Molecular Plant Breeding: Challenges and Opportunities

Yunbi Xu

ccMaize

CAAS-CIMMYT Maize Molecular Breeding Laboratory

y.xu@cgiar.org
OUTLINE

- Molecular Plant Breeding: from Theory to Practice
- Breeding Platforms for G-P-E Typing
- Commercial Breeding and Breeding by Design
- Plant Breeding Driven by Big Data and AI
- Integrated and Open-source Plant Breeding
Population & demand are growing: we are **not** on-track for food security

Global average yield (tons per hectare)

- Wheat
- Maize

Anticipated demand by 2050 (FAO)

[Source: USDA PDS database]
Potential-increasing & Gap-closing

**Potential-increasing**

Unlocking and creating genetic variation
- Variation creation
- Variation coverage
- Functional diversity
- Gene discovery
- Gene transfer
- Genome editing

Improving heritability
- Envirotyping
- Error control
- GE interaction
- Agronomic practice

Heightening selection intensity
- Sample size
- Throughput
- Precision
- Cost

Shortening breeding cycle time
- MAS
- DH
- Integrated

**Gap-closing**

Yield gap 0
- For scientists to conceive and breed potential varieties

Nontransferable technology
- Environmental differences

Yield gap I

Yield gap II

Theoretical potential
Experimental Station yield
Potential Farm yield
Actual Farm yield

Molecular Plant Breeding Platform

Genotyping Platform
- Genotyping of Core Materials
  - Chip-based array
  - Genotyping by sequencing
  - Transcriptome and proteome analysis

Envirotyping Platform
- Water
- Light
- Temperature
- Fertilizer
- Soil

Phenotyping Platform
- Large-Scale Multi-Location Phenotyping
- Yield and quality
- Biotic and abiotic stresses
- Input use efficiency

Breeding Informatics Platform
- Information Collection, Management and Data Analysis
  - Genetic diversity analysis
  - Gene function analysis
  - GWAS/GS/MARS/MAS

Decision Support System
- Generation Advancement + GS/MARS/MAS
- Multi-Environmental Trials
- Novel Germplasm

Revised from Xu et al 2012 Mol Breed 29:833–854
Marker-assisted selection in plant breeding: from publications to practice

Yunbi Xu and Jonathan H. Crouch

*Crop Science 48: 391-407 (2008)*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>IF</td>
<td>1.15</td>
<td>1.15</td>
<td>1.33</td>
<td>1.74</td>
<td><strong>2.02</strong></td>
<td>1.64</td>
<td>1.51</td>
<td>1.48</td>
<td>1.58</td>
<td>1.55</td>
<td>1.63</td>
<td>1.64</td>
<td>1.64</td>
</tr>
</tbody>
</table>

Where are we now?

Google Citation = 731 (12 Nov. 2019)
Major Constraints in Marker-assisted Plant Breeding

Cost - High-efficient and cost-effective platforms
50-90 USD per sample for 55K-SNP chip: each seed company may have 50K to 80K USD budget for genotyping, which can be used for genotyping 1000 samples = 1-3 breeding populations. In this case, breeders may prefer to genotyping their core materials for germplasm evaluation cheaper than phenotyping

Platforms
Genotyping & information platforms: Mol Breeding, Huazhi Rice, Golden Marker, Beijing Compass Biotechnology

Targets - Reducing the cost below phenotyping (10 USD/sample)
The research budget from each company can be used for MAS of 10 or more populations/year
Challenge 1
Transition from basic research to application

Many reports with significant yield increase (up to 30%) but for
- Specific genetic backgrounds (cannot be generalized to other backgrounds)
- Experiments in greenhouses or pots rather than in the field
- Phenotyping based on single plants instead of populations
- Comparison with specific checks or controls (usually wild types instead of commercial varieties)

Application involves wide genetic backgrounds, complex gene by gene and gene by environment interactions, stressful environments, and increasingly improving commercial varieties.
Challenge 2
Transition from simple traits to complex traits

It is very difficult to transfer or pyramid several favorable alleles simultaneously
Challenge 3
Transition from single markers to haplotypes

