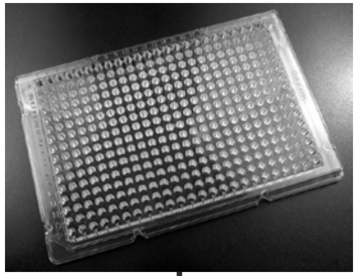
Work package 4b
CatApp Toxicogenomics data - where does it help?

Cat-App final event, Brussels, 6 September 2018

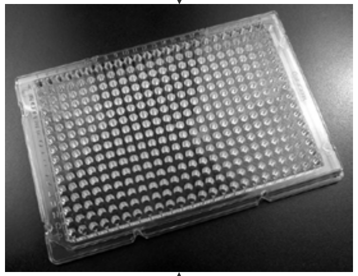
Shu-Dong Zhang, University of Ulster, Northern Ireland

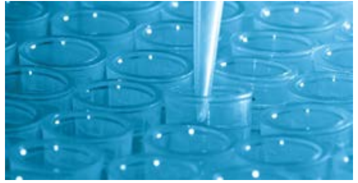
TempO-seq Technology – High-Throughput Targeted Sequencing

Screening plate (384-well) with cells

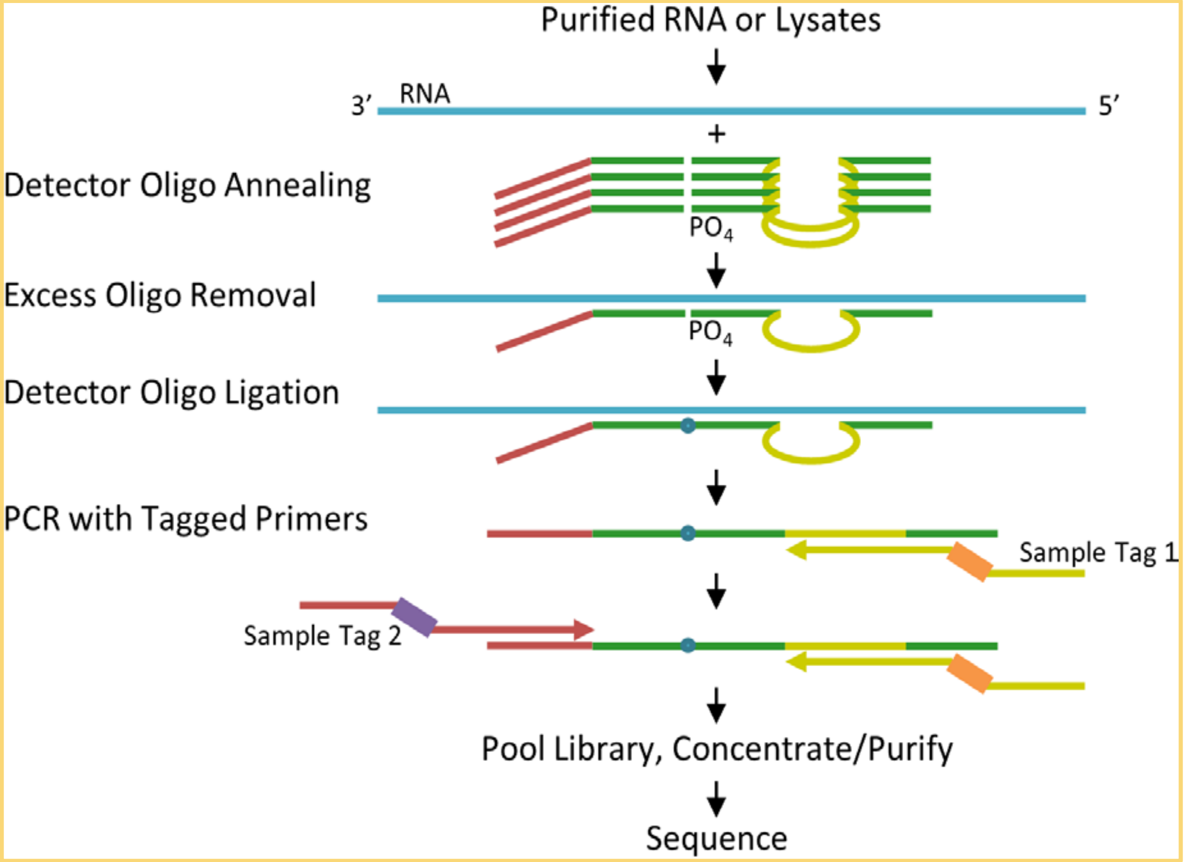


Lysis buffer





Pool of ~3,000 detector oligos
[NTP S1500+]



Yeakley et al. PLoS ONE 12(5): e0178302, 2017

Assay advantages:

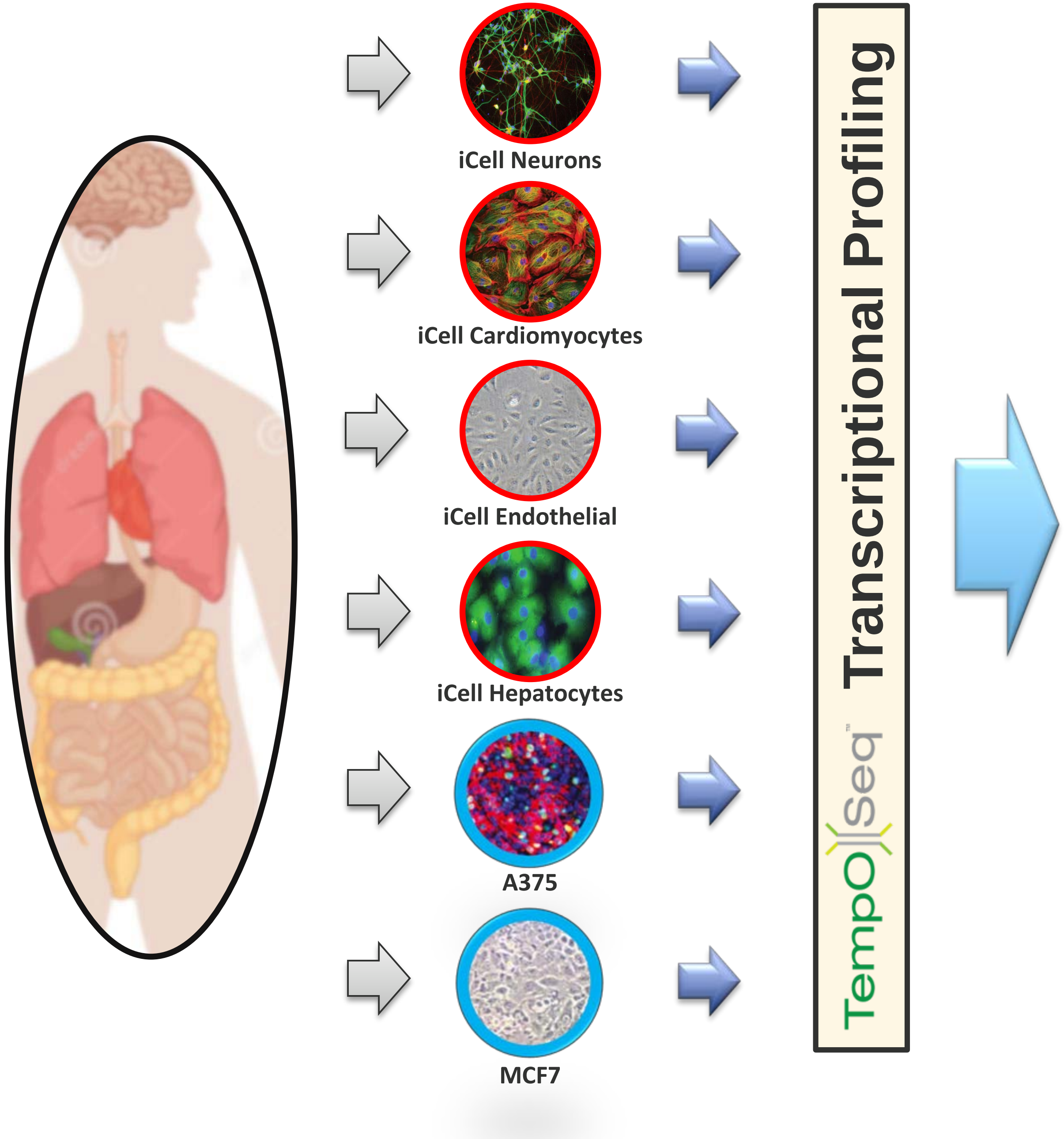
- Works in 384-wells
- No cDNA library prep
- 1000's genes/sample
- High Specificity (probe seq + ligation)

Assay considerations:

- Sequencing depth (per gene per sample)
- Gene selection (targeted set vs whole genome)
- New technology (no large database for comparisons)

