Webinar: Systems biology resources for the insect vector of the citrus greening disease

Surya Saha

Boyce Thompson Institute, Ithaca, NY

ss2489@cornell.edu @SahaSurya

March 5th, 2018

- Please silence your microphones (see bottom left of zoom window)
- We will have a Q & A session at the end but please feel free to ask during the presentation
- Slides: http://bit.ly/ACPv2slides
- Feedback survey: http://bit.ly/ACPv2survey

Acknowledgements

Kansas State University

Sue Brown

Cornell University/BTI

Michelle (Cilia) Heck

USDA/ARS

Wayne Hunter

Robert Shatters

University of California, Davis

Carolyn Slupsky

Indian River State College

Tom D'elia

Citrus Greening Solutions
A USDA NIFA Project

Funding provided by
USDA
United States Department of Agriculture
National Institute of Food and Agriculture

Mueller Lab

Mirella Flores

Prashant Hosmani

Stephanie Hoyt

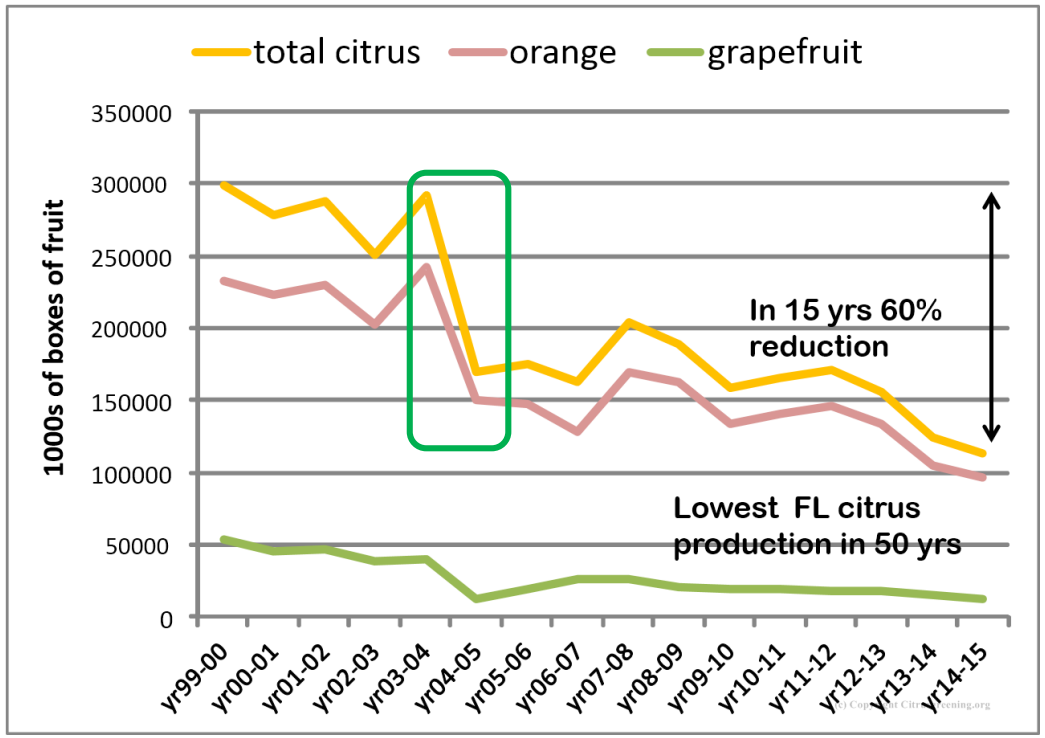
Project Partners

Kansas State University
BTI Boyce Thompson Institute
USDA ARS United States Department of Agriculture Agricultural Research Service
ugr Universidad de Granada
IRSC
UC Davis Food Science and Technology
Los Alamos National Laboratory
Genome Sciences University of Washington
UF IFAS
Center for Public Issues Education

Citrus Greening: Huanglongbing

- Most significant disease of citrus worldwide
- More than $4.5 billion in lost citrus production and more than 8,200 lost jobs (2006/07 to 2010/11)
- Associated with gram negative bacterium *Candidatus* Liberibacter asiaticus (CLas)
- Spread by insect vector, *Diaphorina citri* (Asian citrus psyllid, ACP)

