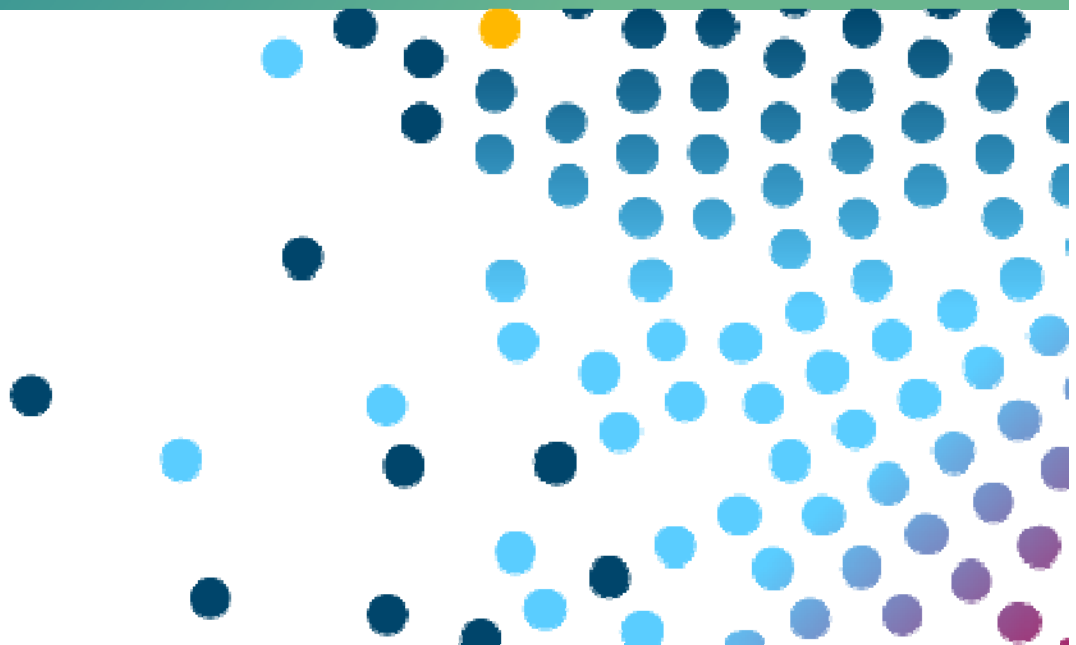
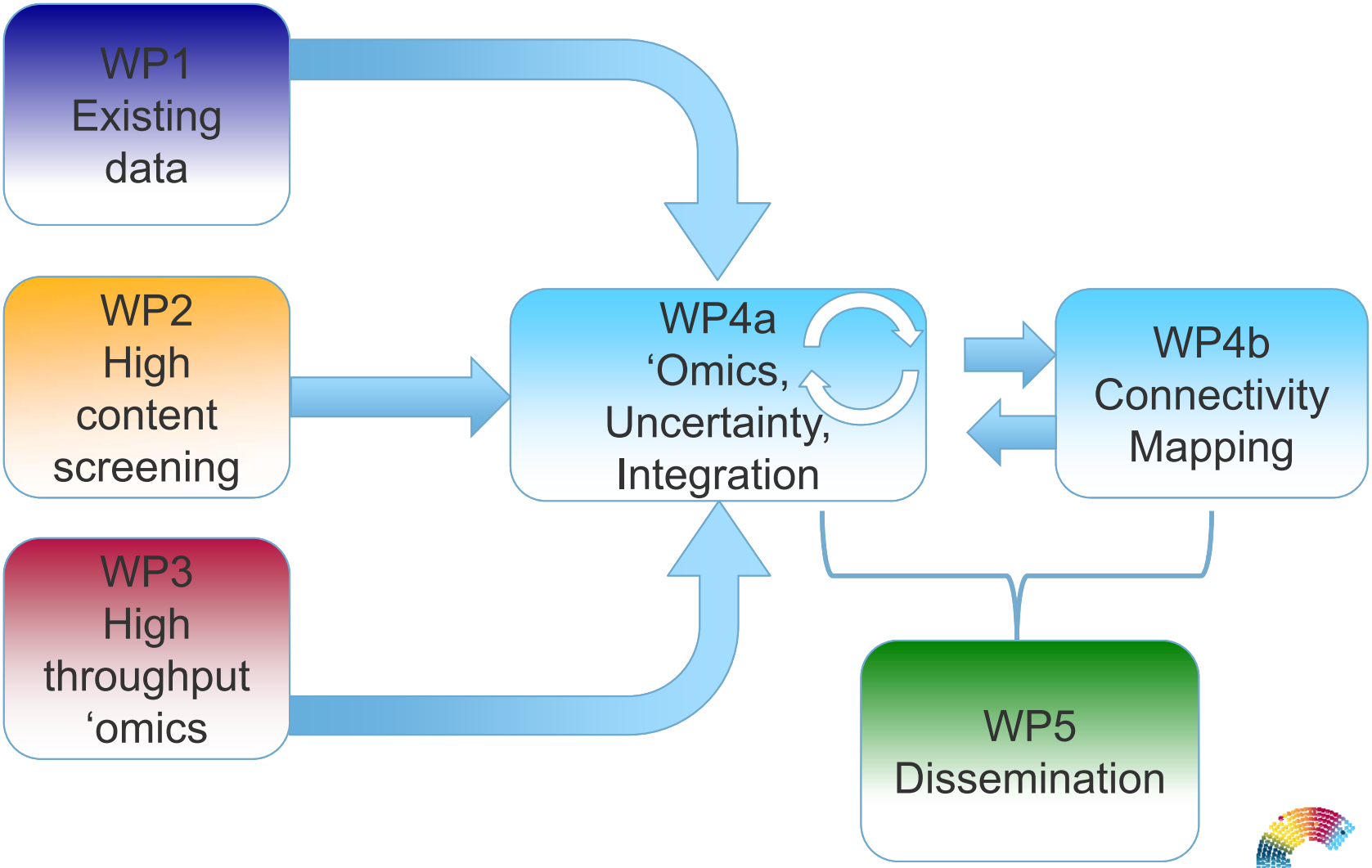
Data Science: Integrative Data Analysis and Visualization

Fred A. Wright
North Carolina State University

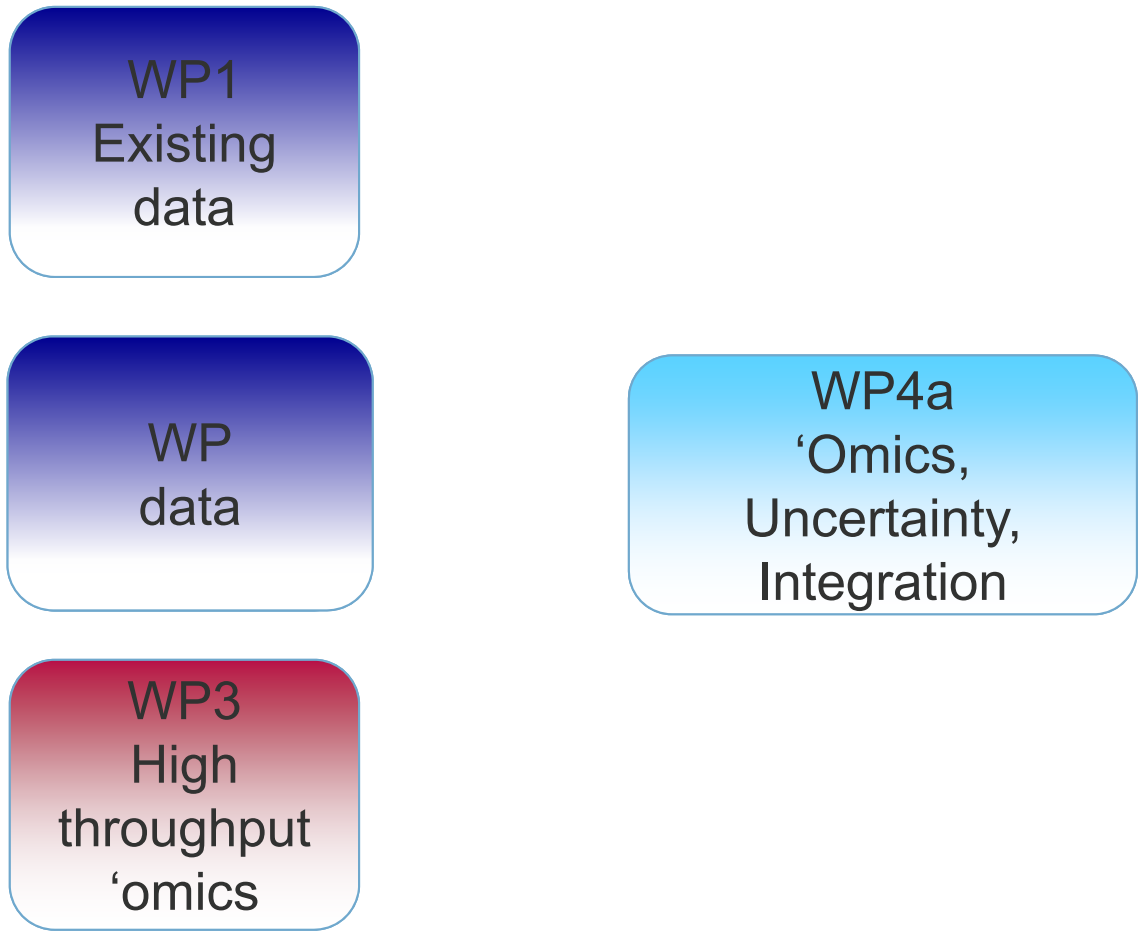


4a.1 | A workflow for data management

Workflow


WP1
Existing
data
WP2
High
content
screening
WP3
High
throughput
‘omics
WP4a
‘Omics,
Uncertainty,
Integration
WP4b
Connectivity
Mapping
WP5
Dissemination

Workflow


WP1
Existing
data
WP
data
WP3
High
throughput
‘omics
WP4a
‘Omics,
Uncertainty,
Integration

Workflow

WP data

DATA DOWNLOAD

- Existing data
  - Manufacturing information
  - Phys/chem properties
  - Analytic Chemistry
  - Existing toxicity data
- High-content screening info
- Transcriptomic data
  - Probe manifests
  - Hash files
  - Sequence files

WP4a ‘Omics, Uncertainty, Integration

Workflow

WP data

DATA UPLOAD

- Sequence count matrices
- P-values
- Points of departure
- Uncertainty estimates
- Various checks and logs
- ToxPi results

