Technology Innovations Provide Both Opportunities and Challenges for Agriculture

Doreen Ware, Ph.D
United States Department of Agriculture ARS

Ware Lab
Mike Campbell
Kapeel Chougule
Nick Gladman
Carol Hu
Yinping Jiao
Vivek Kumar
Sunita Kumari
Young Koung Lee
Zhenyuan Lu
Dimitri Muna
Andrew Olson
Michael Regulski
Josh Stein
Jim Thomason
Peter Van Buren
Bo Wang
George Wang
Liya Wang
Sharon Wei
Lifang Zhang

CSHL
Dick McCombie
Sara Goodwin

USDA-Geneva
Lance Cadle-Davidson
Xia Xu
Jason Londo

Cornell
Qi Sun
Fred Gouker

USDA-ARS, Lubbock TX
Zhanguo Xin
Gloria Burow
Ratan Chopra
John Burke
Chad Hayes

Cinerea B9
Bruce Reisch
Paola Barba
Katie Hyma
Shanshan Yang,
Will Thompson

Flame Seedless
Craig Ledbetter
Rachel Naegele

Concord
Gan-Yuan Zhong

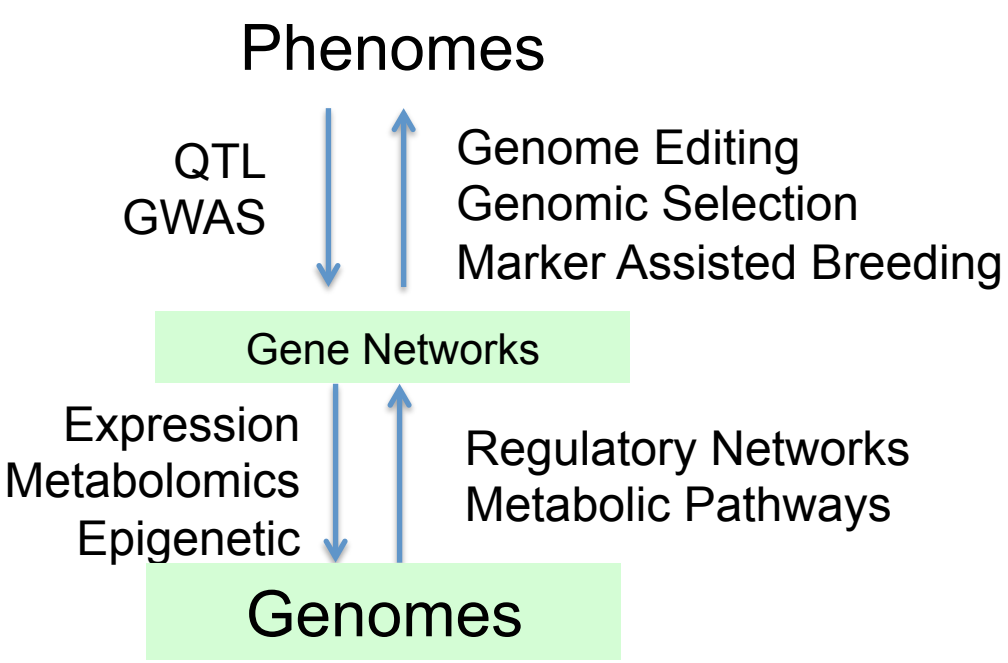
10X genomics
Stephen Williams
Deanna Church

Funding
USDA ARS
NSF
USDA-NIFA
California Table Grape Commission
National Grape and Wine Initiative

Advancing Agriculture Through Collaborative Research on Crop & Model Species

Biology Enabled Agriculture

Genotype X (Environment X Management) = Phenotype



Genome Represents:

- The sum total of an organisms' genetic constitution [Winkler 1920]
- A genetic integration of the environmental past and the potential response to future environments in which an organism exists [Stettler 1998]
- Organisms genetic potential for a phenotype under optimal environmental conditions…..

Biology Enabled Agriculture

Complex Traits: Yield & Quality

Phenomes

QTL
GWAS

Genome Editing
Genomic Selection
Marker Assisted Breeding

Gene Networks

Expression
Metabolomics
Epigenetic

Regulatory Networks
Metabolic Pathways

Genomes

What Has Changed?

- Technology improvements & reduction of cost has lead to the ability to sample and collect data temporally and spatially
- The **Volume, Velocity, Variety,** Complexity of the data has changed
- Single cell, whole organism, geospatial (micro to macro)
- **Biology has shifted from observational science to an information science, and needs to move into a predictive science**
- Resources needed to support this shift has not kept pace
  - Human Resources
  - Community Building
  - Knowledge Management
  - Standards
  - Policies
  - Network
  - Storage
  - Compute
  - Innovation

Biology is an information science

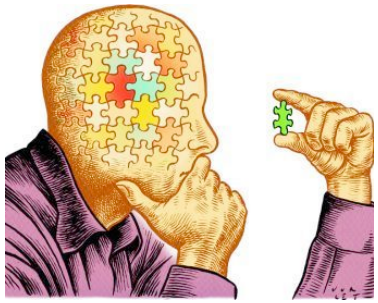
Defining the Critical Needs in Predictive Biology

Experimental hypothesis

Experimentation

Data

$$\int_i^j reasoning\, dx$$



Models & tests

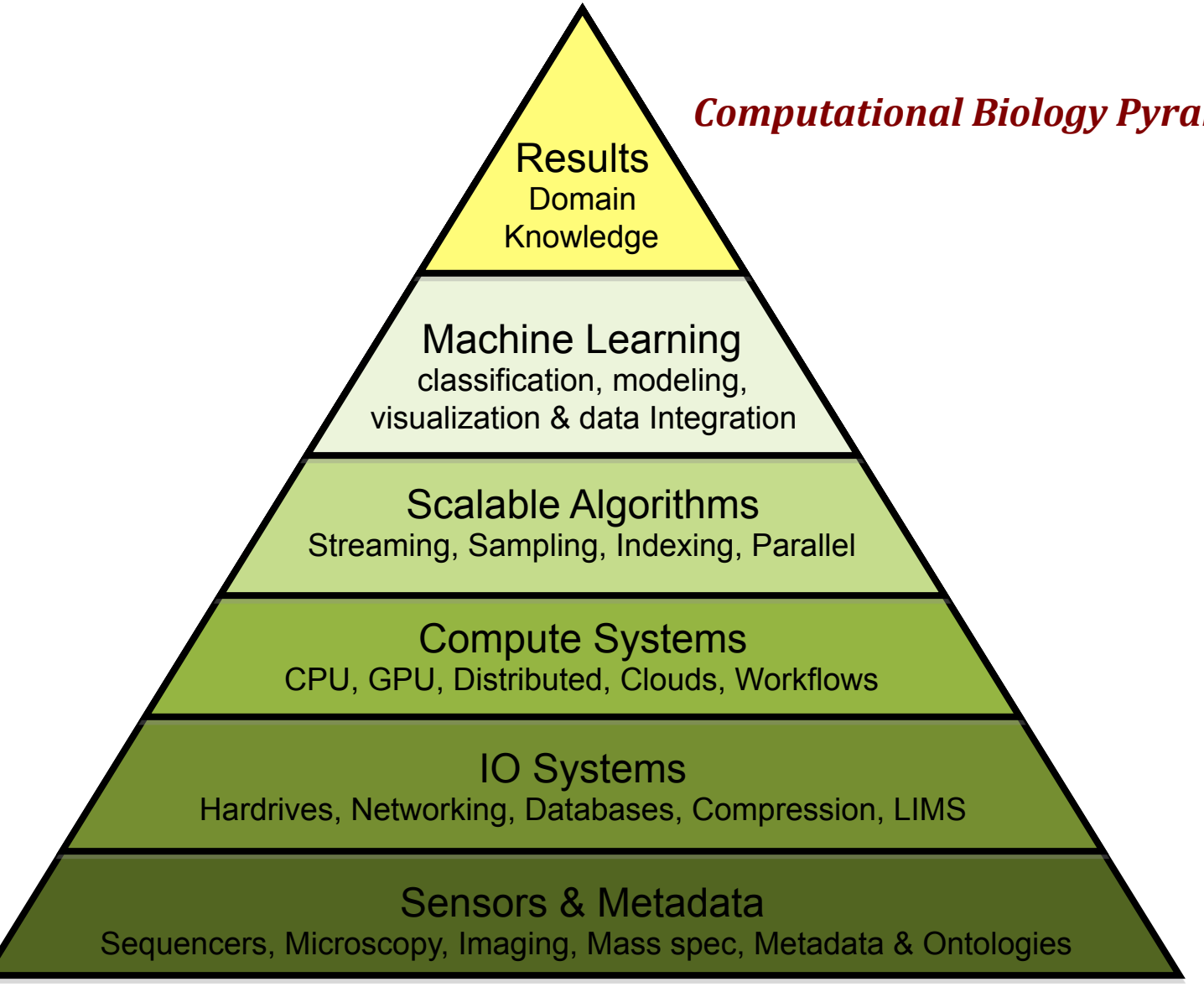
Compute

Dissemination

Courtesy David Weston

Image credits: improveddiagnosis.com, uptowndallascounseling.com; www.udemy.com; www.guynetsystems.com

Biology has transitioned to an information science

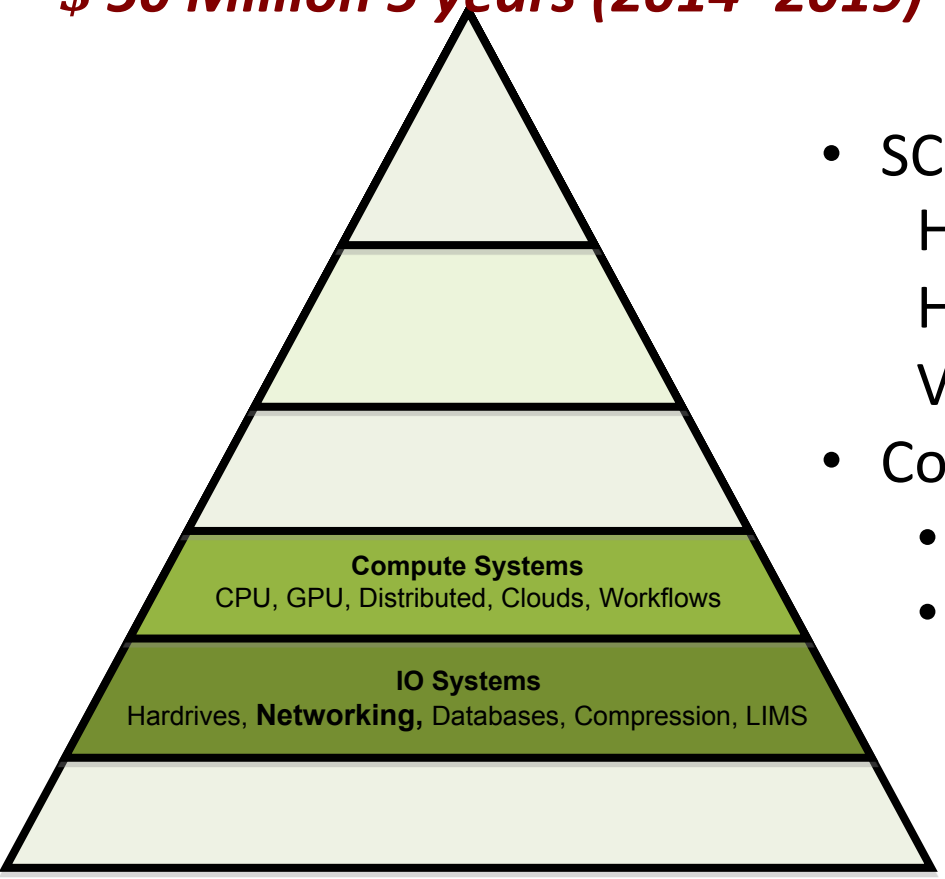


Computational Biology Pyramid

Keystone Big Data in Biology 2014 Stein, Schatz, Ware

ARS Big Data initiative

Computational infrastructure for Agriculture
$ 50 Million 5 years (2014- 2019)



- SCINet/ CERES
  - High-speed network backbone
  - High-performance computing cluster
  - Virtual Research Support Core
- Community Building and Training
  - Science focus workshops
  - Data and Software literacy

Computational Infrastructure for the Life Sciences

CYVERSE™
www.cyverse.org

Access to data storage, cloud & HPC resources for life science community

Gramene
www.gramene.org

EnsemblPlants

Comparative Plant Genome & Pathway Resource

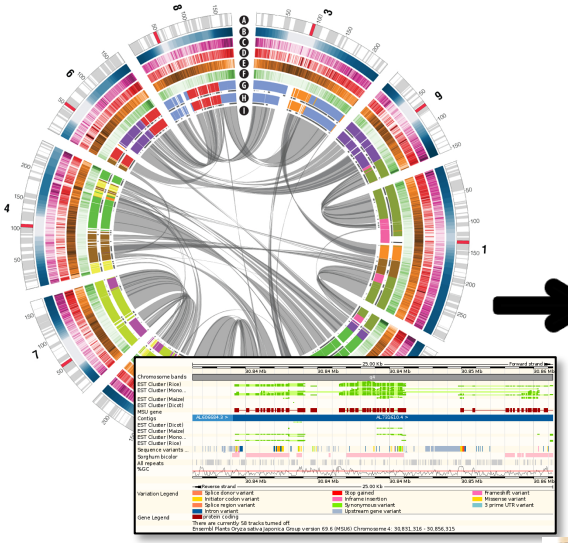
USDA

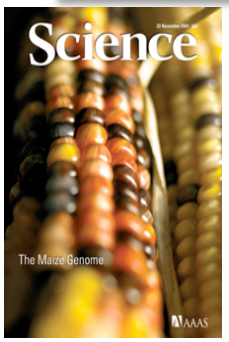
EMBL-EBI

KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase
Kbase.us
Plant, microbes & microbial communities

Moving from a single reference, to population variation, & functional network inference

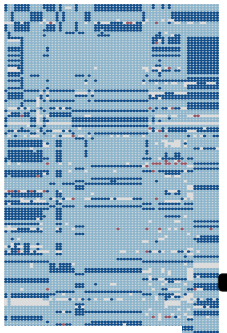
Genomes & Annotations

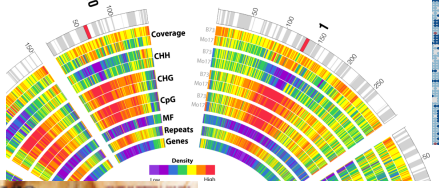


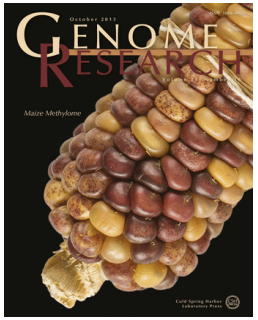


Schnable, Ware et al. Science (2009)
Zhang et al. PLoS Genetics (2009)
Kumari &Ware Plos One (2013)
Olson et al. Plant Genome (2014)
Wang et al Nature Comm. (2016)
Jiao et al. Nature(2016)

