

# Bioinformatics and Agriculture: an old marriage of convenience



Miguel Pérez-Enciso

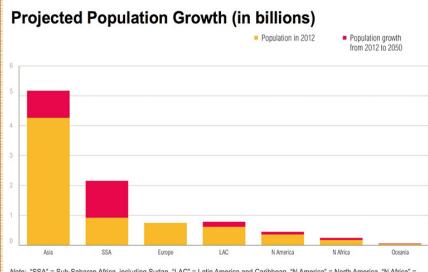
(miguel.perez@uab.es)



## 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.

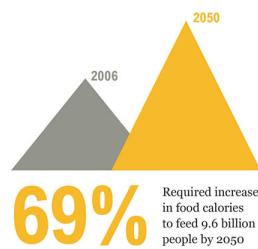
# Global Food Challenge



Note: "SSA" = Sub-Saharan Africa, including Sudan. "LAC" = Latin America and Caribbean. "N America" = North America. "N Africa" Rest of Africa.

WORLD RESOURCES INSTITUTE

Sources: <http://www.jy/rpfMM>



Required increase  
in food calories  
to feed 9.6 billion  
people by 2050

69%

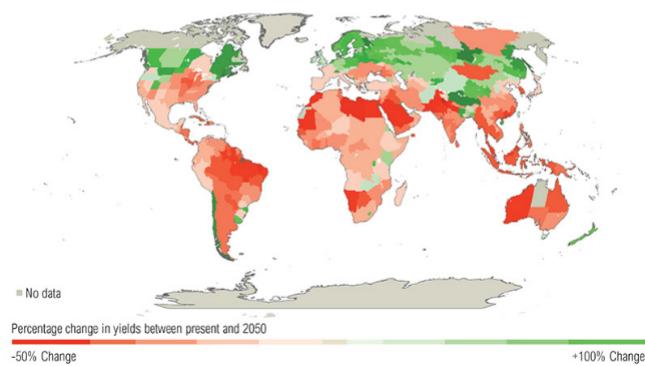
Sources: <http://ow.ly/rpfMN>

<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.

Most studies now project adverse impacts on crop yields due to climate change (3°C warmer world)



WORLD RESOURCES INSTITUTE

Sources: <http://ow.ly/rp1MN>

## The Green Revolution



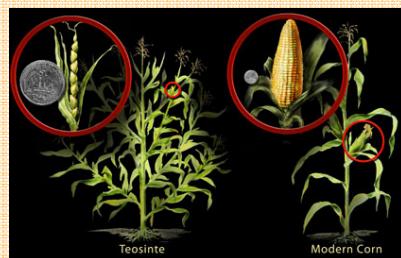
**Norman Borlaug** (Peace Nobel prize, 1970): He developed semi-dwarf, high-yield, disease-resistant wheat varieties in Mexico. During the mid-20th century, Borlaug led the introduction of these high-yielding varieties combined with modern agricultural production techniques to Mexico, Pakistan, and India. 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

