Agricultural Genomics: The Rise of the Genomes

Dr. Doreen Ware, USDA ARS
Agricultural Biotechnology: Emerging Technologies and Insights
January, 27, 2021

Advancing Agriculture Through Collaborative Research on Crop & Model Species
Outline

• Agricultural Drivers
• Maize genome 16 years
  – What’s in a genome
  – Improvements in Sequencing technology we can Continue to evaluate approaches to develop reference assembly and annotations
• Genome/Biology enabled agriculture
  – Sorghum EMS population
    • Forward and reverse genetics
  – Breaking down complex trait: Yeild & Quality
    • Plant architecture: flower
    • Response to environment: water, heat, nitrogen, disease
• Biology & “Big Data”
  – collaborative infrastructure
  – Future
Drivers for Agriculture: Sustainability and Defense

BREEDING FOR 2050 AND BEYOND. Prepare for plant and animal pests and disease while they are still offshore. Design plants for new environments.

CLIMATE CHANGE. Collect and preserve natural diversity.
Maize is the highest world-wide production crop

Reference genomes are foundation tools for ensuring food security & environmental sustainability
Acknowledgements

Maize Diveristy Project (NSF & USDA ARS) ~ 2003- present
USDA ARS Ed Buckler, Sherry Flint-Garcia, Mike McMullen, Jim Holland, Peter Bradbury, Doreen Ware Cornell QiSun, UC Davis Jeff Ross-Ibarra

Maize Genome Project (NSF, DOE, USDA) ~ 2005-2010
University of Washington, Rick Wilson, CSHL: Rob Martienssen, Dick McCombie, Doreen Ware, University of Arizona, Rod Wing, University of Iowa Pat Schnable

B73 Genome & Annotation Improvements V4 (NSF & USDA ARS) ~ 2015-2017
CSHL staff: Yinping Jiao, Bo Wang, Mike Campbell, Josh Stein, Sharon Wei, Doreen Ware, Dick McCombie, Eric Antinou
PacBio: David Rank, Paul Peluso, Jason Chin, Ting Hong, Elizabeth Tseng
BioNano: Alex Hestie, Tiffany Liang, Jinghua Shi. USDA ARS Mike McMullen, Kate Guill, University of Georgia: Kelly Dawe, Jonathan Gent, University of Hawaii: Gernot Presting, Kevin Schneider, Thomas Wolfgruber
Institutes: Cold Spring Harbor Laboratory, USDA ARS, Pacific BioSciences, BioNano, University of Georgia, University of Hawaii

Transcriptome Variation (USDA ARS) 2016-2019
CSHL staff: Yinping Jiao, Bo Wang, Mike Campbell, Josh Stein, Sharon Wei, Doreen Ware, Dick McCombie, Sara Goodwin
PacBio: Elizabeth Tseng, Tyson Clark, Ting Hong, Kevin Eng, Primo Baybayan
Institutes: Cold Spring Harbor Laboratory, USDA ARS, Pacific BioSciences

Maize Nested Association Mapping Panel Reference Assemblies (NSF & USDA ARS) 2018- present
University of Georgia: Kelly Dawe, University of Iowa: Matt Hufford University Minnesota Candy Hirsch
CSHL staff: Josh Stein, Kapeel Chougule Sharon Wei, Doreen Ware, Dick McCombie, Sara Goodwin
PacBio: Emily Hatas, Paul Peluso, Jason Chin, Cortiva: Victor Llaca, Kevin Fengler, Greg May, DNA Nexus: Brent Hannigan, Chai Fungtammasan, Brittanny O’Sullivan, NIH: Adam Phillippy, Serge Koren
Institutes: Cold Spring Harbor Laboratory, USDA ARS, University of Georgia, University of Iowa, Pacific BioSciences, DNANexus, NIH

