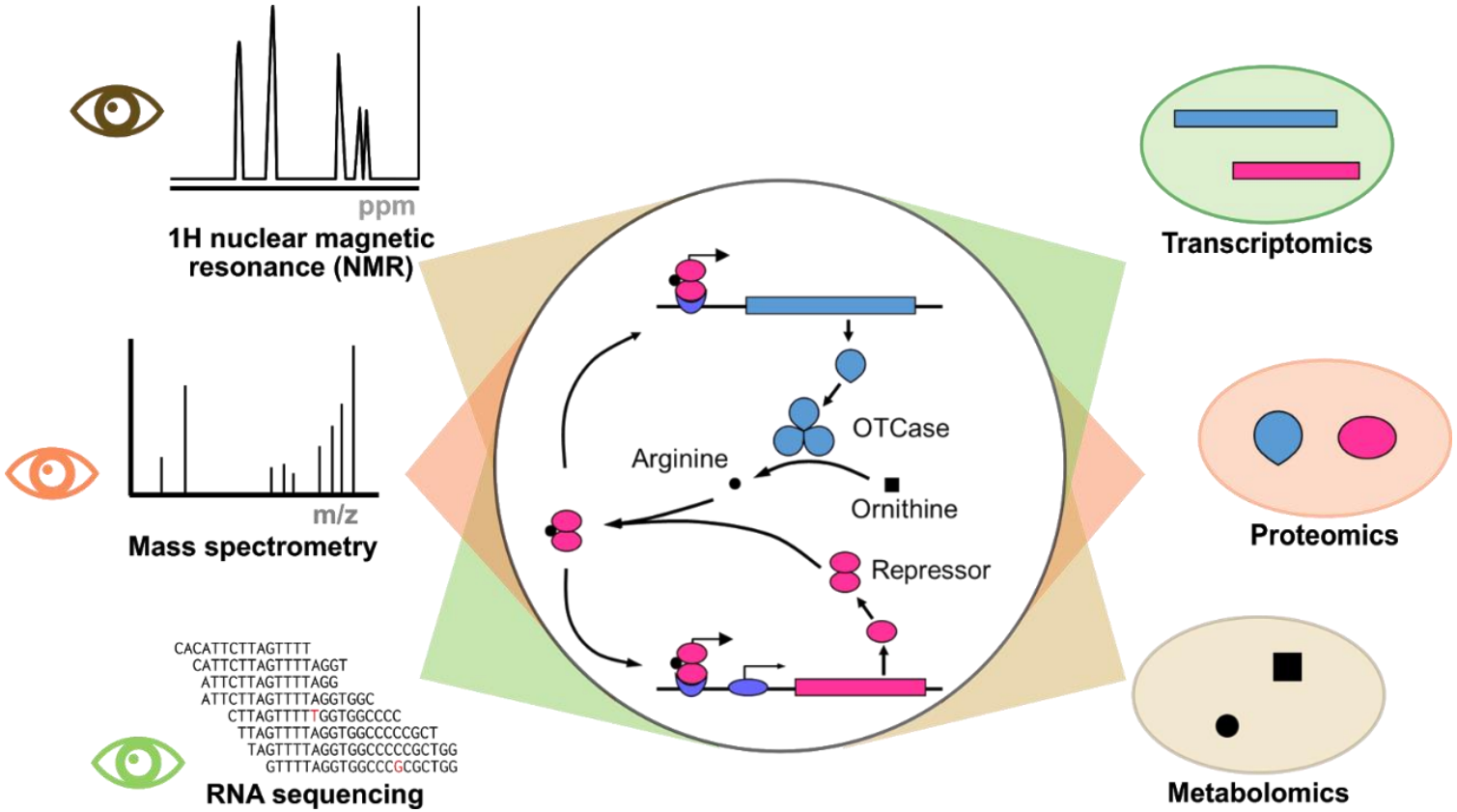
AMIDD Lecture 7: Cellular and omics modelling



Omics data are projections of high-dimensional biological space. It is an *inverse problem* to infer a high-dimensional space from its projections.

Multiscale Modelling of Drug Mechanism and Safety by Zhang, Sach-Peltason, Kramer, Wang and Ebeling, in revision

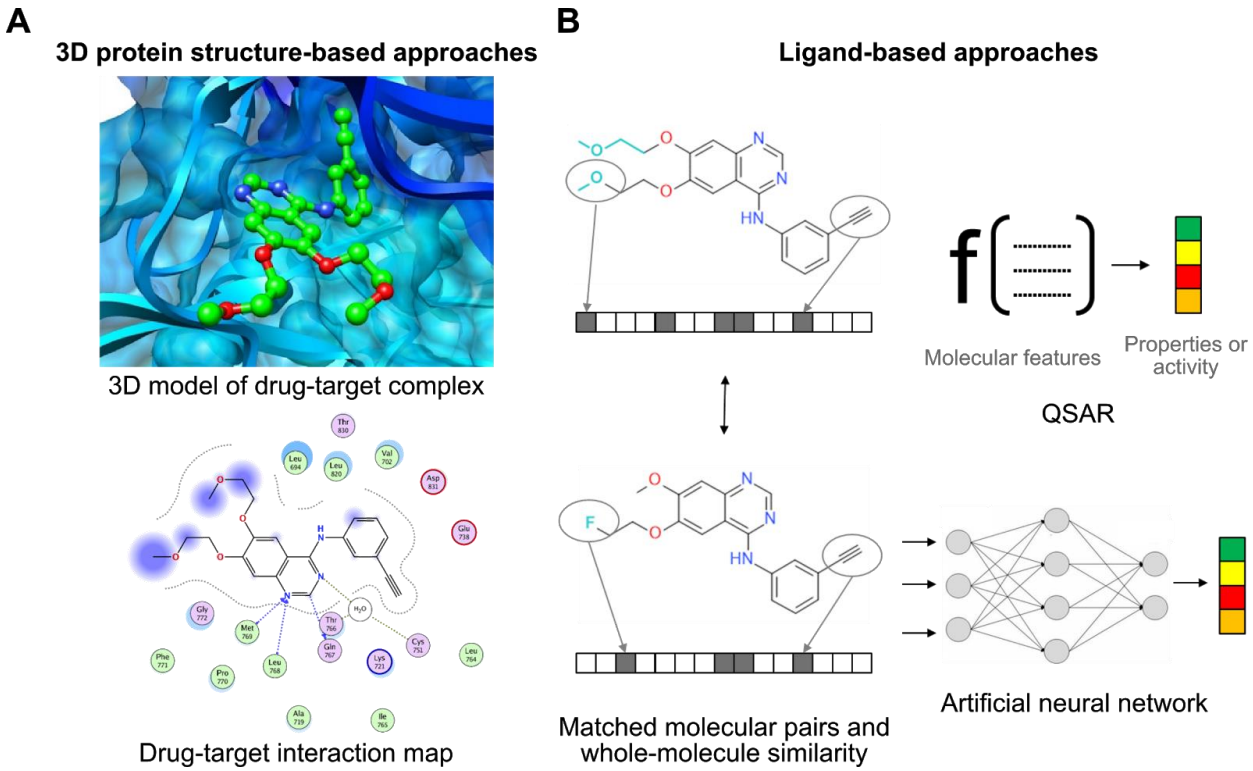
Dr. Jitao David Zhang, Computational Biologist

[1] Pharmaceutical Sciences, Pharma Research and Early Development, Roche Innovation Center Basel, F. Hoffmann-La Roche

[2] Department of Mathematics and Informatics, University of Basel

Contact the author

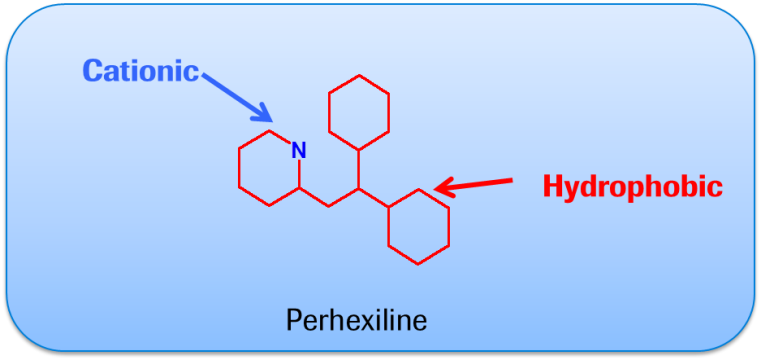
Recapture of the previous lecture



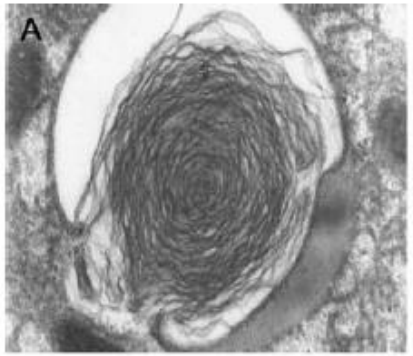
Overview of molecular-level modelling techniques

Drug-induced phospholipidosis is correlated with amphiphilicity

- Phospholipidosis is a lysosomal storage disorder characterized by the excess accumulation of phospholipids in tissues.
- Drug-induced phospholipidosis is caused by cationic amphiphilic drugs and some cationic hydrophilic drugs.
- Clinical pharmacokinetic characteristics of drug-induced phospholipidosis include (1) very long terminal half lives, (2) high volumn of distribution, (3) tissue accumulation upon frequent dosing, and (4) deficit in drug metabolism.



Lüllmann *et al*., Drug Induced Phospholipidosis, *Crit. Rev. Toxicol. 4, 185, 1975*



Anderson and Borlak, Drug-Induced Phospholipidosis,. *FEBS Letters* 580, Nr. 23 (2006): 5533–40.

$$\vec{A} = \sum_i d \cdot \vec{\alpha}_i$$

Fischer *et al.* (Chimia 2000) discovered that it is possible to predict the amphiphilicity property of druglike molecules by calculating the amphiphilic moment using a simple equation.

$\vec{A}$: Caculated amphiphilic moment

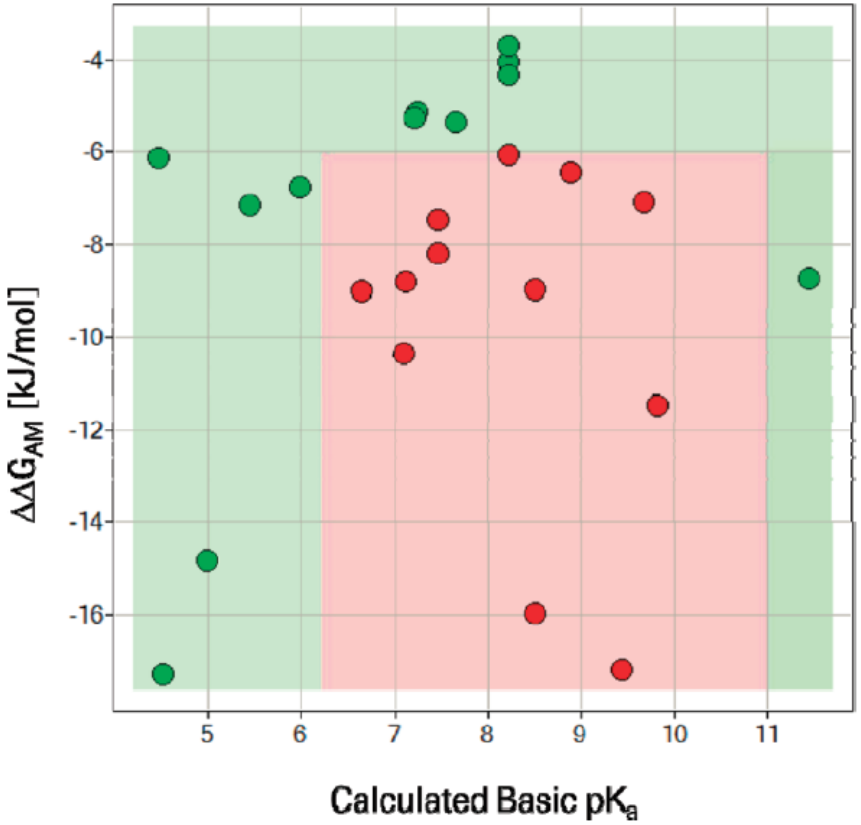
d: distance between the center of gravity of the charged part of a molecule and the hydrophobic/hydrophilic remnant of the molecule

$\vec{\alpha}_i$: the hydrophobic/hydrophilic contribution of atom/fragment *i*

***In silico* calculation of amphiphilicity property may be used to predict phospholipidosis induction potential**

In silico Phospholipidosis prediction

Model Validation from 1999-2004



Plot of amphiphilicity ($\Delta\Delta G_{AM}$) versus calculated basic pK_a for the training set of 24 compounds. The red area defines the region where a positive PLD response is expected, and the green area defines where a negative response is expected according to the tool.

in vitro/ in vivo	in silico/ in vivo	Exp. PC/ in vivo	In silico/ in vitro	n=36
94%	81%	89%	89%	

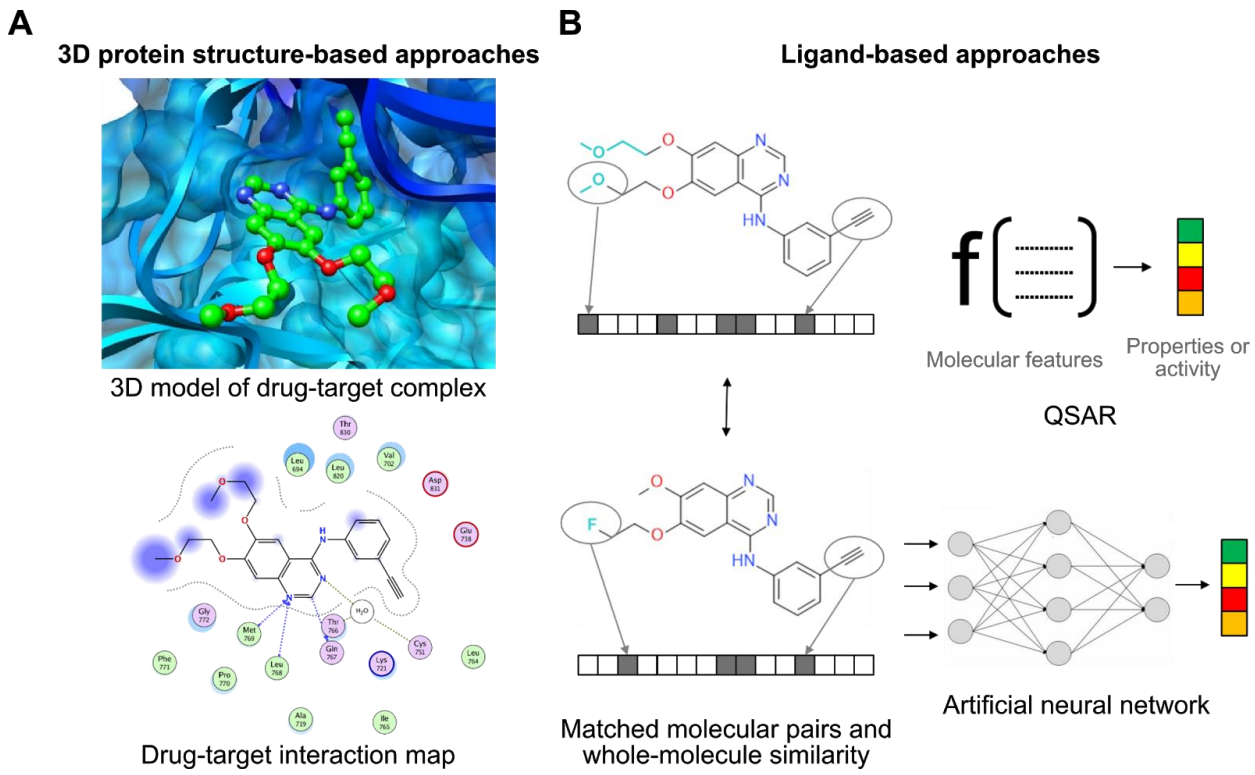
in vitro/in silico			n=422
Accuracy [(TP+TN)/ (P+N)]	**Sensitivity [True Positive Rate]**	**Specificity [True Negative Rate]**	**Precision [TP/(TP+FP)]**
86%	80%	90%	84%

Fischer et al., *J. Med. Chem, 55 (1), 2012*

We gained mechanistic insights of phospholipidosis induction by cationic amphiphilic drugs with the model

Phospholipidosis: lessons learned

- Cationic amphiphilic properties of a molecule is an early marker for safety in drug discovery and early development.
  - Phospholipidosis in dose range finding studies
  - Cardiac ion channel interactions (hERG, natrium channel, …)
  - Receptor binding promiscuity
  - P-gp inhibition
  - Mitochondrial toxicity in case of safety relevant findings, *e.g.* in dose range finding studies
- Extreme basic amphiphilic properties should be avoided because of a higher risk of PLD, QT-prolongation, mitochondrial toxicity. However, basic compounds with moderate amphiphilic properties are still a preferred scaffold for many therapeutic areas (especially CNS).
- **Generally, some safety liabilities, despite complex underlying biological and chemical mechanisms, can be predicted by molecular modelling well, sometimes with surprisingly elegant models!**

