Hrvatsko Društvo za Biotehnologiju

Data science in biotechnology

("Podatkovna znanost" u biotehnologiji)

Želimir Kurtanjek

PBF

28 ožujak 2018

Agenda

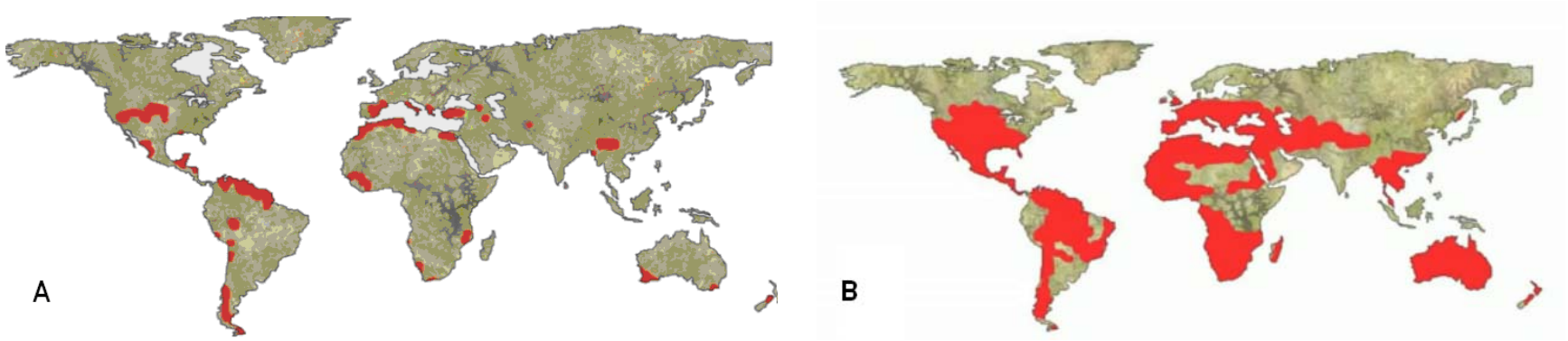
1) Why data science in biotechnology ?
2) EU RDA and project data management
3) Data science vs statistics
   - Data mining, Big data analytics
   - Computer science skills
4) Development (integration of engineering and biology)
5) Deep learning and Decision trees

Examples

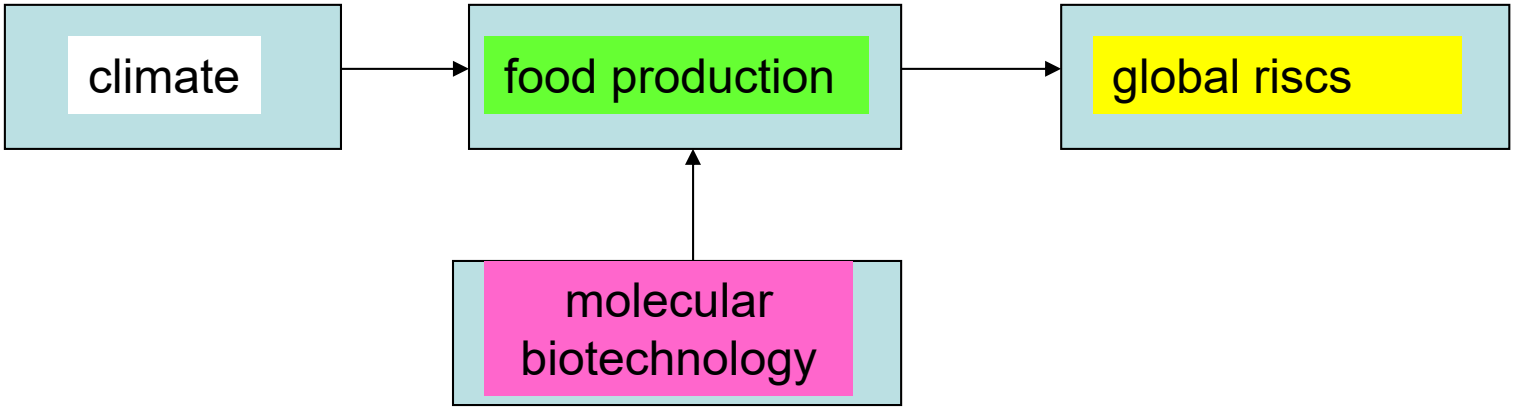
6) GWAS of yeast stress signature
7) DArT wheat DecisionGS
8) Prospects / Conclusions

Discussion

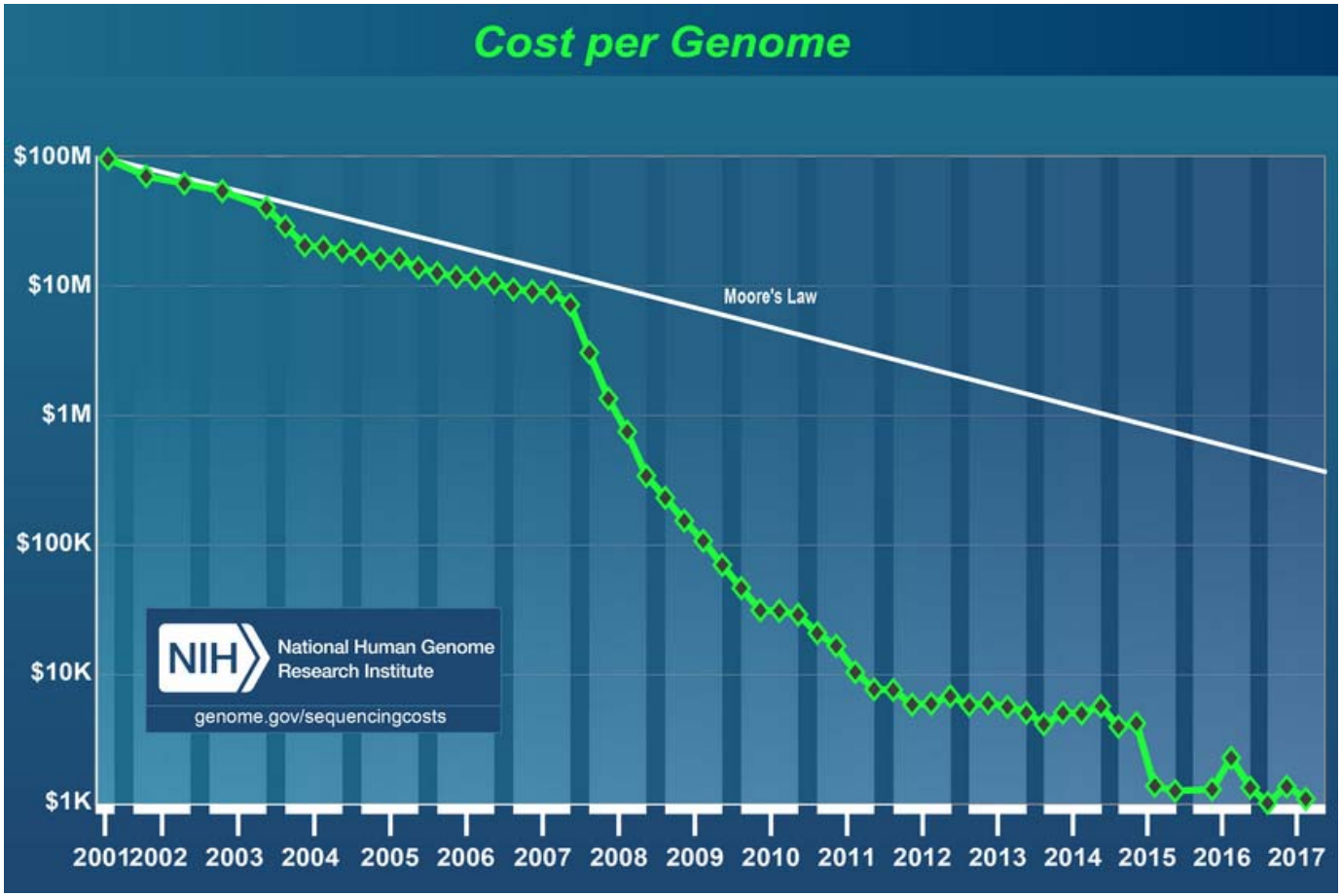
Global climate change presents challenges for biotechnology



Distribution of global aridity land, A: 2011, B: prediction 2050,
(Dai. A., 2011, J. Farrant, 2016)



Biotechnology: Dominance of OMICS data



https://www.genome.gov/images/content/costpergenome_2017.jpg

yeast genome 5700

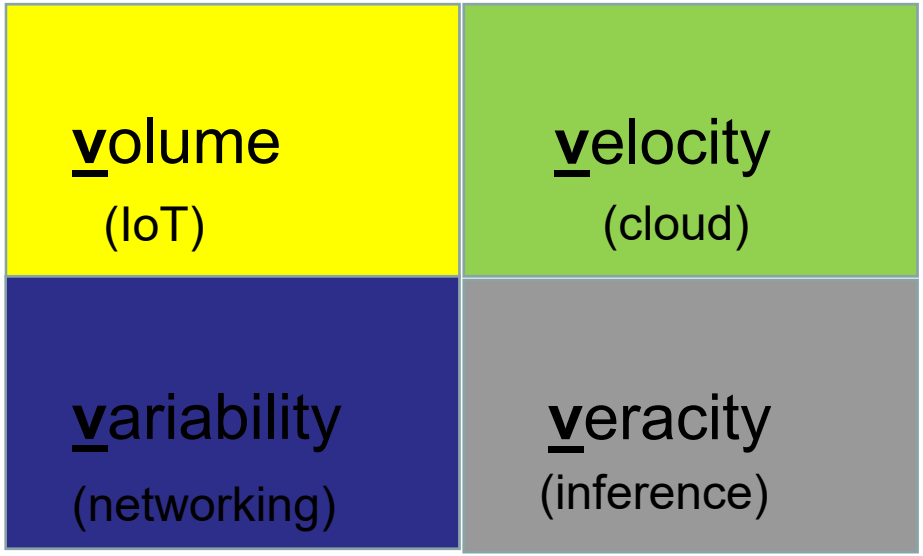
human genome 21 000; protein 100 000; PPI 6.7/ protein

wheat genome 100 000

What are, and why are big data science important in biotechnology?