Gramene www.gramene.org (NSF & USDA ARS) 2002- present NSF IOS-1127112
PI: Doreen Ware, PI (USDA ARS, CSHL) & Pankaj Jaiswal, Co-PI (OSU), Paul Kersey (Ensembl Genomes EBI), Helen Parkinson (ATLAS EBI), Lincoln Stein (Reactome OCIR), Crispin Taylor (ASPB)
Gramene @ CSHL Andrew Olson, Joshua Stein, Sharon Wei, : Marcela Karey Monaco,
Institutes: Cold Spring Harbor Laboratory, Oregon State University, EMBL – European Bioinformatics Institute, Ontario Institute for Cancer Research, American Society for Plant Biologists

USDA ARS Sorghum Functional Genomics and Germplasm improvement: 2015- present
USDA Doreen Ware, Zhanguo Xin, Chad Hayes, Yinghua Huang, Gloria Burow, Ratan Chopra, John Burke
SorghumBase: Nick Gladman, Yinping Jiao, Bo Wang, Kapeel Chougule Andrew Olson. Sharon Wei, : Marcela Karey Monaco Corteva,
Maize Genome is over 12 years old

Schnable, Ware et al. Science (2009)
Sequence genomes provides us the parts list and allows us to see what is the same or different between organisms.

Genes in corn, rice, and sorghum are in similar places in the genome.

_Schnable, Ware et al. Science (2009)_
Triceratops Genome

http://www.jurassicworld.com/creation-lab/
Genomes sequences allow us to see all the variations (mutations) that exist in nature

Letters or Single nucleotides polymorphisms (SNPs)

Gene content or parts list, known as copy number variations (CNVs)

Jumping genes, Transposable Elements (TEs) associated with the regulatory sequence between the Genes

Barbara McClintock
1983 Nobel Prize in Physiology & Medicine
Maize is a “Tale of Two Genomes”

- Ancestors (n = 10)
- Chromosome breakage and re-fusion
- Gene loss
- Return to 10 chromosomes (n = 10)
Evolutionary History of the Maize Genome

Ancestors

A

n = 10

X

B

n = 10

Chromosome breakage and re-fusion

AB

n = 20

Gene loss

C

n = 10

Return to 10 chromosomes

Red = “strong”, dominant genome

Blue = “weak”

Schnable, Ware et.al, Science 2009
Maize, also known as corn, experienced a whole genome duplication and then lost many of the genes. The genes that were kept by corn can tell us about how corn is adapting. Transcription factors, kinases, chromatin modifiers. Not all genomes have the same potential.

Schnable, Ware et.al, Science 2009
Humans Have Limited Molecular Diversity

1.34%

0.09%

Silent Diversity; Zhao, et al. (2001) PNAS
Maize diversity is greater than the difference between human and chimps

Silent Diversity; Tenallion, et al. (2001) PNAS
Individual Maize lines are very different from each other

The SNPs and gene differences affect how corn plants grow

Access to these sequences can accelerate the time it takes to make new lines of corn

Differences come from locally duplicated genes

Images Bo Wang
An additional copy of gene confers tolerance to acidic soil

Maron et. al, PNAS, 2013
Maize genomes are highly variable

- High rate of SNP and **structure variation** in the population
- Structure variations are highly **associated with phenotypic variation**
- **Structural variation** in non-coding region was enriched for phenotypic variation
- **One genome is not enough** to represent the diversity of the population

MAIZE DIVERSITY PROJECT
Meet the Family (2002 - present)

www.panzea.org

COURTESY SHERRY FLINT-GARCIA
26 Reference Assemblies for the Maize Nested Association Mapping (NAM) Population

Adapted from McMullen et al. Science 2009
Maize Nested Association Mapping (NAM) Population

Yu, et al. (2008) Genetics

25 DL

B97  CMI103  CMI228  CML247  CML277  CMI322  CML333  CML52  CML69  Hp301  Il14H  K111  K13  Ky21  M162W  M37W  Mo18W  M571  NC350  NC358  Oh43  Oh7B  P39  Tx303  Tz18

× B73

F1s

SSD

NAM

1

. . .