Annie Kruse



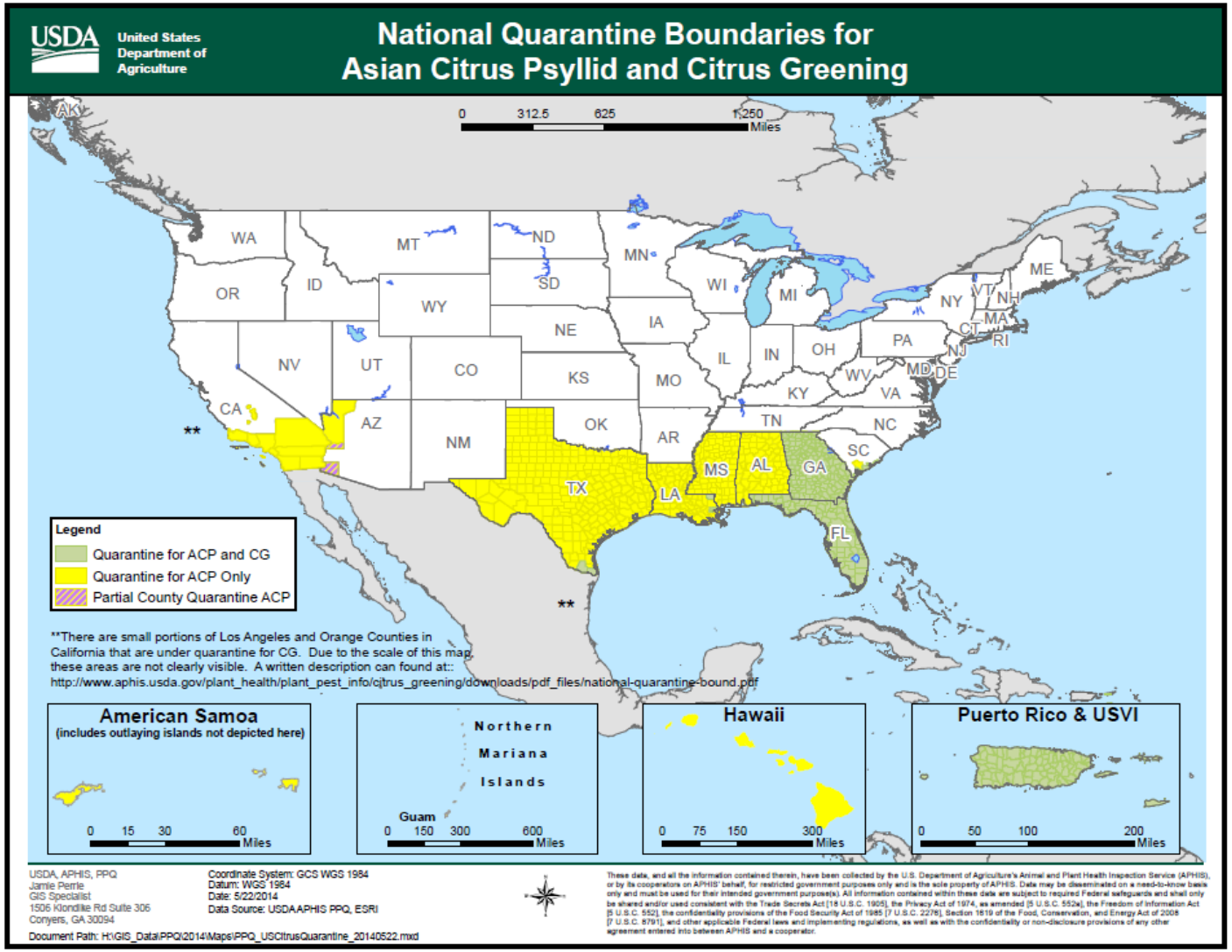
2017

Fresh Citrus at Risk From Devastating Disease

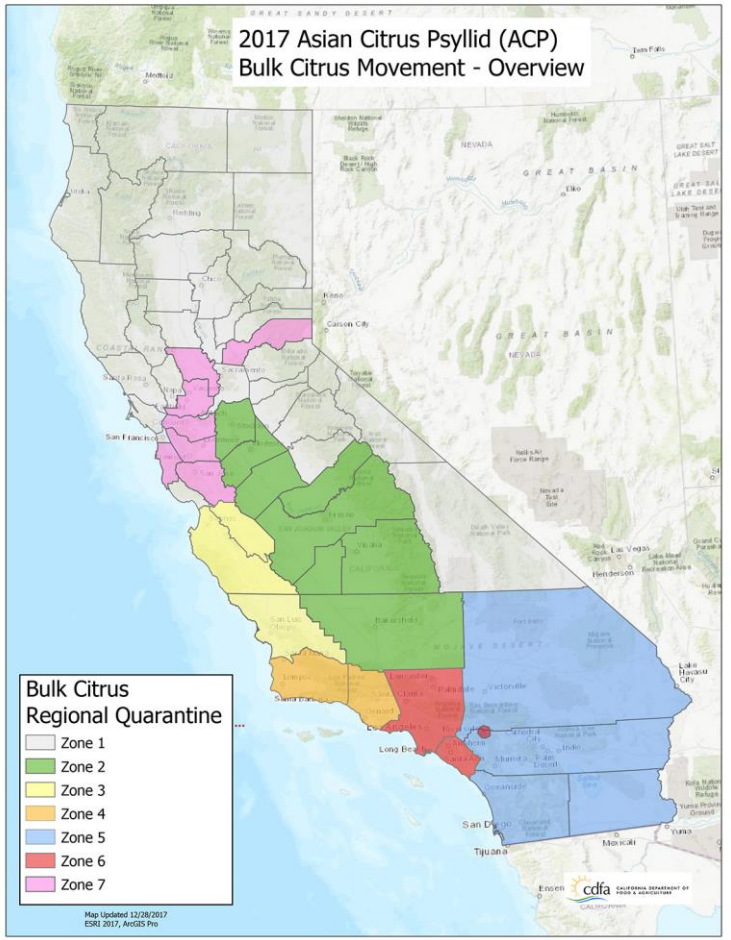
HLB Causes Florida to Fall Behind California in Citrus Production

By Diane Nelson on December 11, 2017 in Food & Agriculture

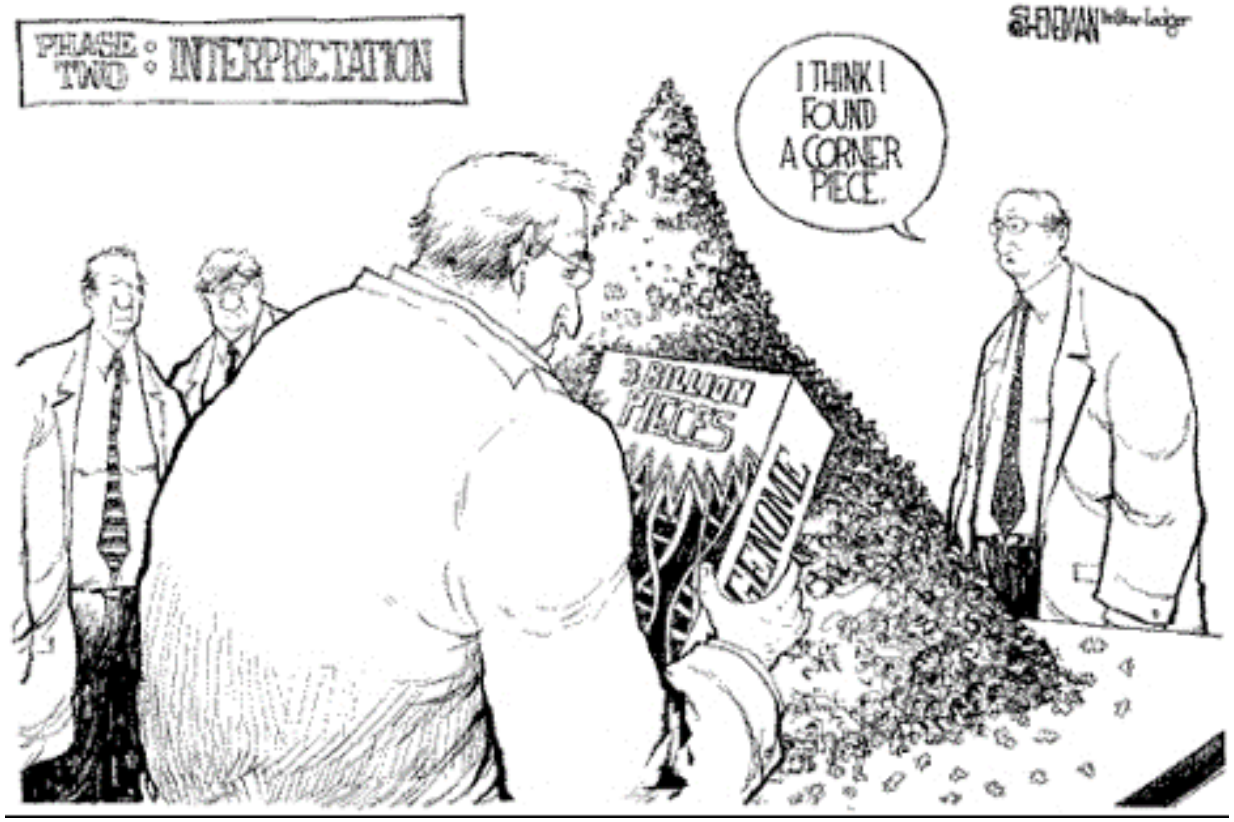
Citrus Greening: Huanglongbing



Asian Citrus Psyllid quarantine in California (Jan 2018)



Current Illumina assembly



http://biobeans.blogspot.com/2012/11/bioinformatics-genome-assembly.html

Genome	Diaci1.1
Contigs	161,988
Total Length	485 Mb
Longest	1 Mb
Shortest	201bp
Ns	19.3 Mb

Scaffold N50: 109,898 bp
Contig N50: 34,407bp

Highly fragmented

Many examples of misassemblies!!

DATABASE
The Journal of Biological Databases and Curation

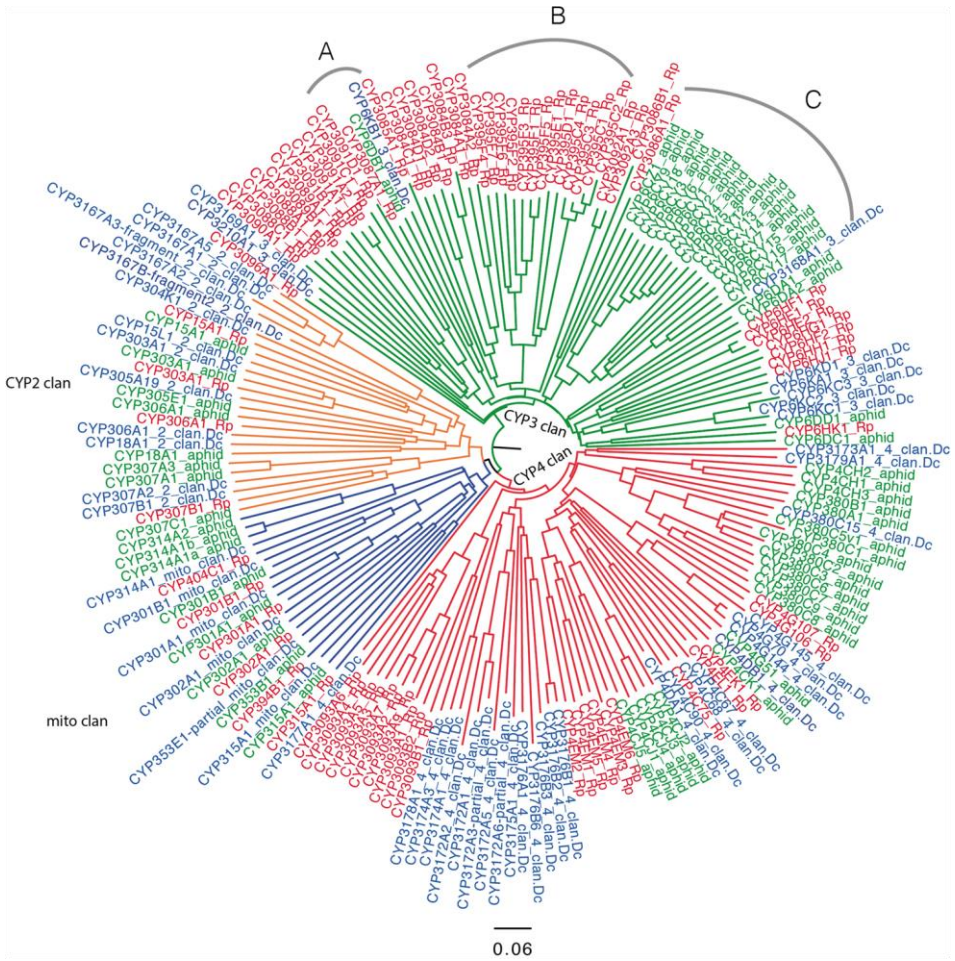
Database, 2017, 1–20
doi: 10.1093/database/bax032
Original article

Original article

Improved annotation of the insect vector of citrus greening disease: biocuration by a diverse genomics community

- Genome Diaci v1.1
- Official gene set v1.0
  - Immune pathway
  - RNAi pathway
  - P450 gene family

- 530 manually curated models
- ~20,000 NCBI predicted models
- MCOT transcriptome v1.1



R. prolixus (red), A. pisum (green) and D. citri (blue)
Four clans of P450s, CYP2 (orange), CYP3 (green), CYP4 (red) and mito (blue) clan are shown in the phylogenetic tree.