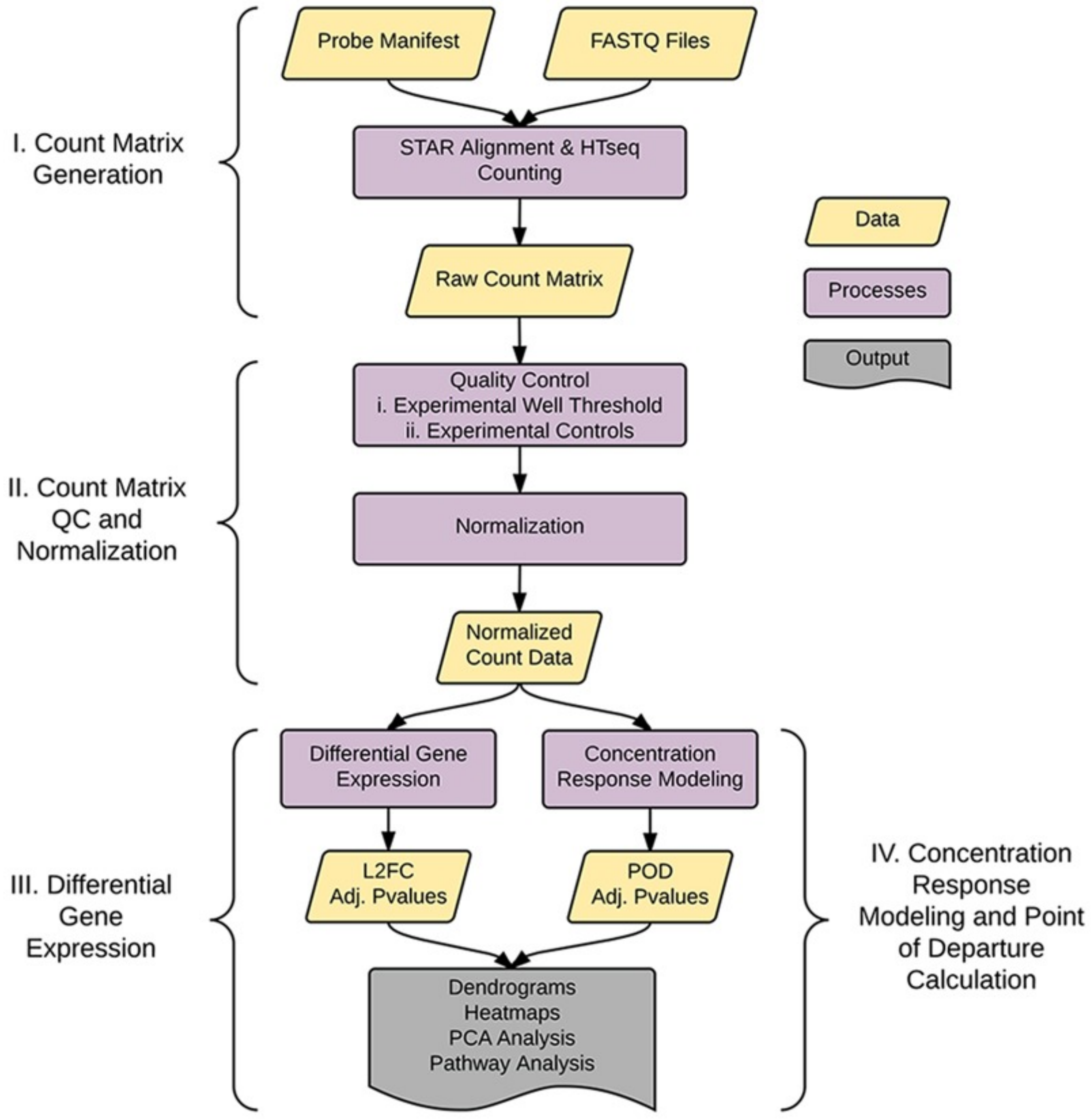
http://biorxiv.org/content/early/2016/01/07/036061

Cat-App Transcriptomics Data – 6 Human Cell Types



- 11,000+ samples from 6 cell types
- 4-point concentration response data
- Differential gene expression for ~3,000 transcripts (targeted analysis)
- Over 35,000,000 data points
- Novel data processing pipeline
- Concentration-response modeling pipeline
- Transcriptomics data can be combined with other data streams