200

5000 RILs
26 Maize NAM founders reference assemblies (2019-2021)

Arun Seetharam -ISU
Margaret Woodhouse – MaizeGDB
Kapeel Chougule -CSHL
Shujun Ou -ISU
Jianing Liu -UGA
Xuehong Wei –CSHL
Zhenyuan Lu –CSHL
Andrew Olson –CSHL
Bo Wang -CSHL
Sharon Wei -CSHL
TingTing Guo -ISU
Rafael Della Coletta -UM
Xianran Li -ISU
John Portwood –MaizeGDB
Kevin Fengler -Corteva
Victor Llaca -Corteva
Amanda Gilbert -UM
Nancy Manchanda -ISU
Samantha Snodgrass -ISU
David Hufnagel -ISU
Sarah Pedersen -ISU
Michael Syring –ISU
Ethy Cannon - MaizeGDB
Carson Andorf -MaizeGDB
Jonathan Gent –UGA
Todd Michael - JCVI
Jianming Yu –ISU
Candice Hirsch -UM
Doreen Ware –CSHL
Matthew B. Hufford -ISU
R. Kelly Dawe -UGA

USDA United States Department Of Agriculture Agricultural Research Service
NSF

CORTeva
agriscience
New assemblies have a vast improvement in the contiguity of the sequence

N50 is the shortest contig length needed to cover 50% of the genome. -> Half of the genome sequence is covered by contigs larger than or equal the N50 contig size.

Matrices | Assembled Contigs
---|---
Total Bases in Assembly | 2,180,413,054
Contig Contiguity (NG50) | 52,409,415
Number of Contigs | 811
Longest Contig | 161,290,055
More than 100,000 genes found in the 26 maize accession

Hufford et al., bioRxiv, 2021
Maize Pan Gene Set

Candy Hirsch
U. Minnesota

Hufford et al, bioRxiv, 2021
Maize Pan Gene Set

Core genes are more likely to be Syntenic

Hufford et al, bioRxiv, 2021
The bulk of the genes in maize are found in other species

Hufford et al, bioRxiv, 2021
Genome contains life history of the species

Teosinte
Zea mays L. ssp. parviglumis

Landraces

Maize, Modern Inbreds
Zea mays L. ssp. mays
Biology Enabled Agriculture

G X (environment + management) = P

Complex Traits: Yield & Quality

Phenomes

- QTL
- GWAS
- Genome Editing
- Genomic Selection
- Marker Assisted Breeding

Gene Networks

- Metabolomics
- Expression
- Epigenetic
- Regulatory Networks
- Metabolic Pathways

Genomes
Profiling new sorghum genetic & phenotypic variation

- Parental line: BTX 623
- >10,000 individual M2 seed pools
- >6,400 M3 seeds obtained and Phenotyping is on-going and need to be expanded
- High quality DNA prepared for all lines

EMS Mutagenesis
- Random
- Single nucleotide change
- >99% GC → AT

Zhanguo Xin
Cropping Systems Research Lab, USDA-ARS, Lubbock TX

Jiao et al. The Plant Cell, 2016
Mutation Detection by whole genome sequencing of 256 mutants for forward genetics

- **Sequencing summary**
  - 20 M3 plants pooled together for sequence to averagely 16X
  - Average whole genome coverage – 86%; gene space coverage -95%

- **Quality control of the population:** 2 contamination lines + 2 sibling lines

- **Mutation detection:**
  - 1,862,560 EMS-induced mutations
  - Sanger sequencing validation rate >98%
  - 7,660 mutation/mutant = 1,798 homozygous + 5,862 heterozygous

- **Mutation Effect:**
  - 22% of mutations are located in genes, covering 95% of Sorghum genes
  - 57% (18,684) of the genes harbor >35,000 disruptive mutations, ~2 disruptive mutations per gene.

Jiao et al. The Plant Cell, 2016
Multiseeded (msd1, msd2, msd3)

Collaborator Zhanguo Xin
USDA ARS, Lubbock
Biology Enabled Agriculture

G X (environment + management) = P

Yield > Flower development > grain number > fertility/branching

MSD1
TCP transcription factor

Biology Enabled Agriculture

G X (environment + management) = P

Yield > Flower development > grain number > fertility/branching

EMS population

Transcription factor

Expression Metabolites

Genomes

Regulatory Networks

TCP transcription factor

Teosinte Branched

Doebley et al., Genetics (1995)
Secrets of Plant Genomes Revealed!

