European Commission

HORIZON 2020

The EU Framework Programme for Research and Innovation

Assessing the toxicity of complex substances: A 21st Century approach to an old problem

Timothy W. Gant (DH/PHE)
Ivan Rusyn (TAMU)
Arlean Rohde (Concawe)

Shu-Dong Zhang (QUB)

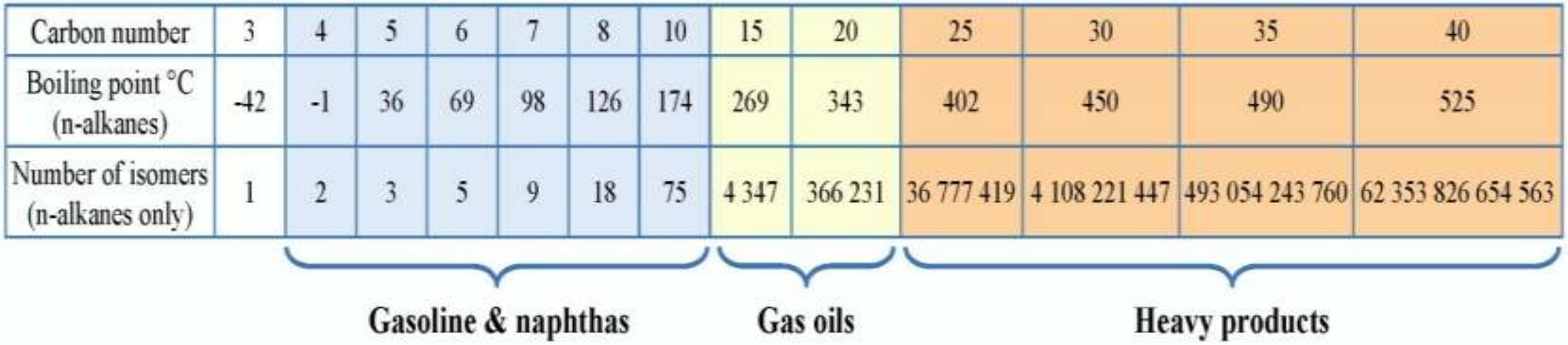
Public Health England

Queen's University Belfast

The challenge

Substances of Unknown or Variable composition, Complex reaction products and Biological materials (UVCB)

$C_{2\text{-}100+}$
$C_{4\text{-}11}$
$C_{8\text{-}15}$
$C_{10\text{-}25}$
$C_{15\text{-}65}$
$C_{20\text{-}100+}$
$C_{35\text{-}100+}$



Carbon number	3	4	5	6	7	8	10	15	20	25	30	35	40
Boiling point °C (n-alkanes)	-42	-1	36	69	98	126	174	269	343	402	450	490	525
Number of isomers (n-alkanes only)	1	2	3	5	9	18	75	4 347	366 231	36 777 419	4 108 221 447	493 054 243 760	62 353 826 654 563
		Gasoline & naphthas						Gas oils		Heavy products			

Do all these differencing mixtures of these products with similar physical chemical properties have the same biological (in)activity?

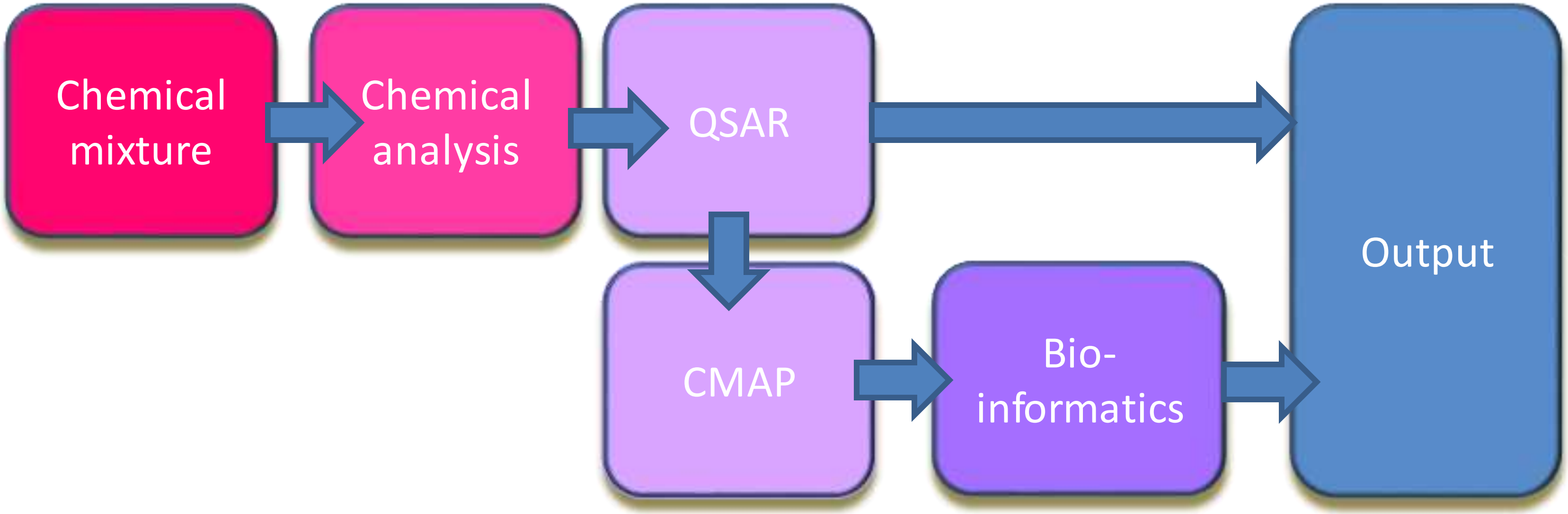
Potential Solutions

Primary Methods

- QSAR
- Mapping of biological activity by gene expression (Cmap)

