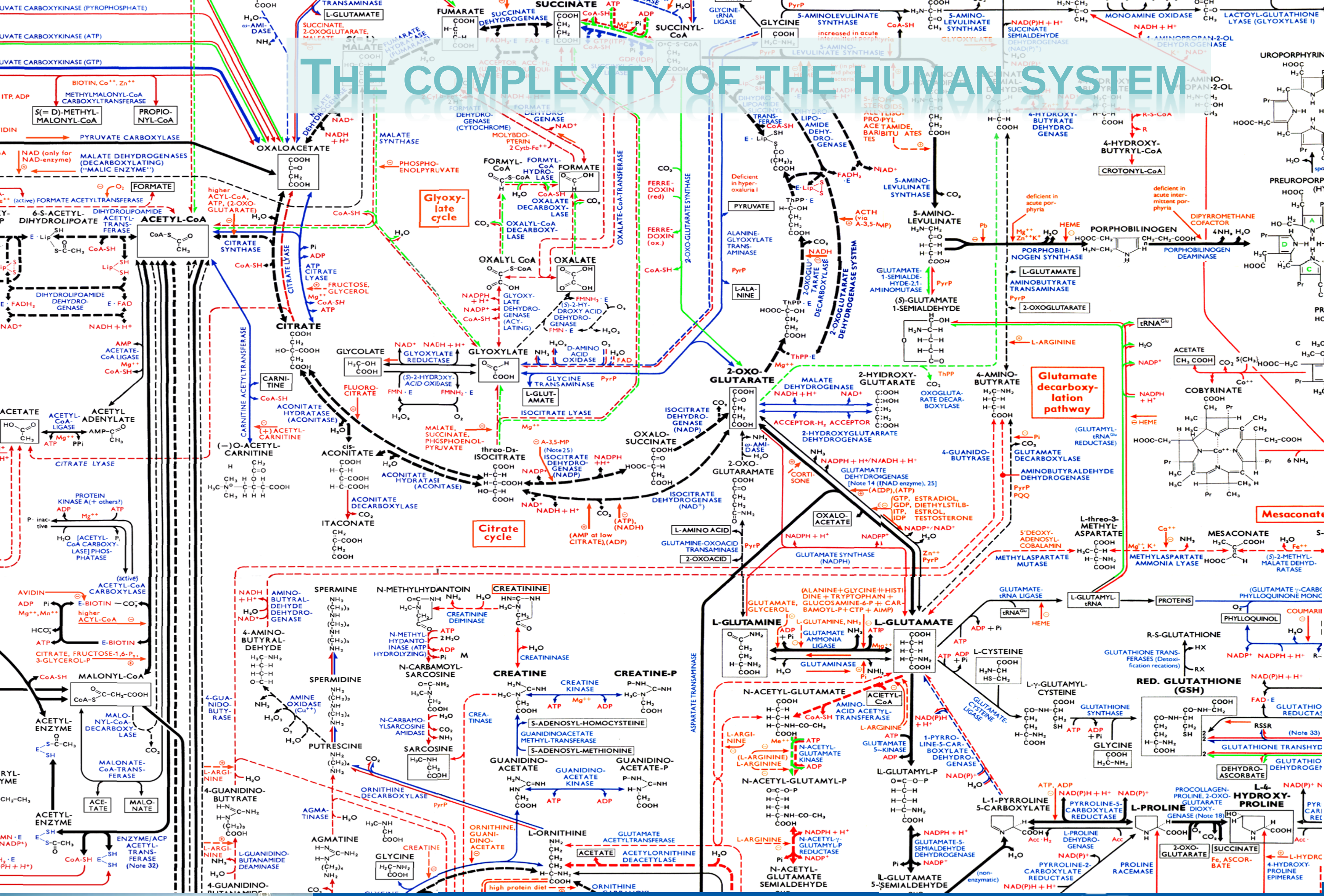
Metabolomics as part of an integrated approach for the identification of predictive markers of Type 2 diabetes