Challenge 4
Transition from associated markers to functional markers

Associated or linked markers
Changes with populations and generations
Validation required before application

Functional markers
Gene isolation and mining
Identification of allelic variation
Functional study
Challenge 5
Transition from simple populations to complex populations

Challenge 6
Transition from examples to wide application

- Standardization
- Reduced costs
- High-throughput
- Automated processes
- Fast and speed processing and delivery
- Shared platforms, facilities, tools and information
- Market-driven (no support by government funds)
OUTLINE

- Molecular Plant Breeding: from Theory to Practice
- Breeding Platforms for G-P-E Typing
- Commercial Breeding and Breeding by Design
- Plant Breeding Driven by Big Data and AI
- Integrated and Open-source Plant Breeding
Three Typing Technologies in Molecular Breeding

G  Geno + typing = Genotyping  基因型鉴定

P  Pheno + typing = Phenotyping  表现型鉴定

E  Environmental + typing = Envirotyping  环境型鉴定

Envirotyping ≈ Environmental Assay

Xu et al. 2012, Mol Breed 29:833–854
Evolution of Genotyping (1980-2010s)

Systems
From gels to chips and sequencing (GBS)

Throughput
From singles to millions

Resolution
10-30 cM to many markers per gene

Cost (per data point)
Several US dollars to 1/1000 cent
# Evolution of Genotyping Platform

**Yunbi Xu (2018)**

<table>
<thead>
<tr>
<th>G1</th>
<th>G2</th>
<th>G3</th>
<th>G4</th>
<th>G5 (final)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gel-base</td>
<td>Gel-base</td>
<td>Solid chip</td>
<td>Liquid chip</td>
<td>Sequencing</td>
</tr>
<tr>
<td>1980s</td>
<td>1990s</td>
<td>2000s</td>
<td>2010s</td>
<td>2020s</td>
</tr>
</tbody>
</table>

| Manual   | Semi-automatic    | Automatic        | Automatic        | Highly automatic |
| Exp. station | Exp. station | Workstation | Workstation | Carry-on |
| Very high-cost | High cost | Low cost | Very low-cost | Almost free-charge |
| Not flexible | Little flexible | Less-flexible | Very flexible | Extremely flexible |
| 100DP/D  | 1000DP/D          | 1M^+DP/D         | 1M^+DP/D         | 1B^+DP/D    |
Final Genotyping Platform: 5G

Automatic sequencing

- Automatic genotyping
- Supra-low cost
- Long reads (to cover the whole chromosome
- Automatic information management
- Integrated with artificial intelligence and robots
- Carry-on

http://www.sohu.com/a/215470854_656712
Genomic Gaps Left by Reference Genomes

- Single genomes that have been sequenced up to 80-90% are used as reference genomes
- In most cases, only one genotype has been sequenced with relatively high resolution, usually representing a domesticated elite variety
- Resequencing indicates that 20-50% of the original reads from different ecotypes cannot be mapped to the reference genome

Maize

50% Hi-seq reads from tropical maize cannot be mapped; while only 20% of the SNPs from landraces can be mapped to the B73 reference (Peter Wenzel, CIMMYT)

Rice

75% of the rice germplasm are indica, and 15-20% of their sequence reads cannot be mapped to the Japonica reference (Kenneth McNally, IRRI)
The Results of Partial Genome Coverage

Single-genome based references provide only a partial genome coverage for a crop species

- Got lost in map-based cloning
- Missing of 40% or more important QTL/genes in AM
- Biased estimation (Ascertainment bias)
  - Genetic diversity
  - Population structure
  - LD and IBD
  - Haplotypes
- MAS
  - Inefficient procedures
  - Unpredictable results

Multiple genome-based references are needed for whole genome strategies
Genotyping by Sequencing Strategies

Whole genome de novo sequencing
It is too expensive to sequence everything at this stage

Genotyping by random sequencing (skim-based GBS)
Highly-demanding for large-scale sequence data availability
Strong bioinformatics support

Genotyping by target sequencing (Skim-based GBS)
GenoPlexs: Multiplex PCR for several to 5K markers
GenoBaits: In-solution capture for 1K to 40 K markers

Multiple sets of SNP markers developed by Genotyping by target sequencing (GBTS)

Developed by CIMMYT-China, Institute of Crop Science (CAAS) and Mol Breeding Co.