Genetic and Epigenetic Variation

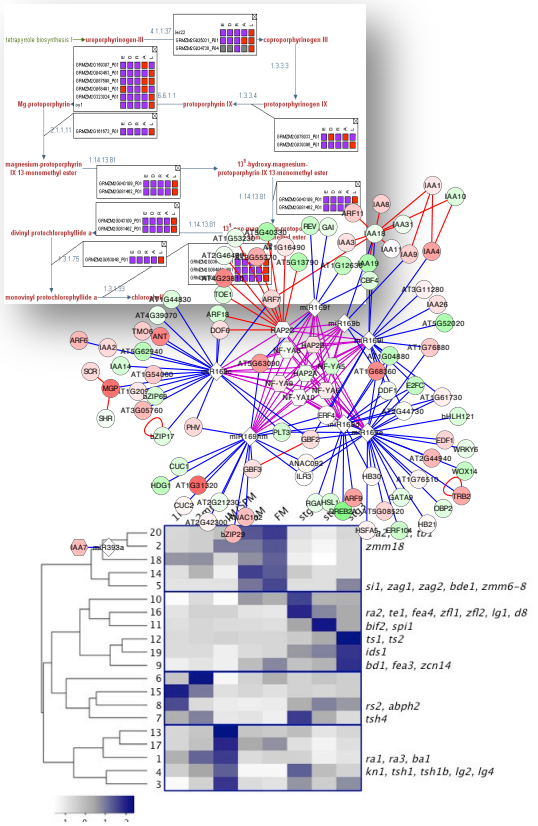






Gore, Chia, et. Al. Science, (2009)
Chia, Song et al. Nature Genetics(2012)
Regulski, Lu et al. Genome Res. (2013)
Jiao et al. Plant Cell (2016)
Wang et al. Nature Com. (2017)
Wang et al. Submitted (2017)

Developmental, Metabolic, & Stress Networks



Gaudinier A et al. Nature methods (2011)
Monaco et al. Plant Genome (2012)
Liu et al, Plos One (2012)
Eveland et al. Genome Res. (2014)
Seaver et al. PNAS. (2014)
Taylor-Teeples M et al. Nature (2015)
Jioa, Lee et al Sumbitted

Populations Contains a High Level of Genetic Diversity

- High rate of SNP and **structure variation** in the population
- Structure variations are **associate with important traits**
- **One genome is not enough** to represent the diversity of the population

PLoS One. 2010 Jan 13;5(1):e8219. doi: 10.1371/journal.pone.0008219.

Rapid genomic characterization of the genus vitis.

Myles S[1], Chia JM, Hurwitz B, Simon C, Zhong GY, Buckler E, Ware D.

Author information

Abstract

Next-generation sequencing technologies promise to dramatically accel genetic mapping of agriculturally important phenotypes. The first step in polymorphism discovery and a subsequent genome-wide assessment o species of interest. In the present study, we provide such an assessmen fruit crop. Reduced representation libraries (RRLs) from 17 grape DNA s with sequencing-by-synthesis technology. We developed heuristic appro validated a subset of these SNPs on a 9K genotyping array. We demons among V. vinifera cultivars, between V. vinifera and wild Vitis species, ar

Chia JM, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, Elshire RJ, Gaut B, Geller L, Glaubitz JC *et al*: **Maize HapMap2 identifies extant variation from a genome in flux**. *Nat Genet* 2012, **44**(7):803-807.

Genome Assembly by Single-Molecule Technology (15,000 bp)

Yinping Jiao

WGS by PacBio 65X (N50= ***15kb***)

De novo assembly

Long single molecule sequencing is a game changer

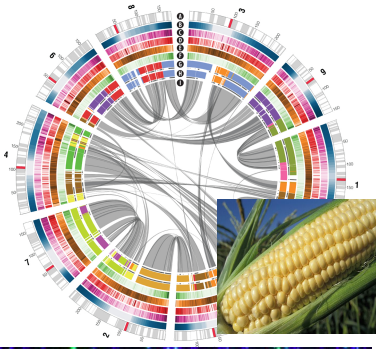
- High quality genomic DNA
- Access to Sequencers
- Access to computes
- Optimization of software/ algorithm
  - Genomic architecture (repeat content, inbred, ploidy….)
  - Computing environments
  - Parameterization

PACIFIC BIOSCIENCES

Gramene

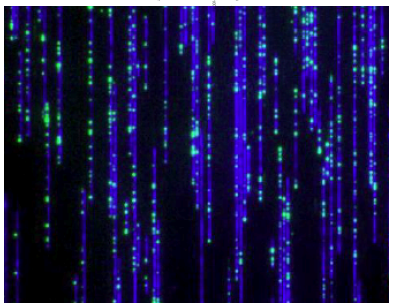
Jiao *et al.*, **Nature**, 2017

Decreasing cost of sequencing increasing computes and data management



WGS by PacBio 65X (N50= 15kb)
1-6 months

De novo assembly
1-6 months



BioNano genome map
Technically challenging

$35 million (2009) Sequencing Centers

BAC library, Sanger sequencing library, finishing libraries,computes

$250- 180 thousand (2016) Sequencing Centers

PacBio long single molecule, Optical map, illumina short read

High quality DNA, Library prep, access sequencer & ***compute

$60- 30 thousand (2017) Pac Bio long single molecule

High quality DNA, Library prep, access sequencer & ***compute

Decrease Sequence Cost | Increase Compute Cost | Improve Assembly Quality | Improve Gene Quality

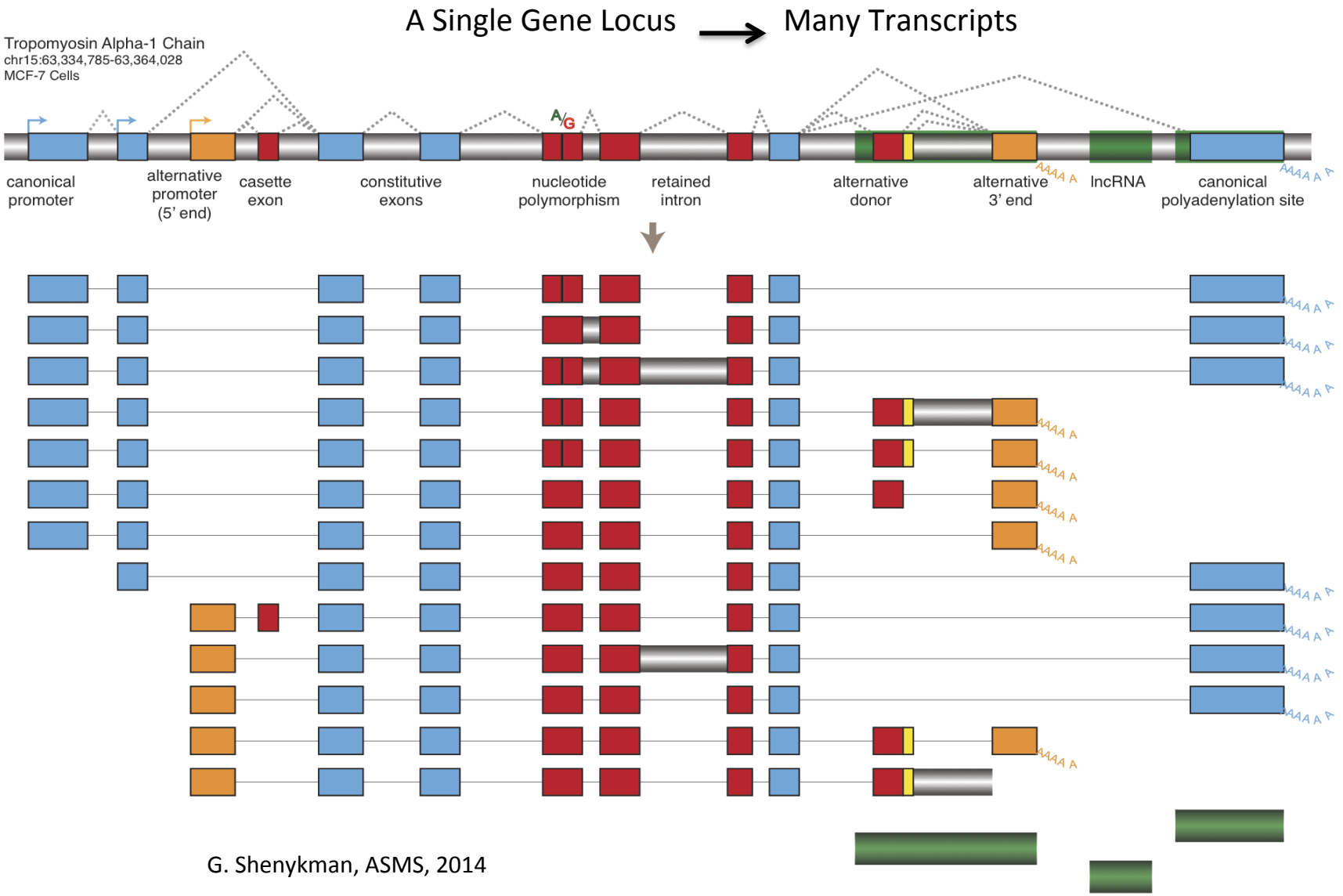
PACIFIC BIOSCIENCES

Gramene

BIONANO GENOMICS

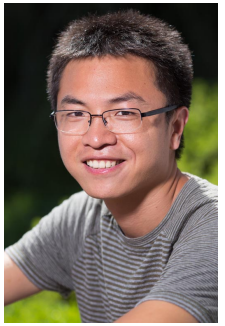
Jiao *et al.*, **Nature**, 2017

Transcript Diversity Contributes to Phenotypic Plasticity

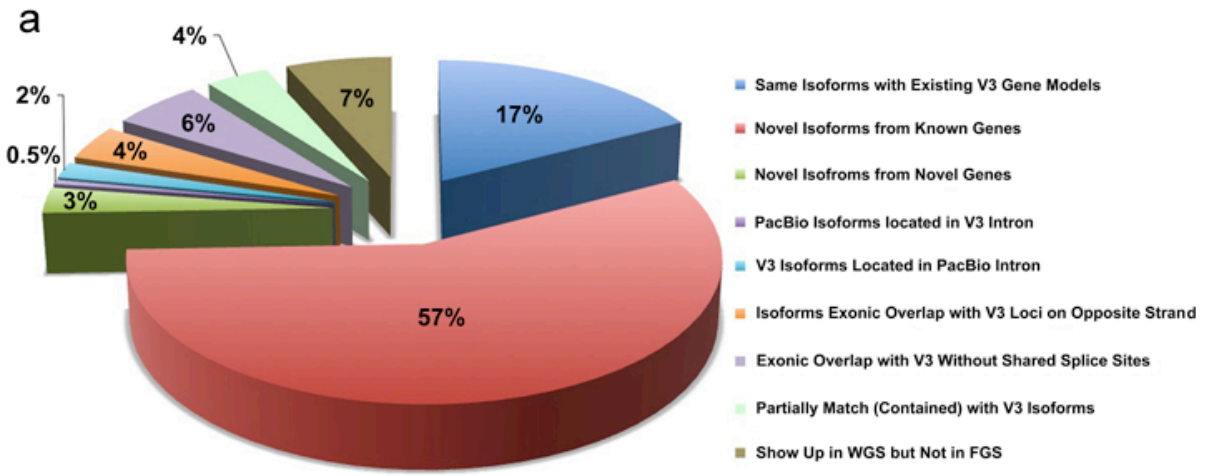


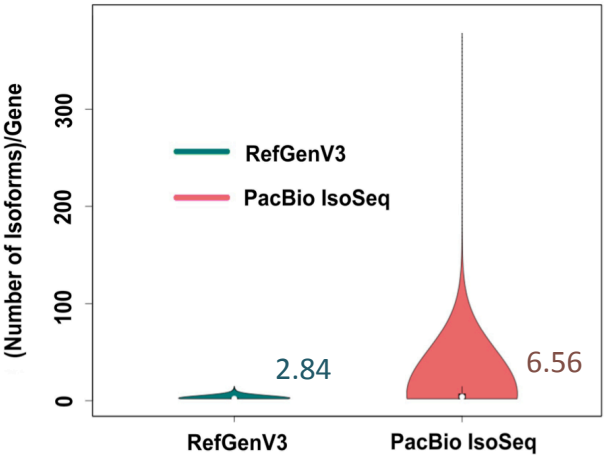
G. Shenykman, ASMS, 2014

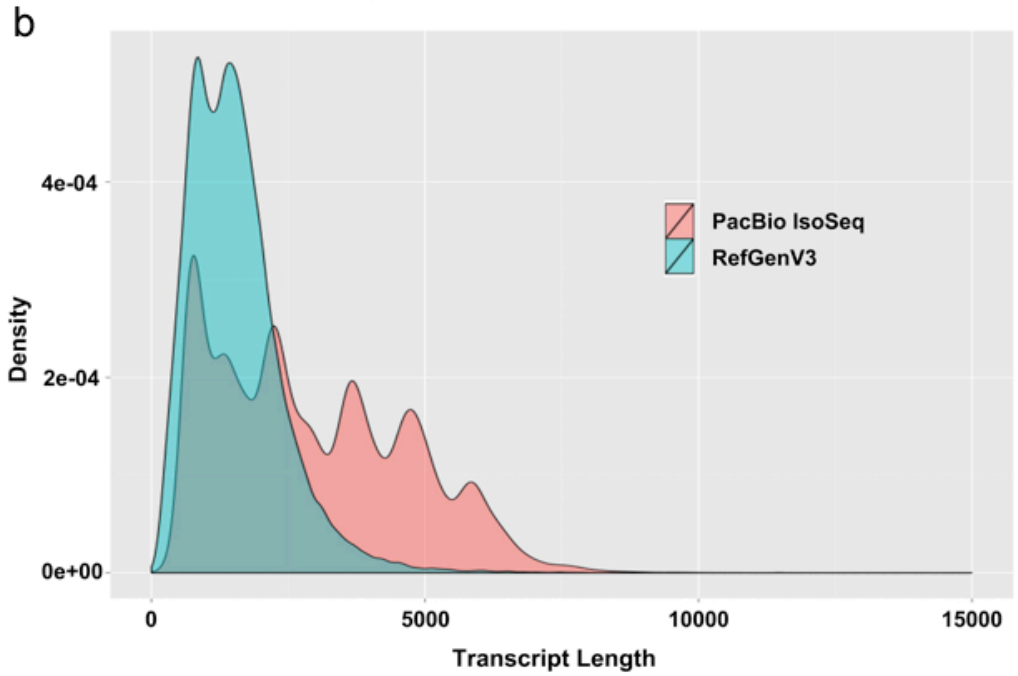
Reference Transcriptomes using PacBio IsoSeq captured longer and novel isoforms