Supporting methods

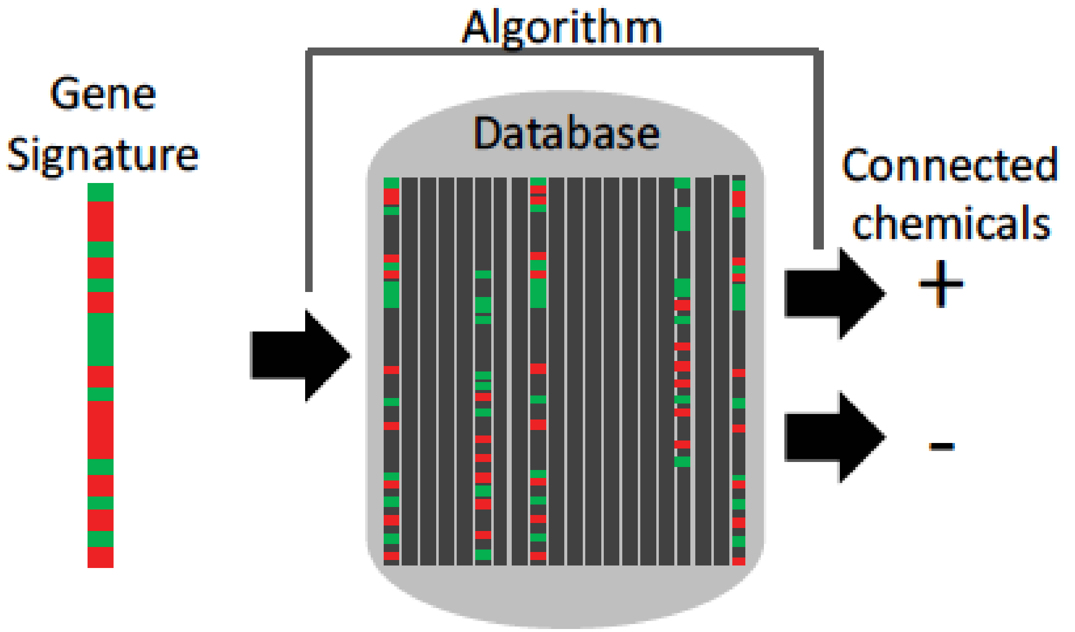
- Chemical analysis
- Bioinformatics



The objective

The CAT-APP framework will provide regulators and registrants with a cost-effective integrative approach to solving the similarity challenges of UVCBs and will define the best practical strategies for overcoming the hesitation to accept the read across and grouping approaches.

Connectivity mapping method (Cmap)



Agonists

Antagonists

Key components

1. Reference Profiles: A set of gene expression profiles, obtained from systematic microarray gene expression profiling.

2. Gene signature: a short list of important genes differentially expressed as a result of particular biological condition or biochemical pertubation.

3. Connection score: a function of a Reference Profile and a Gene Signature. It should reflect the underlying biological connection between them.

The reference set circa 2008

The Broad Connectivity Map 02 base data set

5 cell types : MCF7, PC3, SKMEL5, HL60, and ssMCF7 (ss = serum starved) : 1309 small molecules (many in multiple cell lines)

7056 Affymetrix microarrays on the same basic platform.

HG-U133A
HT_HG-U133A_EA
HT_HG-U133Au

6100 treatment instances, hence
6100 reference gene-expression profiles

The Algorithm

- The Broad uses the Kolmogorov-Smirnov test which looks for equality of distributions - essentially looks in a dataset for a compound with a set of gene expressions whose distribution matches that in the reference set.

- The drawback with this method is that is does not allow a direct statistical evaluation of the matches to reduce false positives.

- In this work, published in BMC Bioinformatics (2008; 9: 258), Shu-Dong Zhang (CAT-APP WP5) then working in my laboratory devised a new connectivity algorithm that uses a Monte-Carlo method to control for false positives. The method is implemented in the software sscMap (freely available).

Connection: Query profile (ordered)

For an ordered gene list of N genes where g_i represents gene (g) number (i) in the signature that score for that gene is the product of its signed rank in the query profile (s) and that in the reference profile (R). Thus the connection score between a query signature of m genes and the reference profile is:

$$C(\mathrm{R},\mathrm{s}) = \sum_{i=1}^{m} R(g_i)s(g_i),$$

If a gene has the same regulation status in the reference and signature queries then it will make a positive contribution to the score. Else the contribution will be negative. If all the genes in the query signature score negatively this indicates a form of antagonism that can be very useful.

Public Health England

The maximum connection score

- The maximum connection (C_{max}) score is then achieved when then genes are ranked in the same order in the query profile as in the reference profile and have the same sign
- The overall connection strength (c) is then a ratio of the connection score over the maximum connection score

$$c=(C/C_{max})$$

For example (ordered)

i	Reference rank and sign	Q_1	Q_2
1	+7	+3	
2	-5		+1
3	+4		
4	+1		-2
5	-2	-2	
6	-3		
7	+6	+1	
8	-10		+3
9	+9		
10	+8		

Q_1=+31 –> agonist (max score = 56)

Q_2 = -37 -> antagonist (max score = -56)

Calculation of significance

Null hypothesis – for a gene signature of length m there is no connection between it and any given reference gene expression signature of length N.

Calculation of significance

A random set of genes of length m is selected from the reference profile

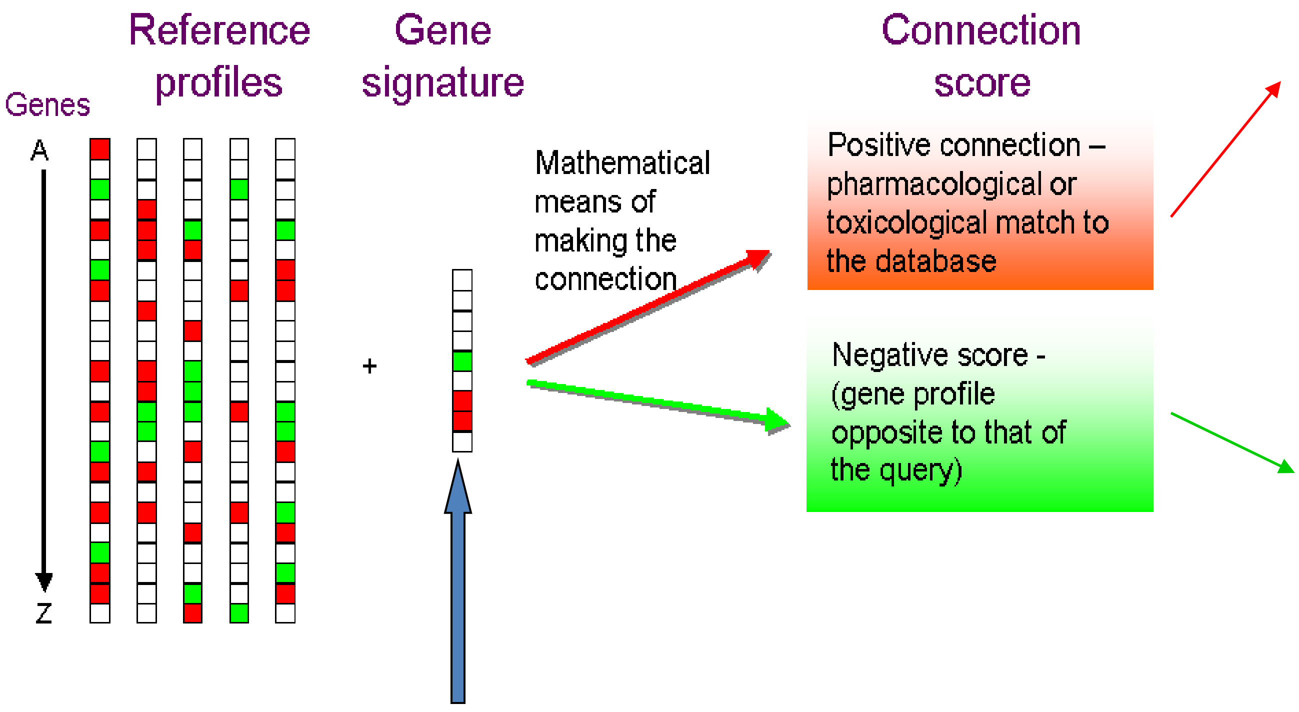
For each random query signature a connection score is calculated. This is repeated many times (typically 10,000) and the proportion of connectivity scores that are greater, or equal, to the observed score in absolute value are an estimate of the two tailed p-value score for a given instance

