Integration & Analysis of High-Throughput Assays in Next Generation Risk Assessment

Logan J. Everett, Ph.D.

Bioinformatics Scientist

Center for Computational Toxicology & Exposure,

Office of Research and Development, U.S. EPA

Research Triangle Park, North Carolina. U.S.A.

With contributions from: Joshua Harrill, Johanna Nyffeler, Imran Shah, and Richard Judson

Disclaimer

The views expressed in this presentation are those of the presenter and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency. Company or product names do not constitute endorsement by US EPA.

About Me

Education

- Bachelors in Computer Science from Binghamton University
- Ph.D. in Genomics & Computational Biology from University of Pennsylvania
- Postdocs at University of Pennsylvania and North Carolina State University

Work Experience

- Worked as Bioinformatics Scientist at Sciome, LLC, supporting research at U.S. National Toxicology Program and in industry
- Joined U.S. EPA Center for Computational Toxicology & Exposure in 2019
- Co-lead the High-Throughput Transcriptomics (HTTr) screening project with Dr. Joshua Harrill

Learning Objectives

- Introduce broad/high-throughput profiling technologies for toxicology
- Understand specific challenges and approaches for analyzing high-dimensional toxicogenomics data
- Potential applications for point-of-departure (POD) estimation, mechanism of action prediction, and read across

Outline

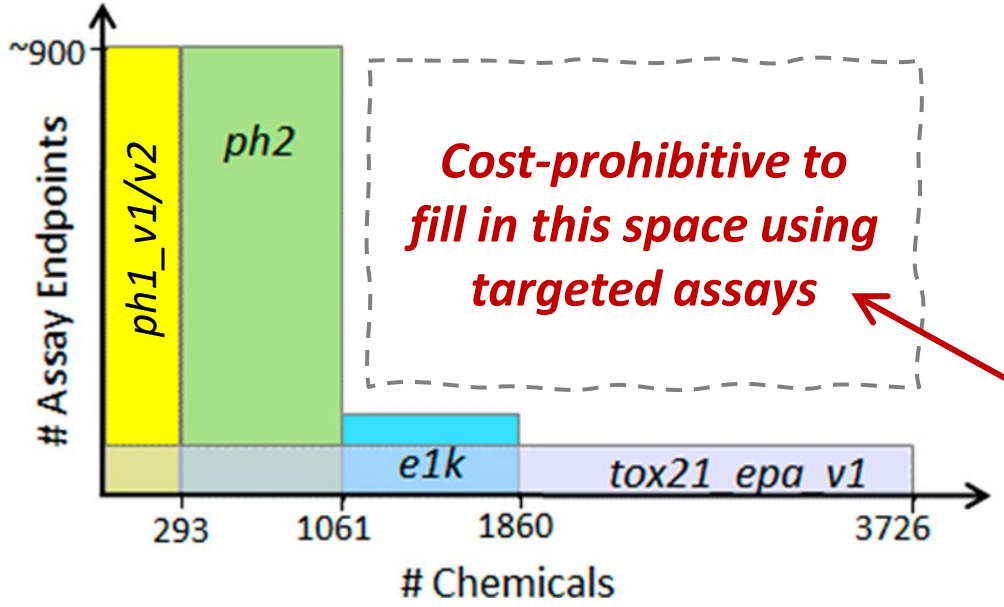
- Overview of High-Throughput Profiling Methods
- Overview of Transcriptomics
- Dose-Response Modeling of Transcriptomic Data
- Connectivity Mapping for Mechanistic Inference
- Cell Painting / High-Content Imaging

High-Throughput Profiling Methods

Broad-Coverage, Non-Targeted Assays

High-Throughput Screening (ToxCast)

Testing Phase	Chemical Set	Unique Chemicals	Assay Endpoints
ToxCast Phase I	*ph1_v1*	310	~700
ToxCast Phase II	*ph1_v2*	293	~200
	ph2	768	~900
	e1k	799	~50
Tox21	*tox21_epa_v1*	3726	~80



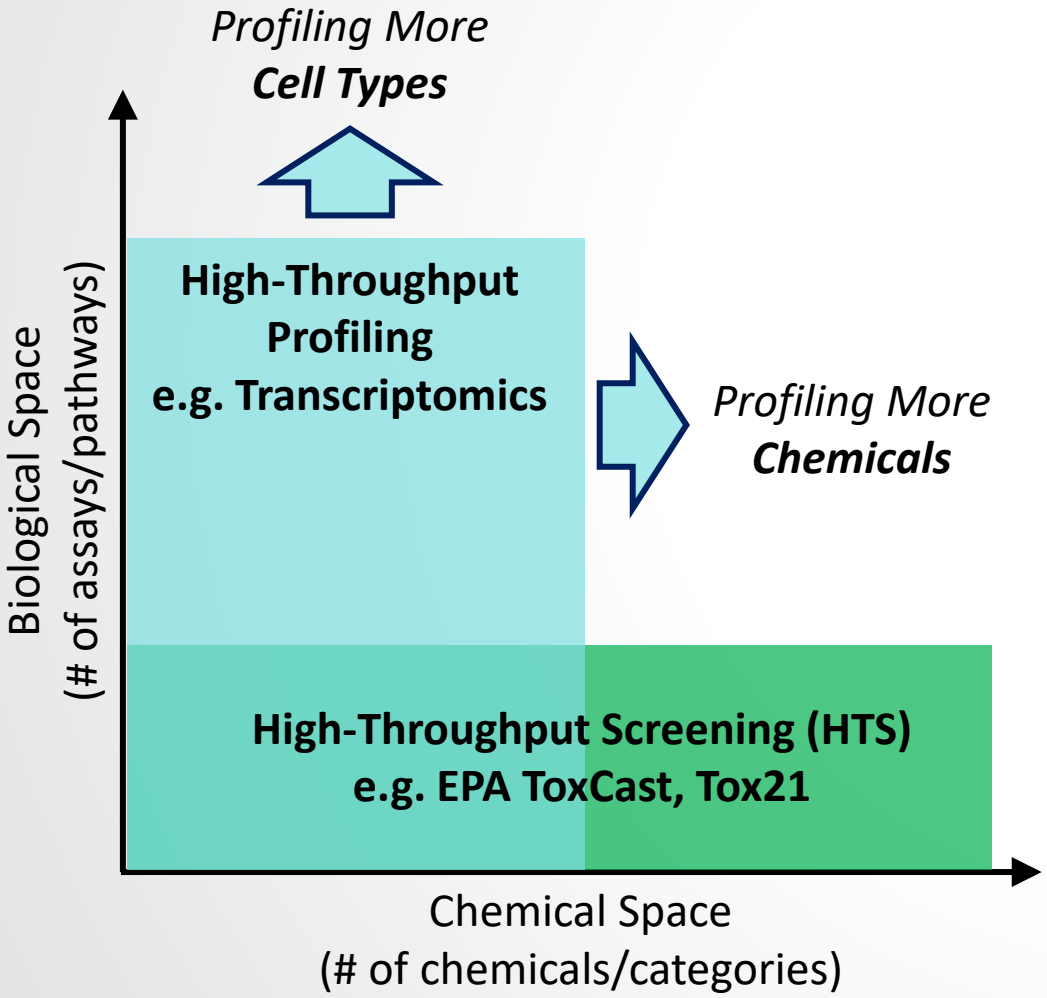
Richard et al. (2016)

ToxCast: EPA-led effort using high-throughput screening (HTS) assays to assess bioactivity and potential toxic effects.

- Expose living cells or isolated proteins *in vitro* to pure chemicals in vehicle/culture media
- Maintain standard library of 1,000s of diverse chemicals
- Mostly targeted assays (*chemical X* → *target Y*) leading to **incomplete coverage of human biological space**
- See: Richard, et al. (2016)
  DOI: 10.1021/acs.chemrestox6b00135

New Strategy for Hazard Evaluation: Improve efficiency and increase biological coverage by using broad-based (i.e. non-targeted) assays that capture many potential molecular and phenotypic responses of human cells to chemical exposure.

High-Throughput Profiling Methods



Definition: Any method that broadly profiles a range of biological pathways, functions, or features, as opposed to targeted testing of a single endpoint, including:

- **Transcriptomics** – profiling thousands of gene mRNA levels
- **Cell Painting** – profiling hundreds of cellular phenotypic features using microscopy
- *Broad batteries of targeted assays*
- *Other 'Omics – e.g. metabolomics, proteomics*

Profiling Across Human Cell Lines

Cell Line	Type / Tissue of Origin	Description/Relevance
MCF-7	Adenocarcinoma / Breast epithelium	Easy to maintain in culture, widely used, responds to a variety of hormonal signals
U-2 OS	Osteosarcoma / Bone	Easy to maintain in culture, responds to a variety of hormonal signals, ideal for cell painting
HepG2	Hepatocellular carcinoma / Liver	Easy to maintain in culture, reflects some aspects of hepatocyte/liver biology, but lacks phase II metabolism
HepaRG	Liver progenitor cells	Can be differentiated into cells that are metabolically competent, more closely resemble primary hepatocytes

Additional cell lines may be needed to detect:

- Hazards to specific organs, e.g. bronchial epithelial cells for inhalation exposures
- Interaction with other targets not expressed in cell lines above

Many Analysis Choices!

No single "best" method for analyzing high-throughput profiling data

- Are you interested in mechanism, or just want a threshold for general bioactivity?
- Is it more important to be **predictive** or **protective** of hazard level *in vivo*?
- What other data is available for the same/analogous chemicals?
- Different technologies require different statistical models, quality control, etc.
- Experimental design *(# of replicates, doses, etc.)* impacts analysis choices!