E Pujos-Guillot, M Brandolini, D Grissa, Y Liu, M Pétéra,

*C Joly, B Lyan, S Czernichow, M Zins, M Goldberg, **<u>B Comte</u>***

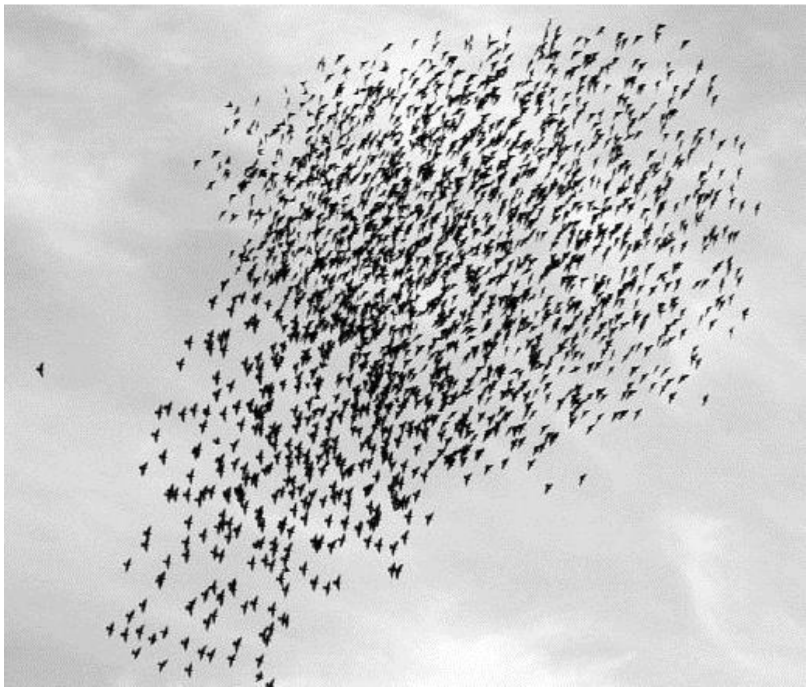
INRA
SCIENCE & IMPACT

THE COMPLEXITY OF THE HUMAN SYSTEM



INRA
SCIENCE & IMPACT

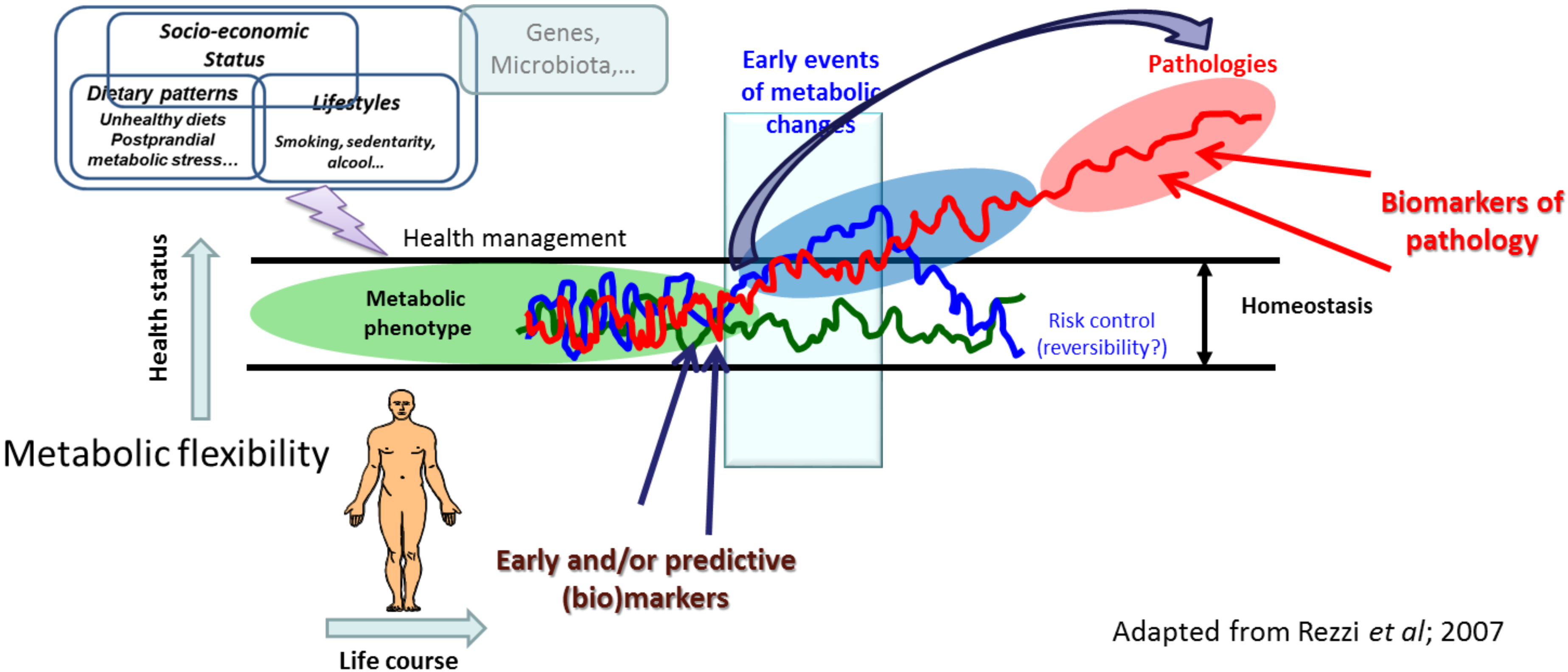
THE COMPLEXITY OF THE HUMAN SYSTEM



One system with
efficient communications

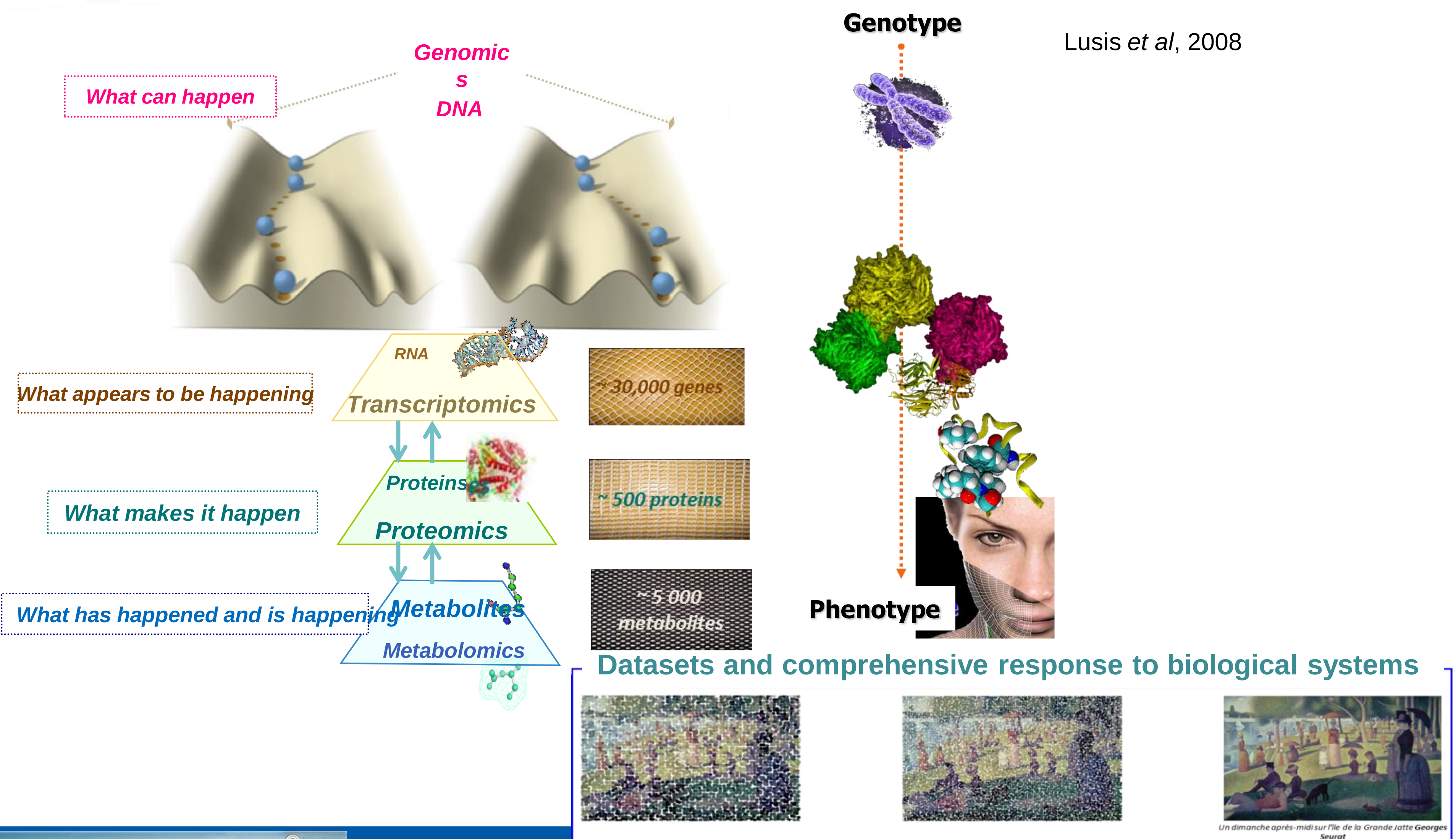
One balanced system
(homeostasis)