Zhang and Gant BMC Bioinformatics (2008) 9:258

Public Health England

Estrogens

Broad Institute database 1
453 Ref profiles
(164 compounds)
5 human cell lines



Reference profiles

Gene signature

Mathematical means of making the connection

Connection score

Positive connection – pharmacological or toxicological match to the database

Negative score - (gene profile opposite to that of the query)

Estrogen

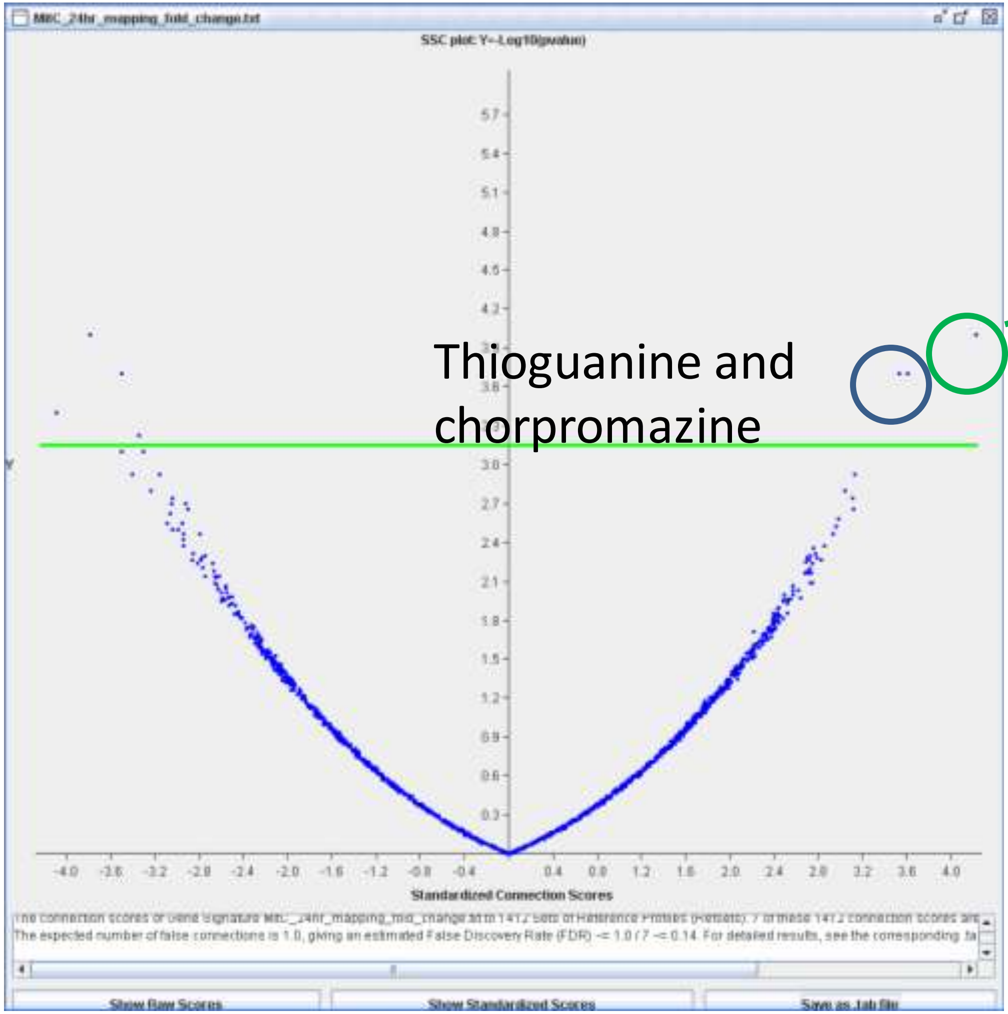
ER agonists:
Estradiol
Alpha-estradiol
Genistein
NDGA
(nordihydroguaiaretic acid)

ER antagonists:
Fulvestrant
Vorinostat
Tamoxifen
Raloxifene

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.