WP4a ‘Omics, Uncertainty, Integration

Workflow

WP data

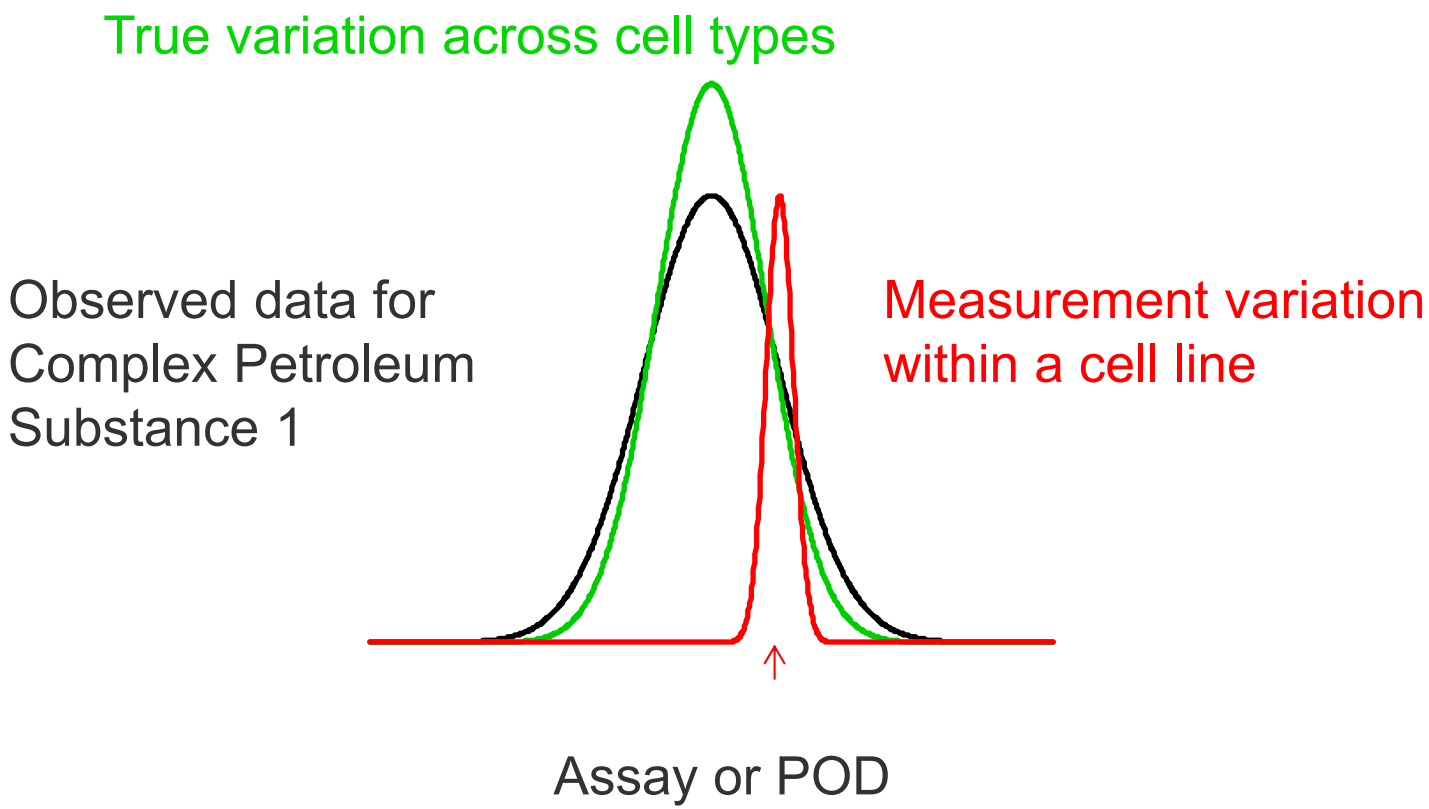
INTERFACE

- Current version based on static web pages and pulldown (WP1)
- Discussions of automated/scripted pulldown
- Possible R API Interface for portions of data that have common format

WP4a ‘Omics, Uncertainty, Integration

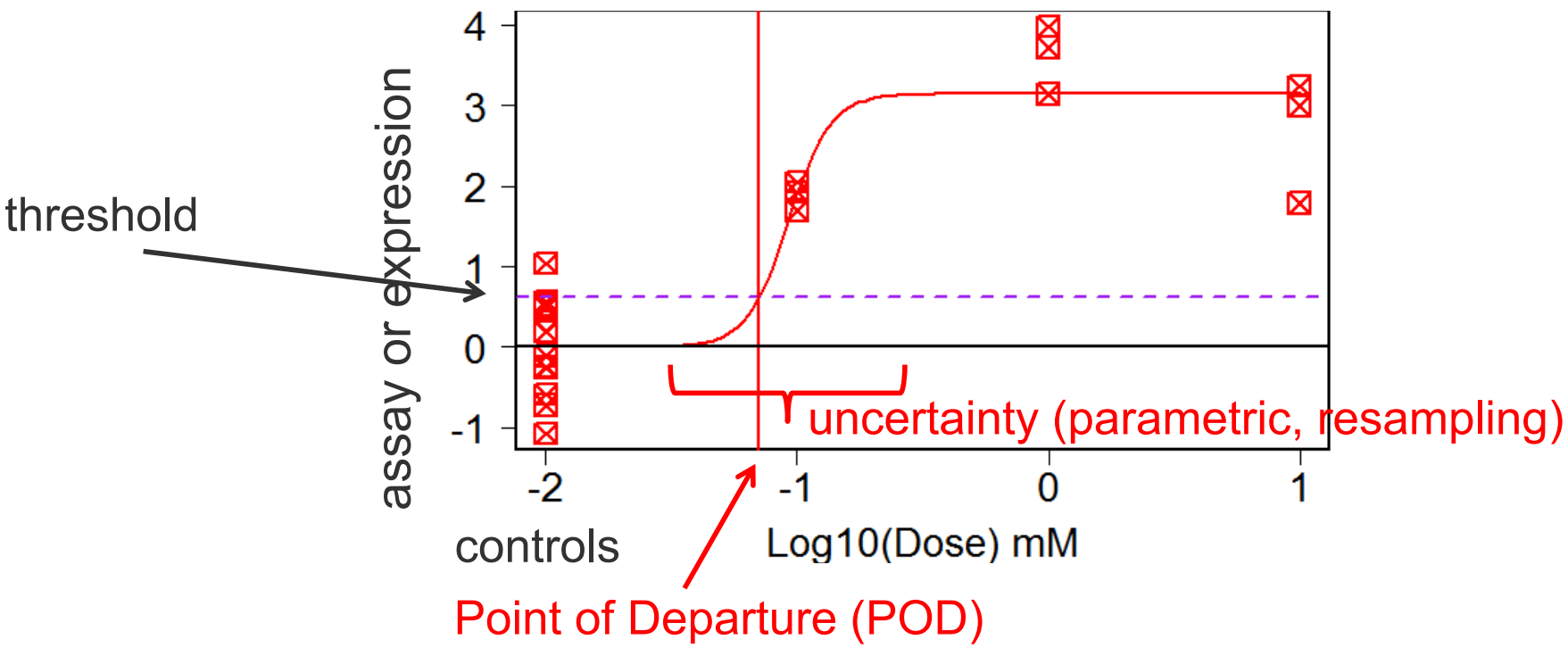
4a.2 Uncertainty and variability analyses

Variability: the general concept for cell line studies

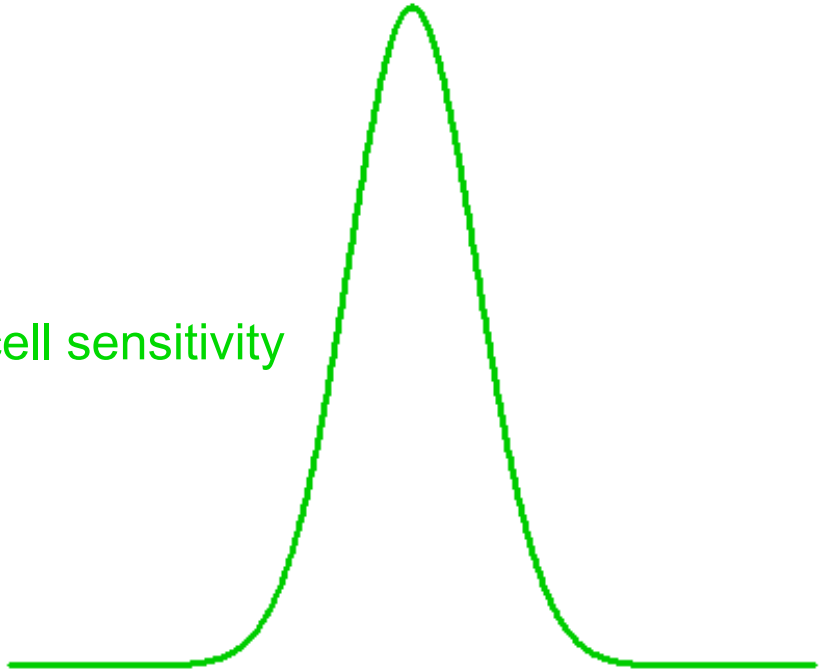


Uncertainty

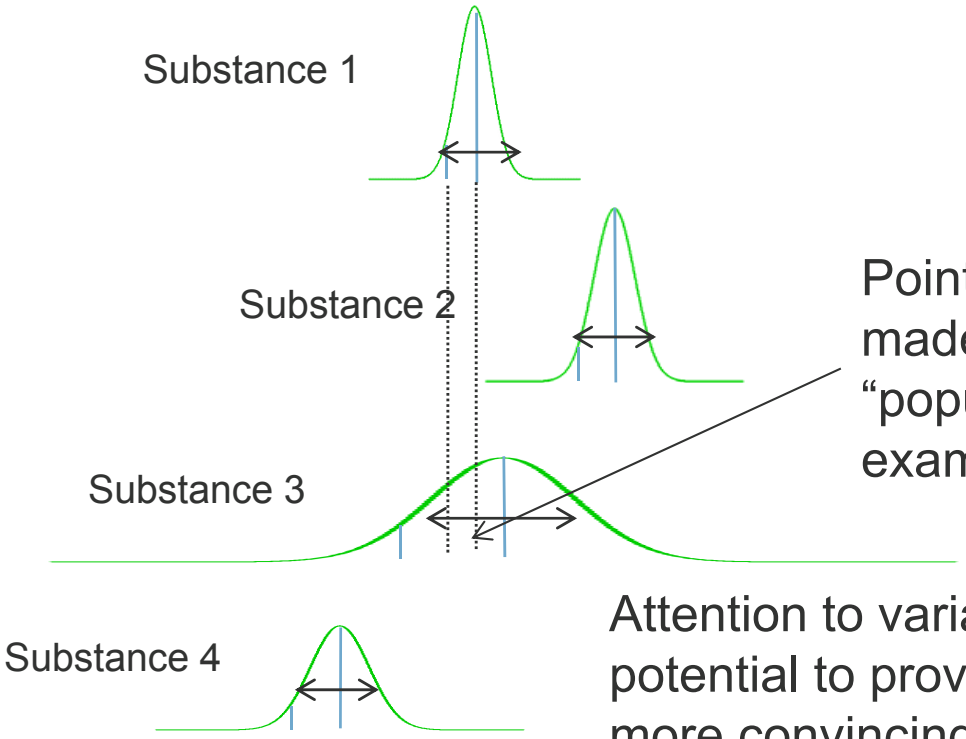
1 cell line, one substance



Variation in cell sensitivity



Cell line variability across substances



Points-of-departure can be made relative to the "population" of cell types examined