Transcriptomic Data Analysis – Pipeline


I. Count Matrix Generation

Probe Manifest

FASTQ Files

STAR Alignment & HTseq Counting

Raw Count Matrix

II. Count Matrix QC and Normalization

Quality Control
i. Experimental Well Threshold
ii. Experimental Controls

Normalization

Normalized Count Data

III. Differential Gene Expression

Differential Gene Expression

L2FC Adj. Pvalues

Concentration Response Modeling

POD Adj. Pvalues

Dendrograms
Heatmaps
PCA Analysis
Pathway Analysis

IV. Concentration Response Modeling and Point of Departure Calculation

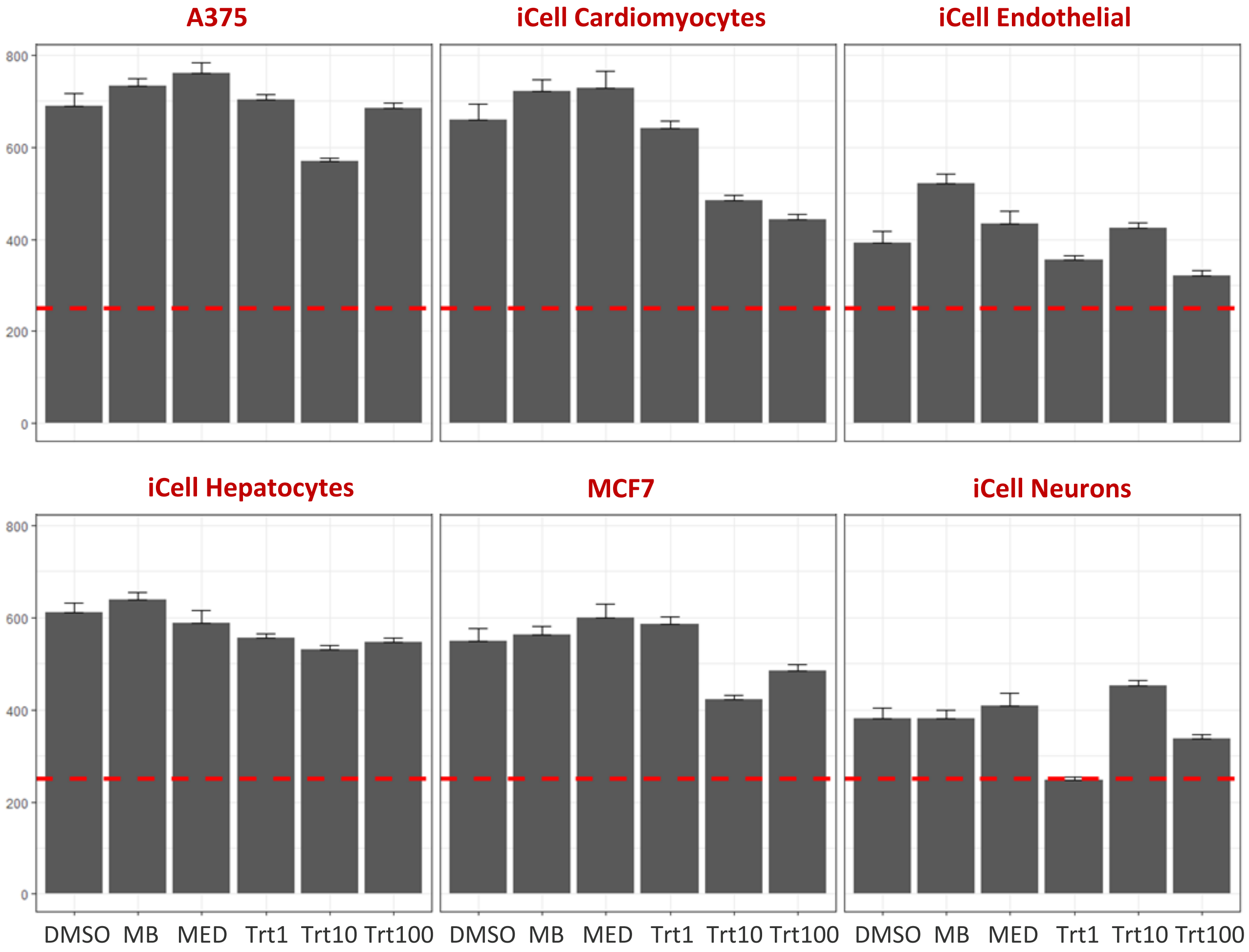
Data

Processes

Output

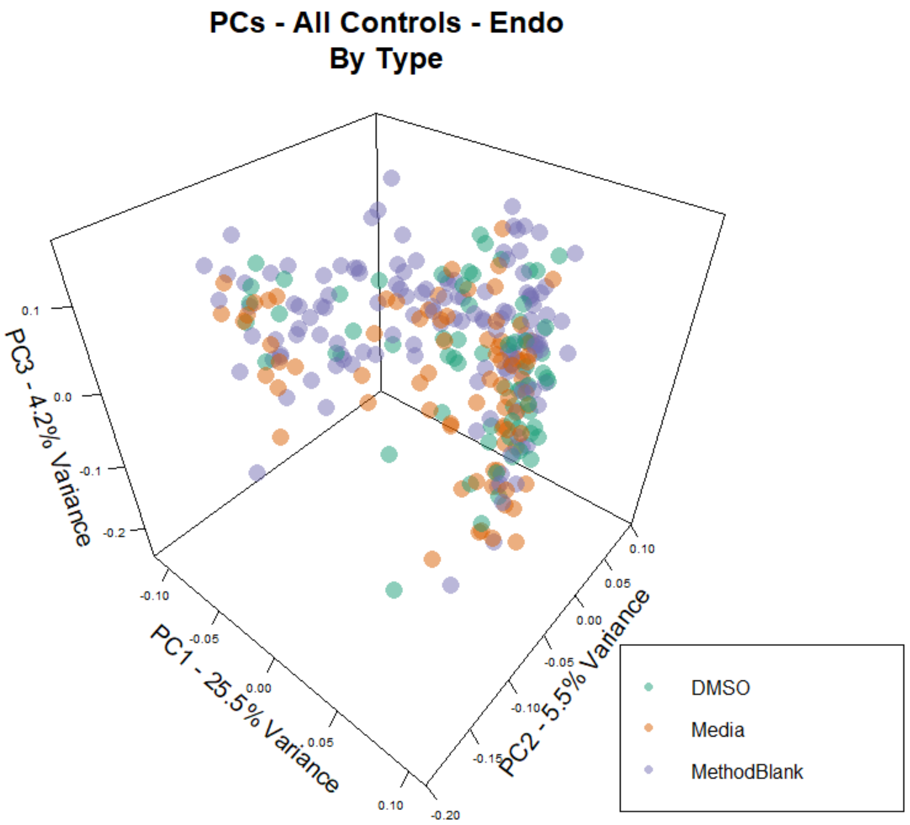
House et al. Front Genet 8:168, 2017

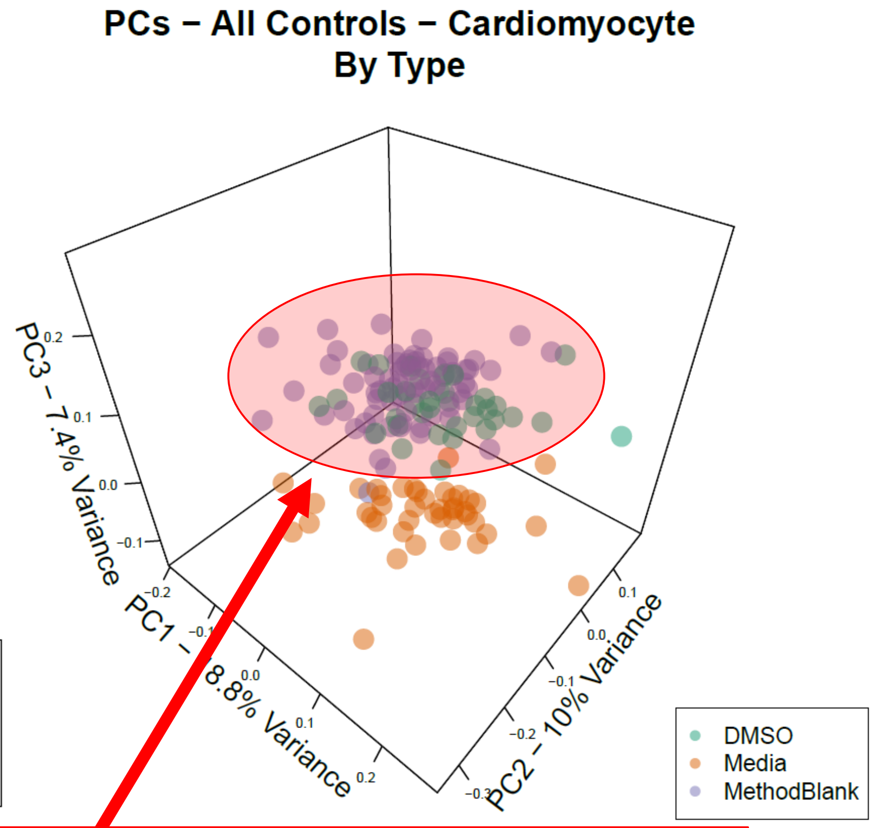
Average number of mapped reads per transcript



Transcriptomic Data Analysis – Quality Control

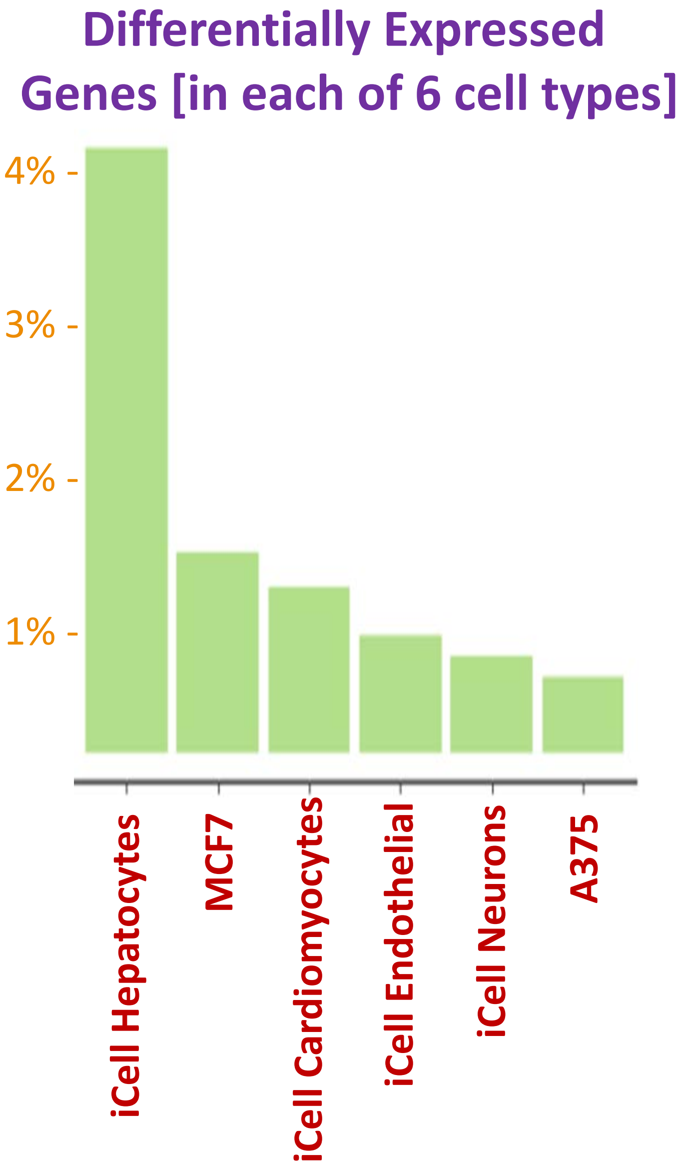
- Raw reads are de-multiplexed and mapped
- Minimum read count is set per gene
- Examination of controls (DMSO, Method blank, media)
- Examination of sequencing library quality



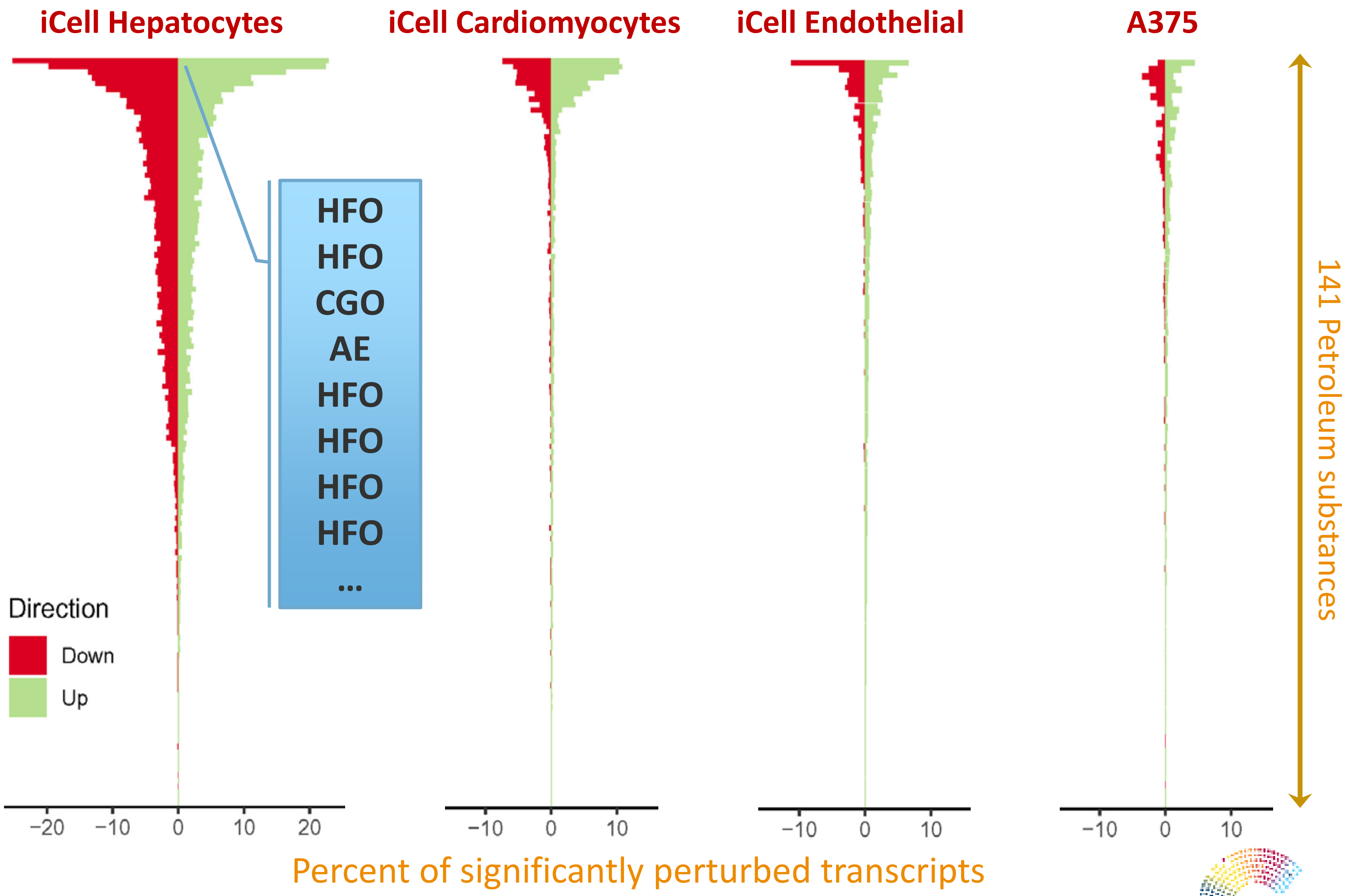


- “method blank” is not different from DMSO
- Vehicle is a proper comparator control for gene expression analysis