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

Mitomycin C 24 hour 50 genes

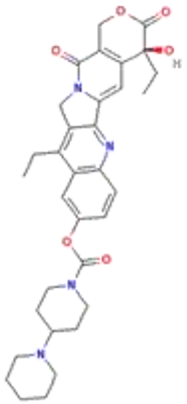


The connection scores of Query Signature MMC_24hr_mapping_fold_change.txt to 1412 Sets of Reference Profiles (Refsets). 7 of these 1412 connection scores are

The expected number of false connections is 1.0, giving an estimated False Discovery Rate (FDR) <= 1.0 / 7 <= 0.14. For detailed results, see the corresponding .ta

Thioguanine and chorpromazine

Irinotecan

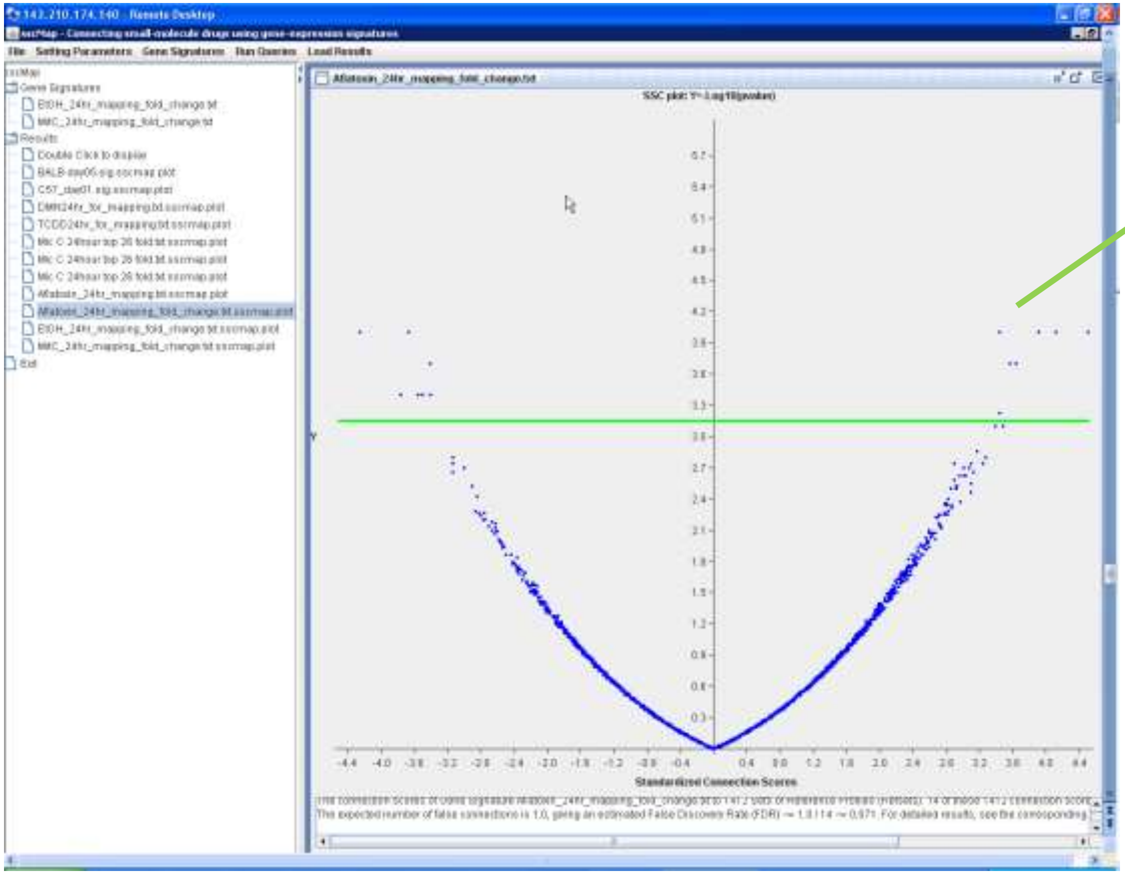


Topoisomerase I inhibitor – leads to double strand DNA breaks

Chorpromazine positive in sister chromatid exchange assay

Public Health England

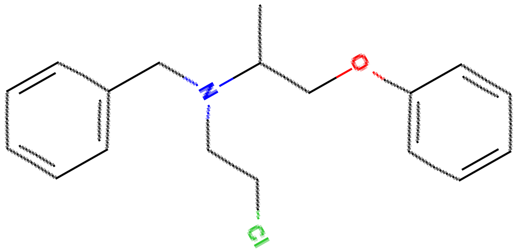
Aflatoxin B1 - 24 hour 50 genes


143.210.174.140 - Remote Desktop
sscMap - Connecting small-molecule drugs using gene-expression signatures
File
Setting Parameters
Gene Signatures
Run Queries
Load Results
sscMap
Gene Signatures
EtOH_24hr_mapping_fold_change.txt
MMC_24hr_mapping_fold_change.txt
Results
Double Click to display
BALB-day05.sig.sscmap.plot
C57_day01.sig.sscmap.plot
DMN24hr_for_mapping.txt.sscmap.plot
TCDD24hr_for_mapping.txt.sscmap.plot
Mic C 24hour top 25 fold.txt.sscmap.plot
Mic C 24hour top 25 fold.txt.sscmap.plot
Mic C 24hour top 25 fold.txt.sscmap.plot
Aflatoxin_24hr_mapping.txt.sscmap.plot
Aflatoxin_24hr_mapping_fold_change.txt.sscmap.plot
EtOH_24hr_mapping_fold_change.txt.sscmap.plot
MMC_24hr_mapping_fold_change.txt.sscmap.plot
Exit
Aflatoxin_24hr_mapping_fold_change.txt
SSC plot: Y=-Log10(pvalue)
5.7
5.4
5.1
4.8
4.5
4.2
3.9
3.6
3.3
3.0
2.7
2.4
2.1
1.8
1.5
1.2
0.9
0.6
0.3
-4.4
-4.0
-3.6
-3.2
-2.8
-2.4
-2.0
-1.6
-1.2
-0.8
-0.4
0.4
0.8
1.2
1.6
2.0
2.4
2.8
3.2
3.6
4.0
4.4
Standardized Connection Scores

The connection scores of query signature Aflatoxin_24hr_mapping_fold_change.txt to 1412 sets of reference profiles (refsets); 14 of those 1412 connection score
The expected number of false connections is 1.0, giving an estimated False Discovery Rate (FDR) = 1.0/14 = 0.071. For detailed results, see the corresponding

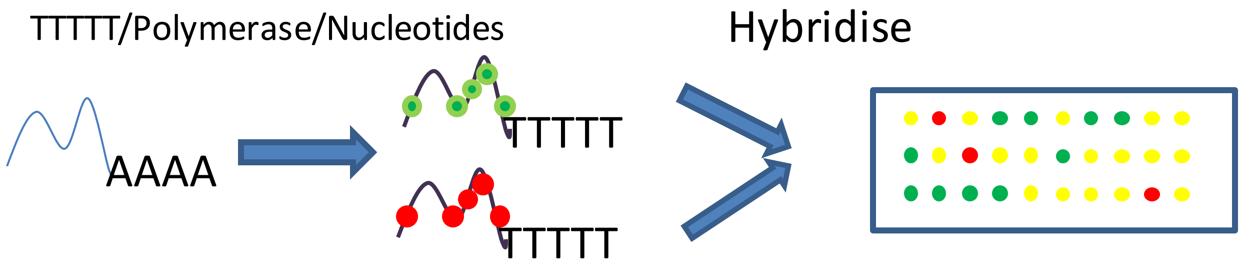
Phenoxybenzamine

GENE-TOX Evaluation A (pre-1980): **Species/Cell Type:**Nonhuman **Assay Type:**In vivo carcinogenicity studies **Assay Code:**CCG+ **Results:**Positive **Panel Report:**EMICBACK/67174; MUTAT RES 185:1-195,1987