Attention to variability has the potential to provide improved and more convincing read-across assessments

Assessment of variability from limited data

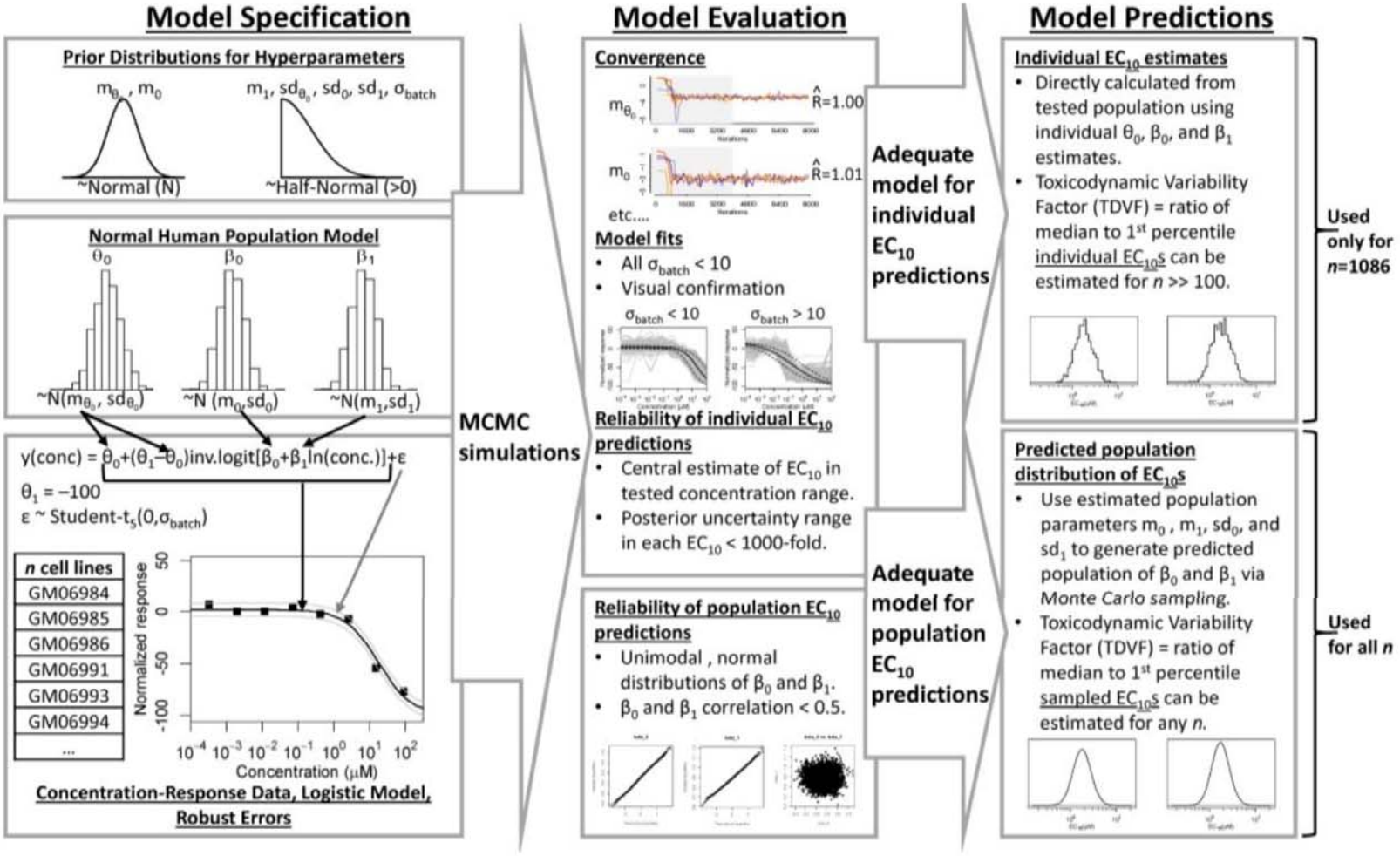
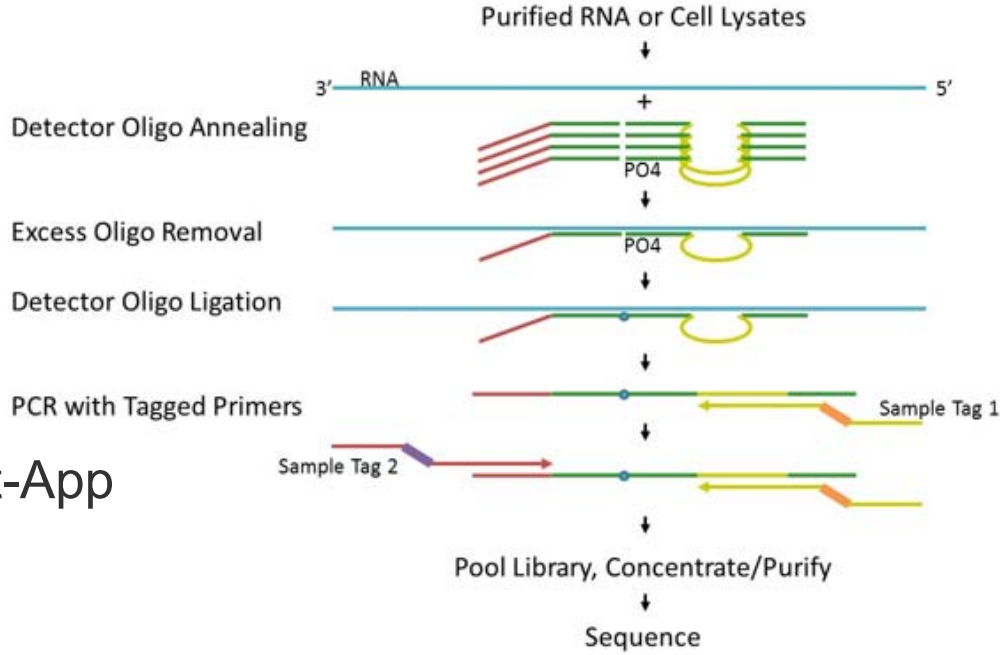


Fig. 1: Bayesian modeling, evaluation, and prediction workflow

Chiu et al. ALTEX. 2016 Dec 13. doi: 10.14573/altex.160825

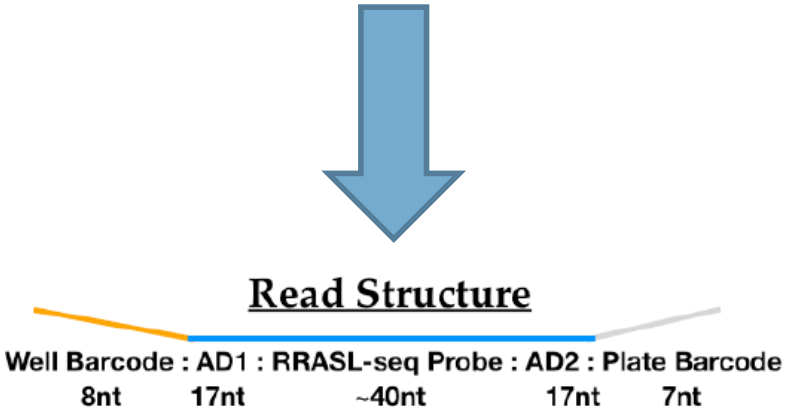
4a.3 | Processing ‘omics data

tempOseq (BioSpyder)



- ~2800 genes/sample for Cat-App
- High specificity
- No variance in product size
- Less expensive than whole-transcriptome RNA Sequencing

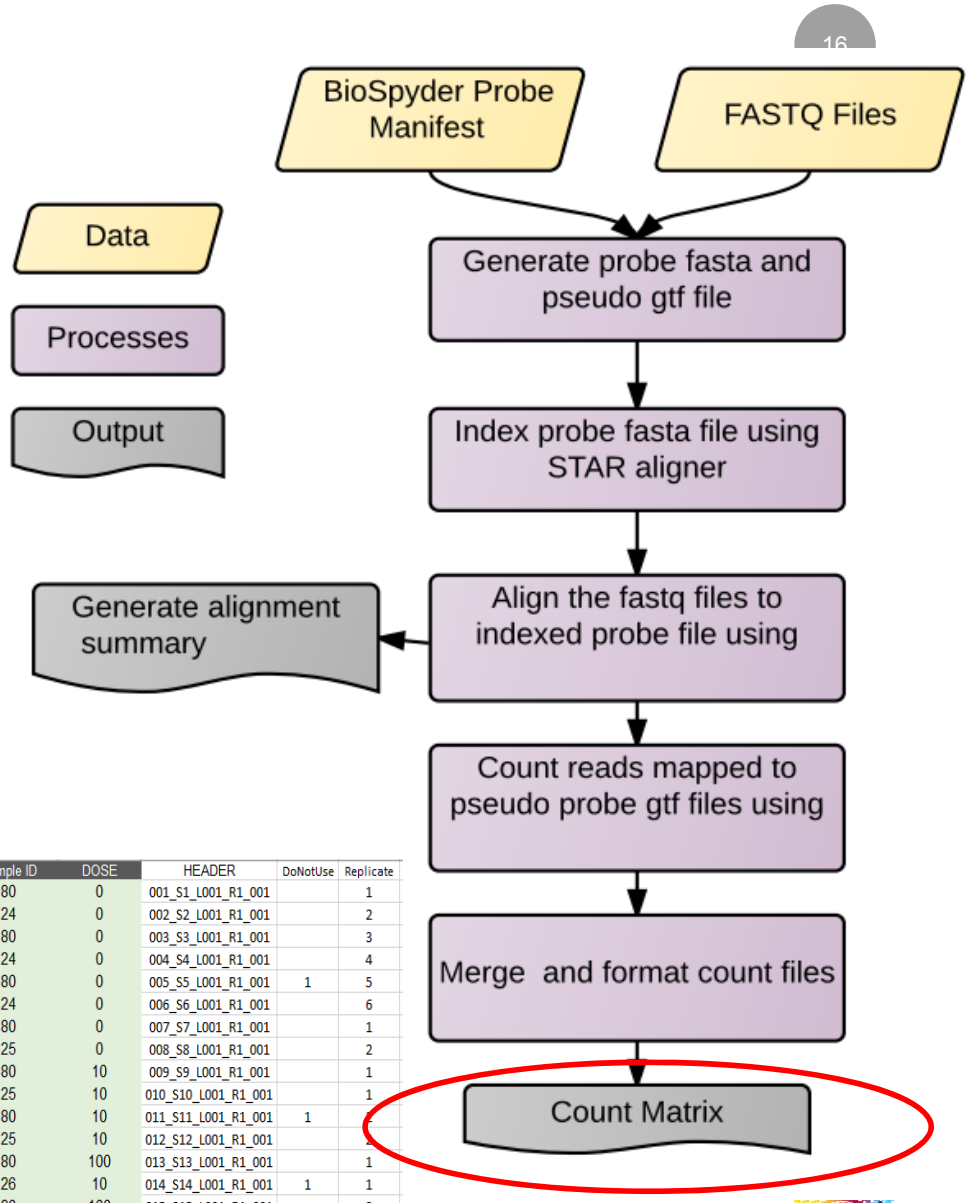
However, a new robust bioinformatics pipeline is needed



