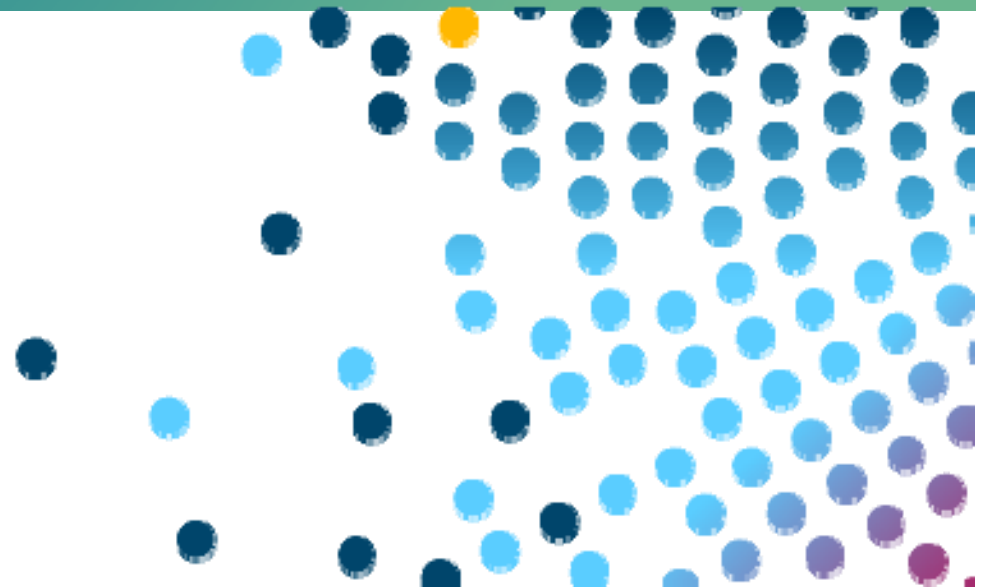


Gene expression connectivity mapping and its application to Cat-App

Shu-Dong Zhang

Northern Ireland Centre for Stratified Medicine

University of Ulster



Outline

- Gene expression connectivity mapping
- Initial success and further development
- Typical applications
- Application to Cat-App
- Discussion



Cat-App Workprogramme

CAT-APP: New technologies to underpin the category approaches and read across in regulatory programmes

Project Management: Hans Ketelslegers / Concawe

Steering: Scientific Committee / Concawe

WP1

Organization of data available on PS (Ivan Rusyn/AgriLife)

- 1.1 Obtain, process and share chemical samples
- 1.2 Collect available records (manufacturing process info., phys./chem. properties, analytical chemistry, existing toxicity data on mammalian, ecotox)
- 1.3 Digitize records into flexible and inter-operable database format

WP2

Toxicity screening (Ivan Rusyn/AgriLife)

WP2.a (Ivan Rusyn/AgriLife)

- High content screening of iPS-derived cells
- hepatocytes, neurons, cardiomyocytes, myo-fibroblasts, endothelial

WP2.b (Tim Gant/PHE)

- Toxicity phenotyping in 10 diverse cell lines (overlapping with LINCS)

WP3

High throughput genomics (Ivan Rusyn/AgriLife)

- 3.1 High-throughput transcriptomics profiling of 13,500 samples for TempO-seq

WP4

Perform data integration and chemical biological read across (Fred Wright/NCSU)

WP 4.a

(Fred Wright/NCSU)

- 4a.1 Coordinate data management and workflow
- 4a.2 Perform uncertainty and variability analyses
- 4a.3 Process and analyse omics data
- 4a.4 Perform ToxPi analysis

WP4.b (Shu-Dong Zhang/ Ulster)

- 4b.1 Perform connectivity mapping
- 4b.2 Develop and apply analysis algorithms to robustness testing, investigate grouping accuracy and profiling cost

WP5

Dissemination, project administration and Outreach (Klaus Lenz/SYNCOM)

- 5.1 Project Dissemination and website
- 5.2 Project Administration
- 5.3 Outreach

Advisory Board

George Daston
Procter & Gamble

Shirley Price
University of Surrey

Chris Rowat
Health Canada

Xiaowei Zhang
Nanjing University

Institute abbreviations:

AgriLife: Texas A&M AgriLife Research - NCSU: North Carolina State University - PHE: Public Health England
Ulster: Ulster University - SYNCOM: SYNCOM R&D consulting GmbH



Work package 4b Connectivity mapping analysis

1. Similarity matrix of the ~ 160 Concawe UVCBs: establishing significant connections among the assayed UVCBs themselves.
2. Performing robustness analysis on the connections and groupings of Concawe substances obtained.

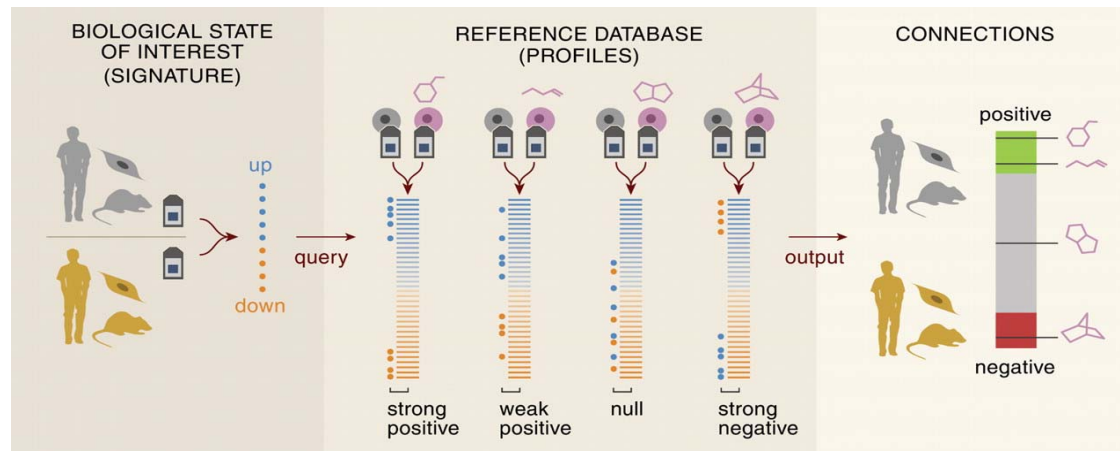




01

Introduction to gene
expression connectivity
mapping

The Lamb Connectivity Map



Key components

1. **Reference Profiles:** A set of gene expression profiles, obtained from systematic microarray gene expression profiling.
(MCF7 and other cell lines, 164 bioactive small-molecule compounds, 564 microarrays, 453 individual reference profiles)
2. **Gene signature:** a short list of important genes, selected by the researchers as a result of some experiments investigating a particular biological condition.
3. **Connection score,** defined as a function of a Reference Profile and a Gene Signature. It should reflect the underlying biological connection between them.