W
h
y

4 V's of „Big data analytics"



Big data in biotechnology are large sets of multi scale intricate complexity of molecular information, phenotypes and exogenous features

The main benefit of „big data analytics" is algorithmic inference

Data science versus statistics

Scientific fields:

Statistics: mathematics

Data science: computer science + applied statistics

Objectives:

Statistics: Hypothesis testing

Data science: algorithmic knowledge inference

Concepts:

Statistics: probability density function

Data science: AI data patterns

[Blei and Smyth 2017] David Blei and Padhraic Symth, “Science and Data Science,” *Proceedings of the National Academies of Sciences*, vol. 114, no. 33, June 2017, pp. 8689-8692.

Needed skills for data science in biotechnology

Field knowledge (technology + bioinformatics)

Mathematics + statistics

Computer software

- **SAS**
- **Phyton**
- **R**
- **++++++**

Examples of big data in biotechnology

Biochemical engineering

1) Yeast growth study under nutrient limitation in chemostat, source Princeton 36 (factors) x 5700 (gene transcriptions) = 2 M numerical data

Food technology (agronomy + bread making technology)

2) Wheat (*Triticum a.*) source CIMMYT

10819(lines)*23747(SNP)*8(phenotypes) = 2 G numerical and descriptive data

Nutrition

3) Obesity mice study source UK Oxford

15 (external) x 2000 (mice) x 10 000 (gene) = 0.3 G numerical and descriptive data

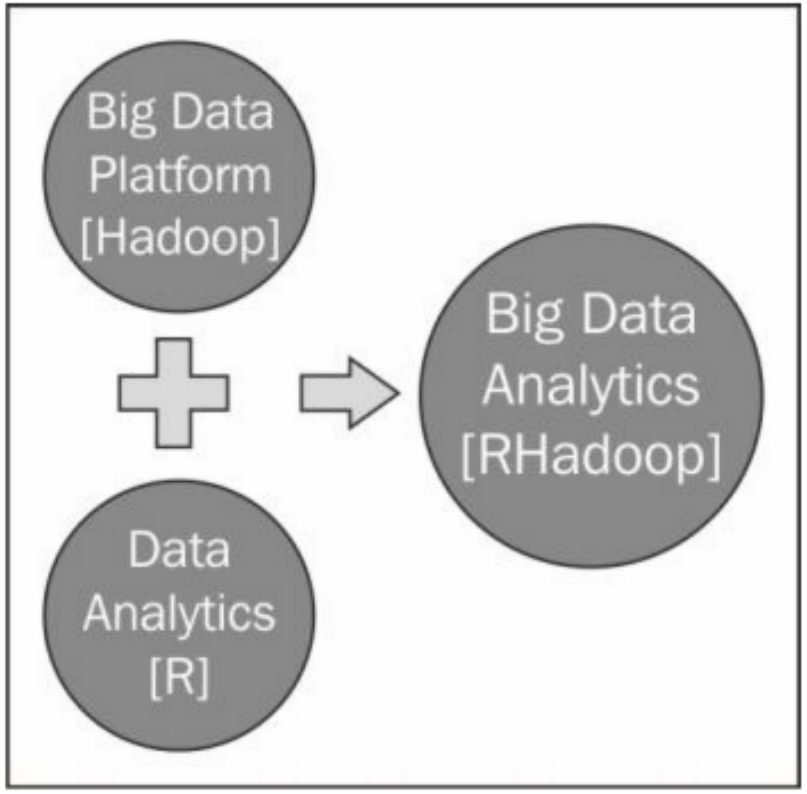
"True" big data studies > 1 T data

Big data (total number of data) is the key factor for generalization and validation in learning and knowledge inference.

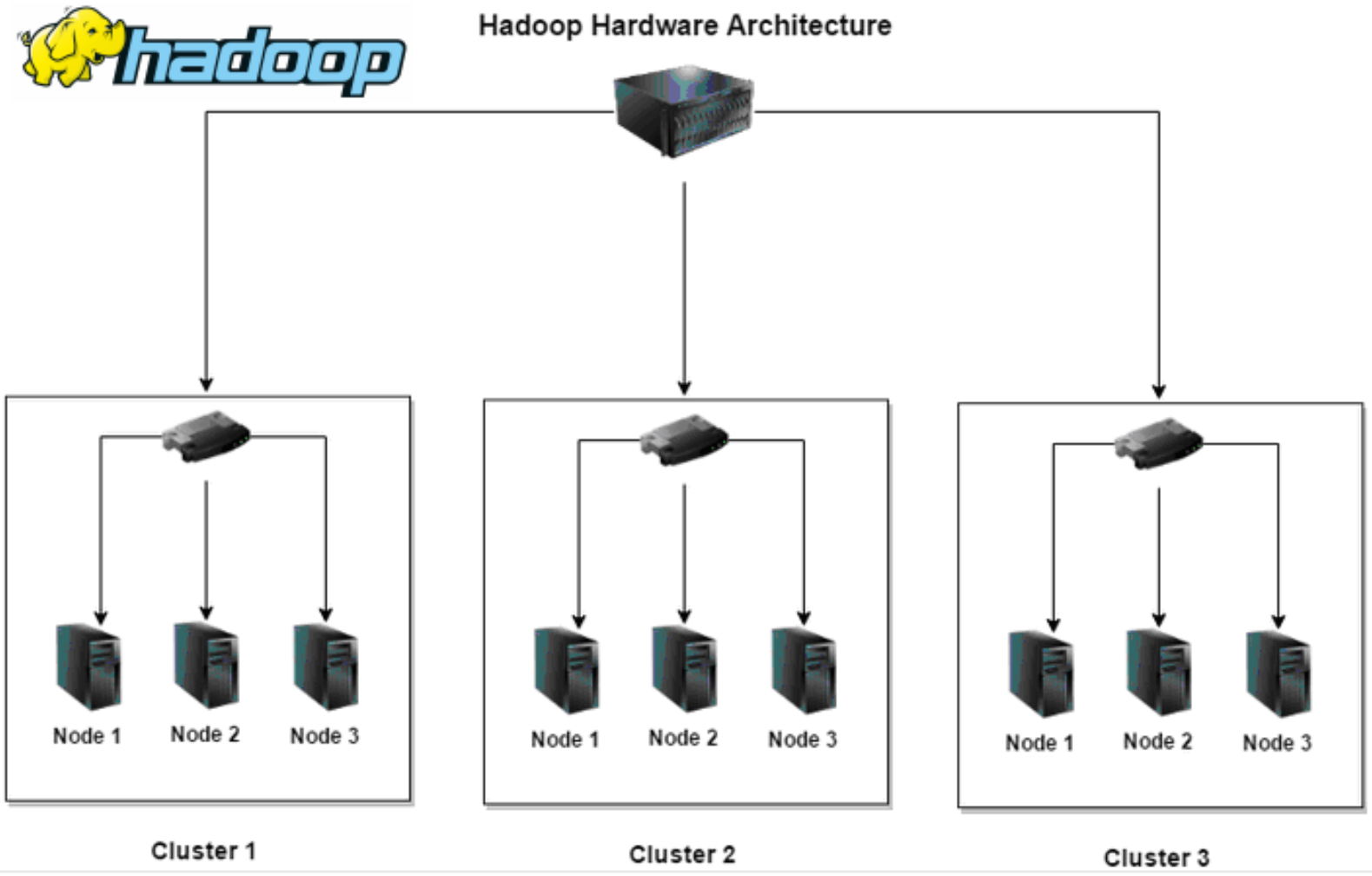
Big Data needs scalable software *R* and parallel computing RHadoop

Apache Hadoop is an open source Java framework for processing and querying vast amounts of data on large clusters of commodity hardware.

Apache Hadoop has become an enterprise ready cloud computing technology. It is becoming the industry de facto framework for Big Data processing.