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

Pipeline I. `temposeqcount`

- Publicly available on github
- Cleans/aligns demultiplexed reads
- Fast!
- Output – raw count matrix into rest of pipeline



Experimental Layout File (Hash)

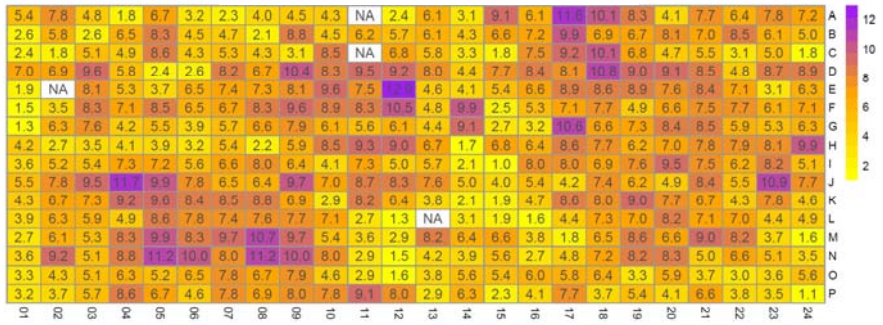
PlateID	FW Primer ID	RV Primer ID	Index Set	96wp position	Index Sequence 1	Index Sequence 2	Treatment	Sample ID	DOSE	HEADER	DoNotUse	Replicate
AA01	F801	R801	A	A01	AAGACTCTT	AAGGTGTTT	MEDIAA	80	0	001_S1_L001_R1_001		1
CA01	F801	R802	C	A01	GCGATGATT	AAGGTGTTT	MEDIAA	24	0	002_S2_L001_R1_001		2
AA02	F801	R803	A	A02	TTGAGTTGT	AAGGTGTTT	MEDIAA	80	0	003_S3_L001_R1_001		3
CA02	F801	R804	C	A02	GACTTCATT	AAGGTGTTT	MEDIAA	24	0	004_S4_L001_R1_001		4
AA03	F801	R805	A	A03	AACACGTCT	AAGGTGTTT	MEDIAA	80	0	005_S5_L001_R1_001	1	5
CA03	F801	R806	C	A03	GTTTATACT	AAGGTGTTT	MEDIAA	24	0	006_S6_L001_R1_001		6
AA04	F801	R807	A	A04	GTATTATTG	AAGGTGTTT	NOCELLS	80	0	007_S7_L001_R1_001		1
CA04	F801	R808	C	A04	TAACGCGTC	AAGGTGTTT	NOCELLS	25	0	008_S8_L001_R1_001		2
AA05	F801	R809	A	A05	AACATACTG	AAGGTGTTT	pirfinidone	80	10	009_S9_L001_R1_001		1
CA05	F801	R810	C	A05	GCAGATGCA	AAGGTGTTT	Nifedipine	25	10	010_S10_L001_R1_001		1
AA06	F801	R811	A	A06	AGATAGCCT	AAGGTGTTT	pirfinidone	80	10	011_S11_L001_R1_001	1	
CA06	F801	R812	C	A06	GAACTGTAT	AAGGTGTTT	Nifedipine	25	10	012_S12_L001_R1_001		2
AA07	F801	R813	A	A07	GGGCAGATC	AAGGTGTTT	pirfinidone	80	100	013_S13_L001_R1_001		1
CA07	F801	R814	C	A07	TCCAGTATC	AAGGTGTTT	Azithromycin	26	10	014_S14_L001_R1_001	1	1
AA08	F801	R815	A	A08	TAAGGTAGC	AAGGTGTTT	pirfinidone	80	100	015_S15_L001_R1_001		2
CA08	F801	R816	C	A08	TCCTCAGCC	AAGGTGTTT	Azithromycin	26	10	016_S16_L001_R1_001	1	2
AA09	F801	R817	A	A09	AAGGATACC	AAGGTGTTT	pirfinidone	80	1000	017_S17_L001_R1_001		1
CA09	F801	R818	C	A09	GTTTGGCAC	AAGGTGTTT	Azithromycin	26	100	018_S18_L001_R1_001		1
AA10	F801	R819	A	A10	GGGTTGTTA	AAGGTGTTT	pirfinidone	80	1000	019_S19_L001_R1_001		2

Pipeline II. Quality Control

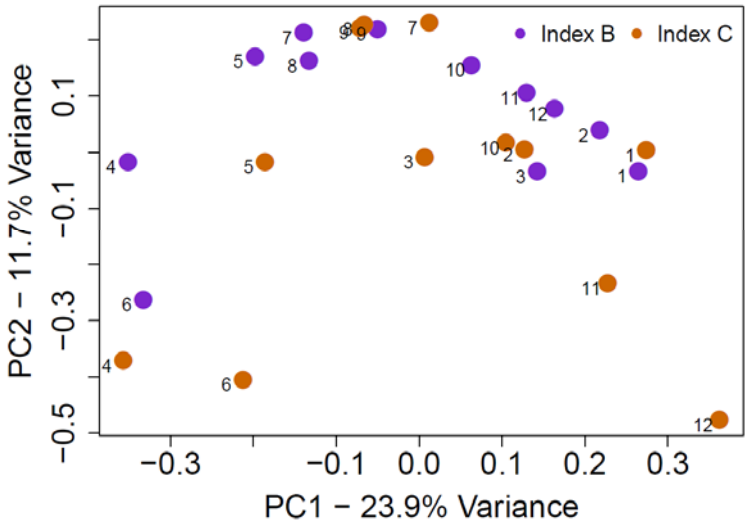
R and DESeq2

- Low well counts
- Other technical issues
- Zero count features
- Experiment-wide normalization

Expression by well



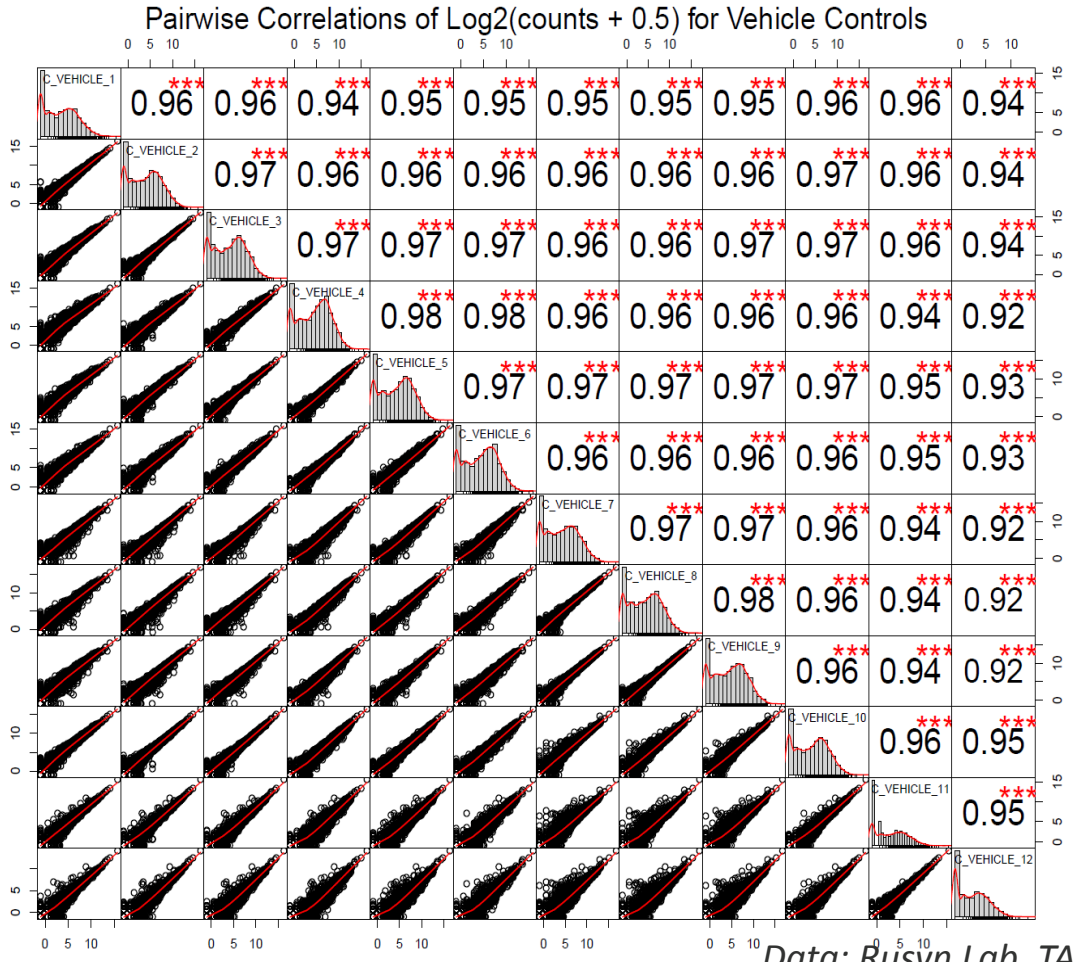
Vehicle Control Principal Components



Data: Rusyn Lab, TAMU

Are the expression data reproducible?

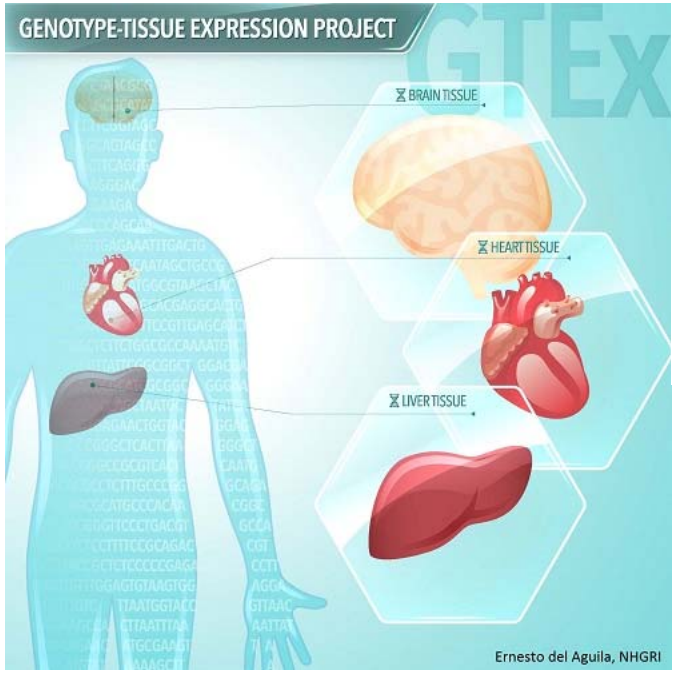
Example data: cardiomyocyte control replicates



Data: Rusyn Lab, TAMU

Are the expression data organotypic?