Transcriptomic Data Analysis – Effect of Petroleum Substances



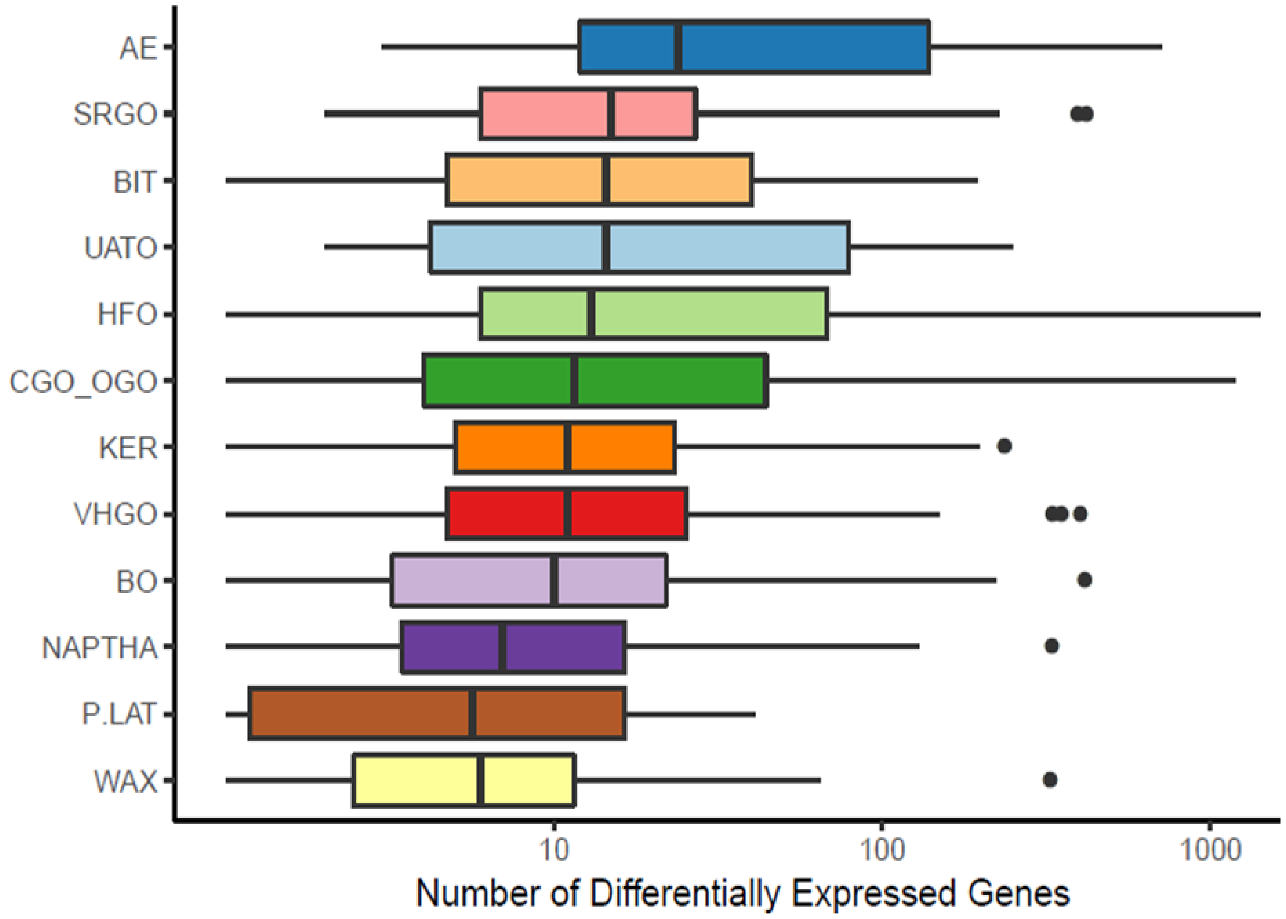
Differentially Expressed Genes [in each of 6 cell types]

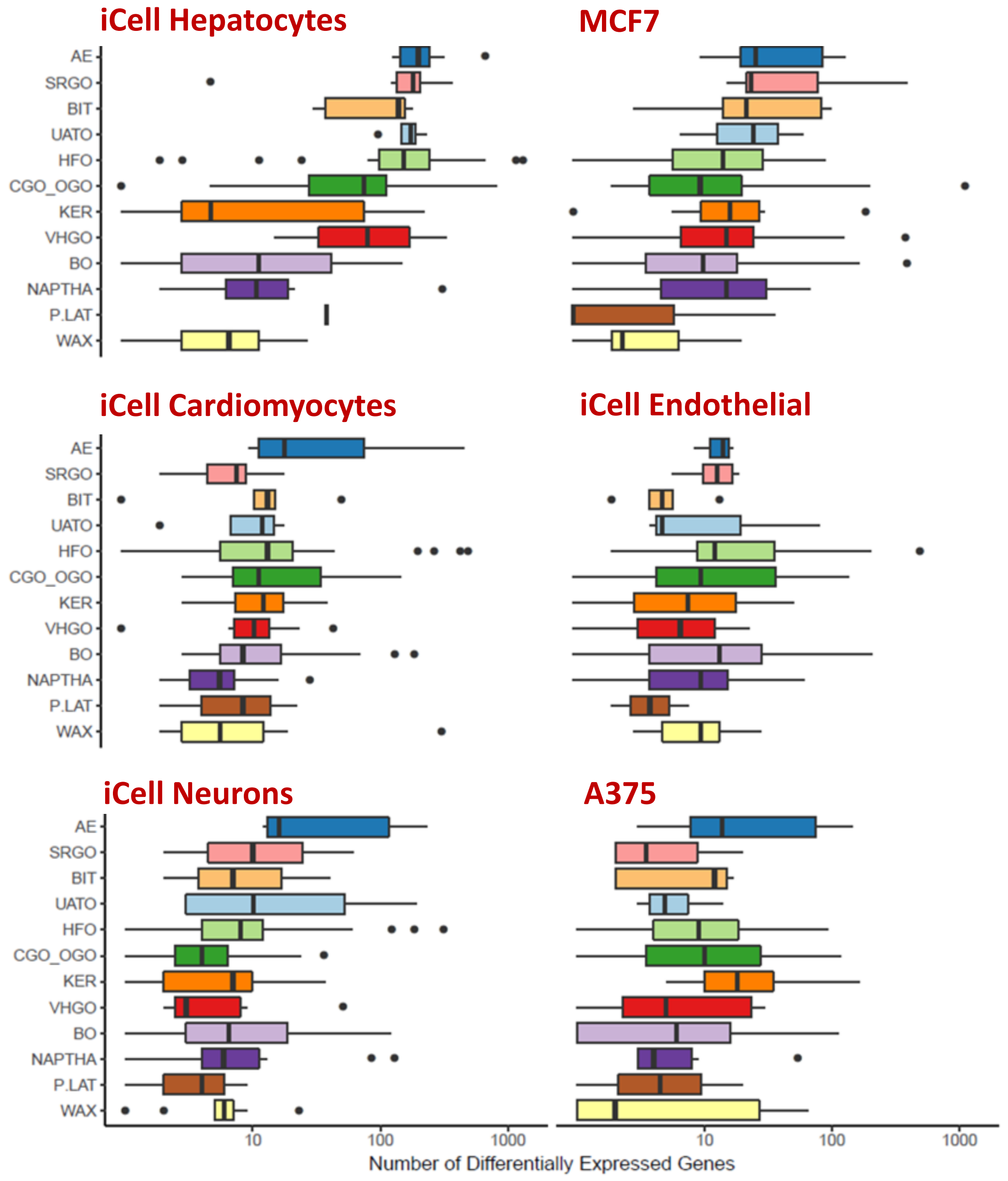


Percent of significantly perturbed transcripts

Transcriptomic Data Analysis – Effect of Petroleum Substances

Transcriptional “Activity “ [all 6 Cell Types]




