Regional Training Course on Molecular Approaches
for Selection of Desired Green Traits in Crops
Jakarta, Indonesia, 4-15 November 2019

Breeding Informatics and Decision Support Tools

Yunbi Xu

ccMaize

CAAS-CIMMYT Maize Molecular Breeding Laboratory

y.xu@cgiar.org

1. Breeding informatics

2. Decision support tools

1.1 Information-driven Plant Breeding

Late 1950s: The first computer network, ARPAnet, was developed.

1980s, universities throughout North America and Western Europe were connected via countrywide networks such as the UK's Joint Academic Network (JANet). Molecular biologists were regularly logging in to central servers to run sequence analysis and transferring data from one machine to another.

Early 1990s, the World Wide Web (WWW) was invented and turned the Internet into the worldwide cultural phenomenon that it is today. The WWW has made the concept of a 'global village' developed by Marshal McLuhan decades ago into a reality.

In 1991, Tim Berners-Lee and Robert Caillou, scientists working at OERN (the Organisation Européenne pour la Recherche Nucléaire: *European Organization for Nuclear Research*) in Geneva, developed the Hypertext Transfer Protocol (HTTP) as a way of linking and cross-referencing documents held on different computers.

Xu 2010 Molecular Plant Breeding, CABI

Data tsunami has come

Considering only DNA sequence data, the volume of biological information is doubling roughly every 6 months.

This is faster than the exponential rate of increase in computing power, as suggested by Moore's law (an empirical observation made long ago that has held until today: the doubling of processor power every 12 months) (Sobral, 2002).

With the development of high-throughput technologies, genotypic information including genetic polymorphisms and gene expression profiling, will increase exponentially

Xu 2010 Molecular Plant Breeding, CABI

ccMaize

What is big data ?

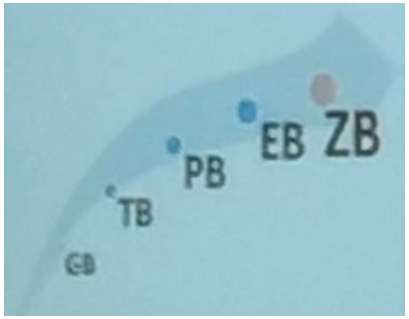
- Big volume: TB; PB; EB; ZB; billions of data points?
- Stored with non-traditional databases ?
- Using data platforms such as Hadoop/Spark?
- Parallel running with multiple machines ?

Many Vs for big data

Volume 数据量	Variety 多样性	Velocity 速度	Value 价值
Veracity 精确	Validity 有效性	Volatility 易变性	Variability 变异性
Visualization 可视化	Vision 想象力	Verbalisers 描述性	.....

Big Data is often described using the five Vs:

Volume, Velocity, Variety, Veracity, Value



https://blog.csdn.net/allenlu2008/article/details/79603476

ccMaize

Volume: Scale of Data

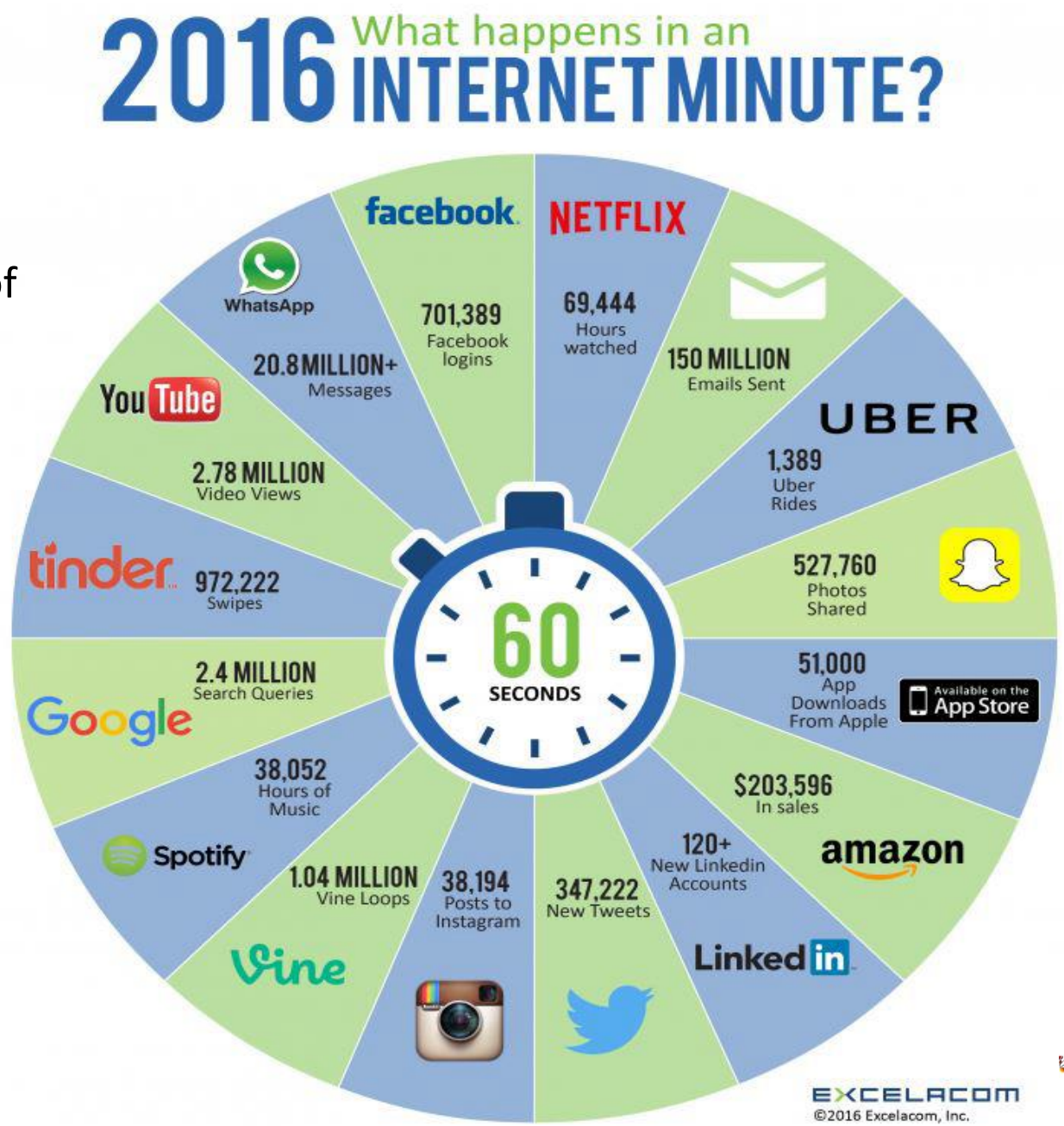
Unit	Value	Size
bit (b)	0 or 1	1/8 of a byte
byte (B)	8	1 byte
kilobyte (KB)	1000^1	1,000 bytes
megabyte (MB)	1000^2	1,000,000 bytes
gigabyte (GB)	1000^3	1,000,000,000 bytes
terabyte (TB)	1000^4	1,000,000,000,000 bytes
petabyte (PB)	1000^5	1,000,000,000,000,000 bytes
exabyte (EB)	1000^6	1,000,000,000,000,000,000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

