Bioinformatics and Agriculture:

Miguel Pérez-Enciso

(miguel.perez@uab.es)

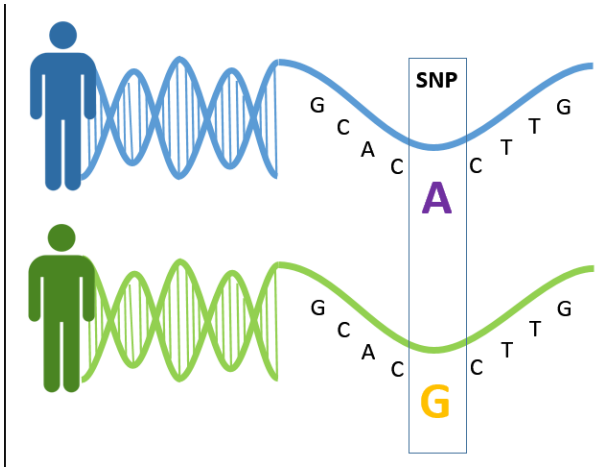
ICREA CRAG UAB Universitat Autònoma de Barcelona

Some Bioinformatics Applications

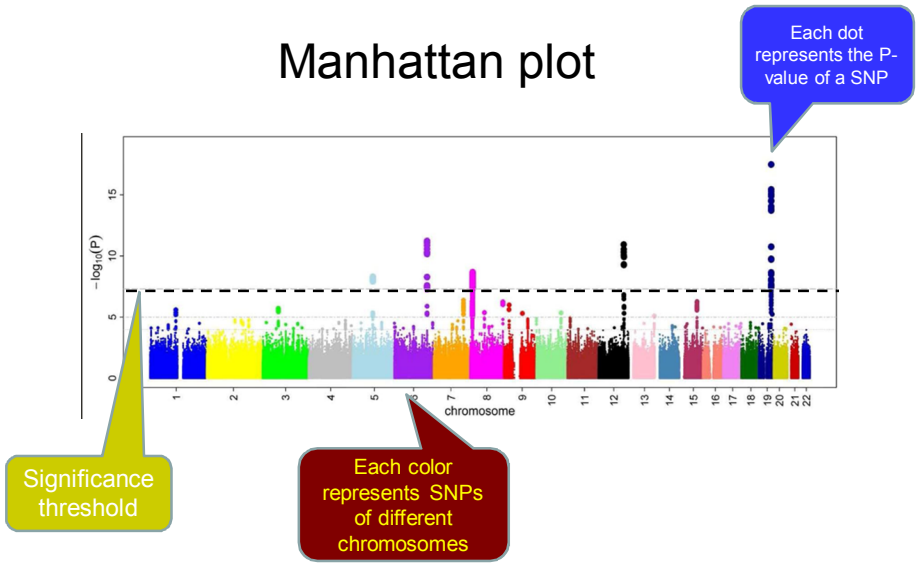
- ✓ Precision agriculture (Big data issues, modeling …)
- ✓ Causal mutation discovery
- ✓ Genomic selection, machine learning
- ✓ Genome assembly
- ✓ Sequence analysis
- ✓ Functional analysis
- ✓ …

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



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

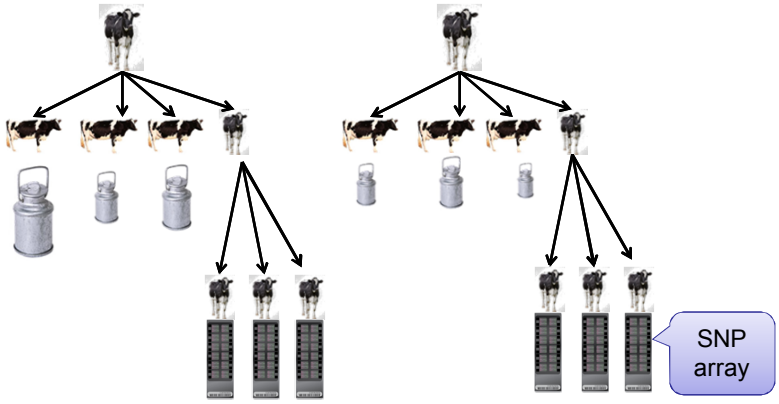
- Score, and cross-validated scores
- Cross-validation generators
- Grid-search and cross-validated estimators

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



SNP data is used to predict future performance.

Genome Assembly

PNAS

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

Jordi Garcia-Mas[a,1], Andrej Benjak[a], Walter Sanseverino[a], Michael Bourgeois[a], Gisela Mir[a], Victor M. González[b], Elizabeth Hénaff[b], Francisco Cámara[a], Luca Cozzuto[c], Ernesto Lowy[c], Tyler Alioto[d], Salvador Capella-Gutiérrez[c], Jose Blanca[e], Joaquín Cañizares[e], Pello Ziarsolo[e], Daniel Gonzalez-Ibeas[f], Luis Rodriguez-Moreno[f], Marcus Droege[g], Lei Du[h], Miguel Alvarez-Tejado[i], Belen Lorente-Galdos[i], Marta Melé[i,2], Luming Yang[g], Yiqun Weng[g,3], Arcadi Navarro[i,j,k], Tomas Marques-Bonet[i,m], Miguel A. Aranda[f], Fernando Nuez[e], Belén Picó[e], Toni Gabaldon[c], Guglielmo Roma[c], Roderic Guigó[c], Josep M. Casacuberta[b], Pere Arús[a], and Pere Puigdomènech[a,1]

[a]Institut de Recerca i Tecnologia Agroalimentàries, Centre for Research in Agricultural Genomics Consejo Superior de Investigaciones Científicas-Institut de Recerca i Tecnologia Agroalimentàries-Universitat Autònoma de Barcelona, 08193 Barcelona, Spain; [b]Centre for Research in Agricultural Genomics Consejo Superior de Investigaciones Científicas-Institut de Recerca i Tecnologia Agroalimentàries-Universitat Autònoma de Barcelona-

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],

Cruz et al. GigaScience (2016) 5:29
DOI 10.1186/s13742-016-0134-5

GigaScience

DATA NOTE Open Access

Genome sequence of the olive tree, *Olea europaea*

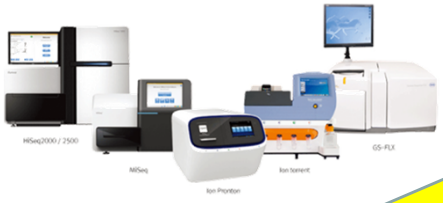
Fernando Cruz[1,2†], Irene Julca[2,3,4†], Jéssica Gómez-Garrido[1,2], Damian Loska[2,3], Marina Marcet-Houben[2,3], Emilio Cano[5], Beatriz Galán[6], Leonor Frias[1,2], Paolo Ribeca[1,2], Sophia Derdak[1,2], Marta Gut[1,2], Manuel Sánchez-Fernández[7], Jose Luis García[6], Ivo G. Gut[1,2], Pablo Vargas[7,8*†], Tyler S. Alioto[1,2,3*†] and Toni Gabaldón[2,3,8,9*†]

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
- …

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.

(DIFFICULT) QUESTION

Do you think Big Data will help to solve causality?