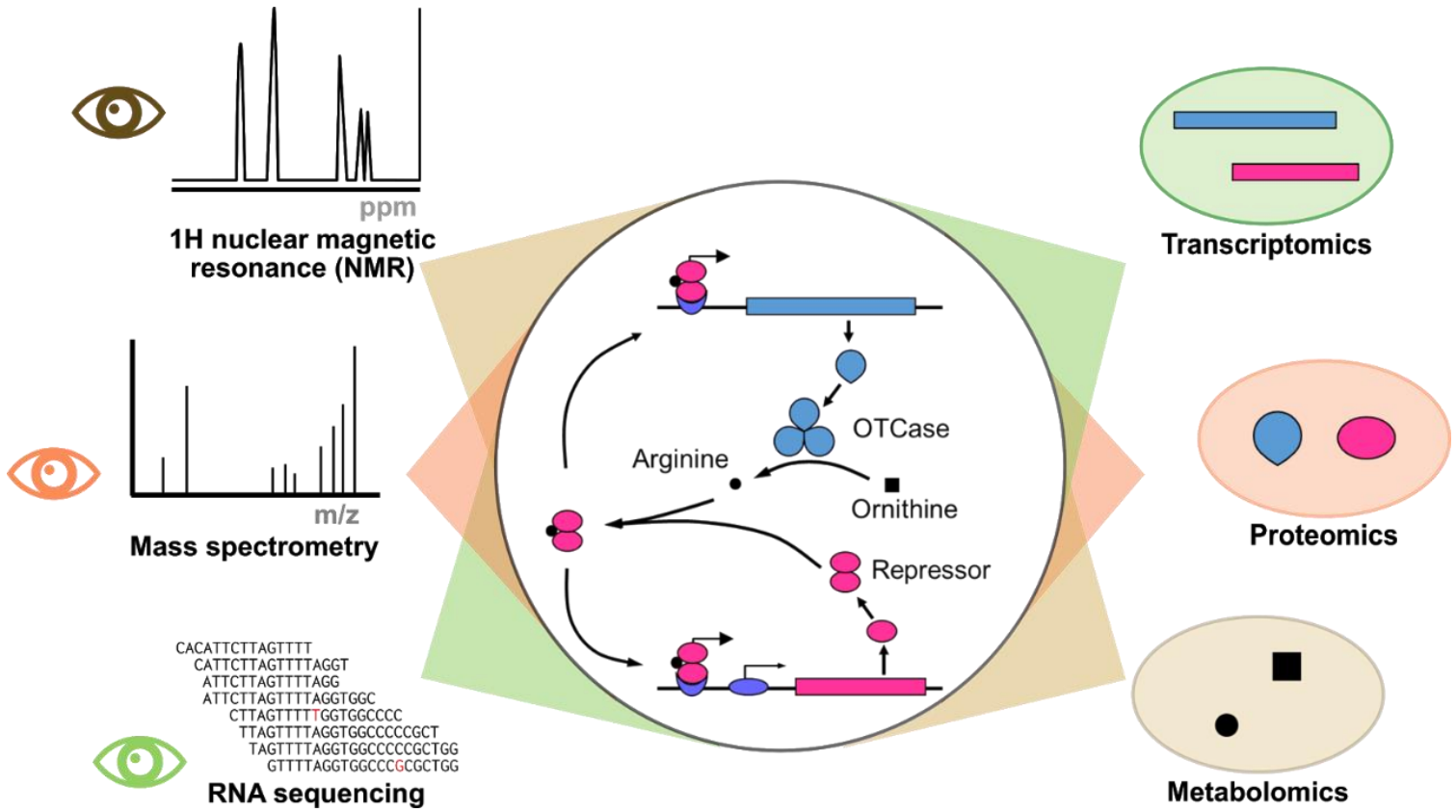


Overview of molecular-level modelling techniques

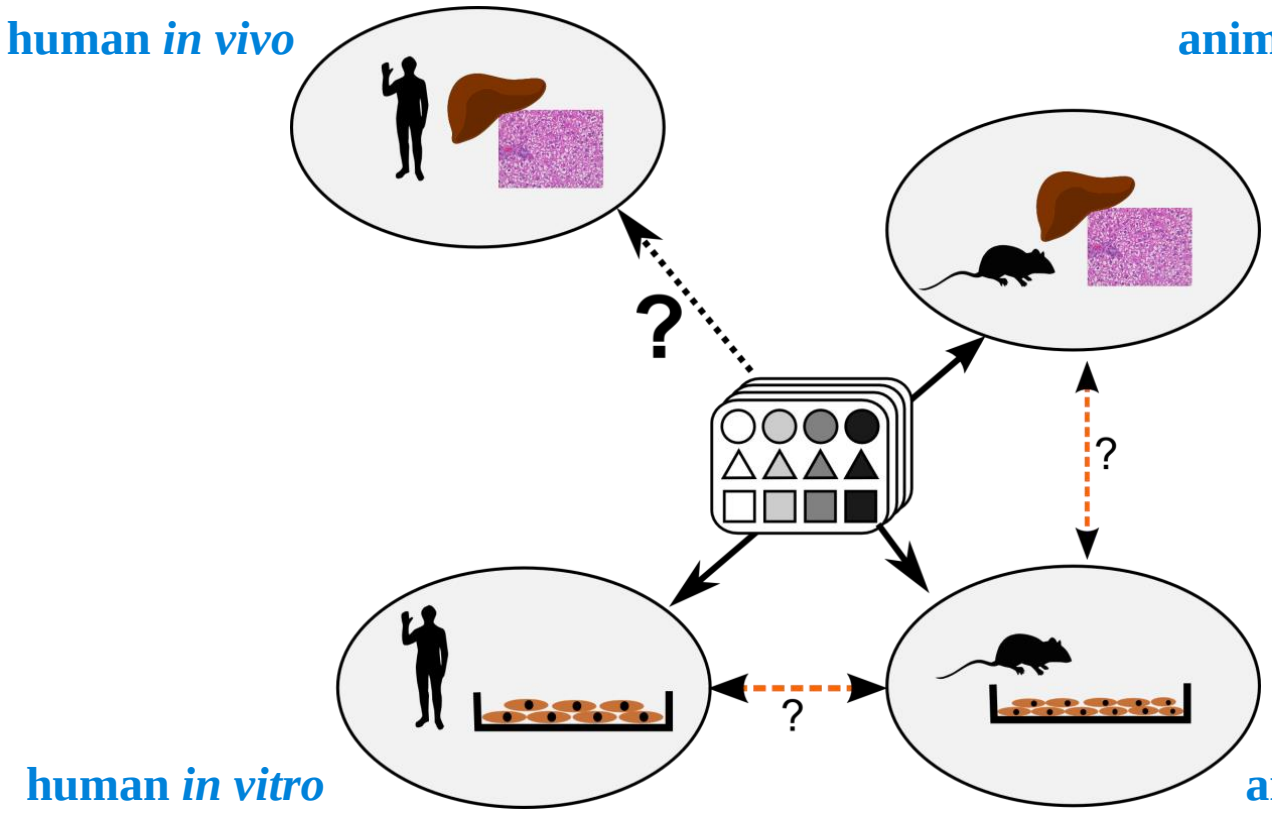
Topics

- **Gene expression profiling: a case study of omics and cellular modelling**
- **Applications for drug safety: TG-GATEs**
- **Applications for drug mechanism: molecular phenotyping**

Omics data are projections of high-dimensional biological space



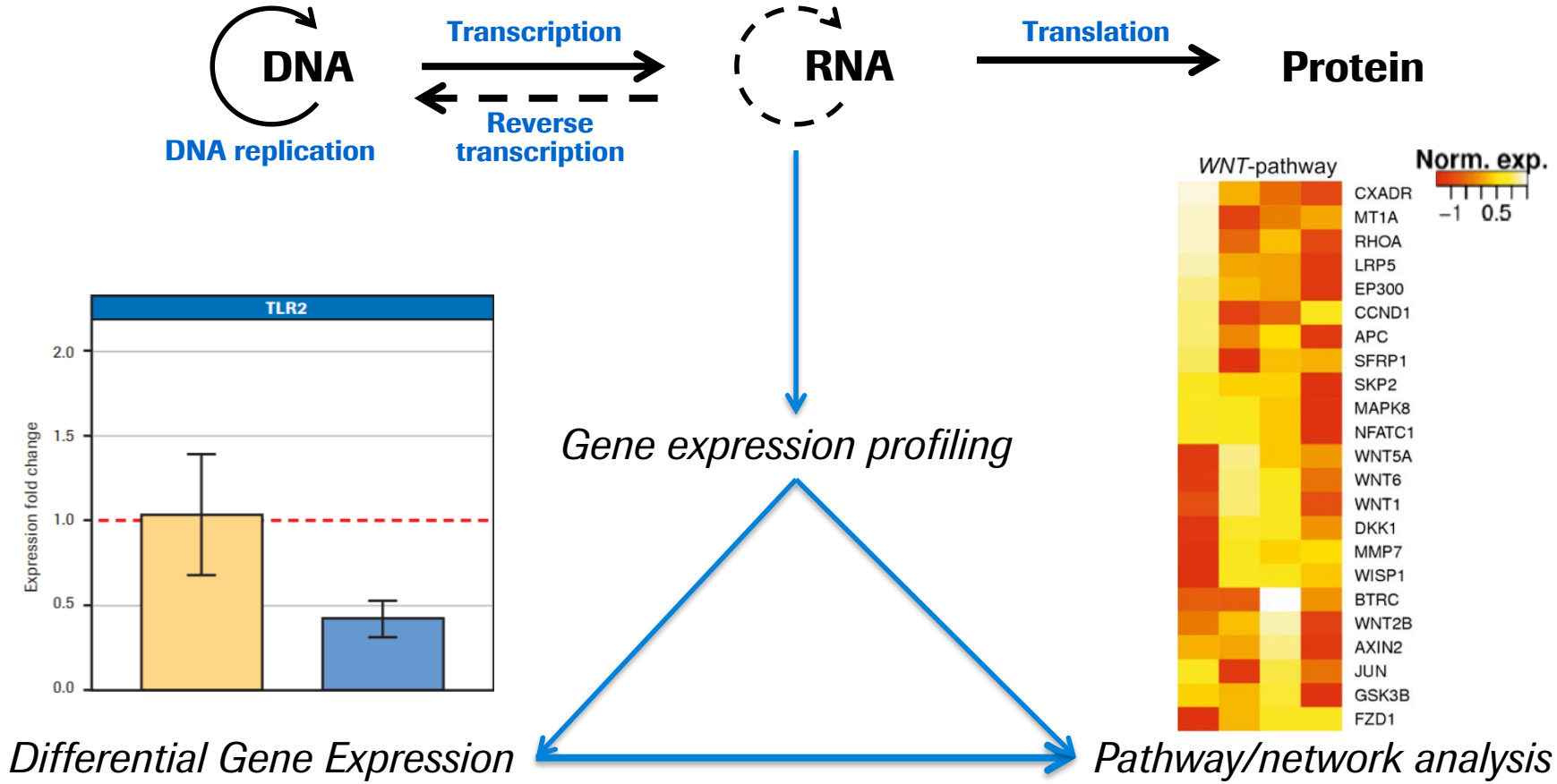
One challenge in drug discovery: non-clinical safety assessment



- Limited *in vitro-in vivo* and cross-species translatability
- Conflict between black-box prediction methods and the need to understand the mode of action

We need better (and interpretable) tools to predict safety profiles of drug candidates

Principles of gene expression profiling



Figures: Wikimedia Commons/Thomas Shafee, CC/Adapted

TG-GATEs: Toxicogenomics Project-Genomics Assisted Toxicity Evaluation system

- **Japanese Consortium 2002-2011**
  - National Institute of Biomedical Innovation, National Institute of Health Sciences, and 15 pharmaceutical companies, including Roche Chugai.
- **Data fully released in 2012 to the public:** Time-series and dose-dependent experiments using 170 bioactive compounds
  - *In vitro* & *in vivo* gene expression profiling, each containing gene expression data of about 20,000 genes
  - *In vitro* PicoGreen DNA quantification assay
  - *In vivo* histopathology in liver and kidney
  - *In vivo* clinical chemistry
- **Total raw data size >2 TB**

170
Compounds

>2000
Cellular assays

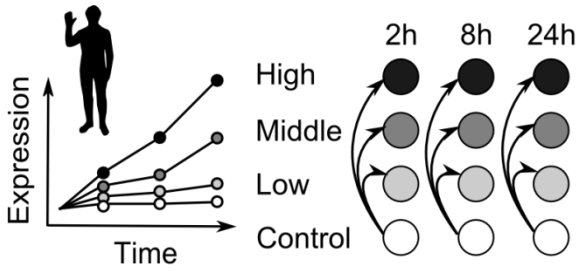
>12000
Pathology records

>24000
Expression profiles

TG-GATEs is a valuable data source to study drug-induced toxicity *in vitro* and *in vivo*

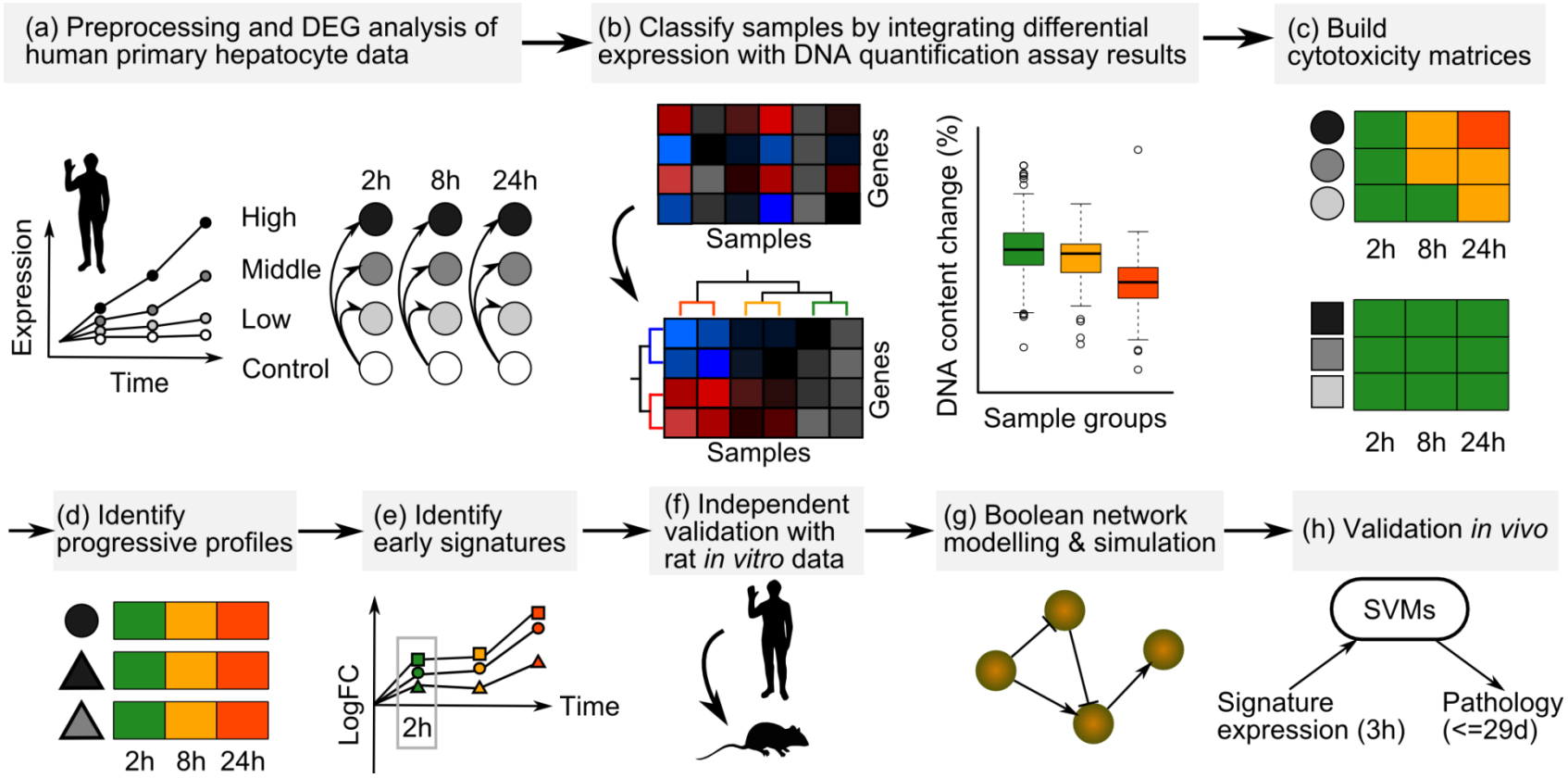
We built a computational pipeline to identify early signatures of toxicity

(a) Preprocessing and DEG analysis of human primary hepatocyte data



We integrate unsupervised learning, regression analysis, and network modelling to reach the goal

We built a computational pipeline to identify early signatures of toxicity (without animation)

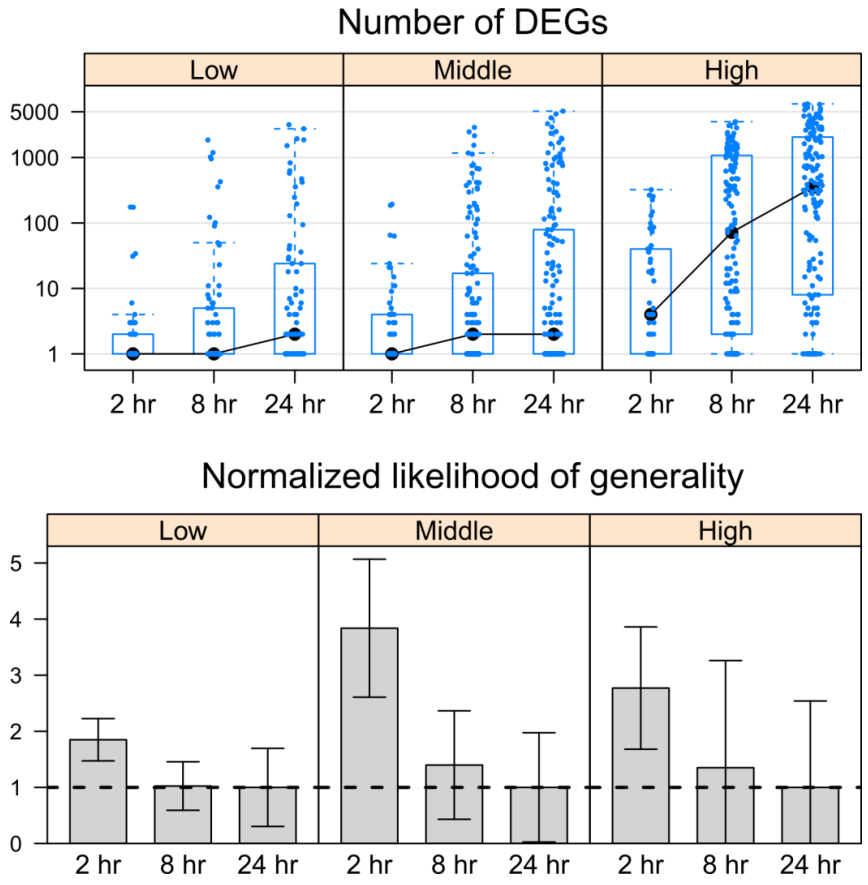


We integrate unsupervised learning, regression analysis, and network modelling to reach the goal