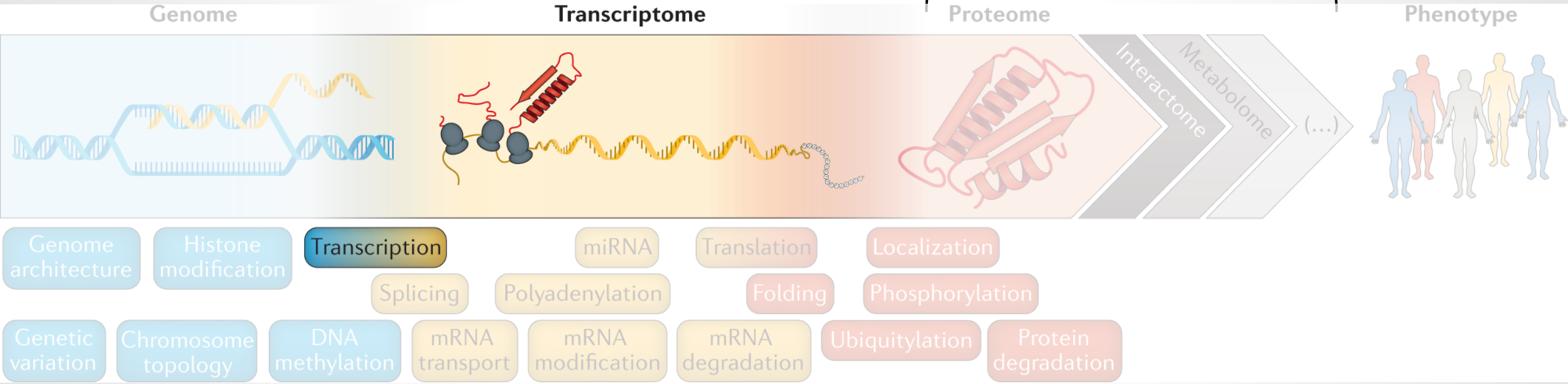
Overview of Transcriptomics

Profiling Genome-Wide Gene Expression

Why Transcriptomics?

Highly dynamic in response to environmental stimuli,
Well-established, cost-effective, high-throughput methods

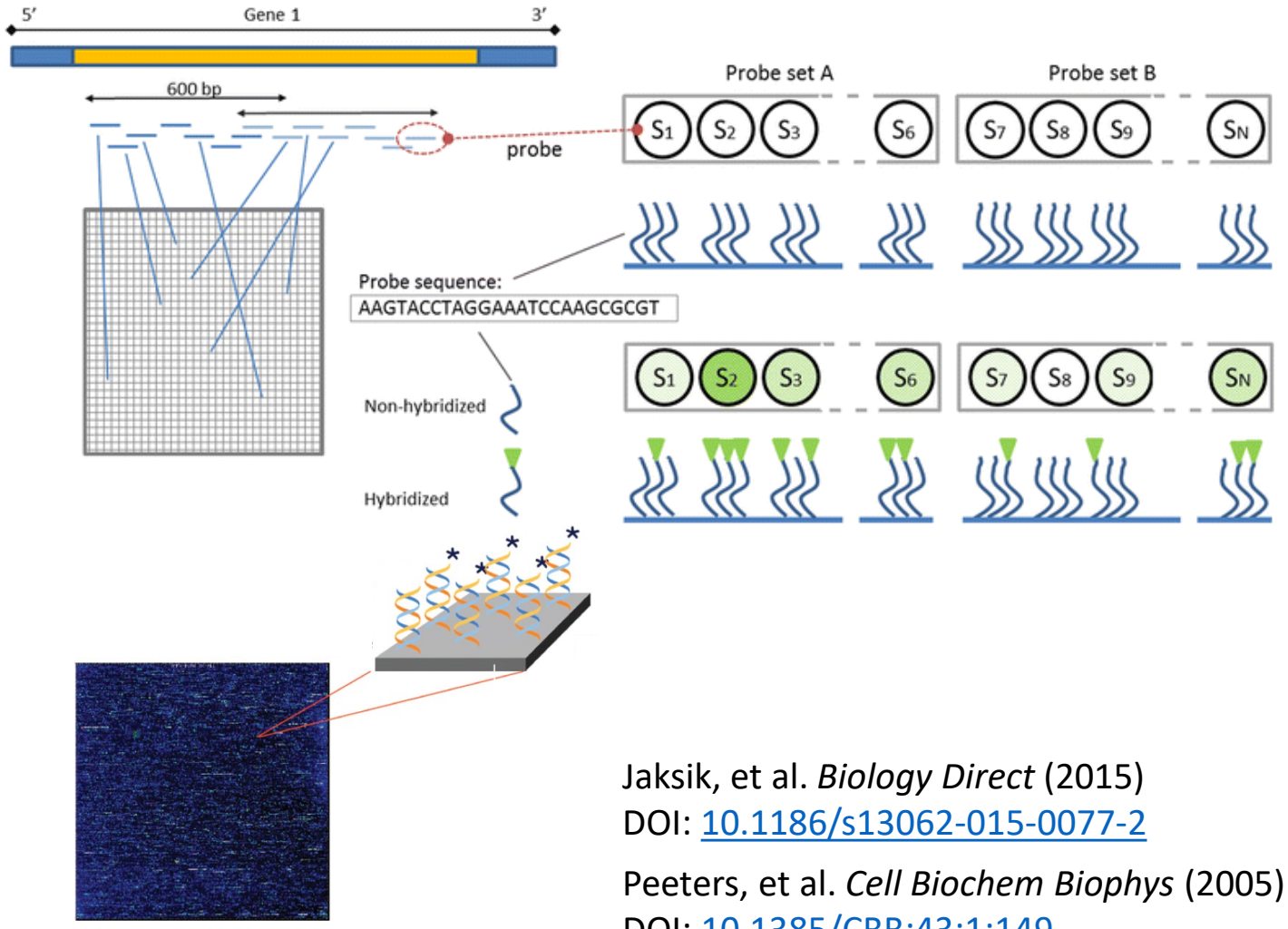
Broad profiling methods are newer
or less cost-effective



Hazardous exposures can cause perturbations
at multiple levels of gene regulation

Buccitelli and Selbach, *Nat Rev Genet* 2020
DOI: 10.1038/s41576-020-0258-4

Microarrays

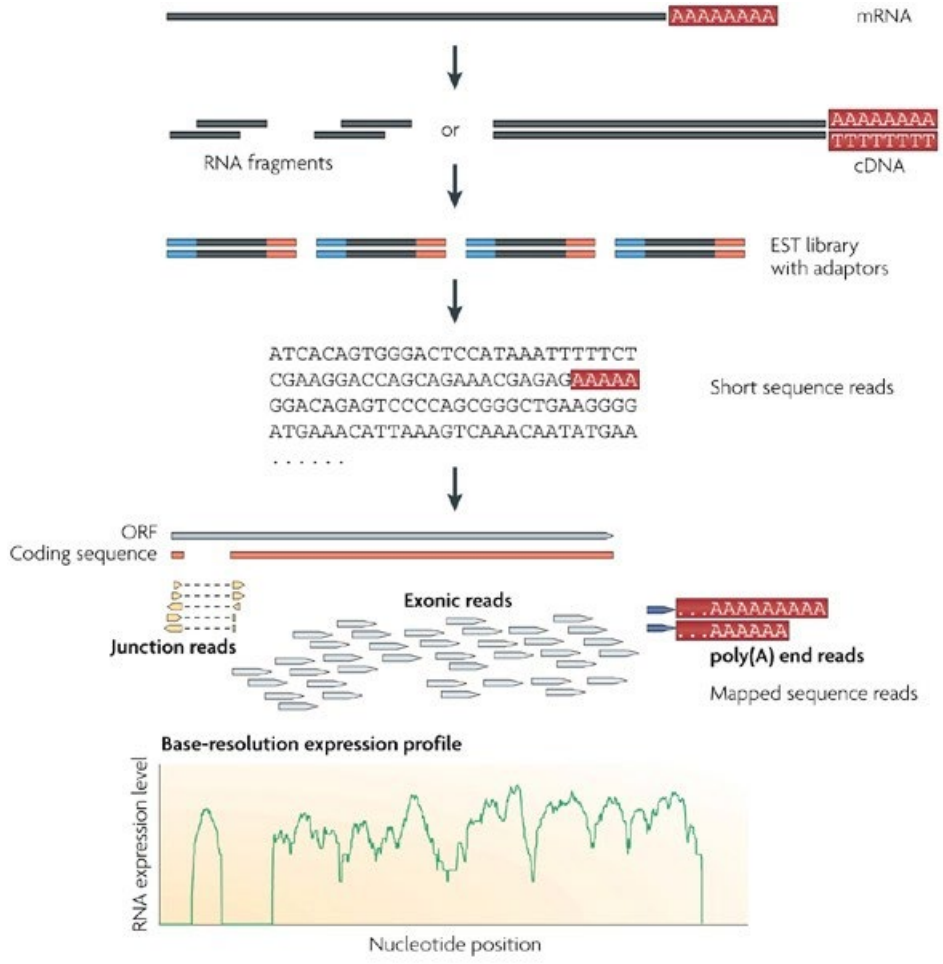


Jaksik, et al. *Biology Direct* (2015)
DOI: 10.1186/s13062-015-0077-2

Peeters, et al. *Cell Biochem Biophys* (2005)
DOI: 10.1385/CBB:43:1:149

- Fragments of expressed transcripts hybridized to pre-defined probe sequences
- Fluorescent labeling of hybridized probes
- **Label intensity** measures relative expression of probe target
- **Limited dynamic range** compared to other technologies
- Many commercially available options, well-established protocols