*Secrets of Plant Genomes: Revealed!* takes viewers on a lively, upbeat journey that explores how these plants got to be the way they are and investigates how we can make better use of them in the future. Plant scientists are hard at work—in the lab, in the field and at the computer—to increase our understanding of nature.

To download each of the parts, click the corresponding download button on the first player.

Introduction

Corn: The Dynamic Genome

Cotton: Building a Better Plant
Biology Enabled Agriculture

\[ G \times (\text{environment + management}) = P \]

\[
\text{Yield} \rightarrow \text{Flower development} \rightarrow \text{grain number} \rightarrow \text{fertility/branching}
\]

EMS population → Genome Editing → Marker Assisted Breeding

Transcription factor, Plant hormones

Expression Metabolites → Regulatory Networks → Metabolic Pathways

Genomes

Nitrogen, soil, and agricultural sustainability

Insufficient Nitrogen fertilizer

Excess Nitrogen fertilizer

Nitrogen deficiencies, LSU (courtesy, Dr. Brenda Tubana)

October 2011. Credit, USGS, NASA.
Biology Enabled Agriculture

$G \times (\text{environment + management}) = P$

Complex Traits: Yield $>$ Fitness $>$ Limiting Nitrogen

NUE

\[ \downarrow \quad \uparrow \]

Transcription factor

\[ \downarrow \quad \uparrow \]

Expression

\[ \downarrow \quad \uparrow \]

Genomes

\[ \downarrow \]

Metabolic Network

Lifang Zhang  Andrew Olson  Christophe Liseron-Monfils  Allie Gaudinier  Siobhan Brady
UC DAVIS  UC DAVIS

Guardinier et al, Nature 2017
Biology Enabled Agriculture

G X (environment +management) = P

Complex Traits: Yield > Fitness > Limiting Nitrogen

NUE

Transcription factor

Expression

Gene regulation

Regulatory Networks

Metabolic Network

Genomes

Lifer Zhang
Andrew Olson
Christophe Liseron-Monfils
Allie Gaudinier
UC DAVIS
Siobhan Brady
UC DAVIS

Guardinier et al, Nature 2017
Biology Enabled Agriculture

G X (environment + management) = P

Complex Traits: Yield > Fitness > Limiting Nitrogen

NUE

Transcription factor

Expression
Gene regulation

Regulatory Networks
Metabolic Network

tDNA population

Genomes

Sufficient N
10mM NO₃

Limiting N
1mM NO₃

Lifang Zhang
Andrew Olson
Christophe Liseron-Monfils
Allie Gaudinier
Siobhan Brady

UC DAVIS

Guardinier et al, Nature 2017
Biology Enabled Agriculture

\[ G \times (\text{environment +management}) = P \]

**Complex Traits:** Yield > Fitness > Limiting Nitrogen

NUE

\[ \downarrow \]

tDNA population

Transcription factor

Expression
Gene regulation

Regulatory Networks
Metabolic Network

Genomes

Sufficient N
10mM NO₃

Limiting N
1mM NO₃

Lifer Zhang
Andrew Olson
Christophe Liseron-Monfils
Allie Gaudinier
Siobhan Brady
UC DAVIS

Guardinier et al, Nature 2017
Changing climate increasing temperature and drought
Epicuticular wax (bloom) of sorghum plays important roles in tolerance of environmental stress.