- Selection and development of 20K SNP markers from the 55K SNP array with improved genome coverage
- Target sequences are enriched by in-solution probes
- Using the same panel of 20K SNP markers to generate 10K, 5K and 1K SNP markers by sequencing at different depths
- Test and validation using two genotype panels:
  - 96 diverse maize germplasm from China, USA and CIMMYT
  - 387 breeding lines generated in CAAS maize breeding programs
- Suitability: 50-40000 SNPs, SSRs or InDels

The system can be used for all organisms (plants and animals)

mSNP and GBTS Markers

mSNP markers
- Multiple SNPs from a single traditional SNP region or one amplicon. Several SNPs are closely linked together

**40K mSNP markers in maize**
- 40K mSNP = 260 K SNPs
- 1 mSNP contains more than 6 SNPs

**1-set GBTS markers for all purposes => multiple sets of SNPs**
Just by sequencing at different depths (at different costs !!!)
- 20K SNPs = > 20K, 10K, 5K and 1K SNPs (Guo et al 2019. Mol Breed 39: 37)
- 40K mSNPs = > 40K, 30K, 20K, 10K, 5K and 1K mSNPs
- 260k SNPs = > 260K, 100K, 50K, 20K, 10K, 5K, 1K SNPs
### Basic Statistics for 40K mSNP Markers from Genotyping by Target Sequencing in Maize

**T – Theory** | **R – Realized**
---|---
Chr | mSNP# | SNP# | Haplotype \(^{(T)}\) | # Haplotype \(^{(R)}\) | SNP/mSNP | Haplotype/mSNP \(^{(T)}\) | Haplotype/mSNP \(^{(R)}\) | Amplicon length (bp)
---|---|---|---|---|---|---|---|---
1 | 6,000 | 39,464 | 5,402,520 | 125,398 | 6.577 | 900.420 | 20.900 | 101.073
2 | 4,676 | 31,280 | 6,748,614 | 102,113 | 6.689 | 1443.245 | 21.838 | 101.654
5 | 4,346 | 28,420 | 6,372,332 | 90,834 | 6.539 | 1466.252 | 20.901 | 101.567
6 | 3,217 | 21,250 | 5,159,326 | 68,036 | 6.606 | 1603.769 | 21.149 | 101.211
7 | 3,389 | 22,353 | 11,895,266 | 140,627 | 6.596 | 3509.963 | 41.495 | 101.596
8 | 3,352 | 21,596 | 4,154,560 | 69,206 | 6.443 | 1239.427 | 20.646 | 101.609
**Total** | 40,000 | 262,958 | 58,045,194 | 912,077 | 6.574 | 1451.130 | 23.107 | 101.444

---

Zifeng Guo, Jianan Zhang, Yunbi Xu et al., in preparation
### Applications of Genotyping by Target Sequencing

<table>
<thead>
<tr>
<th>Applications</th>
<th>40K+</th>
<th>20K</th>
<th>10K</th>
<th>5K</th>
<th>1K</th>
<th>&lt;200</th>
</tr>
</thead>
<tbody>
<tr>
<td>Germplasm evaluation</td>
<td>+++</td>
<td>+++</td>
<td>++</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Classification</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>++</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td>Map construction</td>
<td>+++</td>
<td>+++</td>
<td>++</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Gene mapping/cloning</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>++</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Marker-trait association</td>
<td>+++</td>
<td>++</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Progeny testing</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td>Gene introgression</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>++</td>
<td>+</td>
</tr>
<tr>
<td>Gene pyramiding</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>++</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Plant variety protection</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>++</td>
<td>++</td>
<td>+</td>
</tr>
<tr>
<td>Quality control</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>++</td>
<td>++</td>
<td>+</td>
</tr>
<tr>
<td>Bioassay</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>++</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

Revised from Guo et al 2019. Mol Breed 39: 37
Phenotyping is king

... determining how a genotype associates with its function

Phenotyping determines

- Foot-on-the-ground for all types of omics
  - Useless without association with phenotype
- Function of associated traits
  - Functional genomics
  - Gene identification
  - Marker-trait association (linkage mapping, GWAS)
- Crop improvement procedures
  - Selection
  - Understanding G X E interaction
Design influencing factors for plant phenotyping automation solutions (simplified)