RNA-sequencing



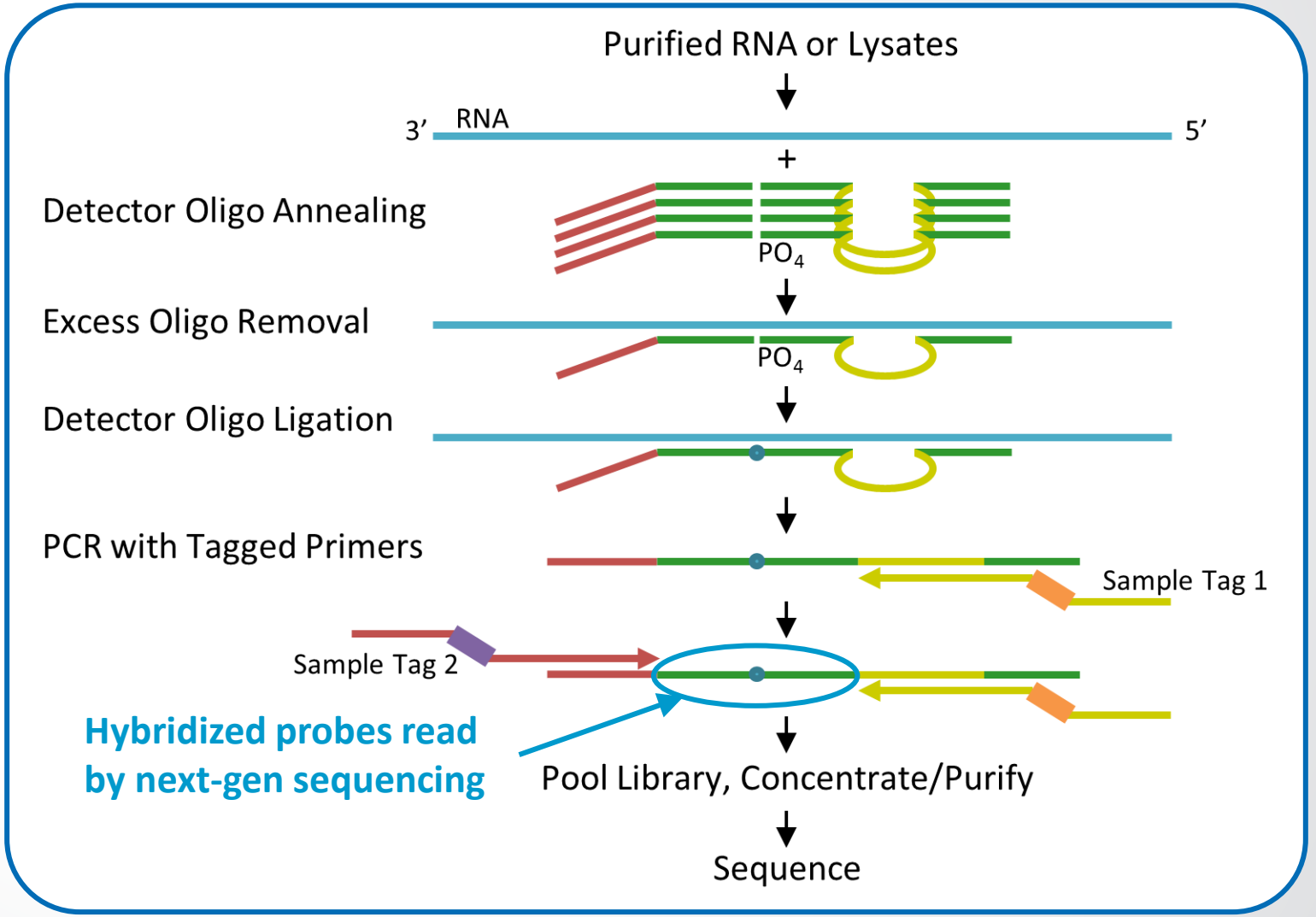
Wang, et al. *Nat Rev Genet* (2009) DOI: 10.1038/nrg2484

- Quantifies mRNA abundance by counting tens of millions of short reads
- Most common platform is Illumina next-gen sequencing
- Greater dynamic range than microarrays
- Captures more information than other technologies, e.g. splicing, lncRNA*
- Requires more complex statistical models for count-based measurements

**Ability to infer additional info depends on type of sequencing library (e.g. paired-end, total RNA)*

Targeted RNA-seq Assay (TempO-seq)

- Next-Gen sequencing of targeted probes hybridized to expressed transcripts
- Scalable: can measure a few thousand genes (S1500+) or up to whole transcriptome
- Captures gene expression at lower cost than RNA-seq or microarrays
- Compatible with raw cell lysates – ***ideal for large-scale screening***



Yeakley, *et al. PLoS ONE* (2017) DOI: 10.1371/journal.pone.0178302

Transcriptomics Knowledgebases

Signature Databases – any collection that links sets of genes to specific biological categories or patterns *(e.g. functions, pathways, responses)*:

- Gene Ontology (geneontology.org) - *Nucleic Acids Res* (2021) DOI:10.1093/nar/gkaa1113
- MSigDB (gsea-msigdb.org) - *Bioinformatics* (2011) DOI:10.1093/bioinformatics/btr260
- Reactome (reactome.org) - *Nucleic Acids Res* (2022) DOI:10.1093/nar/gkab1028

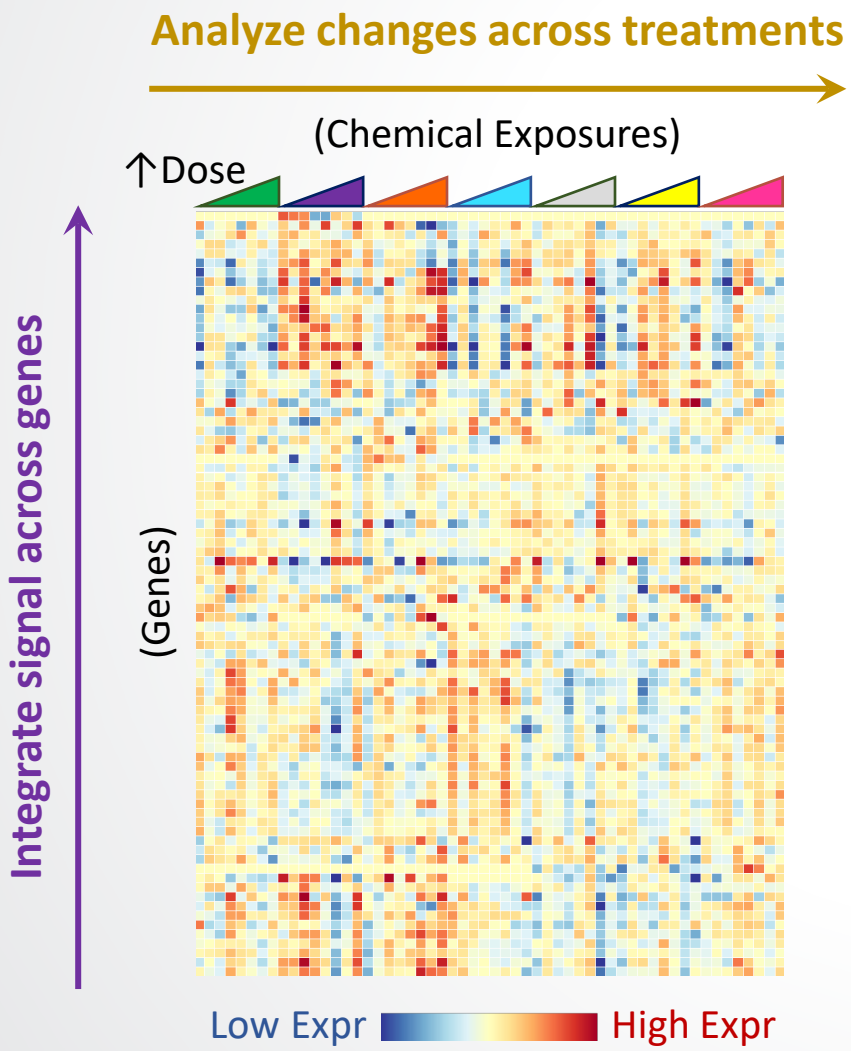
Databases of Toxicogenomic/Transcriptomic Profiles:

- TG-GATES (biosciencedb.jp) - *Nucleic Acids Res* (2015) DOI:10.1093/nar/gku955
- Connectivity Map (clue.io) – *Cell* (2018) DOI:10.1016/j.cell.2017.10.049

General Transcriptomic DBs: Gene Expression Omnibus (NCBI) ArrayExpress (EMBL)

Dose-Response Modeling of Transcriptomic Data

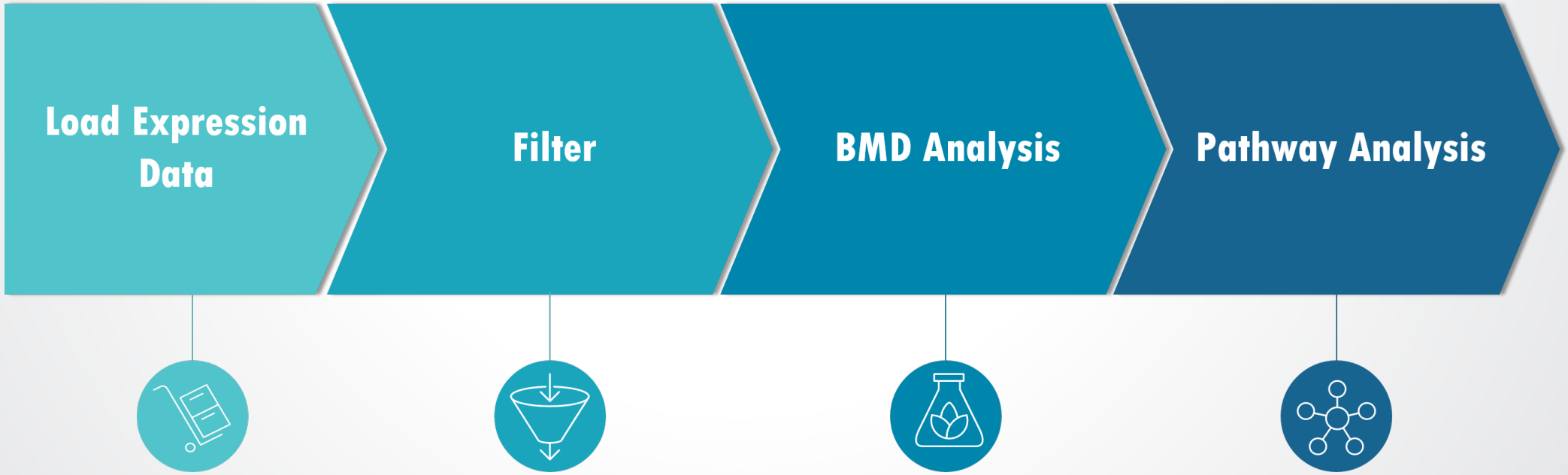
Transcriptomic Dose-Response Models



- Different genes may respond at different doses of a given exposure!
- Need to analyze both:
  - Dose-responsive trends
  - Coordinated changes in gene expression
- Gene-level data noisier in transcriptomics than targeted measurements (e.g. RT-qPCR)
- Dose-response modeling thousands of features (e.g. mRNA levels) leads to computational & statistical challenges