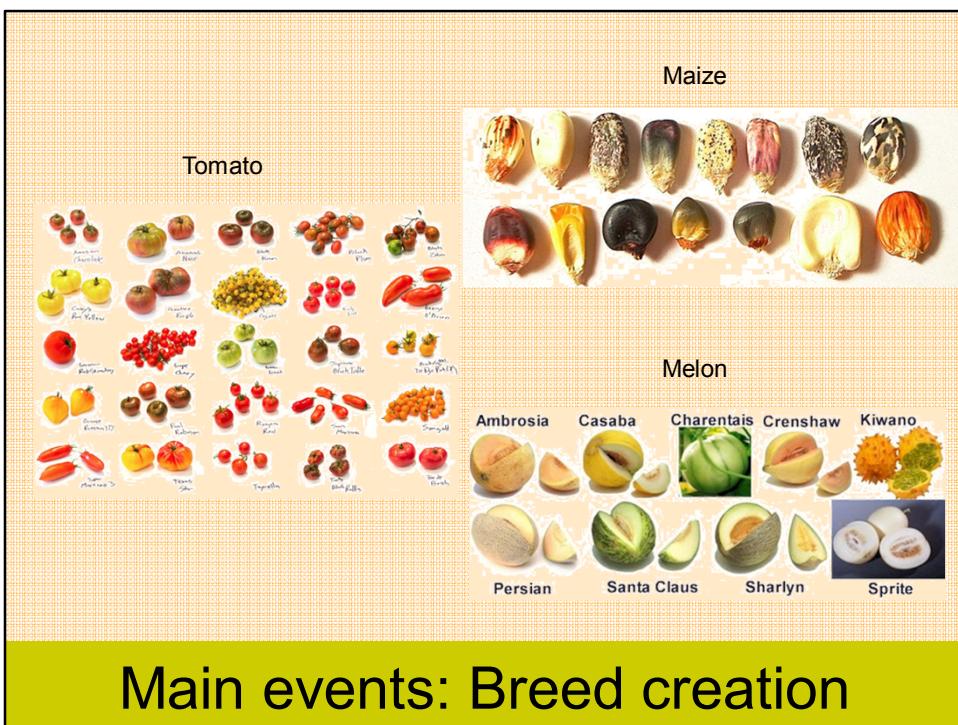
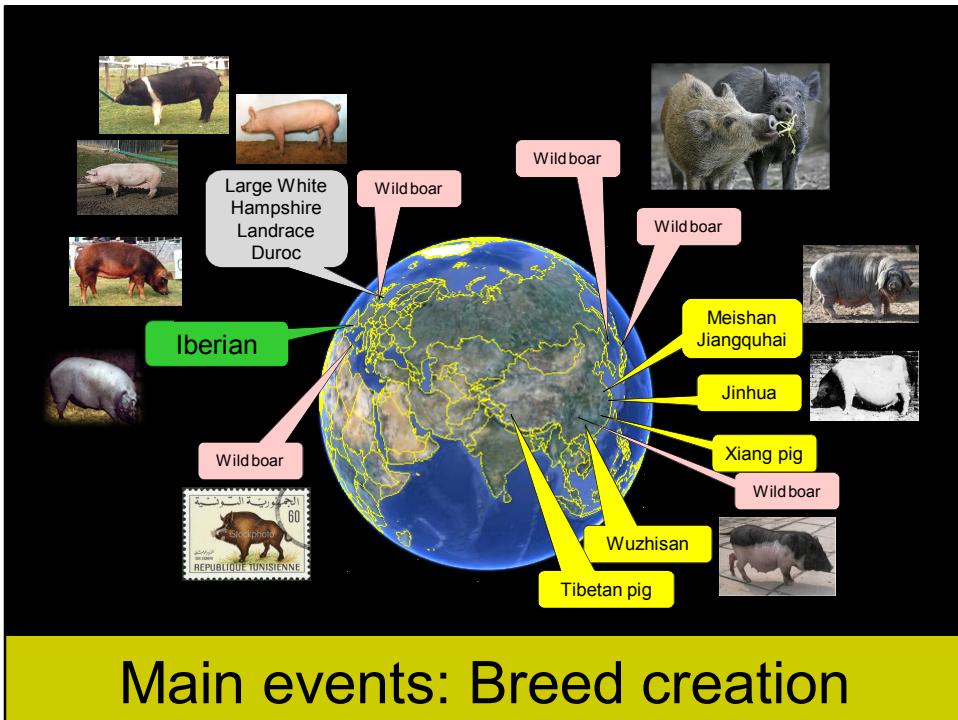