<table>
<thead>
<tr>
<th>Research type</th>
<th>Application</th>
<th>Environment</th>
<th>Plant parts</th>
<th>Crop</th>
<th>Growing medium</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic research</td>
<td>Functional genomics</td>
<td>Laboratory</td>
<td>Seedlings</td>
<td>Field row crops</td>
<td>Micro titer plates</td>
</tr>
<tr>
<td>Applied research</td>
<td>Plant physiology</td>
<td>Research greenhouse</td>
<td>Full plants</td>
<td>Greenhouse crops</td>
<td>Petri dishes</td>
</tr>
<tr>
<td>Screening process</td>
<td>Root phenotyping</td>
<td>Reach-in-chamber</td>
<td>Population</td>
<td>Fruit crops</td>
<td>Trays</td>
</tr>
<tr>
<td></td>
<td>Stress responses</td>
<td>Walk-in-chamber</td>
<td>Fruits pre-harvest</td>
<td>Biofuel crops</td>
<td>Pots</td>
</tr>
<tr>
<td></td>
<td>Treatment responses</td>
<td>Multi-tier chamber</td>
<td>Fruits post-harvest</td>
<td>Vegetables</td>
<td>Tubes</td>
</tr>
<tr>
<td></td>
<td>Product characterization</td>
<td>Production greenhouse</td>
<td>Roots</td>
<td></td>
<td>Rockwool</td>
</tr>
<tr>
<td></td>
<td>Seed phenotyping</td>
<td></td>
<td></td>
<td></td>
<td>In soil</td>
</tr>
<tr>
<td></td>
<td>Germination evaluation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Crop development</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yield prediction</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Crop performance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Harvested product characterization</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Harvested product phenotyping</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Harvested product sorting</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Harvested product quality monitoring</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shelf life</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Revised from Zhiguo Han (2017)

Other ...
Screening numbers + Available budgets + User experience level + Existing infrastructures + Specific objectives + (Client) culture aspects
Greenhouse High-throughput Phenotyping Platform

Plant-to-sensor or Sensor-to-plant

Plant-to-Sensor
- Very detailed measurement
- Different side of view
- Different imaging cameras
- Root phenotyping

Sensor-to-Plant
- Plant do not need move
- High density
- Plant interactions
- More efficient using greenhouse

- Do plant like move?
- Low density

Zhiguo Han (2017)
Field High-throughput Phenotyping Platforms

Sensor-to-plant

Gantry platform
- Sensor installed in Gantry
- Moveable in a certain area (200m x 20m, for example)
- Fully automatic
- High resolution
- Repeated measurement every day

Spidercam platform
- Cable suspended system for field phenotyping
- Large scale (more than 1ha)
- Customized size and sensors
- Manual, automatical or remote control
- High positioning accuracy for time series measurement

Tractor or UGV platform
- Sensor installed in tractor or UGV
- Moveable in a large area (speed 4 km/h, for example)
- Half or fully automatic
- Less resolution
- For large area

UAV platform
- Sensor installed UAV
- Very high speed
- Half or fully automatic
- Much less resolution
- For huge area
- Batteries only last 20 min

Resolution, repeatable

Unmanned ground vehicle (UGV)
Unmanned Aerial Vehicle (UAV)
Continuum of environments for drought-resistance screening

Araus and Cairns 2014

TRENDS in Plant Science
Environmental factors have been treated as a black box. Even if we do not know what is inside the box, we can use it to fill up with:

- Dreams and questions
- Justifications and imaginations
- Facts and doubts
- Current status and future prospects

G X E interaction has been used to explain:

- Anything we cannot explain
- Genotypes showing different phenotypes under different environments
- Different results even due to experimental errors and even mistakes

Trash can
In the era of phenomics

**Envirotyping is queen**

... determining how a phenotype we see reflects its genotype

**Environments determine**

  - Where phenotyping is done
  - Under what specific environments
  - How much they are different from natural conditions
  - How much we can trust on the phenomic data and results
  - Fully, partially, not at all

