Predicting the outcome of chemotherapy through pathway modelling

7th International Conference on Systems Biology
Yokohama - October 11, 2006

http://fre2571.vjf.cnrs.fr/

Charles AUFFRAY
Array s/IMAGE, Genexpress
CNRS and Pierre et Marie Curie University
Villejuif - France

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Deciphering Cellular States of Innate Tumor Drug Responses

Esther Graudens, Virginie Boulanger, Cindy Mollard, Régine Mariage-Samson, Guilaine Grémy, Christine Couillault, Patrick Zaborski, Eric Eveno, Charles Auffray and Sandrine Imbeaud

(2006) Genome Biology 7, R19

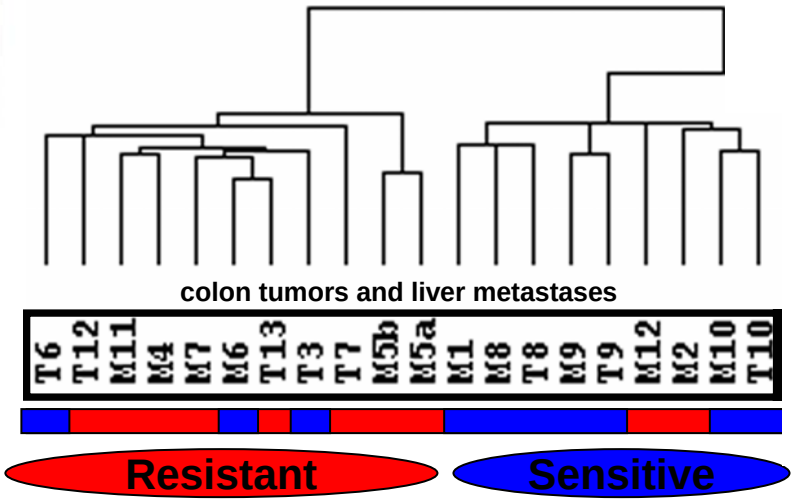
http://genomebiology.com/2006/7/3/R19

Innate Tumor Drug Responses - Experimental Design

Advanced metastatic colorectal (CRC) cancer and drug responses

Conditions: 10 chemo-sensitive and 10 resistant states

Colon tumors	vs.	Colon tumors
Liver metastases		Liver metastases
Adjacent normal colons		Adjacent normal colons



- ✓ **Objectives**: **Innate drug responses** i.e. primary responses (at the presentation of the drugs)
- ✓ **Subsequent Irinotecan (CPT-11)** plus **5-FU/AF** combined chemotherapy
- ✓ **Gene expression profiling** on 11K-cDNA arrays
- ✓ **70 arrays, 3.2×10^6 data points**
- ✓ **Power simulation: >70%, α=0.003**

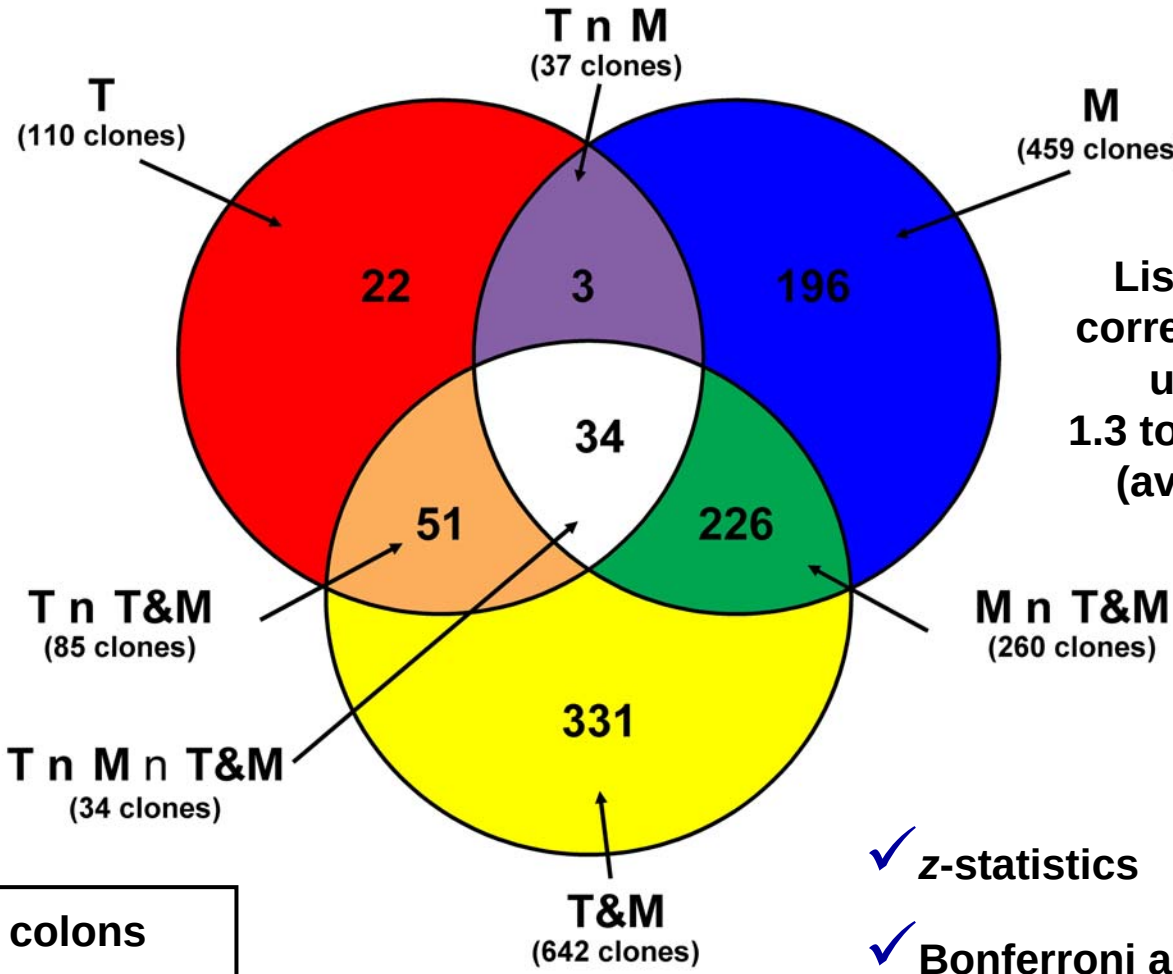
Innate Tumor Drug Responses - Statistical Power

The statistical power is the probability of obtaining statistical significance when true biological differences exist. It is used to verify which subgroups of samples are likely to provide the most comprehensive relevant information and that enough samples are compared to meet the objectives of the study.

Samples	Φ	α	n1	n2	σ expected	Power	σ observed	Power
T	1	0.003	3	5	0.40	0.09	0.27	0.40
M	1	0.003	7	5	0.40	0.27	0.27	0.67
T&M	1	0.003	10	10	0.40	0.70	0.27	0.98

Statistical power (1-ß) for detecting a true 2-fold mean difference between two groups (Φ=1, with base 2 logarithm) at a significance level (α) of 0.003, that accounts for less than 30 false positives in the 10K microarray and a population variability σ=0.4 as previously reported or σ=0.27 as measured in the recorded dataset, n1 and n2 being respectively the resistant and sensitive group sizes.