Big Data computer infrastructure



Apache Hadoop is an open-source software framework for storage and large-scale processing of data-sets on clusters of commodity hardware.

Importance of research data transparency and exchangeability

Robert-Jan Smits EU director-general of research and innovation,

Jean-Claude Juncker proposed Smits as a special envoy on open science at the European Commission, to help push efforts to make **all publicly funded research in Europe freely available by 2020**

Proposition: EU funded project should have **Data management** framework enabling **computer readability (exchangeability)**

https://sciencebusiness.net/framework-programmes/news/eu-research

RDA Groups

- Working Groups
- Interest Groups
- Coordination Groups
- RDA Group Policies
- National Groups

RDA Events

The critical role of university RDM infrastructure in transforming data to knowledge - RDA 11th Plenary Associated event

18 Mar 2018 - 11:45 to 20 Mar 2018 - 11:45

https://www.rd-alliance.org/

Development mile stones (engineering + biology)

1943 MIT
Warren McCullogh
Electrical engineer

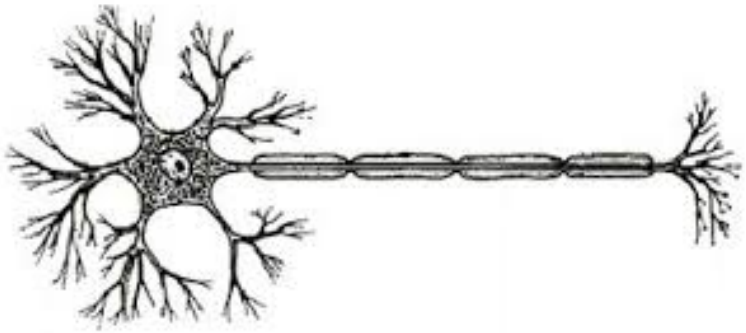
1943 MIT
Walter Pitts
Electrical engineer

1958 Cornell
Frank Rosenblatt
Neural biology

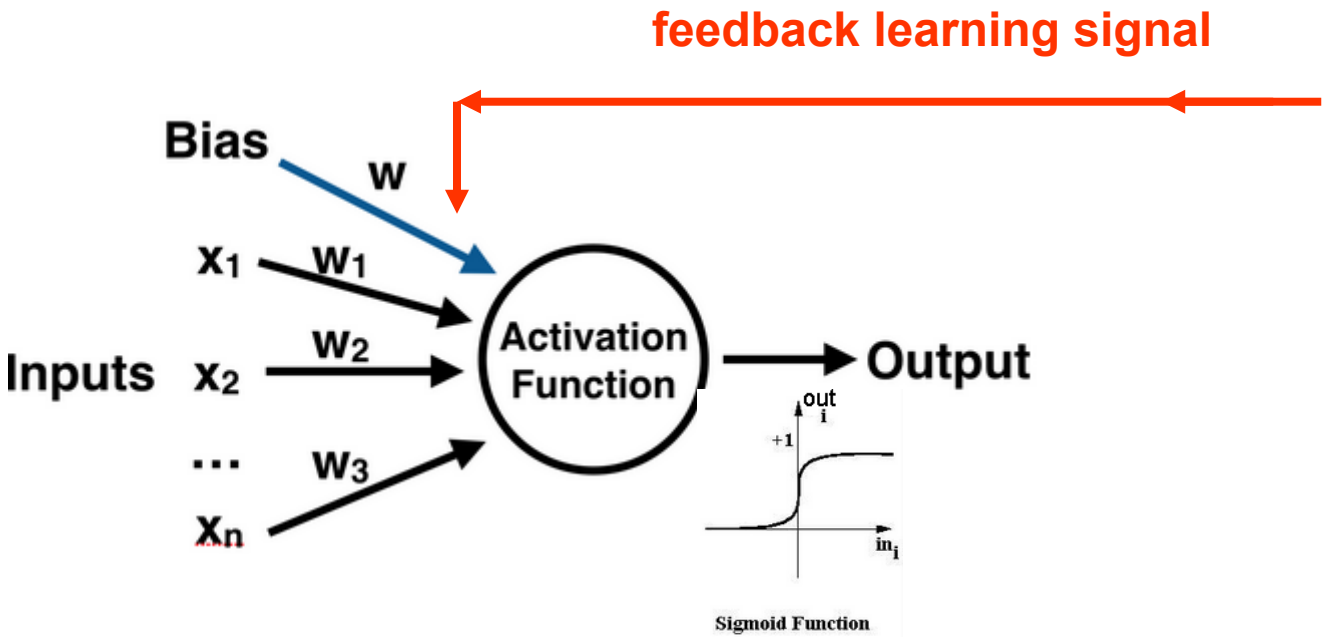
2003 in Dubrovnik
Leo Breiman (Berkley)
mathematican