  - What we can do with phenomic data
  - fun, publication or production

=> Precision phenotyping needs to be coupled with precision envirotyping
Envirotyping: Concept Development

Xu, Y. 2011. From line to space: a 3-D profile of molecular plant breeding. The First Congress of Cereal Biotechnology and Breeding, May 23-27, 2011, Szeged, Hungary (E-typing or environmental assay)

Xu, Y. 2012. Environmental assaying or etyping as a key component for integrated plant breeding platform. Workshop 16: Marker-Assisted Selection, 6th International Crop Science Congress, August 6-10, 2012, Bento Gonçalves, RS, Brazil


Abstract

A bandwagon is an idea, activity, or cause that becomes increasingly fashionable as more and more people adopt it.

... ...

This article reviews and speculates about post-1990 bandwagons in plant improvement, including transgenic cultivars, quantitative trait locus (QTL) mapping, association mapping, genomewide (or genomic) selection, phenomics, envirotyping, and genome editing.
Applications of envirotypic information

A Classification of environments

B Near-iso environments

C Control of experimental error

D Selection of experimental sites

E Agronomic genomics

F Stress prediction

G Precision agriculture
Envirotyping needs to zoom into the plots and individual plants

An example can be found from Masuka et al 2012 Cell

Xu 2016 Theor Appl Genet 129:653–673
Netherlands Plant Eco-phenotyping Centre, NPEC

Housed by Wageningen and Utrecht and co-funded by The Netherlands Organisation for Scientific Research (NWO) for 10 years with 11m euros and total costs up to 22m euros

1. Precision mesocosm -level ECOtron (ECO) plant-plant and plant-microbe

2. Plant-Microbe Interactions phenotyping module (PMI)

3. Multi-Environment climate chamber module (ME)

4. The High-Throughput Phenotyping climate chamber module (HTP)

5. GreenHouse phenotyping (GH) module

6. Open-Field phenotyping (OF) module

April 12, 2018
https://www.wur.nl
https://www.uu.nl
Prepared by Yunbi Xu

Six complementary, experimental modules
China National Phenomics Platform
Shennong (Holy Farmer) Facilities

Grant opening on 27 July 2019, Wuhan
OUTLINE

- Molecular Plant Breeding: from Theory to Practice
- Breeding Platforms for G-P-E Typing
- Commercial Breeding and Breeding by Design
- Plant Breeding Driven by Big Data and AI
- Integrated and Open-source Plant Breeding
Commercial Breeding Pipeline
Structure of Monsanto’s North America corn breeding and breeding technology organization

Partitioned the process into components to enable optimal execution

Schematic of the changes in scale of testing as a maize breeding program moves inbreds from early stages to later stages of hybrid evaluation within a cycle of selection.

Breeding for Facility Agriculture

Crop production: from the field to fully controlled facilities

- Subject to environmental conditions & fluctuations
- Inconsistent growing
- Requires pesticides
- Seasonal operation

- Variable sunlight, temperature, and seasonal conditions
- Year-round operation
- Subject to pests
- Specialized structure
- Water use: reduced 8X

- Consistent yield & quality
- No pest, no pesticides
- Existing structures
- Easily stacked
- Water use: reduced 3X beyond greenhouse

Yunbi Xu (2018) unpublished
Facility Agriculture and Breeding

The Netherlands and Israel: how do the tiny countries feed the world. Greenhouse facilities create the miracle:

- High yielding
- Improved resource use efficiency for all inputs
- Reduced pollution and environmental effects
- Improved quality (consistency, clean, large-scale, timing, off season)

Requirements: quality water and air, sewage disposal, automation, monitoring, and AI.