Innate Tumor Drug Responses
Statistics on Hybridization Differences

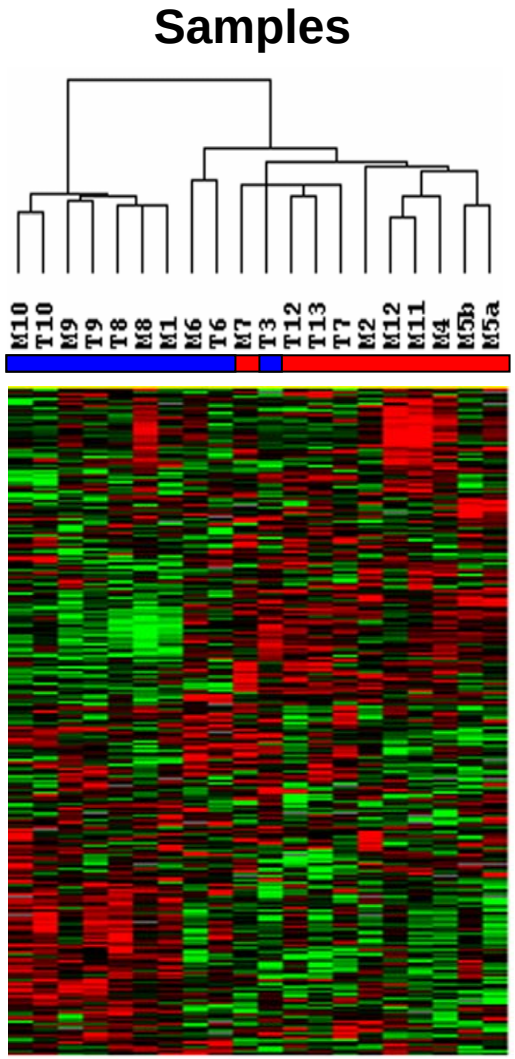


List of 863 clones corresponding to 679 unique genes.
1.3 to 41-fold changes (average 1.7-fold)

T: tumor colons
M: liver metastases

- ✓ *z*-statistics
- ✓ Bonferroni and False Discovery Rate (FDR) *p*-value adjustment
- ✓ α =0.01

Innate Tumor Drug Responses - Process Enrichment



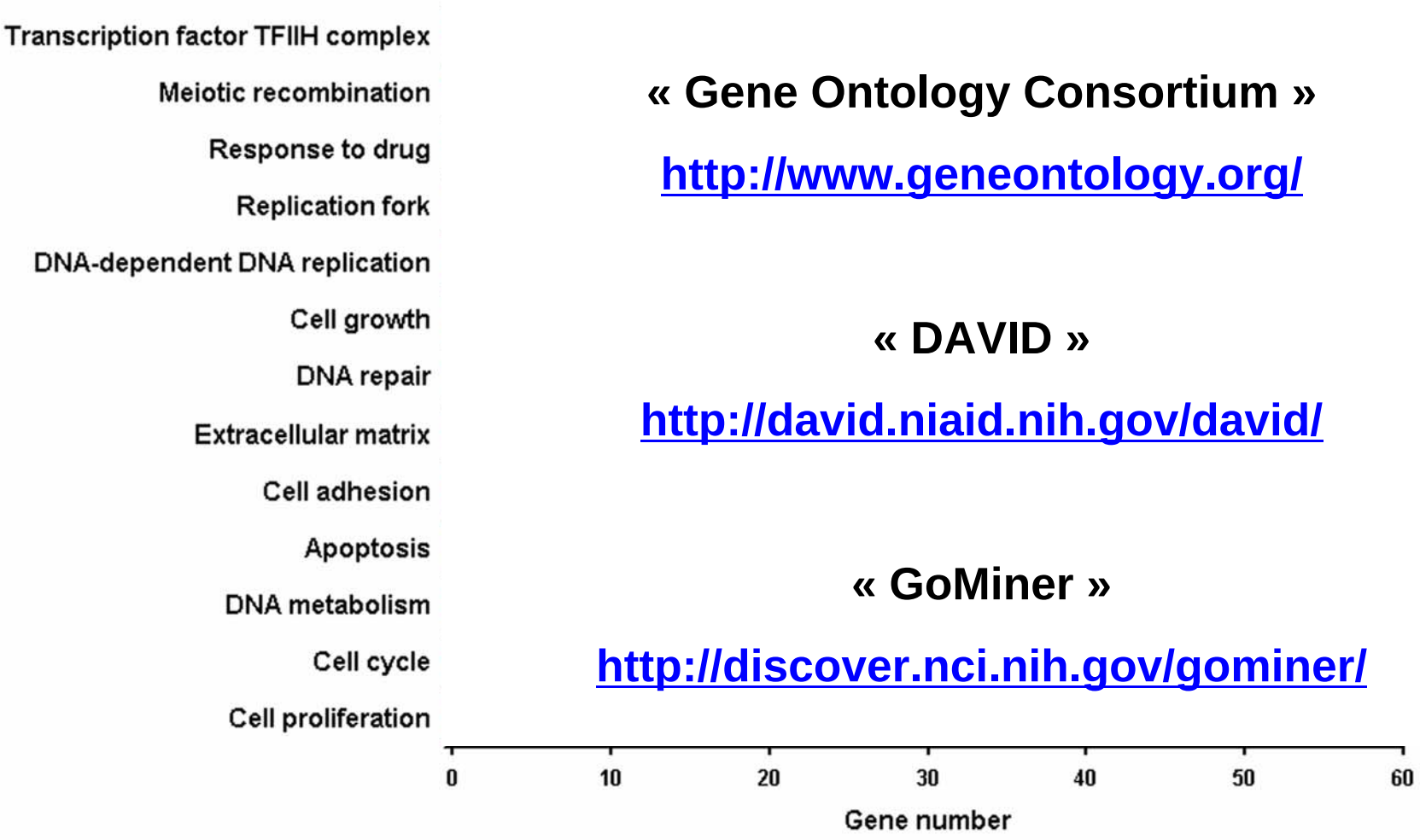
Gene clustering analysis on tumor samples from L863. Top-ranked relevant gene clusters detected using t-statistics with permutation-based adjustment (n=10,000) and α=0.05

Cell cycle
Cell proliferation
Response to DNA damage stimulus
DNA repair

Extracellular matrix (ECM)
Cell adhesion
Cell growth

Innate Tumor Drug Responses - Ontology Enrichment

- Gene Ontology (GO) annotation and controlled-vocabularies
  => 554 terms in total
- DAVID, EASE score (Fisher's adjusted) => enrichment of terms
  => 147 terms, p value ≤ 0.05
- GoMiner, Jacknife's Fisher exact probabilities => up or down modulation



« Gene Ontology Consortium »

