Bioinformatics and Agriculture:
an old marriage of convenience

Miguel Pérez-Enciso

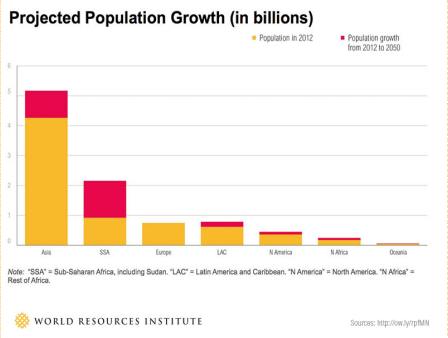
(miguel.perez@uab.es)

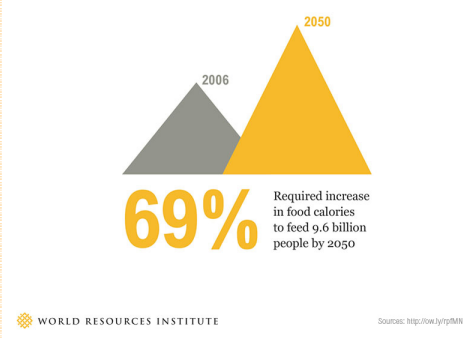
ICREA CRAG UAB Universitat Autònoma de Barcelona

Outline

- The need for increased food with equal or less resources.
- How humans have addressed this so far: domestication and artificial selection.
- How bioinformatics can help.

Global Food Challenge

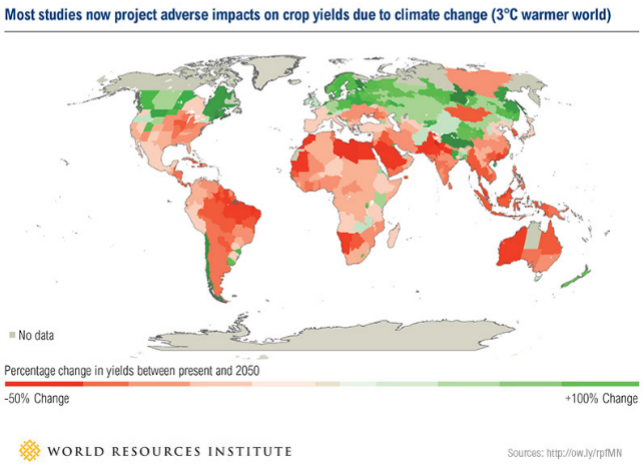




http://www.wri.org/

Climate Change and Water Stress Exacerbate the Challenge

Climate change is expected to negatively impact crop yields, particularly in the hungriest parts of the world, such as sub-Saharan Africa.



The Green Revolution

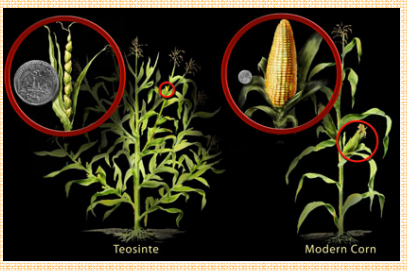
Norman Borlaug (Peace Nobel prize, 1970): He developed semi-dwarf, high-yield, disease-resistant wheat varieties in Mexico. During the mid-20th century, Borlaug led the introduction of these high-yielding varieties combined with modern agricultural production techniques to Mexico, Pakistan, and India. As a result, Mexico became a net exporter of wheat by 1963. Between 1965 and 1970, wheat yields nearly doubled in Pakistan and India, greatly improving the food security in those nations.

Borlaug was often called "the father of the Green Revolution" and is credited with saving over a billion people worldwide from starvation