It is worth observing early

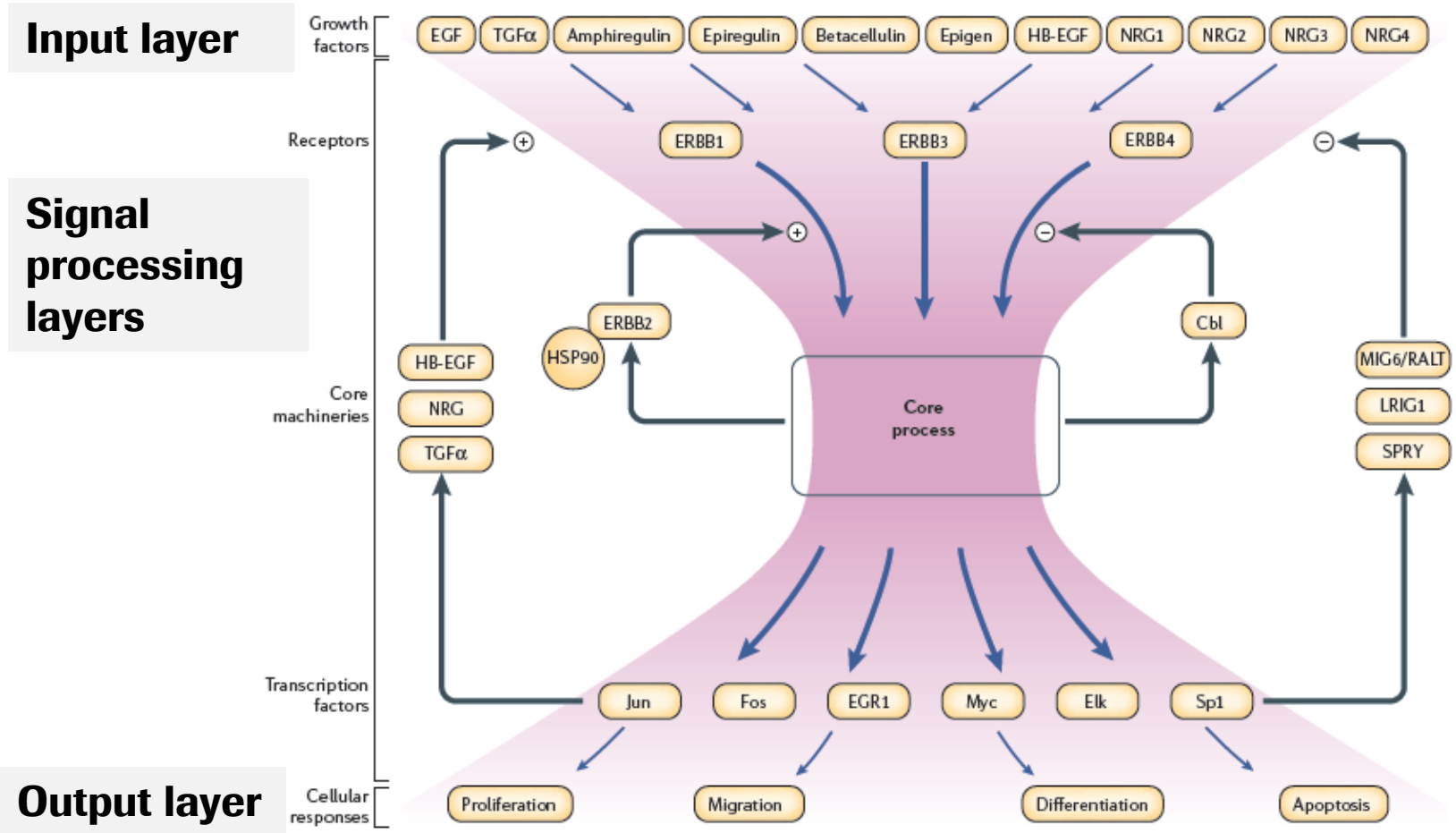
- We found that early-response genes induced 2h after compound administration are more generic (less specific) than late-induced genes: they are more likely to be induced by multiple compounds.

- → We hypothesize that diverse signalling pathways «back-converge» to a few early-response genes, which can be toxicity signatures.



Contrary to common wisdom (at the time), we argue that toxicogenomics should focus on early time points

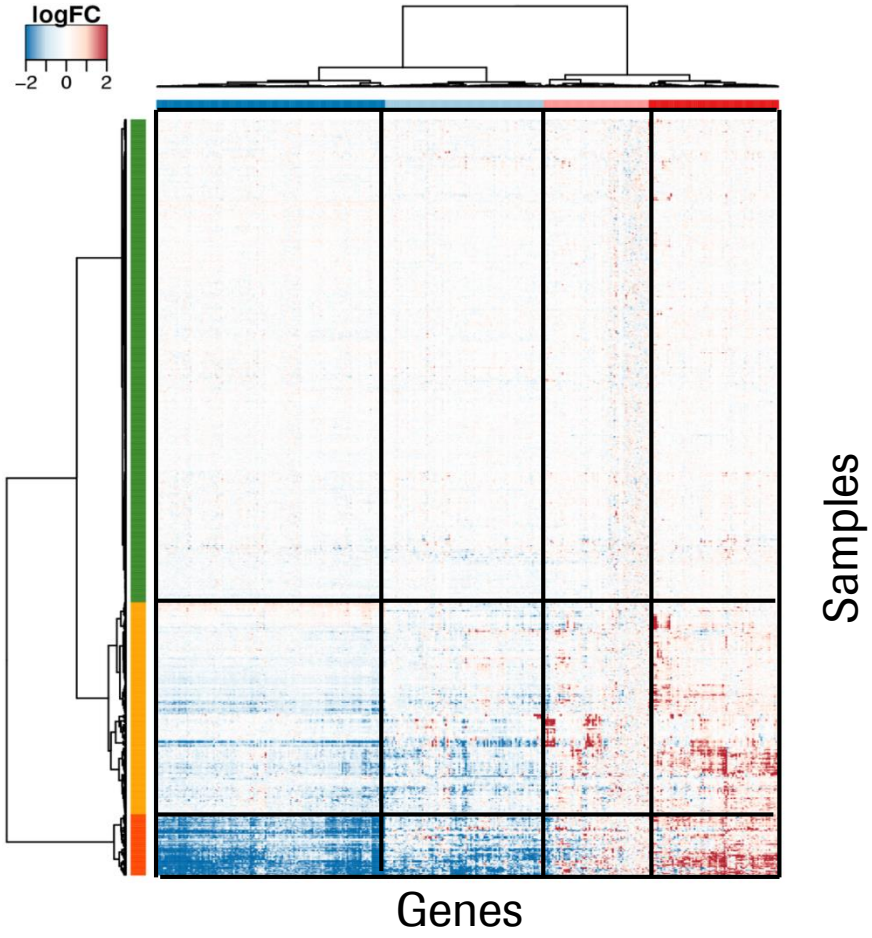
The bow-tie structure of signalling networks as a model that explains the power of early time point

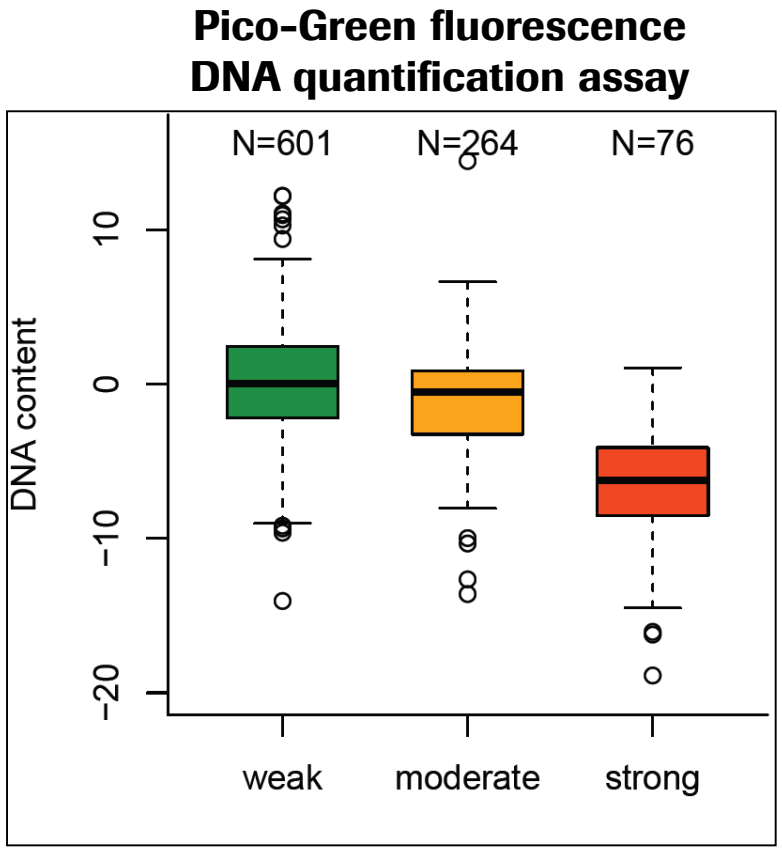


Adapted from Ami Citri and Yosef Yarden, Nature Reviews Molecular Cell Biology (2005)

We hypothesize that signalling pathways that mediate toxicity «back-converge» to a few early-response genes

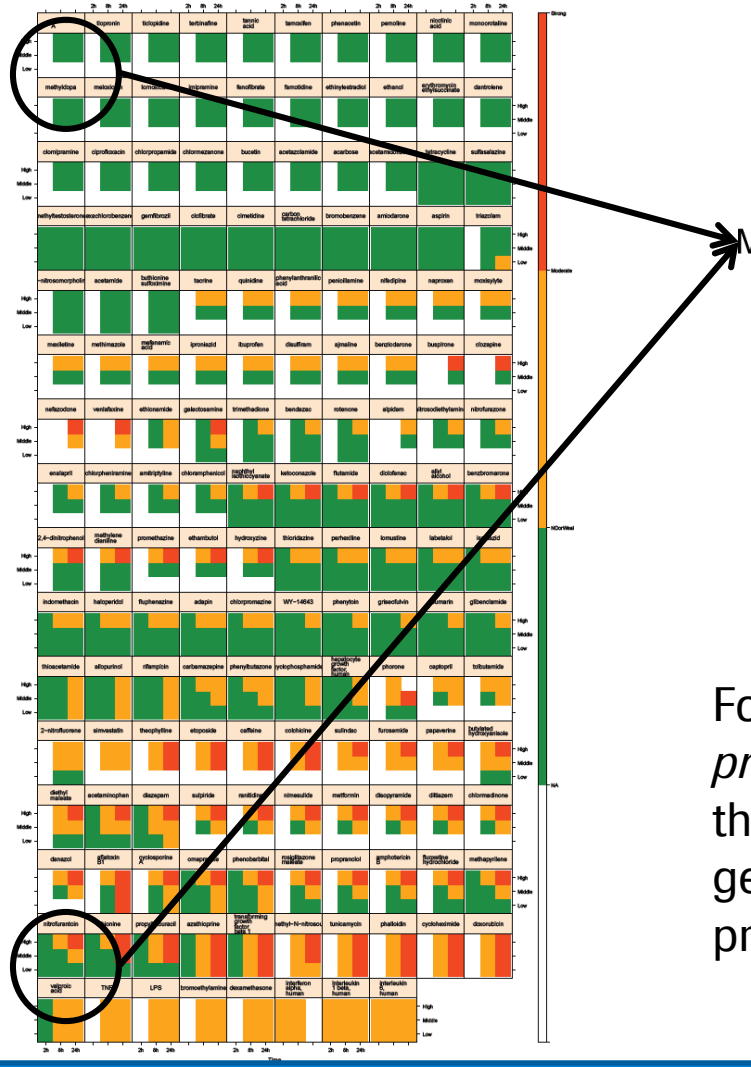
Compound-induced cytotoxicity can be classified into three levels by molecular phenotypes

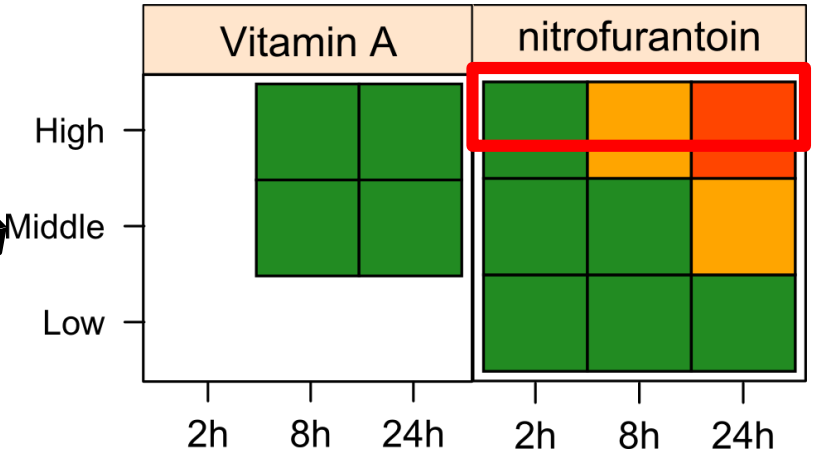




Unsupervised clustering identified groups of compounds associated with cytotoxicity

Cytotoxicity matrix and early signatures identified from progressive profiles *in vitro*

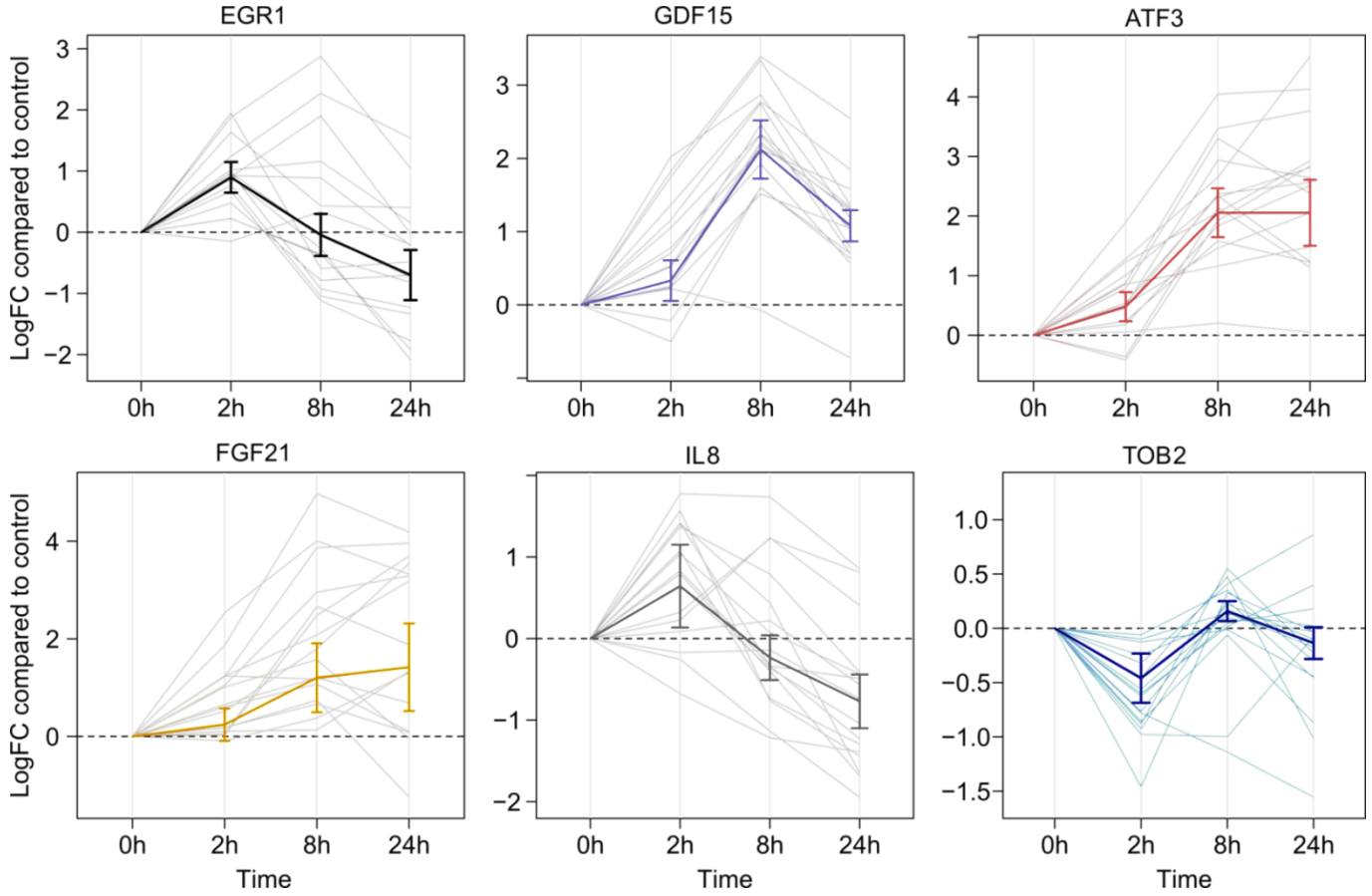




For **predictive signatures,** we focused on *progressive cytotoxicity profiles* (such as the one in the red box) and identified genes commonly regulated at 2h in such profiles.

Unsupervised clustering allowed us to identify progressive cytotoxicity profiles

Expression patterns of early signatures in human



Genes which were consistently and significantly up- or down-regulated at 2h in progressive profiles were chosen as signatures (|logFC|>0.25 & p<0.05). Purely data-driven: no biological knowledge was used for prioritization.