Wild boar

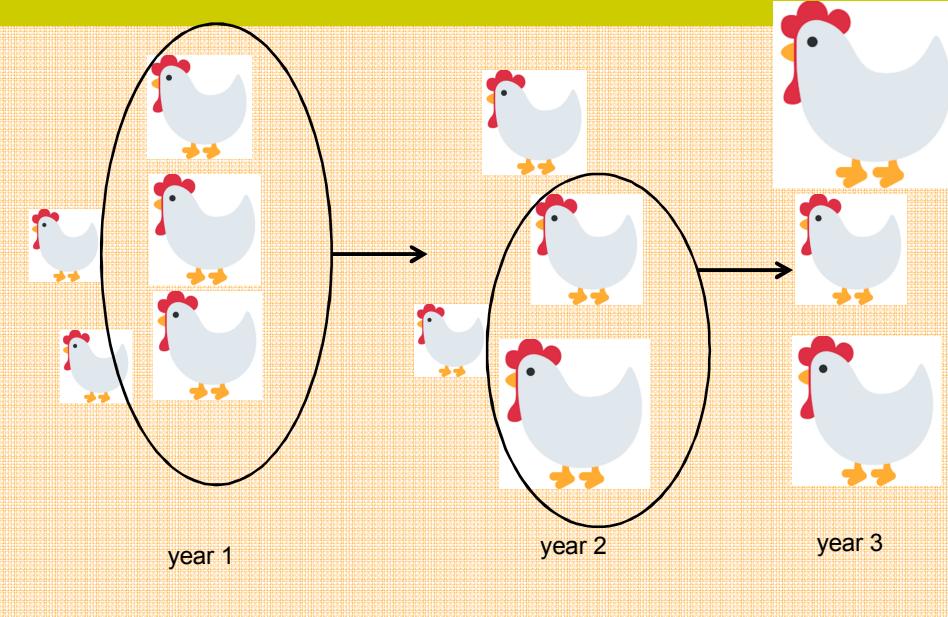


Pig

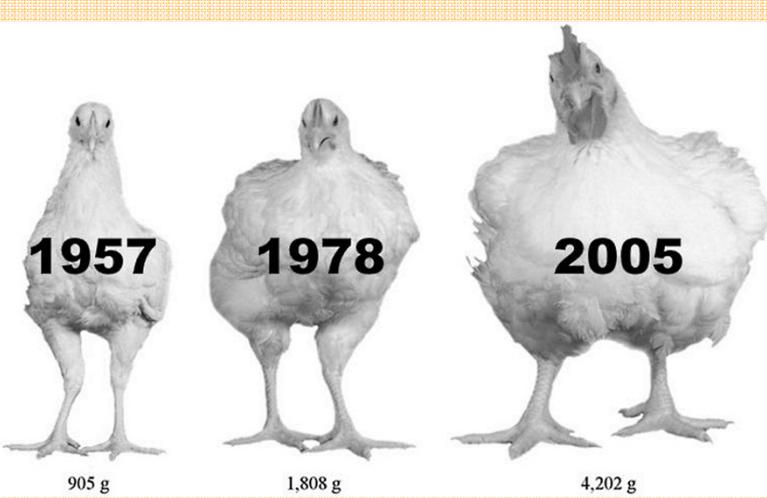
Asia and Anatolia,  
~ 9,000 years ago



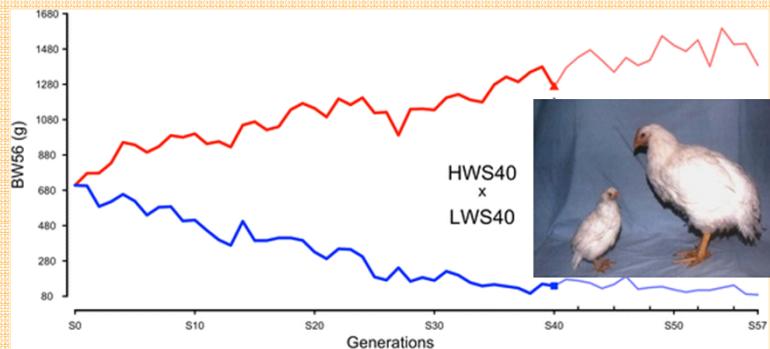
## Main events: Artificial Selection



## Increase in average performance: chicken



## How did selection work in animals?

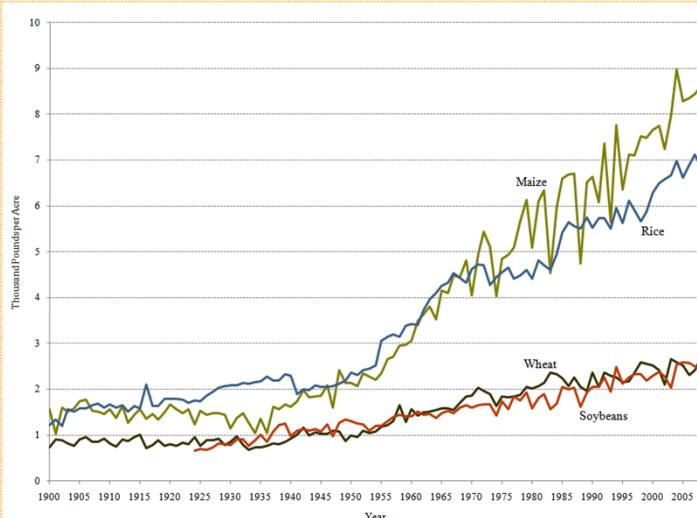


**Long-Term Divergent Selection for Eight-Week Body Weight in White Plymouth Rock Chickens**

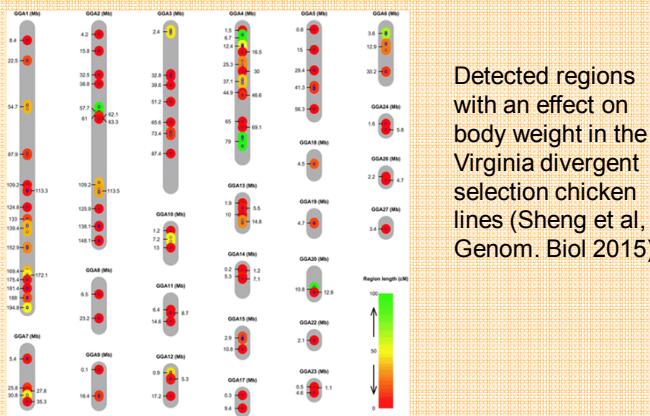
E. A. DUNNINGTON and P. B. SIEGEL

*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**



This is called ‘The infinitesimal model’