Bo Wang

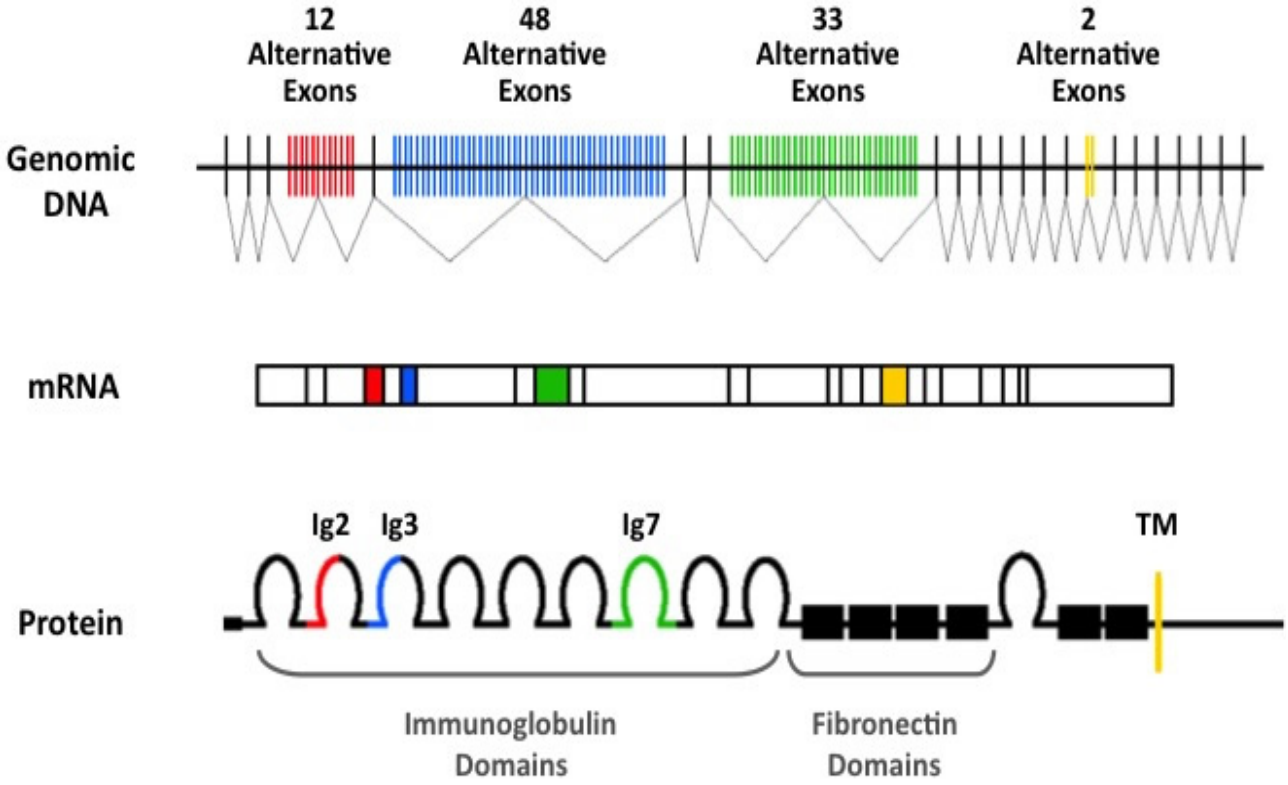






Wang *et al*., **Nature Comm**., 2016

Drosophila DSCAM Gene – 38,000 Isoforms



Schmucker D, et al. 2000. *Cell* 101:671-684

Number of coding-genes in different species:

Drosophila: 13,918

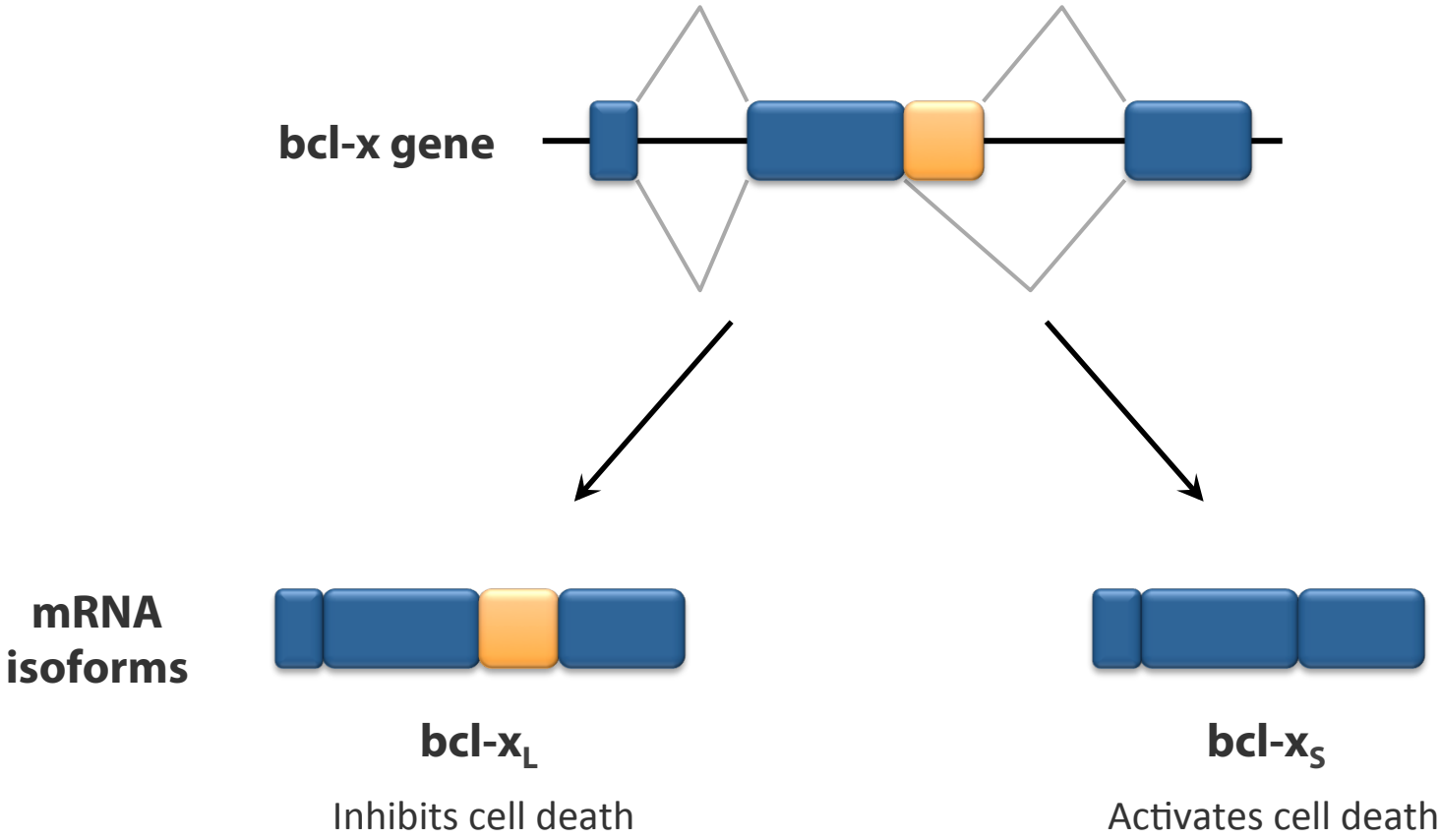
Arabidopsis: 27,416

Maize: ~39,000

Human: 20,296

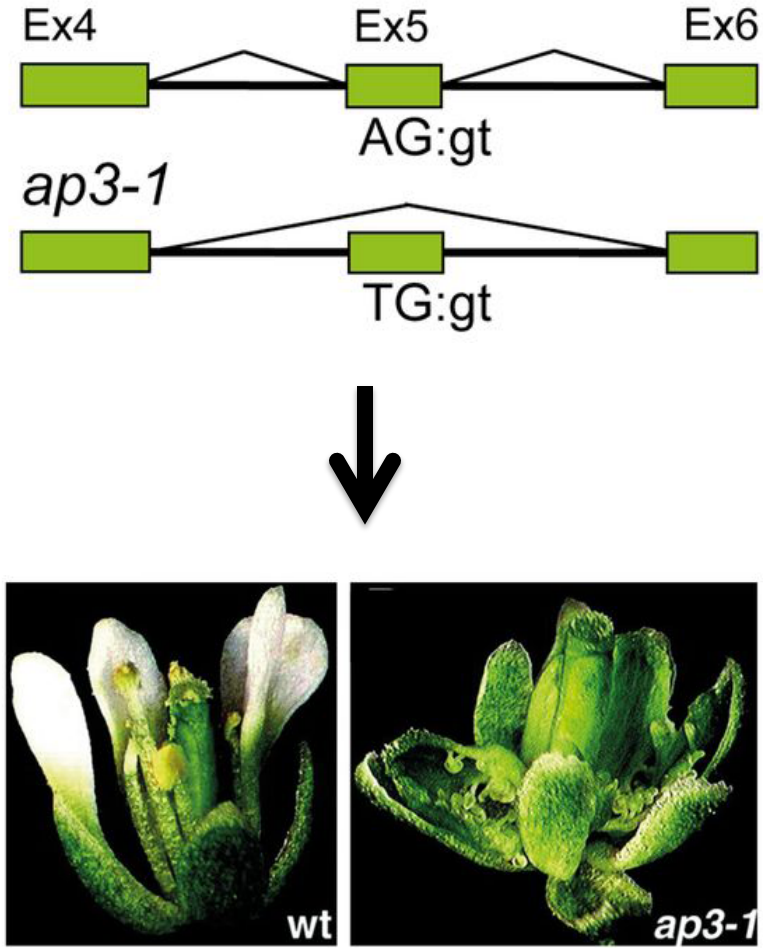
Mouse: 22,528

Mouse One Gene, Two Isoforms with Opposite Effects



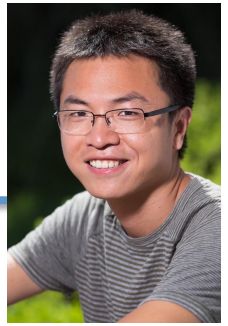
Li *et al.*, Cell Research, 2004

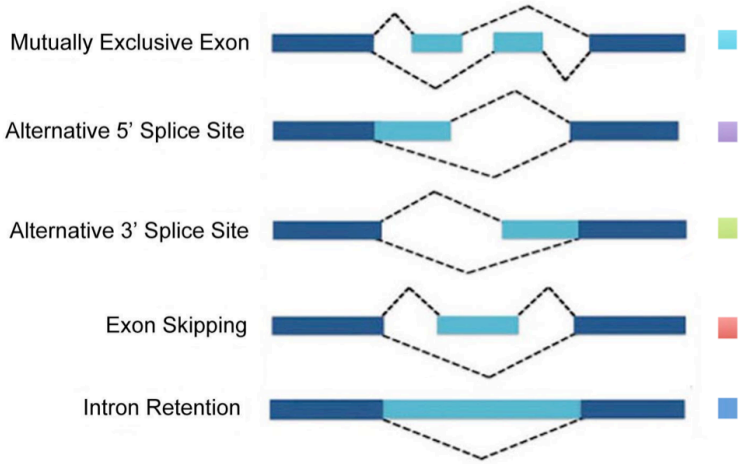
Arabidopsis Point Mutation at the 5' Splice Site Leads to Skipping of exon 5 and a Nonfunctional AP3 Protein

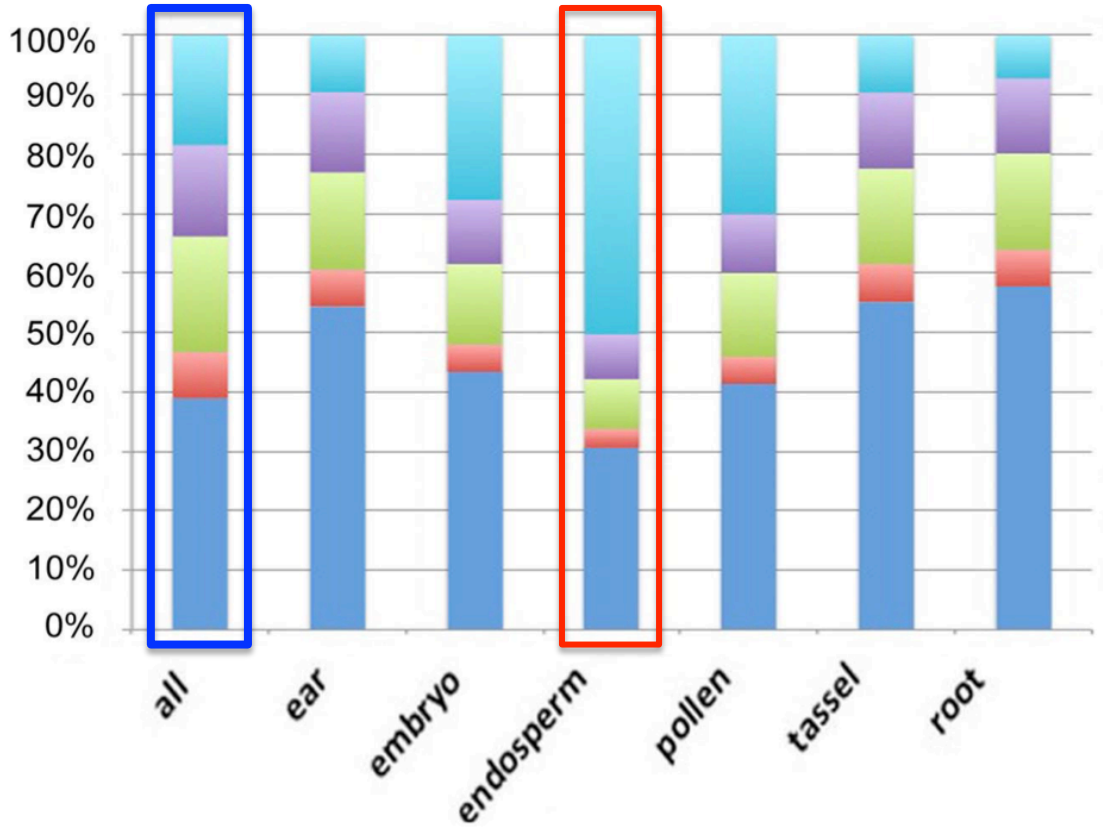


Staiger et al., 2013, Plant Cell

Characterize Alternative Splicing Events Among Different Tissues

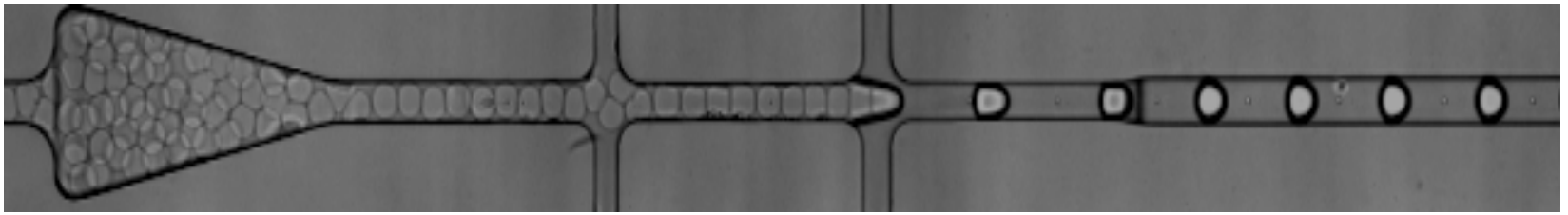






Wang et al. Nature Comm. (2016)