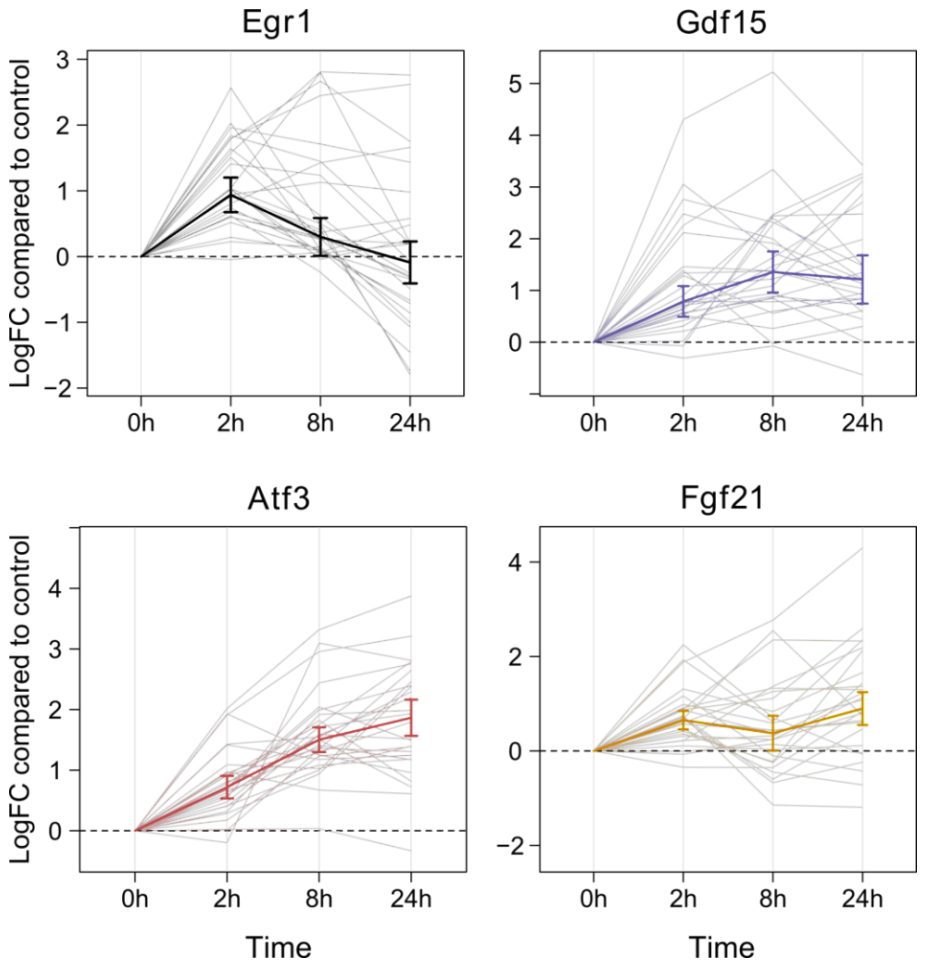
Each thin line represents one treatment, and the thick line represents the average.

Six early gene signatures of cytotoxicity were identified from human *in vitro* data

A consensus signature set of cytotoxicity emerges

Out of six early signatures in human, four are early signatures of progressive profiles in rat: Egr1, Atf3, Gdf15, and Fgf21.

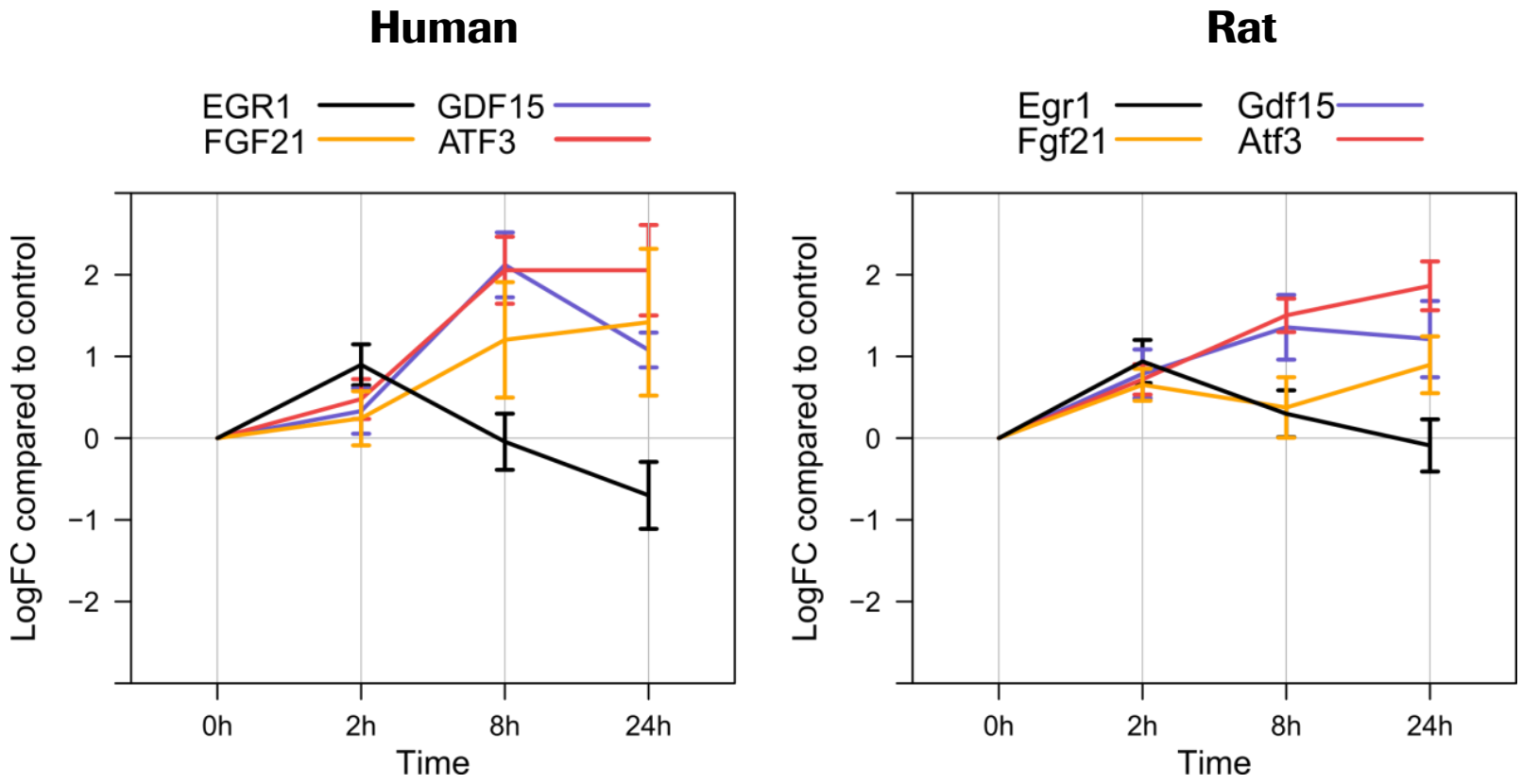
IL-8 does not have rat orthologue; Tob2 shows a similar pattern, but statistically was not significant.



Early-response signatures in rat. Each thin line represents one treatment, and the thick line represents the average. The identification was driven by rat data only.

Out of six human signatures, four were also found in rat *in vitro* data

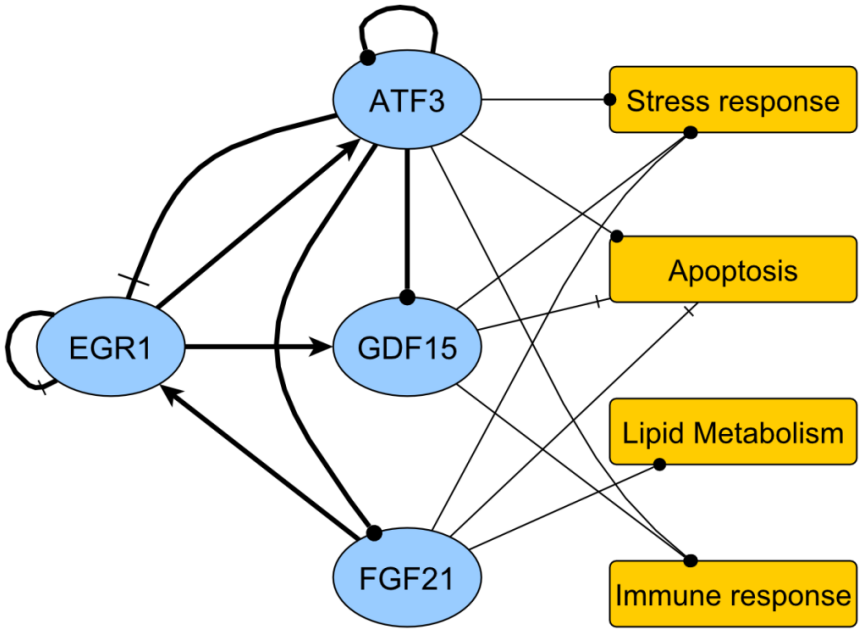
Conserved dynamics of the early signatures in human and rat primary hepatocytes



Lines represent average inductions, and error bars indicate 95% confidence interval of the average induction.

Almost identical regulation profiles in human and rat suggest evolutionary conservation

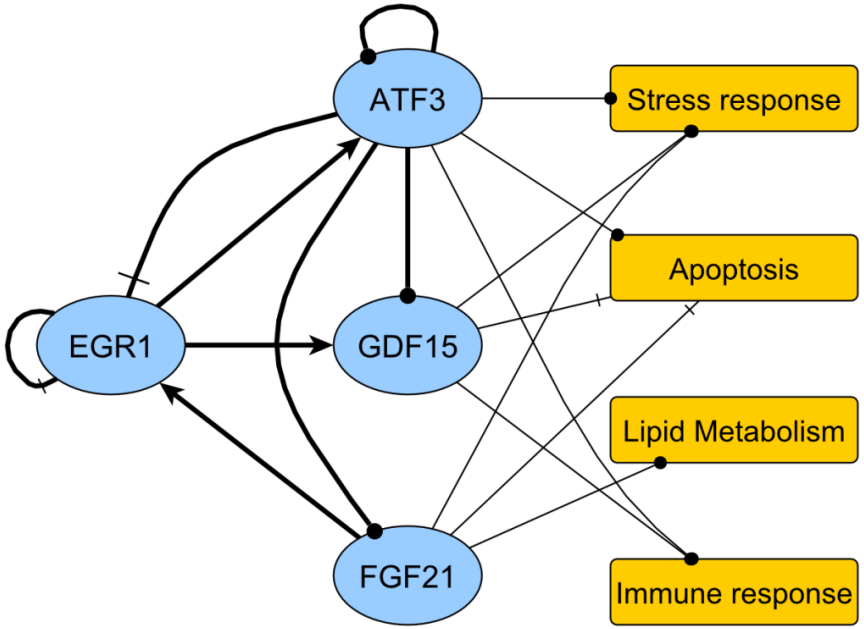
The genes form a functional network of early stress response with conserved structure and conserved dynamics



The early-response signature network, with downstream effects described in literature and annotated in functional databases

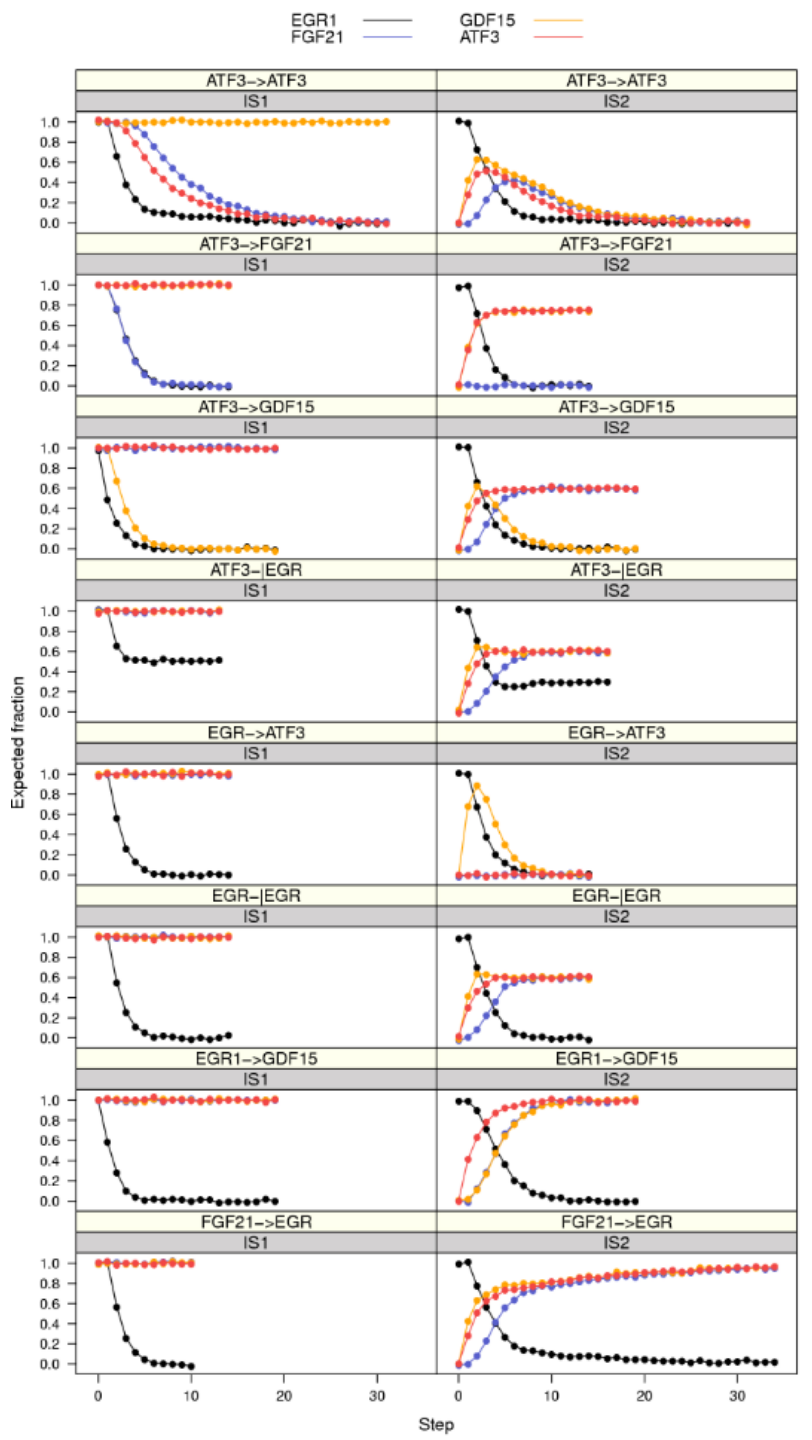
Literature search and functional annotation helped us realize the genes form a functional network

Boolean network modelling



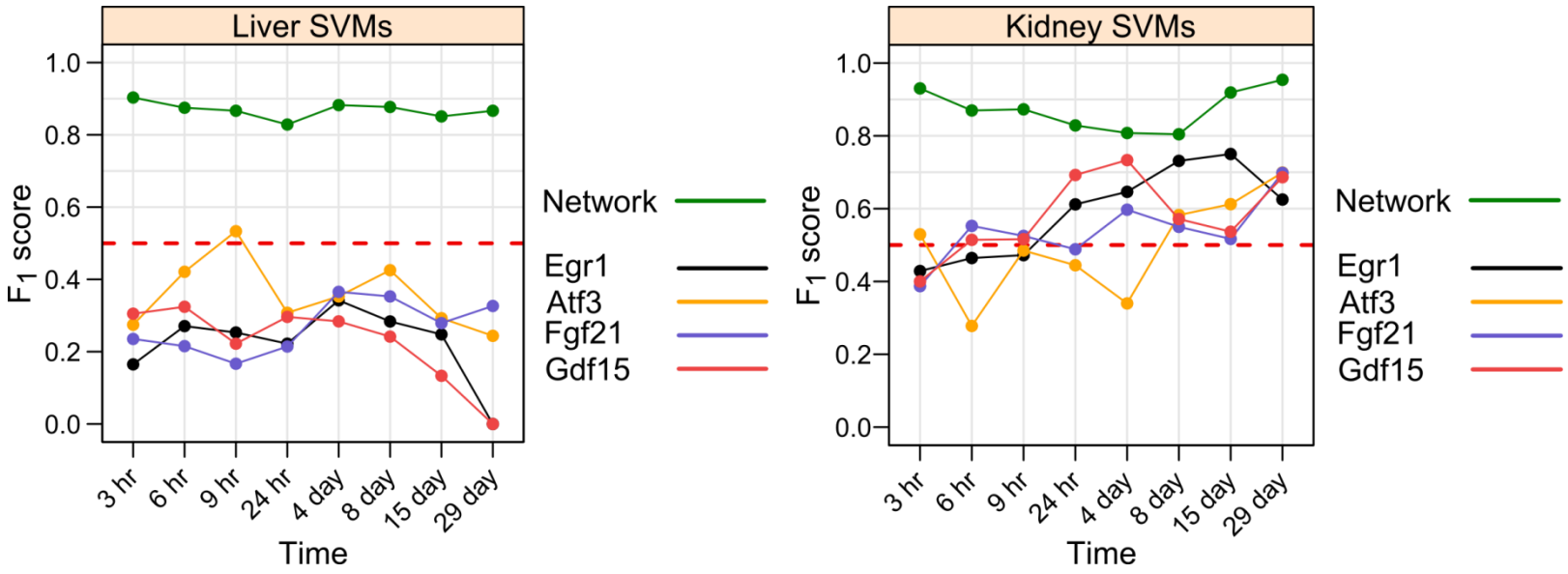
- Boolean network simulation results (Nikolaos Berntenis and Martin Ebeling, BMC Genomics 2013*)* support the hypothesis that **the conserved dynamics of the network in human and rat is encoded in the conserved structure of the network**.
- Permutation results suggest that **ATF3 is an important node to maintain the network dynamics**.

Boolean network modelling revealed that the dynamics is intrinsic to the network



The network finding was translated from *in vitro* to *in vivo*, and from liver to kidney

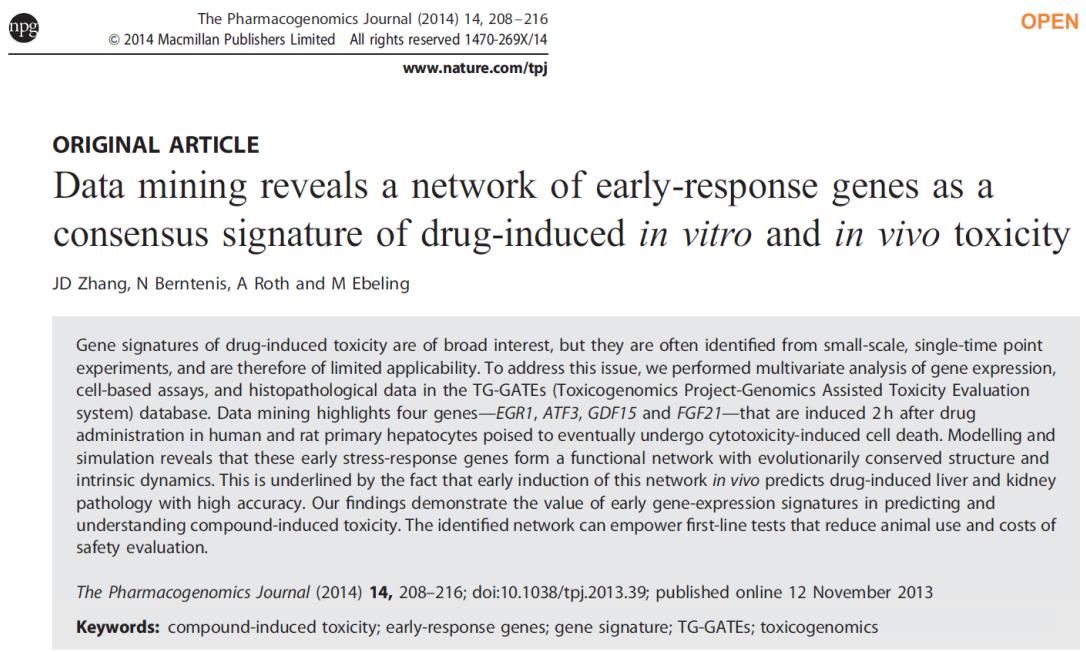
- Support Vector Machines (SVMs) were trained to predict *in vivo* pathology between 3h and 29d using gene expression changes of Egr1, Atf3, Gdf15, and Fgf21 at 3h.
- Profiles were randomly split into training samples (80%) and test samples (20%).
- SVMs are trained by 10-fold cross-validation in training samples. Then they are tested on test samples, which mimic new, unseen data.



