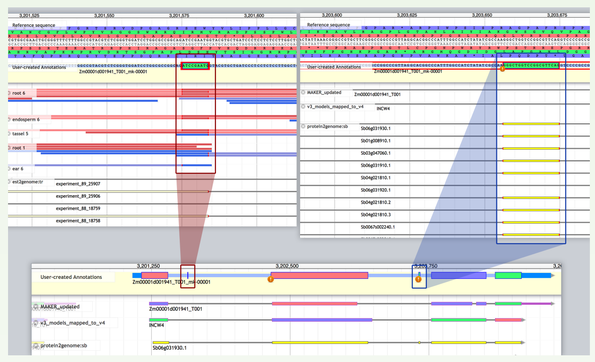


Left. **The Zm00001e002409 gene model has multiple STOP (*) codons in its 5'-end**.

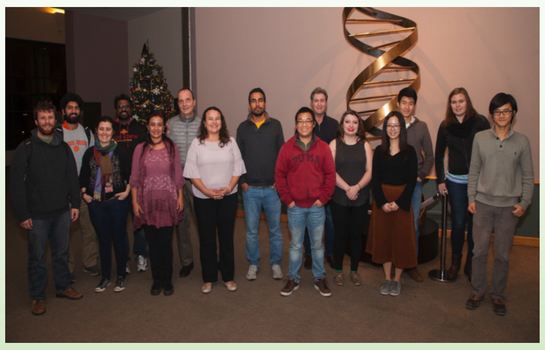
Right. **The Zm00001e002261 gene model is missing a starting Met (M).** Unlike the one upstream of the first amino acid (R) in its protein sequence, the one downstream is conserved in its sorghum, rice and brachypodium orthologs.

Left. **The Zm00001e017759 gene model has an apparent small internal deletion**. Upon closer examination of the mRNA evidence in Apollo, it was concluded it represents an alternatively spliced transcript.

Right. **Apollo curation for the Zm00001d001941 gene model.** Figure taken from Tello-Ruiz *et al* (2019). Experimental evidence supports the existence of two mini-exons not previously identified in the existing V3 and V4 maize models. A conserved 9-nucleotide exon (red shade) and a novel 19-nucleotide exon (blue shade) are supported by protein sequences from sorghum and rice, assembled EST transcripts from ultra-deep sequencing, long Iso-Seq reads combined from six tissues, and/or RNA-seq from root and other tissues.



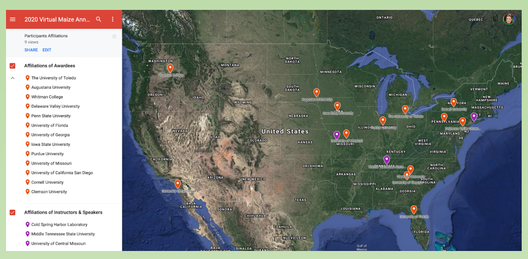
COMMUNITY ANNOTATION TRAINING



Graduate Students - CSHL Plant Genomes 2017

Maize Researchers & Faculty - MGM 2019

Primarily Undergraduate Institutions Faculty - PAG 2019



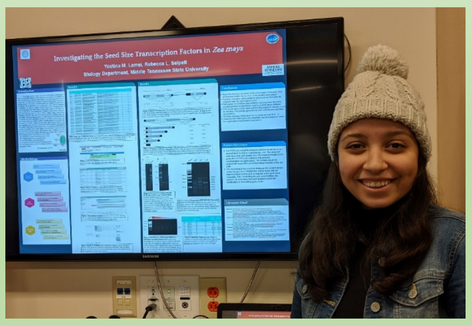
Maize Researchers & Faculty - MGM 2020

CSHL Newsletter: https://www.cshl.edu/teachers-make-genomes-more-useful-from-home/

COURSE-BASED UNDERGRADUATE EXPERIENCES

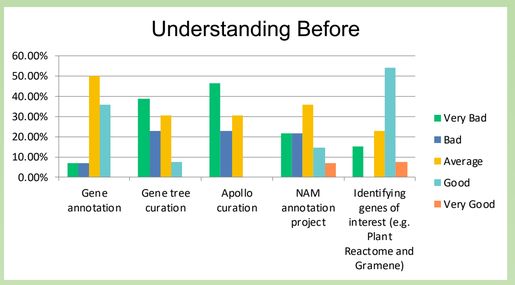
This project has given students great opportunities to increase their knowledge of genome annotation and be part of scientific research working with the same data, tools, and in real time with maize researchers.

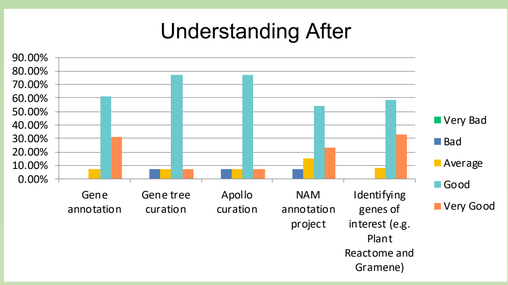
Medium term	Student Quote
Increased self-efficacy	"Personally, I learned more about **my strengths and weaknesses**, time management, flexibility, accommodation, trial and error and the **importance of learning from mistakes.**"
Increased motivation in science	"There is so much that goes on in the background 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."

Rebecca Seipelt, Honors Genetics Class, MTSU (Fall 2019): 18 students (17 female, 1 male)





Kevin Ahern (PhD candidate), Plant Genetics Annotation Lab (Spring 2020), Cornell University (55 students)

Contact information:
marco@cshl.edu
telloruiz@cshl.edu

The project is supported by NSF IOS-1127112, IOS-1445025, PGR-1744001 and USDA