https://commons.wikimedia.org/w/index.php?curid=127051

Main events: domestication

Teosinte → Corn

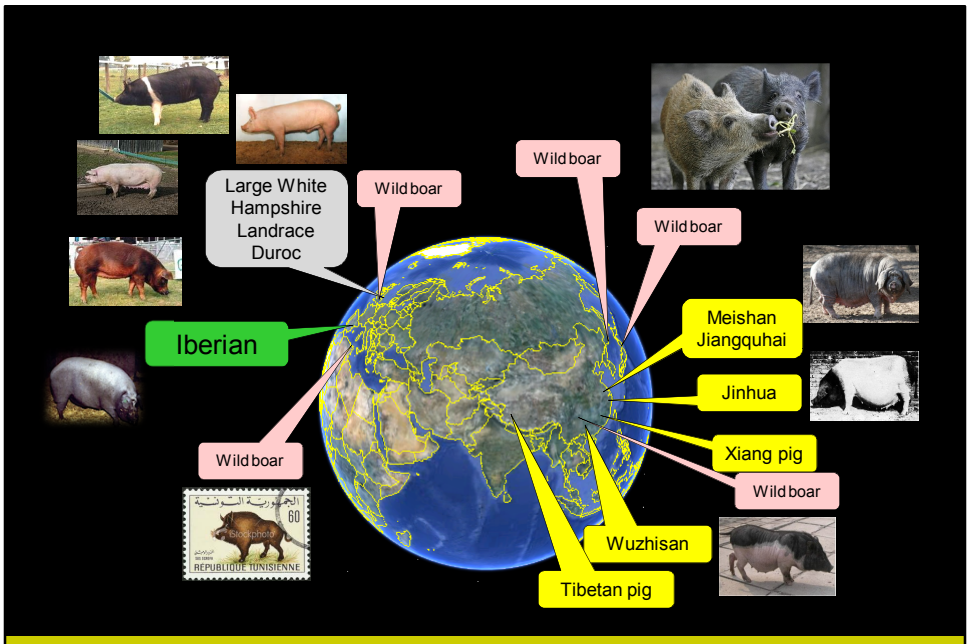


Mesoamerica, ~ 9,000 years ago

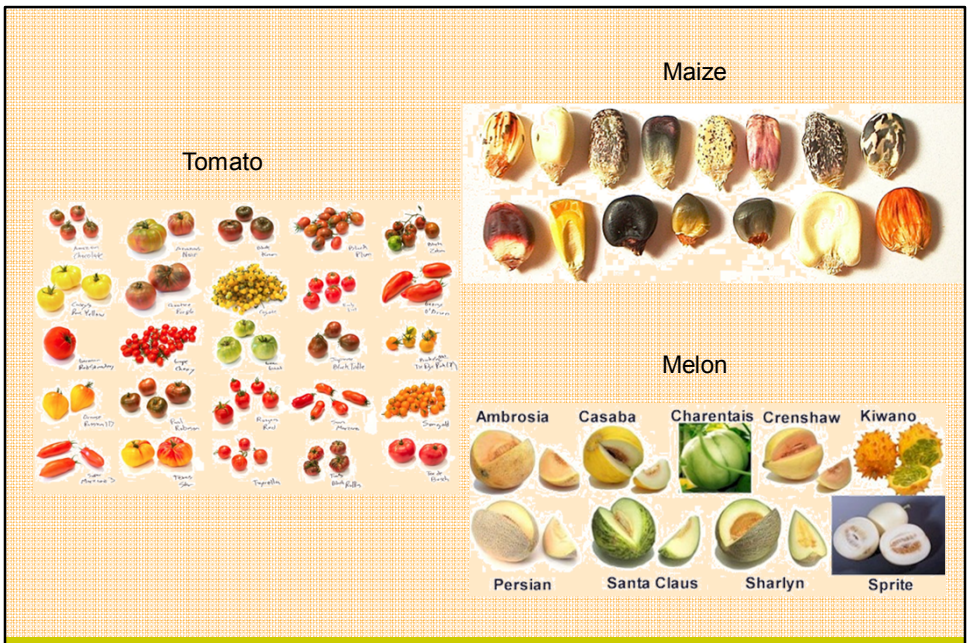
Wild boar

Pig

Asia and Anatolia,
~ 9,000 years ago

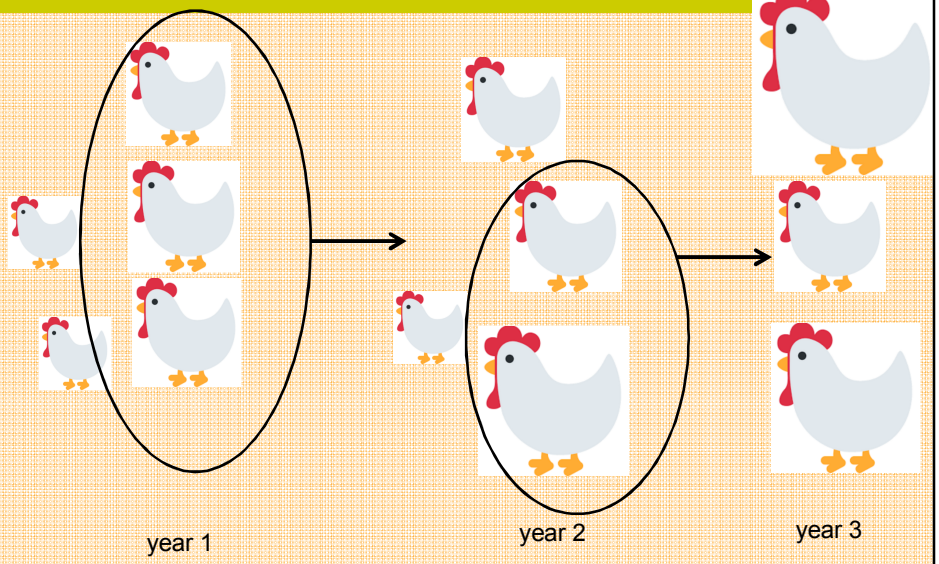


Main events: Breed creation