All information in the world currently available = 7 ZB

ccMaize

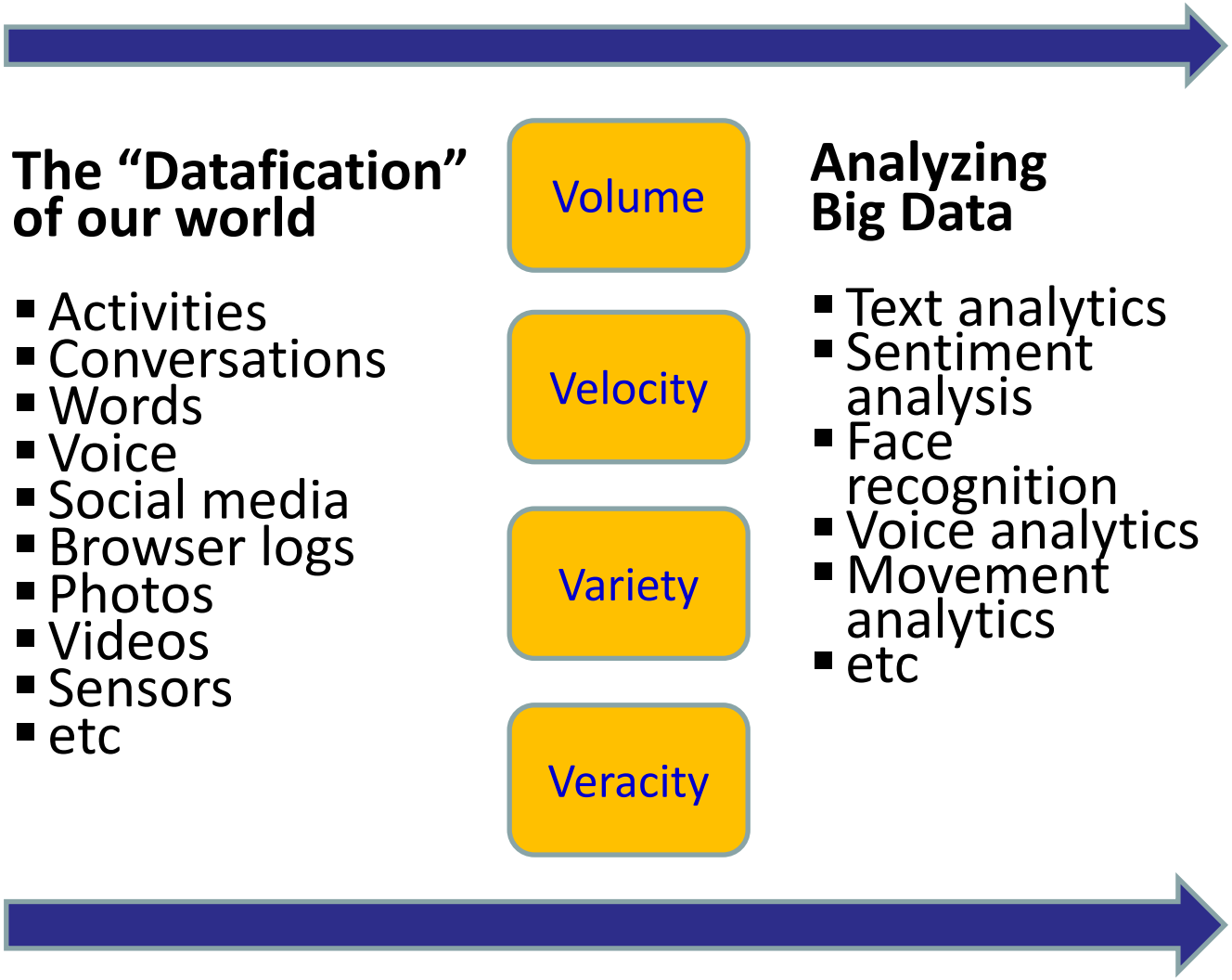
Variety

The different types of data we can use



ccMaize

Value: Turning Big Data into Value



ccMaize

Comparison of big data between plants and humans

	Plant	Human	Impact
Longevity	Short (few weeks to months, seldom years)	Long (tens of years or eternal)	Data acquisition period
Population density	hundreds to 100K per hectare	Several to tens of thousands per km^2	Individual distinct
Population source	Designed, mated or nature	Nature	Population diversity
Reproduction	Quick	Slow	Data accumulation
Value	Low	High	Cost reasonability
Location	Fixed	Mobile	E adaptability
Environments	Nature（逆来顺受）	Nature + artificial	E adaptability
Diseases	Inherited resistance, E-control, external chemical control	Gene modification, surgery, internal medicine therapy	Sample specificity
Data unit	Population-based	Largely individual-based	Sampling

Yunbi Xu (2018) unpublished

ccMaize

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 + envirotype） => four （ phenotype + genotype + envirotype + 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

ccMaize

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

ccMaize

Breeding Informatics in Multinational Seed Companies

- About 25% of total research budget is devoted to breeding informatics
- Development teams are led by breeders or other agricultural scientists, preferably with programming skills
- Development scientists are the interface between breeders and programmers
- These scientists do not manage breeding programs but are devoted full-time to application development
- Applications take 3-6 months to develop
- Life of an application is expected to be no more than 2 years
- Support is available in real time.

ccMaize

Gaps between bioinformatics and plant breeding

Most bioinformatics databases are lacking information on phenotypes, envirotypes, and other organism data.

Explosion of genetic and genomic data for a wide range of plant species has not yet found its way into mainstream plant breeding.

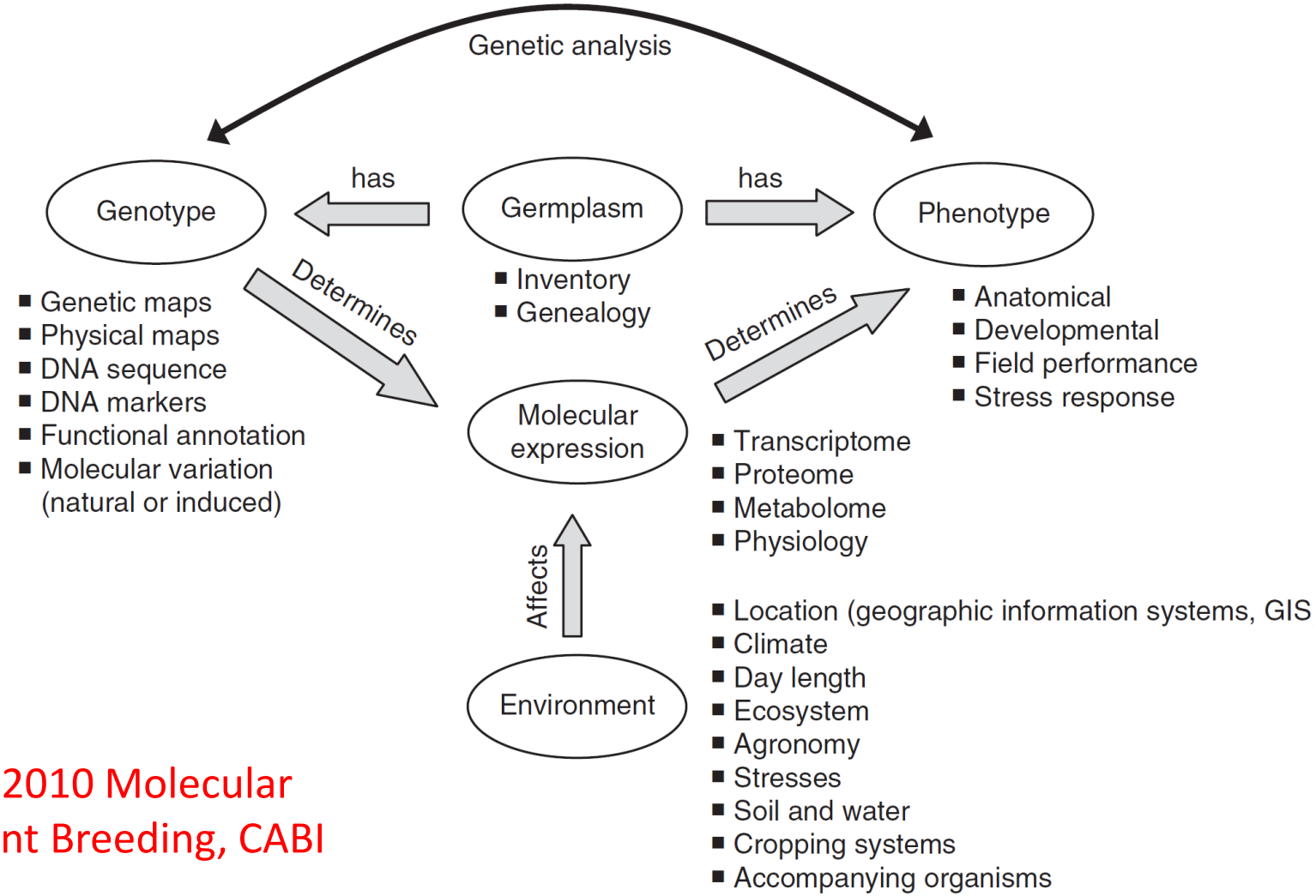
- No idea how or if much of the primary information generated in plant genomics can be applied to real-life breeding situations.
- Breeding requires the integration of information from different sources, usually stored in different databases and managed by different groups.
- Publicly available tools and interfaces available for bioinformatic data are oriented at the cellular/molecular level.
- Much of the genomic research and therefore the publically available data has concentrated on the comparison of genes between species rather than the gene diversity within species required for plant breeding.