iCell Hepatocytes

MCF7

iCell Cardiomyocytes

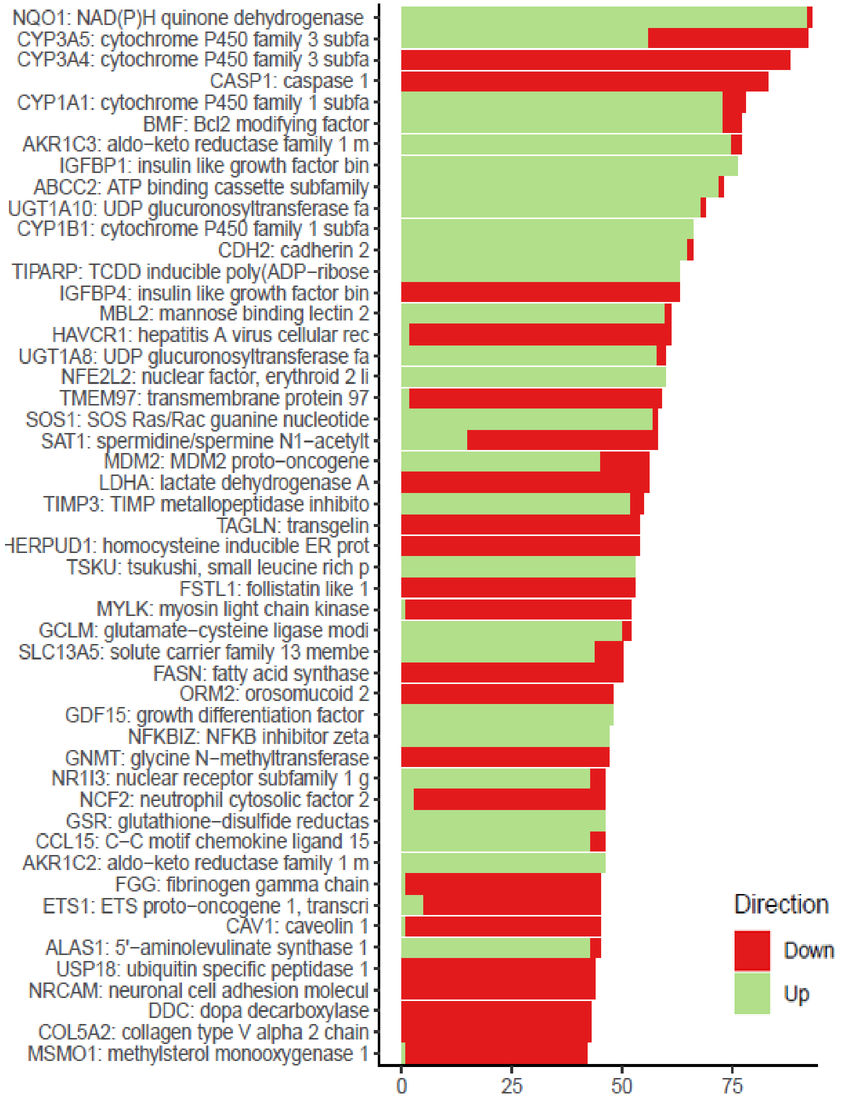
iCell Endothelial

iCell Neurons

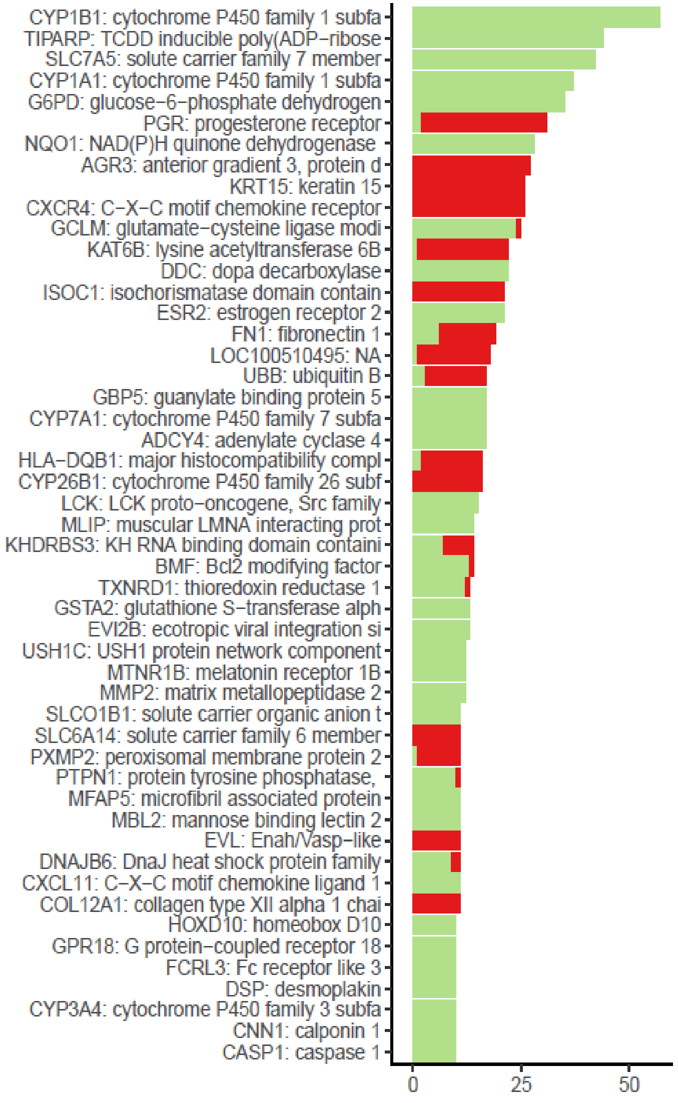
A375

Cell-specific Gene Expression Signature Across all Substances

iCell Hepatocytes

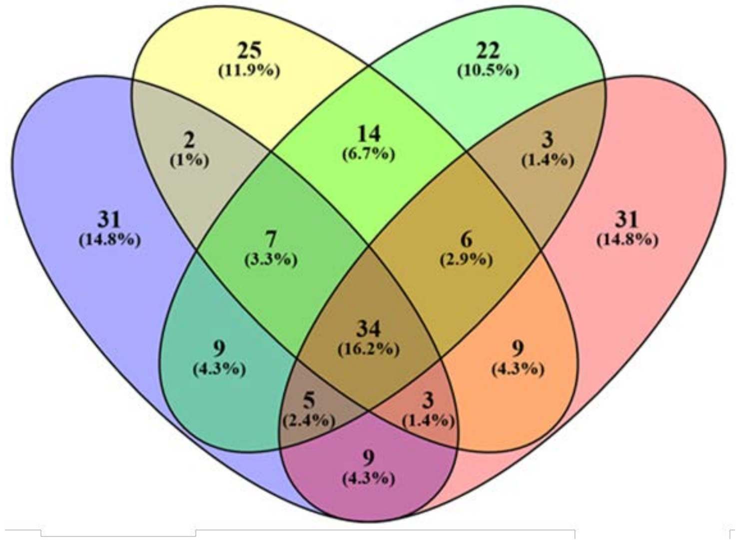


MCF7



Number of petroleum substances that “perturbed” each transcript [top 50 shown]

“Common” Response of 4 HFOs



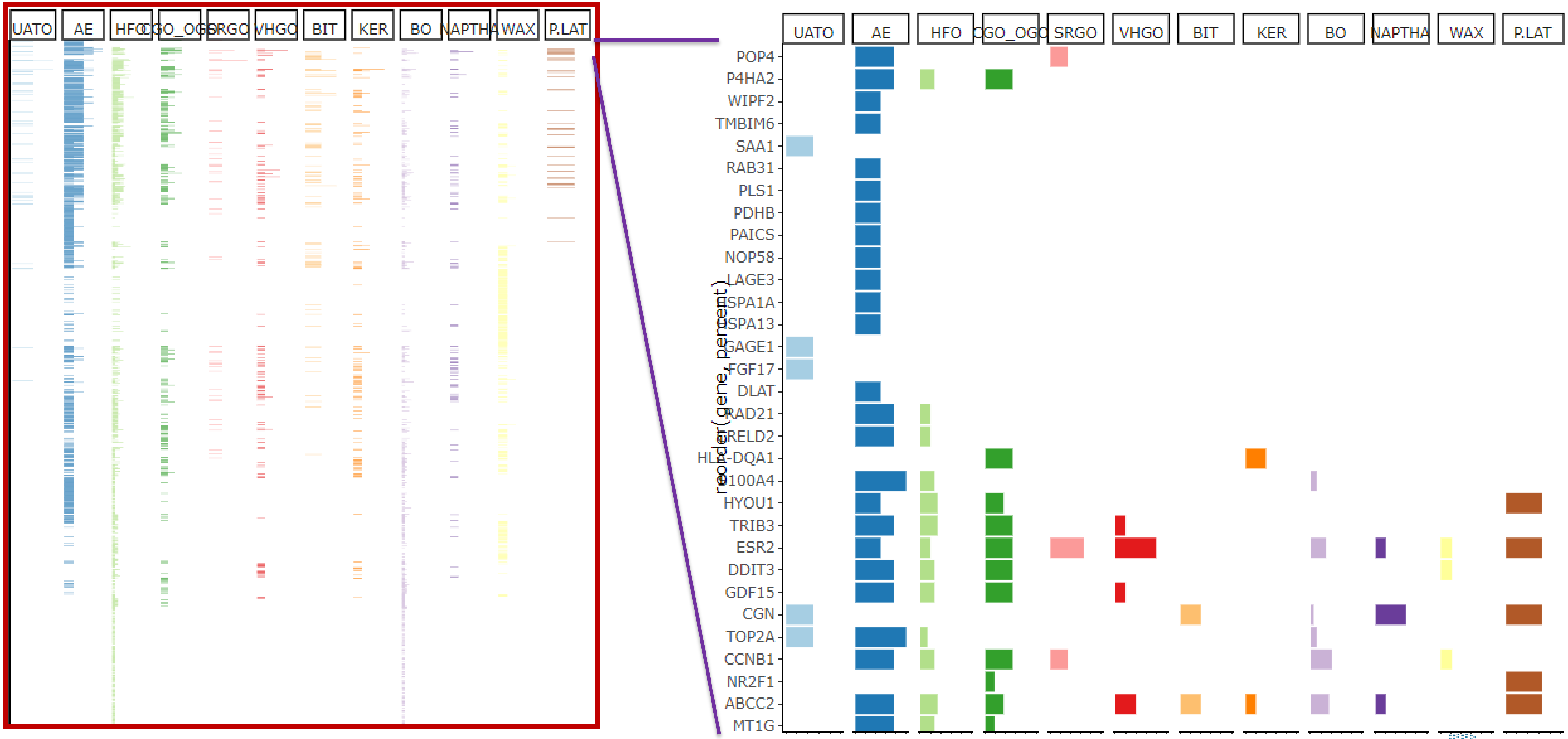
MBL2
CYP1A1
CYP1B1
UGT1A10
TIPARP
CDH2
UGT1A8
CYP3A5
TMEM97
ALAS1

HERPUD1
FSTL1
IGFBP4
NQO1
GDF15
TSKU
CYP3A4
MSMO1
SOS1
TIMP3

IGFBP1
BMF
CYP3A4
ACSL5
MYLK
USP18
EBP
LDHA
TAGLN
AKR1C3

Hepatocyte Gene Expression: Group-Specific “Signatures”?