Breeding: breeding to meet the demands and specific conditions

Yunbi Xu (2018) unpublished
Integrated Product Pipeline for Crops

- Seed companies hope farmers to plant more seeds
- Fertilizer salesmen want farmers to apply more fertilizers
- Chemical companies push farmers to use more chemicals

Nobody really cares about the income and output/input ratio for farmers
Breeding for Mechanization Production

- Increasing scale
- Reducing cost
- Fitting for variety-agronomy-management system
- Meeting the demands of product pipeline
  - Production
  - Storage
  - Processing
  - Marketing
Taking 20 years (1988-2007) to transfer hybrid rice from manual production to mechanization (RiceTec’s experience)

Quality
- Milling quality: whole milling increases from 45% to 60%
- Grain quality: amylose content decreases from over 22-25% to 15-18%

Grain production
- Transplanting to direct seeding

Seed production
- Manual to mechanized
Breeding by Design at different tiers

**Gene design** 基因设计
- Gene knockout, mutation and editing; RNAi
- Single genes: best alleles
- Multiple genes: Best combinations/haplotypes/interactions

**Metabolic pathway design** 代谢途径设计
- Systems biology and breeding
- Metabolic pathway modification and improvement
  - C4 rice
  - High photosynthetic capacity wheat

**Gene network design** 基因网络设计
- Network regulation factors
- Network structure
- Network nodes and edges
Breeding by Design at levels of individuals, populations and species

Trait design Coordination and interaction among traits; Coordination and compensation between sink and source

Morphological design Maximized yield, solar energy utilization, assimilate partitioning, sink-source coordination, population structure, yield stability, stress tolerance, suitability for machine harvest

Population design Optimized population, stabilization, high-efficiency, interaction among individuals and within population

Species design Integrated novel traits from different species; New agricultural crops adapted to different environments and ecological zones

Environment-friendly; Resource-use efficiency; Diverse products; Usage flexibility

Perennial grains (多年生禾谷类作物)
Rain-fed, direct-seeding rice (直播水旱两用稻)
Re-domestication of potato (马铃薯主粮化 黄三文等)
分子模块设计育种
Breeding by molecular modules design

分子模块解析
Molecular Module Dissection

模块耦合
Module Integration

设计型新品种
Designed new varieties
Molecular Plant Breeding: from Theory to Practice

Breeding Platforms for G-P-E Typing

Commercial Breeding and Breeding by Design

Plant Breeding Driven by Big Data and AI

Integrated and Open-source Plant Breeding
What is big data?

- Big volume: TB; PB; EB; ZB; billions of data points?
- Stored with non-traditional databases?
- Using data platforms such as Hadoop/Spark?
- Parallel running with multiple machines?

Many Vs for big data

Big Data is often described using the five Vs:
- Volume, Velocity, Variety, Veracity, Value

https://blog.csdn.net/allenlu2008/article/details/79603476
## Volume: Scale of Data

<table>
<thead>
<tr>
<th>Unit</th>
<th>Value</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>bit (b)</td>
<td>0 or 1</td>
<td>1/8 of a byte</td>
</tr>
<tr>
<td>byte (B)</td>
<td>8</td>
<td>1 byte</td>
</tr>
<tr>
<td>kilobyte (KB)</td>
<td>$1000^1$</td>
<td>1,000 bytes</td>
</tr>
<tr>
<td>megabyte (MB)</td>
<td>$1000^2$</td>
<td>1,000,000 bytes</td>
</tr>
<tr>
<td>gigabyte (GB)</td>
<td>$1000^3$</td>
<td>1,000,000,000 bytes</td>
</tr>
<tr>
<td>terabyte (TB)</td>
<td>$1000^4$</td>
<td>1,000,000,000,000 bytes</td>
</tr>
<tr>
<td>petabyte (PB)</td>
<td>$1000^5$</td>
<td>1,000,000,000,000,000 bytes</td>
</tr>
<tr>
<td>exabyte (EB)</td>
<td>$1000^6$</td>
<td>1,000,000,000,000,000,000,000 bytes</td>
</tr>
<tr>
<td>zettabyte (ZB)</td>
<td>$1000^7$</td>
<td>1,000,000,000,000,000,000,000,000 bytes</td>
</tr>
<tr>
<td>yottabyte (YB)</td>
<td>$1000^8$</td>
<td>1,000,000,000,000,000,000,000,000,000 bytes</td>
</tr>
<tr>
<td>brontobyte (BB)</td>
<td>$1000^9$</td>
<td>1,000,000,000,000,000,000,000,000,000,000 bytes</td>
</tr>
<tr>
<td>nonabyte (NB)</td>
<td>$1000^{10}$</td>
<td>1,000,000,000,000,000,000,000,000,000,000,000 bytes</td>
</tr>
<tr>
<td>doggabyte (DB)</td>
<td>$1000^{11}$</td>
<td>1,000,000,000,000,000,000,000,000,000,000,000,000 bytes</td>
</tr>
</tbody>
</table>