2012
Jennifer Doudna, Berkely
Molecular biology

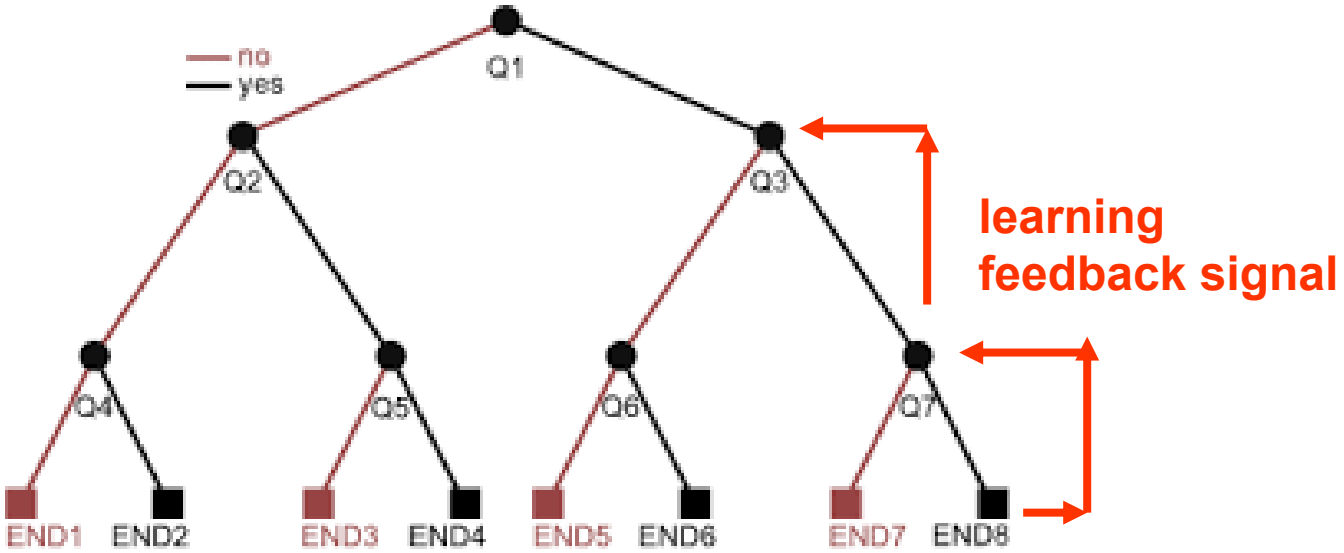
Biological neuron



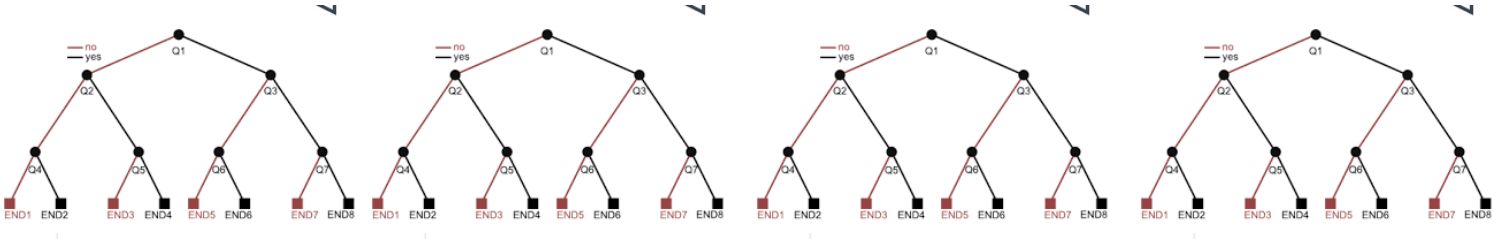
Mathematical neuron



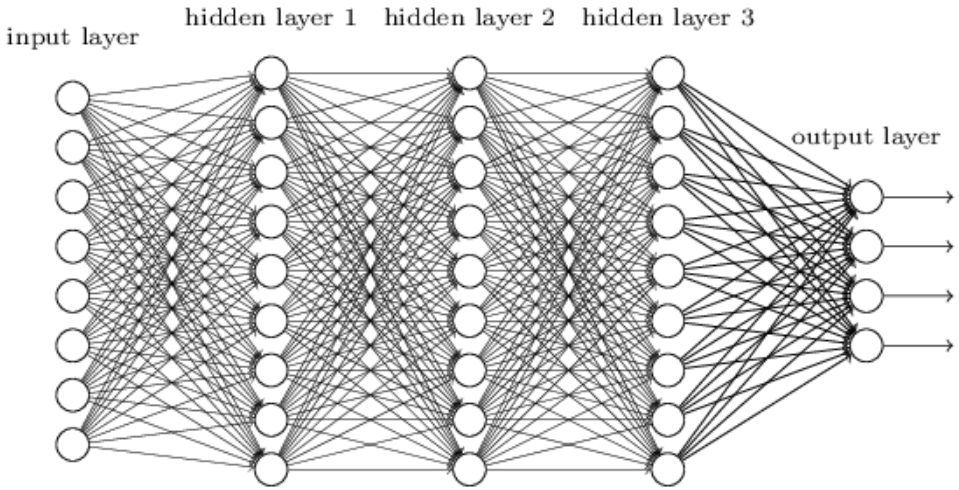
Decision tree

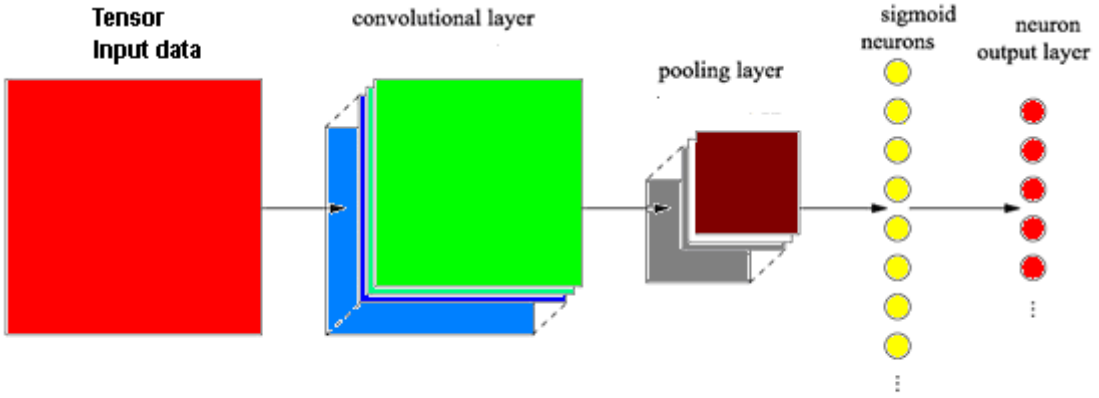


Decision tree forest (network)



Deep learning neural network models





Deep learning structure NN model

EXAMPLES

1. Ž. Kurtanjek

   Genome-wide Big Data analytics: Case of yeast stress signature detection

   The Eurobiotech Journal **1**(4) (2017) 264-270

2. Ž. Kurtanjek

   Wheat DArT based DecisionGS improvement

   (in preparation)

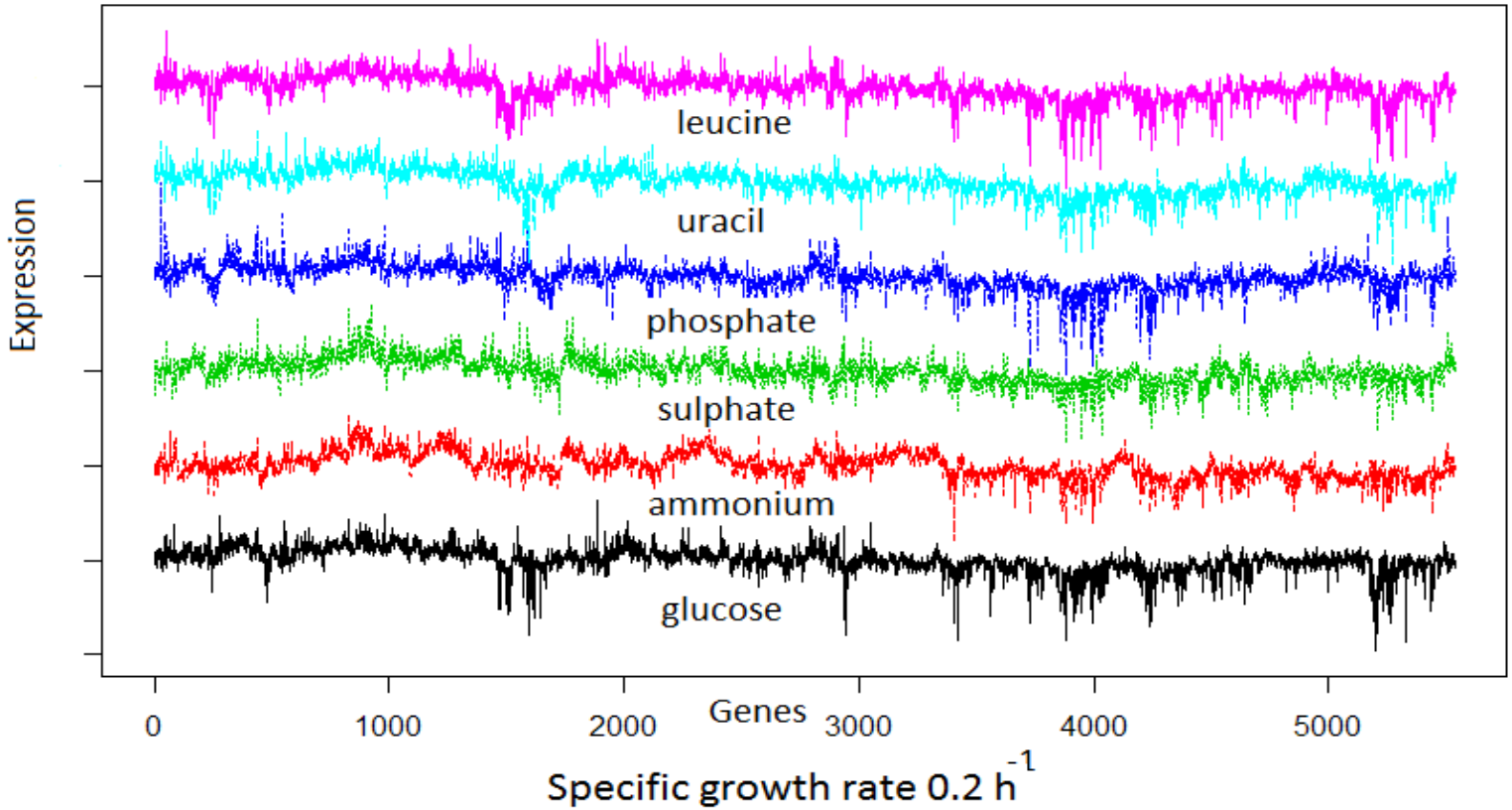
Data source: Lewis-Sigler Institute for Integrative Genomics, Princeton

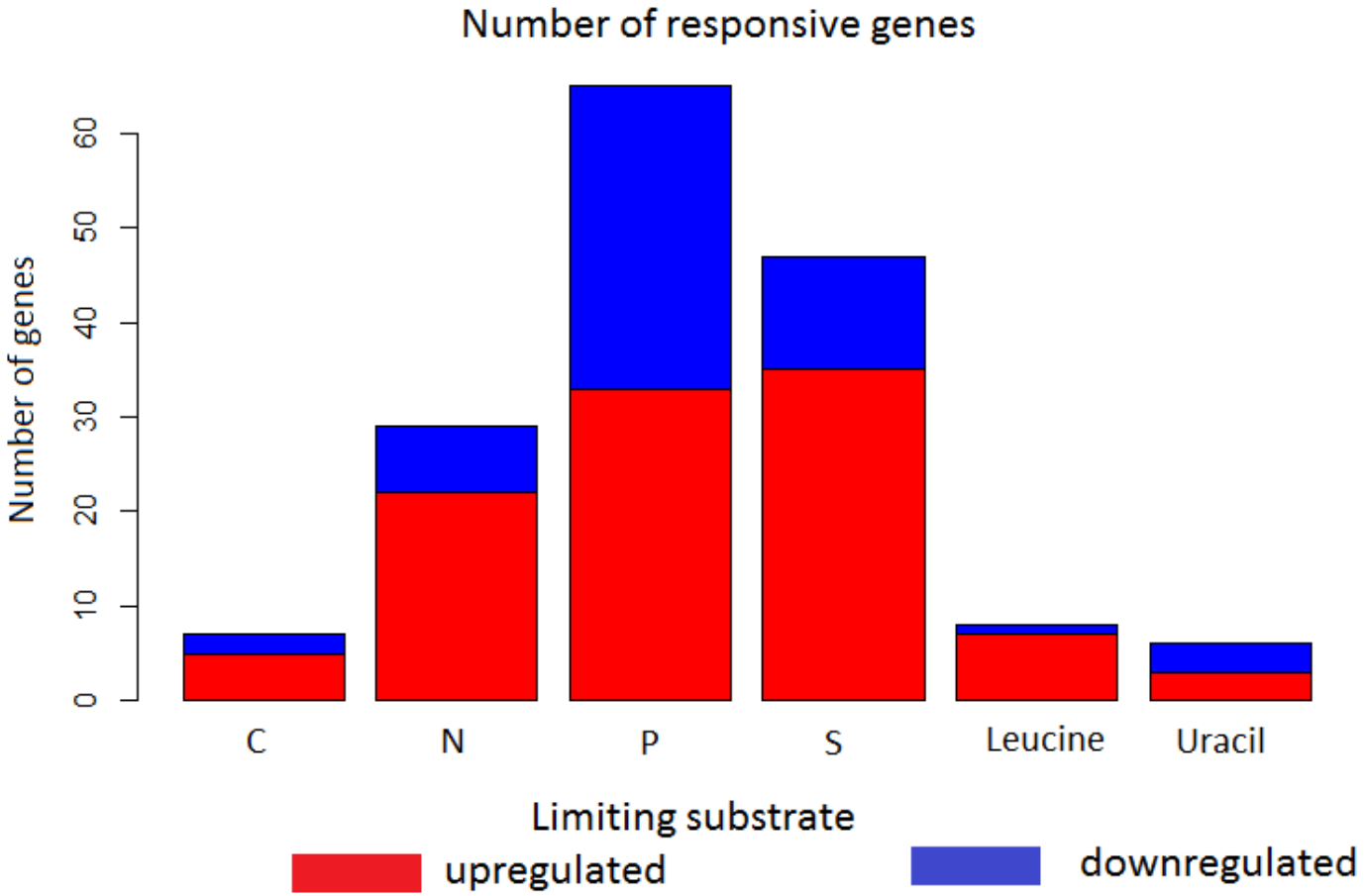
Brauer M.J. et al.Coordination of Growth Rate, Cell Cycle, Stress Response and Metabolic Activity in Yeast. MBoC, Molecular Biology of the Cell, 2008; 19: 352-367.