Revised from Xu 2010 Molecular Plant Breeding, CABI

ccMaize

1.2 Information Collection

Crop biological concepts, relationships and breeding related information



Xu 2010 Molecular Plant Breeding, CABI

ccMaize

Germplasm information

Passport data

- Accession number and/or other numerical identifiers
- Attributes describing the origin (country of origin, collection site, collection expedition, donor institute)
- Botanical classification (scientific name, taxonomic system, crop, regeneration method)
- Breeding information (institute, method and stage)

Pedigree and genealogy

- *P*edigree: ancestral history of a strain and how a particular accession/strain was derived from its parents, including crossing, selfing, backcrossing and selection.
- Genealogy: ancestral relationships among sets of germplasm, which is a more general description of the breeding history of an accession/strain.

Xu 2010 Molecular Plant Breeding, CABI

ccMaize

Selected Internet resources on germplasm resources

Intergovernmental organizations

Commission on Genetic Resources for Food and Agriculture (FAO): http://www.fao.org/ag/cgrfa/
Consultative Group on International Agricultural Research (CGIAR): http://www.cgiar.org/
Convention on Biological Diversity Secretariat: http://www.biodiv.org/
FAO Plant Genetic Resources: http://www.fao.org/ag/cgrfa/PGR.htm
Bioversity International: www.bioversityinternational.org
CGIAR's System-wide Information Network for Genetic Resources (SINGER): http://singer.grinfo.net/
System-wide Information Network for Genetic Resources (SINGER): http://singer.cgiar.org/

National/regional activities

Asian Vegetable Research and Development Center: http://www.avrdc.org/
Information System Genetic Resources: http://www.genres.de/genres-e.htm
Centre for Genetic Resources, The Netherlands: http://www.cgn.wur.nl/UK/
UK Plant Genetic Resource Group: http://ukpgrg.org/
N.I. Vavilov Research Institute of Plant Industry, Russia: http://www.vir.nw.ru/
Southern African Development Community (SADC) Plant Genetic Resources Project: http://www.ngb.se/sadc/sadc.html
United States Department of Agricutlure (USDA) Genetic Resources Information Network: http://www.ars-grin.gov/
Chinese Crop Germplasm Information System: http://icgr.caas.net.cn/cgris_english.html

Non-governmental organizations

Conservation International: http://www.conservation.org/
Global Biodiversity Forum: http://www.gbf.ch/
World Resources Institute: http://www.wri.org/
Genetic Resources Action International (GRAIN): http://www.grain.org/

Xu 2010 Molecular Plant Breeding, CABI

Genetic stocks

Genetic stock: A plant sample that expresses a specific variation (or a specific small set of variations). Genetic stocks are living examples of their underlying genetic variation.

Most frequently used types of plant genetic stocks

- Near-isogenic lines (NILs),
- Single, series or genome-wide-mutants
- Populations with segregating genotypes (recombinant inbred lines (RILs), DHs,
- Introgression lines (ILs)
- Cytogenetic material (primary trisomics, translocation lines, etc)
- Cell culture lines and gene and DNA clones
- Transformants and gene/genome edited lines

Revised from Xu 2010 Molecular Plant Breeding, CABI

ccMaize

Genotypic information

Genotypic information is based on underlying DNA polymorphisms, which can be detected with many different techniques.

Molecular markers

Information associated with PCR-based DNA markers

Xu 2010 Molecular Plant Breeding, CABI

Marker per se
- Marker name and synonyms
- Repeat motif/enzyme and repeat length
- Primer sequence
- PCR protocol (i.e. annealing temperature and number of cycles)
- Expected allele size in a control cultivar or a group of cultivars
- Number of alleles
- Allele frequency (most/least frequent allele)
- Signal strength
- Allele size/range
- Polymorphic information content
- Chromosome location
- Linkage to other markers
- Images and gel pictures
- References
- Source (inventory)
- Patent information
- Historical data (e.g. associated with a trait)
- Project data (date, title, germplasm, reports, etc.)

Marker-derived
- Genetic maps (including haplotype block)
- Physical maps
- Consensus maps
- Comparative maps

Sequences

DNA: Plant genome sequence data have been accumulating from three major sources:

(i) Whole genome sequences
(ii) Genome survey, skim or selective sequences
(iii) ESTs (for all target and full-length cDNA sequences
(iv) Specific sequences: methylation sequences, etc

RNA: gene expression (microarray and RNA-seq)

- What were the taxonomy, sex and developmental stages of the organism?
- What were its growth conditions?
- From which organ and tissue was the sample extracted?
- What protocols were used in sample preparation?

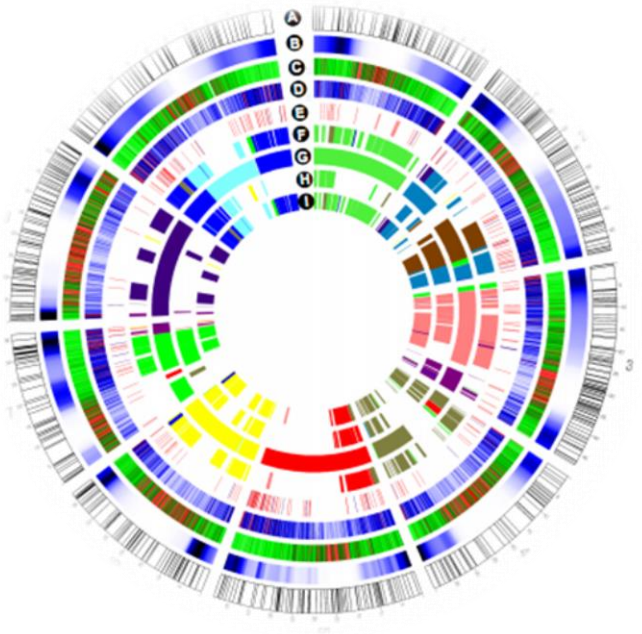
Protein: all or part of a protein or peptide