All information in the world currently available = 7 ZB
Plant breeding data: Multiple sources

Multi-omics data
Multi-phenotypes
Multi-environments
Integrated data
Empirical breeding criteria
Selection indices
Mating design
Combining ability
Heterotic patterns
Parental relationship
Genetic distance
Long-term selection
Growth and development
Dynamic changes
Varietal transition
Quality and nutrients
Abiotic stresses
Biotic stresses
...
...
...

Yunbi Xu (2018) unpublished
Plant breeding is increasingly driven by big data

Data revolution

**Medium**: field book => EXCEL => databases

**Scale**: k => m => b => t

**Dimension**: one (phenotype) => two (phenotype + genotype) => three (phenotype + genotype + envirote) => four (phenotype + genotype + envirotpe + time)

**Throughput**: (data generated in one experiment or unit time): 1 => 100 (1*96) => 10000 (96*96) => 1m (384*3072) => 100M (384*300K)

**Precision**: repeatability, duplicability, compatibility, additivity, predictability

Yunbi Xu (2018) unpublished
Data generation and analysis are costly and time consuming

The real cost of sequencing: higher than you think!
Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB.

Changing costs

CIMMYT and donors are eager to maximize the use and impact of data

Kate Drehe 2013
CIMMYT Science Week
1. Breeding theories
   - GCA, SCA & heritability
   - Simulation
   - Prediction
   - Breeding process
   - Genetic gain
   - Breeding experience
   - Individual & population
   - Heterosis

2. Evaluation and selection
   - Three typing technologies
   - Comprehensive evaluation
   - Data treatment, analysis and mining

3. Breeding procedure
   - Germplasm evaluation
   - Breeding design
   - Parent selection
   - Trait evaluation
   - Field test
   - Variety release

4. Field management
   - Seeding
   - Irrigation
   - Weeding
   - Fertilization
   - Spraying
   - Harvest

5. Breeding information system
   - Collection
   - Sharing
   - Analysis
   - Mining
   - Utilization

Intelligent (Smart) Breeding

Yunbi Xu (2018, unpublished)
AI-assisted breeding: Components

Big data = Foundation for AI development
AI = Four factors: data, computation capacity, algorithm, knowledge
Robot = Way to realized AI

Big data  
\[\text{Robot}\]  
\[\text{Artificial Intelligence}\]  
\[\text{Field test}\]  
\[\text{New varieties}\]  
\[\text{Data, computation capacity, algorithm, knowledge}\]  
\[\text{Modeling – Prediction - Selection}\]  
\[\text{Machine Learning}\]  

Yunbi Xu (2018) unpublished
What we can predict from AlphaGo and AlphaZero? 
From molecular breeding to AI-assisted breeding

AlphaGO vs GS

AlphaFold

- Precision prediction of tertiary protein structure
- AI + Biological Science
- Revolutionized life software

Protein folding game (1 Dec 2018) Predicted 25/43 protein structures, while the second place predicted 3/43 by human

From Lizhi Wang, Iowa State Univ.

Yunbi Xu (2018) unpublished
OUTLINE

- Molecular Plant Breeding: from Theory to Practice
- Breeding Platforms for G-P-E Typing
- Commercial Breeding and Breeding by Design
- Plant Breeding Driven by Big Data and AI
- Integrated and Open-source Plant Breeding
Genomic selection reduces cycle time & cost by reducing frequency of phenotyping
Doubled haploid approaches to breeding-true within two generations

De La Fuente et al 2013 Trends Plant Sci

Pollen from inducer is transferred onto silks of donor plants. Limitations: rate of haploid induction is controlled by quantitative traits in both the inducer and donor.