Methods of generating gene expression signatures

Microarrays



TTTTT/Polymerase/Nucleotides

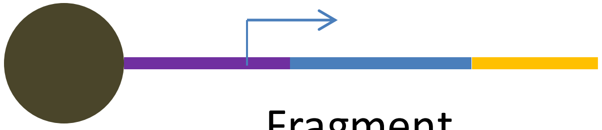
AAAA

TTTTT

TTTTT

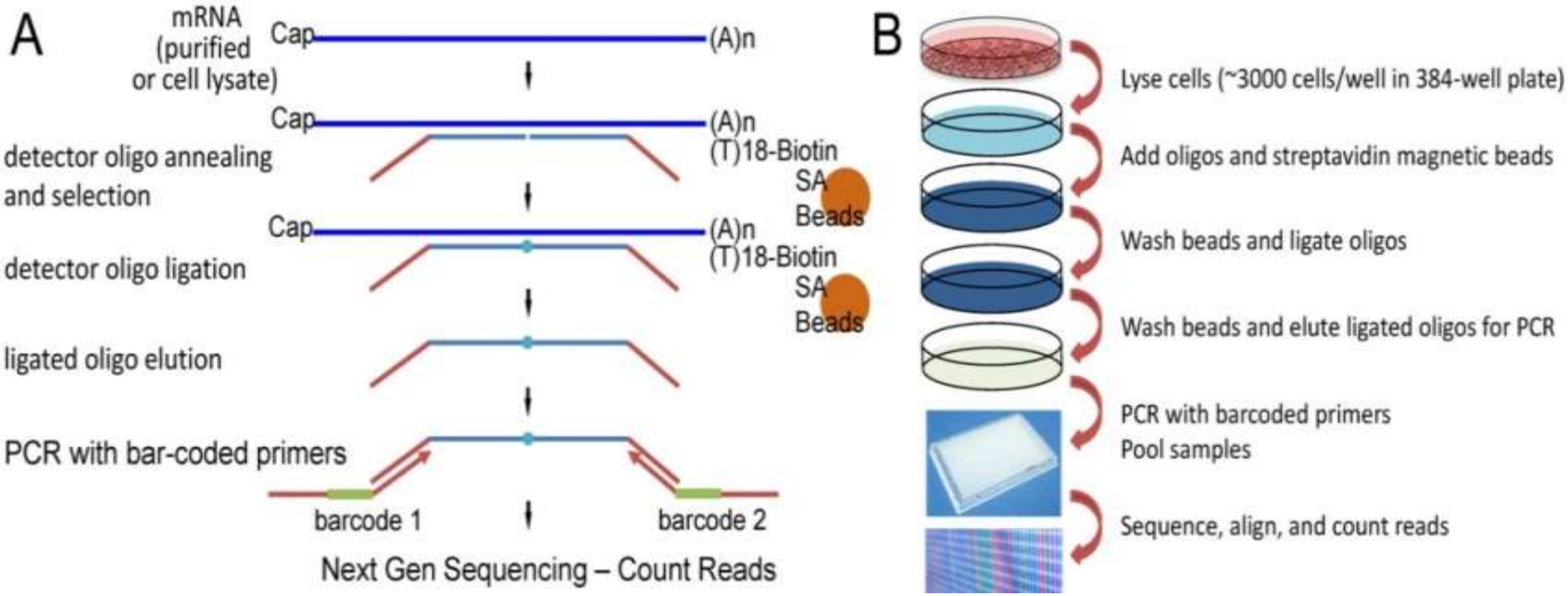
Hybridise

High throughput sequencing.