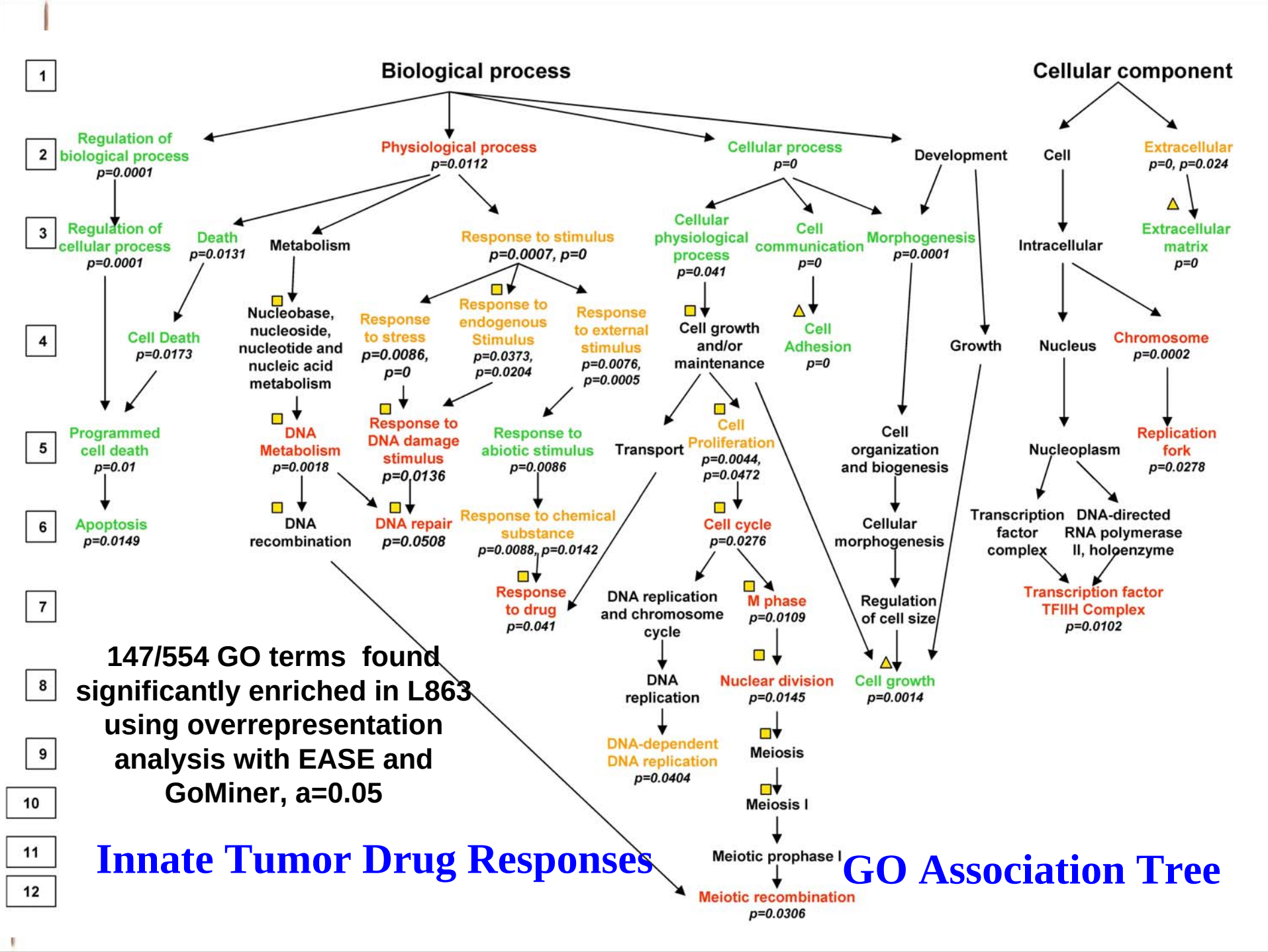
http://www.geneontology.org/

« DAVID »

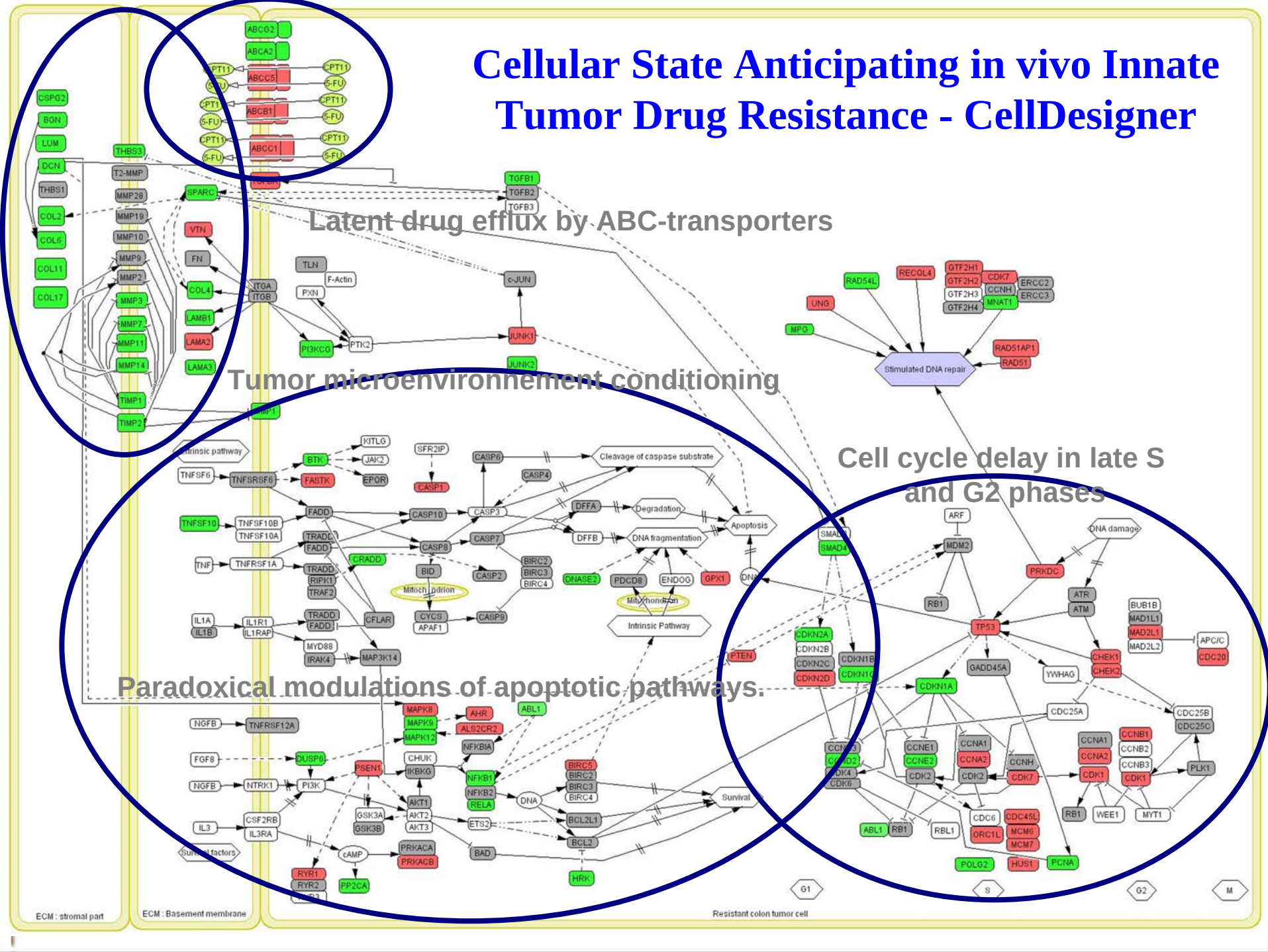
http://david.niaid.nih.gov/david/

« GoMiner »