Annotation by undergraduate students

- 18 students involved
- >250 gene models
- >30 gene families
- 13 gene reports for publication

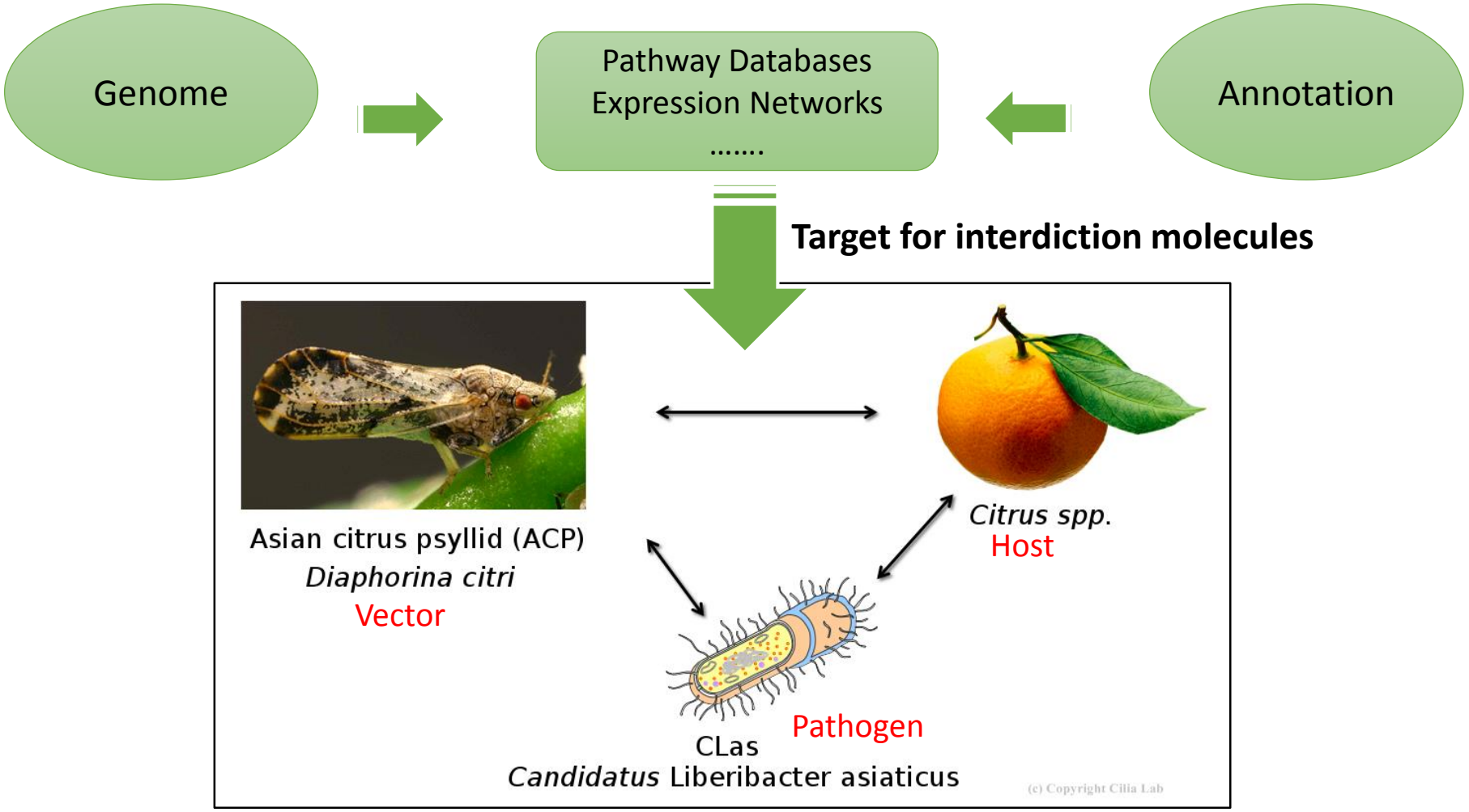
Weekly IRSC Annotation Meetings

Join our community curation project!

Annotation with Web Apollo

Slide: Tom D'elia

Omics resources and databases are required for identification of targets for interdiction

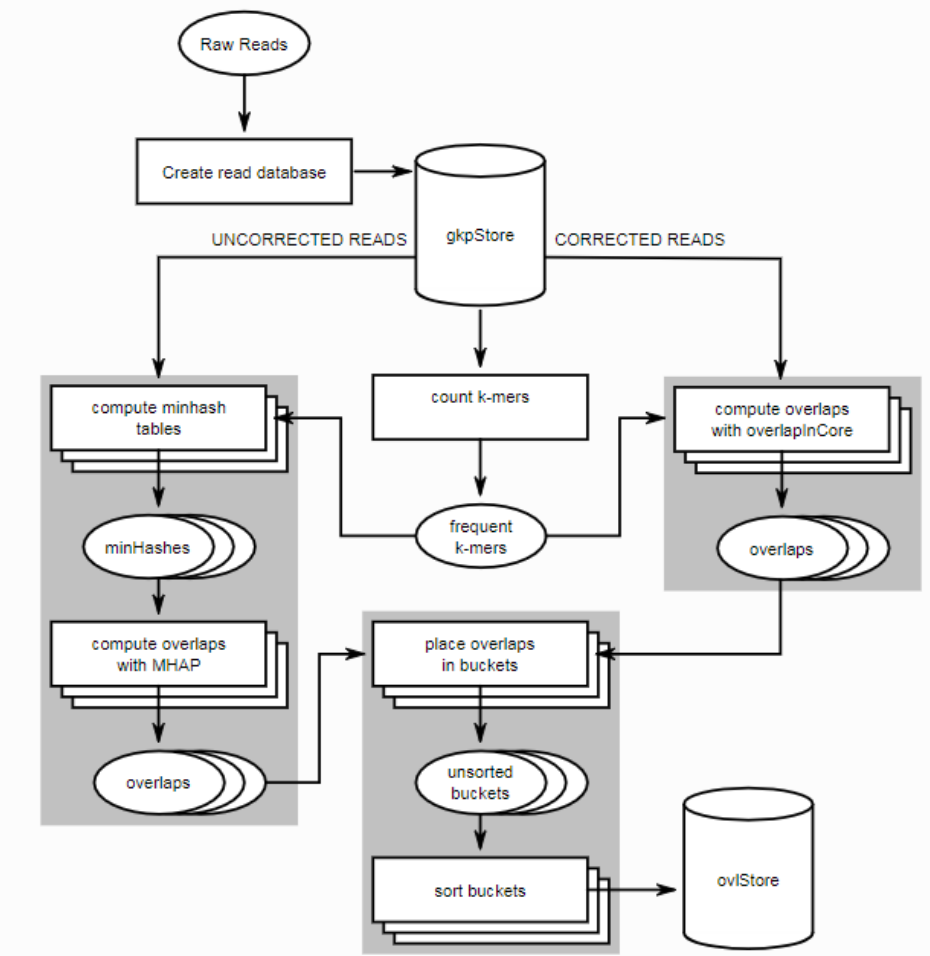


Pacbio assembly

	Error rate 0.013	Error rate 0.015
Number of contigs	7,832	8,030
Total bases	462.8 Mb	493.1 Mb
Longest	1.6 Mb	1.7 Mb
Shortest	4.4 Kbp	5 Kbp
Average length	59.9 Kb	61.4 Kb
Contig N50	85.8 Kb	92.6 Kb