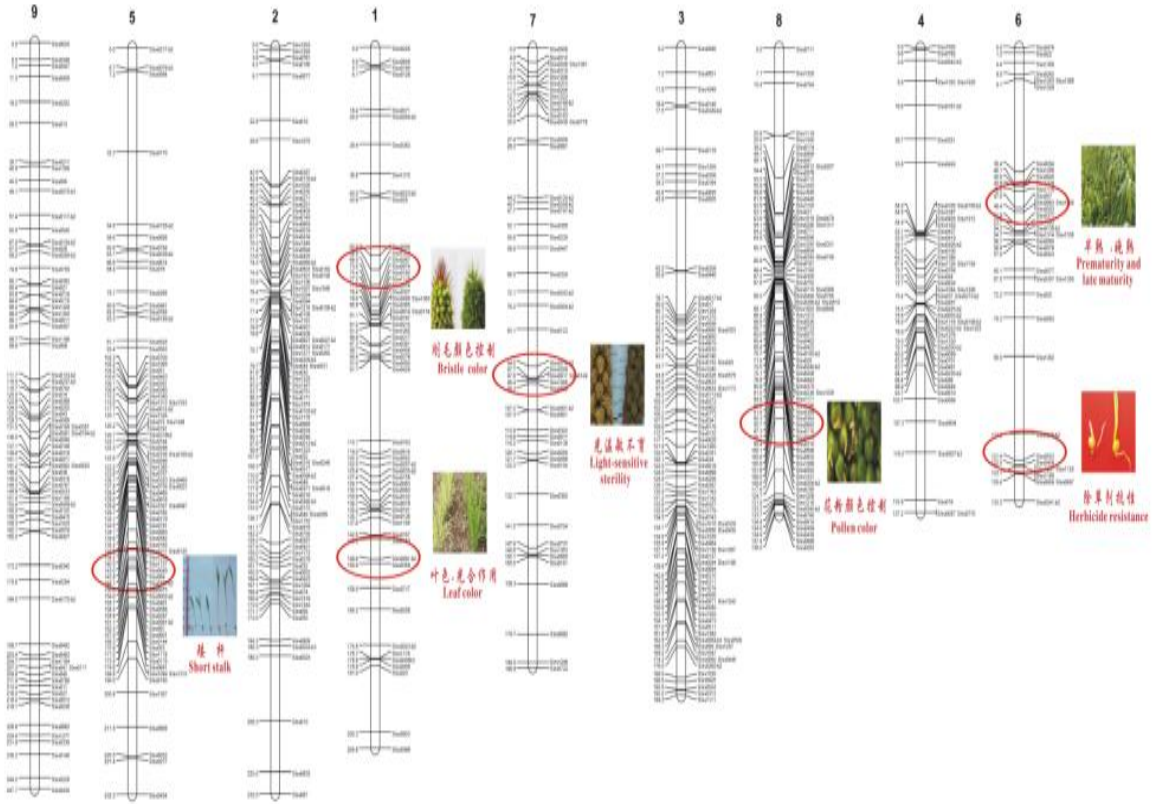
- Amino acid sequences
- N- and C-termini
- Post-translational modifications
- Whole-mass determination
- Sequences predicted from DNA/RNA sequences

Revised from:
Xu 2010 Molecular Plant Breeding, CABI

ccMaize

All Identified Genes/QTL and Associated Information





ccMaize

Phenotypic information

Phenotype: a distinguishable feature, characteristic, quality or physical feature of a developing or mature individual. A phenotype results from the expression of an organism's genetic code, its genotype, as well as the influence of environmental factors and the interactions between the two.

Phenotypic data: all data collected in various genomics and breeding programs, either for basic research or product delivery.

Phenome and phenomics: a collection of traits, while the simultaneous study of such a collection is referred to as phenomics.

Molecular phenotyping: the technique of quantifying pathway reporter genes, i.e. pre-selected genes that are modulated specifically by metabolic and signaling pathways, in order to infer activity of these pathways.

Wikipedia

ccMaize

Envirotypic information

Soil
- Texture
- Water content
- Fertility
- Nutrient content
- Production index

Air
- Pollutants
- Emission of CO_2

Light
- Light intensity
- Day length

Temperature
- Average, maximum, minimum daily temperature
- Effective temperature
- Length of available growing period

Water
- Humidity
- Precipitation
- Ground water
- Water quality
- Potential evapotranspiration

Cropping systems
- Intercropping
- Previous crop

Accompanying organisms
- Root microorganisms
- Weeds
- Pathogens
- Insects

Xu 2010 Molecular Plant Breeding, CABI

ccMaize

1.3 Information Integration

Integration for molecular data

(i) Structural genomics: DNA sequences (to complete genomes) and maps (genetic, physical or cytological);

(ii) Gene expression: mRNA profiling and single gene profiles (Northern);

(iii) Biochemistry: pathways (metabolic and signalling), metabolites, proteomics.

Three ways of information integration

- Link integration: begin their query with one data source and then follow links to related information in other data sources
- View integration: leaves the information in its source database, but builds an environment around the databases that makes them appear to be part of one large system.
- Data warehouse: brings all the data together under one 'roof' in a single database.

Xu 2010 Molecular Plant Breeding, CABI

ccMaize

Data standardization

Benefits of data standardization (Jenkins *et al*., 2005):

(i) Consideration and development of best practice and standard operating procedures, which, in turn enable proper interpretation of experimental results, principled dataset comparison and experiment repetition;

(ii) Standardized reporting of experiments and deposition and archiving of data associated with publications or other standard pieces of work;

(iii) Development of databases and verifiable transmission mechanisms for storage, collection and dissemination of results.

Xu 2010 Molecular Plant Breeding, CABI

ccMaize

Development of generic databases

The increasing number and types of databases and software applications make it more and more difficult for researchers to determine which databases to use.

BioMOBY initiative (http://biomoby.open-bio.org): An international research project involving biological data hosts, service providers and coders whose aim is to explore various methodologies for biological data representation, distribution and discovery.

Generic Model Organism Database : a collaborative project funded by the National Human Genome Research Institute (GMOD; http://gmod.org) to promote the development and sharing of software, schemas and standard operating procedures. The project's major aim is to build a generic organism database toolkit to allow researchers to set up a genome database 'off the shelf'.

Xu 2010 Molecular Plant Breeding, CABI

Use of controlled vocabularies and ontologies

Ontology: Simple definition : the examination of what is meant, in context, by the word 'thing'. An ontology is simply an organized set of concepts about a specified domain, with two components:

(i) an indexed controlled vocabulary of terms (the 'concept');
(ii) information about semantic relationships between these terms.

Globally unique identifiers to standardize the description

Two important ontologies:

- Gene Ontology (GO): the framework for the model of biology (http://www.geneontology.org)
- Plant Ontology (PO): a structured vocabulary and database resource that links plant anatomy, morphology and growth and development to plant genomics data (http://bioportal.bioontology.org)

Gramene-hosted ontologies (http://www.gramene.org) :

- Trait Ontology
- Environment Ontology
- Taxonomy Ontology

Revised from
Xu 2010 Molecular Plant Breeding, CABI

ccMaize

Interoperable query system

There is an increasing need to make existing data from different organisms simultaneously searchable, visible and, most importantly, comparable.