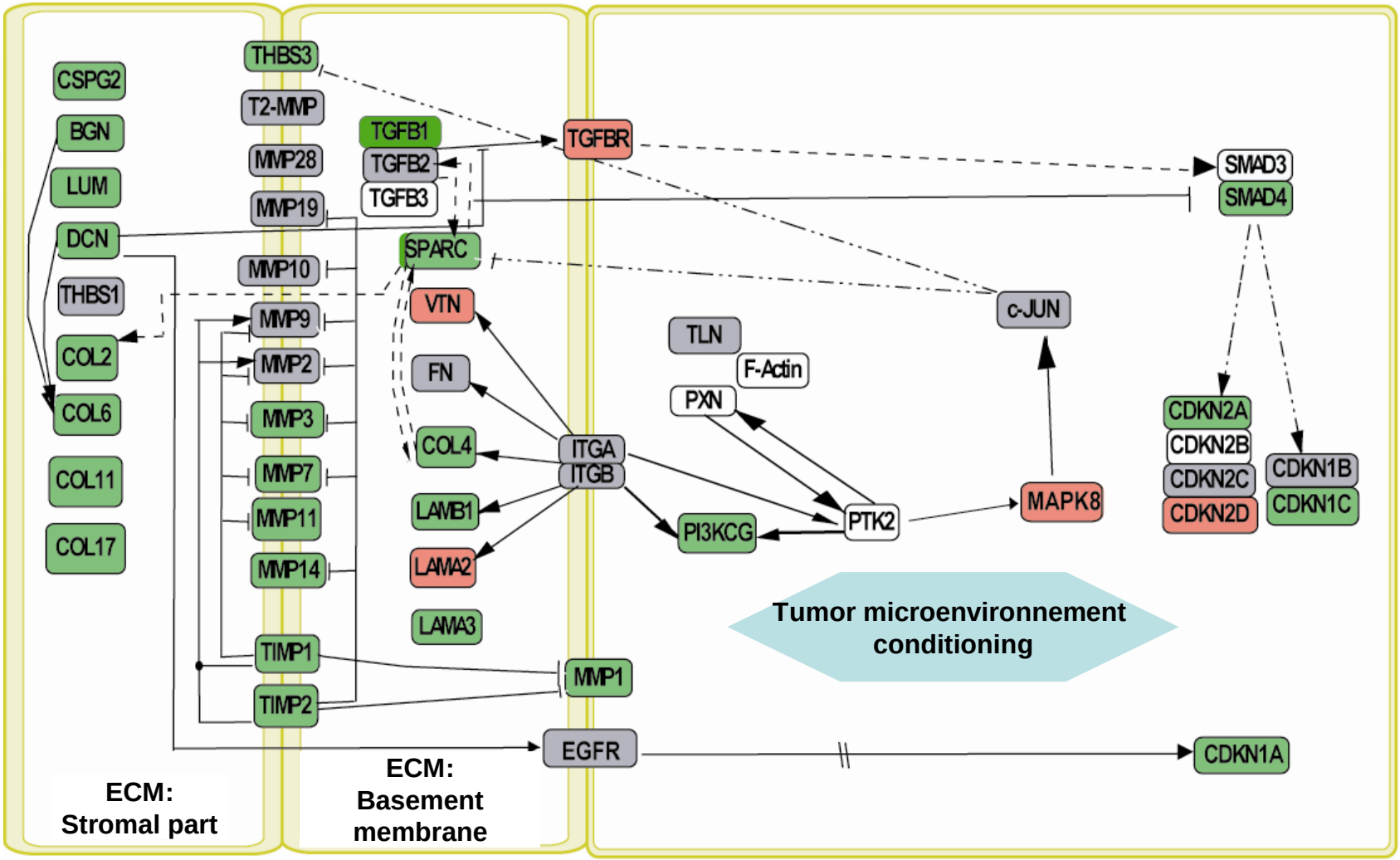
http://discover.nci.nih.gov/gominer/


1
2
3
4
5
6
7
8
9
10
11
12
Biological process
Cellular component
Regulation of biological process p=0.0001
Physiological process p=0.0112
Cellular process p=0
Development
Cell
Extracellular p=0, p=0.024
Regulation of cellular process p=0.0001
Death p=0.0131
Metabolism
Response to stimulus p=0.0007, p=0
Cellular physiological process p=0.041
Cell communication p=0
Morphogenesis p=0.0001
Intracellular
Extracellular matrix p=0
Cell Death p=0.0173
Nucleobase, nucleoside, nucleotide and nucleic acid metabolism
Response to stress p=0.0086, p=0
Response to endogenous Stimulus p=0.0373, p=0.0204
Response to external stimulus p=0.0076, p=0.0005
Cell growth and/or maintenance
Cell Adhesion p=0
Growth
Nucleus
Chromosome p=0.0002
Programmed cell death p=0.01
DNA Metabolism p=0.0018
Response to DNA damage stimulus p=0.0136
Response to abiotic stimulus p=0.0086
Transport
Cell Proliferation p=0.0044, p=0.0472
Cell organization and biogenesis
Nucleoplasm
Replication fork p=0.0278
Apoptosis p=0.0149
DNA recombination
DNA repair p=0.0508
Response to chemical substance p=0.0088, p=0.0142
Cell cycle p=0.0276
Cellular morphogenesis
Transcription factor complex
DNA-directed RNA polymerase II, holoenzyme
Response to drug p=0.041
DNA replication and chromosome cycle
M phase p=0.0109
Regulation of cell size
Transcription factor TFIIH Complex p=0.0102
DNA replication
Nuclear division p=0.0145
Cell growth p=0.0014
DNA-dependent DNA replication p=0.0404
Meiosis
Meiosis I
Meiotic prophase I
Meiotic recombination p=0.0306
147/554 GO terms found significantly enriched in L863 using overrepresentation analysis with EASE and GoMiner, a=0.05
Innate Tumor Drug Responses
GO Association Tree

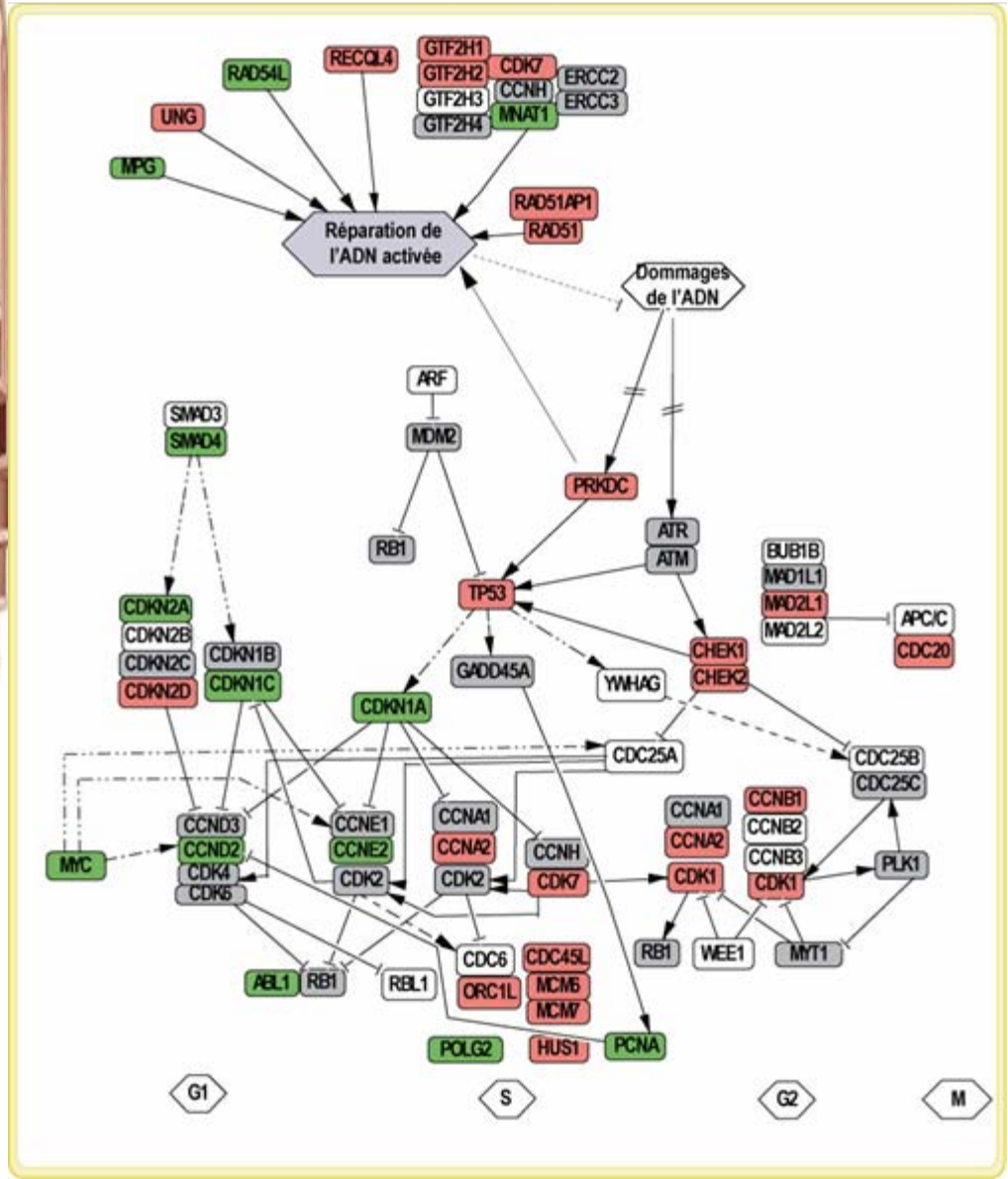
Cellular State Anticipating in vivo Innate Tumor Drug Resistance - CellDesigner