Experiment: chemostat 0.5 L, dilution rate 0.05-0.35 h^{-1}

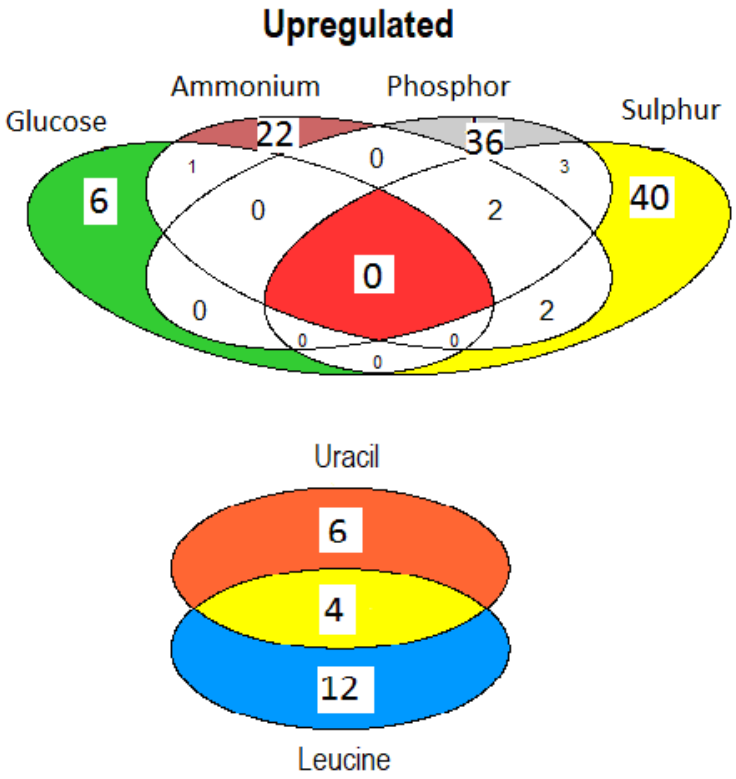
Limiting nutrients: biotic (C,S,N,P) and auxothropic (leucin, uracil)

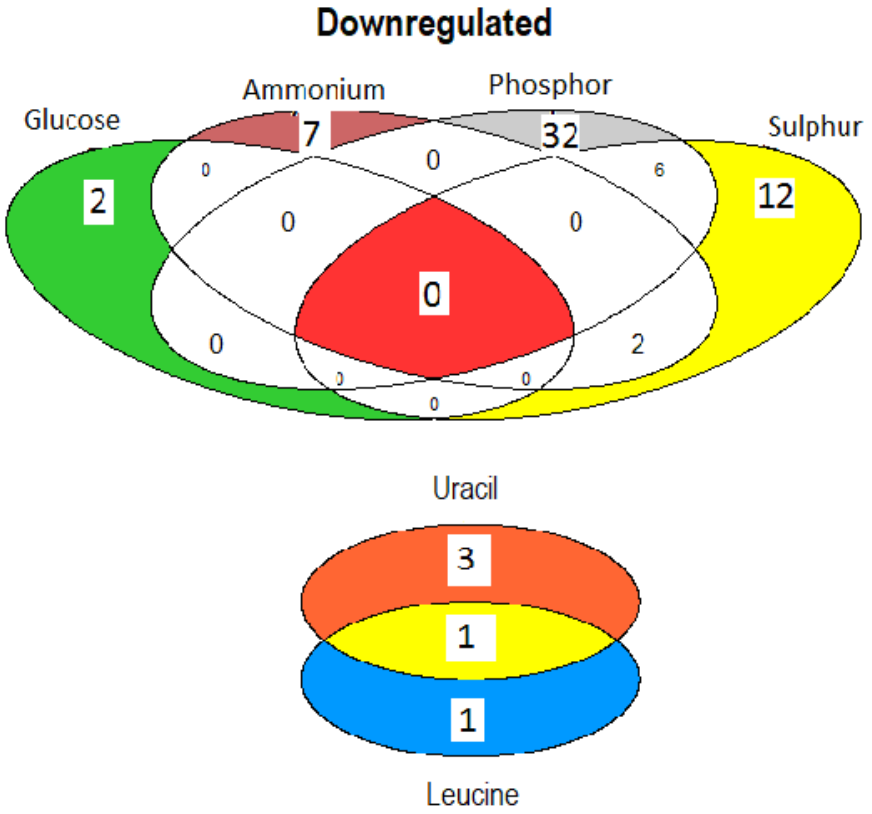
Responses of 5337 genes measured as mRNA abundance