The predictive power of the network is translated from *in vitro* to *in vivo,* and from liver to kidney

Summary of the work with TG-GATEs

- A novel computational pipeline identified four genes – EGR1, ATF3, GDF15, and FGF21 – that are induced as early as 2h after drug administration in human and rat primary hepatocytes poised to eventually undergo cell death.
- Boolean network simulation reveals that the genes form a functional network with evolutionarily conserved structure and dynamics.
- Confirming *in vitro* findings, early induction of the network predicts drug-induced liver and kidney pathology *in vivo* with high accuracy.
- The findings are not only useful for safety assessment, but also inspired the molecular-phenotyping platform.



The Pharmacogenomics Journal (2014) 14, 208–216

www.nature.com/tpj

OPEN

ORIGINAL ARTICLE

Data mining reveals a network of early-response genes as a consensus signature of drug-induced *in vitro* and *in vivo* toxicity

JD Zhang, N Berntenis, A Roth and M Ebeling

Gene signatures of drug-induced toxicity are of broad interest, but they are often identified from small-scale, single-time point experiments, and are therefore of limited applicability. To address this issue, we performed multivariate analysis of gene expression, cell-based assays, and histopathological data in the TG-GATEs (Toxicogenomics Project-Genomics Assisted Toxicity Evaluation system) database. Data mining highlights four genes—*EGR1*, *ATF3*, *GDF15* and *FGF21*—that are induced 2 h after drug administration in human and rat primary hepatocytes poised to eventually undergo cytotoxicity-induced cell death. Modelling and simulation reveals that these early stress-response genes form a functional network with evolutionarily conserved structure and intrinsic dynamics. This is underlined by the fact that early induction of this network *in vivo* predicts drug-induced liver and kidney pathology with high accuracy. Our findings demonstrate the value of early gene-expression signatures in predicting and understanding compound-induced toxicity. The identified network can empower first-line tests that reduce animal use and costs of safety evaluation.

The Pharmacogenomics Journal (2014) **14,** 208–216; doi:10.1038/tpj.2013.39; published online 12 November 2013

Keywords: compound-induced toxicity; early-response genes; gene signature; TG-GATEs; toxicogenomics

Zhang *et al.*, J Pharmacogenomics, 2014

Computational biology and bioinformatics help identifying safer drugs

Looking around and looking forward

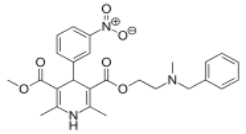
- **Selected further work by external groups**
  - Sutherland *et al.* (Lily), PLOS Comp Biol 2016, confirmed the difficulty to directly translate between different systems
  - El-Hachem *et al.* (U Montreal), Environ Health Perspect 2016, confirmed that early toxicological response occurring in animals is recapitulated in human and rat primary hepatocyte cultures at the molecular level
  - Thiel *et al.*, (RWTH Aachen), PLOS Comp Biol 2017, used physiologically-based pharmacokinetic modeling to characterize the transition from efficacious to toxic doses.
  - Shimada & Mitchison (Harvard), Mol Sys Bio 2019, used machine learning to characterize system-level response to drugs and toxicants in TG-GATEs, and pinpointing underlying molecular mechanisms.
- **What we are doing**
  - Gain molecule-level understanding of drug-induced histopathology
  - Apply the knowledge to accelerate development and reduce attrition rate of new drugs
  - Leveraging stem-cell technology and omics for drug discovery & personalized safety

The four-gene network is not the end, but a start, for the community and for us

Molecular Phenotyping

A workflow to quantify expression of pre-defined pathway reporter genes at early time points after perturbation to infer pathway activities, which may predict late-onset cellular phenotypes

Small molecule



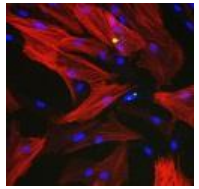
Antibodies

Antisense oligos

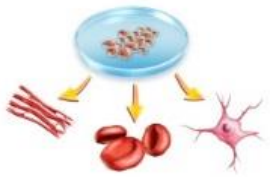
Therapeutic candidates

Early time point (3-12h)

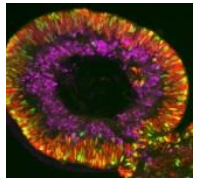
Human *in vitro* disease models



Cell lines/ primary cells

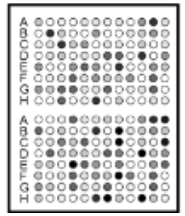


iPS-derived cells (opt. genome editing)



Advanced models

~1000 pathway reporter genes



Next-generation sequencing

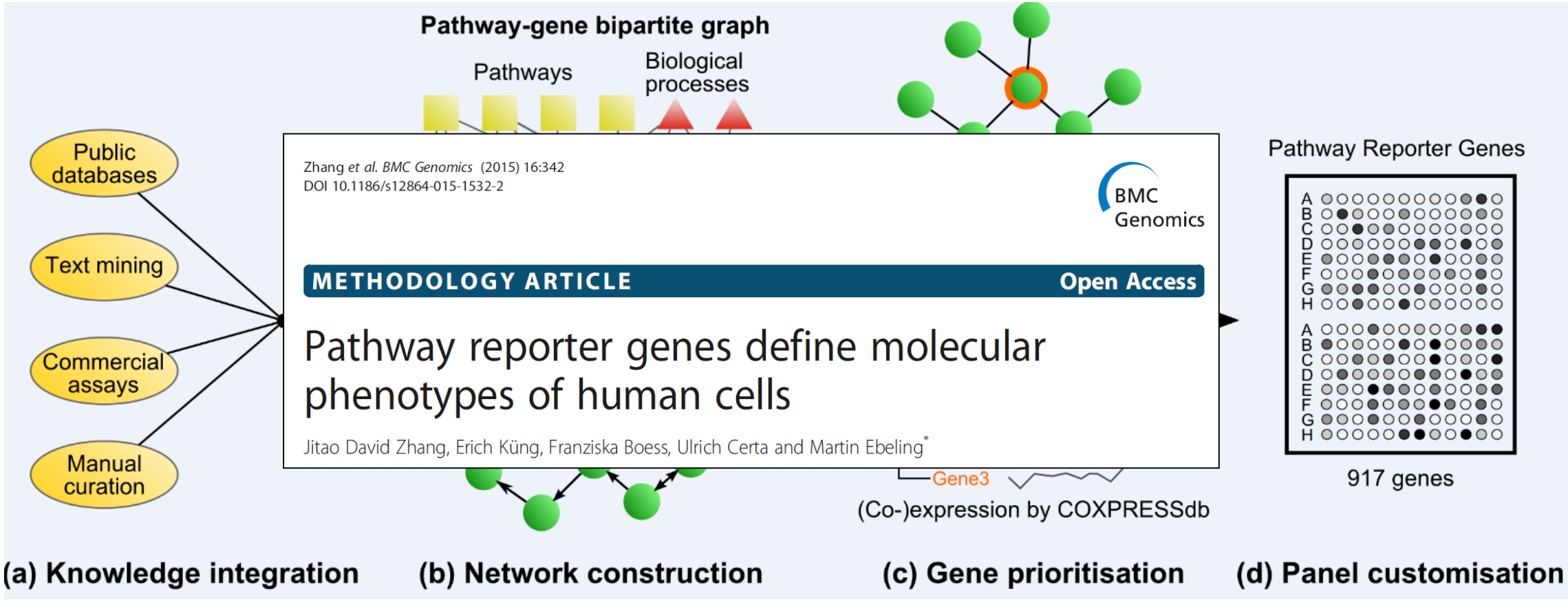
Molecular phenotyping

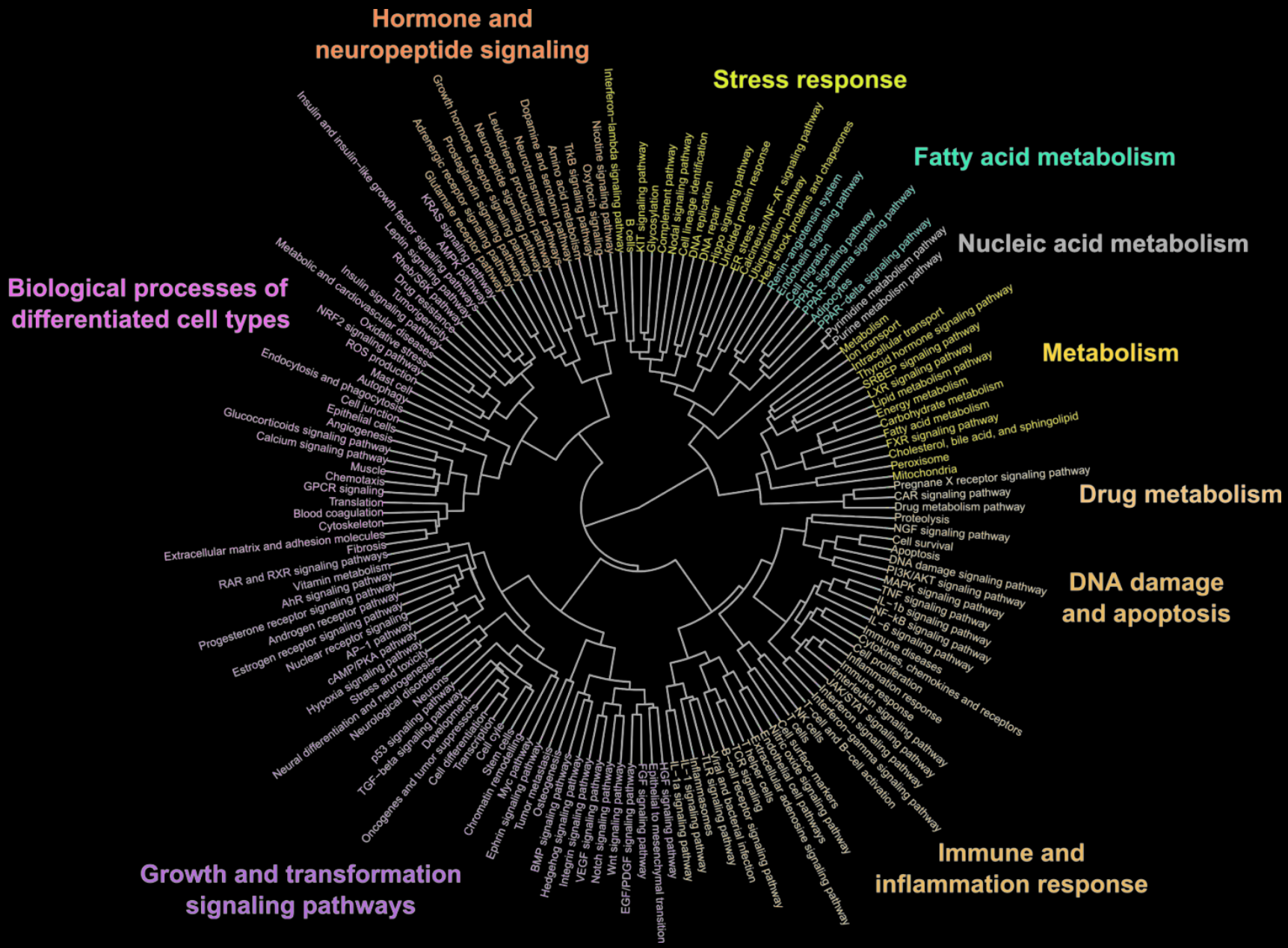
What pathways are perturbed by each compound?

Applications as screening tool:

- **Cluster compounds** based on pathway profiles
- **Detect false-positive hits** in a phenotypic screen
- **Correlate** pathway activity with phenotypic readouts

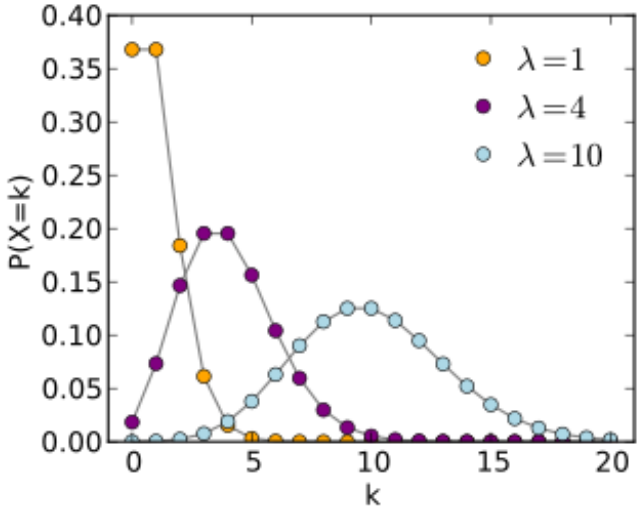
Pathway Reporter Genes




Hormone and neuropeptide signaling
Stress response
Fatty acid metabolism
Nucleic acid metabolism
Metabolism
Drug metabolism
DNA damage and apoptosis
Immune and inflammation response
Growth and transformation signaling pathways
Biological processes of differentiated cell types
Interferon−lambda signaling pathway
B cells
KIT signaling pathway
Glycosylation
Complement pathway
Nodal signaling pathway
Cell lineage identification
DNA replication
DNA repair
Hippo signaling pathway
ER stress
Unfolded protein response
Calcineurin/NF−AT signaling pathway
Ubiquitination pathway
Heat shock proteins and chaperones
Renin−angiotensin system
Endothelin signaling pathway
Cell migration
PPAR signaling pathway
Adipocytes
PPAR−gamma signaling pathway
PPAR−delta signaling pathway
Pyrimidine metabolism pathway
Purine metabolism pathway
Metabolism
Ion transport
Intracellular transport
Thyroid hormone signaling pathway
SREBP signaling pathway
LXR signaling pathway
Lipid metabolism pathway
Energy metabolism
Carbohydrate metabolism
Fatty acid metabolism
FXR signaling pathway
Cholesterol, bile acid, and sphingolipid
Peroxisome
Mitochondria
Pregnane X receptor signaling pathway
CAR signaling pathway
Drug metabolism pathway
Proteolysis
NGF signaling pathway
Cell survival
Apoptosis
DNA damage signaling pathway
PI3K/AKT signaling pathway
MAPK signaling pathway
TNF signaling pathway
IL−1b signaling pathway
NF−kB signaling pathway
IL−6 signaling pathway
Immune diseases
Cytokines, chemokines and receptors
Cell proliferation
Inflammation response
Interleukin signaling pathway
JAK/STAT signaling pathway
Interferon−gamma signaling pathway
Interferon and B−cell activation
NK cells
T cells
Cell surface markers
Nitric oxide signaling pathway
Endothelial cell pathways
Extracellular adenosine signaling pathway
T helper cells
TCR signaling
B−cell receptor signaling pathway
Viral and bacterial infection
TLR signaling pathway
Inflammasomes
IL−1 signaling pathway
IL−1a signaling pathway
HGF signaling pathway
Epithelial to mesenchymal transition
FGF signaling pathway
EGF/PDGF signaling pathway
Wnt signaling pathway
Notch signaling pathway
VEGF signaling pathway
Integrin signaling pathway
Hedgehog signaling pathway
BMP signaling pathways
Osteogenesis
Tumor metastasis
Myc pathway
Ephrin signaling pathway
Chromatin remodeling
Stem cells
Cell Cycle
Transcription
Cell differentiation
Oncogenes and tumor suppressors
TGF−beta signaling pathway
p53 signaling pathway
Development
Neurons
Neural differentiation and neurogenesis
Neurological disorders
Stress and toxicity
Hypoxia signaling pathway
cAMP/PKA pathway
AP−1 pathway
Nuclear receptor signaling
Estrogen receptor signaling pathway
Androgen receptor pathway
Progesterone receptor signaling pathway
AhR signaling pathway
Vitamin metabolism
RAR and RXR signaling pathways
Fibrosis
Extracellular matrix and adhesion molecules
Cytoskeleton
Blood coagulation
Translation
GPCR signaling
Chemotaxis
Muscle
Calcium signaling pathway
Glucocorticoids signaling pathway
Angiogenesis
Epithelial cells
Cell junction
Autophagy
Mast cell
Endocytosis and phagocytosis
ROS production
NRF2 signaling pathway
Oxidative stress
Tumorigenicity
Drug resistance
Metabolic and cardiovascular diseases
Insulin signaling pathway
RhebiSGK pathway
Leptin signaling pathway
AMPK pathway
KRAS signaling pathway
Glutamate receptor signaling
Adrenergic receptor signaling pathway
Prostaglandin signaling pathway
Growth hormone receptor signaling pathway
Neuropeptide signaling pathway
Leukotrienes production pathway
Neurotransmitter pathways
Dopamine and serotonin pathways
Amino acid metabolism
TrkB signaling pathway
Oxytocin signaling
Nicotine signaling pathway
Insulin and insulin−like growth factor signaling pathway

Difference in statistical modelling of microarray data and next-generation sequencing count data

- Microarray data: log-normal distributed, for instance implemented in the *limma* package of R/Bioconductor.
- NGS data: Negative-Binomial distributed (or Poisson with overdispersion), for instance implemented in both *edgeR* and *DESeq2* package of R/Bioconductor.