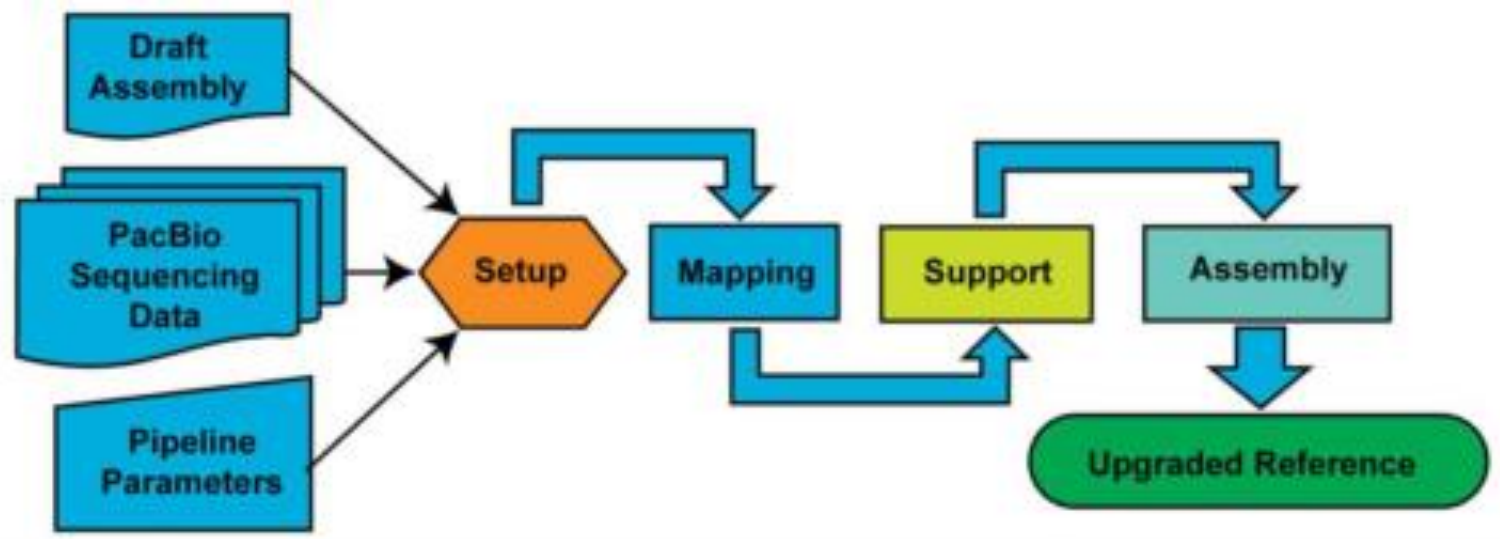
70-80X coverage with 41 SMRT cells
Contiguous assembly with longer contigs
Multiple individuals in DNA sample

Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation



http://canu.readthedocs.io/en/stable/

PBJelly scaffolding



5,290 gap extensions
535 gaps filled
Number of Ns: 0 bp

	Canu assembly	Scaffolded Assembly v1.9
Number of contigs	7,832	8,352
Total bases	462.8 Mb	591.7 Mb
Longest	1.6 Mb	2 Mb
Shortest	4.4 Kb	1.5 Kb
Average length	59 Kb	70.8 Kb
Contig N50	85.8 Kb	115.8 Kb

Chicago scaffolding

	v1.91	v2.0 REFERENCE	v2.0 ALTERNATE
Number of contigs	3,681	1,906	1,751
Total bases	596 Mb	498 Mb	79.1 Mb
Longest	4.2 Mb	4.2 Mb	760.6 Kb
Shortest	1.5 Kb	6 Kb	1.5 Kb
Average length	162 Kb	261.7 Kb	45.2 Kb
Contig N50	620 Kb	749 Kb	75.1 Kb
Ns	5.1 Mb	4.5 Mb	467 Kb

Pilon

Error correction
- DNA sequencing data
- RNA sequencing data

https://github.com/broadinstitute/pilon/wiki

Redundans

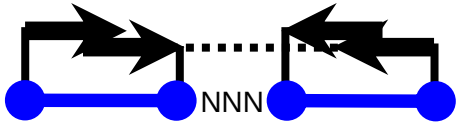
- Duplication removal
- ALT scaffolds

https://github.com/Gabaldonlab/redundans

500ng input DNA from single male psyllid
Duplicated contigs added to alternate assembly

Evaluating the genome assembly

Paired-end RNAseq alignment



	Overall alignment rate	Concordant alignment rate
Diaci 1.1	82%	63%
Diaci 2.0	88%	74%

Average length of aligned coding sequence

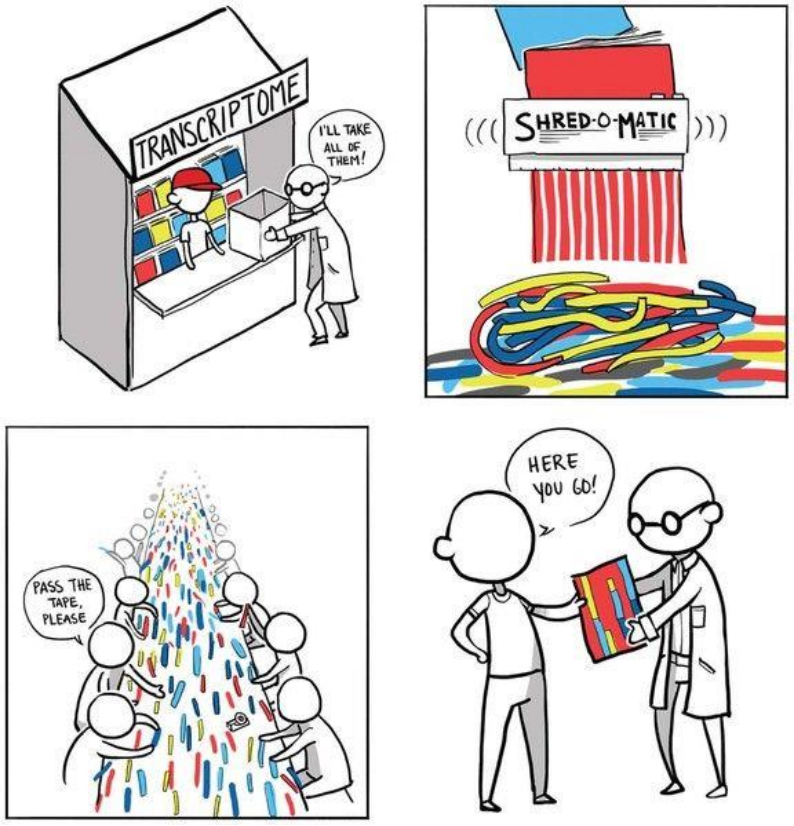
	MCOT	Isoseq (full-length transcripts)
Diaci 1.1	1054 bp	470 bp
Diaci 2.0	1321 bp	699 bp

Benchmarking sets of Universal Single-Copy Orthologs based on a set of 3350 single-copy orthologs from hemipteran species

BUSCO v3

	Complete	Fragmented	Missing
Diaci 1.1	74.8%	0.3%	24.9%
Diaci 2.0	85.2%	0.1%	14.7%

Gene isoform sequencing (Iso-Seq)