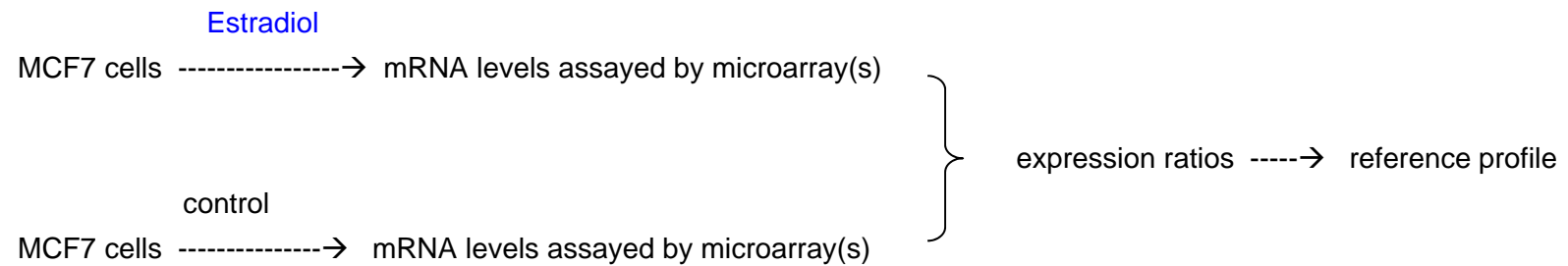
J Lamb et al.,
Science 313,
1929 -1935 (2006)



How to make a reference gene-expression profile

An example:

Cells: MCF7 (Breast cancer epithelial cell line)
Compound: Estradiol
Dose: 0.1 μ M
Duration: 6 hours



The context of Connectivity Map concept

- Using some form of molecular profile to characterize a biological state
- The basis of many practices in medicine
- Various biochemical tests measure the molecular profile of a patient
- To learn about the biological state of the patient (being with a particular disease or healthy).
- Connectivity Map concept is the generalization of molecular profile (bio-marker) measurement.
- The sheer number of different mRNAs measured enables a much richer and complex description of the biological state than a few marker molecules could achieve.
- The specification of the mRNA profiles as measured by high throughput technologies (microarrays, NGS) can provide an adequate description of the biological state.





02

Initial success and
further development

Example: HDAC inhibitors

Keith B. Glaser et al (2003), Molecular Cancer Therapeutics, Vol. 2, 151–163. Gene Expression Profiling of Multiple Histone Deacetylase (HDAC) Inhibitors

Three Cell lines:

T24 (bladder)

MDA435 (breast carcinoma)

MDA468 (breast carcinoma)

Treated with three HDAC inhibitors:

vorinostat

MS-27-275

trichostatin A

The result gene signature: (> 2 -fold, $p < 0.01$)


8 up-regulated genes

5 down-regulated genes

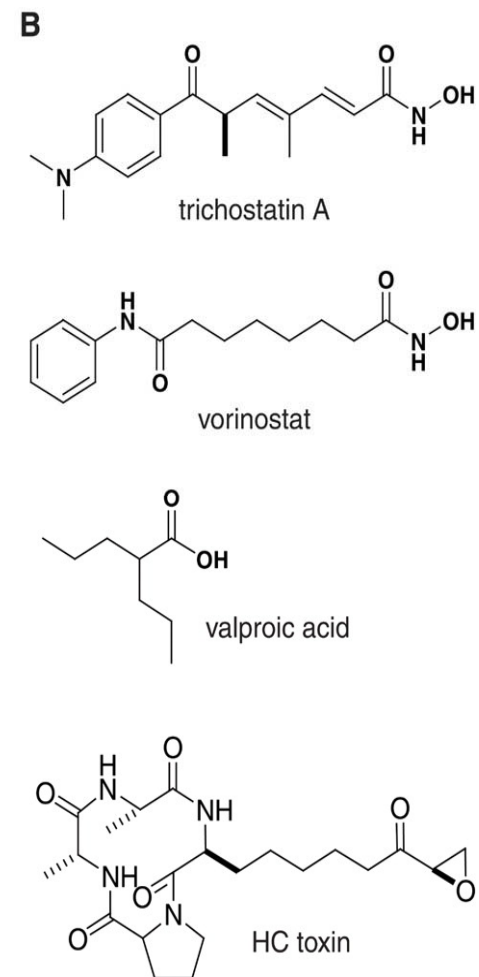
Summary of changes in gene expression for the core set of HDAC inhibitor regulated genes		
A. Summary of all cell lines		
GeneBank accession no.	Gene name	Gene function
Up-regulated		
U09579	<i>p21</i>	Cell cycle regulation
U31875	<i>Hep27</i>	Short-chain alcohol dehydrogenase
M29877	<i>α-Fucosidase</i>	Proteoglycan metabolism
M60750	<i>Histone H2B</i>	Nuclear organization
M63379	<i>TRPM-2</i>	Clusterin-apoptosis
X01703	<i>α-Tubulin</i>	Cytoskeletal structure
X76648	<i>Glutaredoxin</i>	Glutathione dep. DNA synthesis
X76717	<i>Metallothionein 1L</i>	Binds heavy metals
Down-regulated		
D00596	<i>Thymidylate synthetase</i>	DNA synthesis
L24521	<i>TRP</i>	Transformation related protein
L38951	<i>Importin β</i>	Nuclear translocation/shuttling
U70439	<i>APRIL</i>	Acidic protein rich in Leu
X52142	<i>CTP synthase</i>	DNA synthesis



A



rank	perturbagen	dose	cell	score
1	vorinostat [1000]	10 μ M	MCF7	1
2	trichostatin A [873]	1 μ M	MCF7	0.969
3	trichostatin A [992]	100 nM	MCF7	0.931
4	trichostatin A [1050]	100 nM	MCF7	0.929
5	vorinostat [1058]	10 μ M	MCF7	0.917
6	trichostatin A [981]	1 μ M	MCF7	0.915
7	HC toxin [909]	100 nM	MCF7	0.914
8	trichostatin A [1112]	100 nM	MCF7	0.908
9	trichostatin A [1072]	1 μ M	MCF7	0.906
10	trichostatin A [1014]	1 μ M	MCF7	0.893
11	trichostatin A [332]	100 nM	MCF7	0.882
12	trichostatin A [331]	100 nM	MCF7	0.846
13	trichostatin A [448]	100 nM	PC3	0.788
14	valproic acid [345]	10 mM	MCF7	0.743
15	valproic acid [23]	1 mM	MCF7	0.735
16	valproic acid [1047]	1 mM	MCF7	0.733
17	trichostatin A [413]	100 nM	ssMCF7	0.725
18	valproic acid [410]	10 mM	HL60	0.725
19	valproic acid [458]	1 mM	PC3	0.680
33	valproic acid [409]	1 mM	HL60	0.634
39	valproic acid [1020]	500 μ M	MCF7	0.619
52	valproic acid [346]	2 mM	MCF7	0.582
61	valproic acid [1078]	500 μ M	MCF7	0.563
71	valproic acid [629]	1 mM	SKMEL5	0.539
72	valproic acid [347]	500 μ M	MCF7	0.539
73	valproic acid [989]	1 mM	MCF7	0.538
76	valproic acid [433]	1 mM	PC3	0.528
89	trichostatin A [364]	100 nM	HL60	0.507
92	valproic acid [497]	1 mM	ssMCF7	0.501
297	valproic acid [348]	50 μ M	MCF7	0
388	valproic acid [994]	200 μ M	MCF7	0
403	valproic acid [1002]	50 μ M	MCF7	0
419	valproic acid [1060]	50 μ M	MCF7	-0.537



J Lamb et al., Science 313, 1929 -1935 (2006)



A simpler and more robust framework

Zhang & Gant's Improvements upon the Lamb Connectivity Map:

1. A simpler, unified framework for ranking genes
2. More-principled statistical testing procedures
3. Effective safeguard against false connections creeping in
4. Increased sensitivity in picking up real biological connections

Zhang & Gant 2008, BMC Bioinformatics 2008, 9:258.