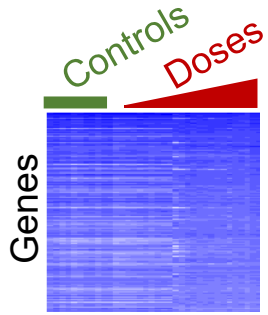
BMDExpress Software

- **Benchmark Dose (BMD):** Lowest dose/conc when an effect exceeds the background response rate
- BMDExpress automates fitting & summarizing multiple models (BMDS software) on many genes
- More information: Phillips, et al. *Bioinformatics* 2019 DOI: 10.1093/bioinformatics/bty878



https://github.com/auerbachs/BMDExpress-2/wiki

BMDExpress Software



- Input all replicates of controls and doses for a **single chemical exposure**
- Compatible with many transcriptomic technologies
- Data should be normalized, samples failing QC removed

Load Expression Data → **Filter** → **BMD Analysis** → **Pathway Analysis**

https://github.com/auerbachs/BMDExpress-2/wiki

BMDExpress Software

Want to model ***only*** those genes that are likely to be dose-responsive:

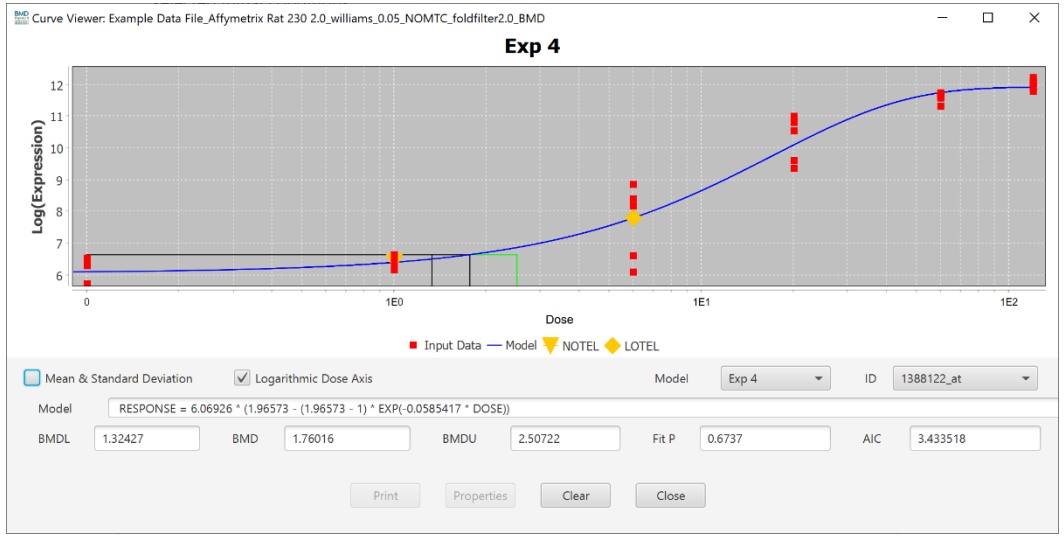
- Remove genes with weak fold-change at all doses vs controls
- Williams Trend Test for monotonicity
- Reduces unreliable curve fits & speeds up analysis

https://github.com/auerbachs/BMDExpress-2/wiki

BMDExpress Software

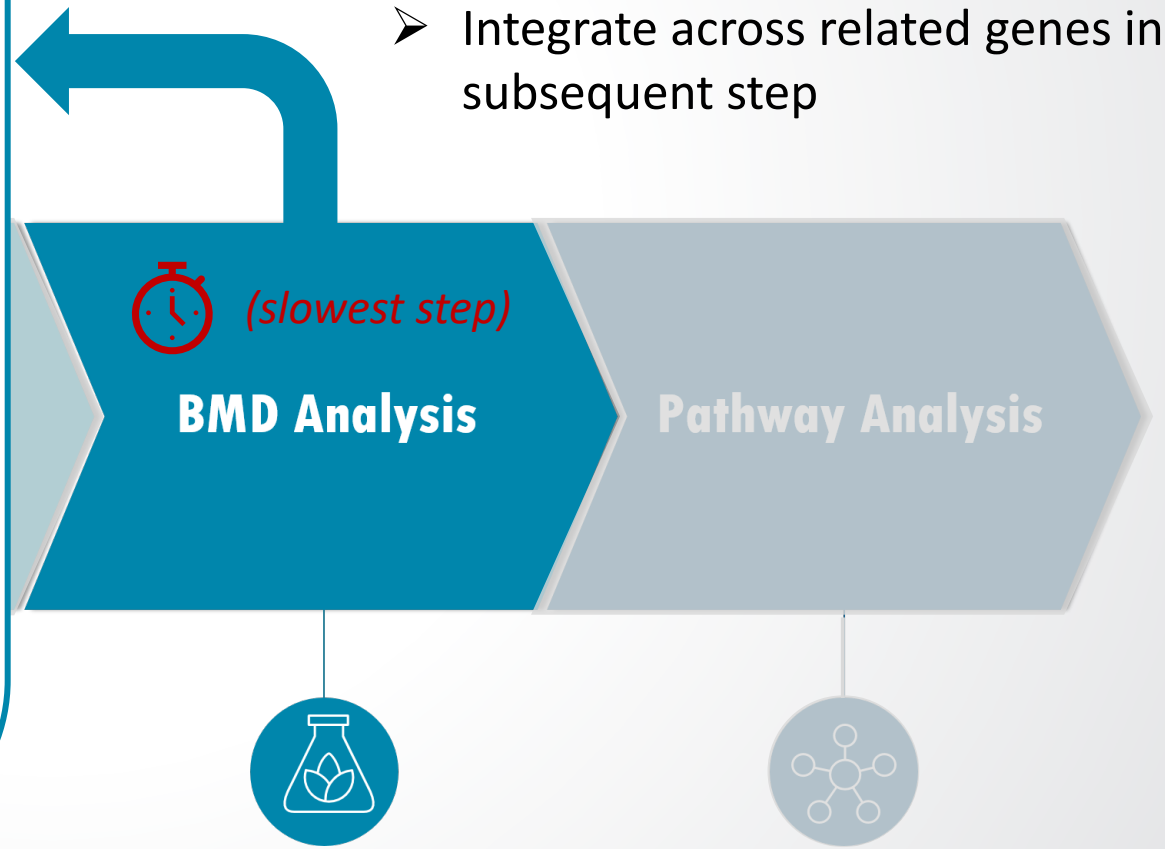
Run independently for each probe/gene:

- Fit multiple curve shapes to data
- Select best-fit model



Gene-level BMD typically computed using ***BMRf* = 1** or **1.349**, many tunable parameters

- Perform dose-response analysis on individual probes/genes
- Integrate across related genes in subsequent step



https://github.com/auerbachs/BMDExpress-2/wiki

BMDExpress Software

Summarize dose-response models for ***biologically related*** sets of gene

- Identify pathways/gene sets with multiple dose-responsive genes
- Category-level POD = median of active gene-level PODs

https://github.com/auerbachs/BMDExpress-2/wiki

BMDExpress Software

Many Pathway Collections:
- Gene Ontology
- Reactome
- MSigDB

Summarize dose-response models for ***biologically related*** sets of gene
- Identify pathways/gene sets with multiple dose-responsive genes
- Pathway-level POD = median of active gene-level PODs
- Further summarize overall POD based most sensitive pathway POD