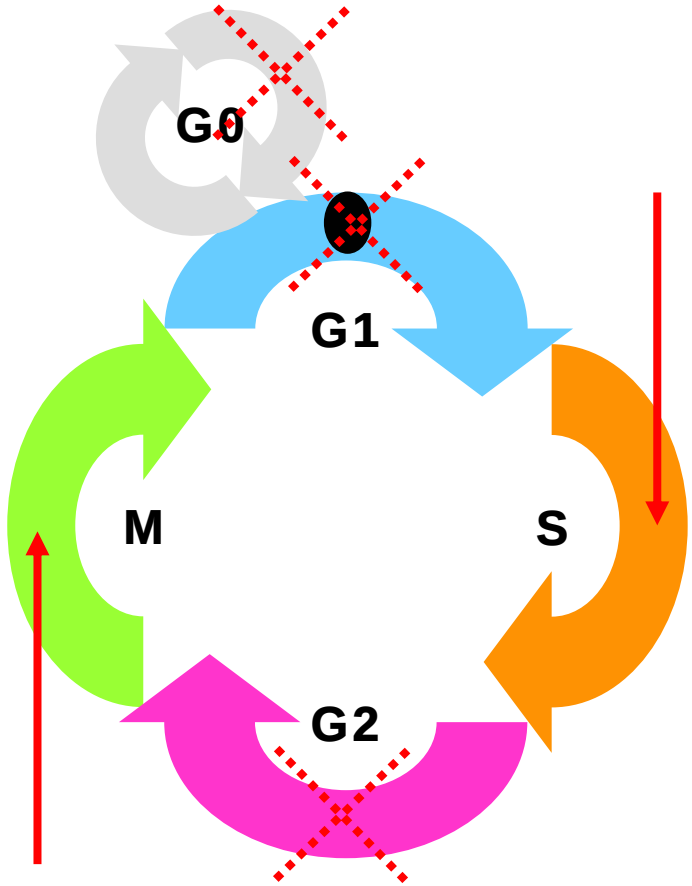


Tumor Microenvironment Conditioning



Cell cycle delays





Innate Tumor Drug Responses
Conclusions and Future Perspectives

First report of cellular states anticipating in vivo innate drug responses in tumor samples collected from colorectal cancer (CRC) patients prior to their exposure to a combined chemotherapy.

The establishment of a functional interaction molecular map represents a starting point that may be helpful to identify by-pass chemotherapy schemes to allow critical therapeutic intervention.

The identification of highly sensitive predictive gene sets may enhance the prognosis of primary tumor responses to subsequent chemotherapy schemes and provide further insights into the molecular characterization of tumor cells.

‘The 39 steps’ in gene expression profiling: critical issues and proposed best practices for microarray experiments

Sandrine Imbeaud and Charles Auffray

(2005) Drug Dicovery Today 10, 1175-1182

www.drugdiscoverytoday.com

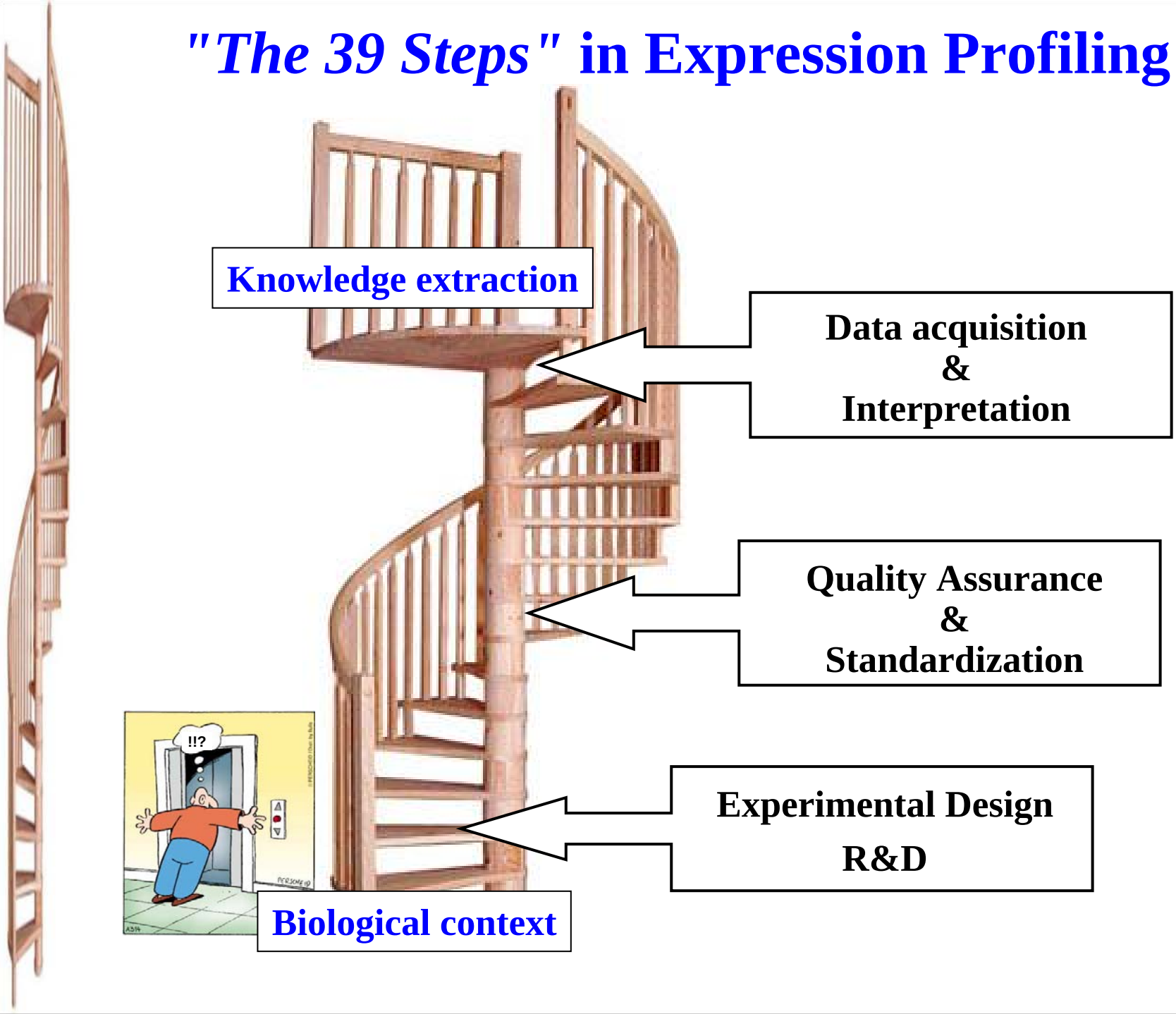
Extracting functional and regulatory order from microarrays