Productive utilization of databases requires interoperability: that is, the precise yet flexible interrelating of information from one database to another. There are, at present, two major impediments to achieving wide-scale interoperability: the state of database protection legislation and computer security issues

Xu 2010 Molecular Plant Breeding, CABI

ccMaize

Redundant data condensing

The availability of complete genome sequences, as well as the flood of other sequence data, is leading to alternative views on how these data can be organized and interrogated.

The ever-increasing size of DNA sequence databases continues to push bioinformatic capabilities and there is a growing need to condense redundant data

Xu 2010 Molecular Plant Breeding, CABI

ccMaize

Database integration

Of particular importance is the ability to attach substantial genomic information to the sequences.

After identifying genes and predicting the proteins they encode, we need to determine

- when and where the proteins are expressed
- how they interact
- how these expression and interaction profiles are modified in response to environmental signals.

Emphasis on the underlying value of genotypic and genomic elements must be balanced with a pheno-centric approach.

- A common data model combining the data with a common gene index
- Integration of ecosystems information (environment and interactions among organisms and populations)

Xu 2010 Molecular Plant Breeding, CABI

1.4 Information Retrieval and Mining

Information retrieval

Data retrieval, document retrieval and text retrieval, each of which has its own body of literature, theory, practice and technologies.
Bibliographic databases*:* referred to the 'abstracting and indexing service' for scholarly literature; expanded to include full-text articles, original data, images and books.

Web of Science is the inclusion of the citations associated with each abstract. It is possible to:

(i) View the abstracts of all articles cited in the original (parent) article;
(ii) Find all articles published since the original (parent) article and those that have cited it; and
(iii) Find all the articles that have cited a particular author.

BIOSIS Previews is made up of two databases: Biological Abstracts and Biological Abstracts/RRM, the latter covering reports reviews and meetings – information not formally published in scientific research journals.

Xu 2010 Molecular Plant Breeding, CABI

Full text of research articles

Access to the full text :

- The journal is published online with OPEN ACCESS
- The publisher has agreed with the database to make the article available
- The individual or individual's library subscribes to the journal The publisher makes the article freely available right after the publication or some time after the publication
- Open-access journals (freely available online for reading, downloading, copying, distributing and using):
  - ✓ PLoS (Public Library of Science, http://www.plos.org/);
  - ✓ Hindawi Publishing Corporation (http:// www.hindawi.com/)

Books and text-rich web sites

Xu 2010 Molecular Plant Breeding, CABI

ccMaize

Pay an attention to what you can retrieval

It is important to assess what qualifies the individual or organization to publish the information and what their motivation for doing so might be. As with any literature research, the information retrieved should be cross-checked and critically evaluated.

https://www.yahoo.com (retrieved at 16-15:00 Beijing Nov 21, 2017):

Genetic	85,800,000
Genetics	54,600,000
Genetica	8,650,000
Genetical	400,000

Transgene	4,240,000
Transgenesis	3,780,000
Transformants	528,000
Transgenics	499,000
Transgenes	436,000

Commercial	41,300,000
Commercialisation	6,480,000
Commercialized	4,500,000
Commercialization	3,790,000
Commercialised	375,000

Kung-Fu	28,600,000
Kongfu	28,600,000
Kong-Fu Cha	13,200,000
Gong-Fu	12,600,000
Kong-Fu	10,500,000
Gong-Fu Panda	8,860,000
Kung-Fu Cha	5,090,000
Kung-Fu Panda	5,810,000
Kong-Fu Panda	977,000
Gong-Fu Tea	517,000
Kung-Fu Tea	416,000
Gongfu	419,000
Gong-Fu Cha	347,000

The best translation:

功夫茶 Kong-Fu Cha

功夫熊猫 Gong-Fu Panda

ccMaize

Google Scholar

Google™ (http://scholar.google.com/) has become popular for literature searches. Searching by authors, key words or authors' affiliation will bring up all related publications (articles, books, etc.) with article title, author list, the number of citations, etc. A full article can be browsed from a provided link. All the articles that cite a specific article can be browsed.

Google Citations

- Which publications received more attention?
- What fields are hot?
- Which scientists have high citations in the relevant fields?
- Which institutes are leading in which fields and their overall performance

Yunbi Xu

关注

CIMMYT/CAAS
在 cgiar.org 的电子邮件经过验证

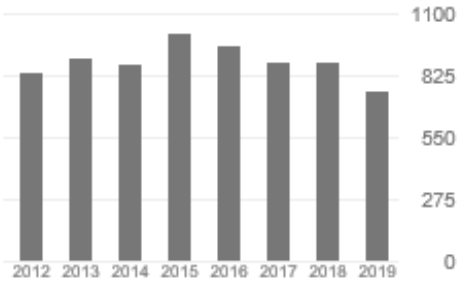
Molecular Breeding

创建我的个人资料

标题	引用次数	年份
Development and mapping of 2240 new SSR markers for rice (Oryza sativa L.) SR McCOUCH, L Teytelman, Y Xu, KB Lobos, K Clare, M Walton, B Fu, ... DNA research 9 (6), 199-207	1788	2002
Development of a microsatellite framework map providing genome-wide coverage in rice (Oryza sativa L.) X Chen, S Temnykh, Y Xu, YG Cho, SR McCouch Theoretical and Applied Genetics 95, 553-567	952	1997
Marker-assisted selection in plant breeding: from publications to practice Y Xu, JH Crouch Crop Science 48, 391-407	731	2008
Microsatellite marker development, mapping and applications in rice genetics and breeding SR McCouch, X Chen, O Panaud, S Temnykh, Y Xu, YG Cho, N Huang, ... Plant Molecular Biology 35, 89-99	720	1997

引用次数 查看全部

	总计	2014 年至今
引用	11499	5354
h 指数	49	38
i10 指数	92	71



Updated on 9:40, Oct 25, 2019 Shihezi, Xinjiang

ccMaize

Information mining

Data mining, or knowledge discovery in databases (KDD), has been described as 'the nontrivial extraction of implicit, previously unknown and potentially useful information.

Plant breeders may want to mine for the following information:

- Germplasm information collected across worldwide institutions;
- Marker–trait associations reported for traits of interest to specific breeding programs;
- Genes that are required for the improvement of traits of agronomic importance through transformation and introgression;
- Molecular markers and marker-related information for the development of MAS tools;
- Alleles, haplotypes, and genomic regions that can be used in breeding by molecular design.

Revised from Xu 2010 Molecular Plant Breeding, CABI

ccMaize

Comparative Informatics

Four basic comparative bioinformatics analyses:

- DNA–DNA conservation: determined by alignment of complete DNA sequence from two plant species
- Syntenic blocks: the identification of segments of the genome in which the order of particular genes is conserved between two species (syntenic blocks)
- Orthologues: another type of comparative analysis focuses on genes and proteins and attempts to identify the orthologous genes in different plants
- Phenomic similarity: phenotypic and physiological similarities can be used

Xu 2010 Molecular Plant Breeding, CABI

ccMaize

Sequence similarity analysis

DNA sequence similarity analysis can be used to trace allele, gene or chromosomal fragments, identify similarities between sequences or genes and align multiple sequences.

Protein sequence analysis includes searching for protein similarity and looking at primary, secondary and tertiary structure.

BLAST (Basic Local Alignment Search Tool) is a set of similarity search programs designed to explore all of the available sequence databases.