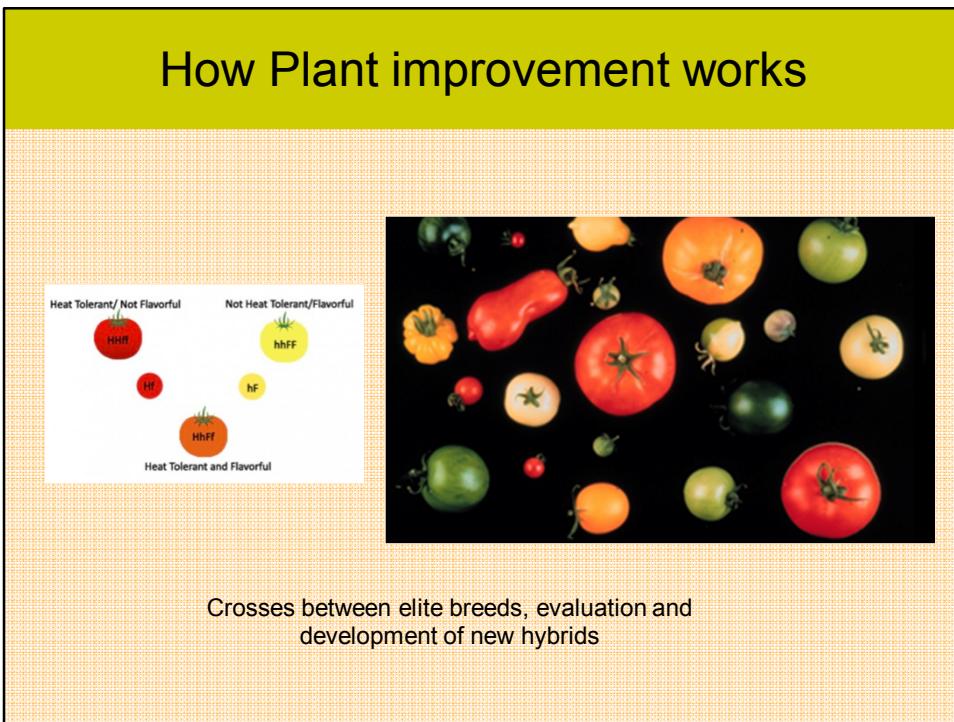


It presupposes that quantitative traits are explained by a large number of genes, each acting individually and of small effect. In addition, quantitative traits are also modified by the environment.

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 16, 1918. Read July 8, 1918. Issued separately October 1, 1918.)

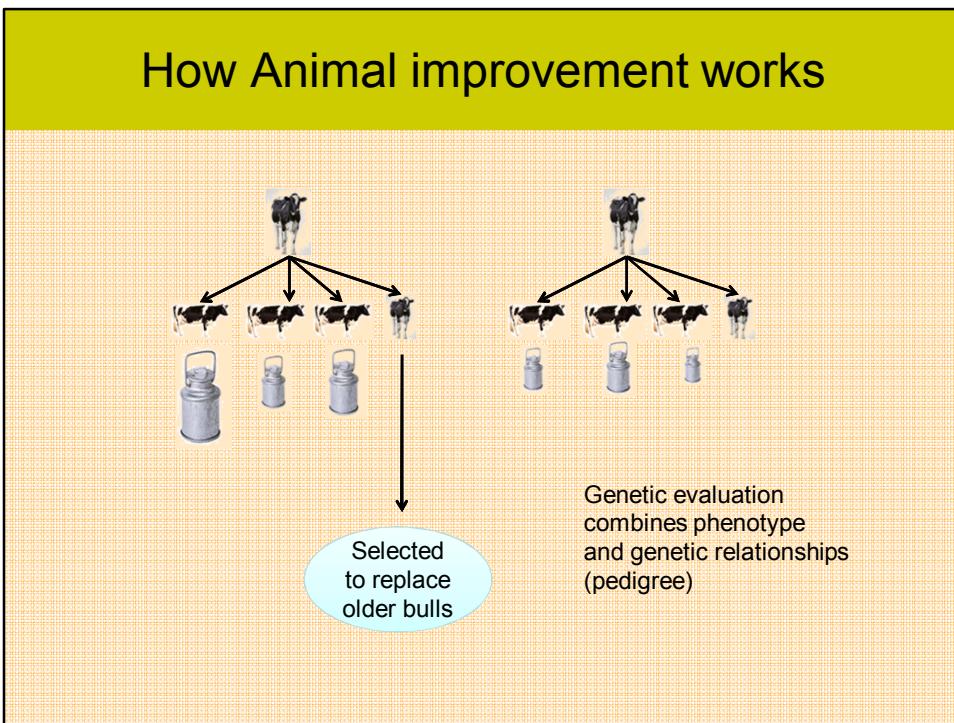
## How Plant improvement works



The diagram illustrates the process of plant improvement. On the left, a genetic trait matrix shows combinations of heat tolerance (H) and flavor (F). The rows represent heat tolerance: 'Heat Tolerant/Not Flavorful' (HhFF), 'Heat Tolerant and Flavorful' (HhFF), and 'Not Heat Tolerant/Flavorful' (hhFF). The columns represent flavor: 'Not Heat Tolerant/Flavorful' (hf), 'Heat Tolerant/Not Flavorful' (Hf), and 'Heat Tolerant and Flavorful' (HhFF). A photograph on the right shows a variety of colorful tomatoes (red, yellow, green) against a black background.