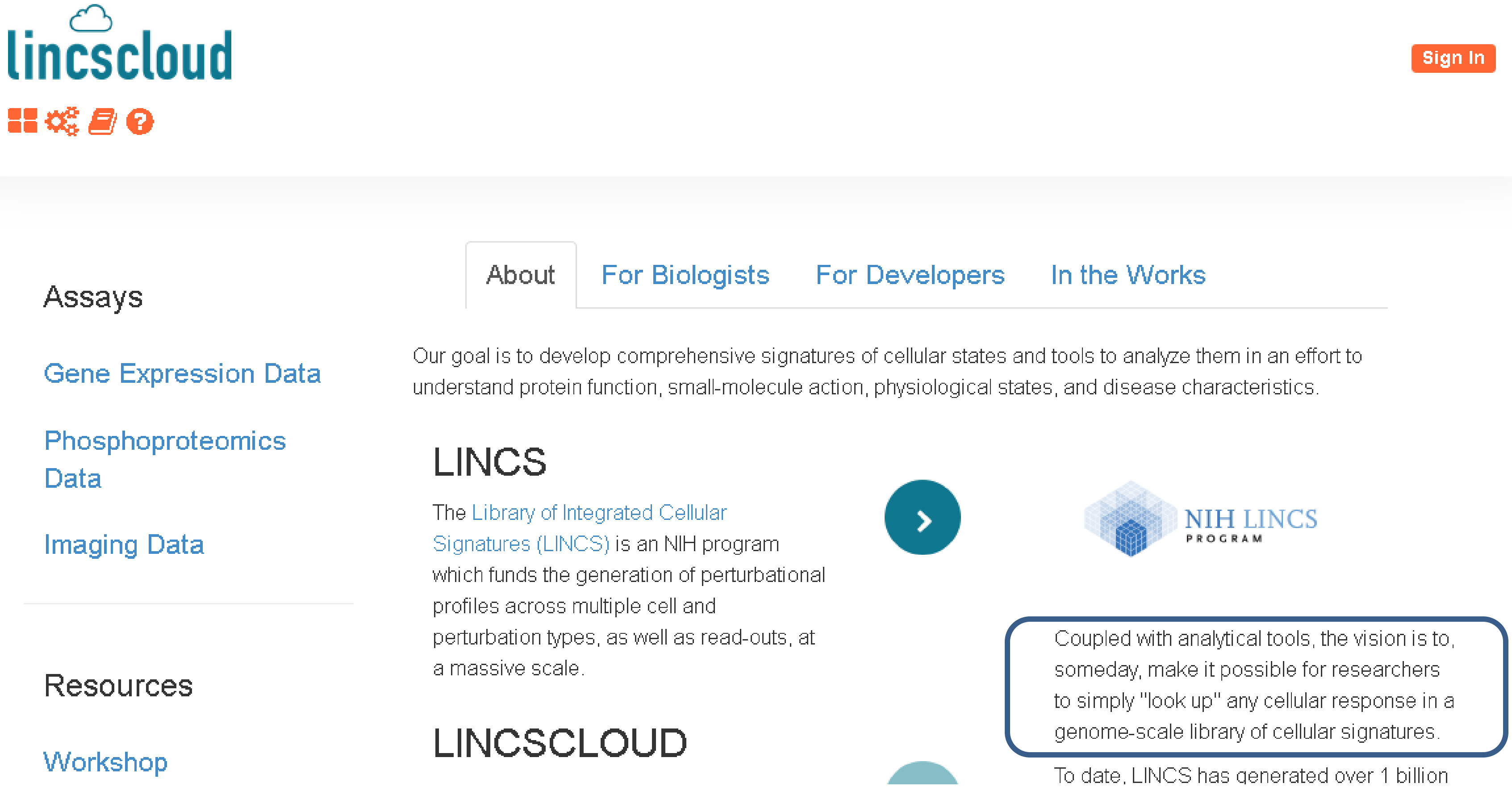
Fragment

TempO-Seq


A
mRNA (purified or cell lysate)
Cap
(A)n
detector oligo annealing and selection
(T)18-Biotin
SA Beads
detector oligo ligation
ligated oligo elution
PCR with bar-coded primers
barcode 1
barcode 2
Next Gen Sequencing – Count Reads
B
Lyse cells (~3000 cells/well in 384-well plate)
Add oligos and streptavidin magnetic beads
Wash beads and ligate oligos
Wash beads and elute ligated oligos for PCR
PCR with barcoded primers
Pool samples
Sequence, align, and count reads

The L1000 Concept
Gene expression is highly correlated. We take advantage of this high degree of correlation to reduce the number of measurements needed to generate meaningful gene expression data for the approximately 20,000 genes in the human genome. By analyzing several query-result pairs from well-known published and unpublished LINCS connections, we determined that a carefully chosen set of 1,000 genes can capture approximately 80% of the information. We call these genes the landmark genes.

LINCS



Assays

Gene Expression Data

Phosphoproteomics Data

Imaging Data

Resources

Workshop

Our goal is to develop comprehensive signatures of cellular states and tools to analyze them in an effort to understand protein function, small-molecule action, physiological states, and disease characteristics.

LINCS

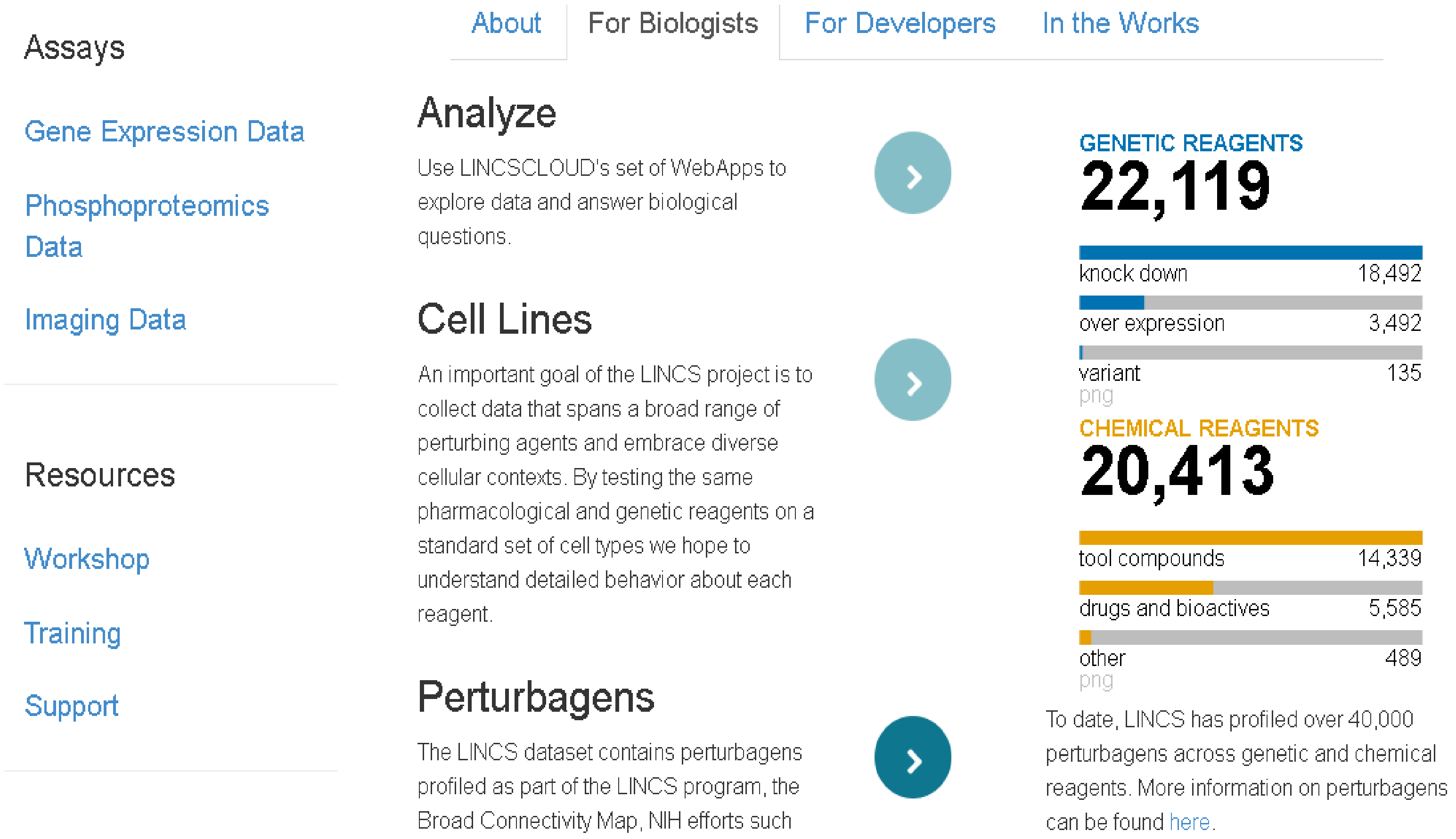
The Library of Integrated Cellular Signatures (LINCS) is an NIH program which funds the generation of perturbational profiles across multiple cell and perturbation types, as well as read-outs, at a massive scale.