The NIH Common Fund GTEx project
53 human tissues, over 7000 RNA-Seq samples



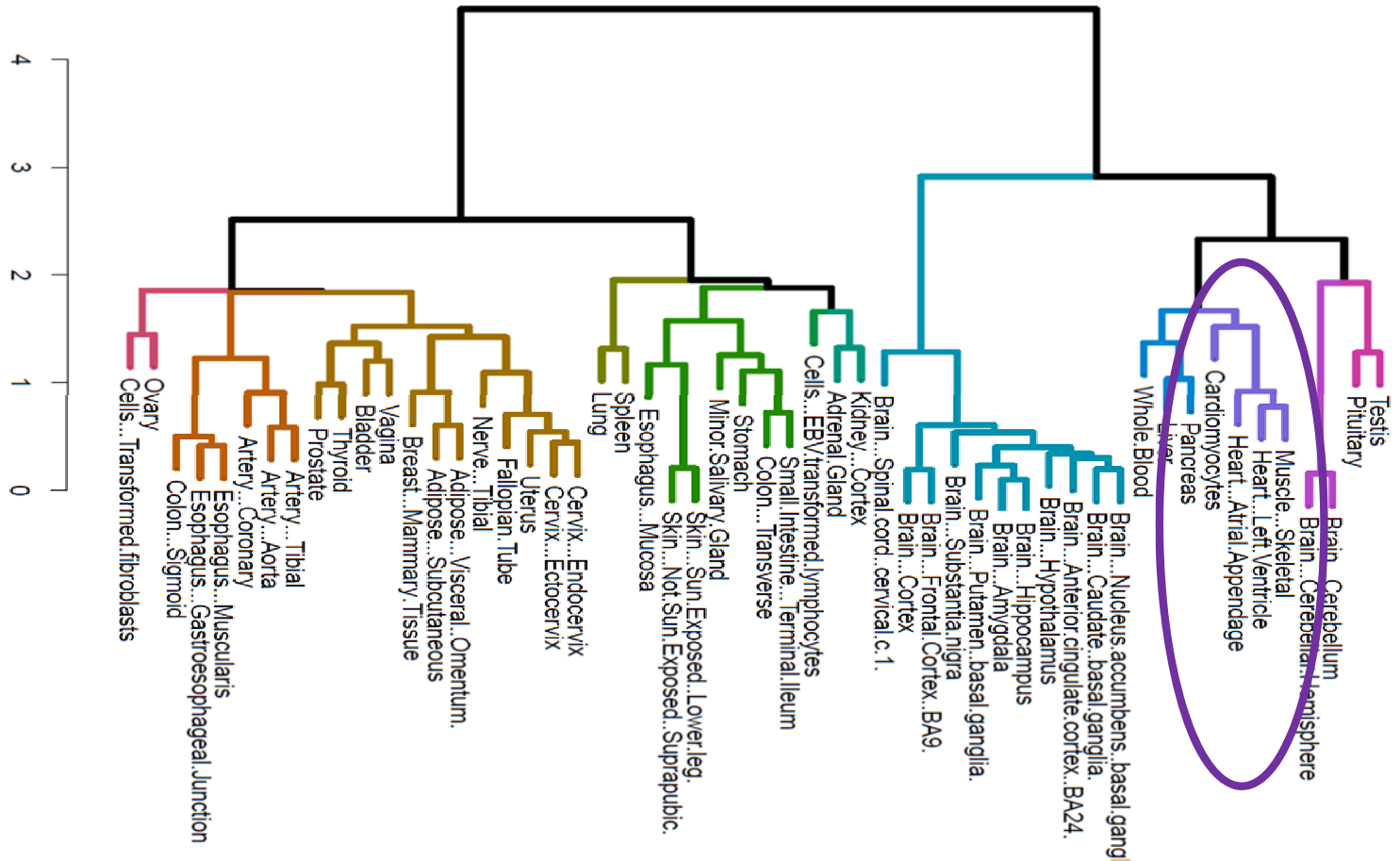
The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans

The GTEx Consortium*†

8 MAY 2015 • VOL 348 ISSUE 6235 sciencemag.org SCIENCE

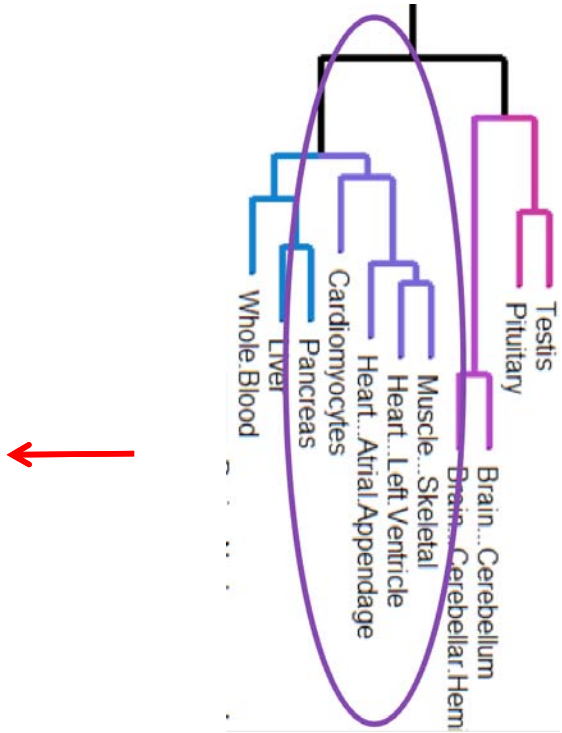
Are the expression data organotypic?

Using iPSCs –Comparison of Cardiomyocytes (TempOSeq) with GTEx (RNA-Seq) for ~2500 genes

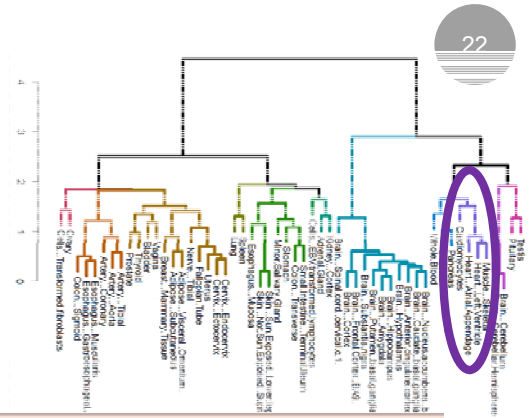


Are the expression data organotypic?

Using iPSCs –Comparison of Cardiomyocytes (TempOSeq) with GTEx (RNA-Seq) for ~2500 genes