10x Platform: Millions of Picoscale Reactions

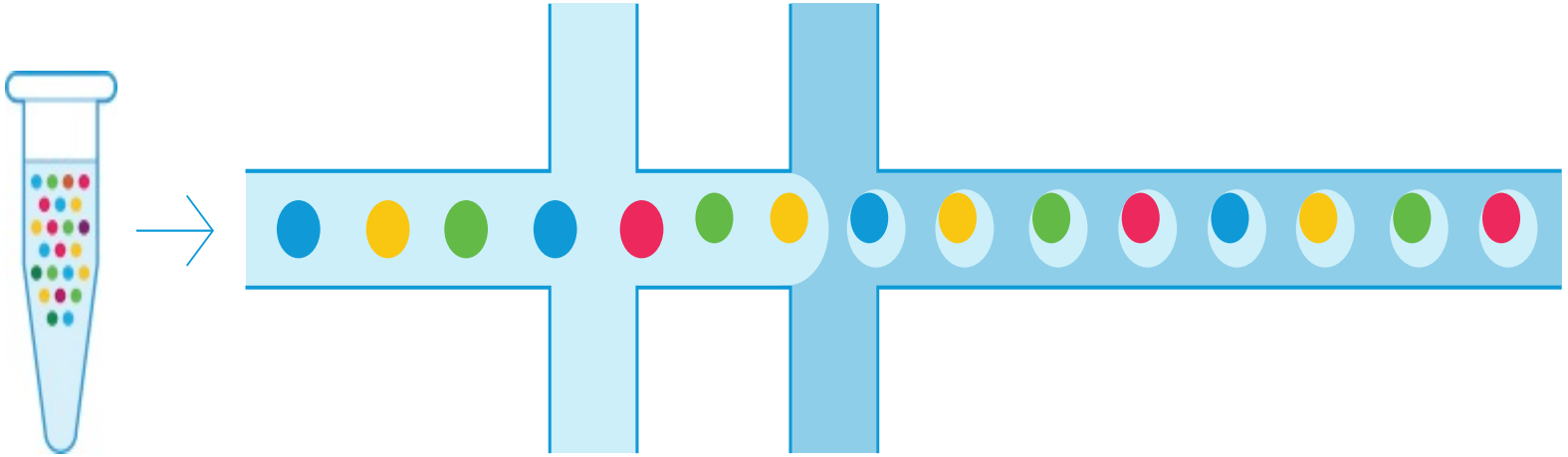


Gel Beads

Sample

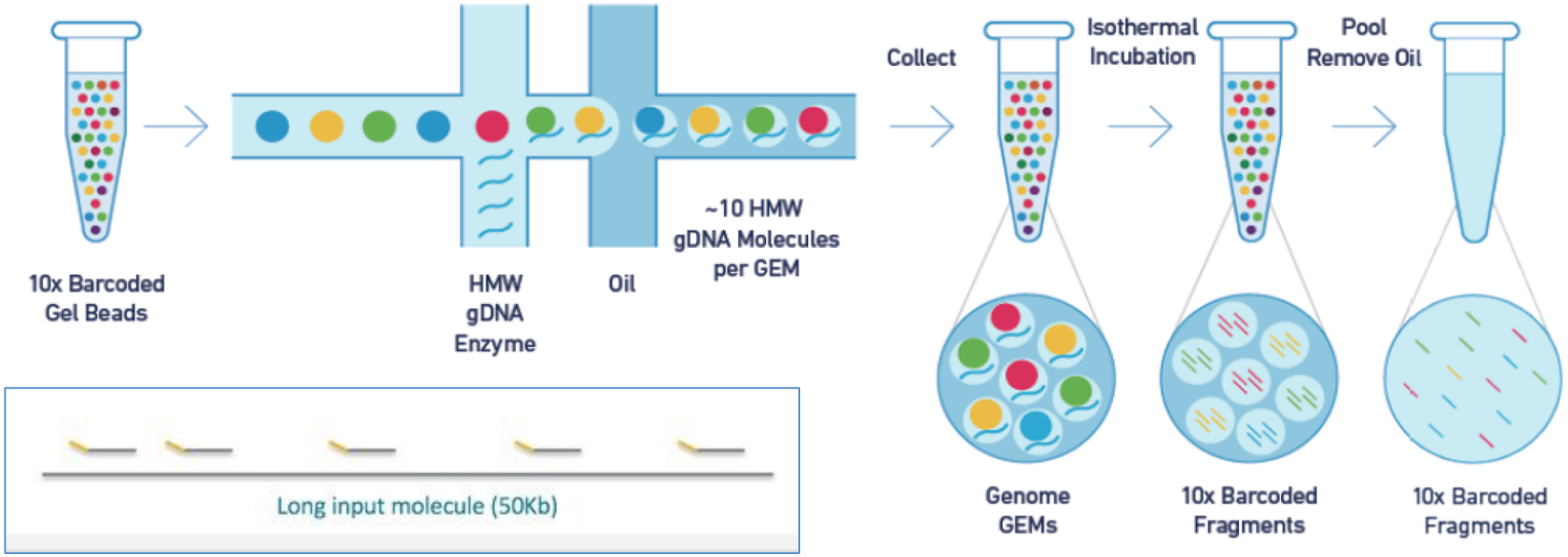
Oil

Droplets with Gel Beads

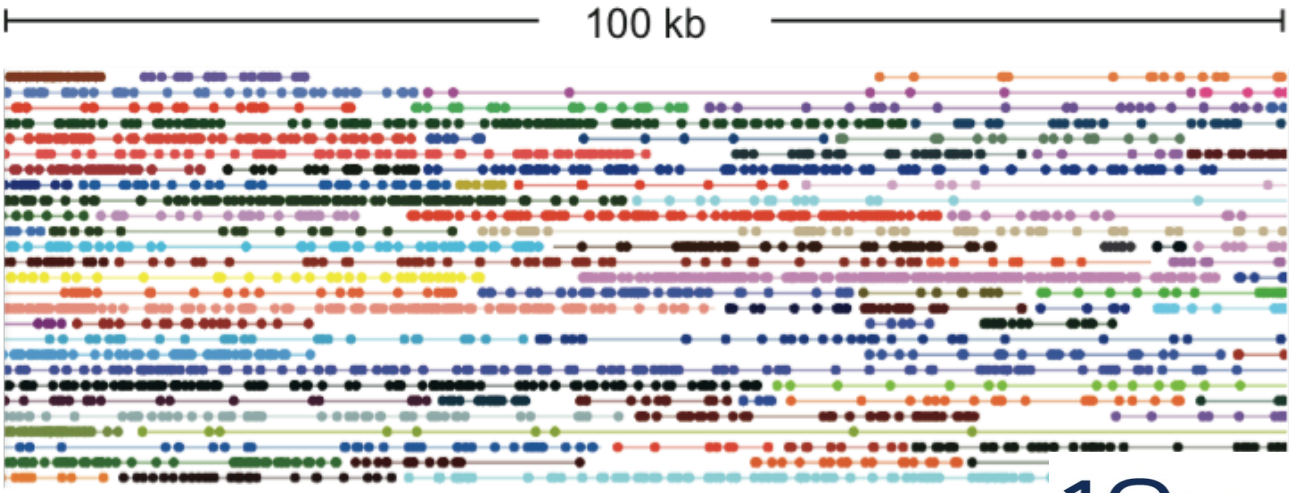


10x GENOMICS®

10x Genomics Linked Long Reads



10X Genomics
Linked Reads



10x GENOMICS®

Hybrid approaches decreased sequencing & computes cost

WGS by 10X
100 X (N50= 75-150 kb)
1-2 weeks

Illumina
Short read sequencing
1-2 weeks

De novo assembly
Supernova
2-6 days

$ 12-3 thousand (2018) Hybrid Approach
library prep, commodity sequencing, compute

Decrease Sequence Cost	Decrease Compute Cost	Reduced Assembly Quality	Reduced Gene Quality

10X GENOMICS®

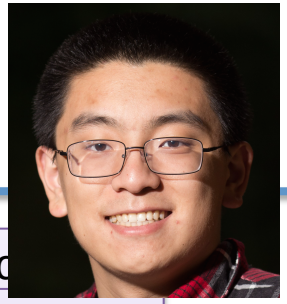
PACIFIC BIOSCIENCES™

Gramene

Campbell, *et al., unpublished*

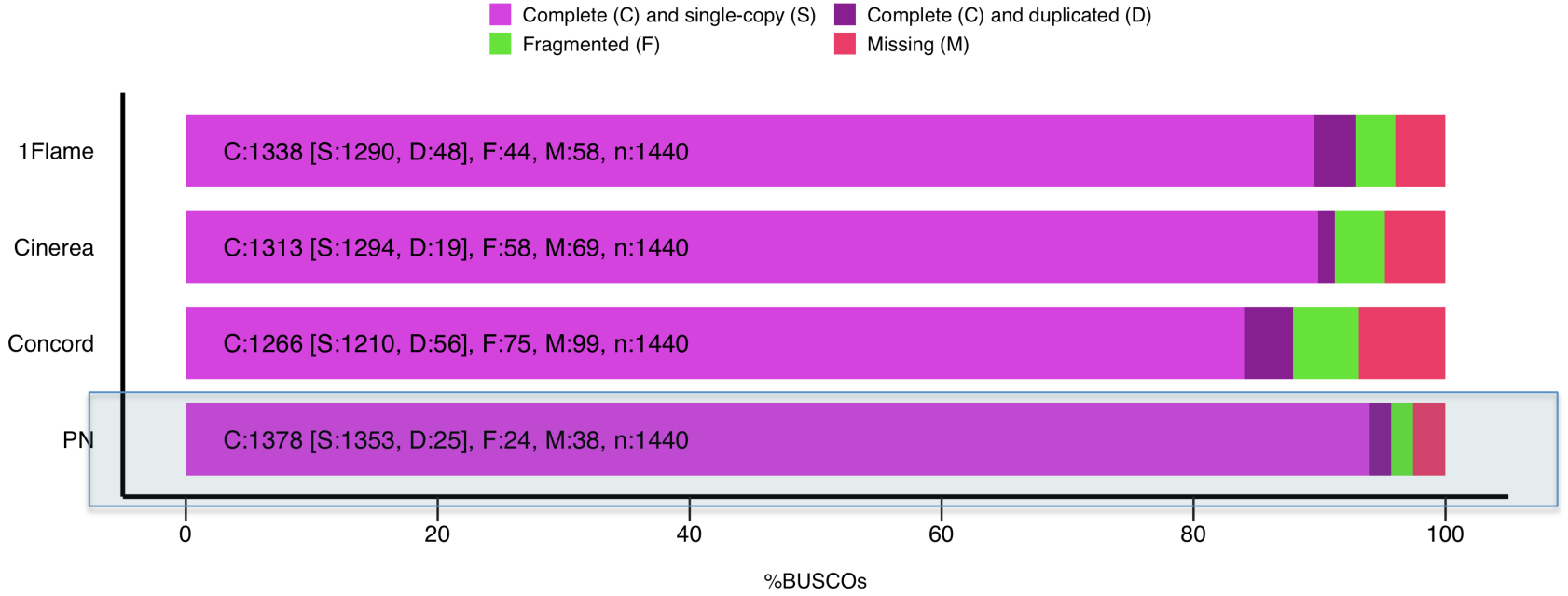
10X Supernova assembly & gene coverage assessment

George Wang



Mike Campbell

	Flame seedless	Cinerea B9	Conc
LONG SCAFFOLDS	2.85 K	4.25 K	5.01 K
EDGE N50	7.48 Kb	6.10 Kb	6.18 Kb
CONTIG N50	42.30 Kb	38.17 Kb	33.05 Kb
PHASEBLOCK N50	445.92 Kb	200.15 Kb	271.94 Kb
SCAFFOLD N50	572.37 Kb	197.15 Kb	191.63 Kb
SCAFFOLD N60	381.11 Kb	143.07 Kb	132.54 Kb
ASSEMBLY SIZE	365.64 Mb	349.92 Mb	382.66 Mb



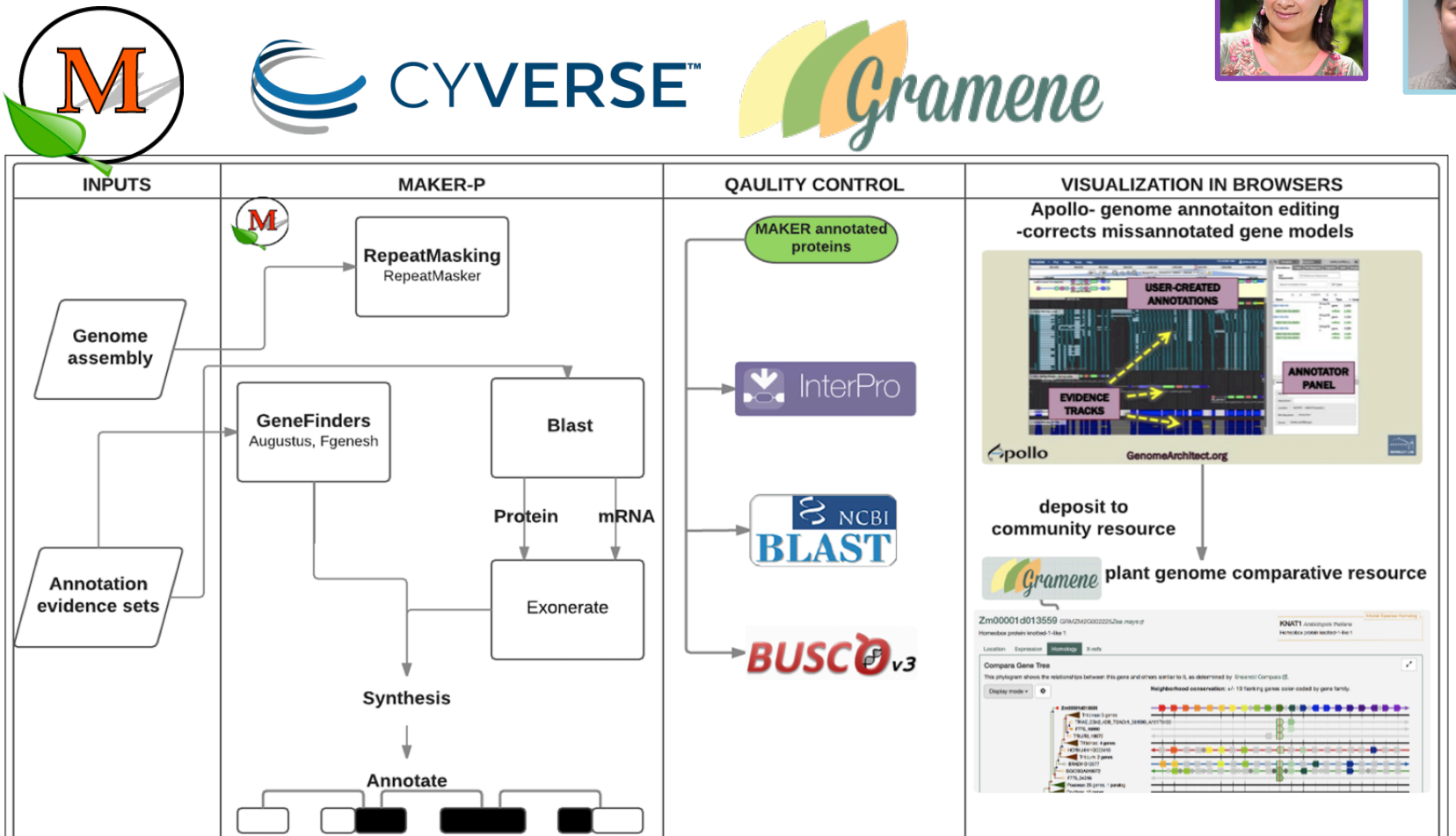
Reproducible workflows and Access to gene annotations