Crosses between elite breeds, evaluation and development of new hybrids

## How Animal improvement works



The diagram illustrates the process of animal improvement. It shows two groups of cattle. The left group has four cattle, and the right group has three cattle. Arrows point from each group to three milk cans below them. A large arrow points down to a light blue oval containing the text: "Selected to replace older bulls". To the right of the oval is a pedigree chart showing a bull at the top, followed by four cows, and then three milk cans at the bottom. The text "Genetic evaluation combines phenotype and genetic relationships (pedigree)" is written next to the pedigree chart.

Selected to replace older bulls

Genetic evaluation combines phenotype and genetic relationships (pedigree)

## Summary so far

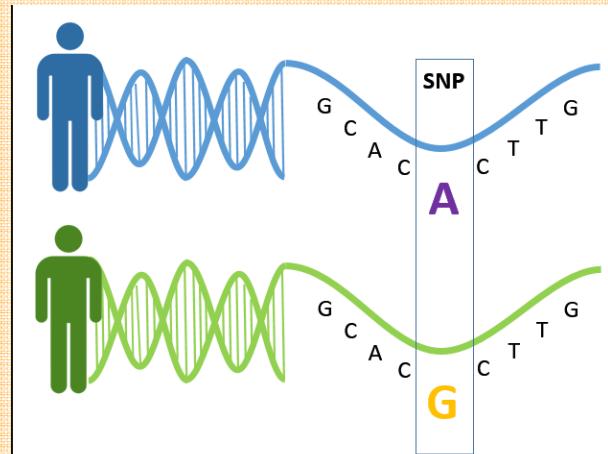
- ✓ Artificial selection is very effective in changing the phenotypes.
- ✓ Many genes involved: no exhaustion of genetic variability (this is called the INFINITESIMAL MODEL).
- ✓ Artificial selection is based on pedigree and phenotypic information.
- ✓ Pedigree information is being replaced by molecular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

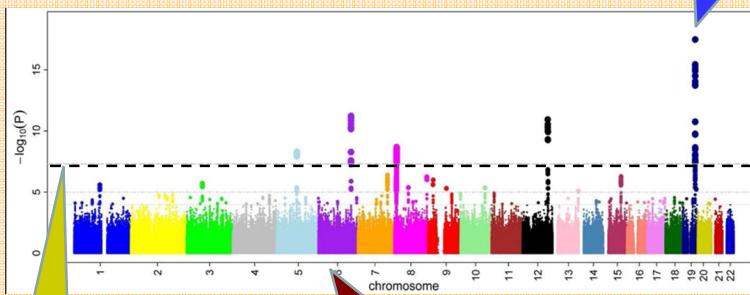
There are millions of small DNA differences between individuals called 'SNPs' (single nucleotide polymorphisms) because they affect a single base



## Causal Mutation Discovery: GWAS

### Manhattan plot

Each dot represents the P-value of a SNP



Significance threshold

Each color represents SNPs of different chromosomes

## 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

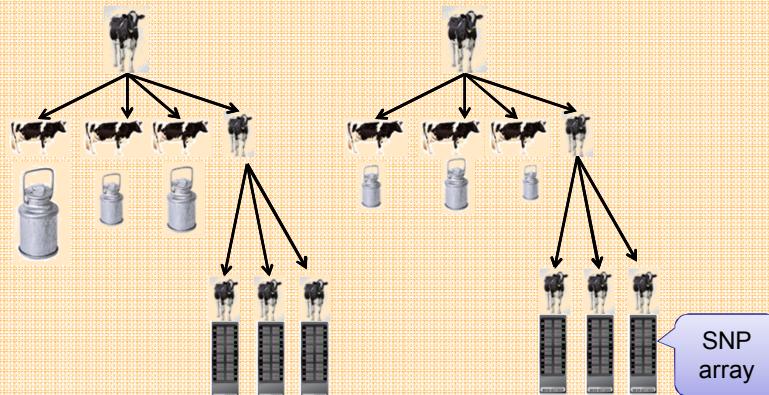
- Score, and cross-validated scores
- Cross-validation generators
- Grid-search and cross-validated estimators