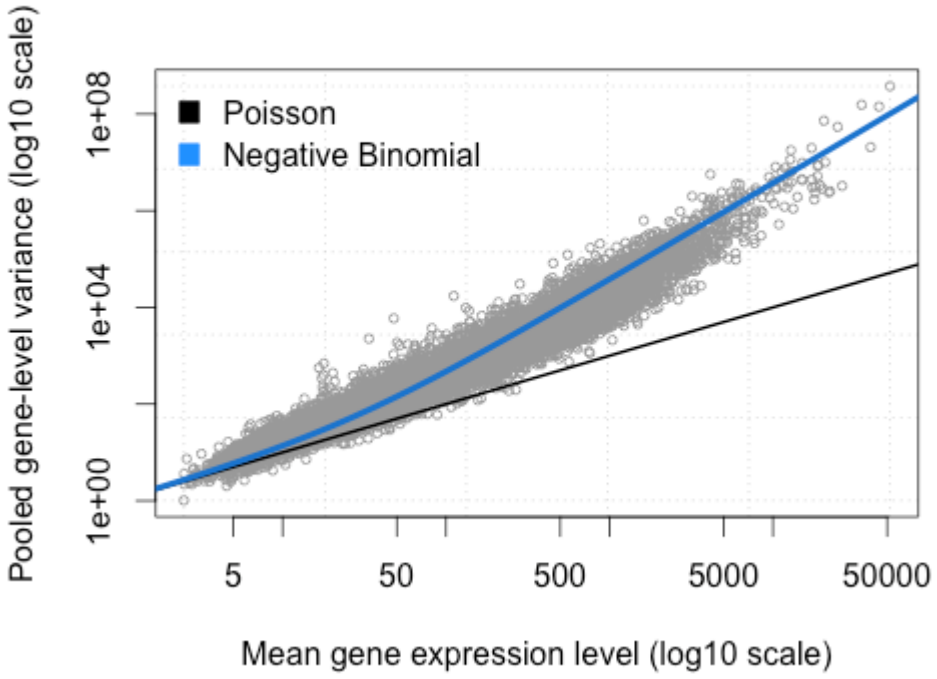
Poisson distribution with three rate parameters, from Wikimedia, reused with the CC Attribution 3.0 license

From Poisson distribution to Negative Binomial Distribution

Two definitions of Negative-Binomial distribution

1. The number of failures seen before getting *n* successes (the inverse of *Binomial Distribution*, which the number of successes in *n* independent trials)

2. Poisson-Gamma mixture distribution, weighted mixture of *Poisson* distributions, where the rate parameter has an uncertainty modelled by a *Gamma* distribution.



Credit of Jesse Lipp, bioramble.wordpress.com

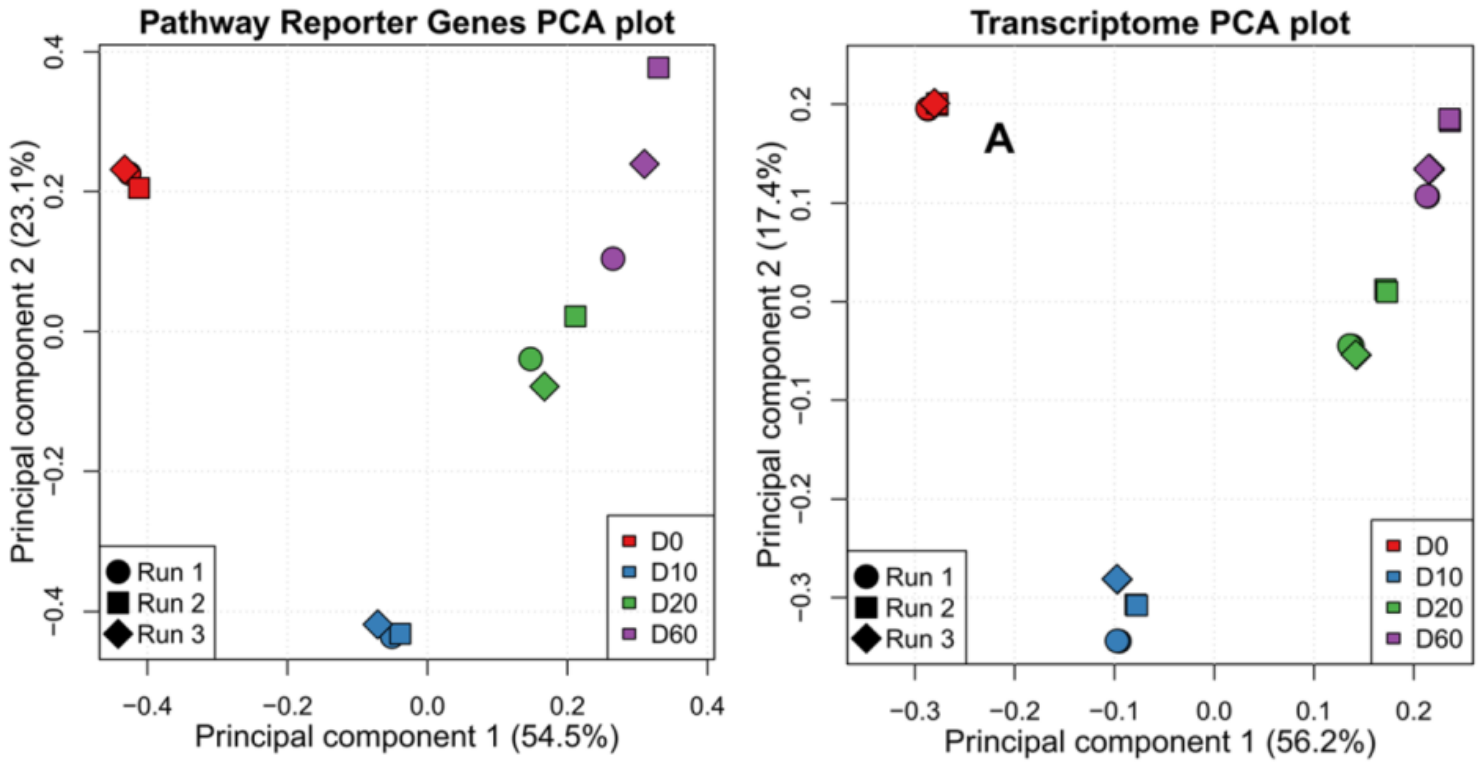
Commonly used dimensionality reduction techniques

- Principal component analysis (PCA)
- t-SNE (t-distributed Stochastic Neighbor Embedding)
- UMAP (Uniform Manifold Approximation and Projection) [A great talk by Leland McInnes, the developer of UMAP, a mathematician, Ph.D. In Profinite Lie Rings]

- For a recent overview of dimensionality reduction techniques and their applications in biology, see Nguyen, Lan Huong, und Susan Holmes. „Ten Quick Tips for Effective Dimensionality Reduction“. *PLOS Computational Biology* 15, Nr. 6 (20. Juni 2019): e1006907. https://doi.org/10.1371/journal.pcbi.1006907.

Data Analysis

1. Global gene expression profiles
2. Differential gene expression
3. Pathway analysis
4. Integrative analysis with other data



Pathway reporter genes faithfully capture global gene expression patterns

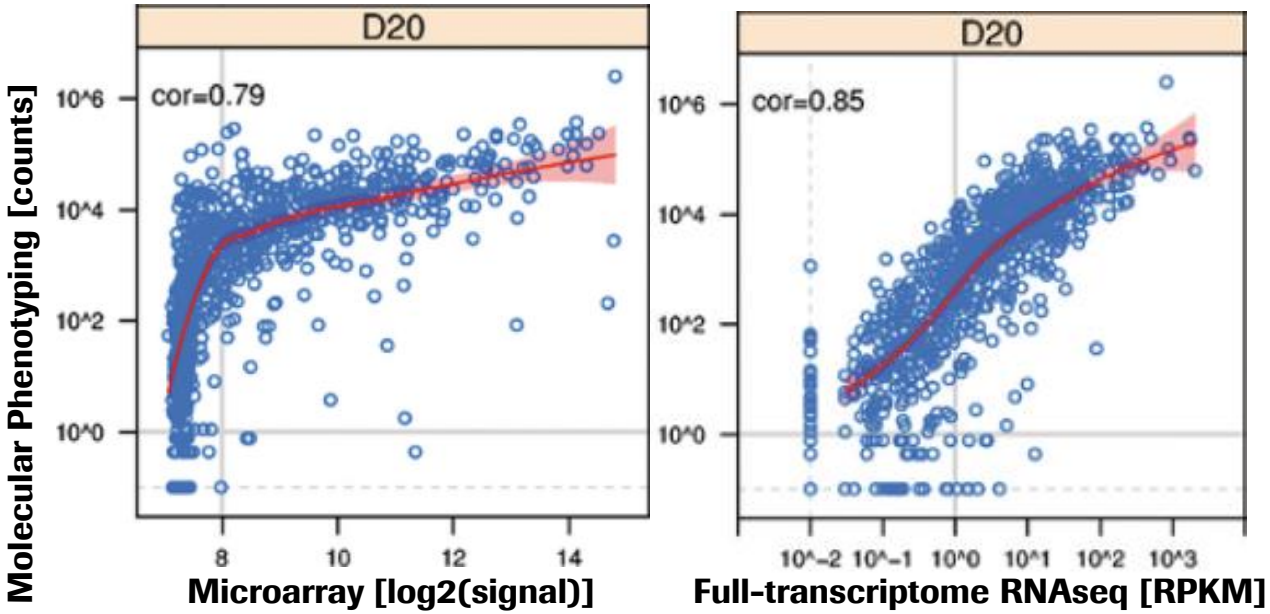
Data Analysis

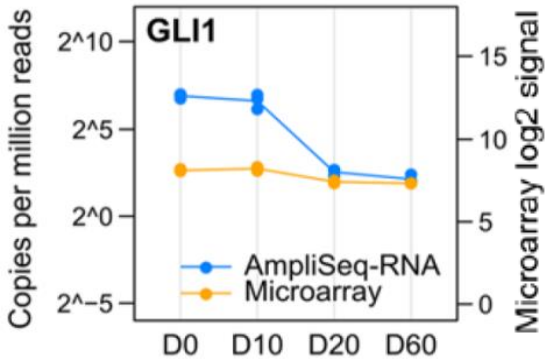
1. Global gene expression profiles

2. Differential gene expression

3. Pathway analysis

4. Integrative analysis with other data





AmpliSeq, like RNA-seq, shows higher dynamic range than hybridization-based platforms

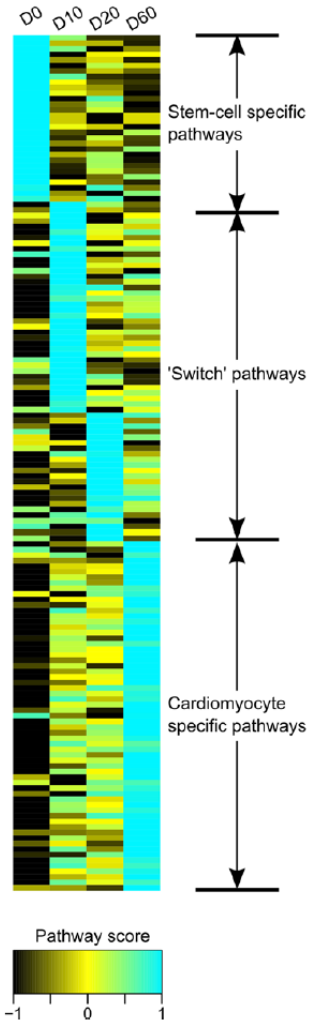
Data Analysis

1. Global gene expression profiles

2. Differential gene expression

3. Pathway analysis

4. Integrative analysis with other data



Activity patterns of 154 human metabolic and signaling networks during differentiation of induced pluripotent stem-cells (iPS) into cardiomyocytes.

Cyan: Pathway is activated

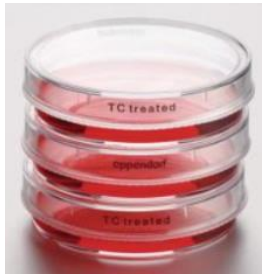
Black: Pathway is suppressed

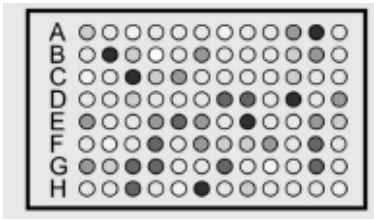
Pathway reporter genes inform about pathway activity patterns

Data Analysis

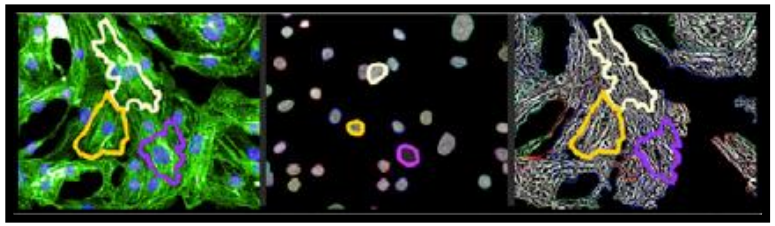
1. Global gene expression profiles
2. Differential gene expression
3. Pathway analysis
4. Integrative analysis with other data

In vitro assay readouts

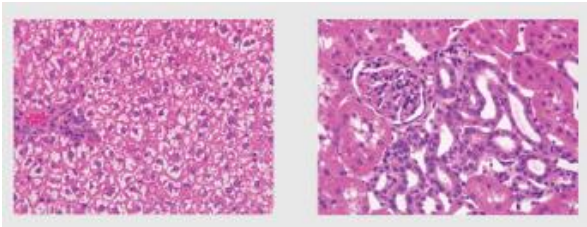




High-content microscopy



Histopathology records

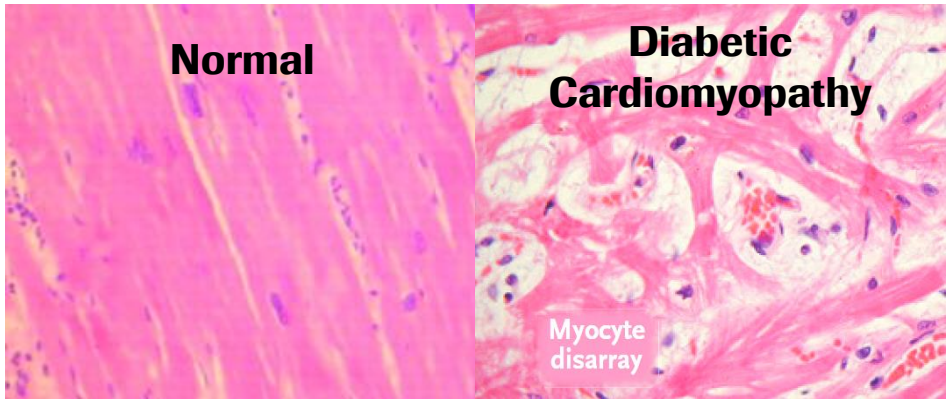


Gene signatures

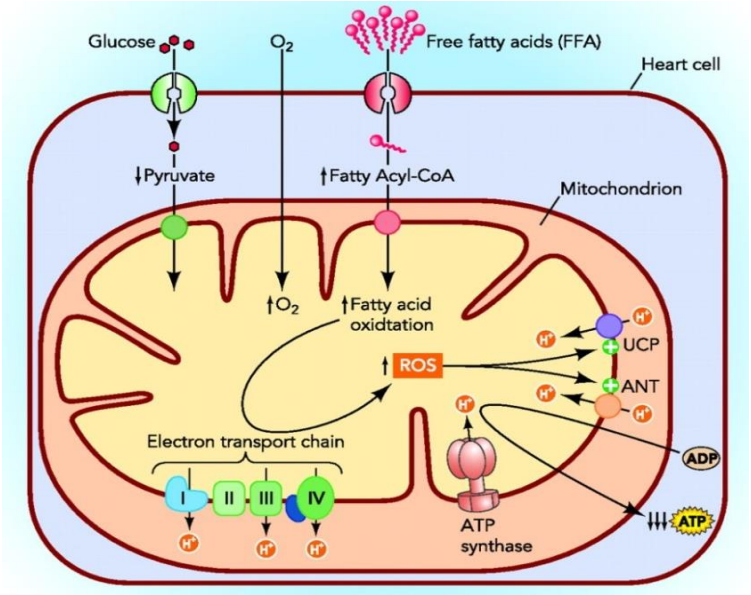
Molecular phenotyping allows data integration

Cell-based model of diabetic cardiomyopathy for phenotypic screening

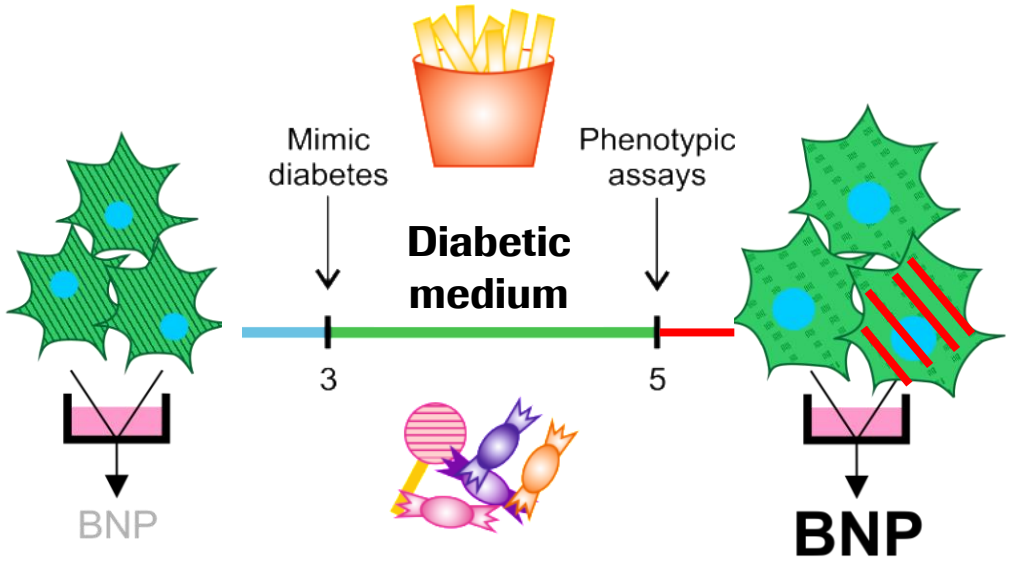
Metabolic dysfunction promotes cardiomyopathy