Sandrine Imbeaud and Charles Auffray

(2005) Mol. Syst. Biol. Msb4100013 E2

www.nature.com/msb/

"The 39 Steps" in Expression Profiling



Knowledge extraction

Data acquisition
&
Interpretation

Quality Assurance
&
Standardization

Experimental Design

R&D

Biological context

"The 39 Steps" in Expression Profiling

Experimental design

1. Objectives Articulation
2. Resources Allocation
3. Study Design
4. Power & Confidence
5. Pilot Collection
6. Stats Design Validation
7. Data Collection

Gene collection

8. Gene Coverage
9. Probe Selection
10. Resources Annotation
11. Confidence Metrics
12. Probe Preparation
13. Controls, QC Metrics

Sample collection

14. Sample Selection
15. Resources Cataloguing
16. Sample Preparation
17. Controls, QC Metrics

Array preparation

18. Instrument Calibration
19. Array Manufacture
20. Quality Controls

Target synthesis

21. Spike RNA Controls
22. Biochemical Reactions
23. Controls, QC Metrics

Hybridization

24. Hybridization/Mixing
25. Washing/Drying
26. Quality Controls

Data transformation

27. Image Acquisition
28. Image Segmentation
29. Data Subtraction/Filter
30. Data Normalization
31. Data Reduction
32. QC Metrics, Descriptive stats of the measures

Knowledge extraction

33. Ratio- or Intensity-Statistics
34. Genes Modules Identification
35. Functional Annotation
36. Networks Definition
37. 'Omics' Integration

(inter-disciplinary validations)

Data storage

38. Data Warehouse

(Data Repositories)

39. Data Integrity, Standards and Maintenance

Experimental Design
What does that mean in practice?

“The appropriate design of an experiment is the key to successful analysis of a problem for without the correct design you will never have the right sort of data”

“To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of” **(Pr. R.A. Fisher, 1938)**

Power simulation

False Negatives

False Discovery Rates

=> Biological inference, Sample size and subgroups definition

=> Statistical risk evaluation

Experimental Design and Power Analysis
Human Universal Reference

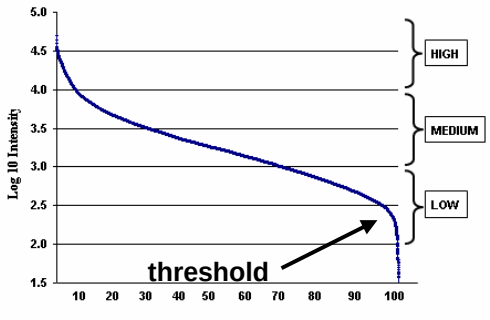
- **Reference Full-factorial Design**

(1) Easily extensive
(2) Simple interpretation
(3) Require less RNA per sample
(4) Insensitive to bad RNA samples

10 Hybs

6 samples : A, B, C, D, E and Ref

- **Human Universal Reference (Stratagene) 10 human cell lines**



One unique batch (N° 1000207)
More than 99% of all the items printed onto the array, including controls, were shown to hybridize:

- 6 % - high intensity
- 64 % - medium intensity
- 30 % - low intensity

More than 88% exhibiting intensities above the failure rate

PubMed
15,860,000 references
October 2005

• Informatics	7,005	(0.04%)	• Bioinformatics	10,126	(144%)
•System	1,199,305	(7.6%)	•Systems biology	549	(0.04%)
• DNA	829,478	(5%)	• Genome	108,470	(12.3%)
• RNA	499,990	(3.1%)	• Transcriptome	1,503	(0.3%)
• Protein	3,178,464	(20%)	• Proteome	5,482	(0.17%)
• Metabolism	4,096,898	(25.8%)	• Metabolome	164	(0.004%)
• Function	6,580,556	(41.5%)	• Physiome	28	(0.0004%)

The Drawbacks of Literature Mining

- ✓ More than 13 Million research articles registered in PubMed
- ✓ No *user's context* search
- ✓ Lack of publication based on *negative observations*
- ✓ A huge number of *false discoveries*

- ✓ **Synonymy**, the many ways in which to refer to the same object

Ex: AMH=MIS – Anti-Mullerian Hormone **and** Mullerian Inhibiting Substance

- ✓ **Polysemy**, the fact that a given word may have multiple meanings

Ex: PIP – prolactin-induced protein **and** phosphatidyl inositol-phosphate

Integrative annotation of 21,037 human genes validated by full-length cDNA clones

Imanishi et al. (2004) PLoS Biol. 2, 856-875

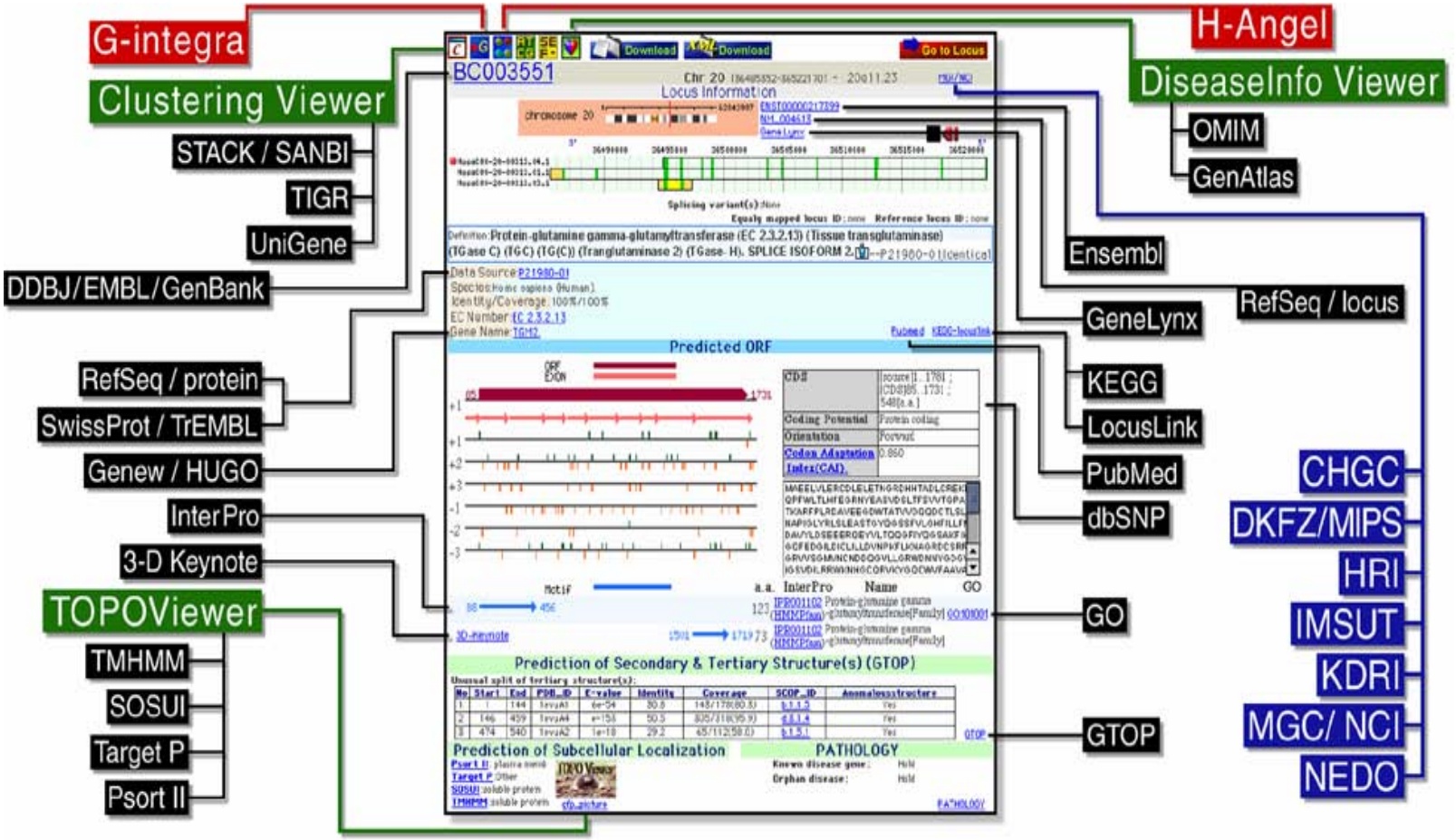
http://www.h-invitational.jp/

The Human ANatomic Gene Expression Library (H-ANGEL), the H-Inv integrative display of human gene expression across disparate technologies and platforms