- *Structural and Functional annotations workflow on JetStream*
- *Access through Cyverse Data Store*
- *Visualization and Access through Gramene*
- *Curation of gene models through Apollo image on Cyverse*





MAKER-P Pipeline Mark Yandell & Carson Holt, U. Utah



Gene Tree Assessment to Support Gene Quality and Pan Gene/ Genome construction

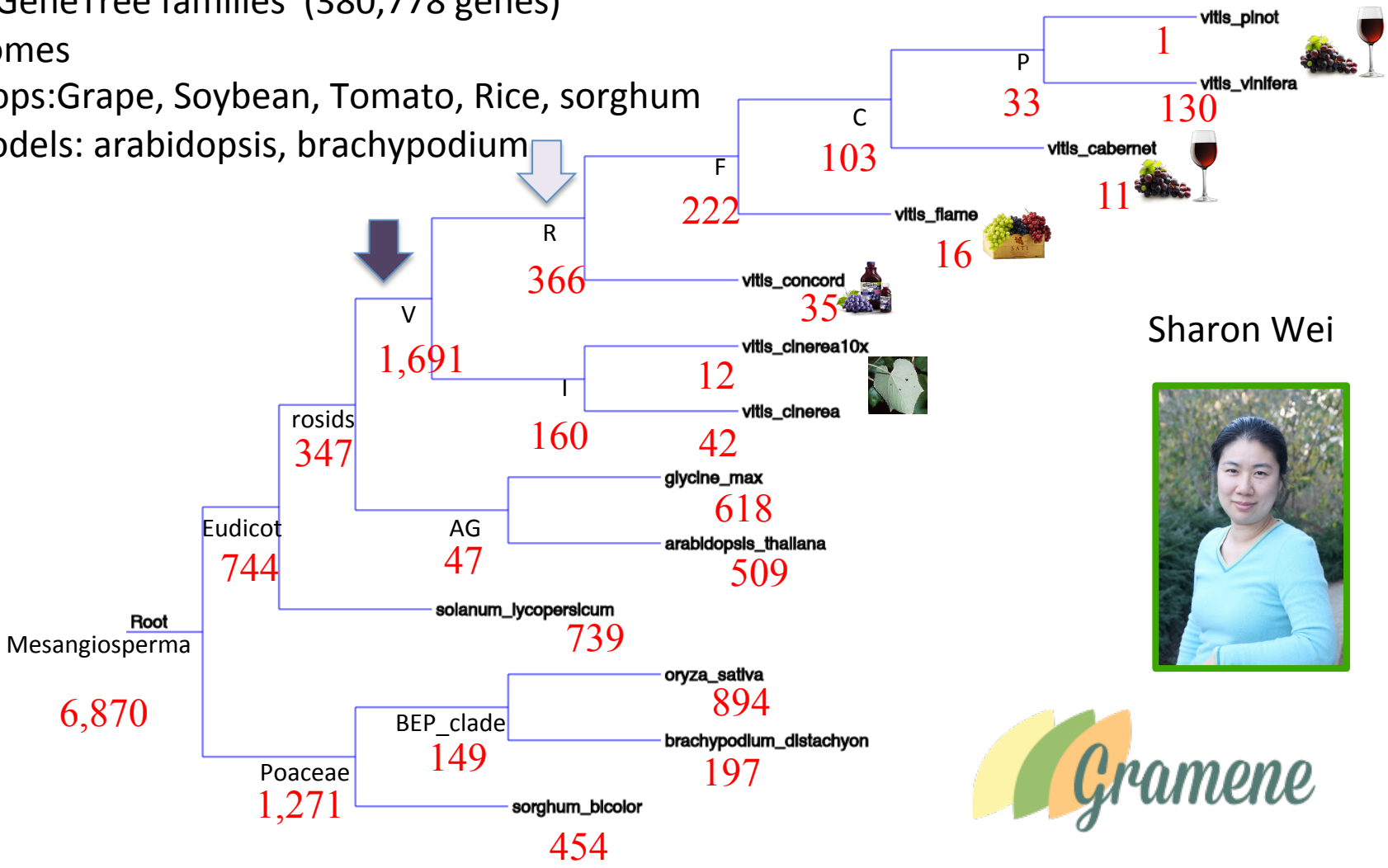
Gene Tree Counts at Ancestral Roots

15,661 GeneTree families (380,778 genes)