THE DYNAMICS OF METABOLIC PHENOTYPES



Adapted from Rezzi *et al*; 2007

GLOBAL INTEGRATIVE APPROACHES



BIOMARKERS

Robust, sensitive and predictive biomarkers of metabolic health status are needed in the fields linked to diseases and nutrition

Explore and Discover
- identify potential differentiating metabolites, genes and proteins

→

Qualification
- candidate metabolites, genes and proteins
- validation analysis
- proof of biomarkers in new studies
- study precision and robustness
- biomarker(s) predictiveness
- test biomarkers in diverse populations

→

Application
- adoption into clinical practice
- continued assessment of economic impact

biomarker discovery — candidate biomarkers — validated biomarkers

Serkova, 2011

1) Can the clinician measure the biomarker?

a) **Accurate and reproducible analytical method(s)**
b) **Pre-analytical issues (including stability) evaluated and manageable**
c) **Assay is accessible**
d) Available assays provide high through-put and rapid turn-around
e) Reasonable cost

2) Does the biomarker add new information?

a) **Strong and consistent association between the biomarker and the outcome or disease of interest in multiple studies**
b) **Information adds to or improves upon existing tests**
c) Decision-limits are validated in more than one study
d) Evaluation includes data from community-based populations

3) Will the biomarker help the clinician to manage patients ?

a) **Superior performance to existing diagnostic tests, or**
b) **Evidence that associated risk is modifiable with specific therapy, or**
c) **Evidence that biomarker-guided triage or monitoring enhances care**
d) Consider each of multiple potential uses (SEE PANEL B)

Morrow *et al*., 2014

INRA
SCIENCE & IMPACT

PREDICTIVE MARKERS OF DIABETES

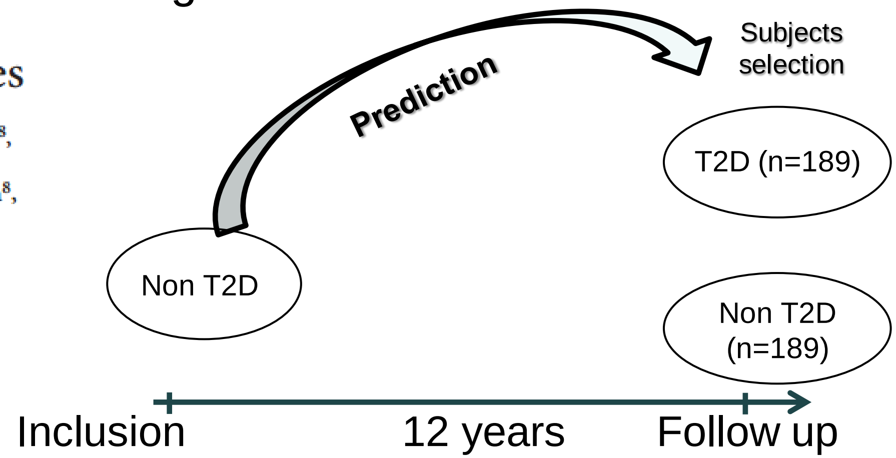
nature medicine

Metabolite profiles and the risk of developing diabetes

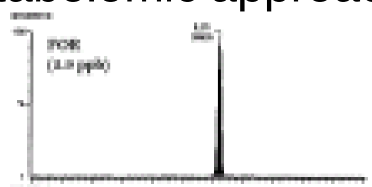
Thomas J Wang[1–3], Martin G Larson[3,4], Ramachandran S Vasan[3,5], Susan Cheng[2,3,6], Eugene P Rhee[1,7,8], Elizabeth McCabe[2,3], Gregory D Lewis[1,2,8], Caroline S Fox[3,9,10], Paul F Jacques[11], Céline Fernandez[12], Christopher J O'Donnell[2,3,8], Stephen A Carr[8], Vamsi K Mootha[8,13,14], Jose C Florez[8,13], Amanda Souza[8], Olle Melander[15], Clary B Clish[8] & Robert E Gerszten[1,2,8]

VOLUME 17 | NUMBER 4 | APRIL 2011 NATURE MEDICINE

Framingham



Targeted metabolomic approach



Supplementary Table 3: Prediction of incident diabetes, with various combinations of amino acids, in high-risk and low-risk individuals

Model Description	"High-risk" sample* -2 log-likelihood ratio	"High-risk" sample* c-statistic
Clinical model: age, BMI, glucose	2.01	0.52
Clincal plus biomarker model, with 3 amino acids (isoleucine, phenylalanine, tyrosine)	27.44	0.65
Clinica plus biomarker model, with 5 amino acids (isoleucine, valine, leucine, phenylalanine, tyrosine)	27.79	0.66

First model (age, sex, body mass index, fasting glucose) denotes the basic clinical model. Higher numbers for log-likelihood ratio indicate better model fit. *High-risk sample includes cases and matched controls from Framingham. Random cohort includes cases and random controls from Framingham. Clinical characteristics of the study samples are shown in Table 1.

Type 2 Diabetes (T2D) and GAZEL

The GAZEL cohort

A EDF-GDF COHORT - over 20,000 participants recruited in 1989;

~ 15,000 men (40-50 y.o.) and ~ 5,600 women (35-50 y.o.)

Different themes of research (i.e. social determinants of health, work conditions and professional exposure) but one main theme: Aging, in the perspective of life course epidemiology

60 % retired

60-75 y.o., over 98 % retired

1989 1990 1998 2002 2004 2009 2013

* *

- Annual questionnaires: occupational and socioeconomic parameters, health status evaluation, and food frequency questionnaires (FFQ, qualitative data)

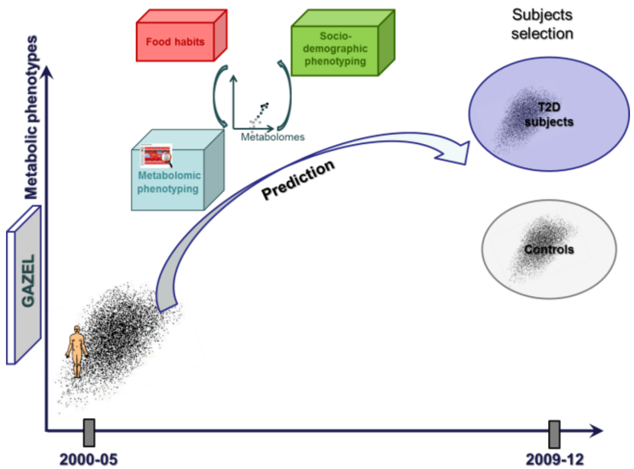
FFQ FFQ FFQ FFQ

Plasma sampling

An integrative ***multidisciplinary approach,*** putting together sociology, epidemiology, nutrition, metabolomics, statistics, and computer science, to ***develop accurate and robust indicators and biomarkers of health status, predictive of T2D development***