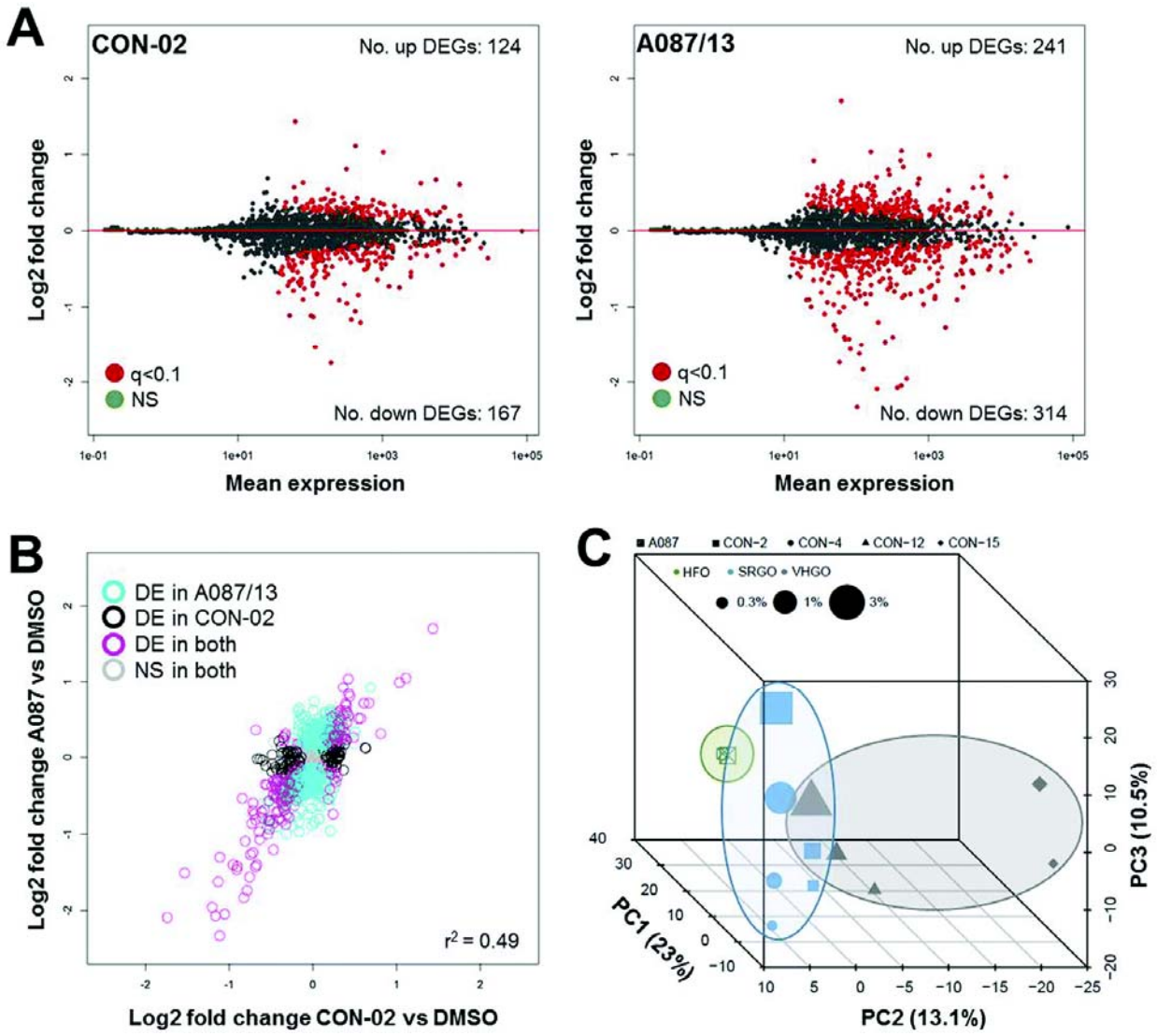


iPSC Cardiomyocytes – TempOSeq Top Expressed Genes among ~2700 genes



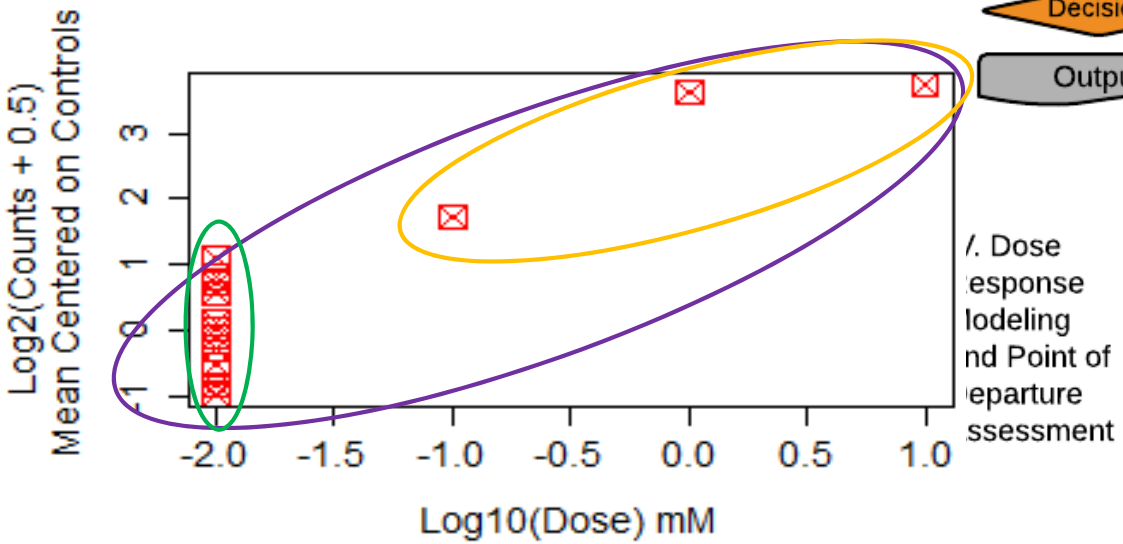
Gene	Alias	Function
TPM1	Tropomyosin 1 (Alpha)	Striated and SM Contractile System
MYL4	Myosin Light Chain 4	Muscle ATPase Motor Protein
ATP5B	ATP Synthase, H+ Transporting, Mitochondrial F1	ATP Synthase Subunit
PLN	Phospholamban	Cardiac Diastolic Function Regulator
MYL9	Myosin Light Chain 4	Regulates Muscle Contraction
CDH2	Cadherin 2	Cell Adhesion Glycoprotein
SLC25A4	Solute Carrier Family 25 Member 4	Mitochondrial Energy Generation
CLIC4	Chloride Intracellular Channel 4	Regulation of Cell Membrane Potential
EEF2	Eukaryotic Translation Elongation Factor 2	Protein Synthesis
ATP5C1	ATP Synthase, H+ Transporting, Mitochondrial F1 Complex, Gamma	Gradient Maintanance Mitochondria
PRKAR1A	Protein Kinase CAMP-Dependent Type I Regulatory Subunit Alpha	cAMP signalling
CNN3	Calponin 3	Muscle Contraction
GAPDH	Glyceraldehyde-3-Phosphate Dehydrogenase	Glycolysis and Metabolism
GNAS	GNAS Complex Locus	Heterotrimeric GTPase (G protein)
PPP1CB	Protein Phosphatase 1 Catalytic Subunit Bet	Cell Division, Glycogen Metabolism

How do we move from simple fold-change to robust dose-response?



Grimm *et al*. Green Chemistry 2016, 18:4407-4419

Flow chart for dose-response
Step 1: Trend detection and statistical flags



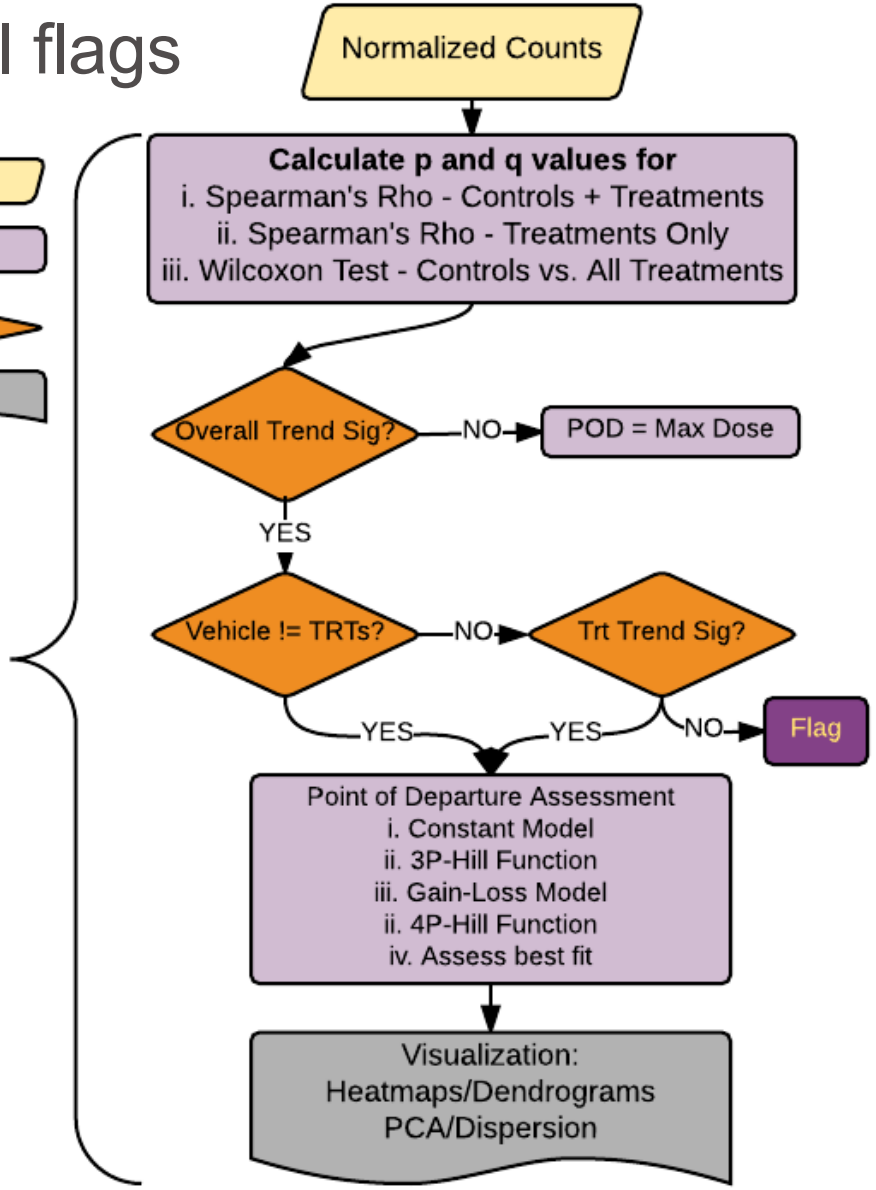
Data Files

Processes

Decision

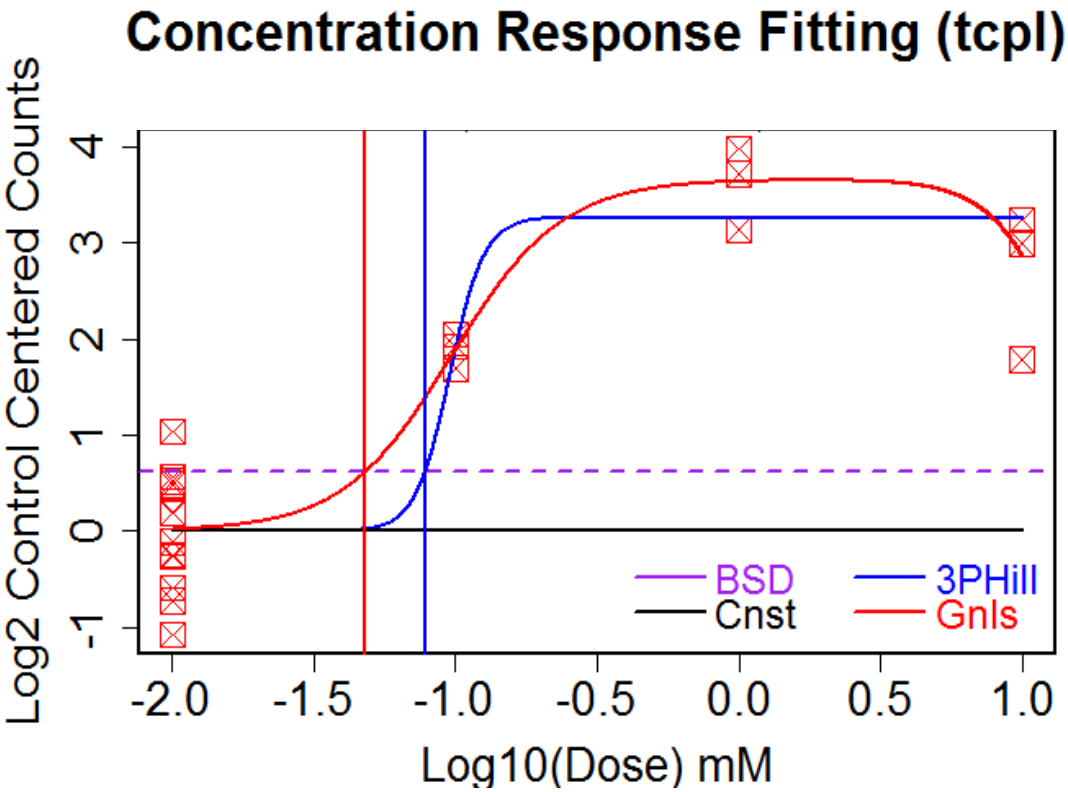
Output

/. Dose
esponse
lodeling
nd Point of
eparture
ssessment



Flow chart for dose-response
Step 2: Fit and assess Point of Departure (POD)

- POD based on +/- SD departures from control
- Also produce EC_{10}, EC_{50}, etc.
- Automated detection of outliers and significance “driven” by controls differing from remainder



4a.4 | ToxPi and integration

Analogue Read-Across with CBRA

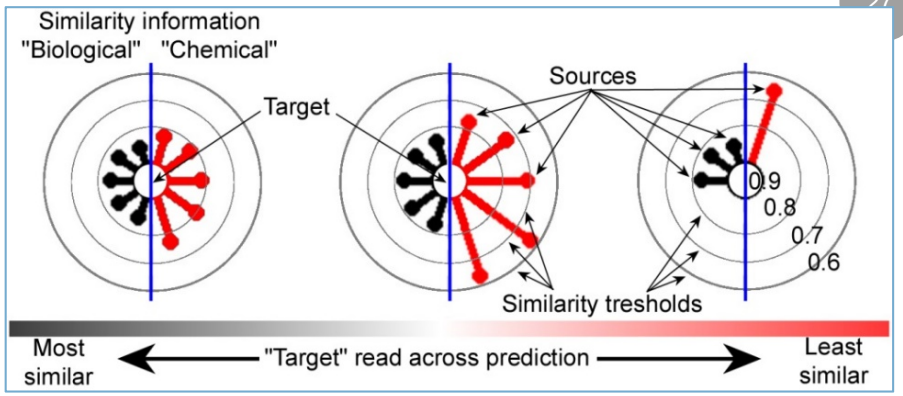
Chemical Research in Toxicology

Article

pubs.acs.org/crt

Integrative Chemical–Biological Read-Across Approach for Chemical Hazard Classification