Coupled with analytical tools, the vision is to, someday, make it possible for researchers to simply "look up" any cellular response in a genome-scale library of cellular signatures.

To date, LINCS has generated over 1 billion

LINCSCLOUD

LINCS status



About | For Biologists | For Developers | In the Works

Assays

Gene Expression Data

Phosphoproteomics Data

Imaging Data

Resources

Workshop

Training

Support

Analyze

Use LINCSCLOUD's set of WebApps to explore data and answer biological questions.

Cell Lines

An important goal of the LINCS project is to collect data that spans a broad range of perturbing agents and embrace diverse cellular contexts. By testing the same pharmacological and genetic reagents on a standard set of cell types we hope to understand detailed behavior about each reagent.

Perturbagens

The LINCS dataset contains perturbagens profiled as part of the LINCS program, the Broad Connectivity Map, NIH efforts such

GENETIC REAGENTS

22,119

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

png

CHEMICAL REAGENTS

20,413

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

png

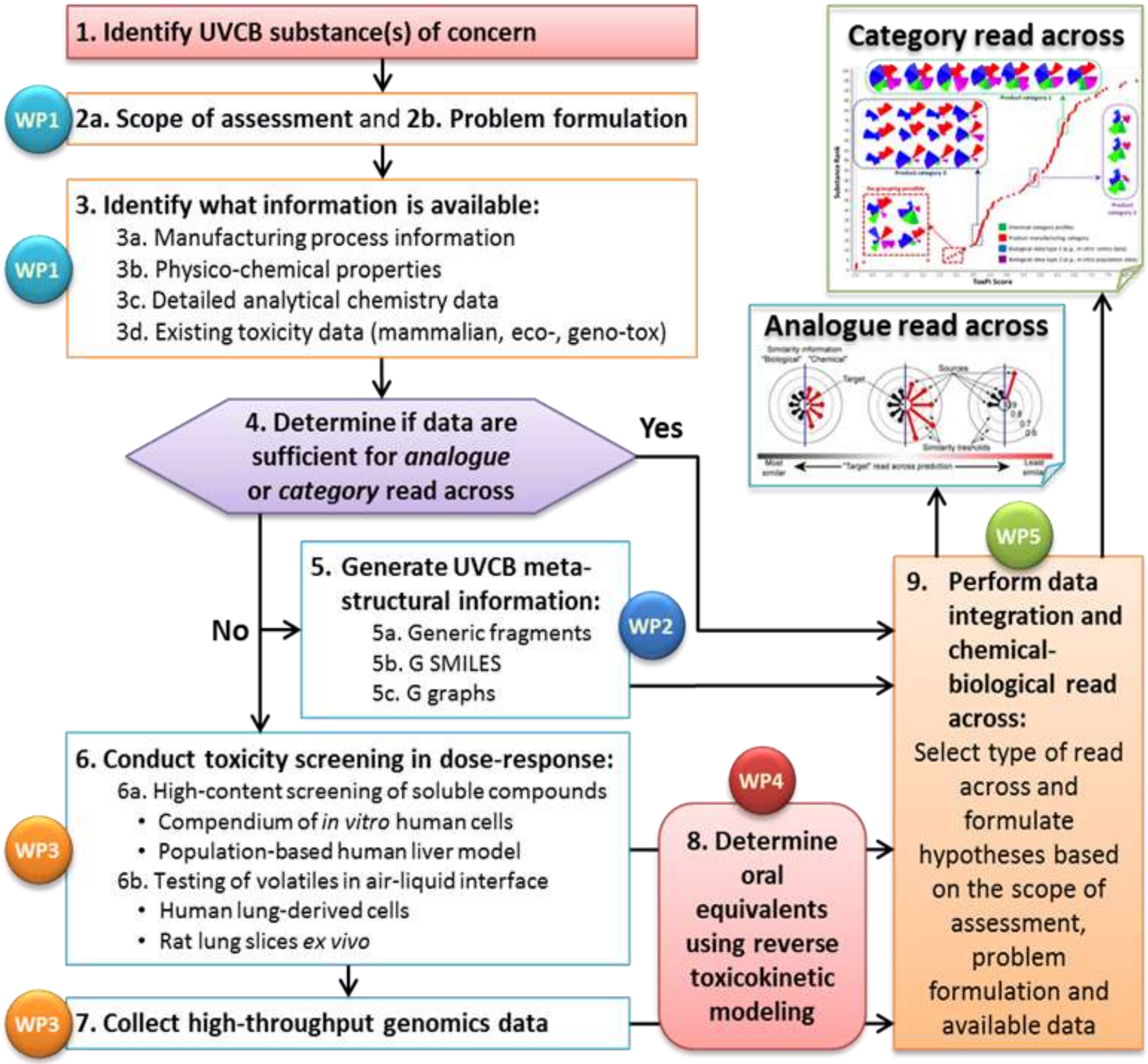
To date, LINCS has profiled over 40,000 perturbagens across genetic and chemical reagents. More information on perturbagens can be found here.