SELECTION OF A SUB-COHORT OF T2D SUBJECTS

DIAPASON :
A CASE-CONTROL STUDY



- ➢ **Inclusion**: 7,537 subjects with biological sampling
- Men, overweight, alive in 2012
- Without T2D

- ➢ **End of the study** (2008-2012):
- **Case** = subjects with **declared** diabetes
- **Controls** paired on:
  - Age (years): [52 ;55[, [55 ;58[and [58 ;64[
  - BMI (kg/m²): [25 ;28[and [28 ;30[
  - Location within 2 classes:
    - * 1-2-3= Nord, Paris/RP, Nord-Est
    - * 7-8-9= Centre, Nord-West, Bretagne

➢ **Case** (n=56):
Men alive in 2012
Mean age: 65.9 $\pm$ 2.9 yr
Mean BMI: 27.1 $\pm$ 1.2 kg/m^2

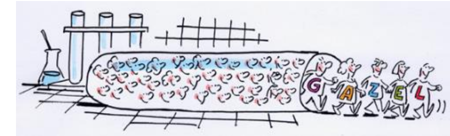
➢ **Controls** (n=56):
Men alive in 2012
Mean age: 65.5 + 2.6 yr
Mean BMI: 26.7 $\pm$ 1.3 kg/m^2

Collected variables: food frequency questionnaires (FFQ), physical activity, alcohol consumption, smoking; possibility to reply to a new questionnaire (food and health)

ANALYSES: Untargeted serum metabolomics

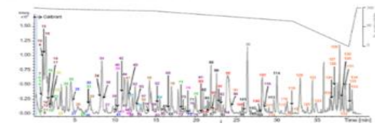
Prediction of T2D development 5 years later

Approaches & Workflow

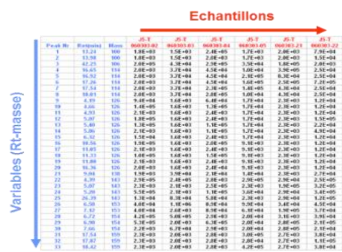


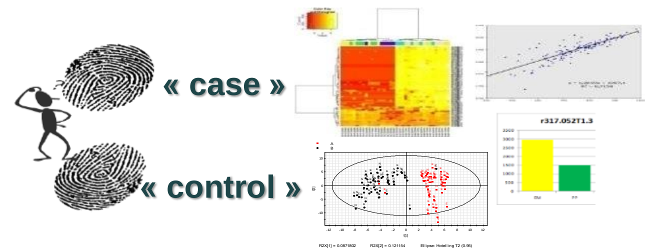
- **Food patterns and trajectories**

- **Untargeted metabolomic approach using mass spectrometry**

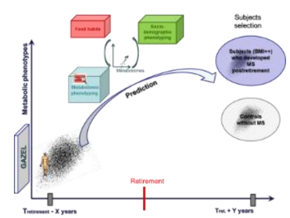


- **Model building and marker discovery**



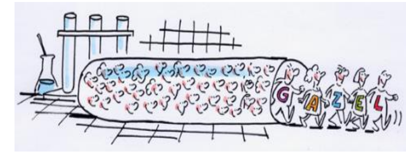


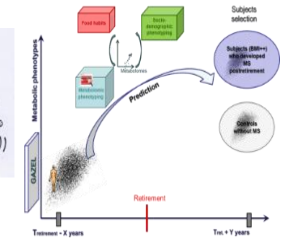
- **Prediction evaluation - Validation**



INRA
SCIENCE & IMPACT

Prediction and Evaluation





➢ Model building using logistic regression

Looking for the respective roles of the variables (X_1, X_2, X_3…)

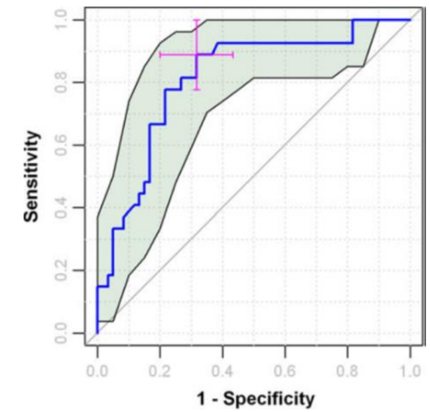
$$Y = f(X_1, X_2, X_3\ldots)$$

Reduction of the number of variables for considering only the most relevant and predictive ones for Y (T2D)

➢ Measurement of prediction accuracy

→ evaluating if the model meets prediction requirements

Sensitivity, specificity,
ROC curve, concordance index



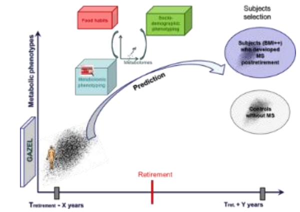
➢ Comparison of the models and to the existing alternatives

→ Does the model bring significant improvement?

Model and accuracy comparisons

	AdaBoostM1 BayesNet	BayesNet	NBTree	Rand.Com. RandomTree	Decorate Trees J48	Simple Logistic	trees LMT
Communs 4	86,21	82,76	82,76	96,55	86,21	82,76	82,76
T-test 9	82,76	82,76	82,76	79,31	86,21	75,86	79,31
CFS 11	96,55	96,55	96,55	93,1	96,55	86,21	86,21
SAM 11	86,21	86,21	86,21	82,76	89,66	96,55	96,55
Entropie 12	96,55	96,55	96,55	96,55	86,21	86,21	86,21
SAM 14	89,66	89,66	89,66	82,76	79,31	86,21	86,21
T-test 36	79,31	79,31	75,86	86,21	82,76	68,97	68,97
CFS 143	86,21	86,21	86,21	68,97	75,86	–	62,07

Subject characteristics



Population with biological sampling
N=7,537

(age: mean=58.8)

Case (N=97) | Potential controls (N=713)

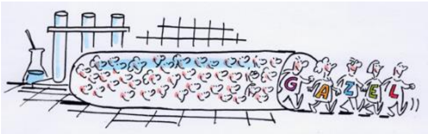
N=56 with biological samples and questionnaires (FFQ) | N=391 with biological samples and questionnaires (FFQ)

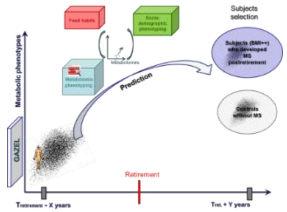
112 retired subjects at sampling (T0, 2004)