Tanino et al. (2005) Nucl. Acids Res. 33, D567-572

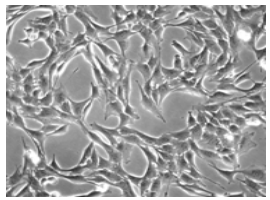
H-Invitational Database

H-InvDB is a comprehensive database integrating annotations of human genes based on human full-length cDNAs



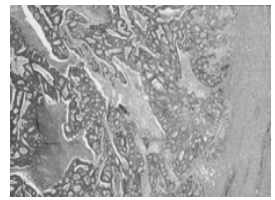
RNA integrity assessment - Imbeaud et *al.*, NAR, 33, e56, 2005

cell lines



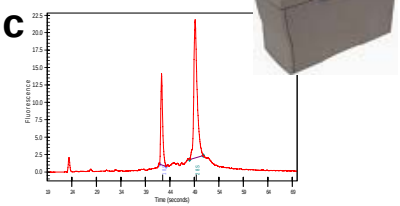
Biological samples

tissue samples



Total RNA

Agilent Bioanalyzer electrophoretic trace



UV absorbance spectrum

Evaluation of RNA purity

i.e. sugar, alcohol, detergent, or DNA contamination

RIN

Evaluation of RNA integrity

RIN number from 1 to 10

The RNA quality metrics

1. Integrity metrics designation per sample => standardization, exchange language
2. Selection of samples for downstream experiments => sample use validation
3. Class of integrity designation => sample use cataloging

Array s/IMAGE - Reproducibility

Samples →

Tech |r| ↓

1	2	3	4	5	6
0,50	0,25	0,13	0,06	0,03	0,02
0,60	0,36	0,22	0,13	0,08	0,05
0,70	0,49	0,34	0,24	0,17	0,12
0,75	0,56	0,42	0,32	0,24	0,18
0,80	0,64	0,51	0,41	0,33	0,26
0,85	0,72	0,61	0,52	0,44	0,38
0,90	0,81	0,73	0,66	0,59	0,53
0,95	0,90	0,86	0,81	0,77	0,74
0,97	0,94	0,91	0,88	0,86	0,83
0,98	0,96	0,94	0,92	0,90	0,89
0.99	0.98	0.97	0.96	0.95	0.94
1,00	1,00	1,00	1,00	1,00	1,00

Use of multiple samples AND multiple technical slides

Array s/IMAGE - Data Analysis Tools

Class Comparison
Class Prediction
Hierarchical (Un)Supervised Clustering
Principal Component Analysis
Controlled Vocabularies
Ontology Mining

- ***t*- and *z*-statistics with permutation and multiple-testing effect correction (Stepdown Bonferroni & FDR)**
- **Significance Analysis of Microarrays (SAM)**
- **1-way & 2-way ANOVA**
- **Binary tree Prediction (compound covariate prediction, diagonal linear discriminant analysis, K nearest neighbors, nearest centroid, support vector machine)**

i.e. ArrayStat (Imaging Research Inc.), XLstat (Addinsoft), Bioconductor (R packages), Genesis (TUG), TMEV (TIGR), BRB ArrayTools (NCI), dCHIP (NIH), GeneTraffic duo (Iobion Informatics) and many others

Genexpress - CNRS UMR 7091 – Villejuif - France

***Principal Investigator*: Charles Auffray**

Functional Genomics

Sandrine Imbeaud

Esther Graudens

Virginie Boulanger

Computional Genomics

Eric Eveno

Régine Mariage-Samson

Biovalidation

Dominique Piatier-Tonneau

Sandrine El Marhomy

Philippe Riou

Partnerships

Patrick Zaborski

Corinne Sébastiani

Inserm – Institut national de la santé et de la recherche médicale

CNRS – Centre National de la Recherche Scientifique

Groupe sanofi aventis – L'essentiel c'est la santé.

C.R.L.C. Centre Val D'Aurelle -Paul Lamarque – Centre Régional de Lutte Contre le Cancer

Université Joseph Fourier

QIAGEN

Agilent Technologies

Amersham Biosciences – Part of GE Healthcare

INTERSCAN SYSTEMES – Les Spécialistes de l'intégration des systèmes codes-barres

Charles AUFFRAY, Laurent NOTTALE and the French SYSTEMOSCOPE Consortium

The SYSTEMOSCOPE International Consortium : Promoting Trans-disciplinary Research in Systems Biology for Health

Scale Relativity in Systems Biology of Muscular and Pulmonary Diseases

French SYSTEMOSCOPE Consortium
Trans-disciplinary Research Program in Systems Biology

Multiscale Functional Networks
in Muscular and Pulmonary Physiopathology

Charles AUFFRAY, biologist, UPMC/CNRS (Coordinator)

Dominique CHARRON, biologist, clinician, INSERM/UDD/UPMC

Jean-Pierre FRANCOISE, mathematician, UPMC

Giuseppe LONGO, mathematician, ENS/CNRS

Jacques MALLET, biologist, UPMC/CNRS

Laurent NOTTALE, physicist, Paris-Meudon Observatory/UDD/CNRS

Christophe PISON, clinician and Valdur SAKS, biologist, INSERM/CHU/UJF

Self-organized living systems: conjunction of a stable organization with chaotic fluctuations in biological space-time.

Auffray, C., Imbeaud, S., Roux-Rouquié, M., and Hood, L.

Philos. Transact. Roy. Soc. Math. Phys. Eng. Sci. (2003) 361, 1125-39.

Living systems have the ability to organize themselves as the result of a conjunction occurring through an interface between the variable part of a mostly stable physical organization, and the stable part of a chaotic network of small fluctuations.

These small fluctuations, which are inaccessible to currently available tools, may be the major determinants of the behaviour of biological systems because they convey collectively the most important part of biological information.

Self-organized living systems: conjunction of a stable organization with chaotic fluctuations in biological space-time.

Auffray, C., Imbeaud, S., Roux-Rouquié, M., and Hood, L.
Philos. Transact. Roy. Soc. Math. Phys. Eng. Sci. (2003) 361, 1125-39.