Cardiomyocyte Gene Expression: Group-Specific “Signatures”?



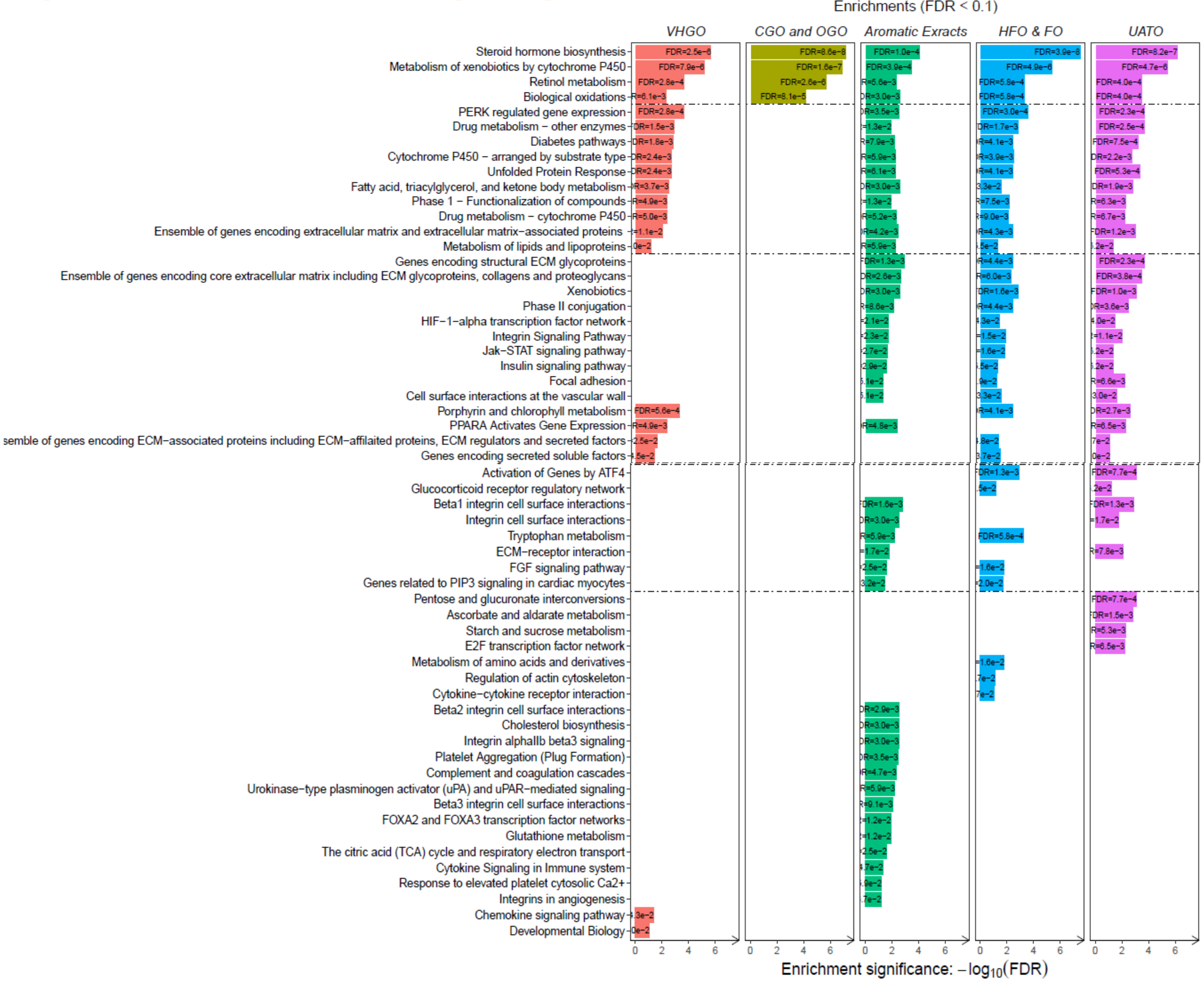
Hepatocyte Gene Expression: Group-Specific "Pathways"?

CYP450 oxidation

Other xenobiotic metabolism

Metabolism, kinase signaling

Cell surface receptors, PPARa

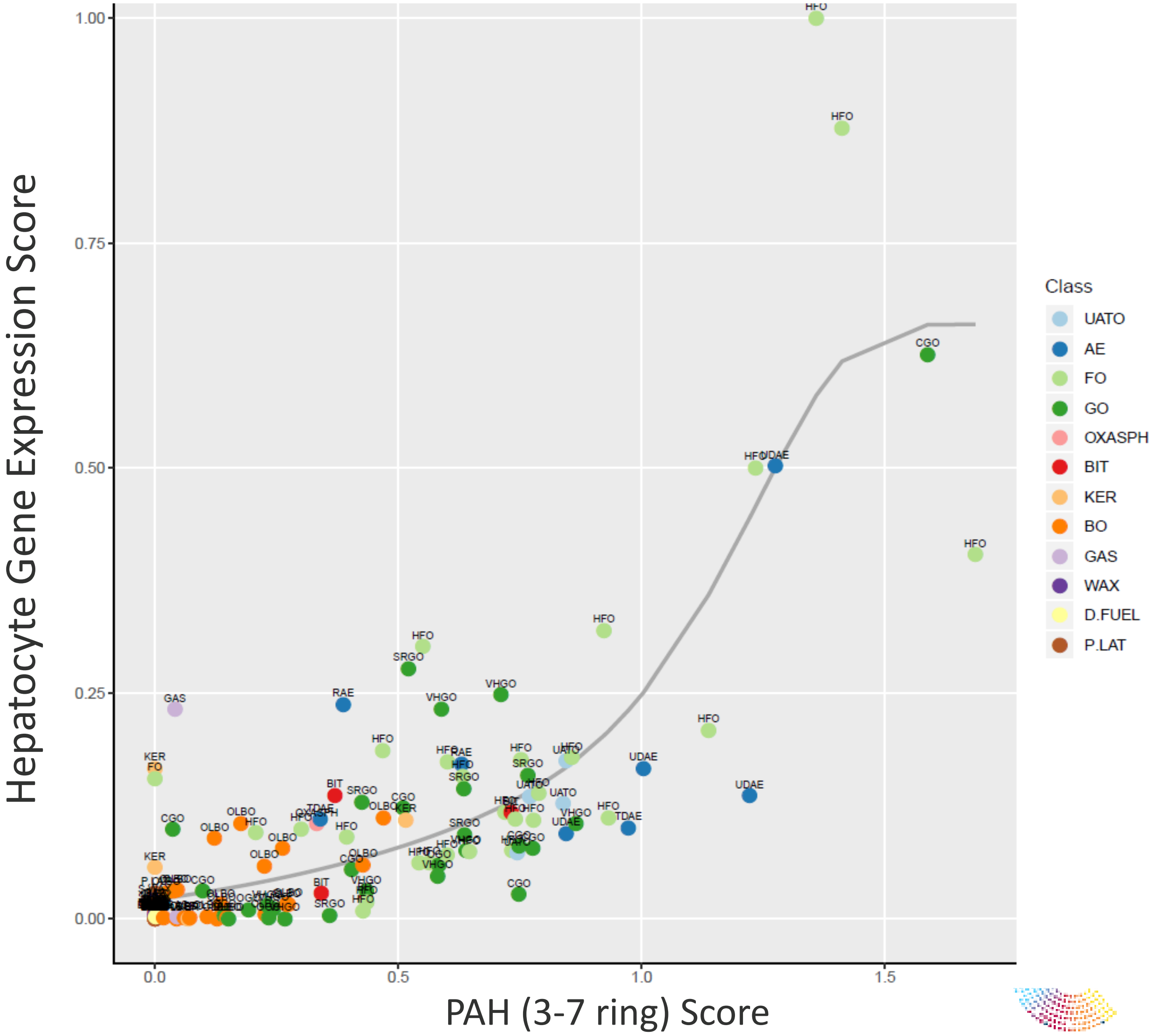


3-7 ring PAH hypothesis for Petroleum Substances

The percentage weight of 3-7 ring PAHs in the UVCB is the most active contributor to the bioactivity observed

Cell-specific Gene Expression Signature and PAH(3-7 ring)