https://github.com/auerbachs/BMDExpress-2/wiki

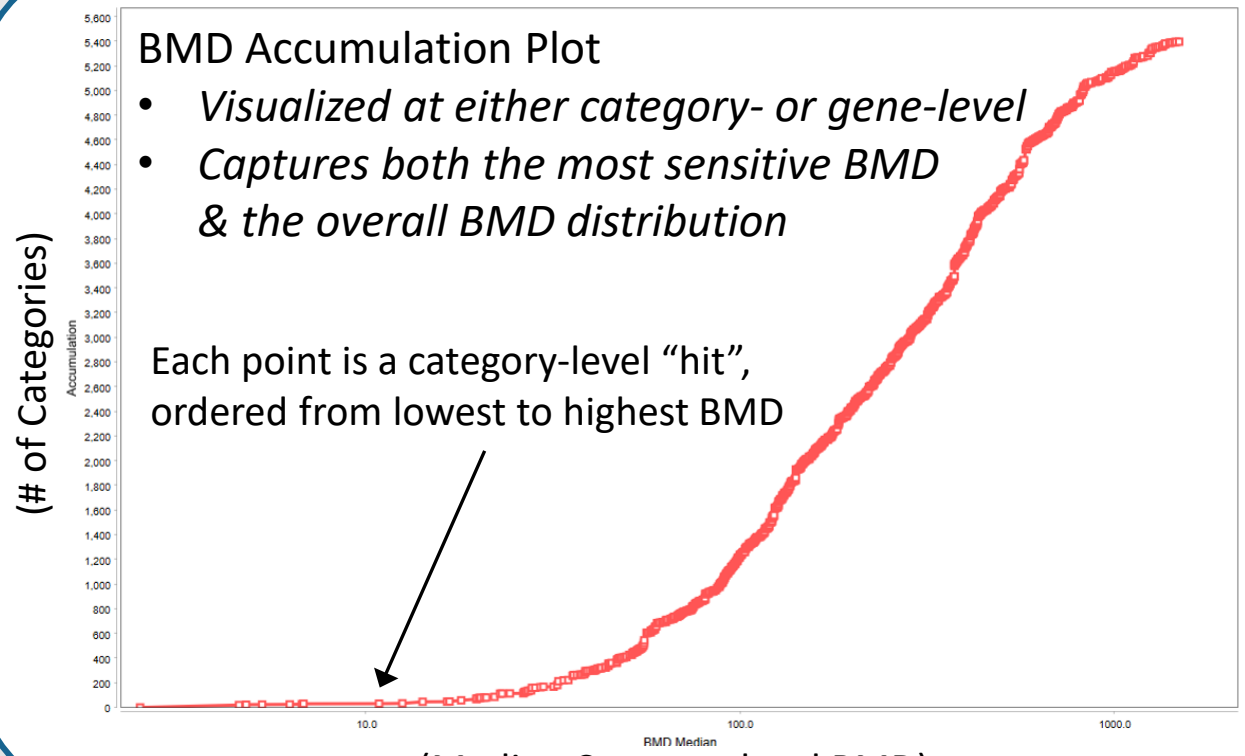
BMDExpress Software

Summarize dose-response models for ***biologically related*** sets of gene

- Identify pathways/gene sets with multiple dose-responsive genes
- Category-level POD = median of active gene-level PODs

BMD Accumulation Plot

- *Visualized at either category- or gene-level*
- *Captures both the most sensitive BMD & the overall BMD distribution*



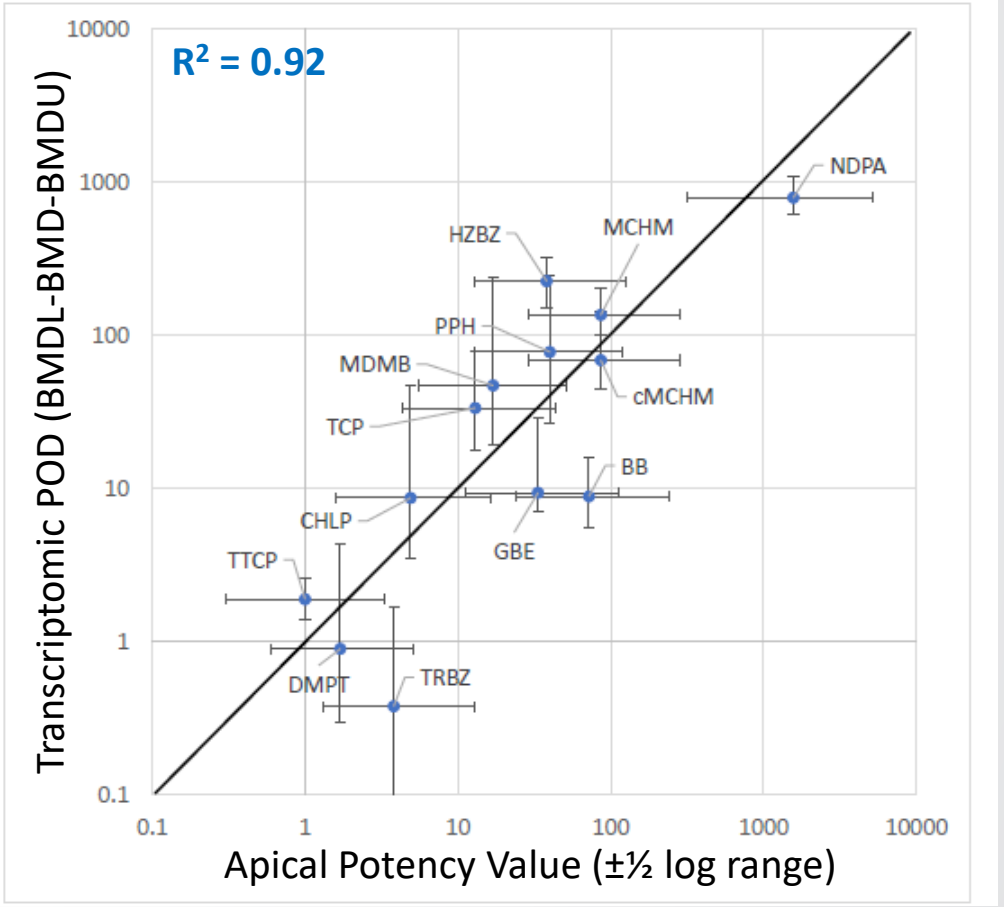
BMD Analysis → Pathway Analysis

https://github.com/auerbachs/BMDExpress-2/wiki

Overall Transcriptomic POD (tPOD)

Common methods for deriving overall tPOD:

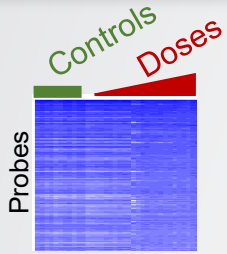
Gene-based Methods	
*N*th Percentile (e.g. 5th %ile) BMD	Reardon, et al. *Tox Sci* 2021
*N*th Lowest (e.g. 25th) BMD	
Pathway/Category-based Methods	
Lowest Active Pathway BMD	Gwinn, et al. Tox Sci 2020
5th Percentile Pathway BMD	Harrill, et al. Tox Sci 2021
Global Methods:	
Distance-based POD (e.g. Mahalonobis Distance)	Nyffeler, et al. *SLAS Discov* 2021



Important to benchmark methods for the intended purpose, e.g. prediction of PODs from animal studies

NTP Research Report, 2018
DOI: 10.22427/NTP-RR-5

Dose-Response Modeling of Gene Sets/Signatures



Concentration Series of Whole Transcriptome Profiles + Vehicle Controls

Love et al. (2014)

DESeq2 Moderated Fold-Changes

Curated Signature Collection

Barbie et al. (2009)

Gene Set Enrichment Analysis

Sheffield et al. (2022)

Benchmark Dose Modeling

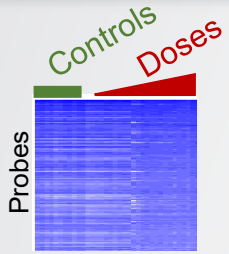
Catalog of gene set signatures with toxicological relevance, annotated for known molecular targets

- **Bioplanet** *(Huang, et al. Front Pharmacol 2019)*
- **CMap** *(Subramanian, et al. Cell 2017)*
- **DisGeNET** *(Piñero, et al. Database 2015)*
- **MSigDB** *(Liberzon, et al. Cell Syst 2015)*

Open Source: github.com/USEPA/CompTox-httrpathway

- EPA/CCTE method for summarizing large-scale transcriptomic screening studies
- Integrates signal across known gene set (a.k.a. signature) ***before*** dose-response modeling

Dose-Response Modeling of Gene Signature Scores



Concentration Series of Whole Transcriptome Profiles + Vehicle Controls

Love et al. (2014)

DESeq2 Moderated Fold-Changes

Curated Signature Collection

Barbie et al. (2009)

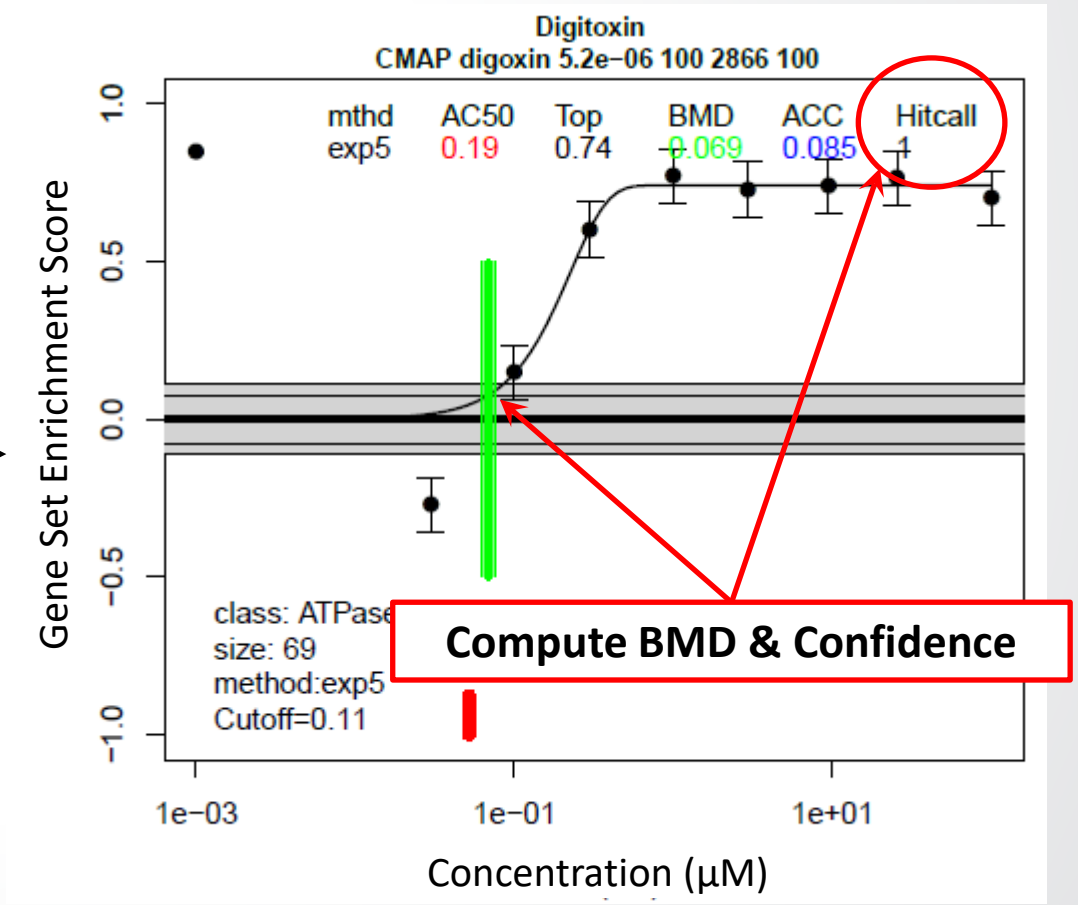
Gene Set Enrichment Analysis

Sheffield et al. (2022)