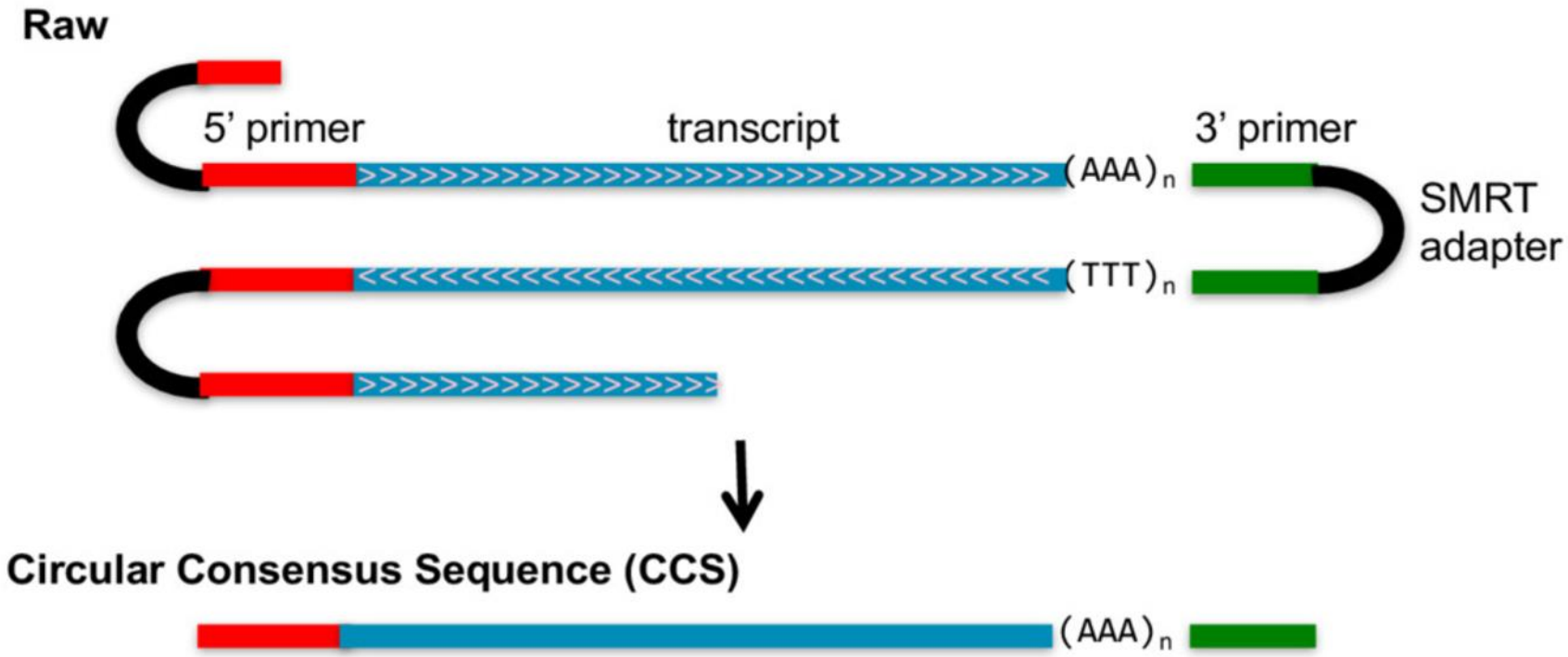
Korf 2013

Accurate gene models are necessary for targeting assays

- Majority of genes are alternatively spliced to produce multiple transcript isoforms.
- Iso-Seq generates full-length cDNA sequences (full-length transcripts and gene isoforms).

Current MCOT (*de novo* and genome-based) transcriptome is useful but fragmented

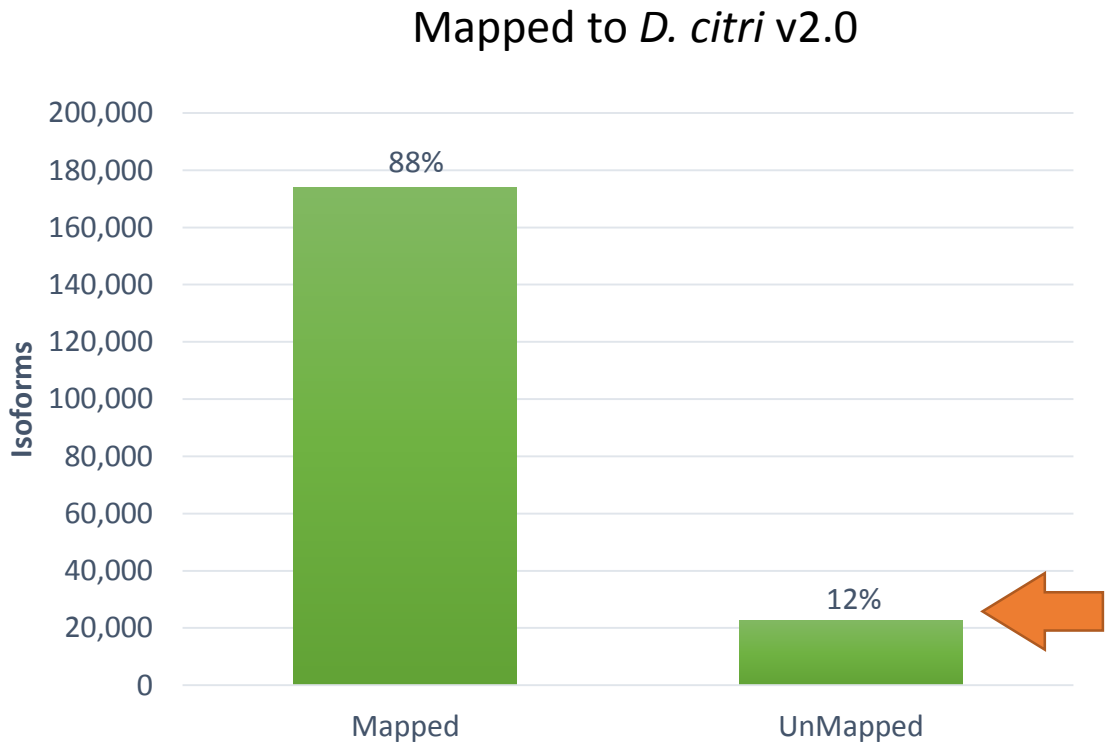
Sequencing full-length gene isoforms



Full-Length = 5' primer + 3' primer + polyA

Iso-Seq transcriptome

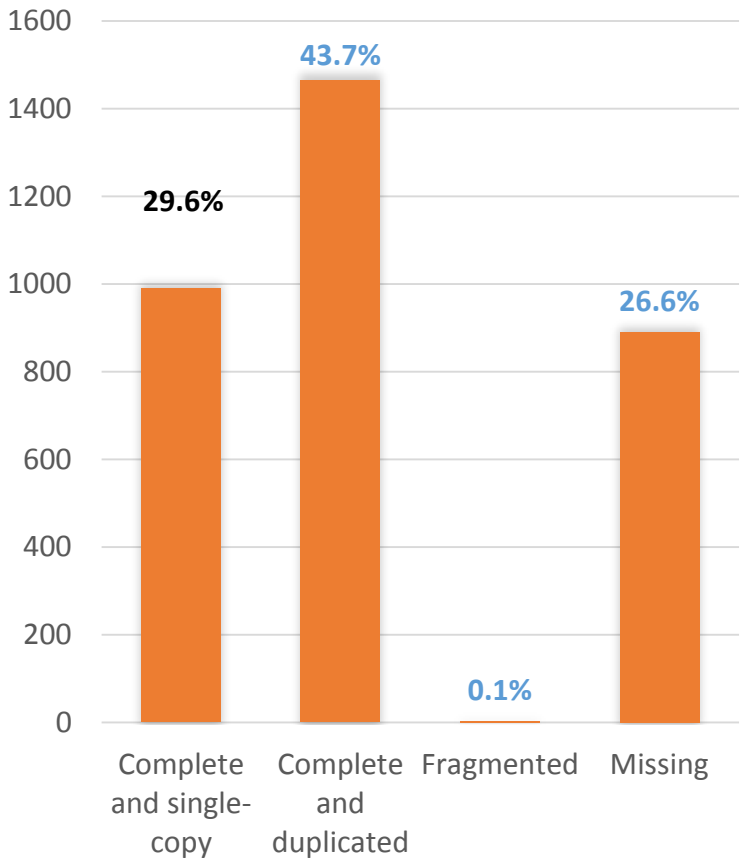
	Counts
Number of genes	14,768 (30,562 in MCOT)
Number of isoforms	52,223
Average number of isoforms/gene	3.53
N50	2.8 Kb
Longest	9.7 Kb