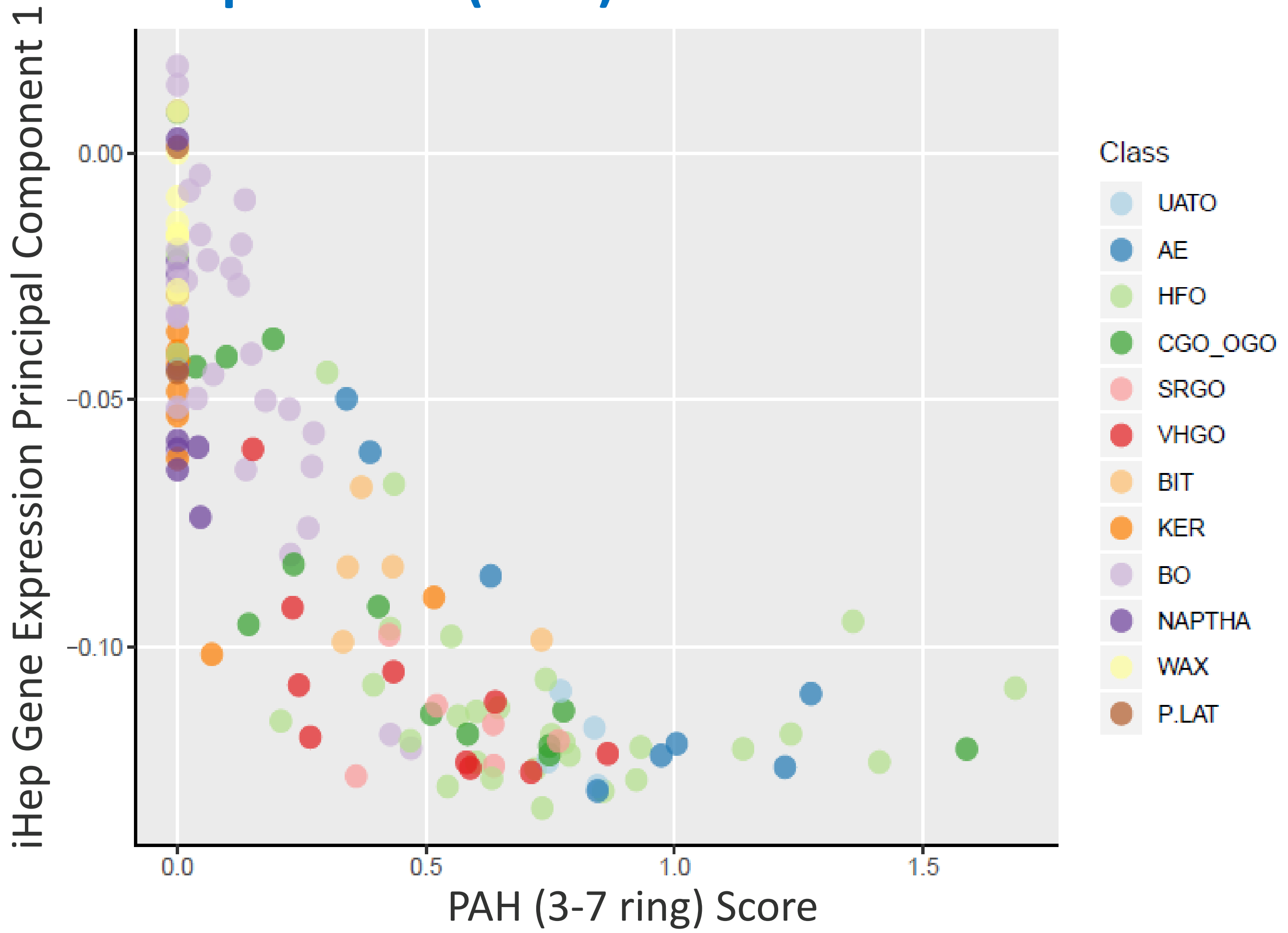
Cell type	Correlation of gene expression with PAH content
A375	0.24
iPSC CM	0.11
iPSC ENDO	0.18
iPSC HEP	**0.75**
MCF7	0.20
iPSC NEUR	0.09
All Cells	0.47



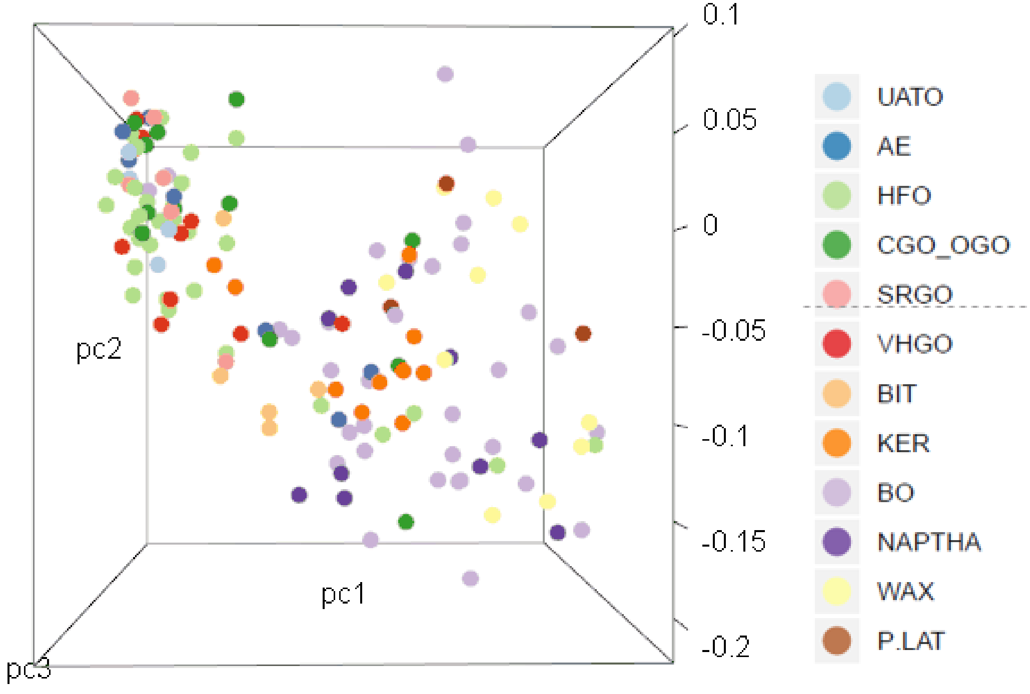
Relationship between Hepatocyte gene expression and PAH(3-7)

Almost all of gene expression signal in Hepatocytes correlates with PAH (3-7 ring) content

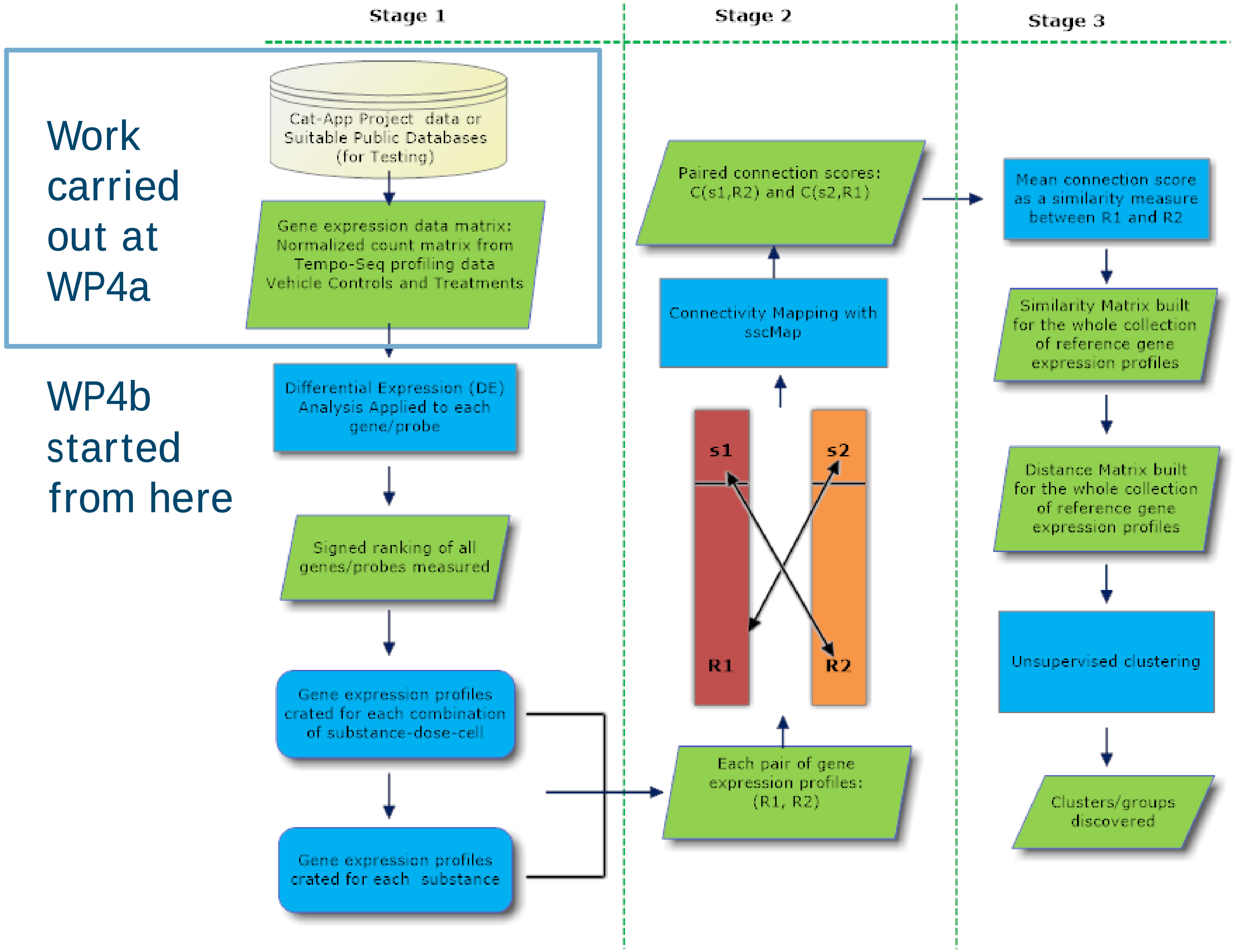
Spearman (rank) correlation = -0.85



Principal components (PC1-3) of the “Residual” gene expression



WP4b: Gene expression connectivity mapping based work flow



Stage 1

Stage 2

Stage 3

Work carried out at WP4a

WP4b started from here

Cat-App Project data or Suitable Public Databases (for Testing)