BLAST services: NUCLEOTIDE BLAST; PROTEIN BLAST; TRANSLATED BLAST; GENOMIC BLAST pages (human genome, eukaryotes, microbial genomes)

specialized BLAST pages

- VECSCREEN, a BLAST-based detection of vector contamination
- IGBLAST, for analysis of immunoglobulin sequences in GenBank®
- GEO BLAST, for gene expression data
- SNP BLAST

Xu 2010 Molecular Plant Breeding, CABI

ccMaize

1.5 Information Management Systems

Gap between breeding and molecular biology in IMS:

- Many of the systems have been developed independently for phenotypic and genotypic data and used by two different groups of scientists.
- Breeding information has been managed in most breeding programs by using relatively simple tools with less training required but not suitable for data management and statistical analysis when multiple-resource data are incorporated.
- Insufficient communication between breeders/agronomists and geneticists/ molecular biologists has resulted in limited hardware, database and software support, compared to those established in the biotechnology and IT industries.
- Understanding and communication between IT scientists and breeders and between facility developers and breeders, is also lacking. This contributes to underdevelopment of information systems designed for plant breeding.
- Many breeding companies and institutions, especially in developing countries, are lacking the personnel and facilities for information management.

Xu 2010 Molecular Plant Breeding, CABI

ccMaize

Laboratory information management systems (LIMS)

LIMS manage data, samples, laboratory users, instruments, standards and support laboratory functions such as invoicing, plate management, sample tracking and work flow automation.

To move the whole process of information collection, management, analysis, decision making, review and release into the workplace:

- Instruments are integrated in the lab network.
- Laboratory personnel perform calculations, review and document results using online information and electronic lab notebooks connected to the LIMS.
- Management can supervise the lab process, react to bottlenecks in workflow and ensure regulatory requirements are met.
- Laboratory participants can place work requests and follow up on progress, review results and other documentation.

There is a huge difference between working with 300 plants and that with 30,000 plants

Revised from
Xu 2010 Molecular Plant Breeding, CABI

Future needs for informatics tools

- multi-year trial data: environment classification and yield stability analysis;
- heterotic pattern analysis and heterotic pooling;
- genotype-by-environment interaction;
- germplasm characterization and classification;
- marker characterization and genetic mapping;
- identification and quantification of novel genetic variation;
- molecular and functional analysis of genetic diversity and evolutionary process;
- association of genotypes with phenotypes through linkage genetics;
- simulation and modelling of gene networks and biological interactions;
- heterosis and combining ability analysis and hybrid performance prediction;
- statistical methods for molecular data of multiple sources.

Xu 2010 Molecular Plant Breeding, CABI

ccMaize

1. Breeding informatics

2. Decision support tools
   決策支撐工具

Decision Support Tools

Decision support tools: Introduction

Breeding information management

Support to breeding activities

Breeding simulation

CGIAR practice

Integrated plant breeding

ccMaize

Bottlenecks in Molecular Plant Breeding

- Need of integrated plant breeding platform
- Establishment of molecular breeding networks
- Improved precision phenotyping and envirotyping
- Effective and efficient information management, analysis and delivery

Integration, Large-scale and Standardization
育种的集成化、规模化、 程序化

ccMaize

Why do we need decision support tools in plant breeding?

To provide the best compromise between time, cost and genetic gain, we need to

- Identify new sources of beneficial genetic variation and develop robust marker–trait associations;
- Manage and manipulate large amounts of genotype, pedigree, phenotype and envirotype data;
- Select desirable recombinants through an optimum combination (in time and space) of genotypic, phenotypic and envirotypic information;
- Develop breeding systems that minimize population sizes, number of generations and overall costs while maximizing genetic gain for traditional and novel target traits.

Revised from Xu 2010 Molecular Plant Breeding, CABI

ccMaize

Which Decision Support Tools Do We Need?

Germplasm management, evaluation, and enhancement
Breeding population management and improvement
- Building up heterotic patterns
- Prediction of hybrid performance
- Marker-assisted inbred and synthetic creation

Genetic map construction
Marker-trait association identification and validation
Marker-assisted selection methodologies and implementation
Genotype by environment interaction analysis
Intellectual property right and plant variety protection
Breeding design through simulation and modeling

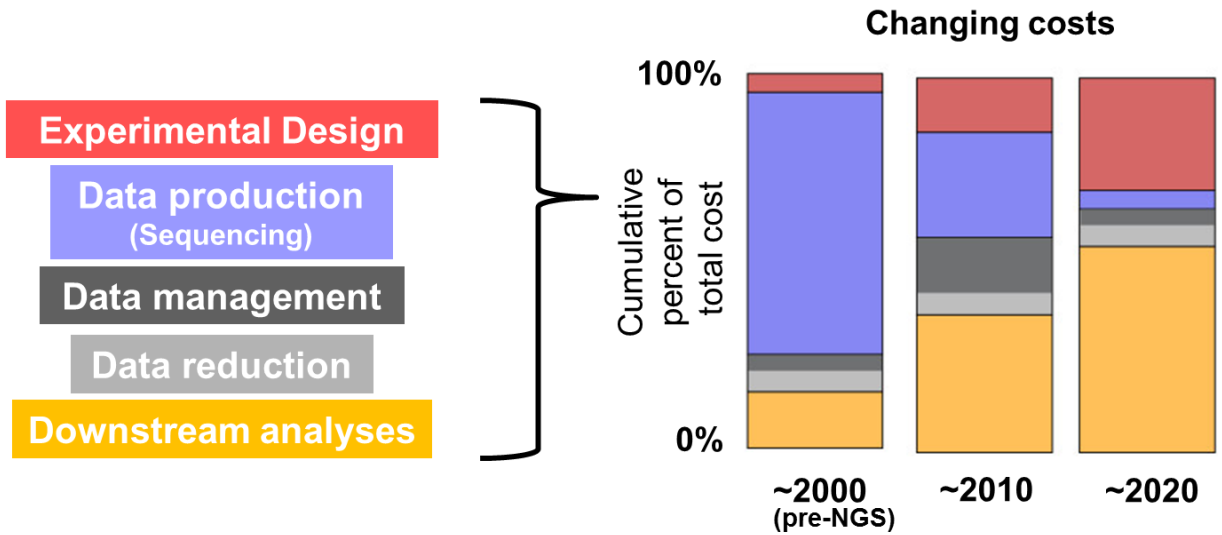
Xu 2010 Molecular Plant Breeding. CABI Publisher

Data generation and analysis are costly and time consuming

Genome Biol. 2011 Aug 25;12(8):125. doi: 10.1186/gb-2011-12-8-125.

The real cost of sequencing: higher than you think!

Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB.