	Controls	Case	p
Age, years	56.6±2.6 (56)	58.9±3.0 (56)	0.54
Weight, kg	80.4±7.2 (56)	79.9±6.2 (56)	0.90
BMI, height/m^2	26.7±1.3 (56)	27.1±1.2 (56)	**0.046**
Waist/Hip ratio	93.3±4.5 (56)	95.7±3.7 (56)	**0.005**
Systolic blood pressure, mmHg	129.1±12.5 (56)	136.5±13.7 (56)	**0.004**
Diastolic blood pressure, mmHg	77.9±8.4 (56)	80.2±8.4 (56)	**0.005**
Total cholesterol, mM	5.7±0.7 (56)	5.6±0.9 (56)	0.41
HDL-cholesterol, mM	1.5±0.4 (54)	1.5±0.3 (56)	0.51
Triglycerides, mM	1.0±0.4 (56)	1.2±0.5 (56)	0.06
Fasting glucose, mM	5.5±0.5 (56)	6.6±1.3 (56)	**4.9E-9**

Case at T1 (2009):
100% T2D,
70% hypertension,
76% dyslipidemia

Food Patterns





❖ **Selection of 9,042 men in GAZEL (15,011)**

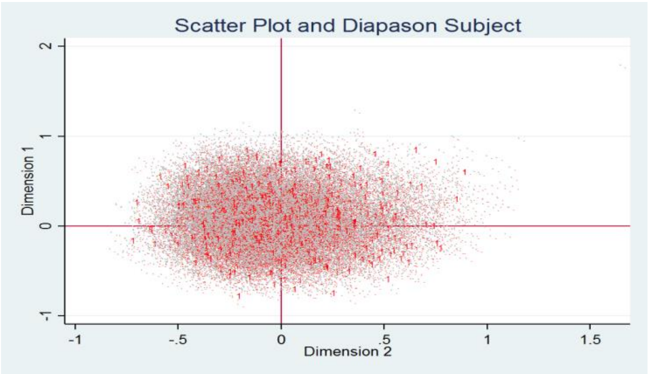
Different criteria (alive in 2012, answered to at least 2 FFQ between 1998, 2004 and 2009…)

Determination of relevant and not redundant food items in the annual questionnaires: 22 food items

Red Meat, Poultry, Fish, Cooked Meat, Eggs, Fried Food, Fat type, Light Products, Milk, Dairy, Cheese, Bread, Vegetables, Raw vegetables, Starchy Food, Fruits, Desserts, Pastries, Sugar, Coffee, Sweet Beverages, Wine

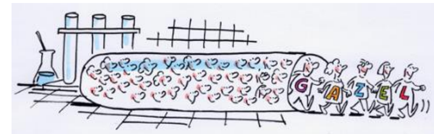
❖ **Construction of Foods Patterns with Multiple Correspondence Analysis (MCA) on the GAZEL cohort**

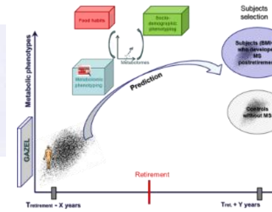
- Development of MCA on the 1998 variables and construction of Food Patterns
- Identification of food patterns



Scatter plot of the eligible subjects of GAZEL and the sub-cohort study

Food Patterns





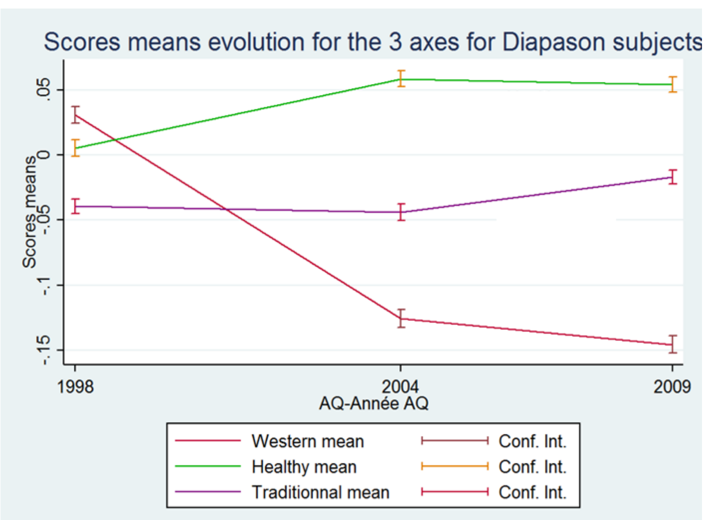
3 food patterns: ‘Healthy’, ‘Western’ and ‘Traditional French’

- Healthy: healthy foods (vegetables, fruits…)
- Western: fat and sweet products
- Traditional French: cooked and traditional French diet (rich in wine, cheese, vegetables, low in dairy, fish and fruits)

❖ Projection of the variables in 2004 and 2009 on the MCA in 1998

- Analysis of the evolution of the mean score between 1998 and 2009

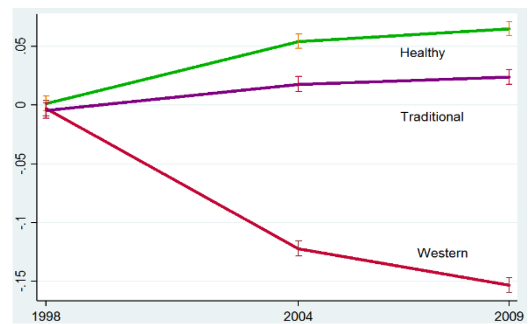
Food trajectories of the sub-cohort followed the same trends than those of the whole cohort



Healthy

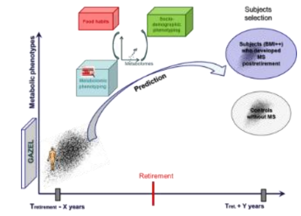
Traditional

Western



Improvement of food patterns with time. Subjects of the sub-cohort are following the same trajectories than those of Gazel

Prediction of evolution towards T2D



Using food items, socioeconomics and clinical data in 2004

- ❖ More than 80 variables (categorical & quantitative) were considered
- ❖ Logistic regressions were performed to pre-select relevant variables ($p < 0.25$)

⇨ 32 variables were obtained with BMI as cofactor:

10 food items (ex: raw vegetables, dessert, bred…),
10 socio-economic variables (ex: diploma, income…),
12 clinical parameters (ex: hypertension, weight, TG,…)

- ❖ Stepwise selection (AIC criteria) to identify predictive variables:

⇨

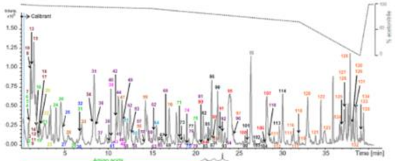
	Pr(>Chi)
Vegetables	0.0068**
Raw vegetables	0.029*
Sugar	2.80E-5***