13 genomes

Crops:Grape, Soybean, Tomato, Rice, sorghum

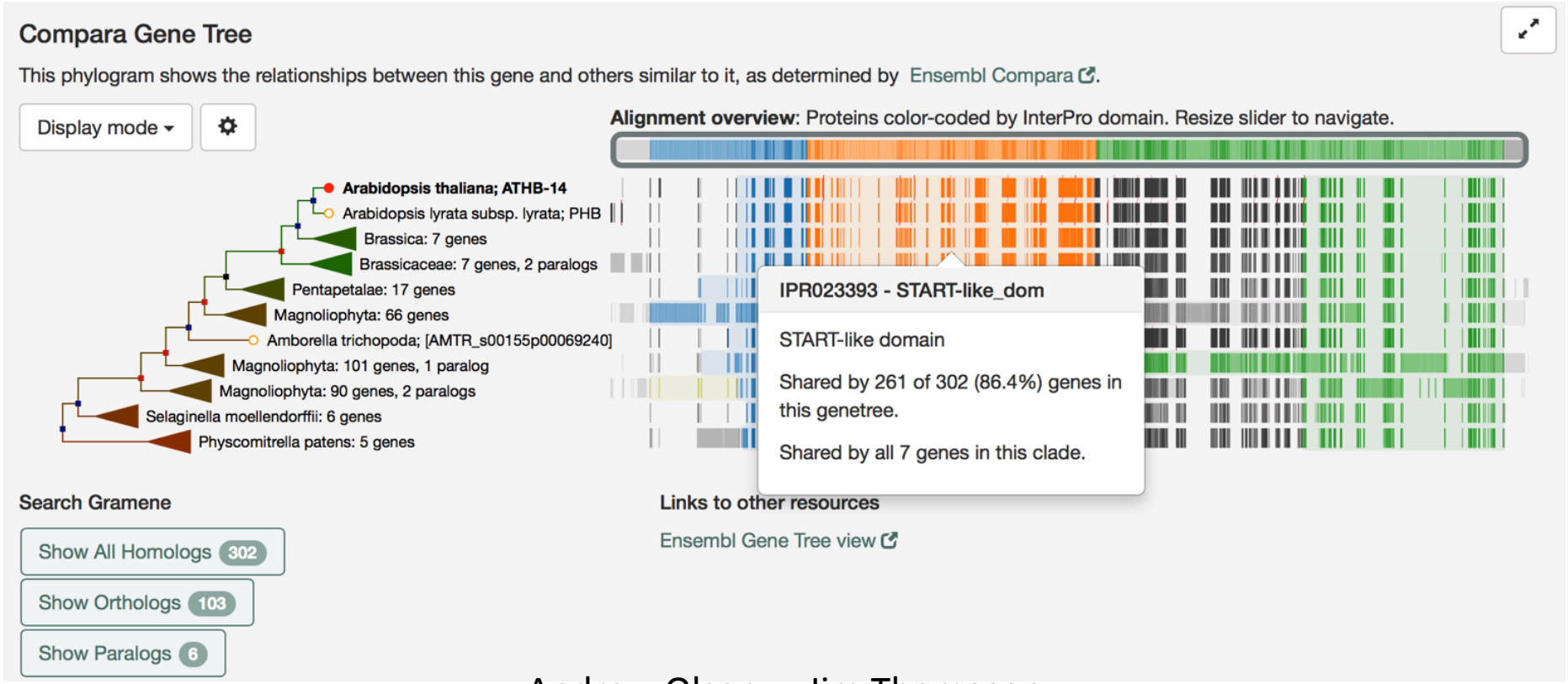
Models: arabidopsis, brachypodium



Sharon Wei

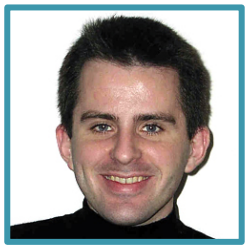
Gramene

Protein based alignment overview highlighting function domain



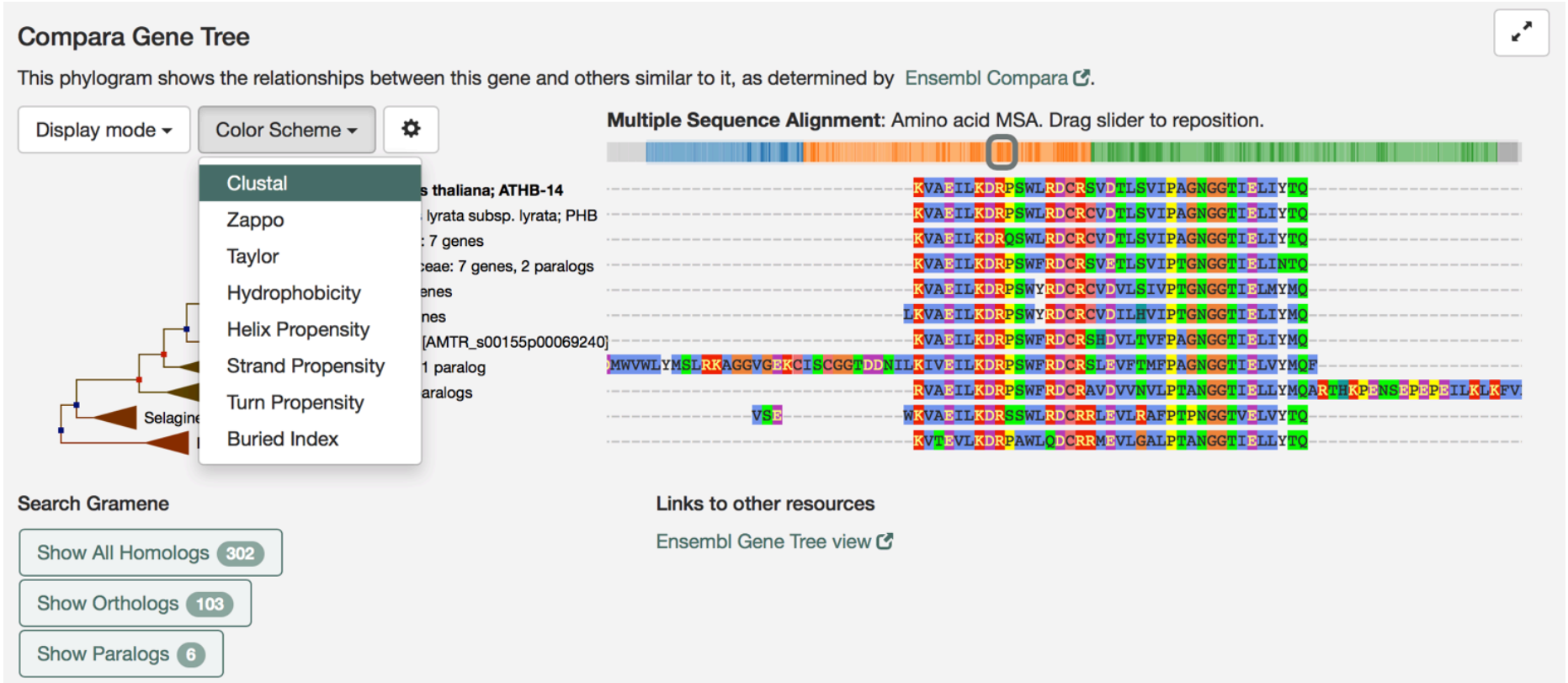
Andrew Olson

Jim Thomason



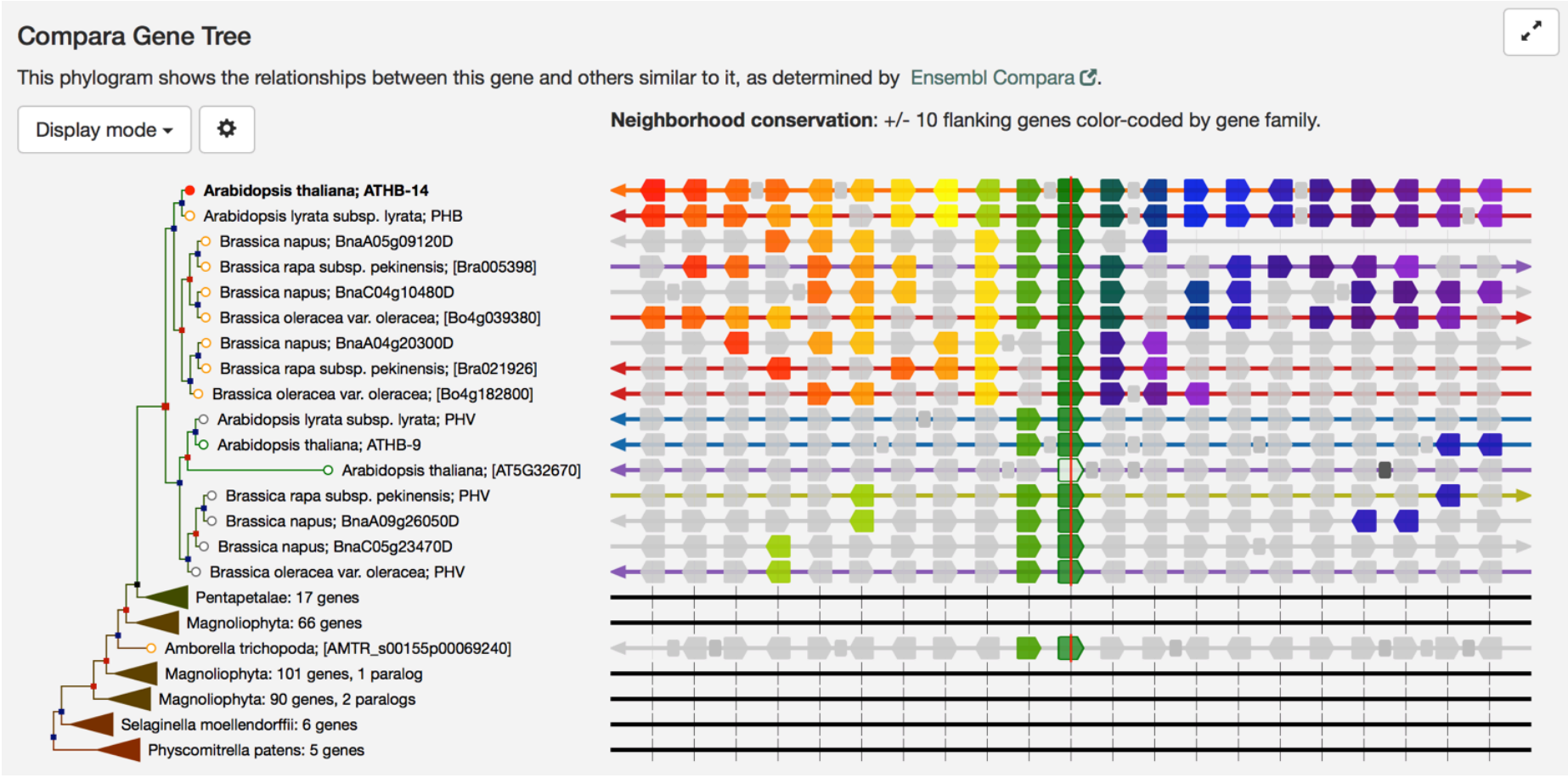
Gramene

Multiple sequence alignment



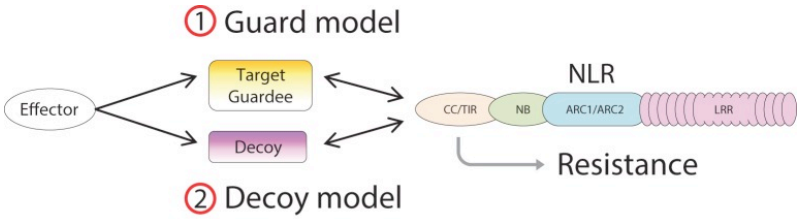
Gramene

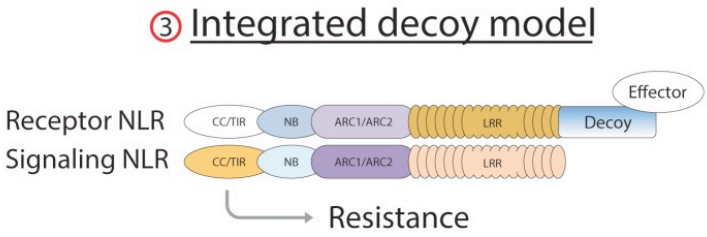
Neighborhood conservation



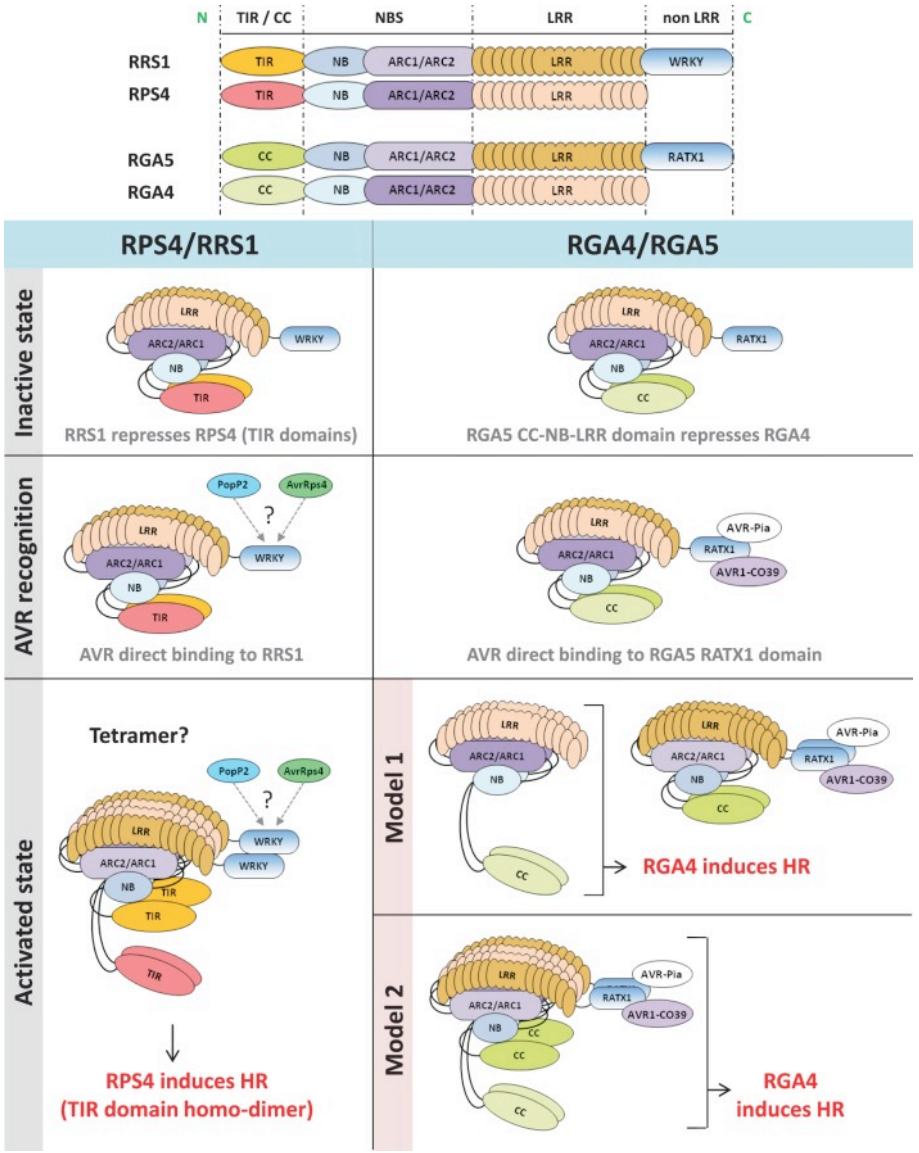
Gramene

Coupled R-genes and the Integrated Decoy Hypothesis





Thomas Kroj
(INRA/CIRAD)



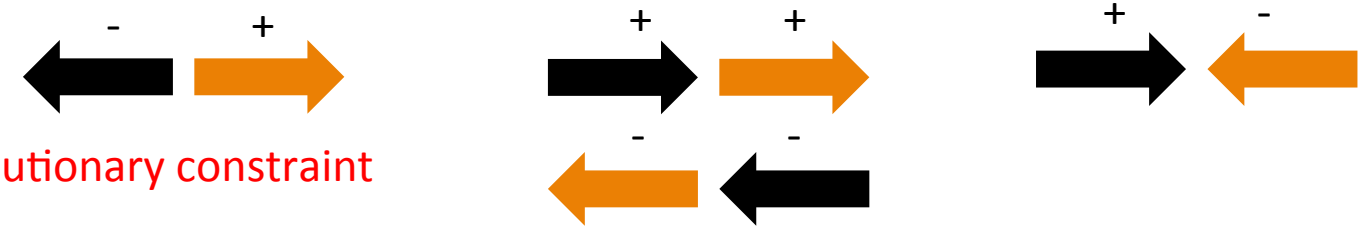
Cesari et al. Front Plant Sci. 2014; 5: 606.

Genome architecture and evolutionary constraint to identify genes in wild rice genomes

Josh Stein

	Head-to-Head (H2H)	Head-to-Tail (H2T)	Tail-to-Tail (T2T)
Expect	25%	50%	25%

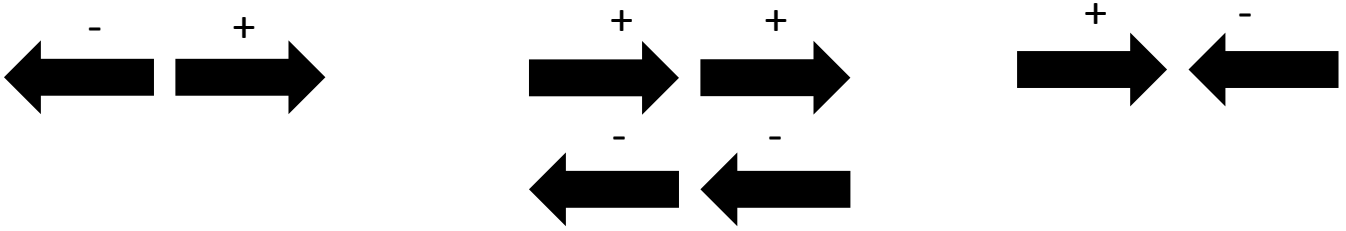
Heterogeneous



Evolutionary constraint

n = 311	45%	37%	18%
Synteny (%)	72%	55%	61%
Unusual domains	30%	15%	21%

Homogeneous

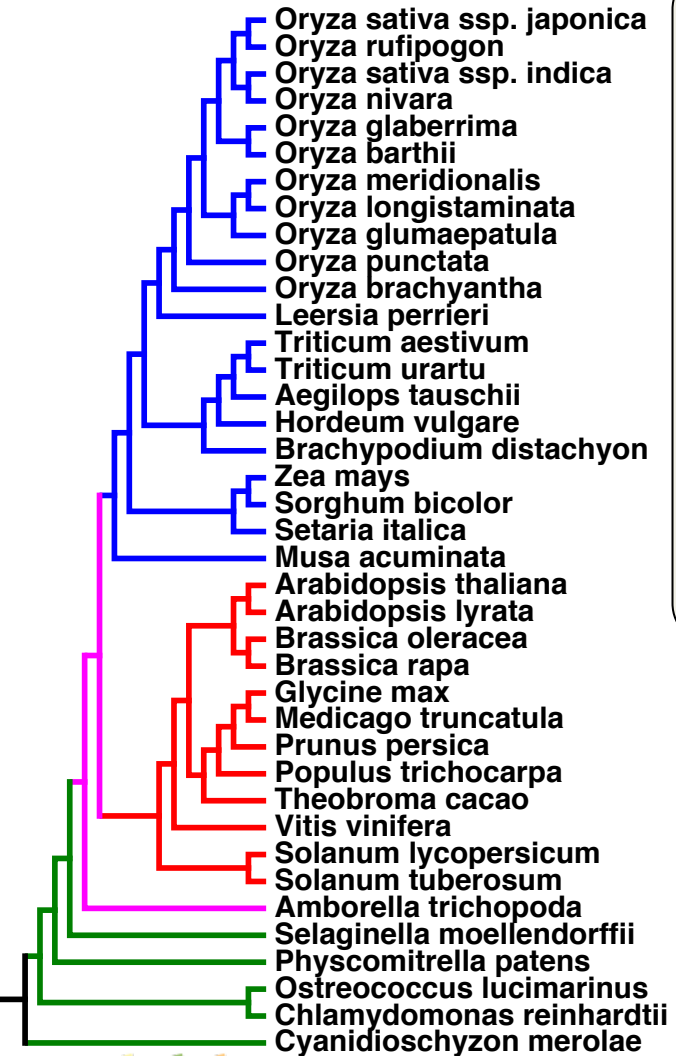


n = 895	17%	74%	9%
Synteny (%)	83%	65%	60%
Unusual domains	20%	9%	3%

Chi-square test, $P < 0.0001$

Stein et al, *Nature Genetics 2018*

Gramene Adds Value to Plant Genomes



Annotation Pipelines

- Repeats/TE's
- Genes
- EST/cDNA
- InterPro domain
- Gene Ontology (GO)
- Variant Effect Prediction

Comparative Analysis

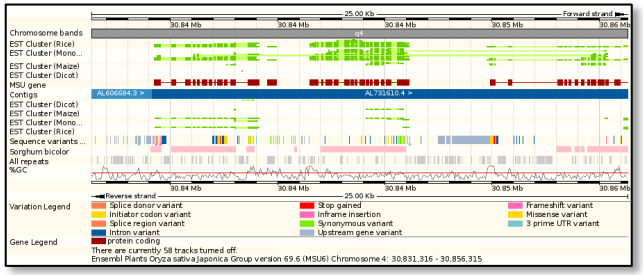
- Whole Genome Alignment
- Phylogenetic Gene Trees
- Ortholog/Paralog calling
- Synteny mapping

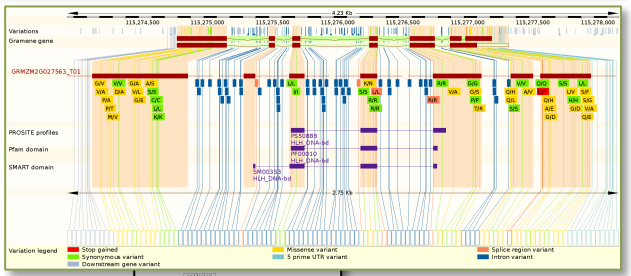
Pathway Curation & Projection

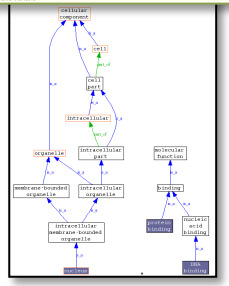
Programmatic Access:

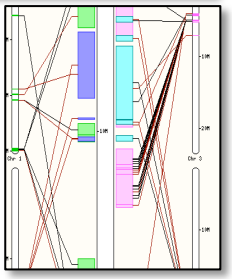
Gramene API
Ensembl API & RESTFUL interface
Reactome API & RESTFUL interface
BioMart
Public mysql server