Diabetic cardiomyocyte metabolism

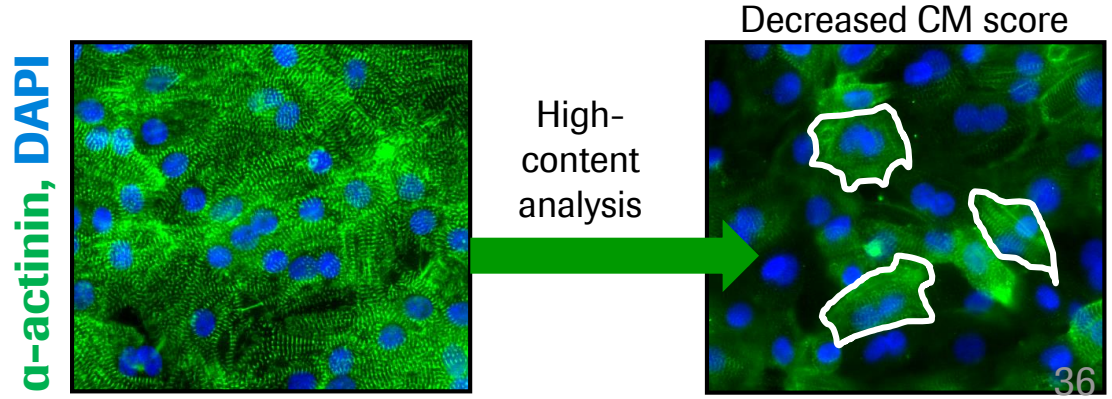


iPS-derived cardiomyocyte model



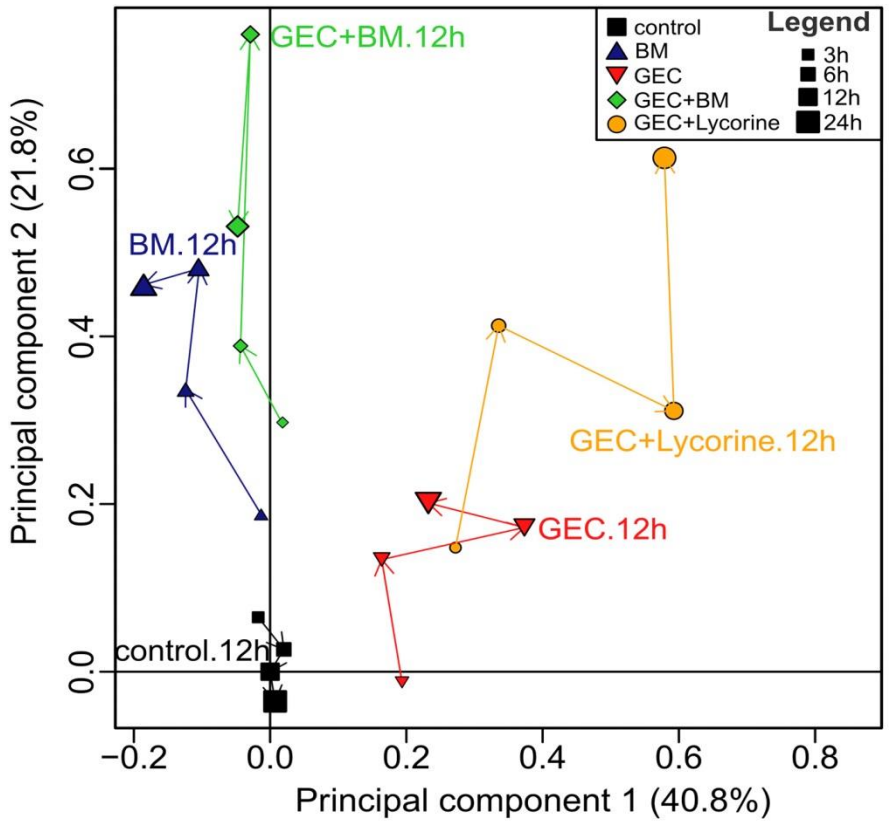
Glucose, Insulin, Fatty Acids

Cortisol, Endothelin-1



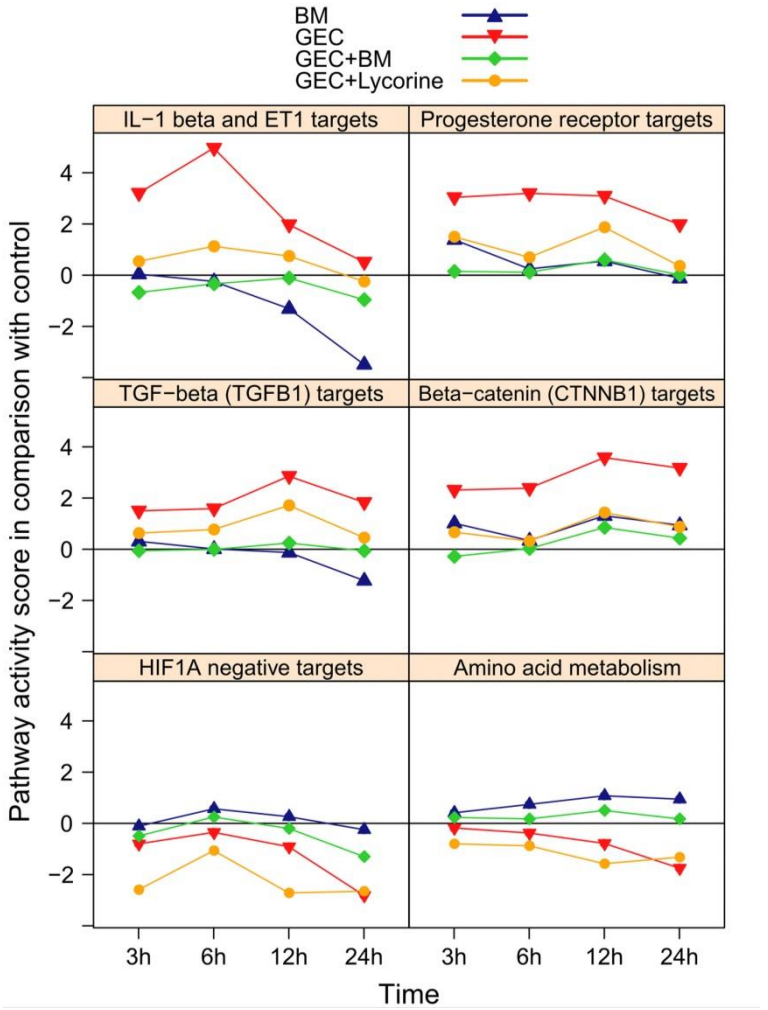
Determining the optimal timepoint for molecular phenotyping

Maximal dynamic range at 12 hours



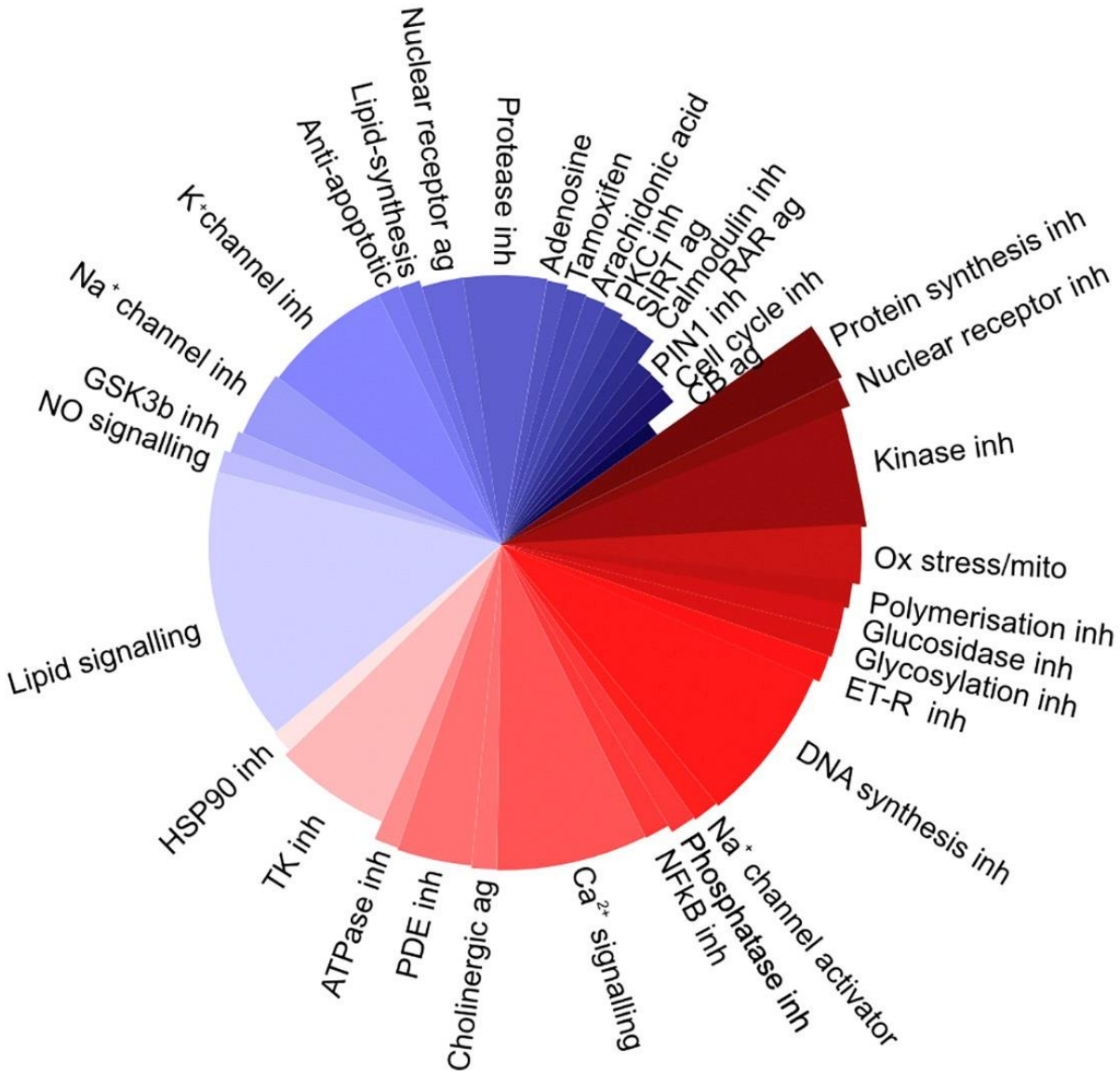
Phenotypic screening performed after 48 hours

Lycorine and positive control manipulate similar pathways

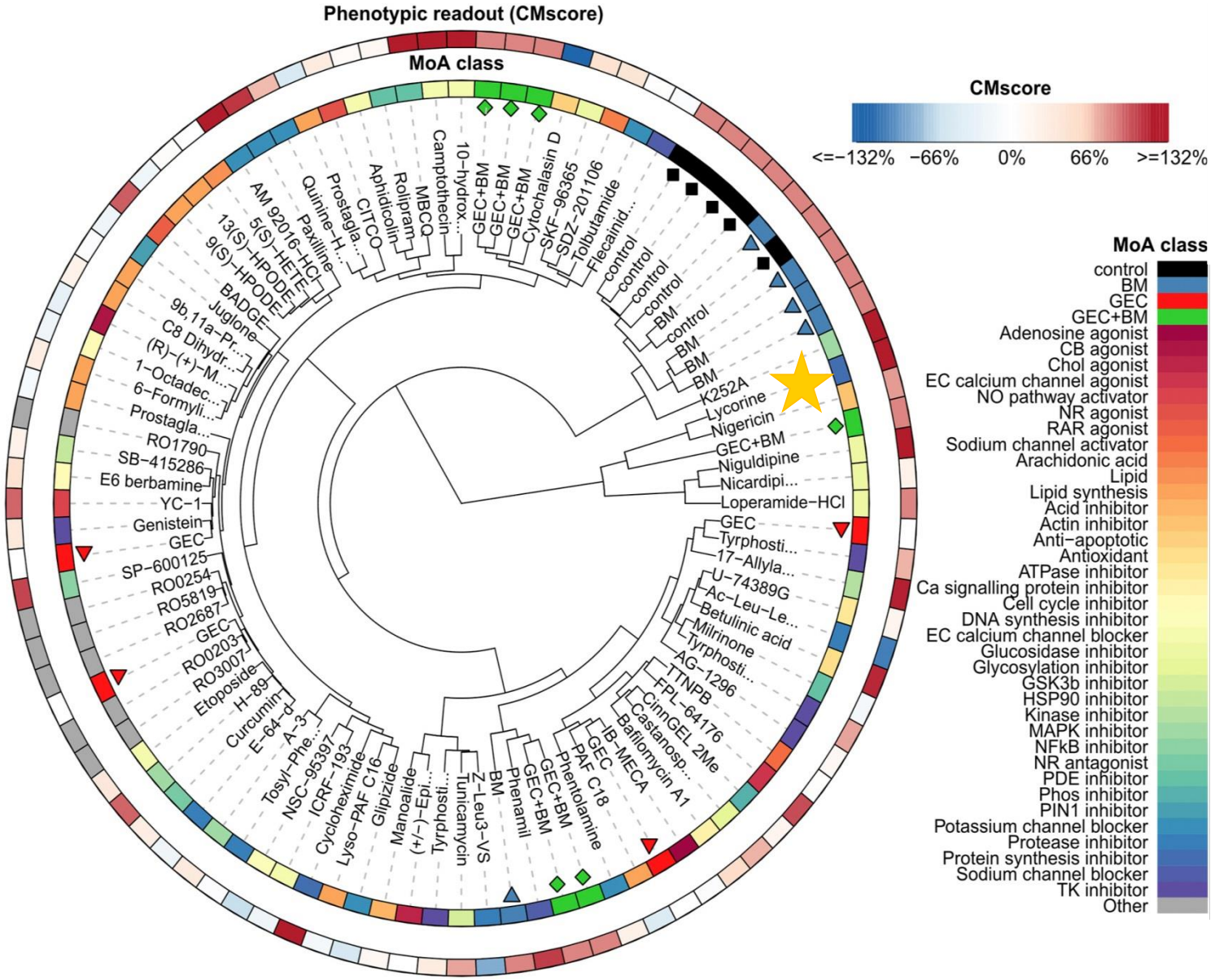


Mechanistic enrichment of screening data

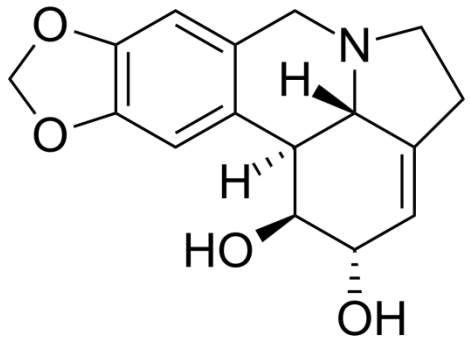
Integration of molecular and phenotypic information



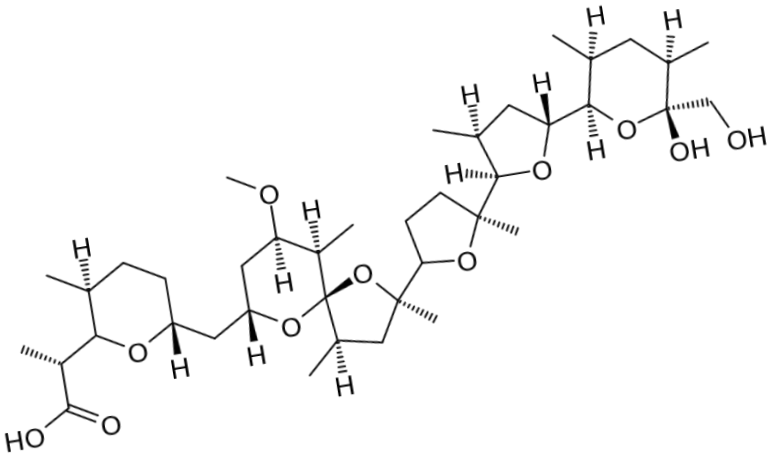
Integration of molecular and phenotypic information