Reference Profiles:

- ✓ To treat up- and down-regulated genes in a unified framework, on a equal footing basis.
- ✓ To give a signed rank to each gene, with the most differentially expressed genes (either up or down) receiving the highest ranks.

Connection strength:
$$C(\mathbf{R}, \mathbf{s}) = \sum_{i=1}^n R(g_i) s(g_i),$$

$$\text{Connection Score} = \frac{\text{Connection Strength}}{\text{Maximum possible connection strength}}$$

Statistical testing procedures:

The null hypothesis H_0 :

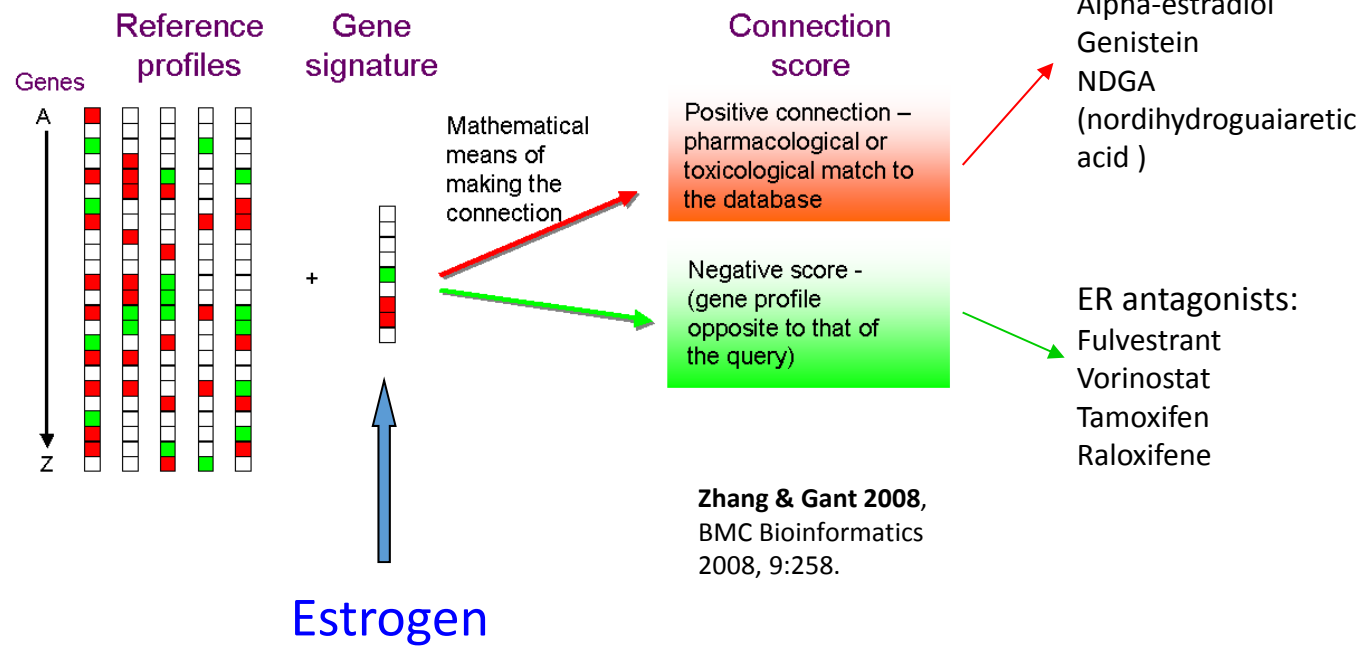
\mathbf{R} is a reference gene expression profile, \mathbf{s} is a gene signature. The null hypothesis H_0 states that there is no underlying biological connection between \mathbf{R} and \mathbf{s} , and that \mathbf{s} is merely a random gene signature.

The calculation of p-value:

After observing a connection score between a reference profile \mathbf{R} and a gene signature \mathbf{s} , we generate a large number (10,000) of random gene signatures, and calculate the 10,000 connection scores. The proportion of scores higher than the observed score (in absolute values) is the p value.



453 Ref profiles
(164 compounds)
5 human cell lines



Fujimoto et al (2004), Estrogenic activity of an antioxidant, nordihydroguaiaretic acid (NDGA), Life Sciences, 74, 1417-1425, where NDGA has been shown to have estrogenic activity and able to elicit an estrogen-like response.



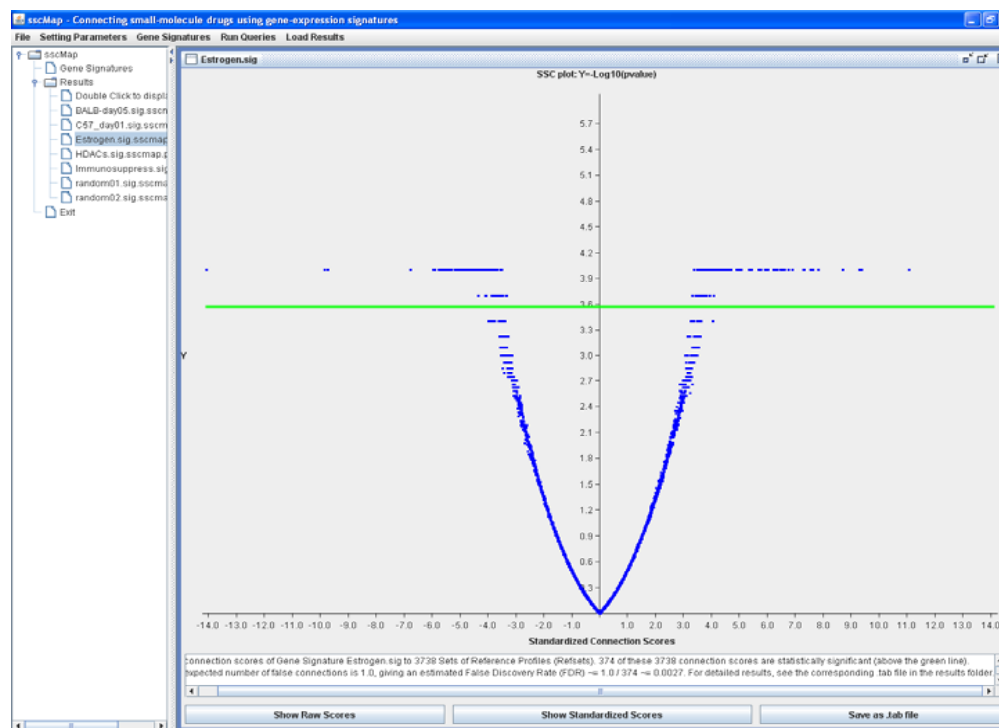
sscMap: A Java software for the connectivity mapping

<http://purl.oclc.org/NET/sscMap>

Bundled with 6100 individual reference profiles, covering over 1000 compounds.

The implementation of the improved framework

Zhang & Gant 2009, BMC Bioinformatics 2009, 10:236.