Benchmark Dose Modeling

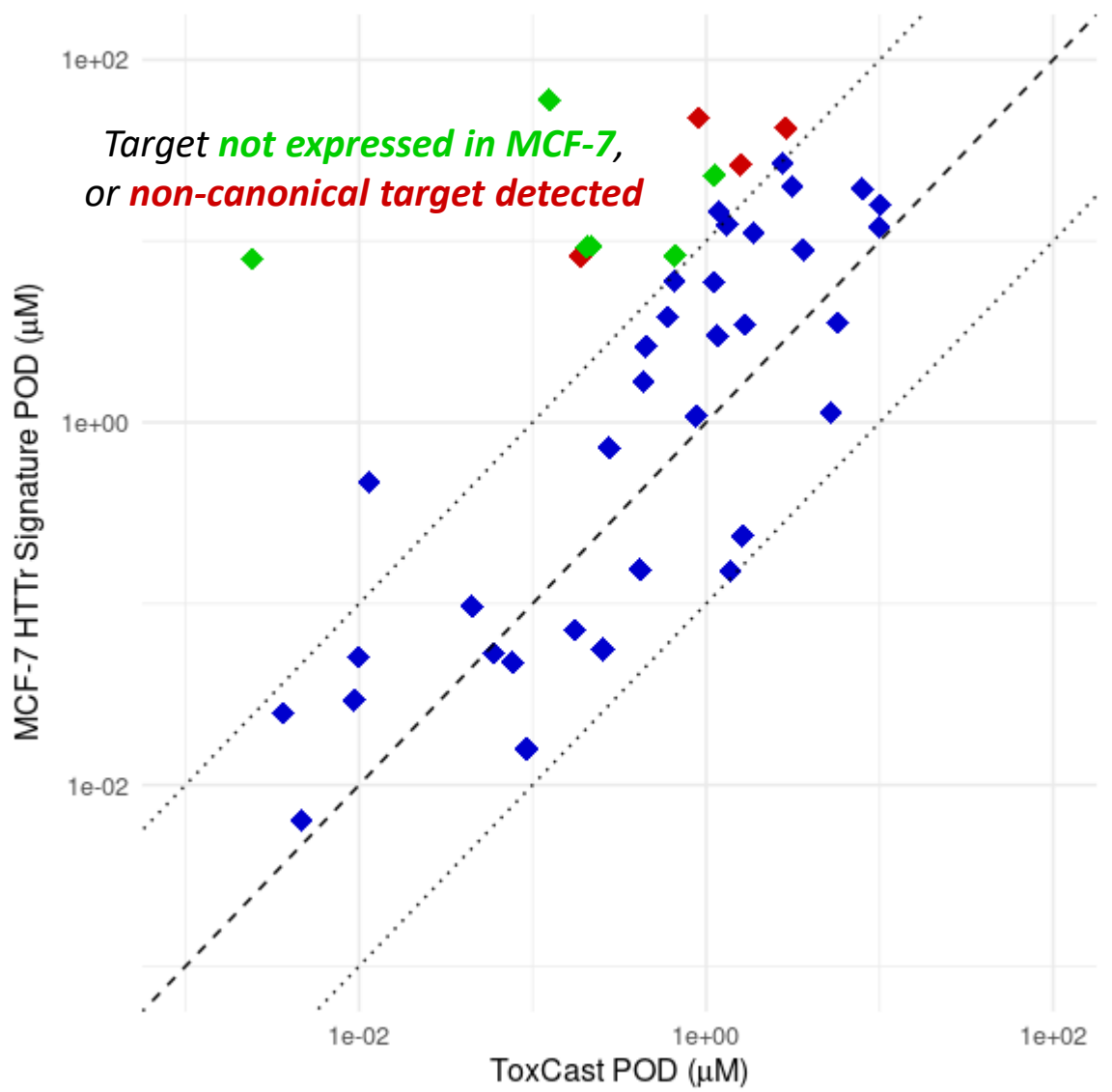
Overall tPOD = 5th percentile of active signature BMDs

(Hitcall captures confidence that signature is truly dose-responsive)



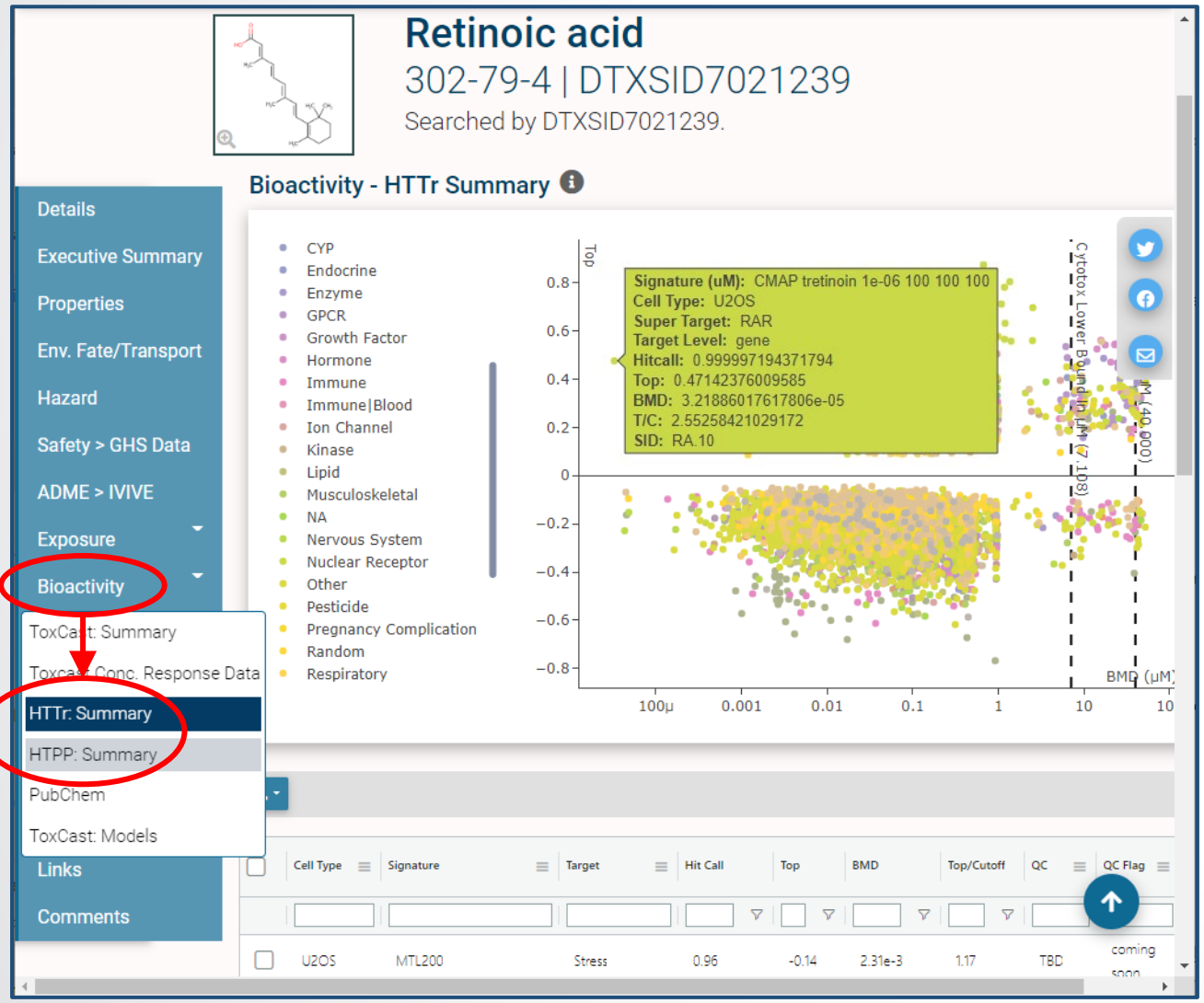
CRAN.R-project.org/package=tcplfit2

tPODs Are Concordant With ToxCast



- Pilot study of 44 well-characterized chemicals in MCF-7 cells, 6h exposure
  Harrill, et al. *Toxicol Sci* (2021)
  DOI: 10.1093/toxsci/kfab009
- Compared transcriptomic PODs to previous ToxCast targeted assay results *(multiple cell types, assays, and exposure lengths)*
  Paul-Friedman, et al. *Toxicol Sci* (2020)
  DOI: 10.1093/toxsci/kfz201
- Signature-based PODs are highly concordant with ToxCast results for the majority of test chemicals in pilot study

EPA High-Throughput Profiling Results



- Now available on EPA's CompTox Chemicals Dashboard! (comptox.epa.gov)
- Signature-level results viewable and downloadable (HTTr)
- Cell Painting results will also be available (HTPP)

Alternate Dose-Response Modeling Methods

Many other analysis methods proposed, this is an active area of research!