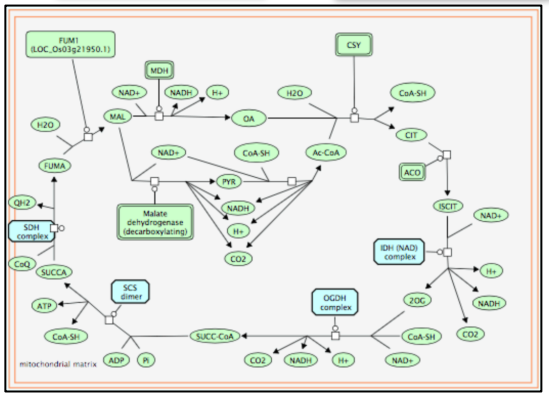
Graphical User Interface











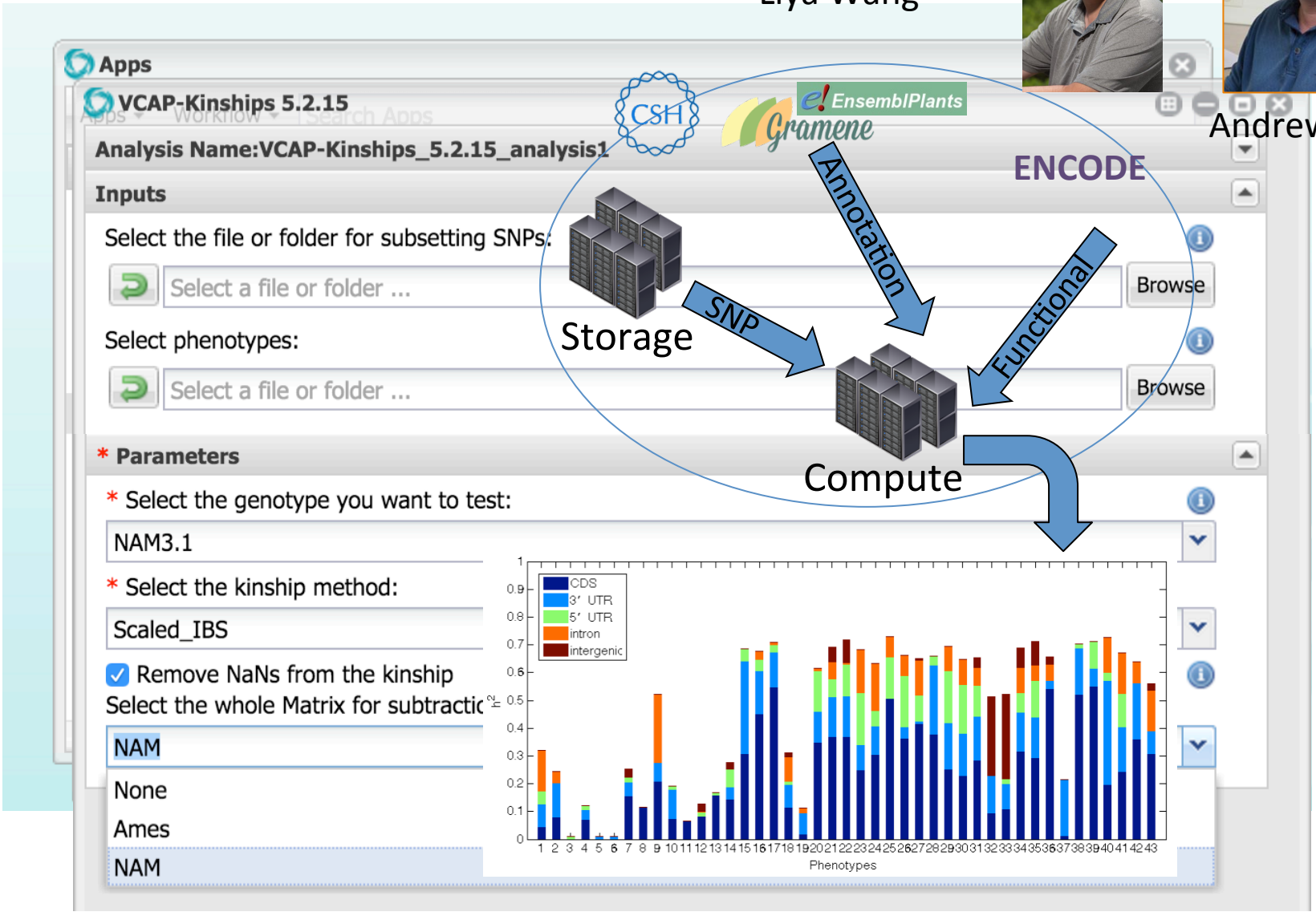
Gramene
www.gramene.org

EnsemblPlants

Heritability Testing Incorporating Genomic Features

Liya Wang

Andrew Olson

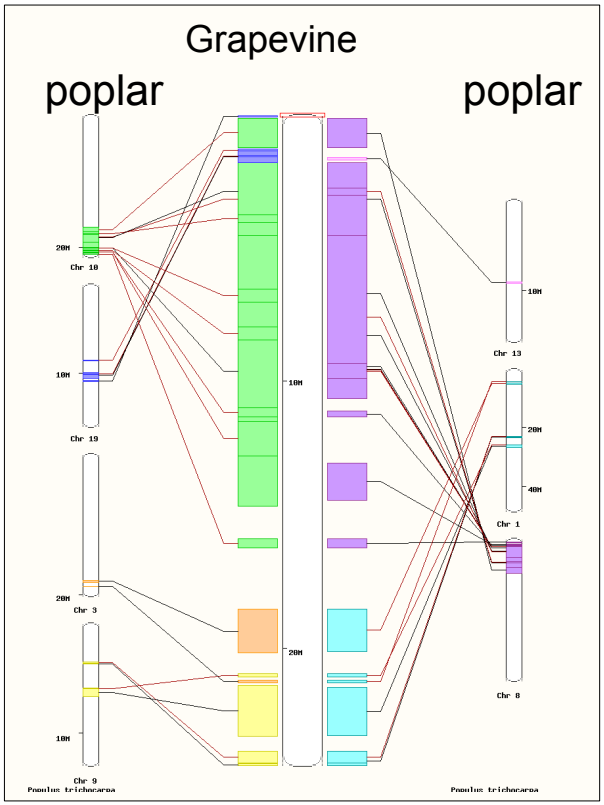


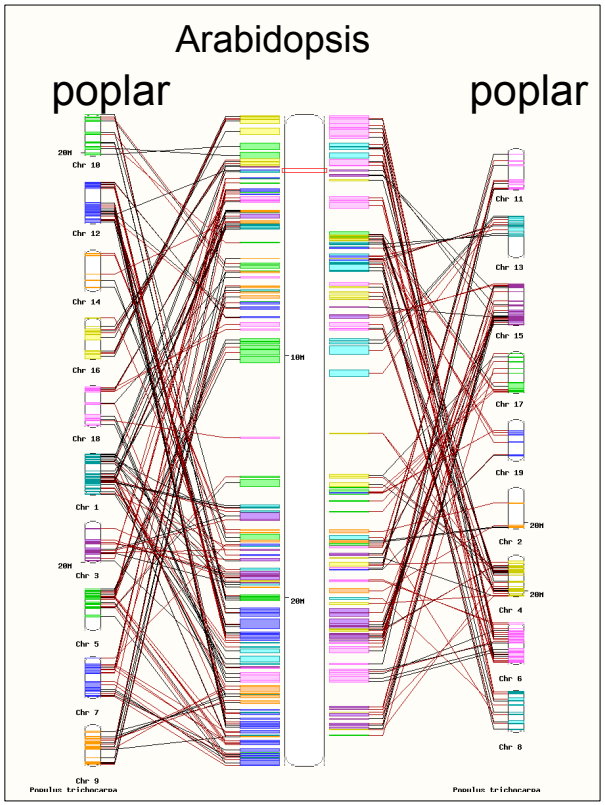
Comparative Genomics

- Increased use of grapevine as reference genome for Eudicot

Example: view whole genome duplication in Poplar; infer homoeologues

Support phenotype projections across species:
Seed size, seed number, fruit color, heat tolerance





Vitis vinifera

- Excellent evolutionary reference
- No whole genome duplication since Eudicot split

Arabidopsis thaliana

- Excellent for functional annotation
- 2 lineage-specific genome duplications/ reorganization

Gramene - Exploring Function through Comparative Genomics and Network Analysis

NSF IOS 1127112 (2012- 2017)

Transnational collaboration





Doreen Ware, PI (USDA ARS, CSHL)
Michael Campbell, Kapeel Chougule, Yinping Jiao, Sunita Kumari, Andrew Olson, Joshua Stein, Marcela K. Tello-Ruiz, Jim Thomason, Peter van Buren, Bo Wang, Sharon Wei

Pankaj Jaiswal, Co-PI (OSU)
Noor Al-Bader, Justin Elser, Matthew Geniza, Parul Gupta, Sushma Naithani, Justin Preece

Paul Kersey / **Robert Petryszyk** (EMBL-EBI)
Dan Bolser, Christopher Grabmuller, Chuang Kee Ong, Dan Staines, Brandon Walts / Maria Keays, Alfonso Muñoz-Pomer Fuentes, Laura Huerta Martínez

Lincoln Stein (OICR)
Peter D'Eustachio (NYU); Guanming Wu, Robin Haw, Joel Weiser, Sheldon McKay; Antonio Fabregat (EBI)

Crispin Taylor (ASPB) Patty Lockhart; Weijia Xu (TACC), Amit Gupta(TACC)

CSH Cold Spring Harbor Laboratory

EMBL-EBI

OICR Ontario Institute for Cancer Research

OSU Oregon State University

American Society of Plant Biologists

USDA

NSF

Gramene - Exploring Function through Comparative Genomics and Network Analysis

NSF IOS 1127112 (2012- 2017)

Transnational collaboration









Doreen Ware, PI (USDA ARS, CSHL)
Michael Campbell, Kapeel Chougule, Yinping Jiao, Sunita Kumari, Andrew Olson, Joshua Stein, Marcela K. Tello-Ruiz
Jim Tho

Pankaj
Noor Al
Sushma

Paul K
Dan Bo
Staines
Pomer

Lincoln
Peter D
Weiser,

Crispin
Amit Gu

- Partnering is not always easy, but absolutely necessary
- Sharing data and open data is not easy but is necessary and requires standards, policy, infrastructure and incentives
- Training and education are often unfunded, under appreciated

CSH Cold Spring Harbor Laboratory
EMBL-EBI
OICR Ontario Institute for Cancer Research
OSU Oregon State University
American Society of Plant Biologists
USDA
NSF