Yen Low,[†,‡] Alexander Sedykh,[†] Denis Fourches,[†] Alexander Golbraikh,[†] Maurice Whelan,[‡] Ivan Rusyn,[*,‡] and Alexander Tropsha[*,†]



Category Grouping with ToxPi

BIOINFORMATICS APPLICATIONS NOTE

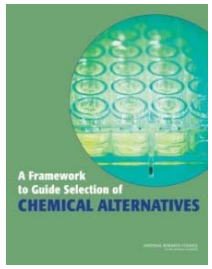
Vol. 29 no. 3 2013, pages 402–403
doi:10.1093/bioinformatics/bts686

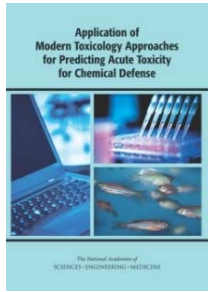
Systems biology

Advance Access publication November 29, 2012

ToxPi GUI: an interactive visualization tool for transparent integration of data from diverse sources of evidence

David M. Reif[1,*], Myroslav Sypa[2], Eric F. Lock[2], Fred A. Wright[3], Ander Wilson[1], Tommy Cathey[4], Richard R. Judson[1] and Ivan Rusyn[2]

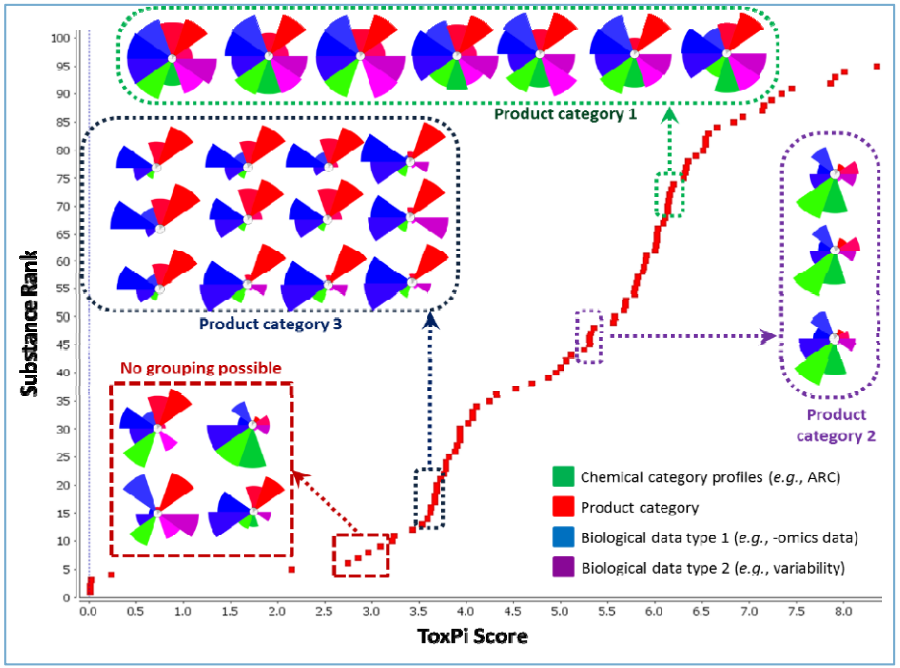




Endocrine Profiling and Prioritization of Environmental Chemicals Using ToxCast Data

David M. Reif,[1] Matthew T. Martin,[1] Shirlee W. Tan,[2] Keith A. Houck,[1] Richard S. Judson,[1] Ann M. Richard,[1] Thomas B. Knudsen,[1] David J. Dix,[1] and Robert J. Kavlock[1]

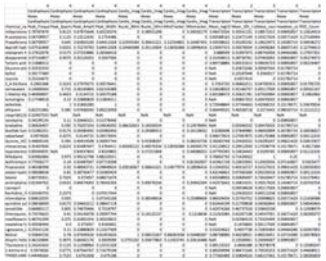
[1]National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, USA; [2]Office of Science Coordination and Policy, Office of Pollution Prevention, Pesticides and Toxic Substances, U.S. Environmental Protection Agency, Washington, DC, USA



Clustering: Profile similarity

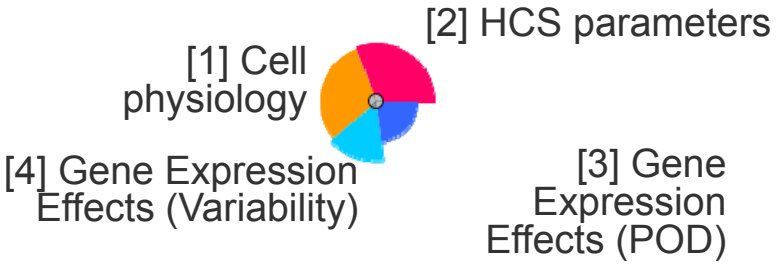
UVCBs with similar profiles may be interspersed throughout an overall rank distribution

1. Gather data

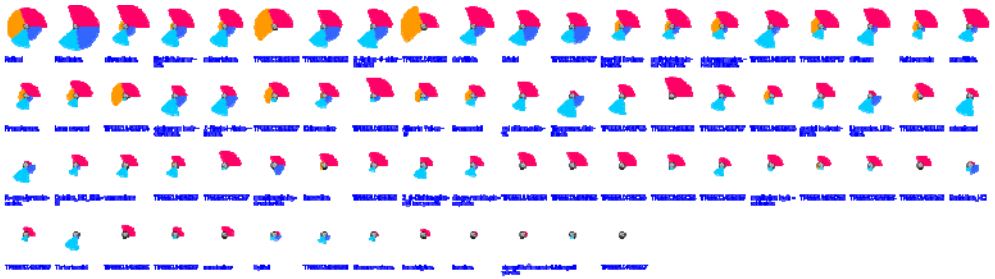


2. Load into GUI

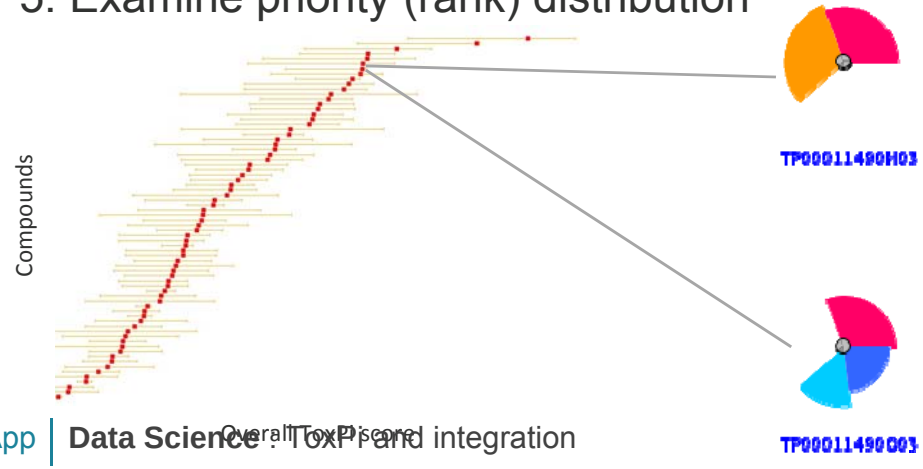
3. Apportion data into ToxPi model slices



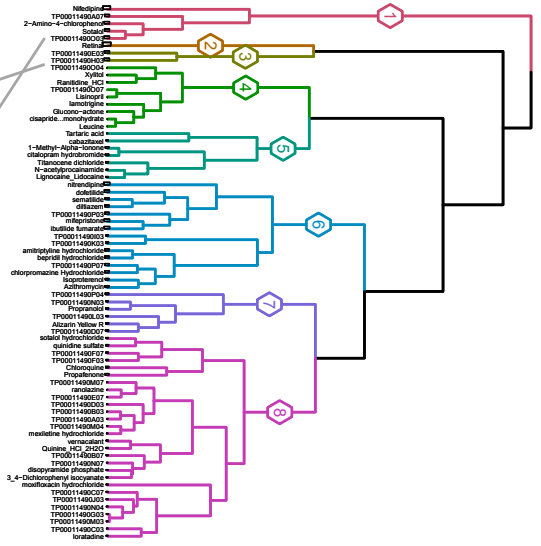
4. Examine profiles (sorted by rank)



5. Examine priority (rank) distribution



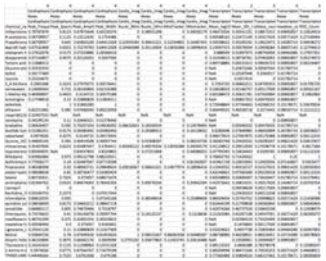
6. Examine profile similarity via clustering



Clustering: Profile similarity

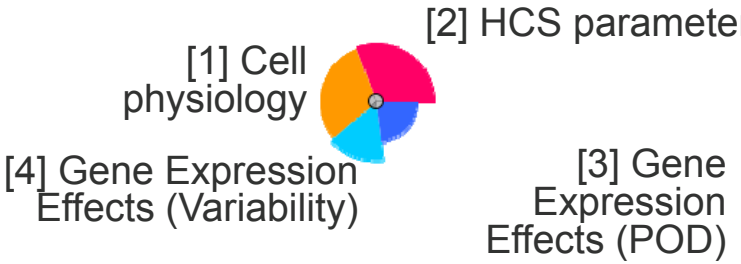
UVCBs with similar profiles may be interspersed throughout an overall rank distribution

1. Gather data

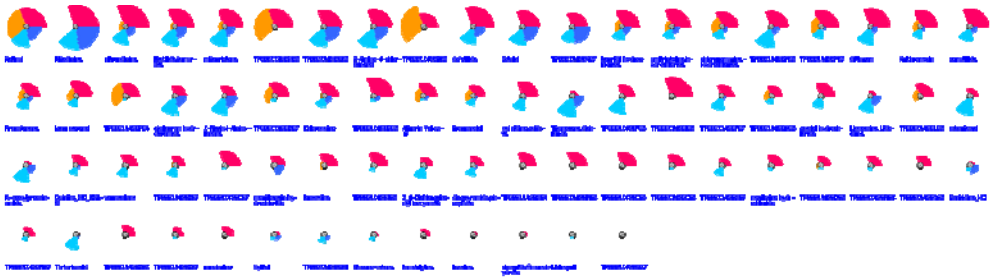


2. Load into GUI

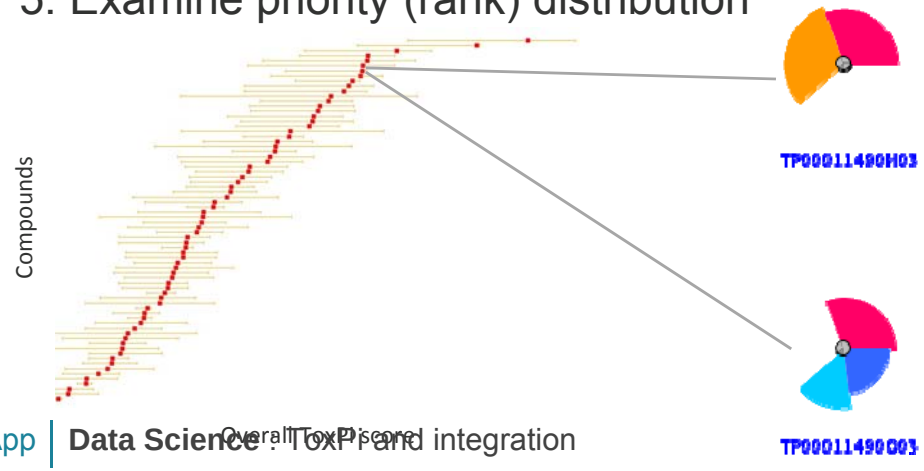
3. Apportion data into ToxPi model slices