Three fronts to push forward

- Methodological and algorithmic advancement
- Novel applications: phenotypic targeting and manipulation
- Software and Tools: High performance computing model: multi-core, GPU, Cluster



Methodology and software development

O'Reilly PG, Wen Q, Bankhead P, Dunne PD, McArt DG, McPherson S, Hamilton PW, Mills KI, Zhang SD. **QUADrATiC**: scalable gene expression connectivity mapping for repurposing FDA-approved therapeutics. BMC Bioinformatics. 2016 May 4;17(1):198.

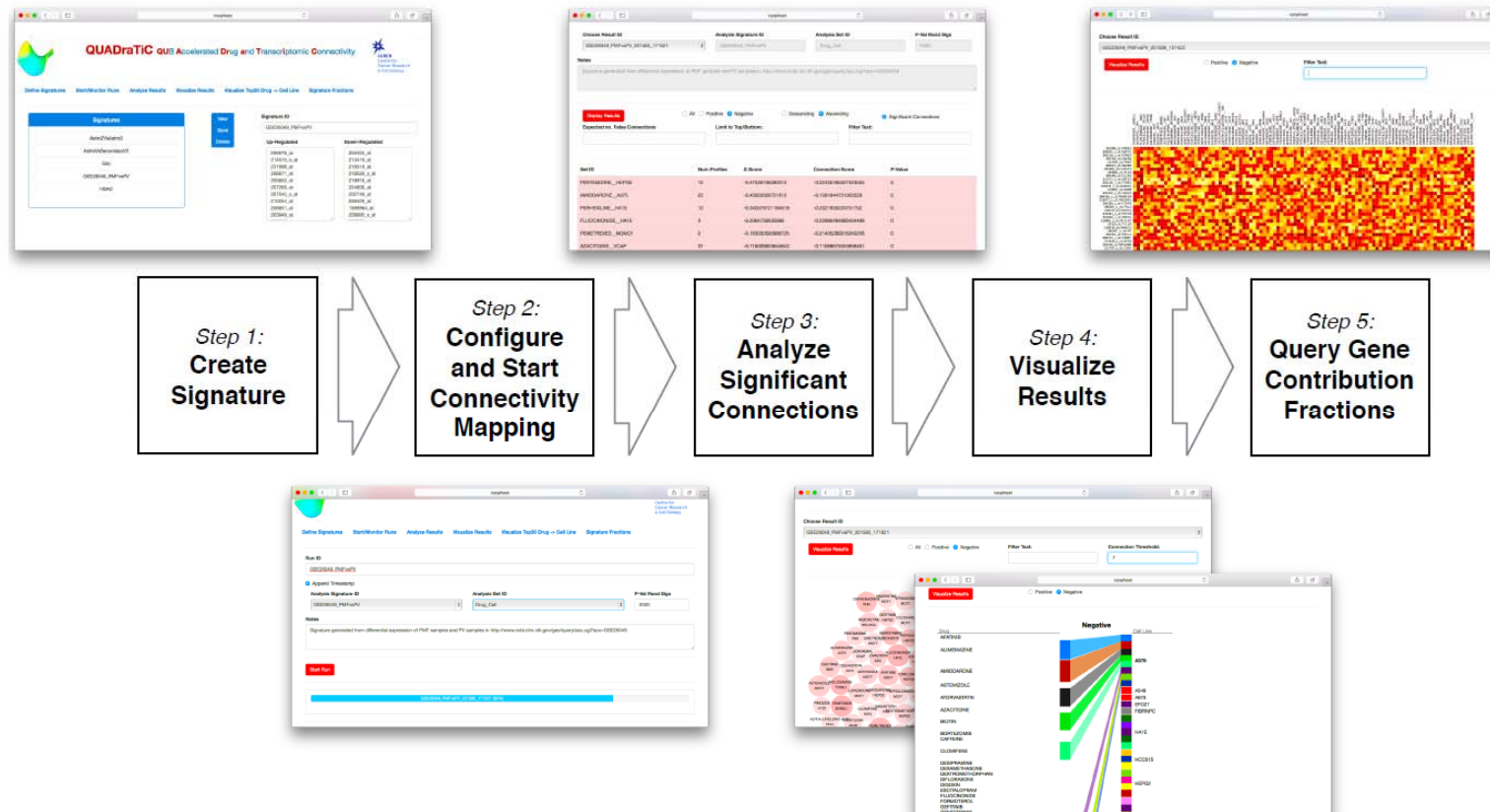
McArt DG, Bankhead P, Dunne PD, Salto-Tellez M, Hamilton P, Zhang SD. **cudaMap**: a GPU accelerated program for gene expression connectivity mapping. BMC Bioinformatics. 2013 Oct 11;14:305.

McArt DG, Dunne PD, Blayney JK, Salto-Tellez M, Van Schaeybroeck S, Hamilton PW, Zhang SD. Connectivity Mapping for Candidate Therapeutics Identification Using Next Generation Sequencing **RNA-Seq Data**. PLoS One. 2013 Jun 26;8(6):e66902.

McArt DG, Zhang SD. Identification of candidate small-molecule therapeutics to cancer by **gene-signature perturbation** in connectivity mapping. PLoS One. 2011 Jan 31;6(1):e16382..



Connectivity mapping to LINCS FDA reference profiles





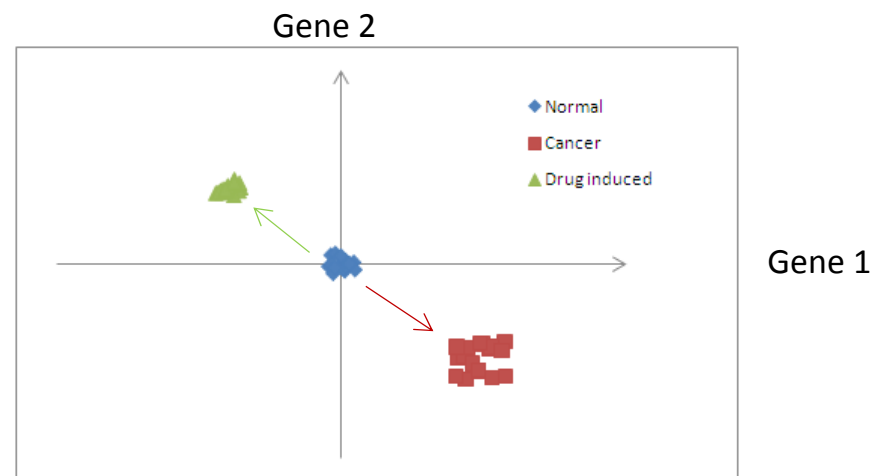
03 | Applications

The Principle of Phenotypic targeting

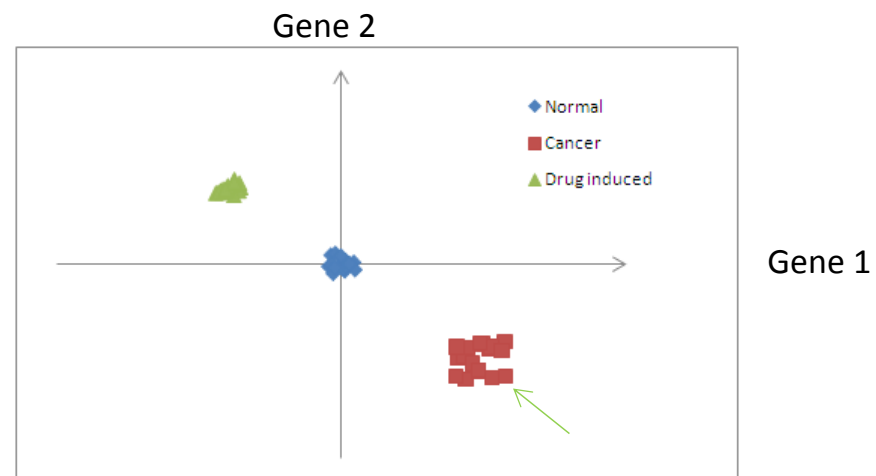
- Biological states can be adequately characterised by gene-expression profiles
- A phenotype (biological state) represented by localized points in the high-dimensional gene-expression space.
- Use connectivity mapping to identify compounds that can move a state point towards its desired locations.