	Pr(>Chi)
Monthly income in 2002	0.091

	Pr(>Chi)
Glycemia	2.92E-9***
Waist/Hip (W/P) ratio	0.019*

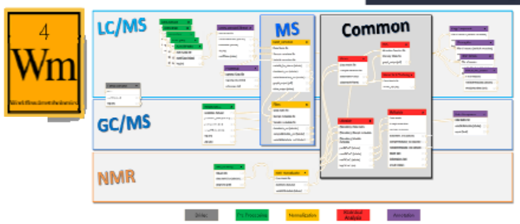
UNTARGETED METABOLOMICS APPROACH

➢ MASS SPECTROMETRY METABOLOMIC ANALYSES ON SERUM SAMPLES



Monitoring thousands of metabolites

Workflow4Metabolomics.org



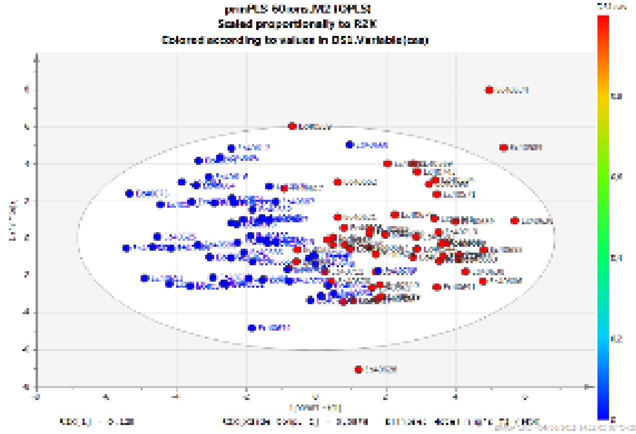
➢**Final data set:**

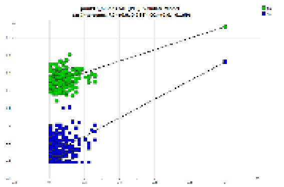
Number of ions (variables): 1,195 (ESI pos), 208 (ESI neg)

Number of samples (subjects): 111 (1 outlier, 1 Case)

ANOVA analyses after BH correction	Factor T2D Cofactor BMI		Factor BMI
Nb significant ions ‘Case’ ESI pos	**52**	13	**41**
Nb significant ions ‘Case’ ESI neg	**10**	4	**12**

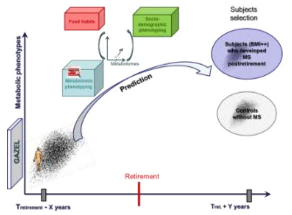
PLS after UV-scaling on significant ions (without BH correction)





Q^2=0.39

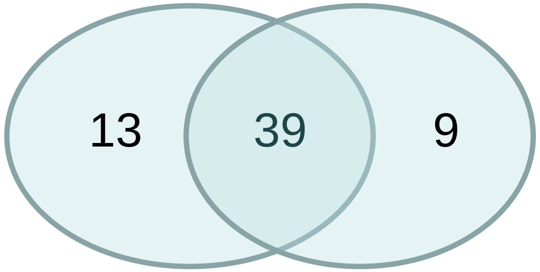
Feature selection methods



Esi+: 48 ions

Esi-: 10 ions

Features



1,195 ions

Anova BH

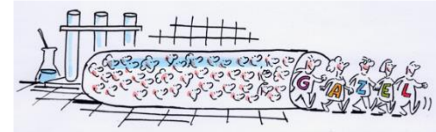
Random forest

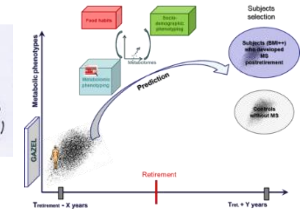
- **Logistic regressions on 58 ions:**

5 metabolites identified as predictive

	Pr(>Chi)
p148	0.016*
p167	0.058
p198	0.015*
p268	0.0002***
p288	0.037*

Model comparisons





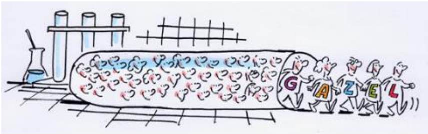
	AUC	95% CI	Misclassification (%)	False positive	False negative
W/H ratio + glycemia	0.74	0.66-0.82	26%	16	13
5 metabolites	0.82	0.75-0.89	18%	10	10
W/H ratio + glycemia + 4 metabolites	0.89	0.83-0.95	11%	5	7

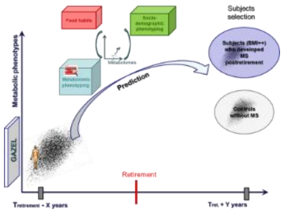
- Metabolomics data showed a better prediction capacity: more integrative

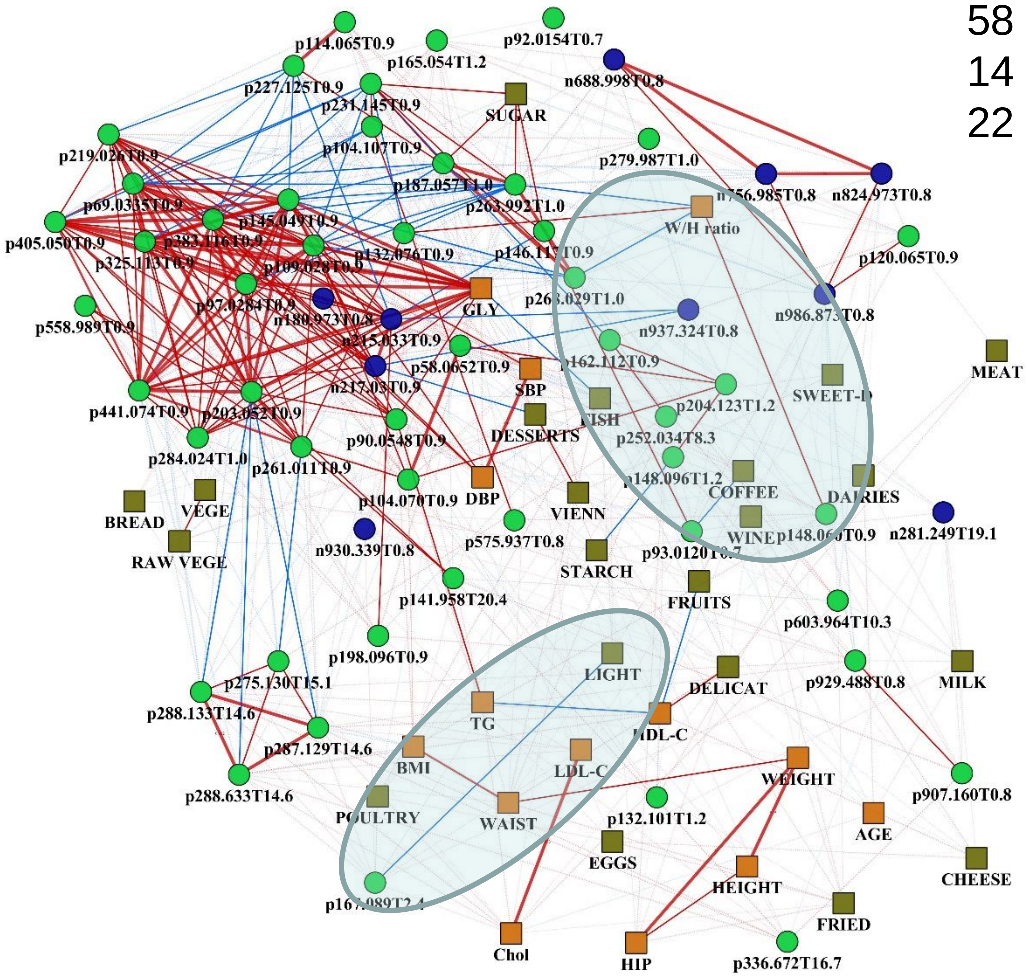
- The best predictive model is including metabolomics and clinical data
  The 4 metabolomic markers in the integrative model are different from the 5 previous ones