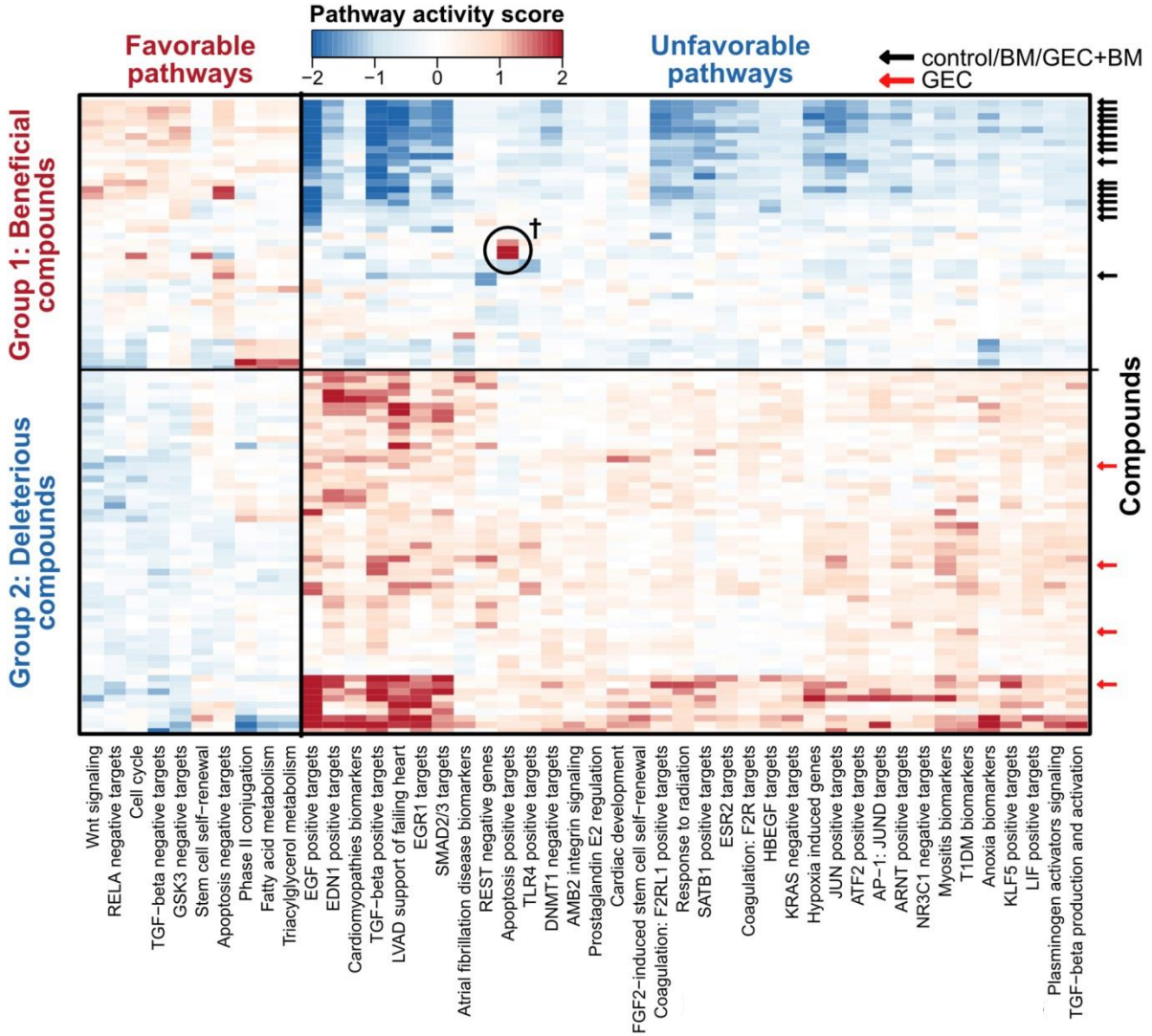
Lycorine: Protein synthesis inhibitor



Nigericin: Potassium ionophore



Hierarchical clustering separates compounds and pathway responses



Pathways associated with CM score

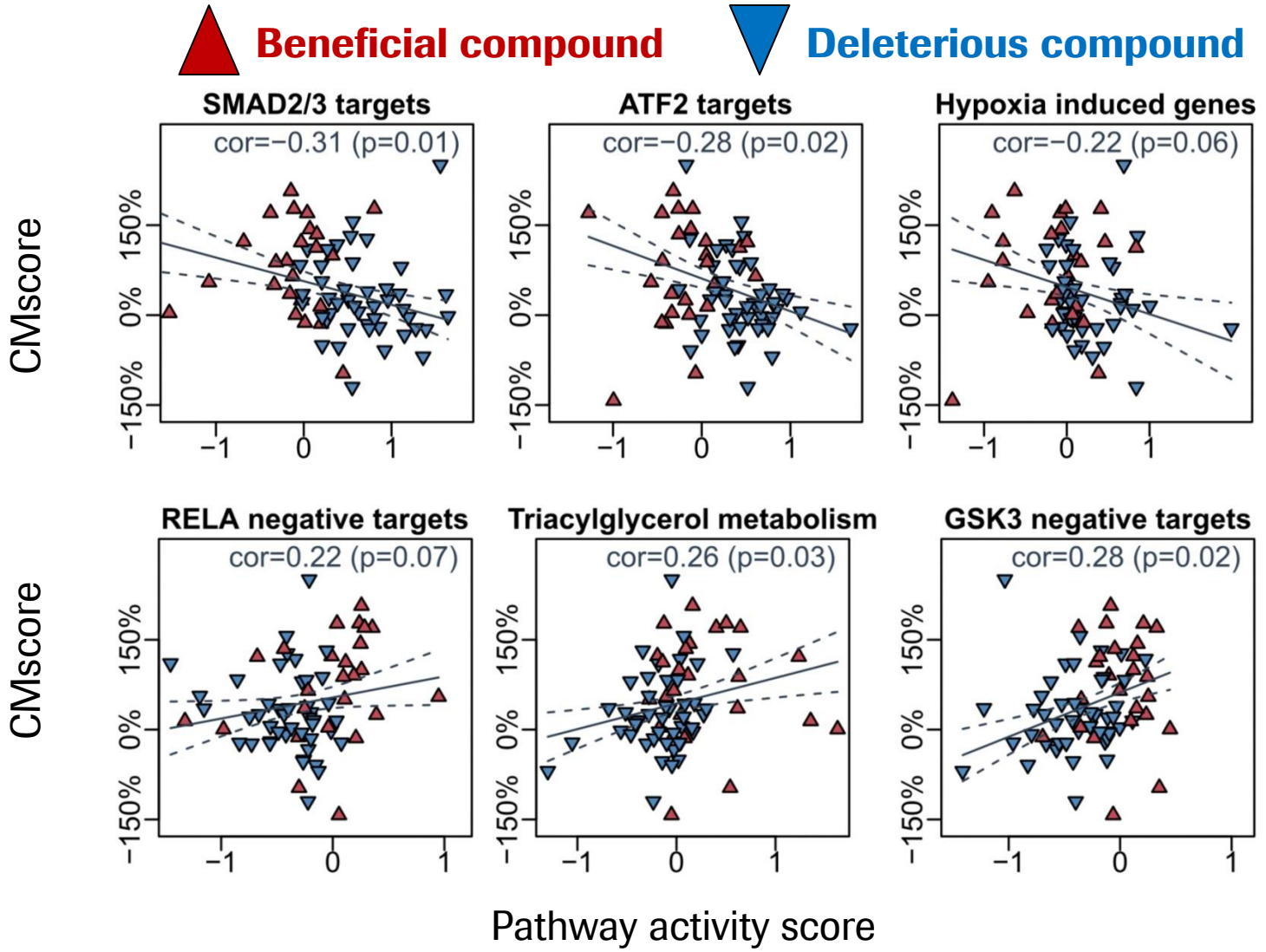
Beneficial compounds regulate **favorable** pathways positively and **unfavorable** pathways negatively

Beneficial compounds have **higher** CMscore

Deleterious compounds regulate **unfavorable** pathways positively and **favorable** pathways negatively

Deleterious compounds have **lower** CMscore

Beneficial compounds generate specific pathway signatures



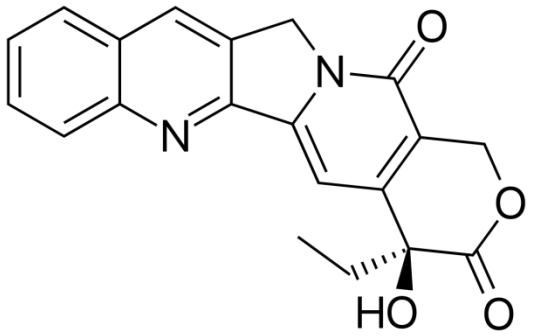
Beneficial compounds negatively regulate

Beneficial compounds positively regulate

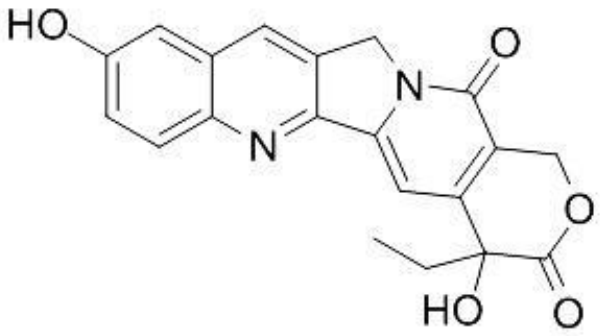
Pathway signatures can be monitored during screening campaigns for maintained beneficial mechanistic effects

Molecular phenotyping allows filtering of undesirable molecules

Camptothecin

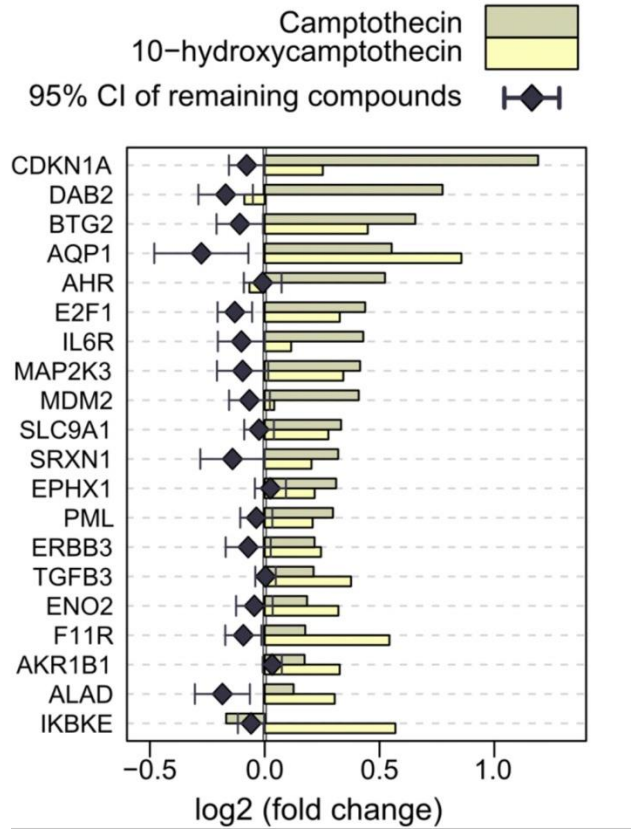


10-hydroxycamptothecin



- Topoisomerase inhibitors
- Produced high CMscore in the phenotypic assay
- Identified as 'hits'
- Cluster with beneficial compounds

Induce target genes of apoptosis



Compounds with undesirable signatures can be eliminated from further testing

Beneficial compound signatures are downregulated in cardiomyopathy samples

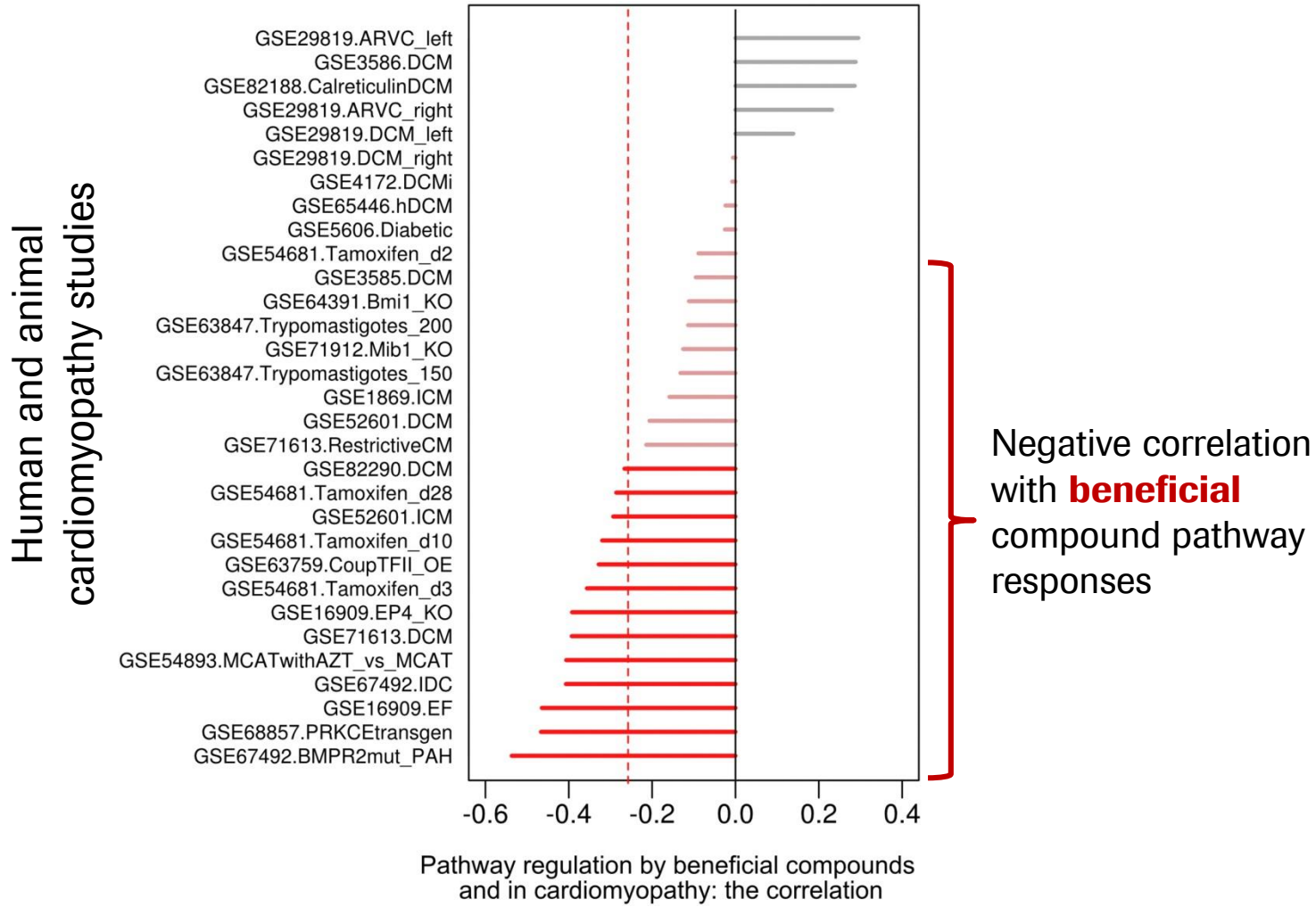
Data Mining of GEO

20 datasets from human and murine cardiomyopathy

Generate **surrogate molecular phenotype** by analyzing panel expression

Pathway analysis of surrogate MP data

Correlation with beneficial and deleterious compound pathway responses



Beneficial compound signatures are downregulated in cardiomyopathy samples

Data Mining of GEO

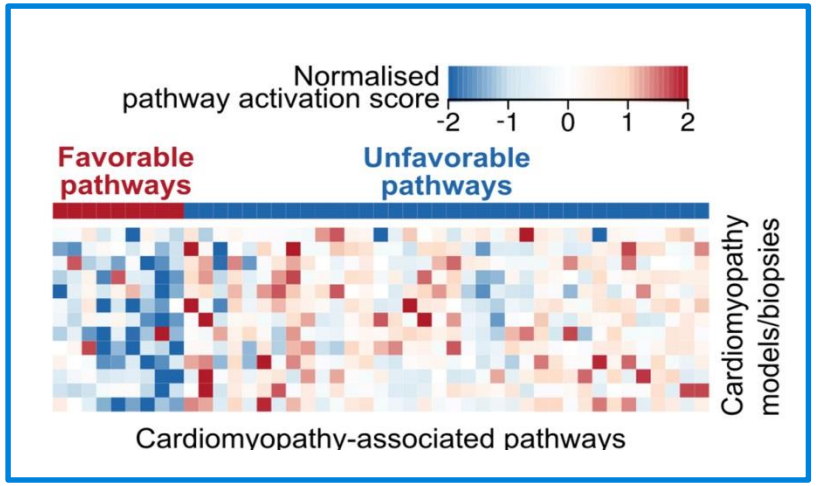
20 datasets from human and murine cardiomyopathy

Generate **surrogate molecular phenotype** by analyzing panel expression

Pathway analysis of surrogate MP data

Correlation with beneficial and deleterious compound pathway responses

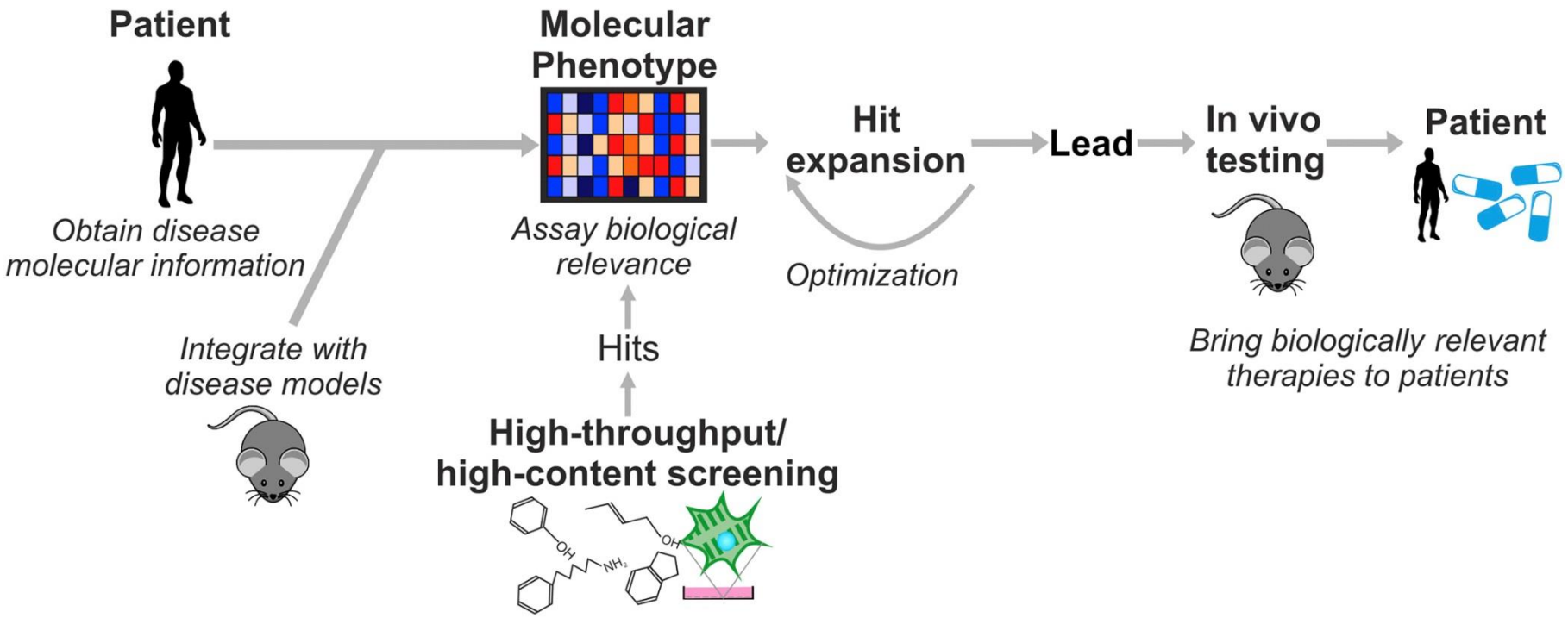
Favorable pathways are downregulated



Unfavorable pathways are enriched

Molecular phenotyping can enrich screening campaigns to select compounds with profiles with biological relevance to patients

Molecular Phenotyping can enrich Phenotypic Drug Discovery



1. MP provides mechanistic validation of hits in successive screening campaigns
2. MP enables undesirable and false-positive hits to be eliminated
3. MP brings biological relevance to screening assays by integrating patient information

Summary

- **Gene expression profiling: a case study of omics and cellular modelling**
- **Applications for drug safety: TG-GATEs**
- **Applications for drug mechanism: molecular phenotyping**

- **Current research topics**
  - Single-cell sequencing
  - Genome editting
  - Microbiome
  - High-content cellular imaging
  - Integrative modelling