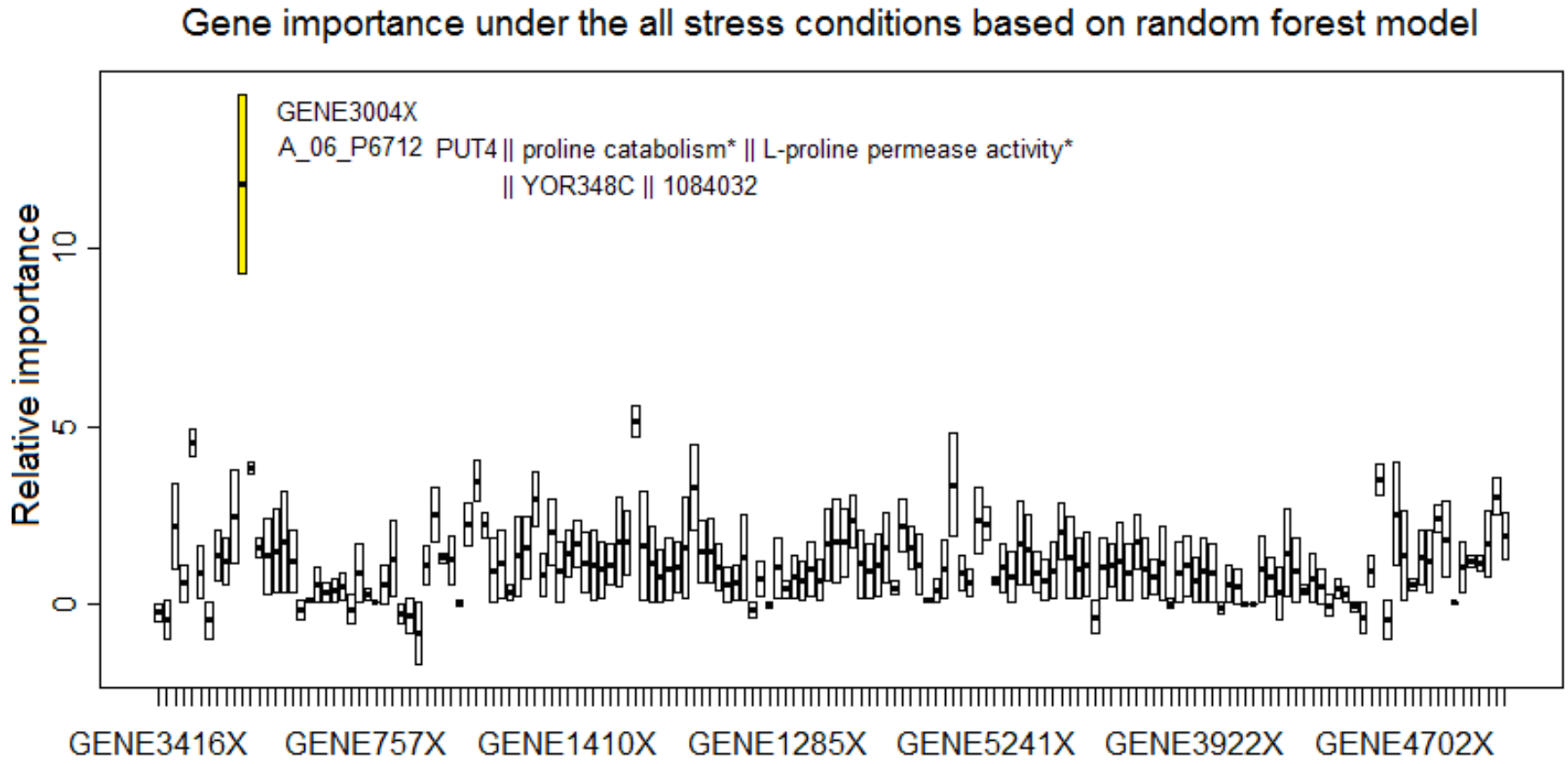



Number of responsive genes
Number of genes
0
10
20
30
40
50
60
C
N
P
S
Leucine
Uracil
Limiting substrate
upregulated
downregulated

Extraction of the key responsive genes by quantile q(x) separation





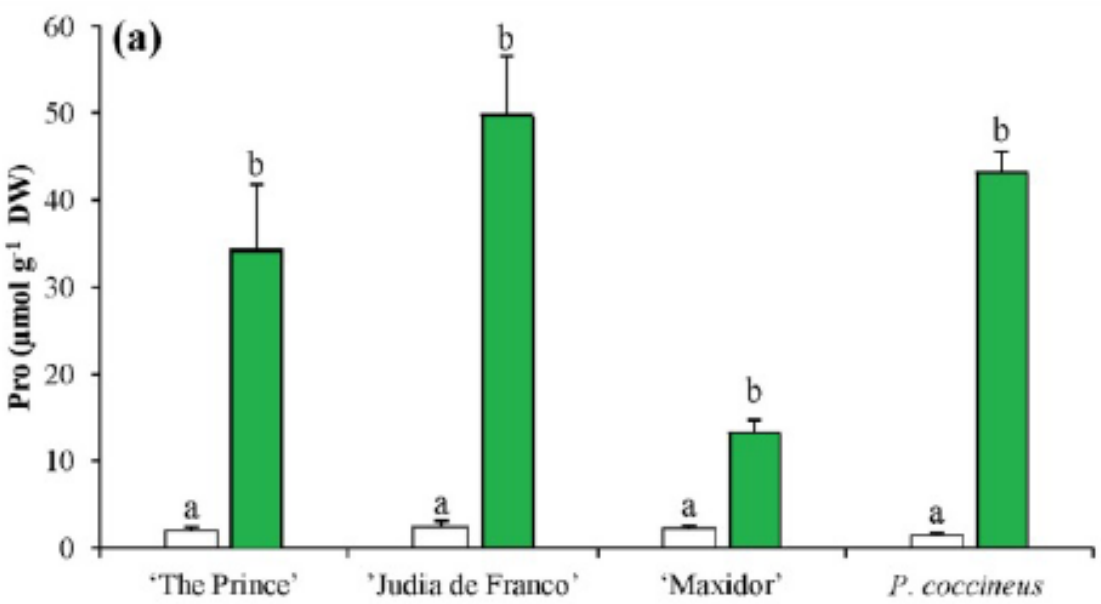


Applied is quantile random forest for stress signature

Literature experimental confirmations

M. Morosan et. al. The Eurobiotech J. **1**(3) 2017

Plant proline response to abiotic stress



Kaino T., Takagi H.,. Proline as a Stress Protectant in the Yeast *Saccharomyces cerevisiae*, Biosci. Biotechnol. Biochem; 2009: 73(9); 2131-2135

Tsolmonbaatar A., Hashida K., Sugimoto Y., Furukawa S., Takagi H. Isolation of baker'syeast mutants with proline accumulation that showed enhanced tolerance to bakingassociated stresses, Int J Food Microbiol., 2016; 238; 233-240.

Hayat S., Hayat Q., Alyemeni M.N., Wani A.S., Pichtel J., Ahmad A..Role of proline under changing environment, Plant Signal. Behav.; 2012; 7(11); 1456-1466.

James M. Phang, 2017, ANTIOXIDANTS & REDOX SIGNALING
Mouse Cancer Genetics Program, Center for Cancer Research, National Cancer Institute at Frederick, NIH, Frederick, Maryland

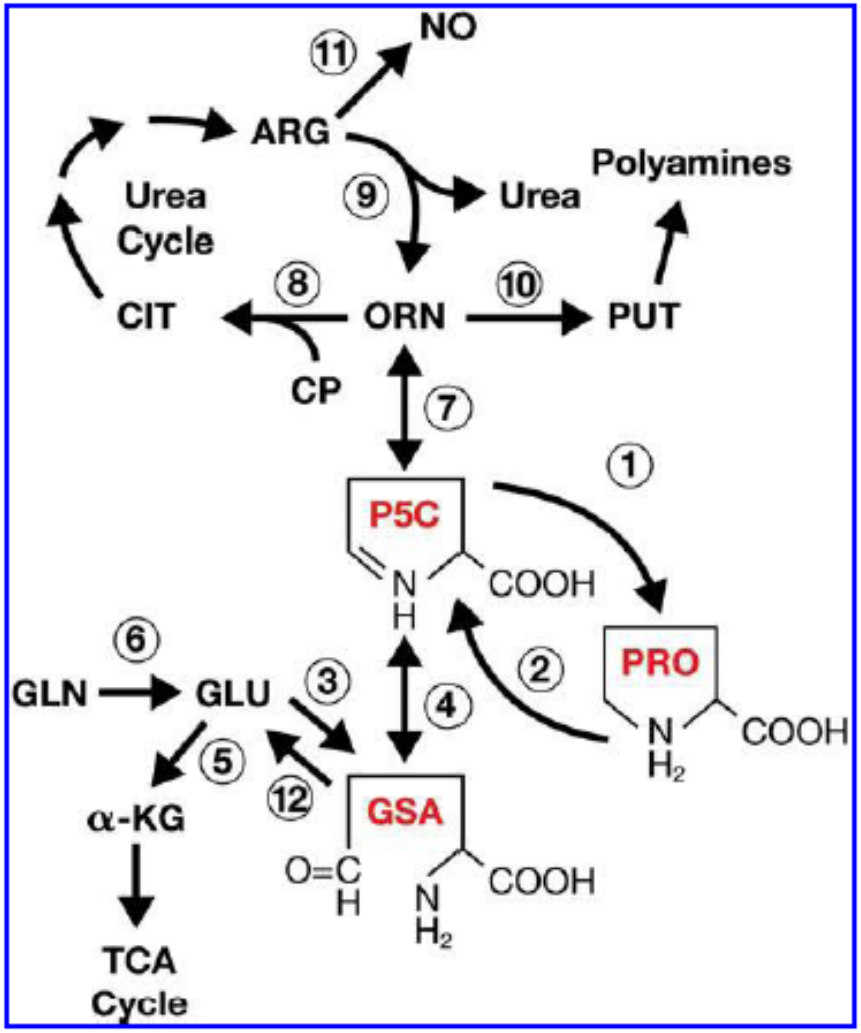
Proline Metabolism in Cell Regulation and Cancer Biology: Recent Advances and Hypotheses

Significance: It is increasingly clear that proline metabolism plays an important role in metabolic reprogramming, not only in cancer but also in related fields such as aging, senescence, and development.

Future Directions: The proline metabolic axis can serve as a scaffold on which a variety of regulatory mechanisms are integrated. Once understood as a central mechanism in cancer metabolism, proline metabolism may be a good target for adjunctive cancer therapy.