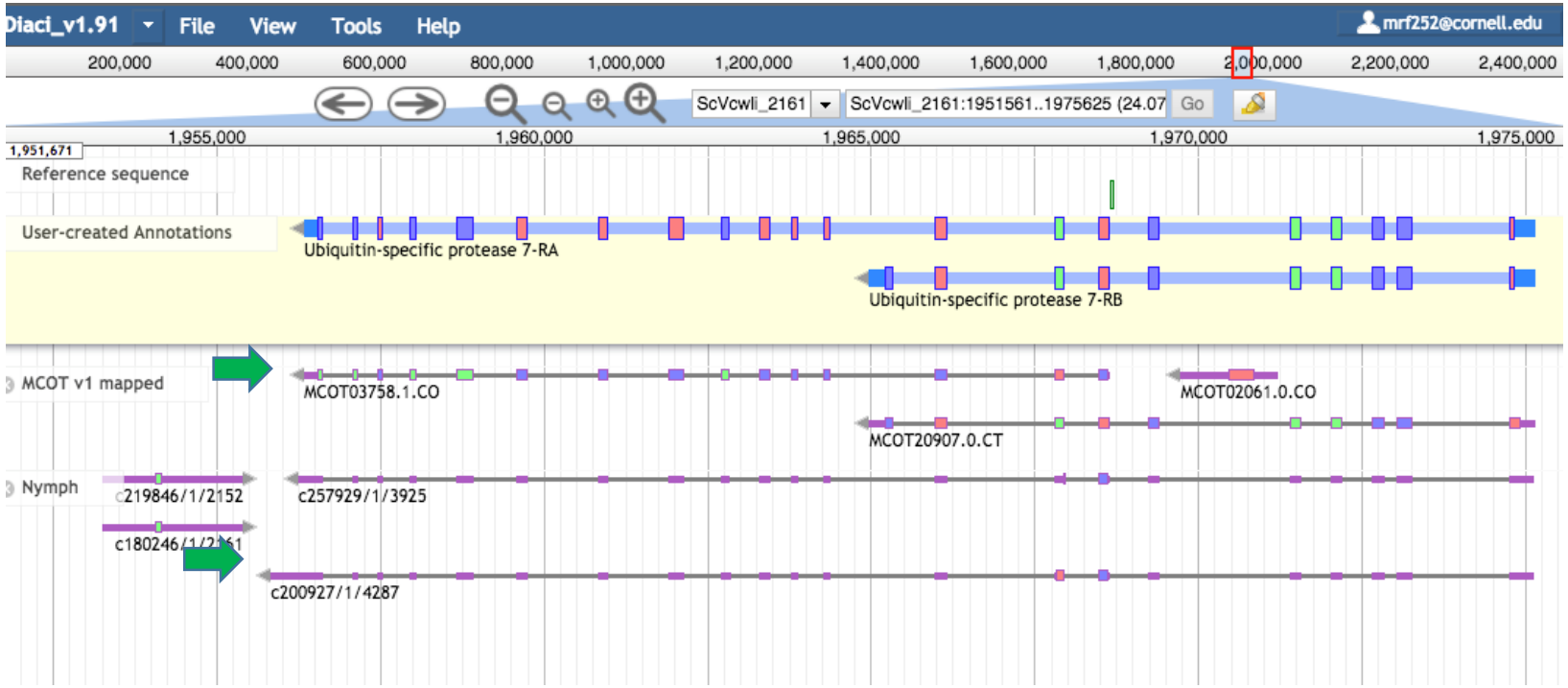
Total isoforms: 196,419

Isoseq provides a comprehensive (*de novo* and genome-based) transcriptome with full-length transcripts and a range of isoforms

Iso-Seq identifies split genes



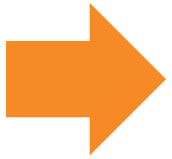
Hemiptera: **3,320** BUSCOs



Official Gene Set v2.0
Structural annotation pipeline

Repeat masking

- RepeatModeler
- Protein masking
- RepeatMasker



mRNAseq analysis

- mRNAseq mapping
- Stringtie genome guided transcriptome
- Pacbio Isoseq mapping
- portcullis
- transDecoder
- mikado

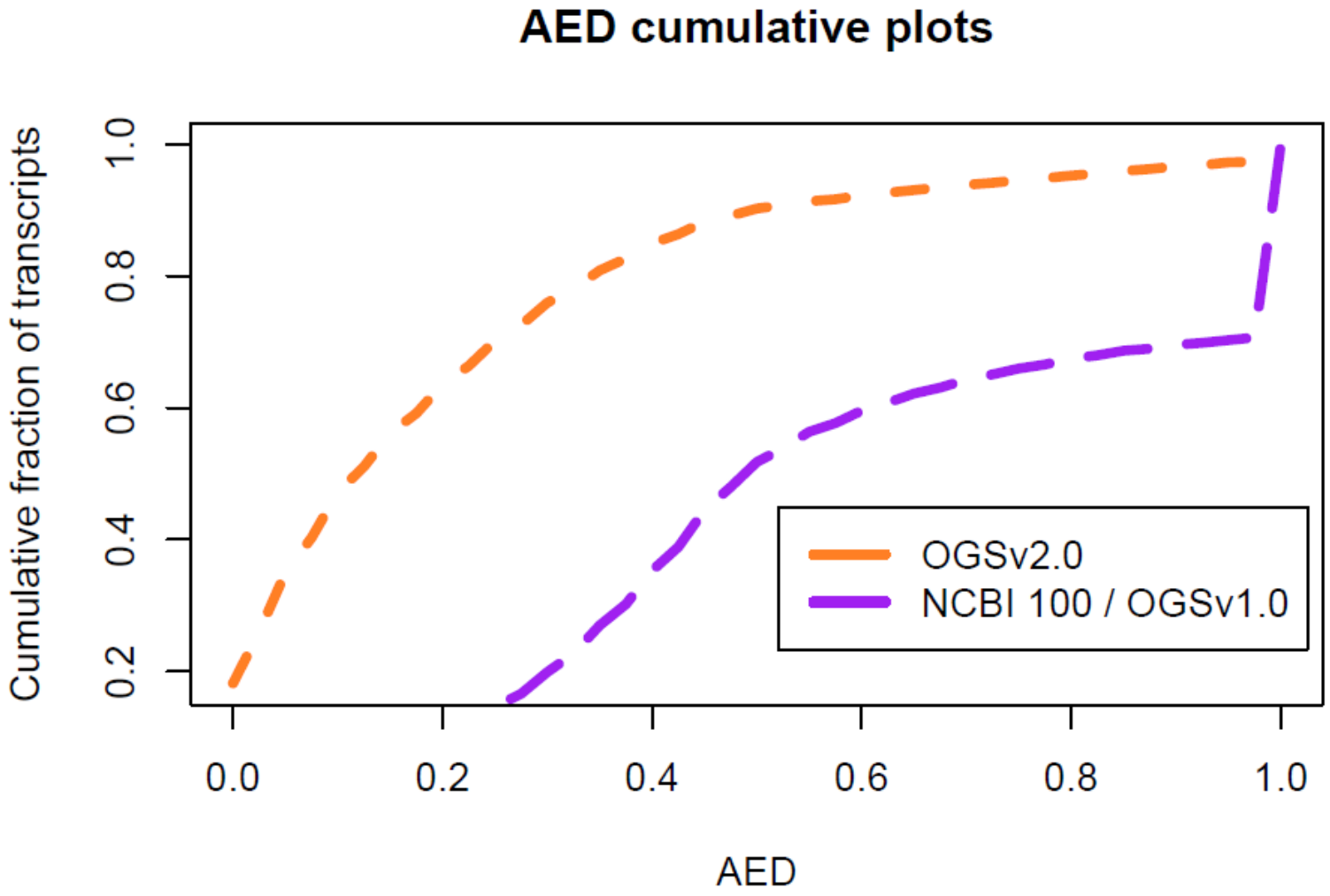
Gene prediction

- SNAP training
- Augustus training with Braker2
- MAKER

RNAseq data sources

- Michelle (Cilia) Heck (USDA)
- Carolyn Slupsky (US Davis)
- Wayne Hunter (USDA)
- Public data from the Citrusgreening community

Evidence based annotation



Annotation Edit Distance

AED= 0 complete support

AED =1 lack of support

AED provides a means to evaluate quality of annotations from RNAseq and ortholog evidence

eAED is the AED edit distance at an exon level and is inferred from RNAseq only

Improvements in OGS v2.0 compared with OGS v1.0 / NCBI 100

	OGS v1.0/NCBI 100	OGS v2.0
Number of genes	20,245	20,793
Avg. gene length	8,936.89	11,780.41
Number of mRNAs	20,996	25,292
Exons per mRNA	5.58	7.06
5’ UTRs	514	15,381
3’ UTRs	422	16,507

Maker quality string in GFF file

Name=DcitrM000195.1.2;_AED=0.10;_eAED=0.10;**_QI=252|1|1|1|1|1|14|2036|737**;
Note=Myotubularin (AHRD V3.11 *** tr|A0A1S4F3M7|A0A1S4F3M7_AEDAE). Similar to MCOT06033.0.CT XP_008468956.1. AED 0.10

1. Length of the 5 UTR
2. Fraction of splice sites confirmed by an transcript/EST alignment
3. Fraction of exons that overlap an transcript/EST alignment
4. Fraction of exons that overlap transcript/EST or Protein alignments
5. Fraction of splice sites confirmed by a gene prediction
6. Fraction of exons that overlap a gene prediction
7. Number of exons in the mRNA
8. Length of the 3 UTR
9. Length of the protein sequence produced by the mRNA

Cantarel, Genome Research 2008. Table 2. Maker quality index summary

Official Gene Set v2.0 naming convention

For each isoform

Protein: Dcitr **P** XXXXX.Y.Z

XXXXX is gene number, Y is version and Z is isoform number

mRNA transcript: Dcitr **M** XXXXX.Y.Z

Coding sequence: Dcitr **C** XXXXX.Y.Z

Future

Non-coding gene will be Dcitr **R** XXXXX.Y.Z

AHRD functional annotation

Manually curated alternative transcript names have -RA, -RB, -RC,……

Atg16 Autophagy-related protein 16-1-RA. (AHRD V3.11 *** Guanine nucleotide-binding protein subunit beta-2-like protein tr|A0A0J7P1M8|A0A0J7P1M8_LASNI). Similar to MCOT13489.0.CO XP_008476251.1. AED 0.44