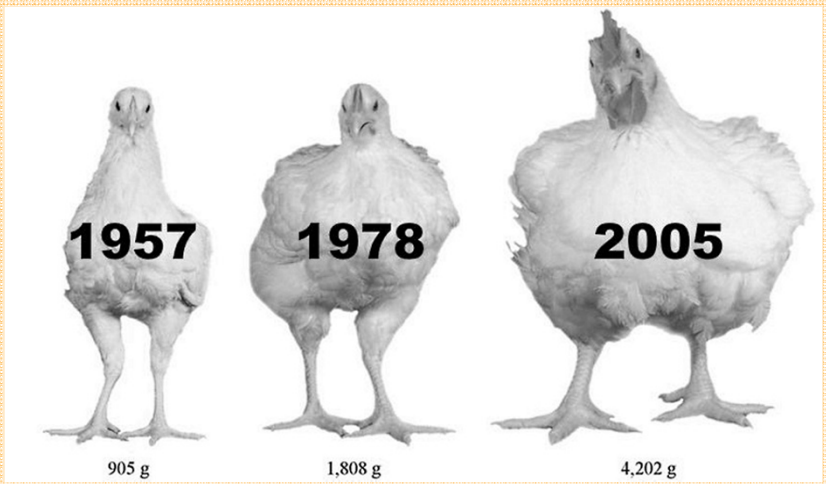


Main events: Breed creation

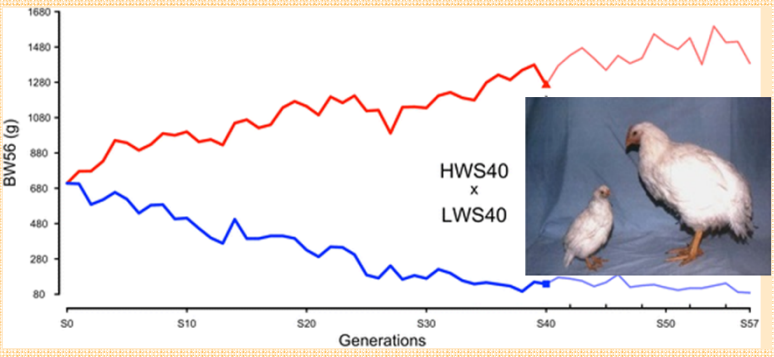
Main events: Artificial Selection



Increase in average performance: chicken



How did selection work in animals?

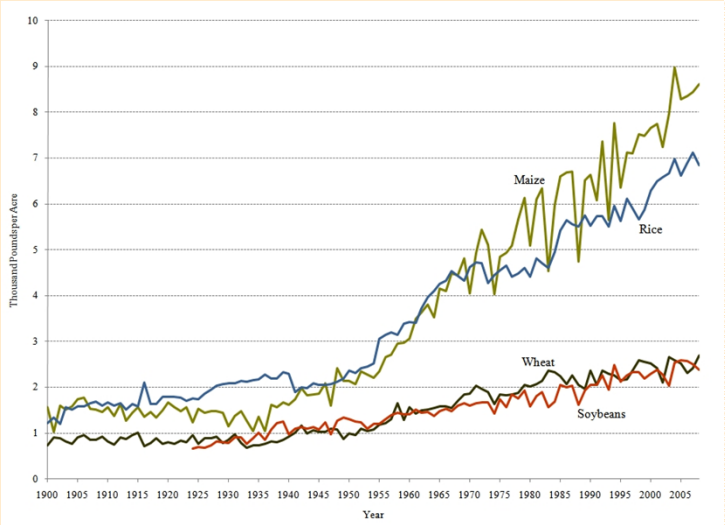


Long-Term Divergent Selection for Eight-Week Body Weight in White Plymouth Rock Chickens