Bayesian Methods:

- BIFROST – Reynolds, et al. *Comp Tox* (2020) DOI: 10.1016/j.comtox.2020.100138
- BBMD – Shao & Shapiro, *EHP* (2018) URL: benchmarkdose.com
- ToxicR – Wheeler, et al. *Environmetrics* (2022) CRAN: ToxicR

Integration across genes using latent variables:

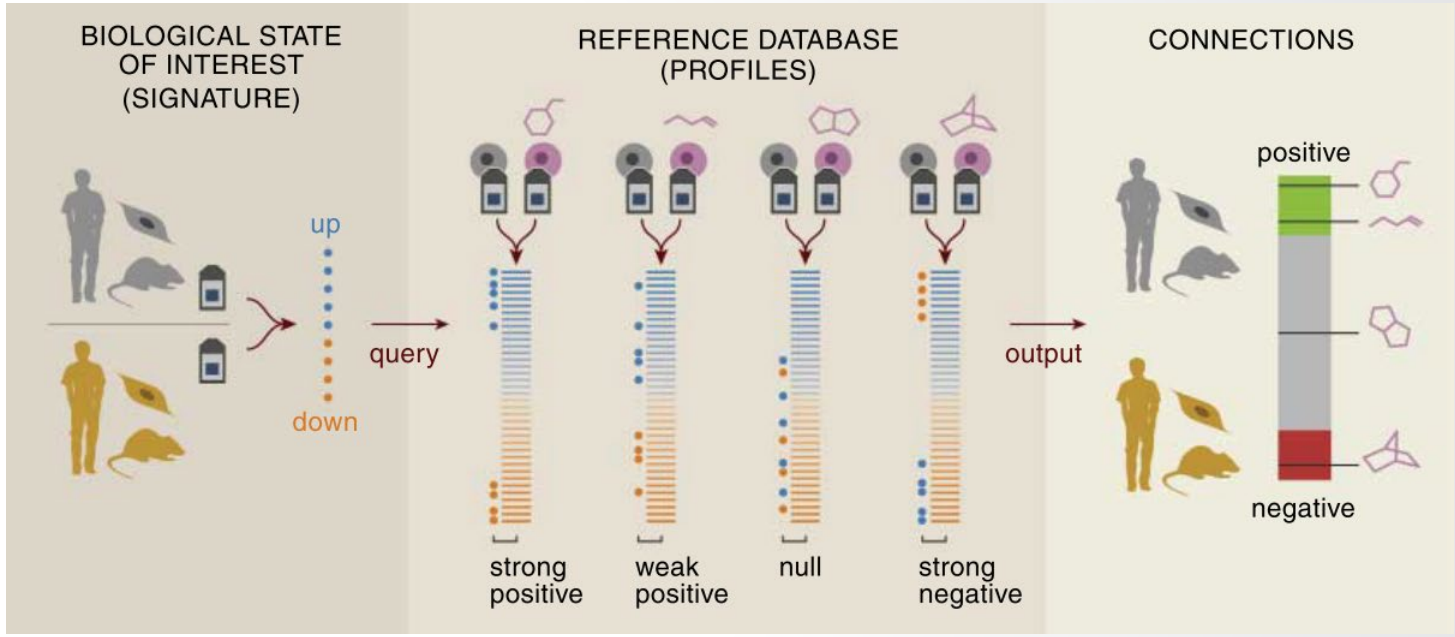
- Basili, et al. *Chem Res Tox* (2022) DOI: 10.1021/acs.chemrestox.1c00444

Connectivity Mapping

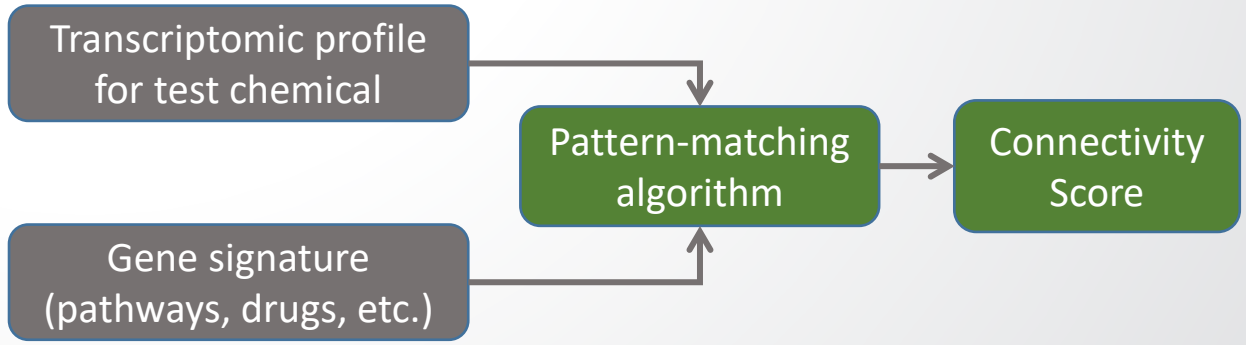
Inference of common mechanism/effects by transcriptomic similarity

Inferring Mechanism thru Connectivity Mapping

- Postulates that diverse biological states can be “fingerprinted” by the universal language of genes
- Similarity between transcriptomic profiles and genes signatures suggest common mechanisms
- Increasing utility in toxicology to infer mechanism by similarity
  DeAbrew *et al.*, 2016
  Wang *et al.*, 2016
- Web-based tool: https://clue.io/ (Broad Institute)



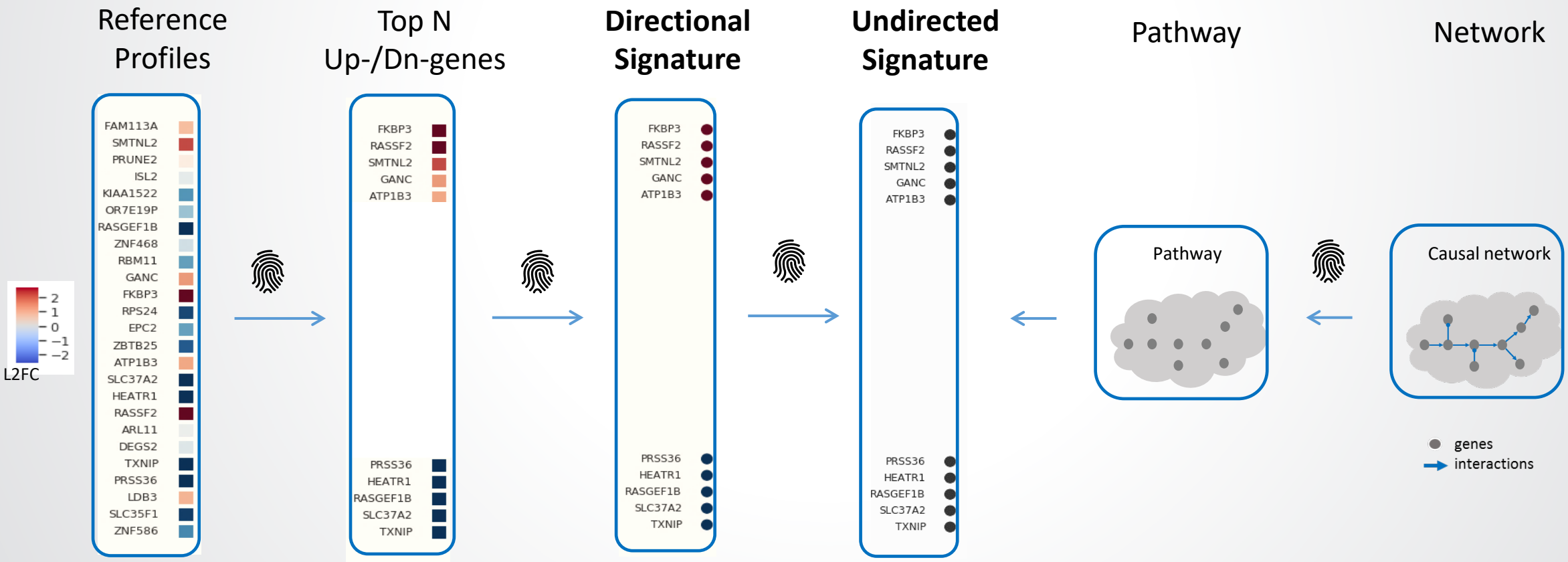
Lamb *et al.*, Science 2006 DOI: 10.1126/science.1132939