Biology Enabled Agriculture



Complex Traits: Development

Multiseed/ Branching

EMS population

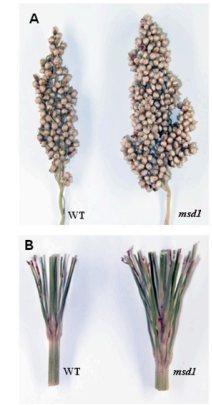
Genome Editing
Marker Assisted Breeding

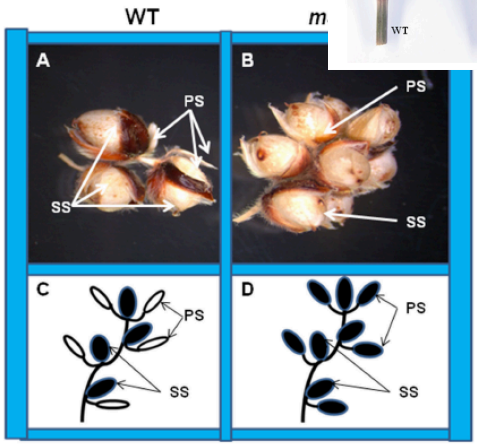
Transcription factors, Plant hormones

Expression

Regulatory Networks
Metabolic Pathways

Genomes





USDA-ARS, Lubbock T
Zhanguo Xin
Gloria Burow
Ratan Chopra
John Burke
Chad Hayes

Biology Enabled Agriculture

Complex Traits: Heat tolerance

Heat tolerance

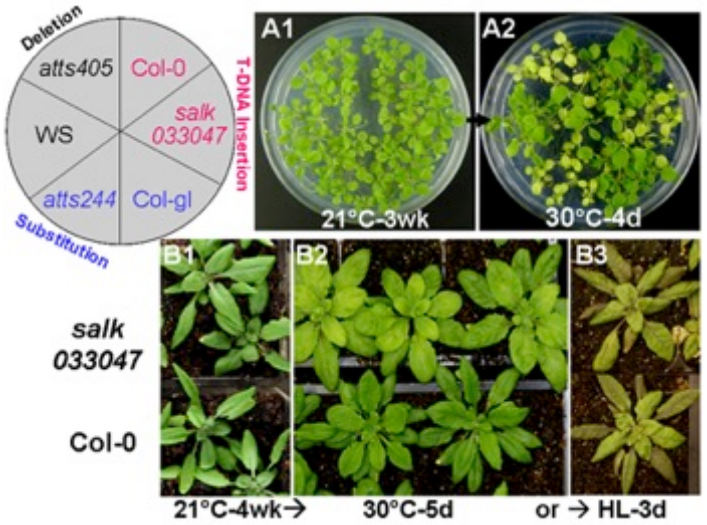
EMS population

ATP dependent protease

Expression

Regulatory Networks

Genomes



Ftsh11 identified in a model plant

Biology Enabled Agriculture

Complex Traits: Heat tolerance

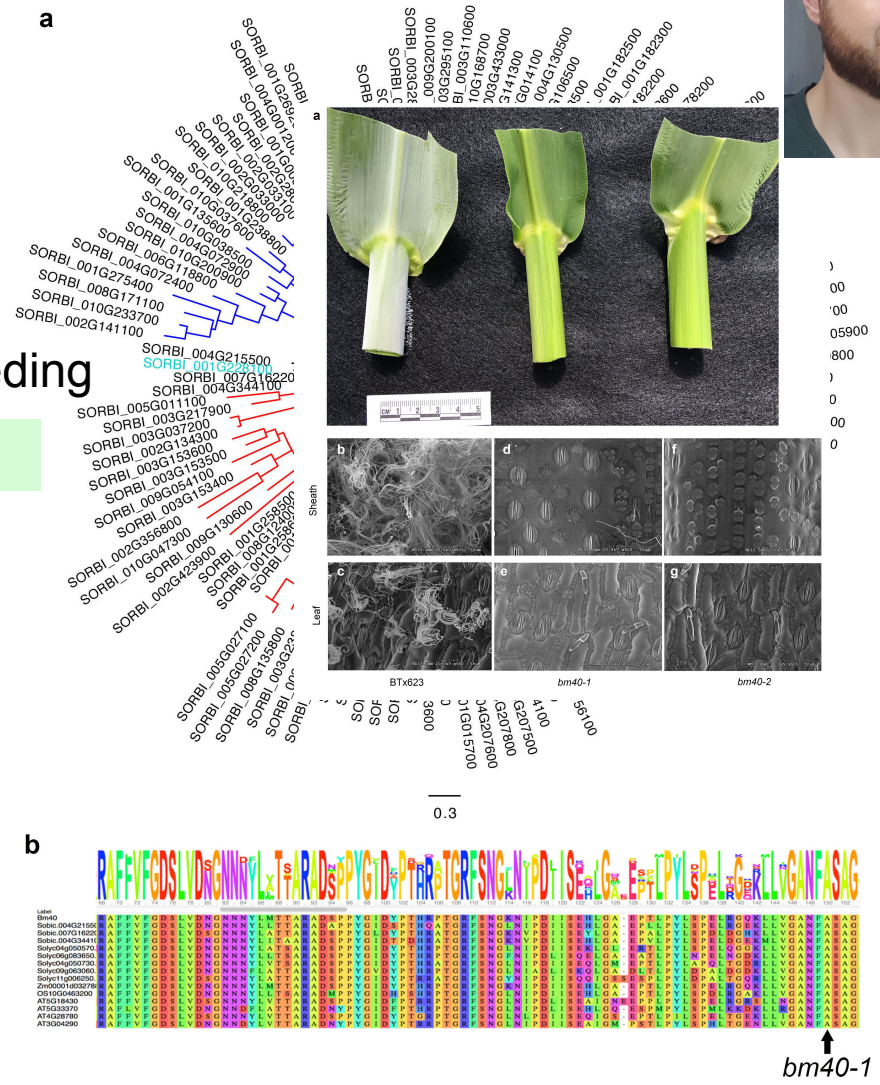
Heat tolerance

Genome Editing
Genomic Selection
Marker Assisted Breeding

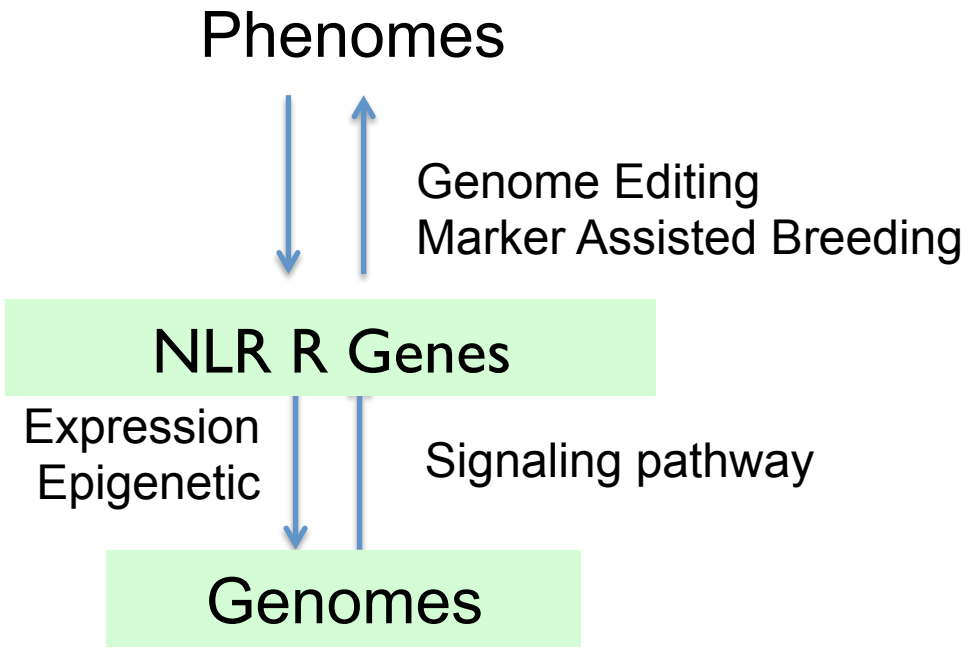
Long chain fatty acid, ATP dependent protease

Metabolic Pathways

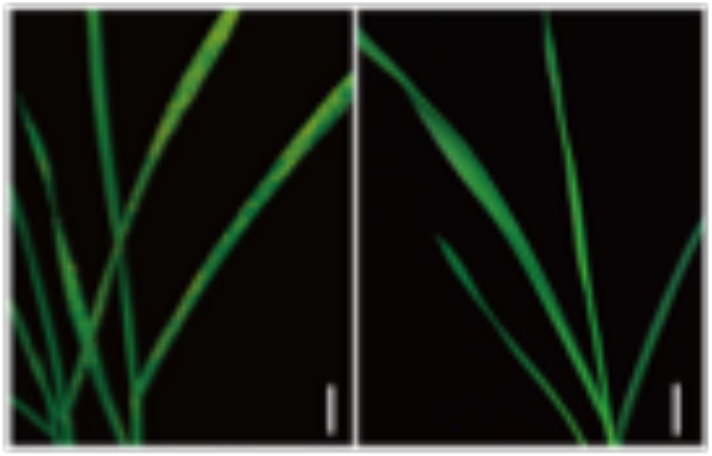
Genomes



Biology Enabled Agriculture

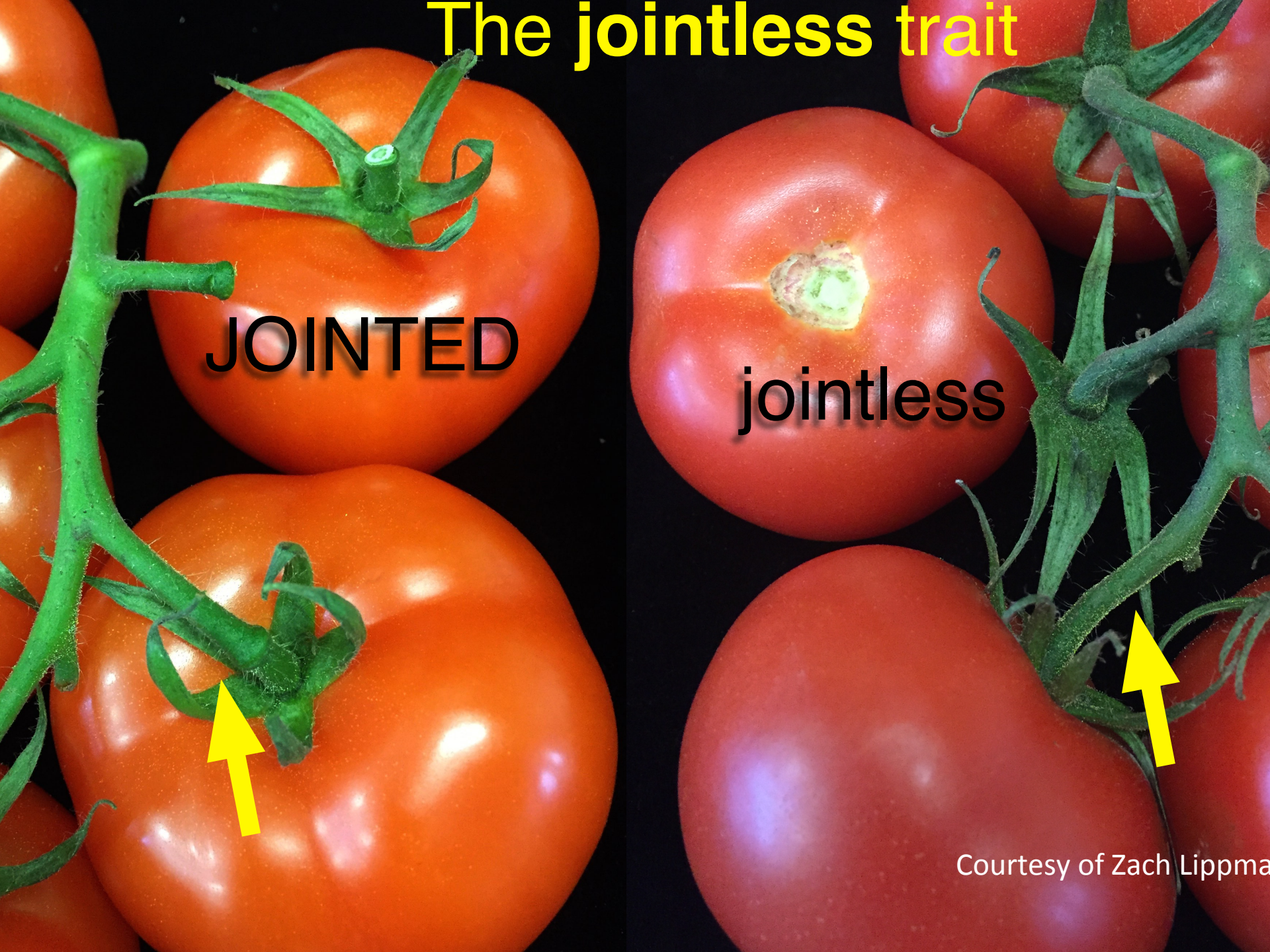


Disease Resistance



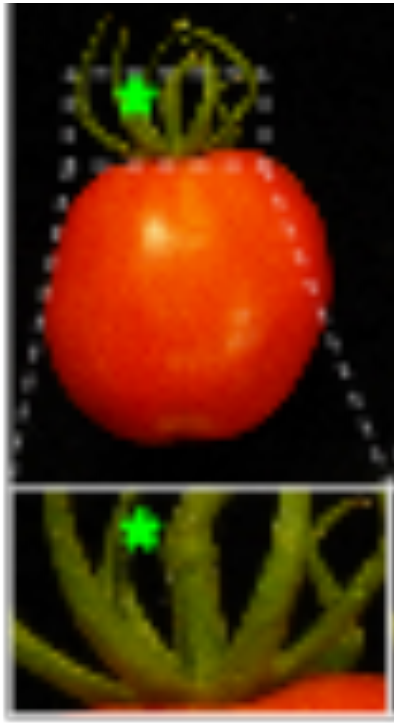
Nature has not given us enough Wang et al., 2014. Nat Biotech

CRISPR One-step Powdery Mildew resistance in Wheat

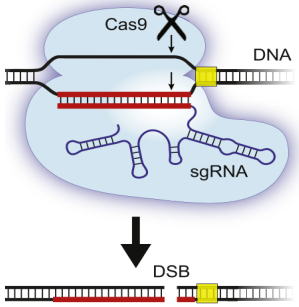

The jointless trait
JOINTED
jointless
Courtesy of Zach Lippma

ONE-STEP, ANY VARIETY!!

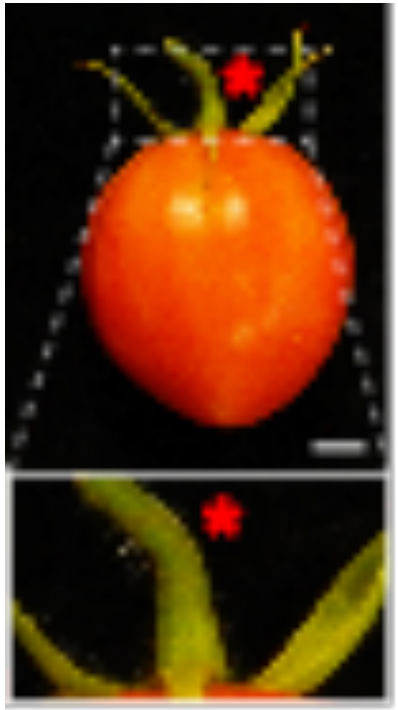
normal gene
normal activity



CRISPR

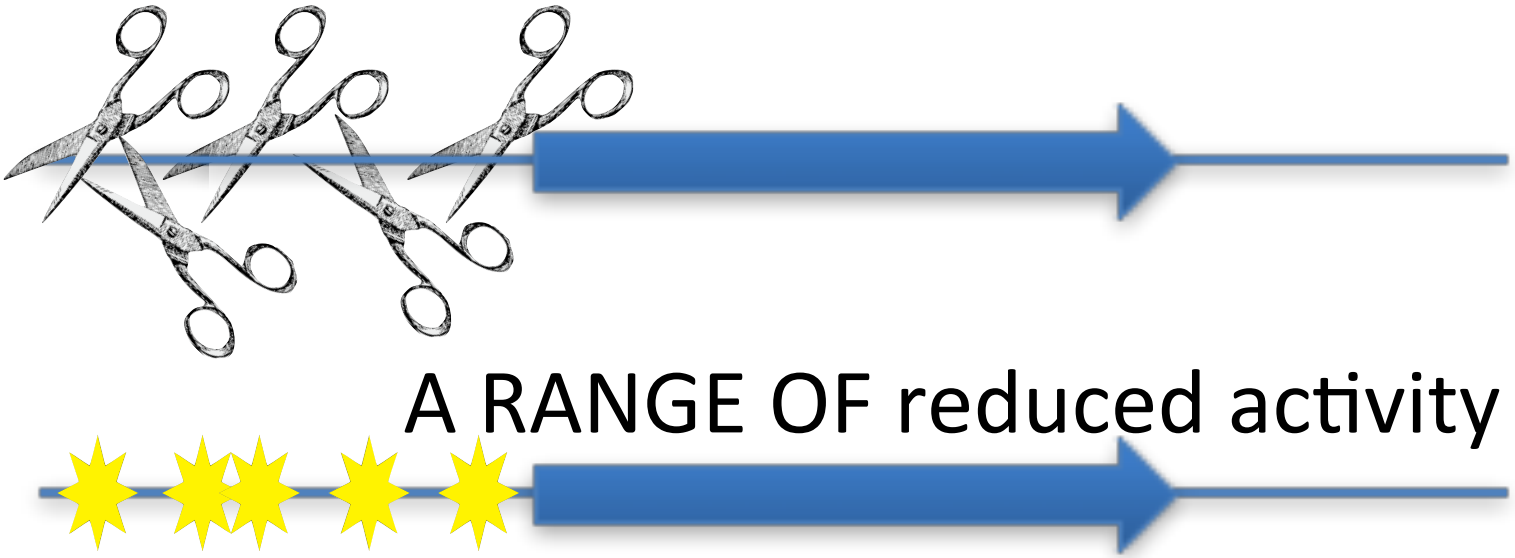


mutated gene
ZERO activity



Courtesy of Zach Lippman

DIRECTED mutation of "CONTROLLING REGION" of a gene



CRISPR has the power to enhance breeding by rapidly customizing and optimizing crop productivity

Genome Biology provides the insights to drive the translation

Technology Innovations Provide Both Opportunities and Challenges for Agriculture

Ware Lab
Mike Campbell
Kapeel Chougule
Nick Gladman
Carol Hu
Yinping Jiao
Vivek Kumar
Sunita Kumari
Young Koung Lee
Zhenyuan Lu
Dimitri Muna
Andrew Olson
Michael Regulski
Josh Stein
Jim Thomason
Peter Van Buren
Bo Wang
George Wang
Liya Wang
Sharon Wei
Lifang Zhang

CSHL
Dick McCombie
Sara Goodwin

USDA-Geneva
Lance Cadle-Davidson
Xia Xu
Jason Londo

Cornell
Qi Sun
Fred Gouker

USDA-ARS, Lubbock TX
Zhanguo Xin
Gloria Burow
Ratan Chopra
John Burke
Chad Hayes

Cinerea B9
Bruce Reisch
Paola Barba
Katie Hyma
Shanshan Yang,
Will Thompson

Flame Seedless
Craig Ledbetter
Rachel Naegele

Concord
Gan-Yuan Zhong

10X genomics
Stephen Williams
Deanna Church

Funding
USDA ARS
NSF
USDA-NIFA
California Table Grape Commission
National Grape and Wine Initiative

Advancing Agriculture Through Collaborative Research on Crop & Model Species