- In a model including food items and metabolomics: correlations between them filtered one of them → Preference on metabolomic markers

CORRELATION NETWORK





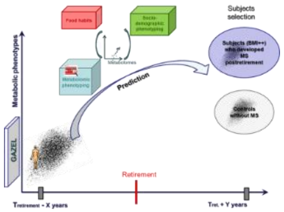


58 ions,
14 clinical parameters,
22 food items

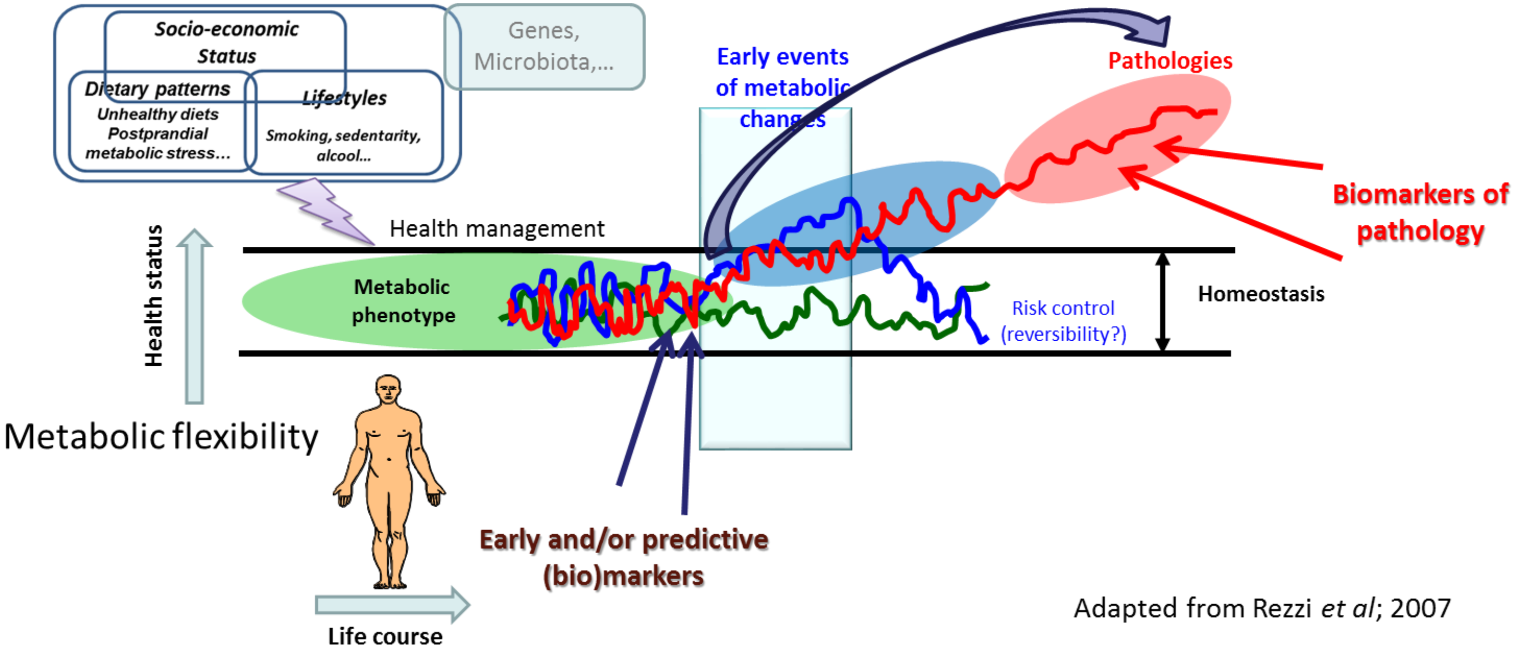
Exclusion of metabolites by correlations:
p268 & p148 correlated with W/H ratio