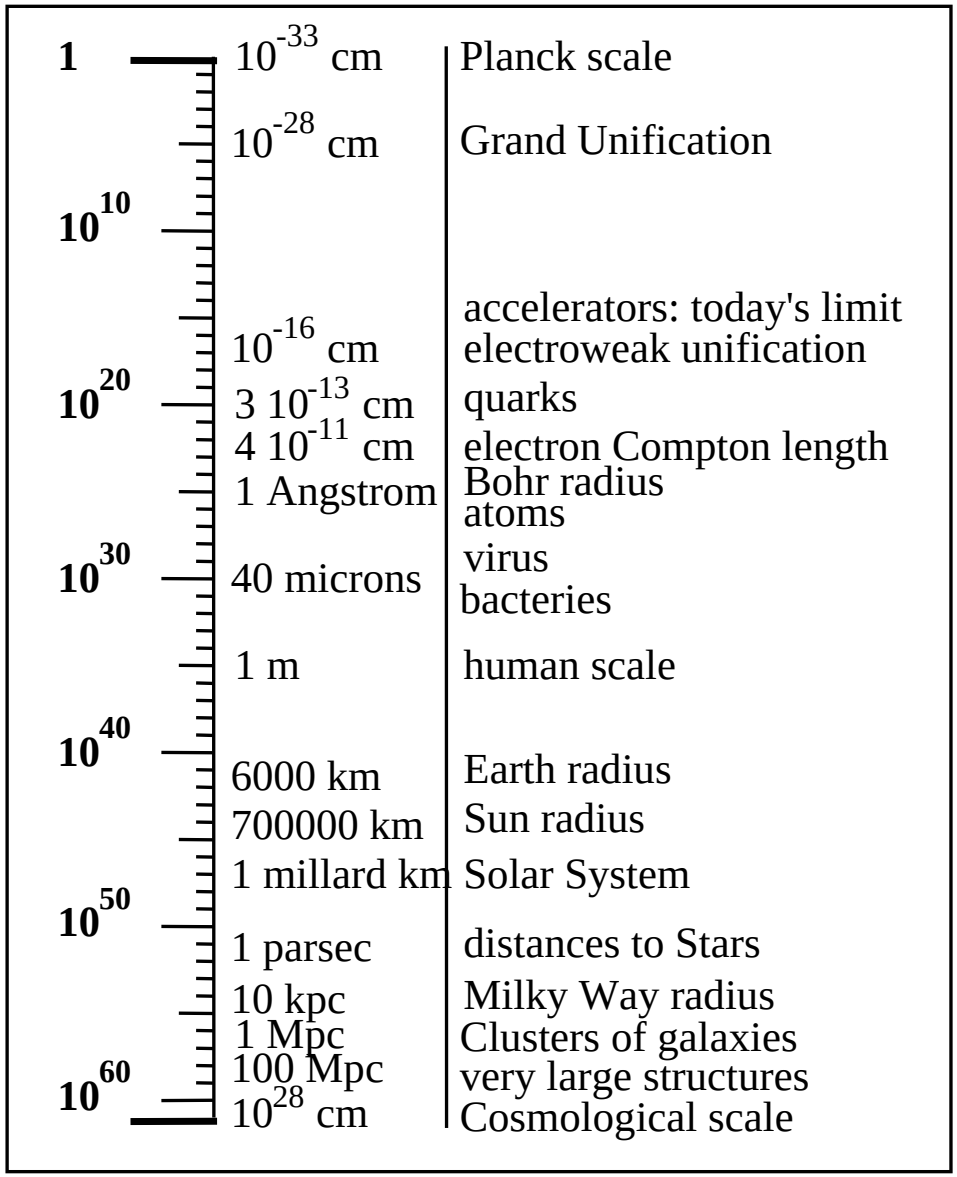
Complex biological systems operate in a space with a variable number of dimensions.

Detection of small changes of low intensity signals will require the development of a new conceptual and practical framework combining in an iterative mode systemic modelling of biological systems, to generate hypotheses, together with a high level of standardization of high-throughput platforms enabling reliable cross comparisons, to test them.

The Theory of Scale Relativity : Non-Differentiable Geometry and Fractal Space-Time

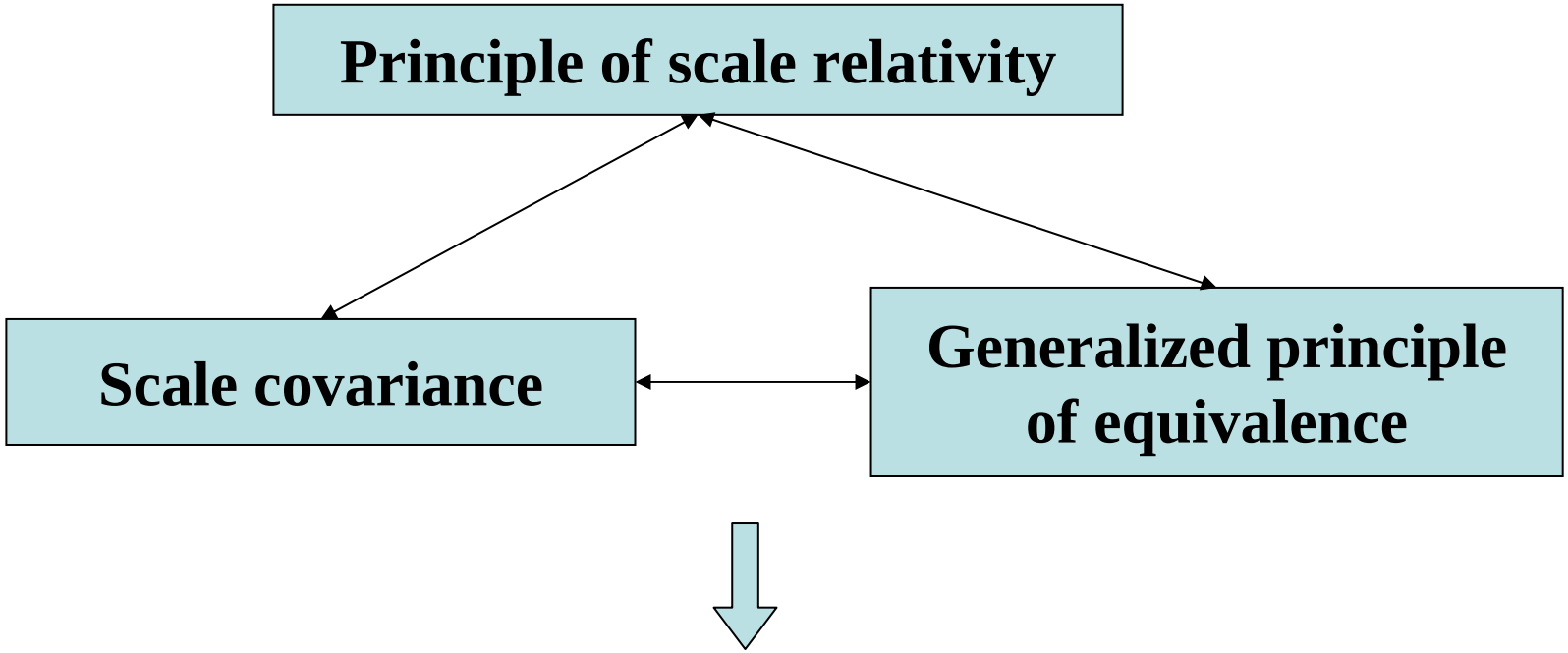
Nottale, L. (2004) CASYS'03 American Institute of Physics Conference Proceedings 718, 68-95

(http://wwwusr.obspm.fr/~nottale/)



Scales in nature

The Theory of Scale Relativity : Non-Differentiable Geometry and Fractal Space-Time
Nottale, L. (2004) CASYS'03 American Institute of Physics Conference Proceedings 718, 68-95 (http://wwwusr.obspm.fr/~nottale/)



Linear scale-laws: "Galilean"	Linear scale-laws : "Lorentzian"	Non-linear scale-laws:
self-similarity,	varying fractal dimension,	general scale-relativity,
constant fractal dimension,	scale covariance,	scale dynamics,
scale invariance	invariant limiting scales	gauge fields