Selection of haploid kernels is difficult and done manually by using the R1-ny marker which pigments the cap of the aleurone and embryo. Limitations: colored corn, popcorn, colorless allele

Haploid seedlings are treated with colchicine to artificially double chromosome number. Limitations: colchicine is carcinogen, techniques vary, success rate is genotype specific

Haploid plants are self pollinated to develop DH line. Limitations: plants are weak, plants are typically chimeric and shed little pollen, no guarantee that genotype of interest will be successfully self pollinated

DH lines
Speed Breeding Accelerates Generation Time

Grown in a controlled environment room with extended photoperiod (22 hours light/2 hours dark)

Speeding reproduction process by regulating all genes that affect plant growth and development

(Masong Zhou 2019)

MAS: Best combination of genes affecting plant growth and development

Integrated Transgenic Breeding

Why integrated transgenic breeding:
- Option of using transgenic lines suitable for tissue-culture;
- Transgenic lines are usually bad for agronomic traits;
- Removing somaclonal variation

- Simultaneous conventional and transgene integration
- Transferring 2-7 transgenes into several hundreds of inbreds/varieties
- Grow 2-3 generations per year
- Developed inbreds/varieties are highly similar to the recurrent parents
- Supported by about 100M data points per year

Kunsheng Wu (2017) National Maize Genetics and Breeding Conference
The Excellence in Breeding Platform delivers its agenda through five modules

- **Breeding program excellence**  Joining efforts to de one and promote a standard breeding program performance management system to monitor success and highlight investment needs of breeding programs targeting the developing world.

- **Trait discovery and breeding tools and services**  A common platform to share tools, information and training modules on how to successfully incorporate new approaches into the breeding process, from trait discovery to cultivar development.

- **Genotyping and sequencing**  Broker access to genotyping services at reduced cost, assess the latest advances and support breeding programs to optimize the use of genotyping in their work.

- **Phenotyping**  Adapt cutting edge phenotyping approaches for routine use in breeding programs, broker access to phenotyping capacities and expertise, and share and improve infrastructure.

- **Bioinformatics and data management**  Harness the power of genotype, phenotype and other data by providing access to integrated bioinformatics tools and biometrics support.

Envirotyping needs to be included
Open-source breeding strategy

We can learn from: Large-scale commercial breeding programs typically operate as a coordinated network of breeding programs.

Xu et al 2019 Plant Communications
Open-source breeding in China

Maize molecular breeding initiatives in China supported by integrated plant breeding platforms

Tongzhou International Seed GS Breeding Initiatives
Genotyping 100 GS populations supported by Beijing governmental funds

One + Eight Breeding Initiatives
Including one institute (Institute of Crop Science, CAAS) and eight seed companies, with genotyping cost subsidized by Ministry of Agriculture and Rural Affairs

Jiusuo Breeding Initiatives (Proposal)
Seed companies in the winter nursery Sanya, China, through fully open source breeding by sharing everything
Breeding revolution driven by increased data collection

**4-D profile of plant breeding**

Revised from Xu 2016 Theor Appl Genet 129: 653–673
Final Molecular Plant Breeding Products

- Genotypes and genes
- Optimized gene networks
- Alleles or haplotypes
- Specific genomic regions
General References


Molecular Plant Breeding

Yunbi Xu

Forewords by
Dr. Norman Borlaug, Nobel Prize laureate
Dr. Ronald L. Phillips

Chapters reviewed by
Over 30 worldwide experts

Book reviewed by
Crop Science
Field Crops Research
Experimental Agriculture

Translated into Chinese and Persian

Serving as a worldwide textbook and references

ISBN: 978 1 84593 392 0
2010  734 pages
Acknowledgements

Maize Molecular Breeding Laboratory
CIMMYT-CAAS Joint International Research Center for Applied Genomics and Molecular Breeding

International Maize and Wheat Improvement Center (CIMMYT)
Institute of Crop Science, Chinese Academy of Agricultural Sciences (CAAS)

Funding
The Bill and Melinda Gates Foundation
The CGIAR Research Program MAIZE
National Key Research and Development Program of China
The Agricultural Science and Technology Innovation Program, CAAS

Contact: y.xu@cgiar.org
WeChat: molecularbreeder