Food items: p167 correlated with light product consumption

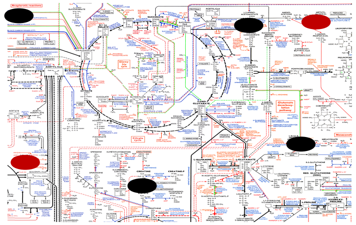
Conclusion



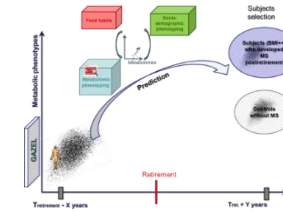
Few predictive biomarkers



Adapted from Rezzi *et al*; 2007



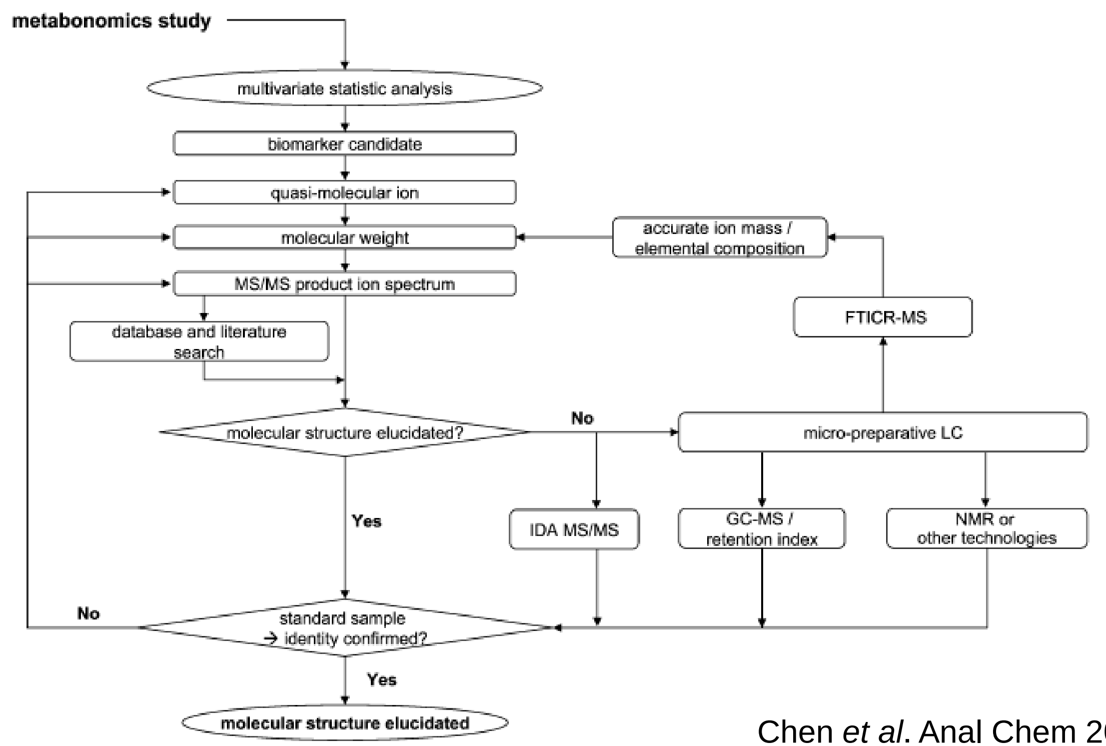
TOWARDS DATA INTERPRETATION…



- Predictive metabolites currently being identified
- 48 signals to be annotated to map data on metabolic pathways and networks, and generate hypotheses on pathophysiological mechanisms

METABOHUB

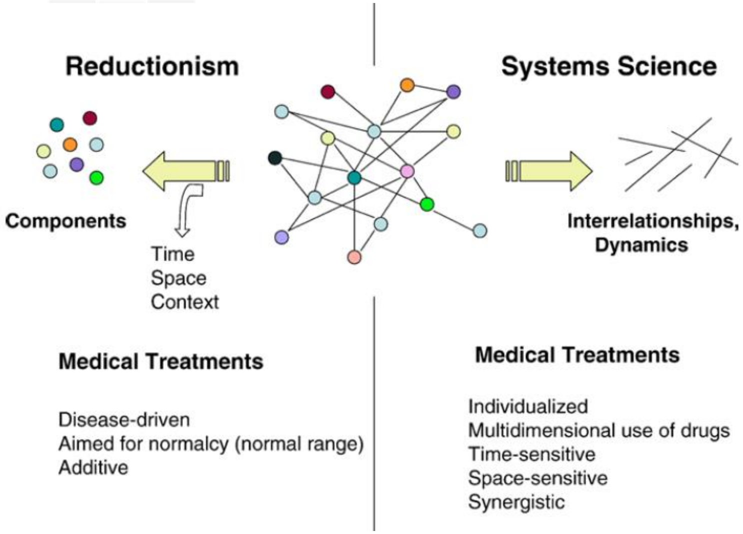
1,085 reference spectra in DB
(2500 compounds)



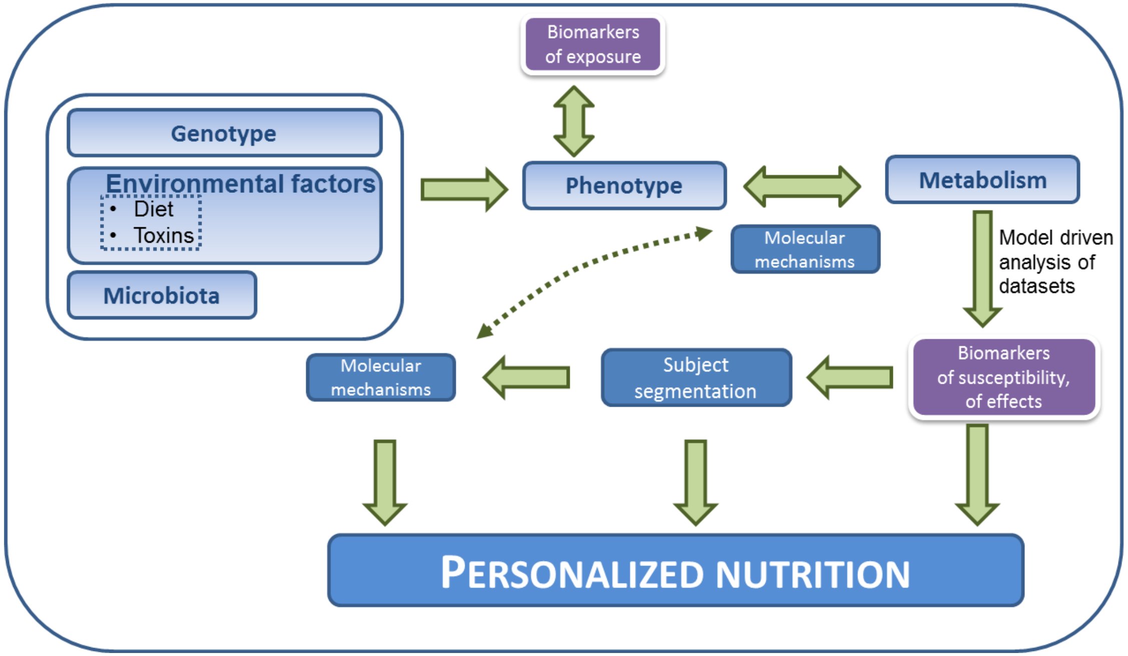
Chen *et al*. Anal Chem 2008

INRA
SCIENCE & IMPACT

… AND SYSTEMS BIOLOGY



Ahn *et al*., 2006



Adapté de Mardinoglu & Nielsen, 2011

ACKNOWLEDGMENTS

Plate-Forme EXPLORATION du METABOLISME des gènes aux métabolites

JUR1019 - Unit of Human Nutrition

Mathieu RAMBEAU
Dhouha GRISSA
Julien BERTRAND
Yunfei LIU

Estelle PUJOS-GUILLOT

Mélanie PÉTÉRA
Bernard LYAN
Charlotte JOLY
Delphine
Frank GIACOMONI
Anthony FERNANDEZ
Jean-Bruce MBA

Epidemiology

Sébastien Czernichow, Marie Zins, Marcel Goldberg

Marie Plessz (ALISS)
Charlotte Dion (ALISS)

Sociology

Séverine Gojard (ALISS)

- ***Metatoul***: Jean-François Martin

Statistics

Mohamed Hanafi

Nantes

Artificial intelligence

Amedeo Napoli, Jérémie Bourseau

Financial support:

Metaprogramme DID'IT AlimH INRA

Thank you for your attention