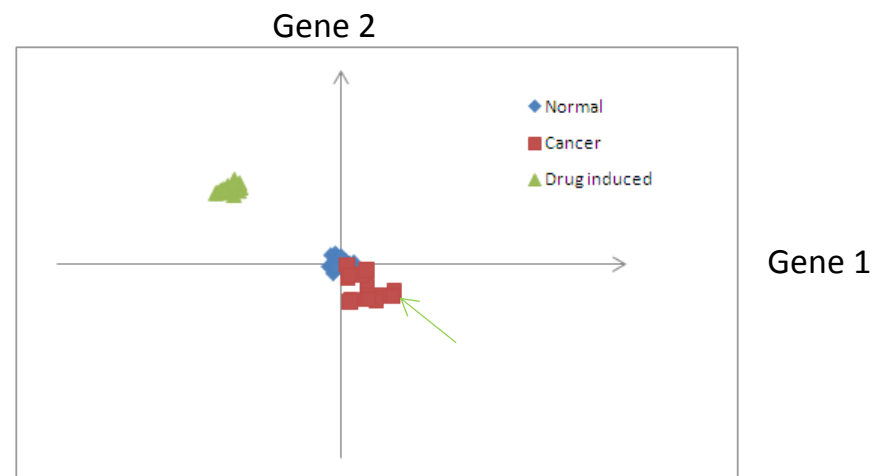
Schematic view: a two-gene world



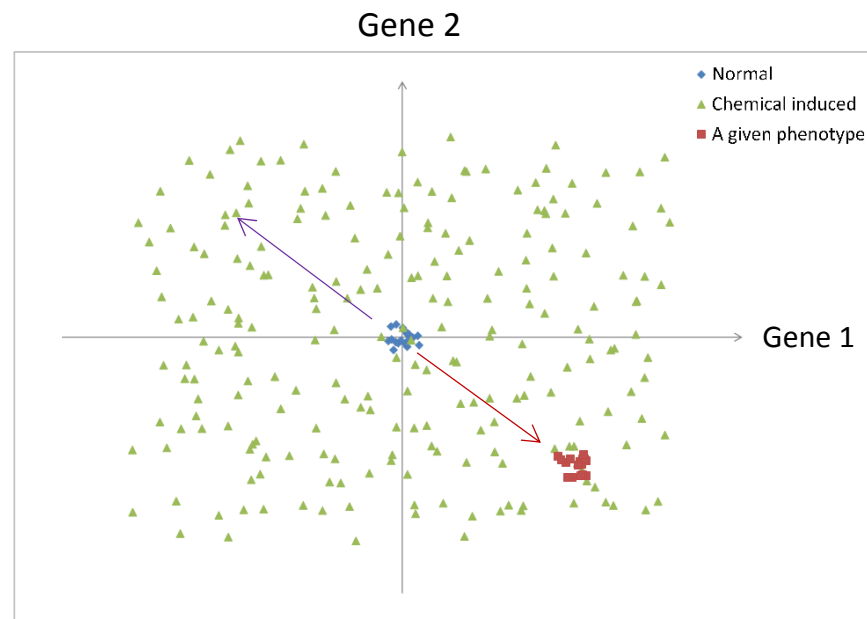
Schematic view: a two-gene world



Schematic view: a two-gene world



The general principles of phenotypic targeting: A Schematic view in a two-gene world



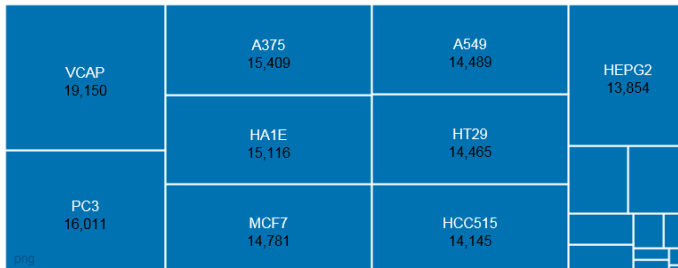
Recent data release from the Broad Institute LINCS project

GENETIC REAGENTS

22,119

knock down	18,492
over expression	3,492
variant	135

png



CELLULAR CONTEXTS

77

cancer	59
primary	10
other	8

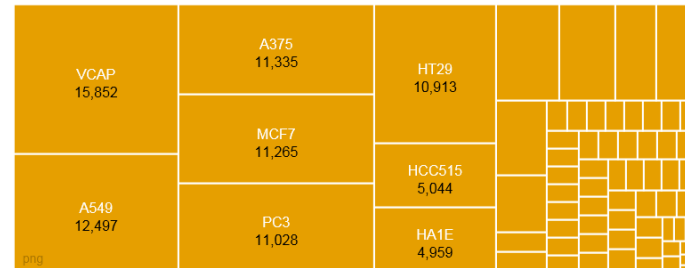
png

CHEMICAL REAGENTS

20,413

tool compounds	14,339
drugs and bioactives	5,585
other	489

png



Gene signature constructions with three major types of analyses

- Differential expression analysis: typically two condition comparisons. Disease vs Normal, Resistant vs sensitive
- Co-expression analysis: Closest correlates of key genes of interest; new players of biological processes or pathways.
- Survival analysis: Integrating clinical and genomics data for Prognostic markers identification.



Differential expression based analyses and applications: Disease vs Normal

Ramsey et al 2013, Stem Cells. 31(7):1434-45.