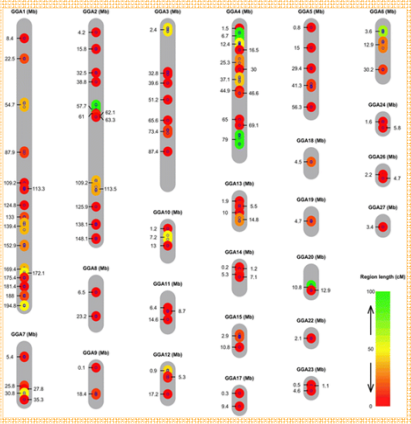
E. A. DUNNINGTON and P. B. SIEGEL

Department of Animal and Poultry Sciences, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061-0306

Increase in average performance: cereals



IMPORTANT!!! Many genes affect the traits of interest



Detected regions with an effect on body weight in the Virginia divergent selection chicken lines (Sheng et al, Genom. Biol 2015)

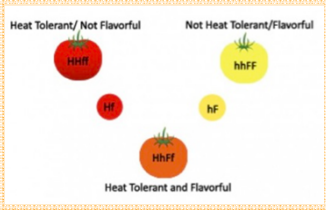
This is called 'The infinitesimal model'

It presupposes that quantitative traits are explained by a large number of genes, each acting individually and of small effect. In addition, quantitative traits are also modified by the environment.

XV.—**The Correlation between Relatives on the Supposition of Mendelian Inheritance.** By **R. A. Fisher**, B.A. *Communicated by* Professor J. Arthur Thomson. (With Four Figures in Text.)

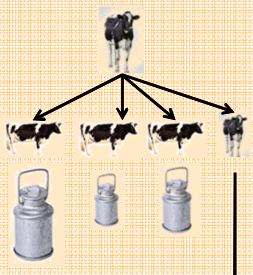
(MS. received June 15, 1918. Read July 8, 1918. Issued separately October 1, 1918.)

How Plant improvement works

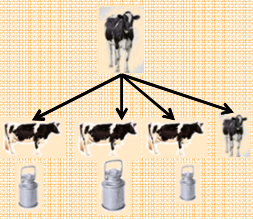


Crosses between elite breeds, evaluation and development of new hybrids

How Animal improvement works



Selected to replace older bulls



Genetic evaluation combines phenotype and genetic relationships (pedigree)

Summary so far

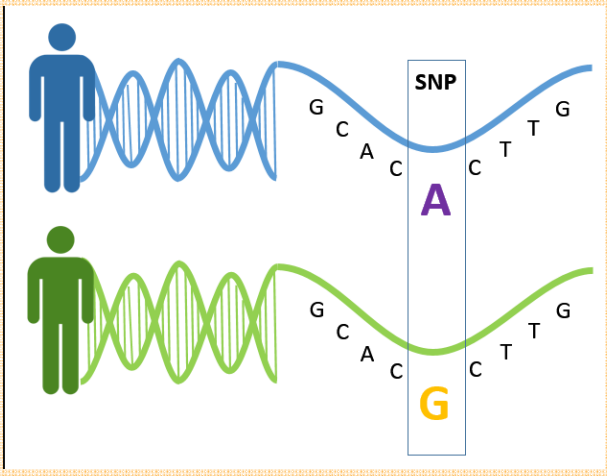
- ✓ Artificial selection is very effective in changing the phenotypes.
- ✓ Many genes involved: no exhaustion of genetic variability (this is called the INFINITESIMAL MODEL).
- ✓ Artificial selection is based on pedigree and phenotypic information.
- ✓ Pedigree information is being replaced by molecular information.

Some Bioinformatics Applications

- ✓ Precision agriculture (Big data issues, modeling …)
- ✓ Causal mutation discovery
- ✓ Genomic selection, machine learning
- ✓ Genome assembly
- ✓ Sequence analysis
- ✓ Functional analysis
- ✓ …