<table>
<thead>
<tr>
<th>Gene in Arabidopsis</th>
<th>Sorghum gene</th>
<th>Amino acid change</th>
<th>Mutant Id</th>
</tr>
</thead>
<tbody>
<tr>
<td>CER6</td>
<td>Sobic.001G453200</td>
<td>E159K</td>
<td>ARS20</td>
</tr>
<tr>
<td>KCS12</td>
<td>Sobic.004G341300</td>
<td>R189C</td>
<td>ARS73</td>
</tr>
<tr>
<td>CER5</td>
<td>Sobic.009G083300</td>
<td>P581L</td>
<td>ARS73</td>
</tr>
<tr>
<td>CER5</td>
<td>Sobic.009G083300</td>
<td>L244F</td>
<td>ARS20</td>
</tr>
<tr>
<td>CER6</td>
<td>Sobic.006G020600</td>
<td>A133T</td>
<td>ARS205</td>
</tr>
<tr>
<td>KCS7</td>
<td>Sobic.002G268300</td>
<td>P449S</td>
<td>ARS31</td>
</tr>
<tr>
<td>KCS4</td>
<td>Sobic.002G268500</td>
<td>A49V</td>
<td>ARS31</td>
</tr>
<tr>
<td>KCS20</td>
<td>Sobic.005G168700</td>
<td>R303Q</td>
<td>ARS185</td>
</tr>
<tr>
<td>CER1</td>
<td>Sobic.001G222700</td>
<td>L100F</td>
<td>ARS185</td>
</tr>
</tbody>
</table>

Jiao et al. The Plant Cell, 2016
Biology Enabled Agriculture

G X (environment +management) = P

Complex Traits: Yeild > Stress > Heat tolerance

Heat tolerance

Fitness impact  ↓  EMS population

ATP dependent protease  ↓  Regulatory Networks

Genomes

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

Ftsh11 identified in a model plant
Biology Enabled Agriculture

G X (environment + management) = P

Complex Traits: Stress Resilience > Water Use efficiency

EMS population

Genome Editing
Marker Assisted Breeding

Long chain fatty acid, ATP dependent protease

Mass Spec

Metabolic Pathways

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

Jiao et al. Front. Plant Science 2018
Climate change impacts disease pressures

- Yield loss: Pesticide spraying increases direct cost and impacts to the environment, and human health

- Disease resistance "R"genes (NLR) are rapidly evolving and often seen in cluster. Good candidate for structural variation

- Pan Genomes: High quality reference assemblies to support characterization of core and dispensible (adaptive) genes

Sorghum Leaf spot
http://texassorghum.org/wp-content/uploads/2015/10/Fig.-2.jpg

Resistance to Southern Leaf Blight
Kump et al. Nat Genet 2011
Sugarcane aphids

• Since 2013 in the US sugarcane aphids have been causing enormous damage to sorghum crop

• Tx2783 has high resistance to SCA

Collaborators: Yinghua Huang (USDA-ARS), Zhanguo Xin (USDA-ARS), Chad Hayes (USDA-ARS)

https://www.myfields.info/pests/sugarcane-aphid

SCA locus mapped to the top of chromosome 6

Wang et al., bioRxiv, 2021
Sorghum sugarcane aphid tolerant TX2783 reference assembly

447 PacBio contigs (25.6 Mb contig N50)

Wang et al., bioRxiv, 2021
1. Tx2783 191KB insertion black box
2. All of the genes in the cluster are disease resistance genes
3. Complex region with local duplications
4. Inversion
Biology Enabled Agriculture

Yield > Biotic Stress > Disease Resistance

Genome Editing
Marker Assisted Breeding

NLR R Genes

Expression
Signaling pathway

Genomes

TX2783 191 kb structural variant (SV) containing a cluster of R Genes

Resequencing data from 62 accessions identified the SV is segregating at a low frequency in these lines.

Wang et al., *bioRxiv*, 2021

Yinpeng Jiao
Zhuanguo Xin
Bo Wang
Liya Wang
Exploring Genomes with an eye toward breeding

Biology Enabled Agriculture

Genomes  X (Environment +Management)  = Phenotypes

Association Analyses
Genes with major affect

Regulatory and Inference Networks
Many genes with minor contributions associated with a complex trait

Comparative analysis
Transfer information across species

Improved understanding of biological mechanisms & systems will improve breeding models and support for genetic engineering
Decreasing cost of sequencing leads to increasing computes and data management