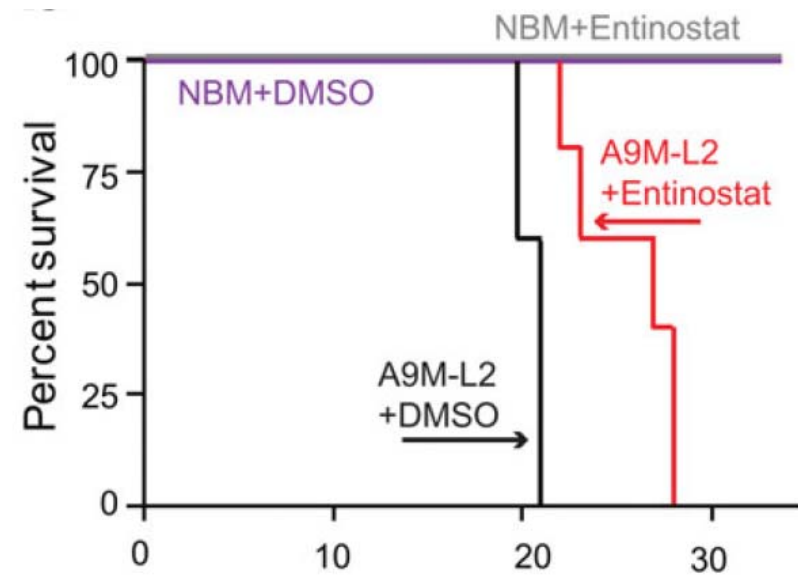
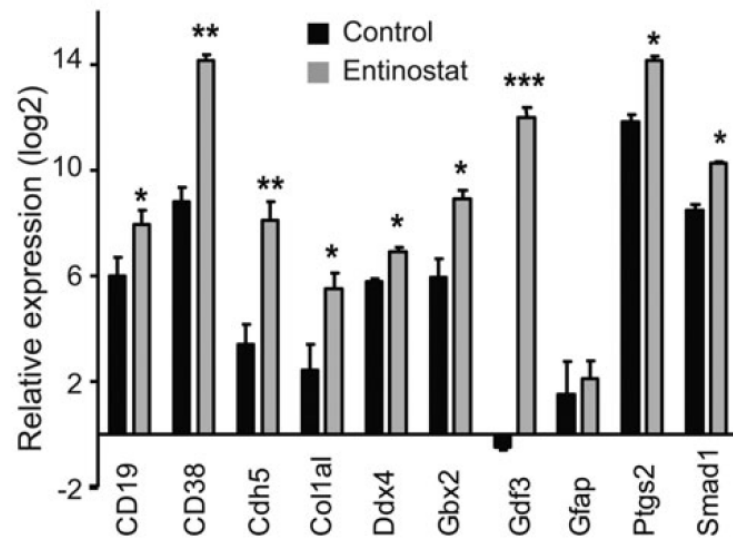
STEM CELLS®

TRANSLATIONAL AND CLINICAL RESEARCH

Entinostat Prevents Leukemia Maintenance in a Collaborating Oncogene-Dependent Model of Cytogenetically Normal Acute Myeloid Leukemia

JOANNE M. RAMSEY,^a LAURA M.J. KETTYLE,^a DANIEL J. SHARPE,^a NUALA M. MULGREW,^a GLENDA J. DICKSON,^a
JANET J. BIJL,^b PAMELA AUSTIN,^b NADINE MAYOTTE,^b SONIA CELLOT,^b TERENCE R.J. LAPPIN,^a SHU-DONG ZHANG,^a
KEN I. MILLS,^a JANA KROSL,^b GUY SAUVAGEAU,^{b,c,d} ALEXANDER THOMPSON^a





Differential expression based analyses and applications: Disease vs Normal

Liberante et al 2016, Oncotarget. 7(6):6609-19.

www.impactjournals.com/oncotarget/

Oncotarget, Vol. 7, No. 6

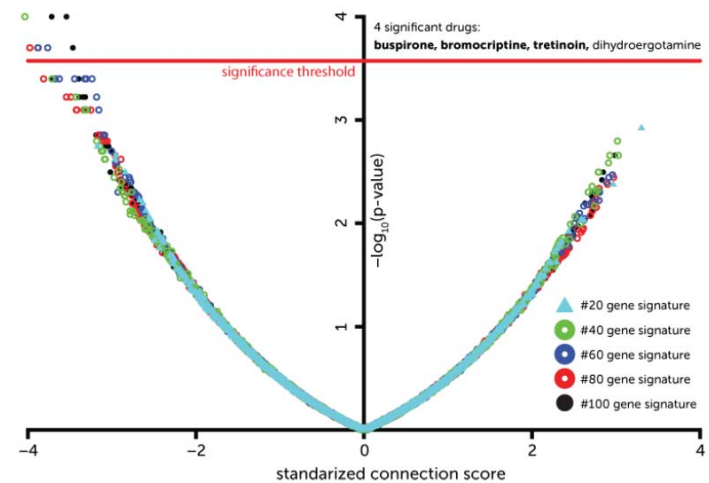
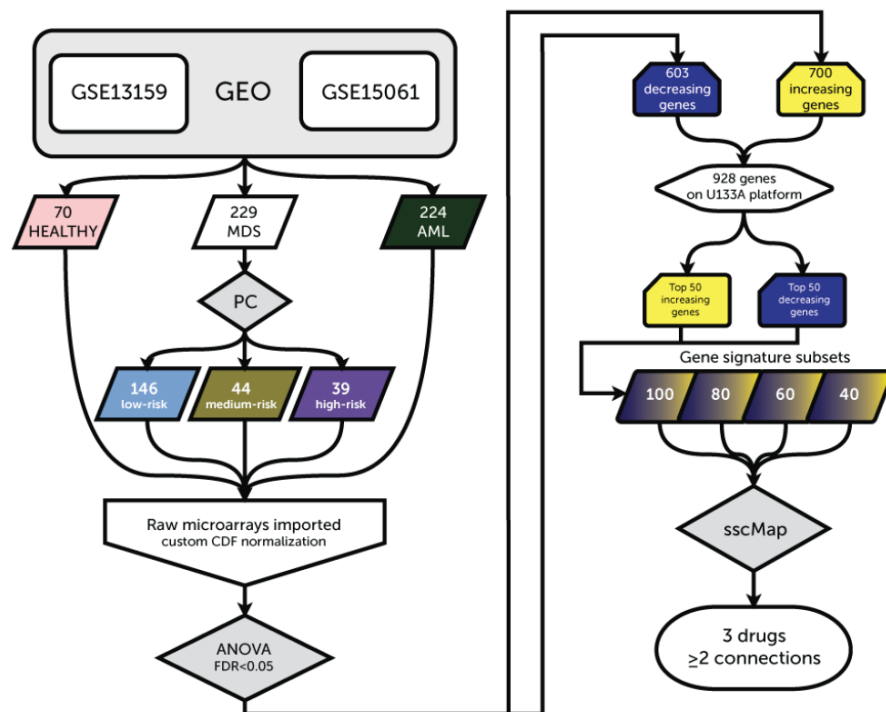
Identification and validation of the dopamine agonist bromocriptine as a novel therapy for high-risk myelodysplastic syndromes and secondary acute myeloid leukemia

Fabio Giuseppe Liberante¹, Tara Pouryahya¹, Mary-Frances McMullin¹, Shu-Dong Zhang^{1,*}, Kenneth Ian Mills^{1,*}

¹Centre for Cancer Research and Cell Biology (CCRCB), Queen's University Belfast, Belfast, United Kingdom

*These authors contributed equally to this work





Co-expression based analyses and applications: Seed genes

Malcomson et al 2016, PNAS 113(26):E3725-34.

Connectivity mapping (ssCMap) to predict A20-inducing drugs and their antiinflammatory action in cystic fibrosis

Beth Malcomson^{a,1}, Hollie Wilson^{a,1}, Eleonora Veglia^b, Gayathri Thillaiyampalam^c, Ruth Barsden^a, Shauna Donegan^a, Amal El Banna^a, Joseph S. Elborn^a, Madeleine Ennis^a, Catriona Kelly^d, Shu-Dong Zhang^{c,d}, and Bettina C. Schock^{a,2}

PNAS PLUS



The utility of an effective connectivity map

1. Developing new biological hypotheses following the leads of novel connections
2. Chemical screening process, drug development pipeline
3. Identify novel pharmacological and toxicological properties of candidate compounds
4. Predictive toxicology
5. Discover new uses for old drugs
6. Extensible to other 'omics technologies, eg. Protein profiles, metabolite profiles.