### 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

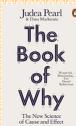
## What is the main direction today: Genomic selection



- 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**



**The genome of melon (*Cucumis melo* L.)**

Jordi Garcia-Mas<sup>1</sup>, Andrej Berkajik<sup>2</sup>, Walter Sanseverino<sup>2</sup>, Michael Bourque<sup>3</sup>, Gladys Mir<sup>4</sup>, Victor M. Gonzalez<sup>2</sup>, Elizabeth Henaff<sup>2</sup>, Francisco Climent<sup>5</sup>, Luca Cozzuto<sup>6</sup>, Ernesto Lowy<sup>7</sup>, Tyler Alstro<sup>8</sup>, Salvador Capella-Gutiérrez<sup>2</sup>, Jose Blanca<sup>2</sup>, Joaquín Calafate<sup>9</sup>, Pello Zárrate<sup>10</sup>, Daniel González-Ibáñez<sup>11</sup>, Luis Rodríguez-Moreno<sup>12</sup>, Marcus Droege<sup>13</sup>, Lei Du<sup>14</sup>, Miguel Álvarez-Tejedor<sup>15</sup>, Belén Gómez-Plaza<sup>16</sup>, Marta Gómez-Plaza<sup>16</sup>, Luming Yang<sup>17</sup>, Yiqun Weng<sup>17</sup>, Arcadi Navarro<sup>18</sup>, Jordi Garcia-Mas<sup>1</sup>, Josep M. Casas<sup>19</sup>, Ferran S. Roig<sup>19</sup>, Toni Galalde<sup>20</sup>, Guglielmo Renna<sup>21</sup>, Roderic Guigó<sup>22</sup>, Josep M. Casas<sup>23</sup>, Pere Arús<sup>24</sup>, and Pere Pujolménec<sup>25</sup>

<sup>1</sup>Institut de Recerca i Tecnologia Agroalimentàries, Centre for Research in Agricultural Genomics Consell Superior d'Investigacions Científiques-Institut de Recerca i Tecnologia Agroalimentàries-Universitat Autònoma de Barcelona-Barcelona-Universitat de Barcelona, 08193 Barcelona, Spain; <sup>2</sup>Centre for Research in Agricultural Genomics Consell Superior d'Investigacions Científiques-Institut de Recerca i Tecnologia Agroalimentàries-Universitat Autònoma de Barcelona





Cruz et al. *GigaScience* (2016) 5:29  
DOI 10.1186/s13742-016-0136-4

**DATA NOTE** **Open Access**

Genome sequence of the olive tree, *Olea europaea*

Fernando Cruz<sup>1,2</sup>, Irene Juárez<sup>1,3</sup>, Jésica González-García<sup>1,2</sup>, Damián Loska<sup>1,2</sup>, Marina Martínez-Houben<sup>1,2</sup>, Esteban Canvi<sup>1</sup>, Belén Galán<sup>1</sup>, Leonor Fria<sup>1,2</sup>, Pablo Ríosca<sup>1,2</sup>, Sophie Derdik<sup>1,2</sup>, Marta Gómez<sup>1,2</sup>, Manuel Gómez-Pérez<sup>1,2</sup>, José Luis García<sup>1</sup>, Ivo G. Gut<sup>1</sup>, Pablo Vergara<sup>1,2</sup>, Tyler S. Alton<sup>1,2</sup> and Toni Galalde<sup>1,2,3,4,5</sup>

**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.



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

Sequence can be used for numerous applications:

## 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.