4,329 / 25,292 are Unknown proteins but have evidence support

DcitrM001235.1.1 AED 0.06

PFAM domains for 6,610 genes

GO terms for 4,080 genes

Quality of functional annotation

Automated Assignment of Human Readable Descriptions (AHRD)

AHRD-Version 3.11 Quality score (***)

DcitrM000265.1.1 Aldehyde dehydrogenase (* * *)

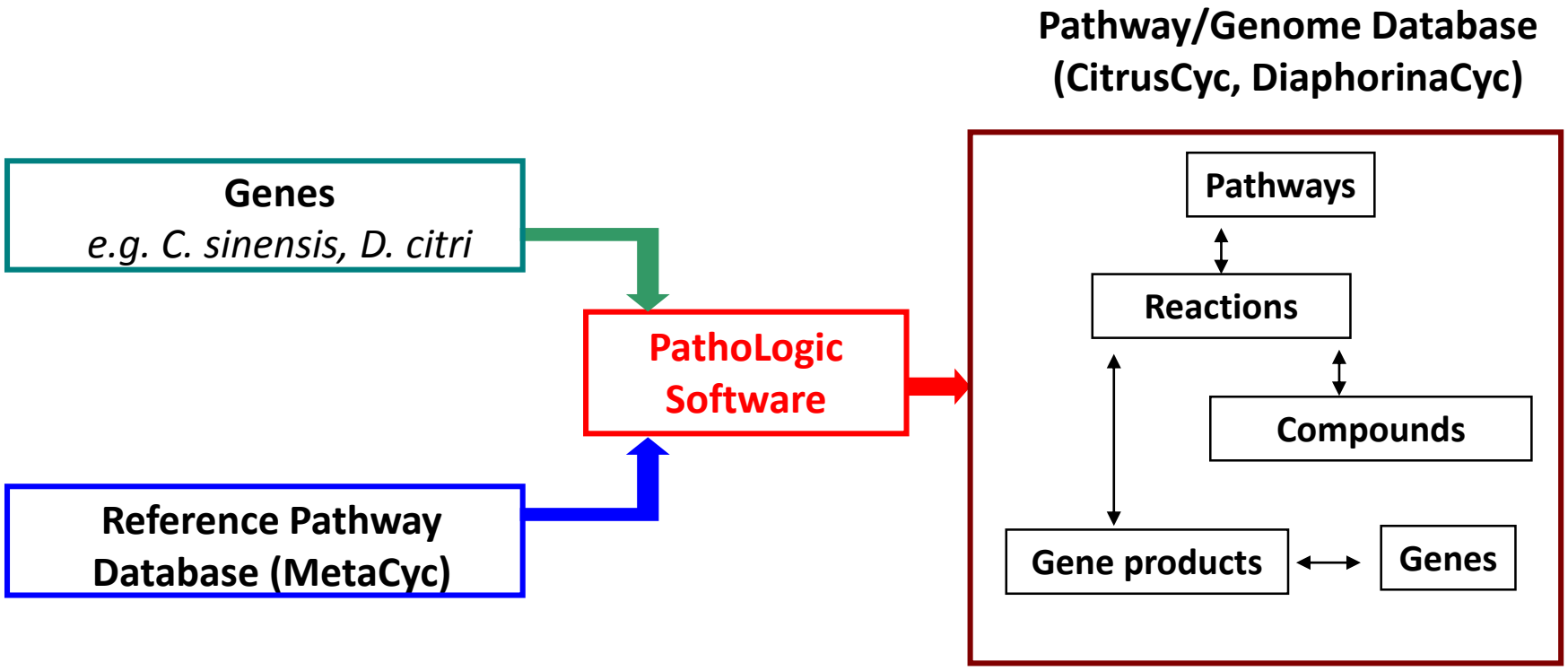
DcitrM000885.1.1 NIF3-like protein 1 (* - *)

DcitrM001075.1.1 Transcription factor TFIIIB component B (- - *)

Position	Criteria
1	Bit score of the blast result is >50 and e-value is <e-10
2	Alignment of the blast result is >60%
3	Human Readable Description score is >0.5

"AHRD's quality-code consists of a three character string, where each character is either **'*'** if the respective criteria is met or **'-'** otherwise."

Metabolic Pathway Database Construction



- Predicts metabolic pathways
- Predicts which genes code for missing enzymes in metabolic pathways
- Infers transport reactions from transporter names

Source: Peter Karp (SRI)

DiaphorinaCyc cellular overview with RNAseq data

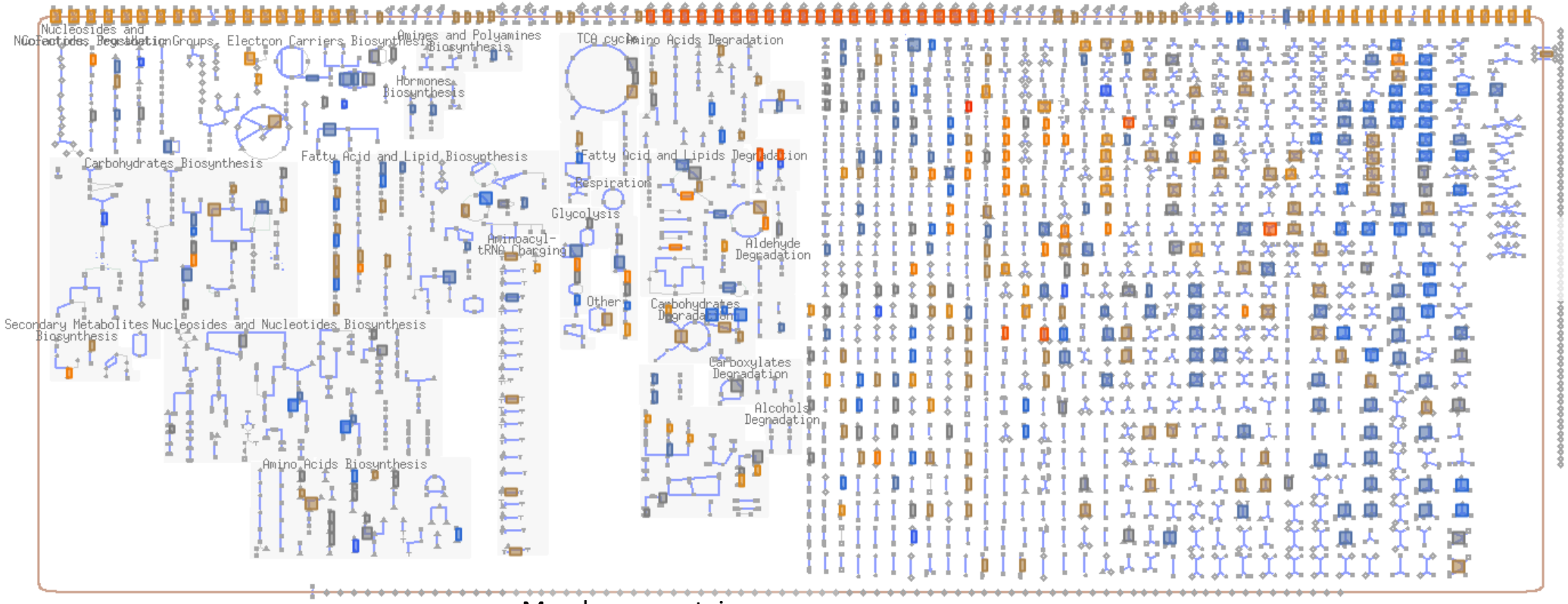
Pathways: 171
Enzymes: 3,507 (was 2,857)

Transport Reactions: 17
Proteins: 25,295 (was 12,548)