04 | Application to Cat-App

Apply connectivity mapping to Cat-App

1. Connectivity mapping scores as similarity measure for the substances
2. Unsupervised clusters as newly discovered categories
3. Supervised classification to assign substances to known categories.
4. Connectivity mapping scores as slices into the ToxPi framework.
5. Per-cell line connectivity scores as slices in a genomics-only ToxPi.



Work package 4B Connectivity mapping analysis

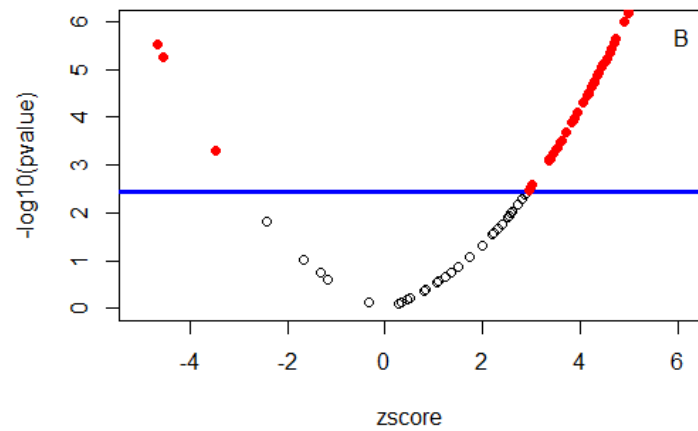
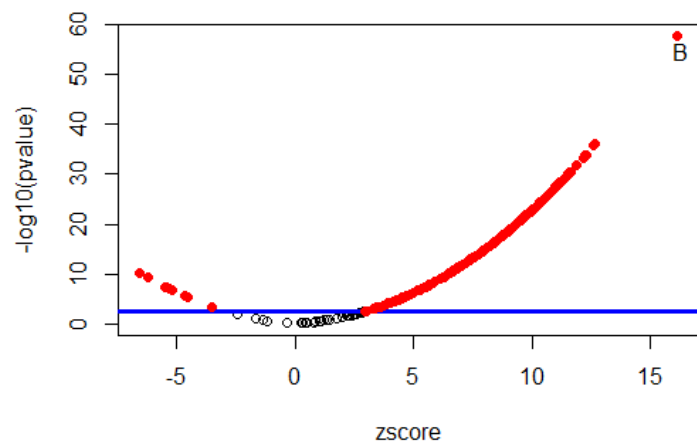
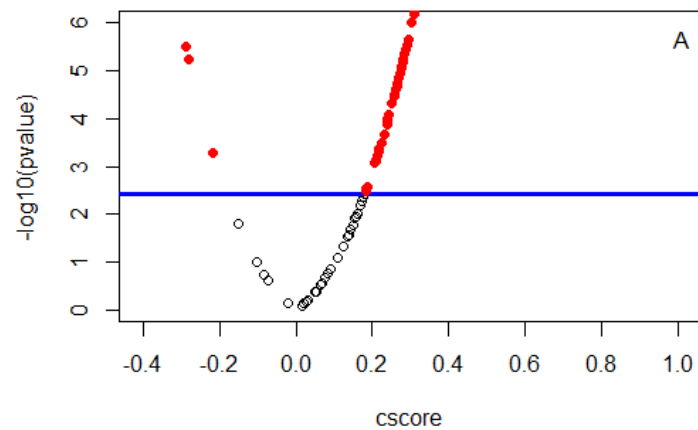
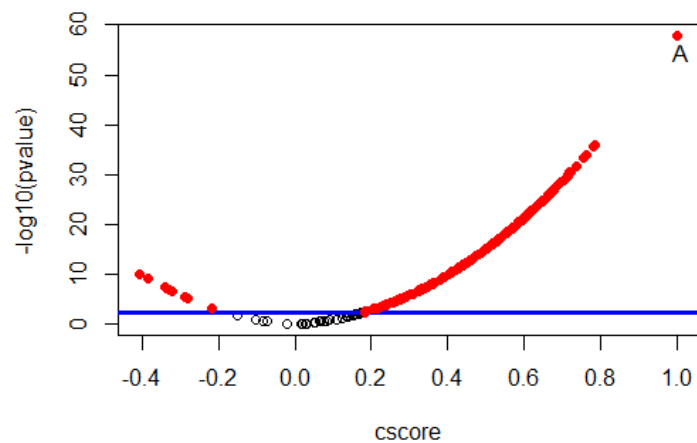
1. Similarity matrix of the ~ 160 Concawe UVCBs: establishing significant connections among the assayed UVCBs themselves.

(a) Creation of the expression profiles for the ~ 160 petroleum substances.

Substance	concentration	Cell line	time point
-----------	---------------	-----------	------------

(b) Clustering at the substance level or individual profile level (substance-cell line-concentration combinations)





Work package 4B Connectivity mapping analysis

2. Performing robustness analysis on the connections and groupings of Concawe substances obtained.

- Using a suitable public dataset to develop and test the algorithms and analysis pipeline.
- Applying the tested procedures to Cat-App data.



Full Z-Score Data Set
~1.3 million (= N) instances



Select M instances
randomly as
reference set



X



Extract
Landmark
Values



X*



Select M instances
randomly as
test set



Y



Extract
Landmark
Values



Y*



Summary

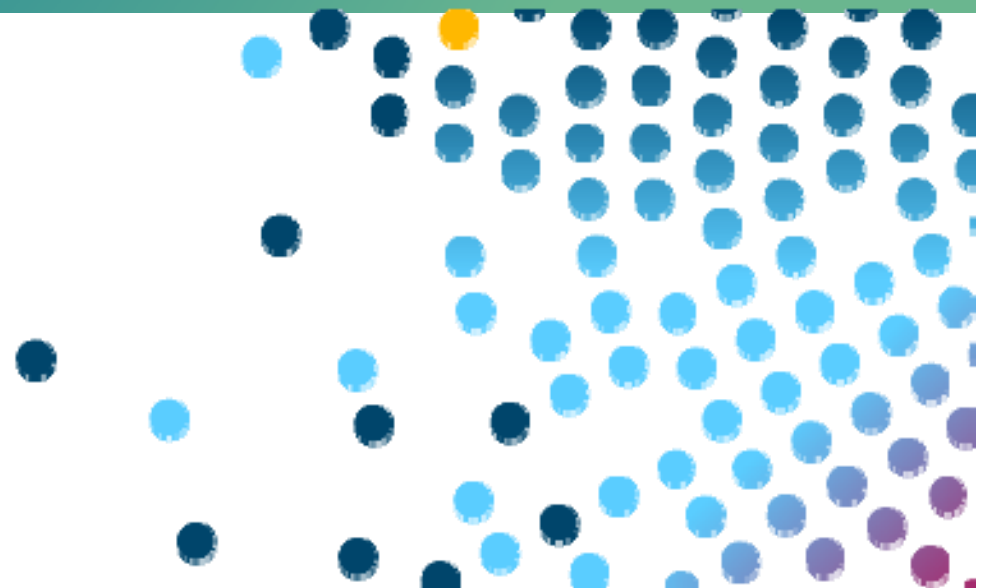
- Gene expression connectivity mapping
- Initial success and further development
- Typical applications
- Application to Cat-App
- Discussion



Boulevard du Souverain, 165 B
1160 Brussels
Belgium

T. +32 2 566 91 60
F. +32 2 566 91 81

www.concawe.eu



Work package 4B Connectivity mapping analysis

1. Similarity matrix of the ~ 160 Concawe UVCBs: establishing significant connections among the assayed UVCBs themselves.