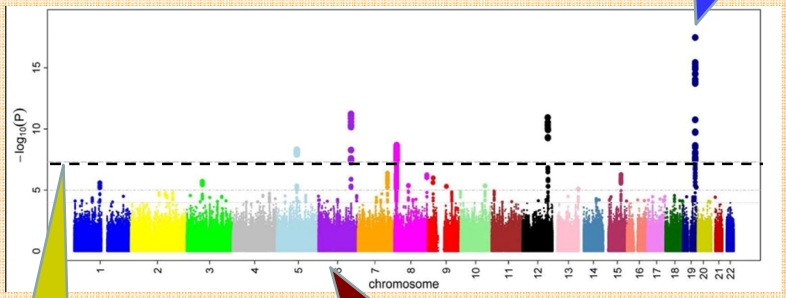
Causal Mutation Discovery: Genome Wide Analysis Study (GWAS)

There are millions of small DNA differences between individuals called 'SNPs' (single nucleotide polymorphisms) because they affect a single base



Causal Mutation Discovery: GWAS

Manhattan plot



Causal Mutation Discovery: WARNING

- Small effect
- Strong disequilibrium
- Many candidates
- Small samples

It is going to be VERY difficult to prove causality

Machine learning, genomic selection

Machine learning is the discipline that develops algorithms allowing computer learning from data and making predictions on future data, e.g., to predict likelihood that someone suffers from a disease based on molecular information from healthy and affected people.

Supervised learning: predicting an output variable from high-dimensional observations

- Nearest neighbor and the curse of dimensionality
- Linear model: from regression to sparsity
- Support vector machines (SVMs)

Model selection: choosing estimators and their parameters

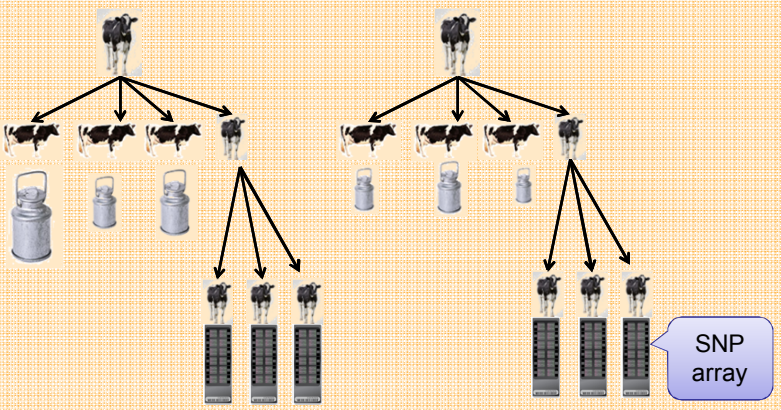
- Score, and cross-validated scores
- Cross-validation generators
- Grid-search and cross-validated estimators

Unsupervised learning: seeking representations of the data

- Clustering: grouping observations together
- Decompositions: from a signal to components and loadings

scikit-learn
Machine Learning in Python

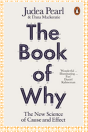
What is the main direction today: Genomic selection



- Pedigree replaced by molecular (SNP based) relationships.
- DNA can be obtained right after birth, and we do need to wait for the cows to produce milk.
- Similar to a multivariate regression.

(DIFFICULT) QUESTION

Do you think Big Data will help to solve causality?



Genome Assembly

Full Paper

Whole Genome Sequencing of Turbot (*Scophthalmus maximus*; Pleuronectiformes): A Fish Adapted to Demersal Life

Antonio Figueras[1,*], Diego Robledo[2], André Corvelo[3,4], Miguel Hermida[5],

An active area of research, practical and evolutionary interest
Huge variety in genome sizes and complexities:

Rice	0.4 Gb	Mammals	3.2 Gb
Maize	2.5 Gb	Chicken	1.2 Gb
Wheat	17 Gb	Fishes	0.3 -130 Gb
Conifers	25 Gb	Turbot	0.5 Gb

Sequence Analysis

- New (> 2008) sequencing technologies have revolutionized genomics. There are already databases with thousands of human complete genomes.
- Bioinformatics is critical to profit from the deluge of data.

Sequence can be used for numerous applications:

- Polymorphism discovery.
- Transcriptome analysis (RNAseq)
- Epigenome analysis.
- Motif, nucleosome discovery (ChIP-seq, Dnase1-seq…)
- Metagenome
- …

Final remarks

- ✓ Large number of species of interest, with wide genome size ranges and genomics knowledge.
- ✓ Specific issues in agriculture: genomic selection, genome assembly, ...
- ✓ Wide genetic variability.
- ✓ Possibility of carrying out ambitious experiments.

Take home message

- ✓ There are no simplistic responses to complex scenarios.
- ✓ Bioinformatics can give you an initial advantage but more important is to have good questions.
- ✓ Learn as much Statistics as you can.