4. Examine profiles (sorted by rank)

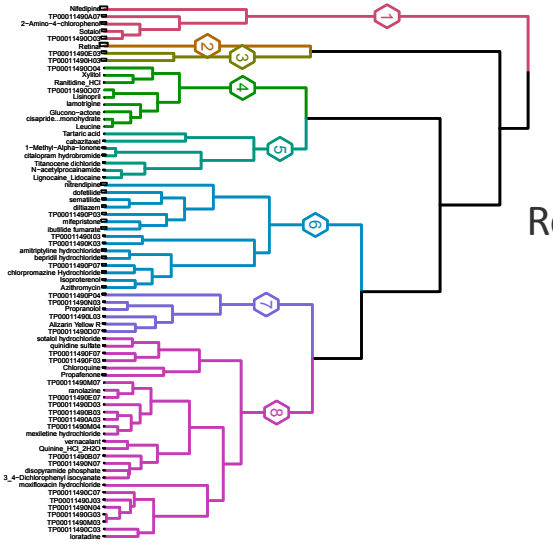


5. Examine priority (rank) distribution

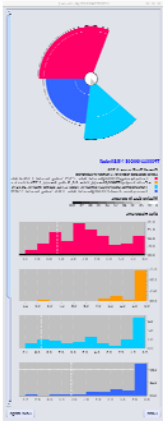


- These two substances are adjacently ranked, due to similarity in <u>overall</u> ToxPi scores
- However, their profiles indicate different reasons for achieving similar overall scores
- We can apply formal clustering methods for analysis of profile similarity

Additional work to enhance ToxPi interpretability and robustness for read-across

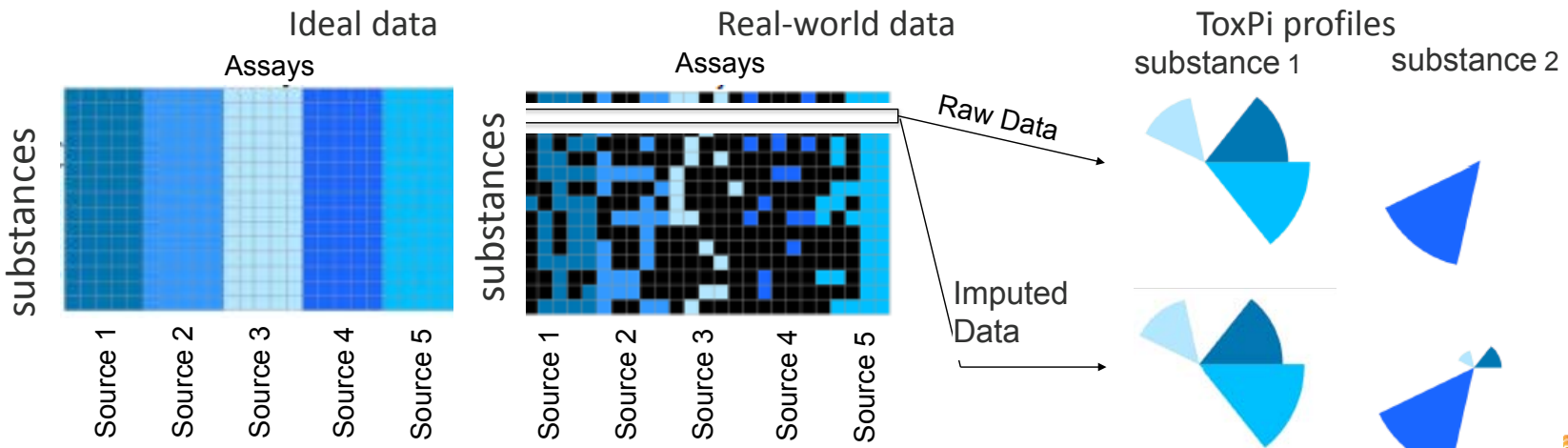


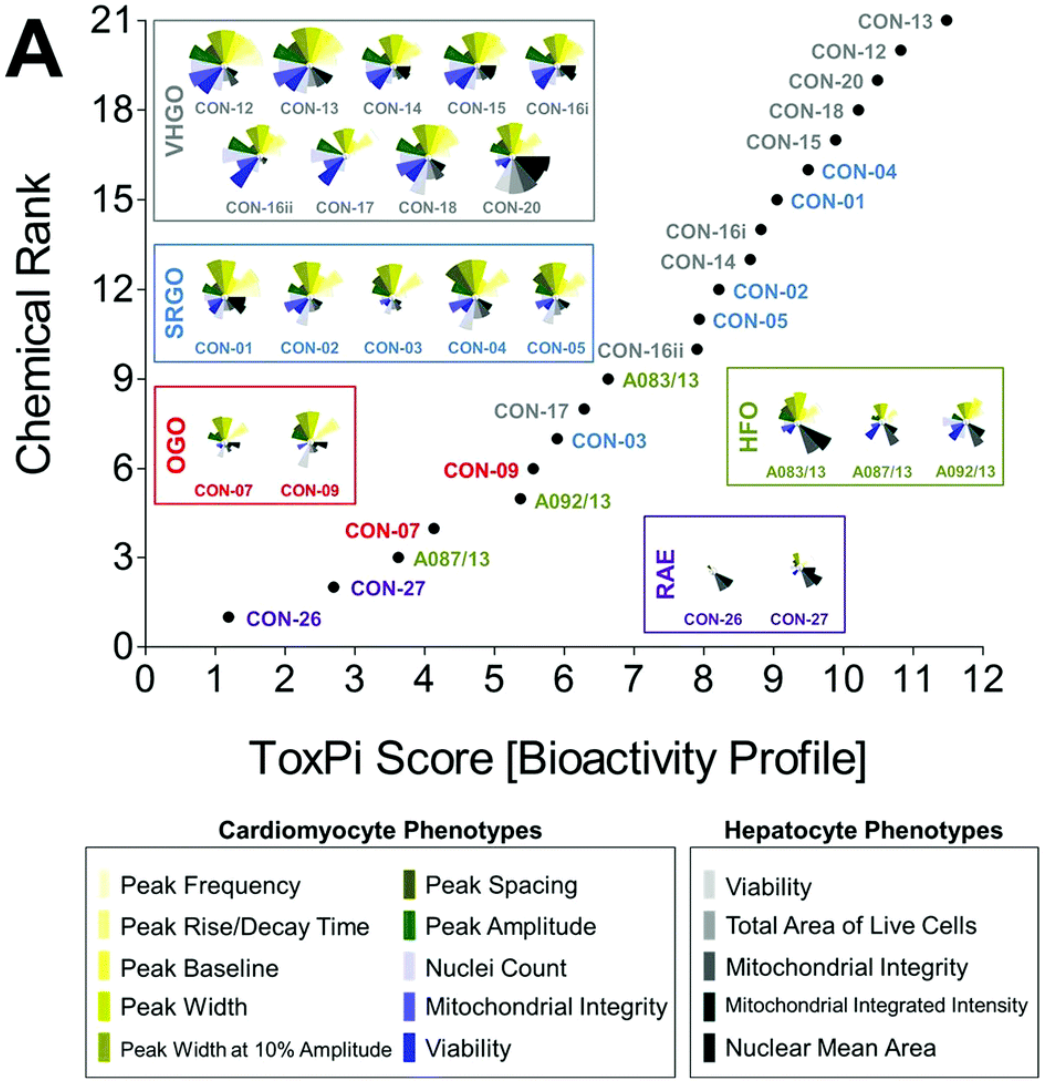
Robustness of clustering

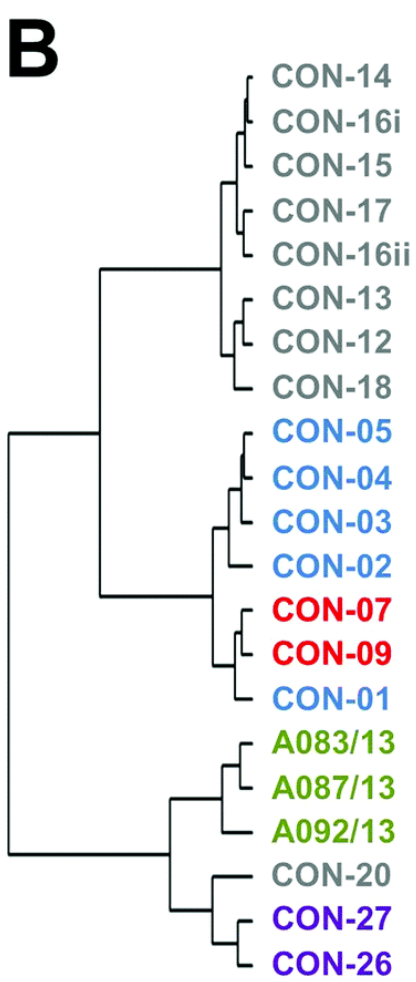


The new GUI has options for showing where a given substance falls within overall distributions for component data within slices

Methods to impute missing data







Grimm *et al*. Green Chemistry 2016, 18:4407-4419

Acknowledgments

John House (North Carolina State University)
Dereje Jima (North Carolina State University)
David Reif (North Carolina State University)
Denis Fourches (North Carolina State University)
Yi-Hui Zhou (North Carolina State University)
Ivan Rusyn (Texas A&M University)
Weihsueh Chiu (Texas A&M University)
Fabian Grimm (Texas A&M University)
Shu-Dong Zhang (University of Ulster)
Hans Ketelslegers (Concawe and ExxonMobil)
Tim Gant (Public Health England)
Klaus Lenz (SYNCOM)
Russell Thomas (US EPA-NCCT)
Peter Shepard (BioSpyder)

Funding:
NIH/NIEHS
EPA/NCER
European Petroleum
Refiners Association
(Concawe division)

Boulevard du Souverain, 165 B
1160 Brussels
Belgium

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

www.concawe.eu