James M. Phang, 2017, ANTIOXIDANTS & REDOX SIGNALING

PROLINE METABOLISM AND CANCER



Proline metabolic pathway

DecisionGS (Decision tree genetic selection) by boosted decision tree forest models based on DArT (Diversity Array Technology) for wheat selection improvement

Data source: CIMMYT

599 breeding lines

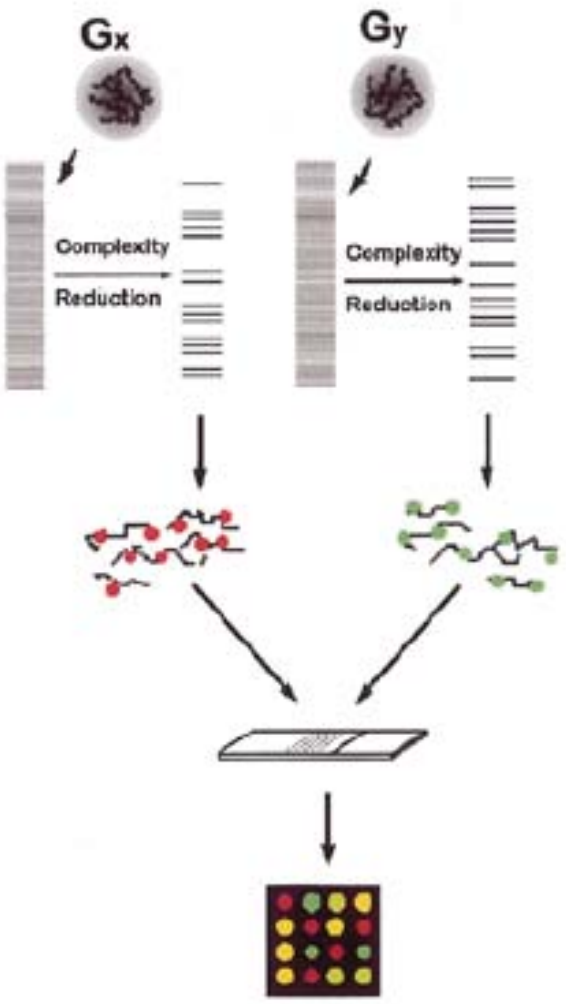
1279 DArT signatures

8 phenotypes

1) thousand-kernel weight (TKW),
2) test weight (TW),
3) grain hardness (GH),
4) grain protein (GP),
5) SDS sedimentation (SDS),
6) grain yield per square meter (GYSM)
7) plant height (PHT)
8) days to maturity (DTM)

DArT genotypization
Diversity Arrays Technology

Developed at Diversity Arrays Technology Pty Ltd, Canberra, Australia



High throughput genetic fingerprinting microarray technology platform for analysis of DNA polymorphism (SNP)

Low cost, minimal sample, complexity reduction

Applications

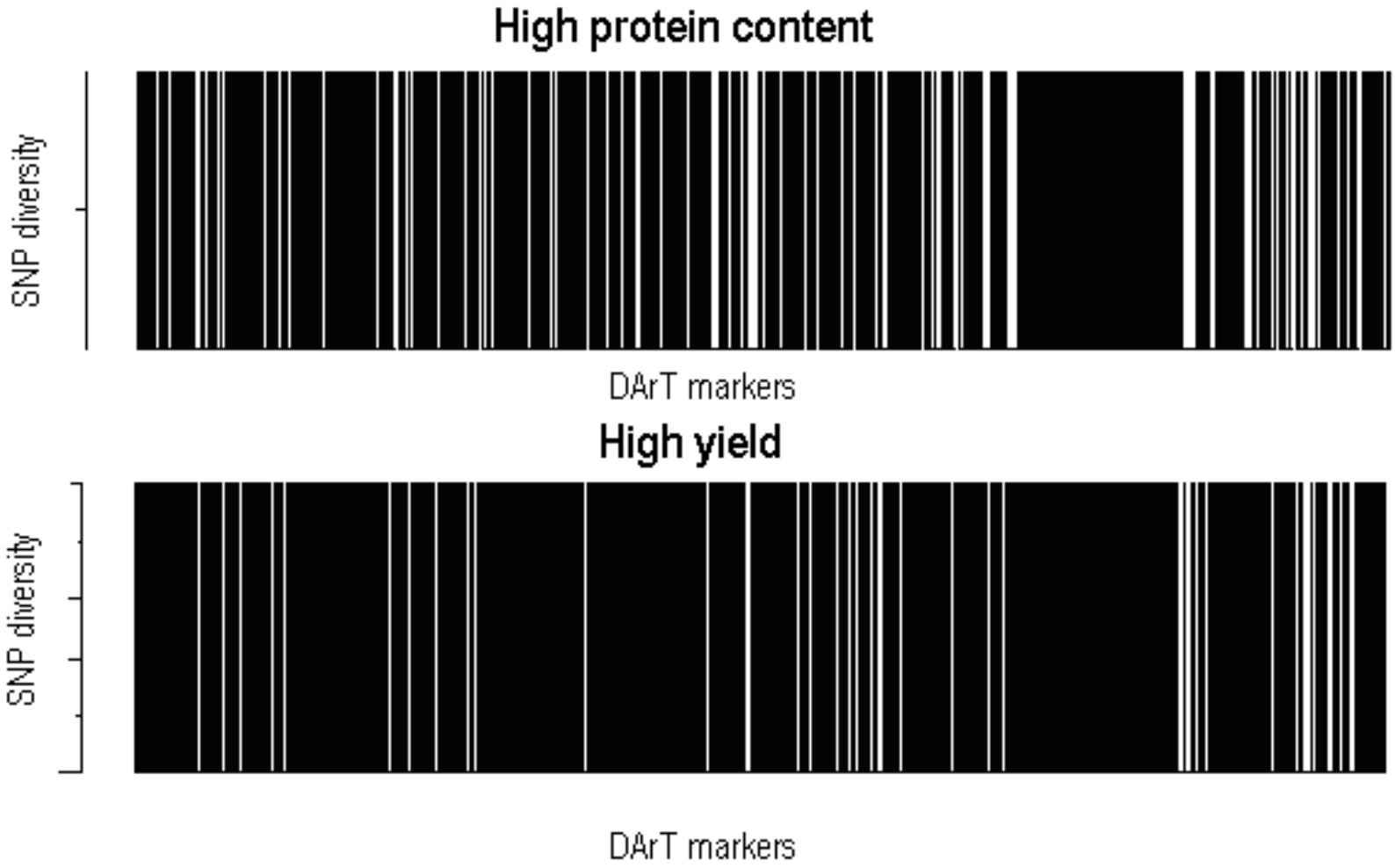
QTL Quantitative Trait Location
Genome + environment synergism

Genome mapping
GWAS big data analysis

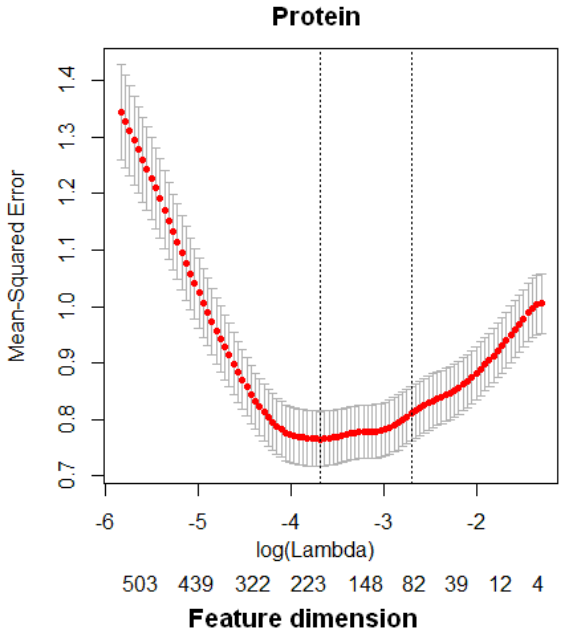
cost 0.2 $ / marker

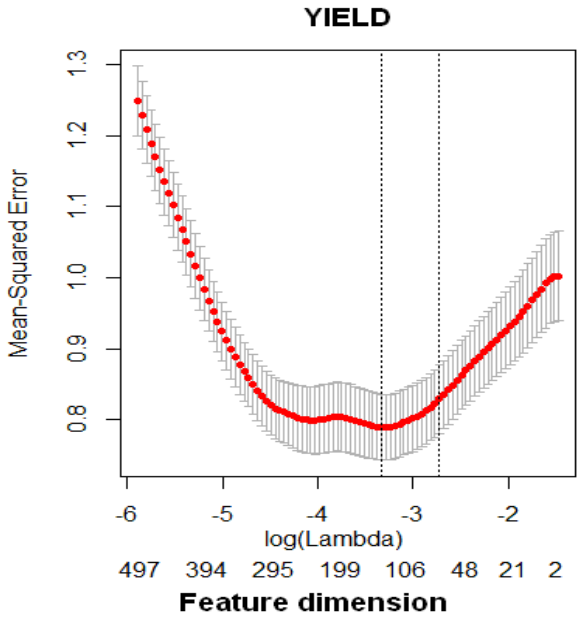
250 $ / genome

A. Kilian, Nucleic Acid Research, 2001, **29** (4) e25

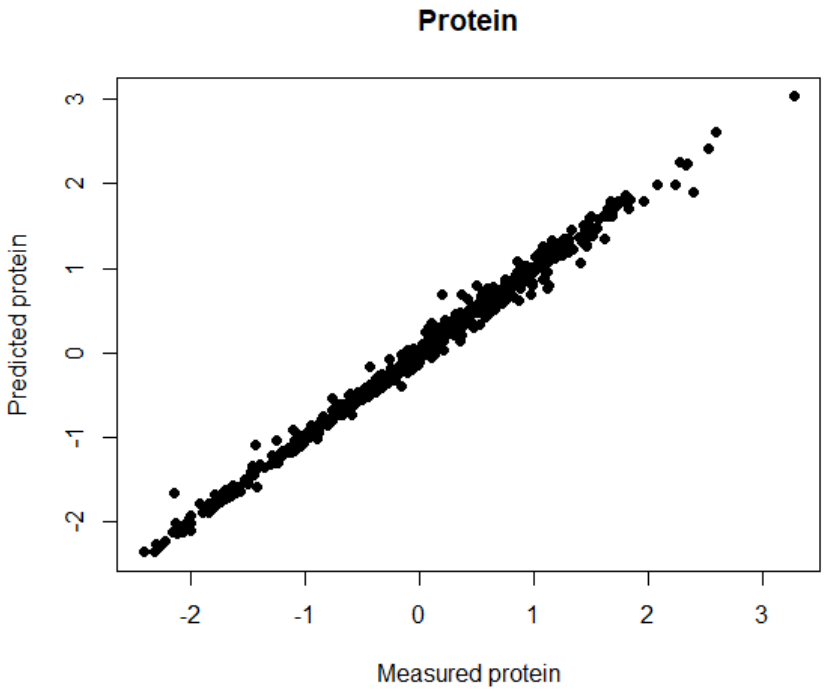

High protein content
SNP diversity
DArT markers
High yield
SNP diversity
DArT markers