Gene expression data matrix: Normalized count matrix from Tempo-Seq profiling data Vehicle Controls and Treatments

Differential Expression (DE) Analysis Applied to each gene/probe

Signed ranking of all genes/probes measured

Gene expression profiles crated for each combination of substance-dose-cell

Gene expression profiles crated for each substance

Each pair of gene expression profiles: (R1, R2)

Connectivity Mapping with sscMap

Paired connection scores: C(s1,R2) and C(s2,R1)

Mean connection score as a similarity measure between R1 and R2

Similarity Matrix built for the whole collection of reference gene expression profiles

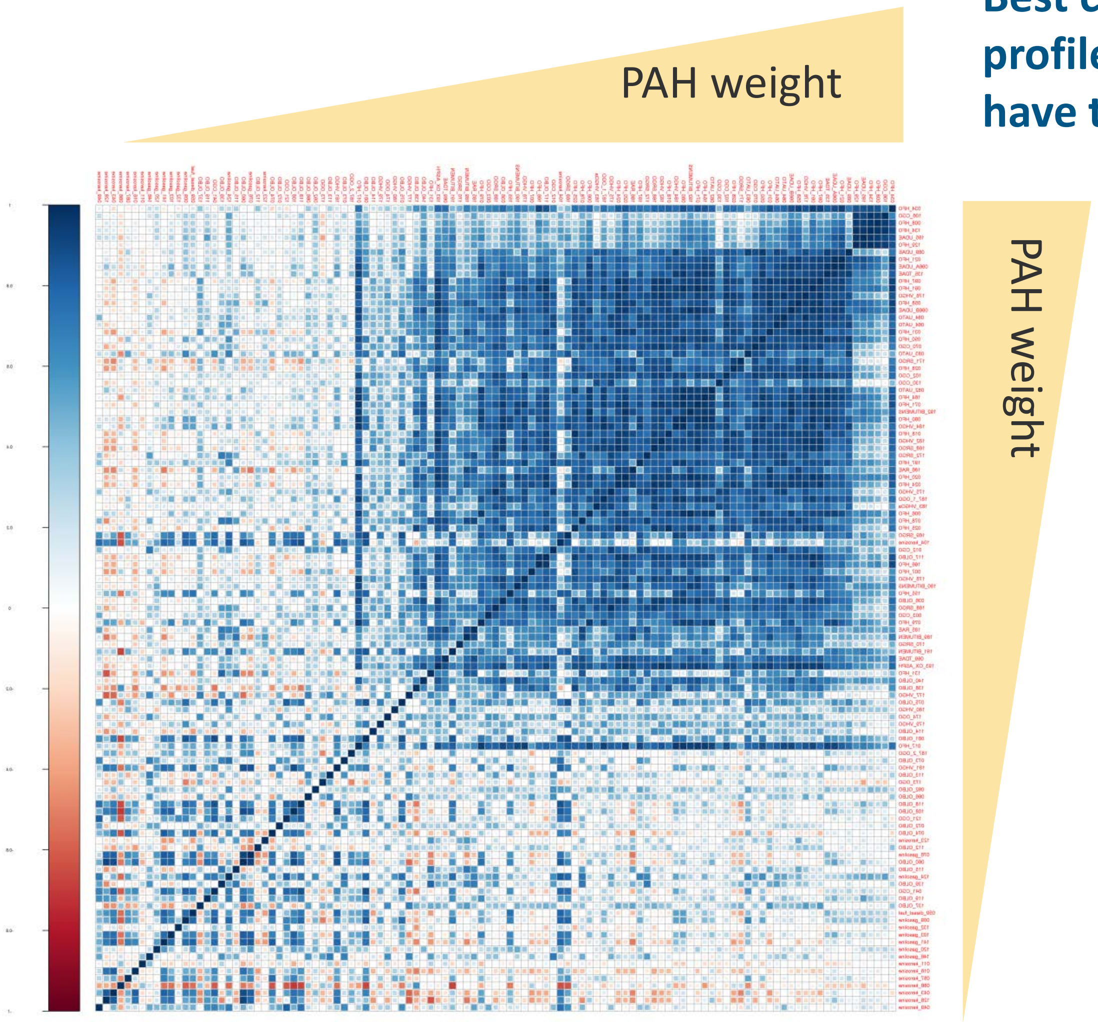
Distance Matrix built for the whole collection of reference gene expression profiles

Unsupervised clustering

Clusters/groups discovered

Heatmap of C-map scores for each UVCB, pairwise ordered by PAH weight

Best correlation between gene-expression profiles of the UVCBs occur with those that have the highest PAH weight



PAH weight

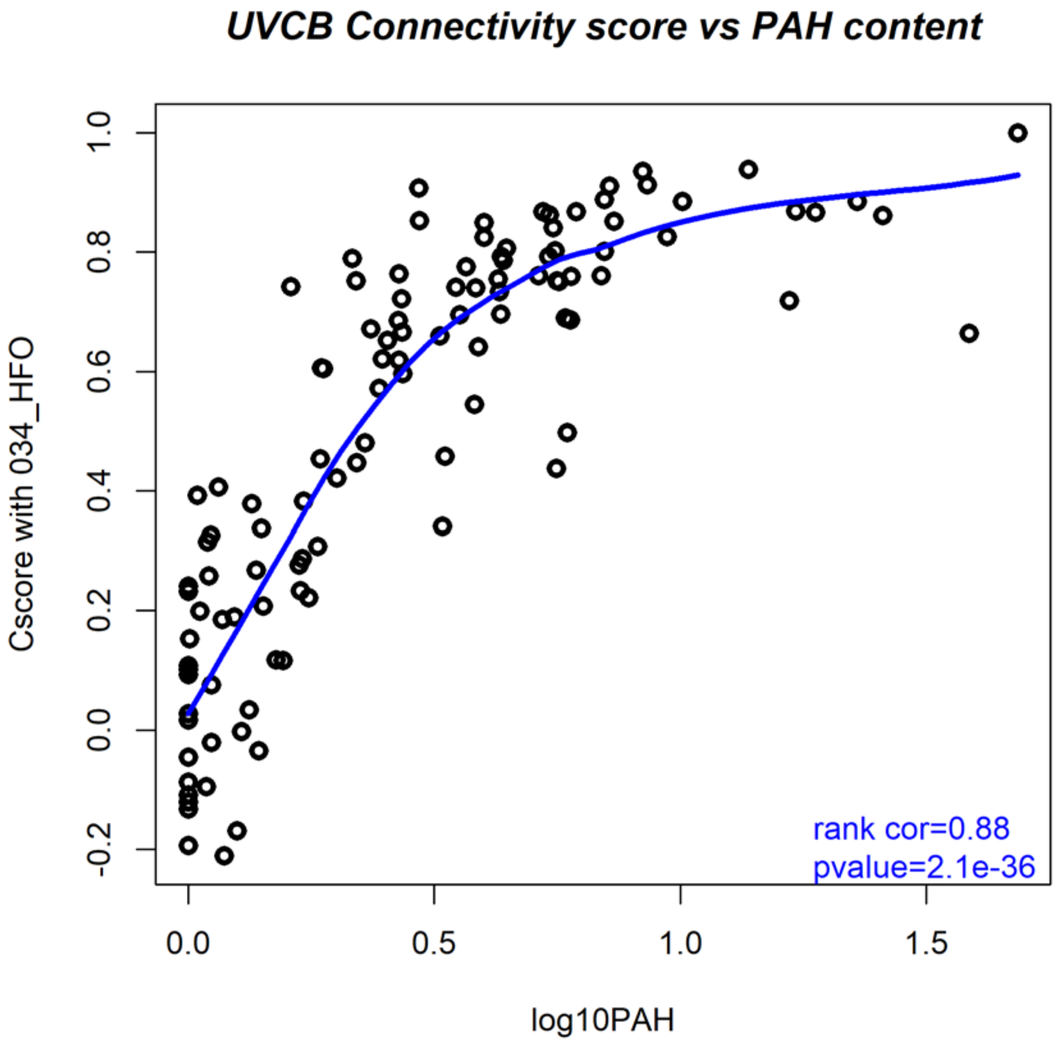
PAH weight

Heatmap of the correlation scores ordered by the log 10 PAH[(Wt% * (Ring 3-7 PAHs)+1]

Connectivity score example for HFO

Example for 034_HFO, which is the one with the highest 3-7 ring PAH content amongst the UVCBs (top right in the heatmap, previous slide)

This sample induces the highest number of differentially expressed genes



Overall conclusions on transcriptomics data

- Gene expression data are useful for elucidating mechanisms behind biological responses
- High throughput gene expression profiling methods are needed to handle the large number of samples and experimental conditions that are needed for grouping
- In this instance, the gene expression data provide support that 3-7 ring PAH content is driving most of the bioactivity
    - *Hepatocytes were most responsive (highest differential gene expression)*
    - *Highest expressed genes were involved in PAH metabolism related pathways*
    - *Other cell-types did not provide additional mechanistic information*
- This was confirmed by the correlation between 3-7 ring PAH content with gene expression
- Connectivity mapping are a useful tool for comparing multiple gene expression profiles from multiple UVCBs
    - *Gene expression profiles between substances with high PAH level are similar*
- Gene expression data were overall not as informative as cell-based endpoint measurement
    - *Generated transcriptomics data are still informative to add further support to the PAH hypothesis and adding a mechanistic component in combination with other data*

Thank you for your attention

Shu-Dong Zhang

sd.zhang@ulster.ac.uk

Cat-App

www.concawe.eu/cat-app