The Cmap method has not really been tested to day with mixtures or complex substances

LINCS Cells

Name	Type	Description
MCF7	Cancer cell line	Breast
PC3	Cancer cell line	Prostate
A549	Cancer cell line	Lung
A375	Cancer cell line	Melanoma
HEPG2	Cancer cell line	Liver carcinoma
VCAP	Cancer cell line	Vertebral metastasis
HCC515	Cancer cell line	Lung
HT29	Cancer cell line	colon
HEK293T	Cancer cell line	Human embryonic kidney cells
HL60	Cancer cell line	Promyelocytic leukemia
HA1E	Immortalized normal	kidney epithelial
ASC	Adipocyte	Pennington Biomedical Research Center
SKL	Primary	Skeletal myocyte
PHH	Primary	Hepatocyte
NPC	Primary	iPS-derived neural progenitor
NEU*	Primary	Terminally differentiated neuron
H9	Stem cell line	Human embryonic stem cells
H9-derived NP		

The CAT-APP plan



1. Identify UVCB substance(s) of concern

WP1 2a. Scope of assessment and 2b. Problem formulation

WP1 3. Identify what information is available:
- 3a. Manufacturing process information
- 3b. Physico-chemical properties
- 3c. Detailed analytical chemistry data
- 3d. Existing toxicity data (mammalian, eco-, geno-tox)

4. Determine if data are sufficient for *analogue* or *category* read across

Yes

No

WP2 5. Generate UVCB meta-structural information:
- 5a. Generic fragments
- 5b. G SMILES
- 5c. G graphs

WP3 6. Conduct toxicity screening in dose-response:
- 6a. High-content screening of soluble compounds
  - Compendium of *in vitro* human cells
  - Population-based human liver model
- 6b. Testing of volatiles in air-liquid interface
  - Human lung-derived cells
  - Rat lung slices *ex vivo*

WP3 7. Collect high-throughput genomics data

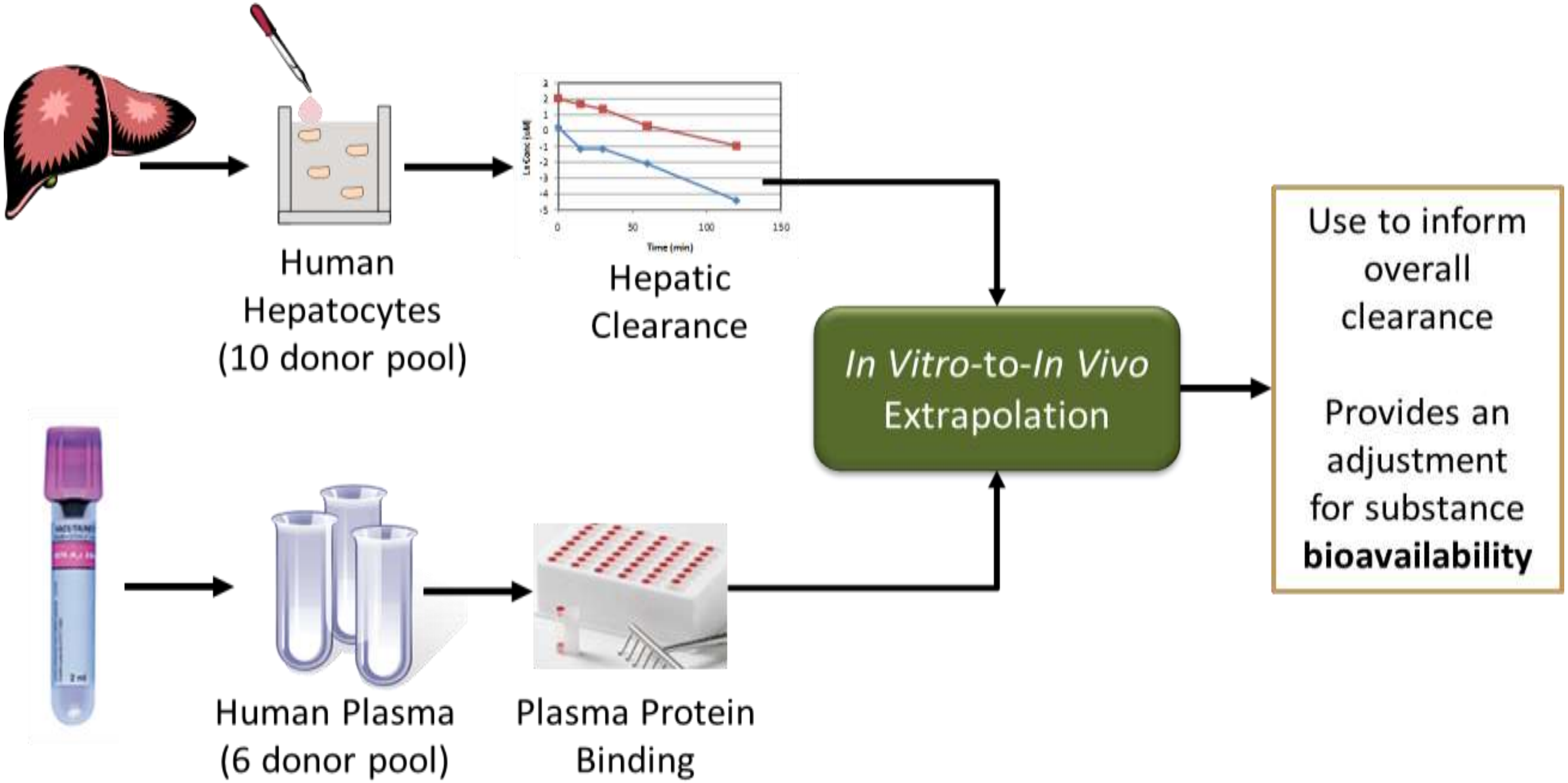
WP4 8. Determine oral equivalents using reverse toxicokinetic modeling

WP5 9. Perform data integration and chemical-biological read across: Select type of read across and formulate hypotheses based on the scope of assessment, problem formulation and available data

Analogue read across

Category read across

Getting the cell dose right



Human Hepatocytes (10 donor pool)

Hepatic Clearance

Human Plasma (6 donor pool)

Plasma Protein Binding

In Vitro-to-In Vivo Extrapolation

Use to inform overall clearance

Provides an adjustment for substance **bioavailability**

A concentration curve or four concentration points will be used in the cells.

CAT-APP cell types

- **35 established cell lines**

Advantage: available and correlate with LINCS cells; likely to respond similarly to generic chemicals allowing CMAP; open to all

Disadvantage: Limited metabolisms; do not map population variability; liquid exposure

- **50 - Human iPSC derived hepatocytes**

Advantage: metabolism; population variability

Disadvantage: more expensive and difficult to work with ; liquid exposure

- **Lung slices**

Advantage: metabolism; relevant exposure route

Disadvantage: complex cell systems; difficult and expensive to work with and design exposure routes.

Key questions

- Does the method work for complex mixtures?
- Will the output be acceptable for regulation?
- How long does the query signature need to be?
- Which cell system performs best?
- Does the output correlate with QSAR predictions?
- Is there population variance?