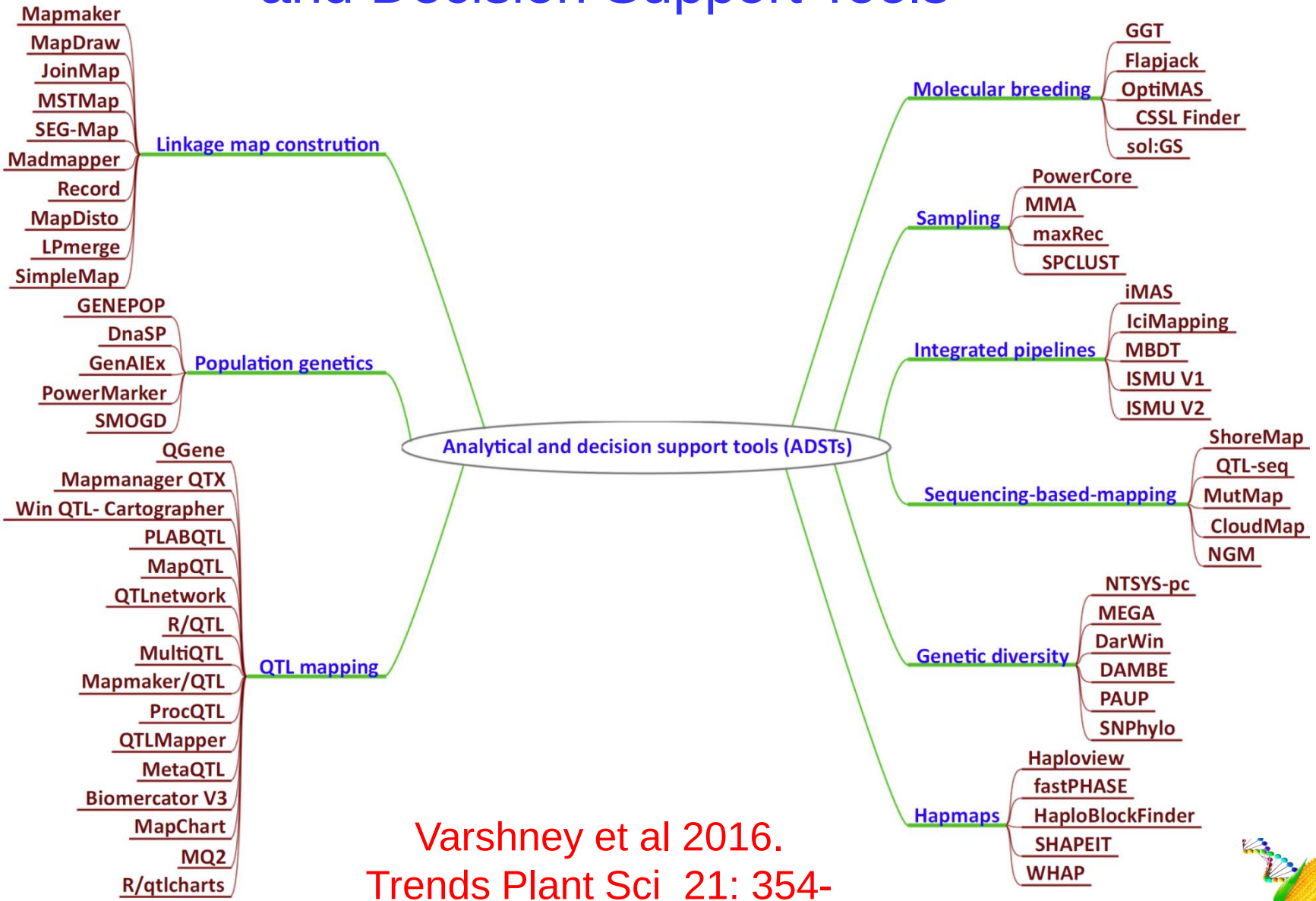


CIMMYT and donors are eager to maximize the use and impact of data

Kate Drehe 2013
CIMMYT Science Week

ccMaize

The Most Popular Analytical and Decision Support Tools



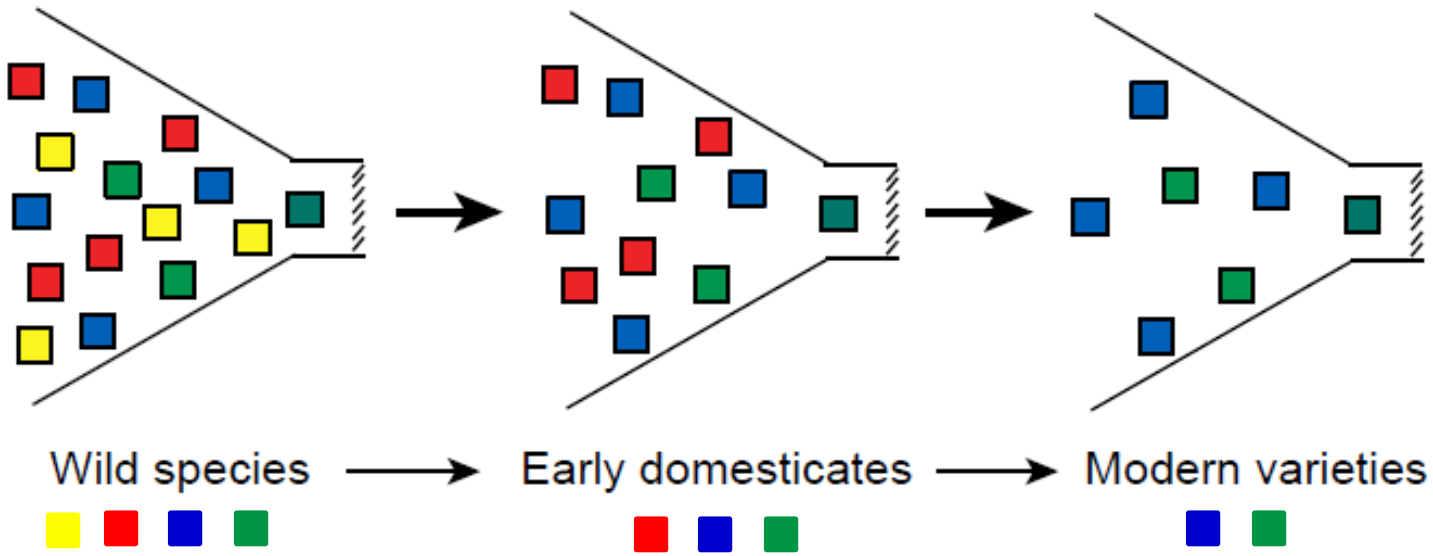
Varshney et al 2016. Trends Plant Sci 21: 354-363

ccMaize

Management of genetic diversity

ccMaize

Monitoring Geneflow from ancestors to cultivars



Genetic bottlenecks imposed on crop plants during domestication and through modern plant-breeding practices.

Tanksley and McCouch 1997

ccMaize

Available breeding simulation tools

- **QuLine**, a computer software that simulates breeding programs for developing inbred lines
- **QuHybrid**, a computer software that simulates breeding programs for developing hybrids
- **QuMARS**, a computer software that simulates marker-assisted recurrent selection and genome-wide selection

FieldLab

An Android application for paperless data collection in the field and laboratory

- Reduce the time and effort in consolidating and encoding the observation data gathered
- Avoid errors in recording observation data
- Improve data accessibility

Samsung Galaxy Tablet

- Higher computing
- Wider screen

(W x H x D) 4.74 x 7.48 x 0.47 inches

From Guoyou Ye, 2013, IR

ccMaize

FieldLab

- Import/Create study workbook
- Manage study
- Data entry form with validation, range input and look-up values on scoring
- Read barcode label to search for record
- Manage trait to measure
- Manage images captured and audio recorded
- Export data in workbook format

From Guoyou Ye, 2013, IRRI

ccMaize

CGIAR Excellence in Breeding (EiB)Platform 2017-2022

http://excellenceinbreeding.org

Excellence in Breeding PLATFORM

HOME | FUNDING & EXECUTION | TARGET BUDGET | PROPOSALS | CONTACT

Tools and services that create synergies and accelerate genetic gains of breeding programs targeting the developing world

CGIAR Excellence in Breeding Platform

Through the **Excellence in Breeding Platform** the CGIAR intends to modernize breeding programs targeting the developing world for greater impact on food and nutrition security, climate change adaptation and development.