(a) Creation of the expression profiles for the ~ 160 petroleum substances.

Substance	concentration	Cell line	time point
-----------	---------------	-----------	------------

(b) Clustering at the substance level or individual profile level (substance-cell line-concentration combinations)



Work package 4B.1 Connectivity mapping analysis

(a) Creation of the expression profiles for the ~ 160 petroleum substances.

- Factors: Substance, Cell line, Concentration, time point
- Input data: Processed and normalized gene expression data from WP4a
- Output: signed ranking of all genes assayed
- Methods: For individual dosage: ranking genes by log ratio and p-value (if available)

For multiple dosage (dose-response data), ranking genes by their correlations with dosage, and statistical significance (p-value); or by maximum log ratio across doses.



Work package 4B.1 Connectivity mapping analysis

(b) Clustering at the substance level or individual profile level (substance-cell line-concentration combinations)

- Connectivity mapping scores used as a similarity measure
- For a given cell line, a similarity matrix is constructed for all the substances
- An overall similarity matrix is constructed by averaging across cell lines
- Clustering be performed using the overall similarity matrix



Work package 4B Connectivity mapping analysis

2. Performing robustness analysis on the connections and groupings of Concawe substances obtained.

- Using a suitable public dataset to develop and test the algorithms and analysis pipeline.
- Applying the tested procedures to Cat-App data.



Work package 4B.2 Connectivity mapping analysis

2. Robustness analysis: suitable public dataset to use

- Gene expression profiling data.
- Perturbgen vs control design, ideally with replicates.
- Share a common set of transcripts measured.



Work package 4B.2 Connectivity mapping analysis

2. Robustness analysis: Methodology

- Simulation based approach
- Controlled levels of random noise
- Gene signature perturbations
- Effect of reduced gene expression space



Full Z-Score Data Set
~1.3 million (= N) instances



Select M instances
randomly as
reference set



X



Extract
Landmark
Values



X*



Select M instances
randomly as
test set



Y

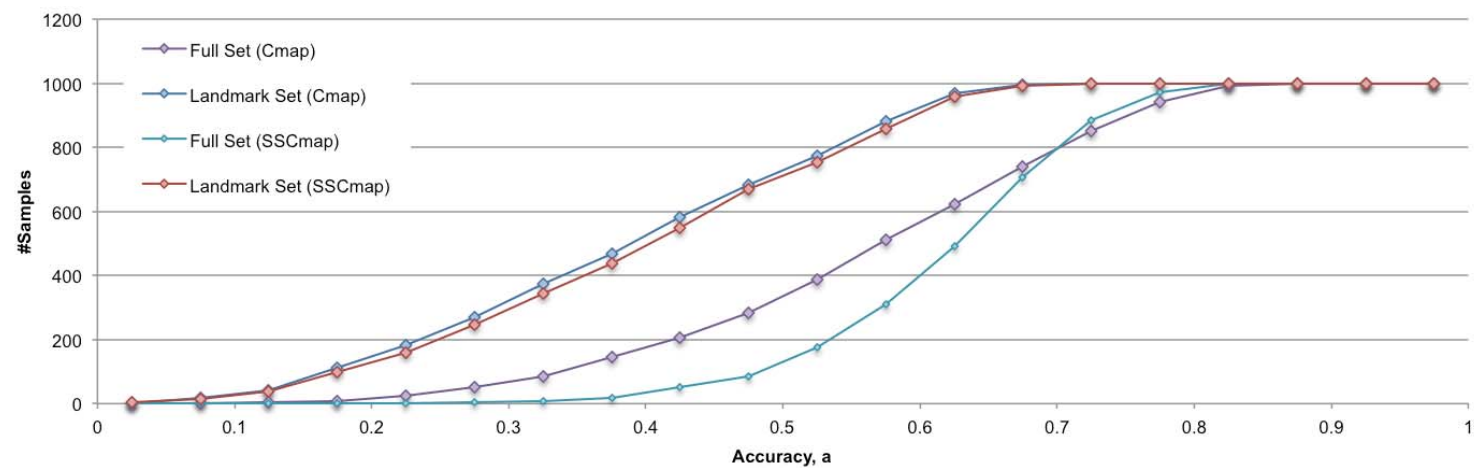
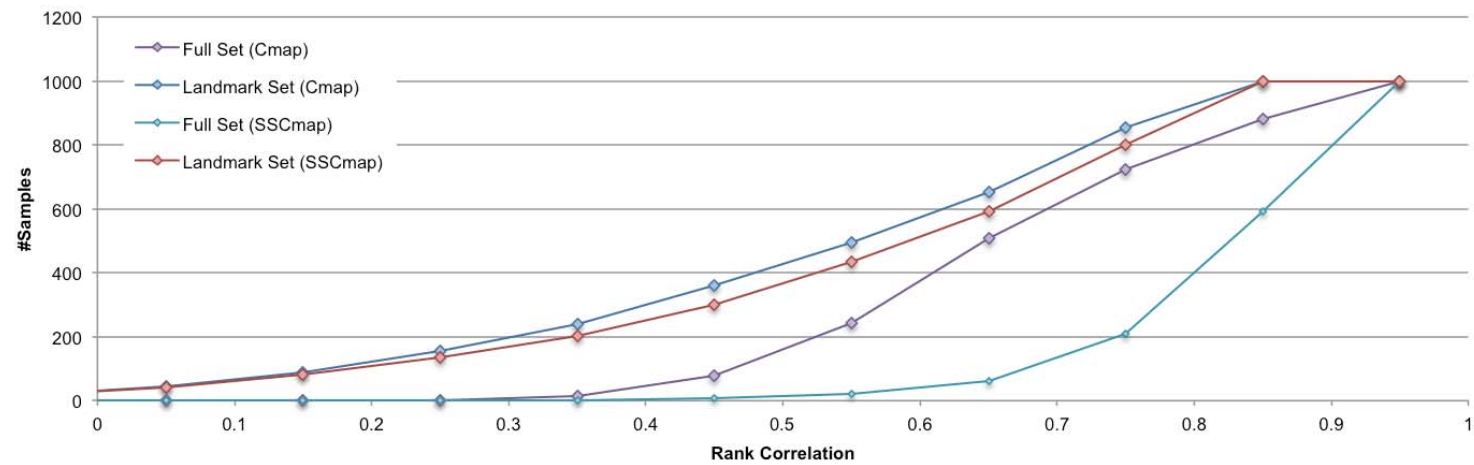


Extract
Landmark
Values



Y*





Boulevard du Souverain, 165 B
1160 Brussels
Belgium

T. +32 2 566 91 60
F. +32 2 566 91 81

www.concawe.eu