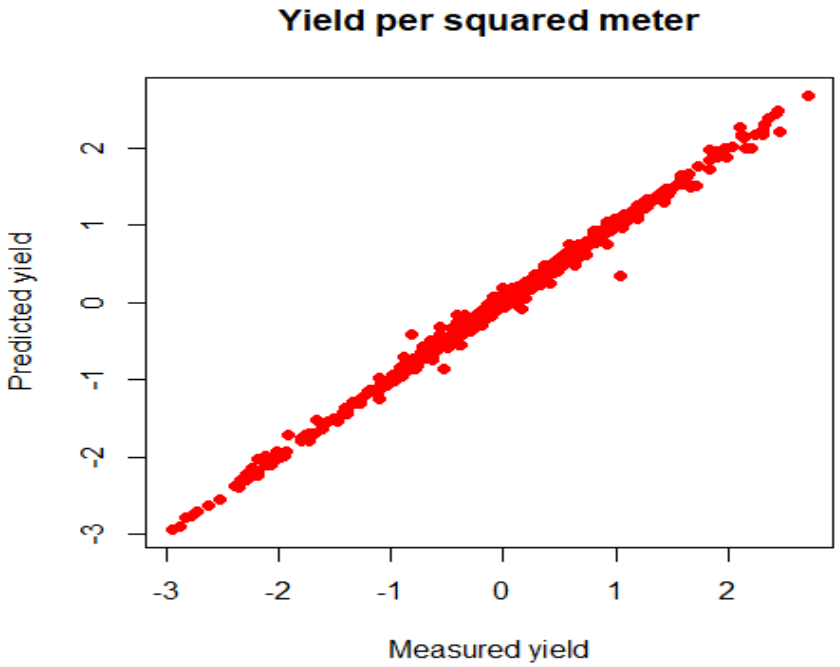
Determination of significant DArT signatures by elastic nets models for protein and yield

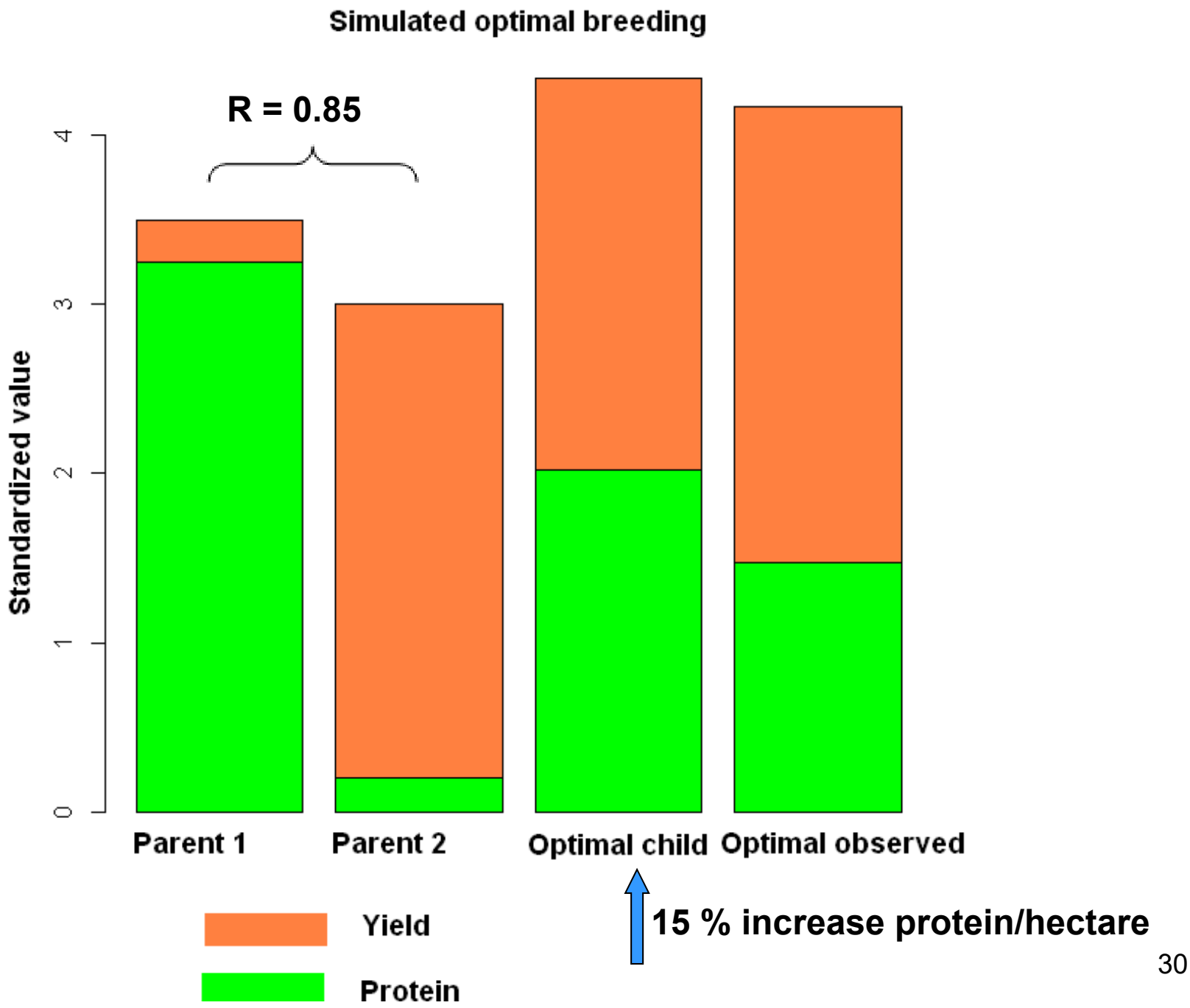




Validation of DecisionGS models for protein and yield






Simulated optimal breeding
R = 0.85
Standardized value
0
1
2
3
4
Parent 1
Parent 2
Optimal child
Optimal observed
Yield
Protein
15 % increase protein/hectare

Prospects of CRISPR and Data science for food quality improvement and production

Gartland K., J. Gartland, Contributions of biotechnology to meeting future food and environmental security needs, The EuroBiotech Journal, 2 (1) 2018, 1-8

Waltz, E., “Gene edited CRISPR mashroom escapes USDA regulation *Nature*, **532** (4) (2016) 293.

USDA: CRISPRE edited mushroom did not trigger USDA oversight because it does not contain foreign DNA from ‘plant pests’ such as viruses or bacteria.

The common white button mushroom (*Agaricus bisporus*) has been modified to resist browning.

The common white button mushroom has been modified to resist browning. Knocked out one of six PPO genes — reducing the enzyme’s activity by 30%.

K. Gartland (private communication): There are in development about 35 CRISPR improved food products (agriculture, animals, fish, fruits, vegetables).

In 2017 U.S. Department of Agriculture gave the green light to a version of the plant *Camelina sativa* (kupusnjača sjetveni podlanak), an important oilseed crop that had been genetically engineered using CRISPR to produce enhanced omega-3 oil.

Jennifer Doudna (Berkely):

Within the next few years, this new biotechnology will give us higher-yielding crops, healthier livestock, and more nutritious foods. Within a few decades, we might well have genetically engineered pigs that can serve as human organ donors…we are on the cusp of a new era in the history of life on earth—an age in which humans exercise an unprecedented level of control over the genetic composition of the species that co-inhabit our planet.

Conclusions

Two mile stones developments in biotechnology

1) **CRISPR Cas9 DNA and RNA editing**
2) **AI data science application in life science**

Challenges for CRISPR and DArT in biotechnology:

1) **Food production in global climate change**
2) **DArT and GWAS can be effectively applied in metagenomics for bioprospecting and environment protection**

Propositions:

1) **Basic information and computer skills in AI and Data science should be a part of MS and PhD levels of education in biotechnology**
2) **Integration of data and open access is needed for science projects**