$50 \text{ million (2009)}$ Sequencing Centers
- BAC library, Sanger sequencing library, finishing libraries, computes

$250-180 \text{ thousand (2016)}$ Sequencing Centers
- PacBio long single molecule, Optical map, illumina short read
- High quality DNA, Library prep, access sequencer & ***compute

$90-50 \text{ thousand (2017)}$ Sequencing Centers
- PacBio long single molecule, Optical map, 10X illumina short read
- High quality DNA, Library prep, access sequencer & ***compute

$45-25 \text{ thousand (2018)}$ Local/ Sequencing Centers

$2-6 \text{ thousand (2021)}$ Long single molecule

---

Schnable, Ware et al, Science 2009
Jiao et al., Nature, 2017
Ou et. al, Genome Biology 2020
Liu et al, Nature Comm 2020
Wang et. al, submitted 2021
Hufford et. al, submitted 2021
Biology has transitioned to an Information Science

Big Data” Biology Pyramid
Quantitative Biology Technologies

- **Results**
  - Domain Knowledge

- **Machine Learning**
  - classification, modeling, visualization & data Integration

- **Scalable Algorithms**
  - Streaming, Sampling, Indexing, Parallel

- **Compute Systems**
  - CPU, GPU, Distributed, Clouds, Workflows

- **IO Systems**
  - Hardrives, Networking, Databases, Compression, LIMS

- **Sensors & Metadata**
  - Sequencers, Microscopy, Imaging, Mass spec, Metadata & Ontologies

*Keystone Big Data in Biology 2014  Stein, Schatz, Ware*
Artificial Intelligence (AI) is allowing us to reimagine how we approach science.
Artificial Intelligence (AI) is allowing us to reimagine how we approach science

“AlphaFold is a once in a generation advance, predicting protein structures with incredible speed and precision. This leap forward demonstrates how computational methods are poised to transform research in biology and hold much promise for accelerating the drug discovery process.”

ARTHUR D. LEVINSON
PHD, FOUNDER & CEO CALICO, FORMER CHAIRMAN & CEO, GENENTECH
Wild Grapes have natural resistance to disease.
Wild Grapes come with baggage they have shed with domestication

Unpleasant taste

This, (2006); Feechan, et.al (2015)
**Plant NLR R-genes**

**Pathogen elicitor**

**Gene-for-gene hypothesis**

---

Imagine a world where you can begin to model protein ligand models for disease resistance genes!

Proteins are essential to life, supporting practically all its functions. They are large complex molecules, made up of chains of amino acids, and what a protein does largely depends on its unique 3D structure. Figuring out what shapes proteins fold into is known as the "protein folding problem", and has stood as a grand challenge in biology for the past 50 years. In a major scientific advance, the latest version of our AI system AlphaFold has been recognised as a solution to this grand challenge by the organisers of the biennial Critical Assessment of Protein Structure Prediction (CASP). This breakthrough demonstrates the impact AI can have on scientific discovery and its potential to dramatically accelerate progress in some of the most fundamental fields that explain and shape our world.

A protein’s shape is closely linked with its function, and the ability to predict this structure unlocks a greater understanding of what it does and how it works. Many of the world’s greatest challenges, like developing treatments for diseases or finding enzymes that break down industrial waste, are fundamentally tied to proteins and the role they play.
The Next Green Revolution will be “Data” and “Design” driven

Agriculture has transition to a Data Science. Massive data generation—genotypes, phenotypes, soil & environment.

Data now exceeds human capacity to formulate and test hypotheses about gene function, regulatory networks, and predictive agriculture.

Develop new approaches and systems that can supply new hypotheses for researchers for crop breeding, fermentation systems, solar energy capture, pest and disease management.

Adapt what has evolved in nature and design the space nature missed

Norman Borlaug
Nobel Peace Prize 1970
While CRISPR has the power to enhance breeding by rapidly customizing and optimizing crop productivity, Genome Biology provides the insights to drive the translation.
Unraveling Complex Traits using Genomic, Genetic, Systems & Computational Approaches

Advancing Agriculture Through Collaborative Research on Crop & Model Species