Drawing from innovations in the public and private sector, the Platform will provide access to **cutting-edge tools, services and best practices, application-oriented training and practical advice.**

ccMaize

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.

ccMaize

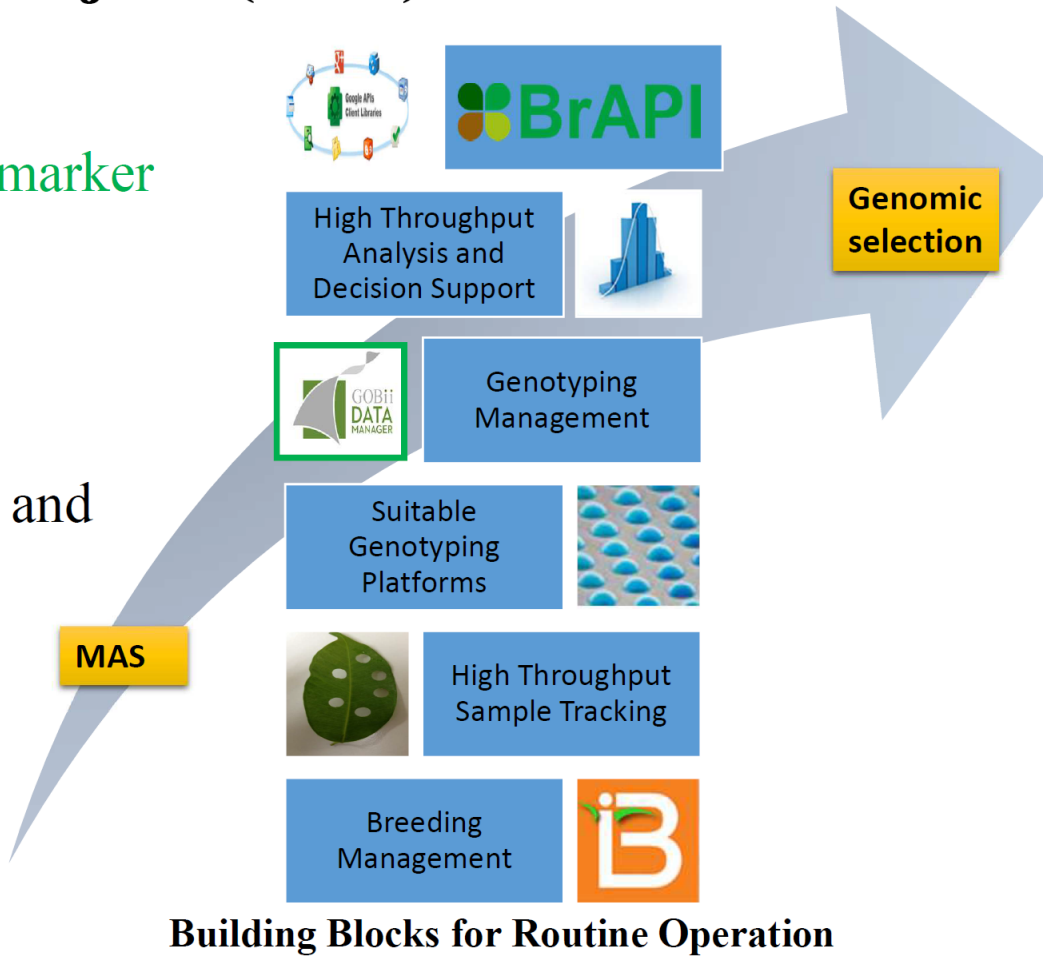
GOBii Project (1st *P*)

Mission:

Transform breeding by enabling marker and genomic–assisted selection in **routine** breeding application

Scope:

Putting systems and tools in place and **connecting** existing fundamental building blocks



Building Blocks for Routine Operation

ccMaize

GOBii Products (2nd *P*)

Marker Portal

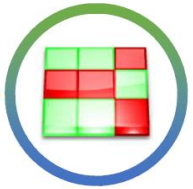
Portal containing all deployed GOBii tools
http://asia.gobii.org:8081/gobii-portal/

GS in Galaxy

Genomic selection and genome wide association mapping workflows in Galaxy

GOBii Data Manager

Large scale genomics data management system



Flapjack Modules

Pedigree verification
Marker assisted backcrossing
Forward breeding

TimeScope

System administrative tool for deleting unwanted data in GDM

Data QC

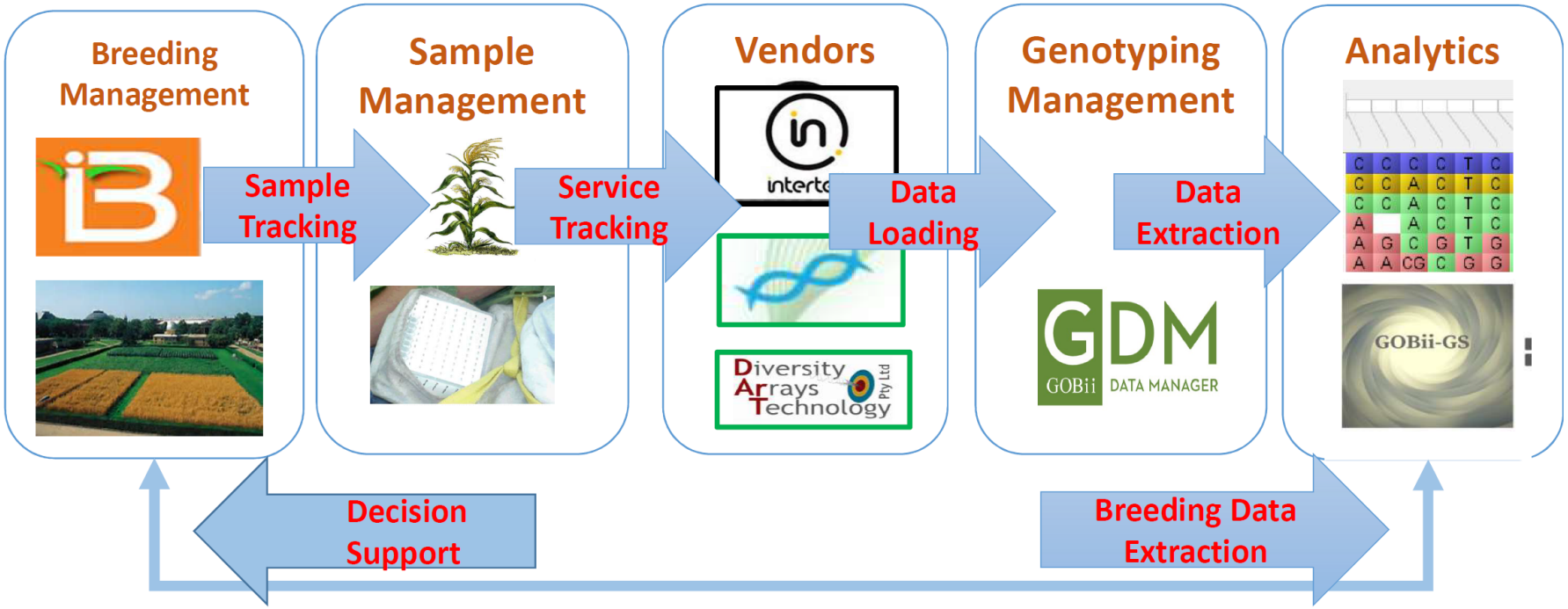
Plugin in KDCompute for checking quality of data loaded into GDM

Star Gao, GOBii Training Course – China, Beijing 2019

ccMaize

GOBii Products (2nd *P*)

End to End Data Management



Star Gao, GOBii Training Course – China, Beijing 2019

ccMaize

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: moleculardbreeder