Transporters: 87
Compounds: 1193

Membrane proteins

http://ptools.citrusgreening.org/overviewsWeb/celOv.shtml



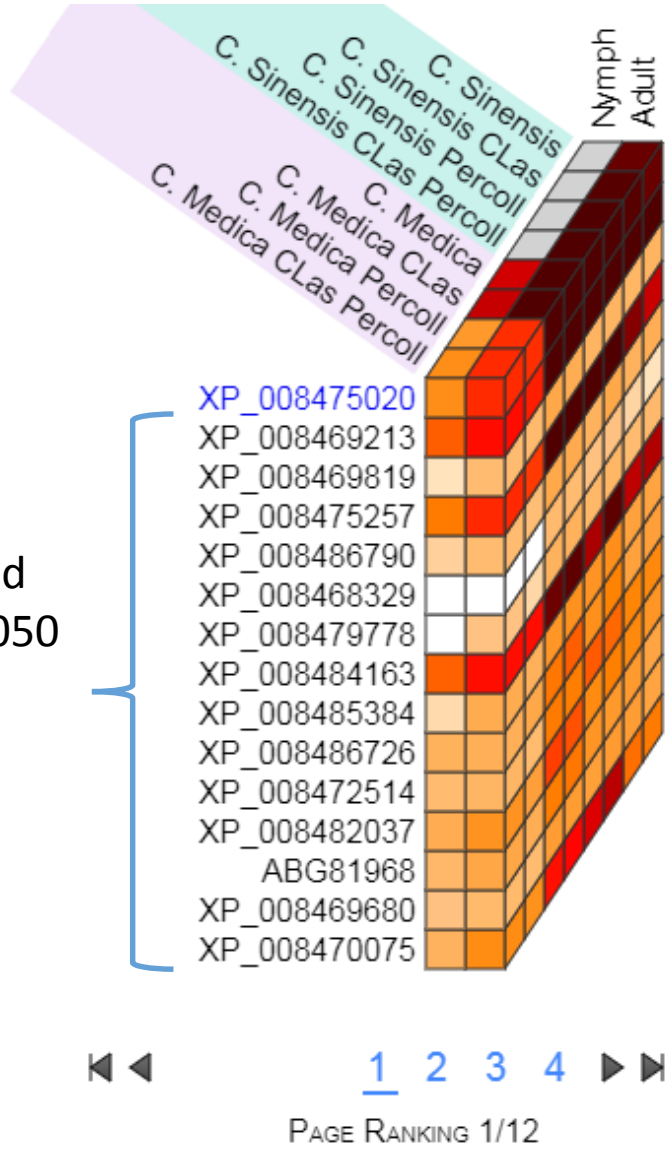
Secretory proteins

Membrane proteins

Cellular Overview of *Diaphorina citri* overlaid with RNAseq expression counts

Kruse et al. 2017

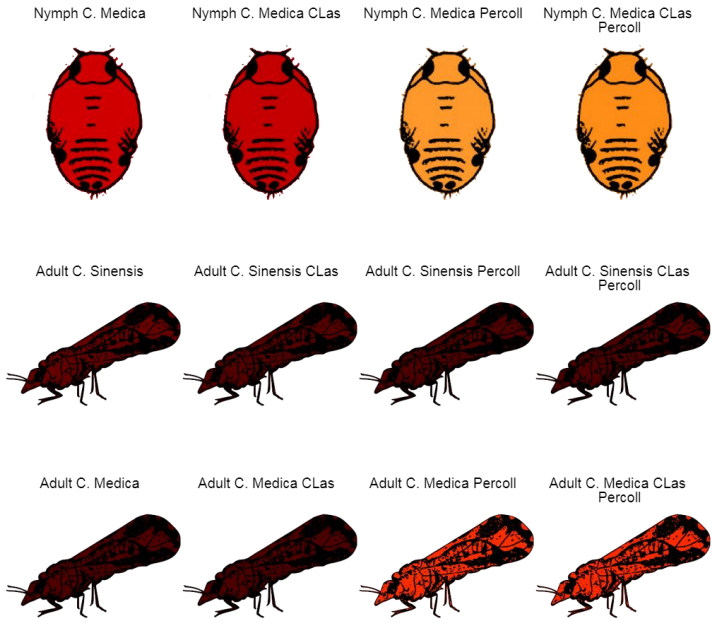
Psyllid Expression Network



Genes correlated with XP_00847050 ATP synthase subunit beta

Hosts: *C. medica* and *C. sinensis*
Treatment: Percoll gradient fractionation
Stages: Nymph and adult
Conditions: Clas+ and healthy

Colored by level of expression



Citrusgreening.org

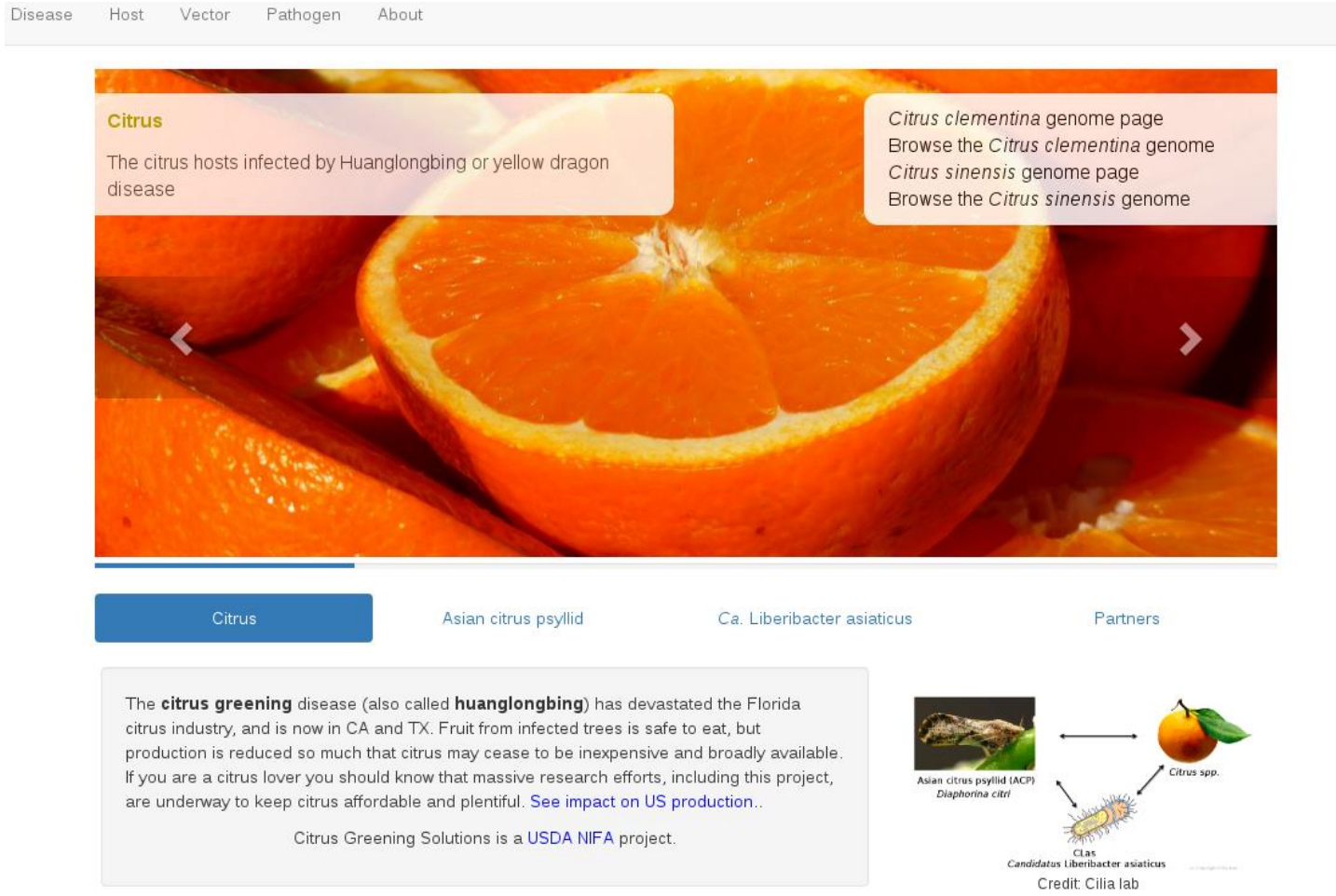
https://citrusgreening.org/

Host, Vector and Pathogen(s)
- Research highlights
- Blast Databases
- Gene pages
- Genome browser – Jbrowse
- Metabolic pathway database
- Annotation Editor – Apollo
- Psyllid Expression Network (PEN)
- FTP site for download

Disease background

News, Publications, Links

Social Media



Improved genome and annotation will expedite identification of targets for interdiction

Genome
Pacbio v2.0

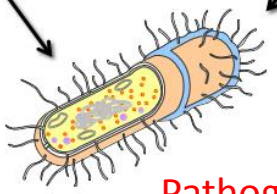
Pathway Databases
Expression Networks
…….

Annotation
Isoseq

Target for interdiction molecules

Asian citrus psyllid (ACP)
Diaphorina citri
Vector

Citrus spp.
Host



CLas
Candidatus Liberibacter asiaticus
Pathogen

Future work

- Transcriptome assembly with OGSv2.0, Pacbio Isoseq and Illumina transcripts
- Non-coding genes (lncRNAs, etc.)
- Proteomics evidence for OGS v2.0 genes
- Manual curation to create OGS v2.1
- New RNAseq datasets for Psyllid Expression Network

Feedback survey http://bit.ly/ACPv2survey

Thank you!!

@Citrusgreening
https://www.facebook.com/citrusgreening