Imran Shah

“Fingerprinting” Bioactivity via Gene Signatures

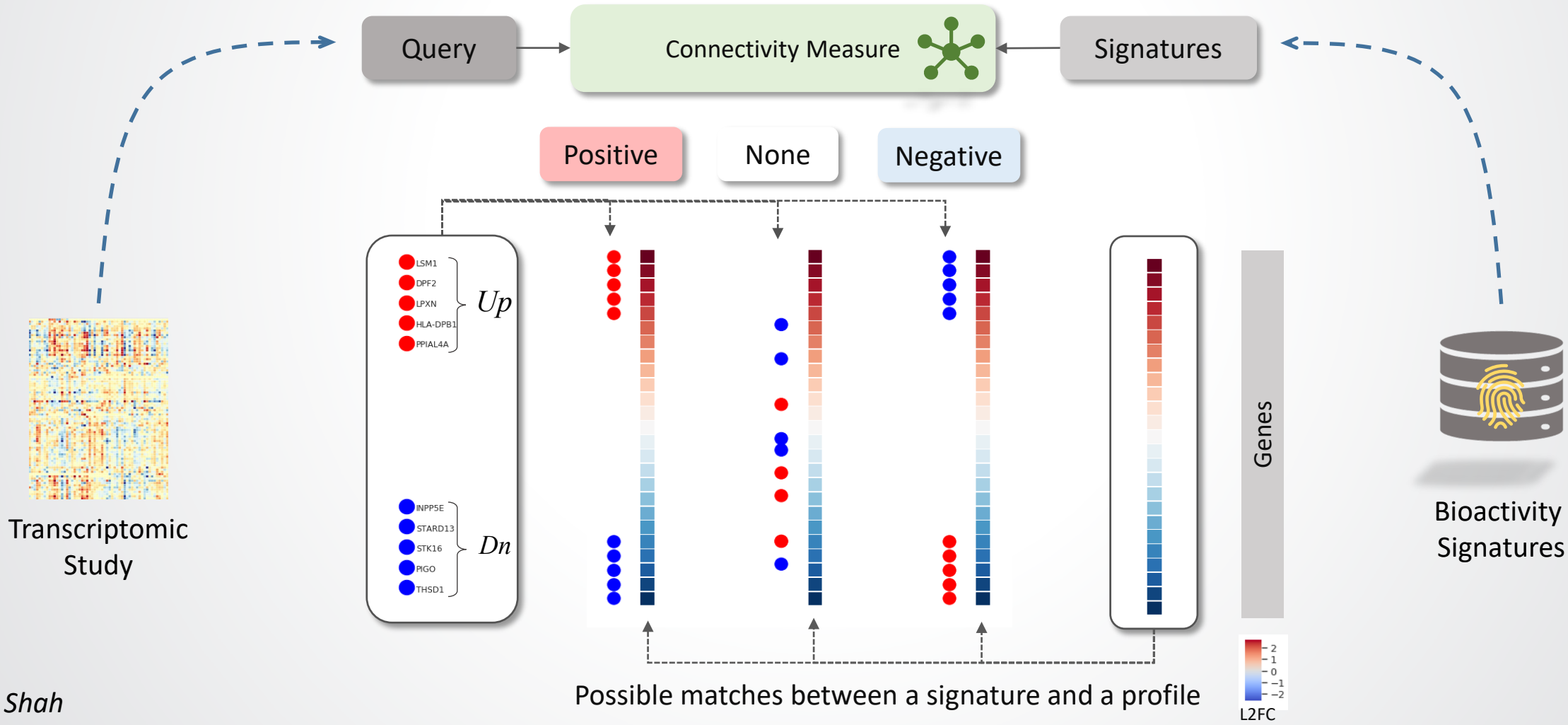
Signatures can be derived from specific gene expression profiles as well as conceptual models of biological pathways & networks, different representations are also possible.



Imran Shah

Connectivity Mapping with Gene Signatures

Multiple connectivity "score" functions proposed



Imran Shah

Applications of Connectivity Mapping

- **Group chemicals by Mode of Action (MOA)**, e.g.:
  De Abrew, et al. *Tox Sci* (2016) DOI: 10.1093/toxsci/kfw058

- **Predict mechanism** based on similarity to reference chemicals, e.g.:
  Wang, et al. *BMD Genomics* (2016) DOI: 10.1186/s12864-016-2406-y

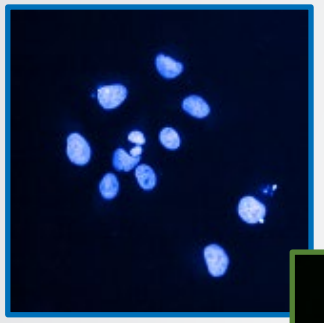
- Select chemicals for **Read Across analysis**, e.g.:
  De Abrew, et al. *Toxicology* (2019) DOI: 10.1016/j.tox.2019.05.008
  GenRA: Helman, et al. *ALTEX* (2019) DOI: 10.14573/altex.1811292

Cell Painting

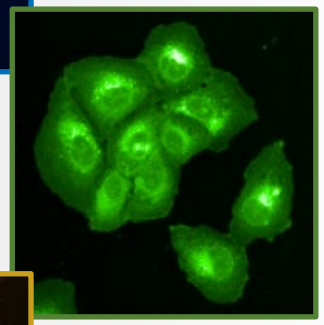
High-Throughput Phenotypic Profiling (HTPP) of cells using High-Content Imaging

Cell Painting with Multiple Markers

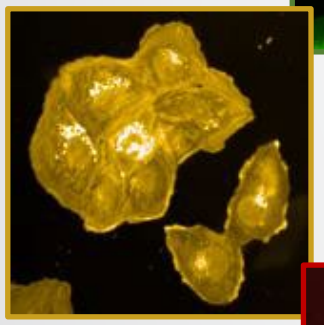
DNA



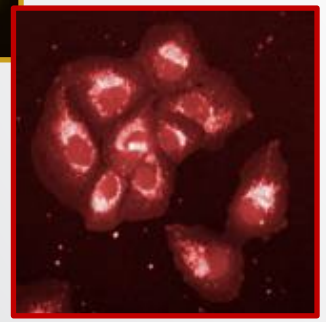
RNA + ER



Golgi + membrane + actin skeleton



Mitochondria

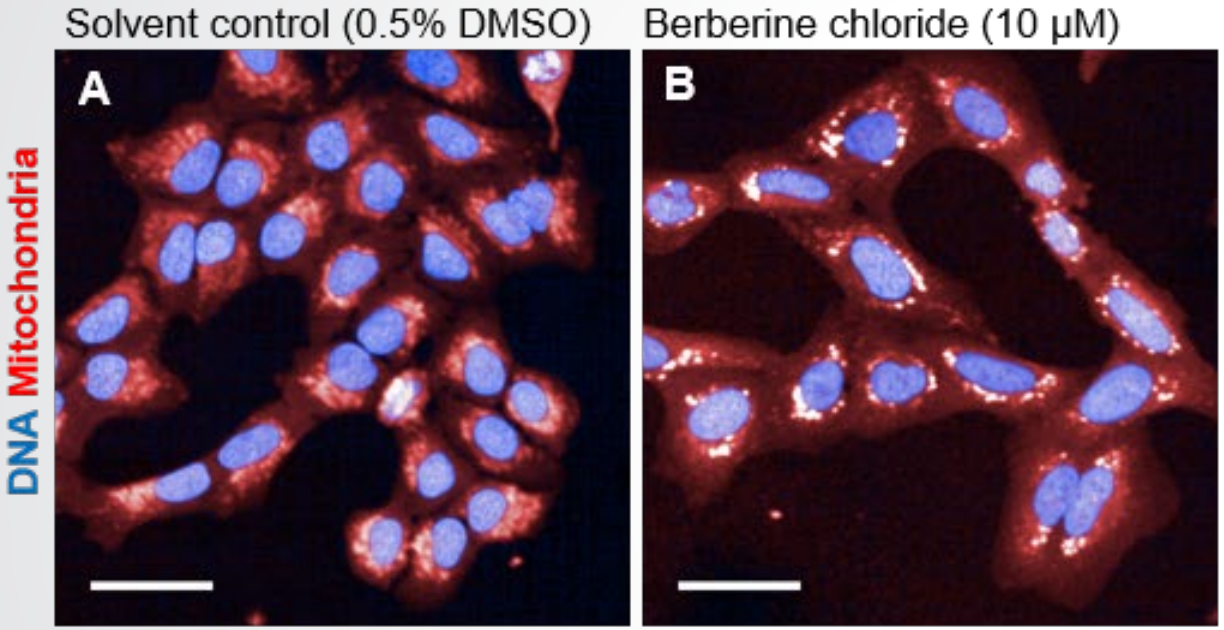


- Measures a large variety of phenotypic features in fluoroprobe labeled cells *in vitro*.
- Originally introduced in Bray, et al. *Nat Protoc* (2016) DOI: 10.1038/nprot.2016.105

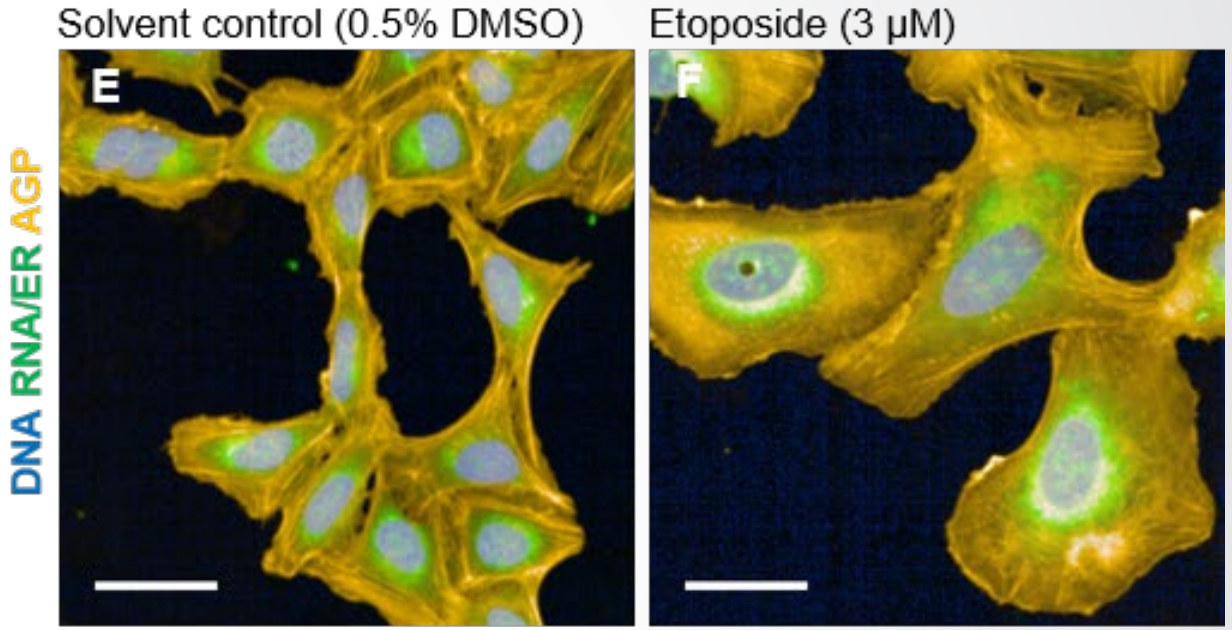
Marker	Cellular Component	Labeling Chemistry
Hoechst 33342	Nucleus	Bisbenzamide probe that binds to dsDNA
Concanavalin A – AlexaFluor 488	Endoplasmic reticulum	Lectin that selectively binds to α-mannopyranosyl and α-glucopyranosyl residues enriched in rough endoplasmic reticulum
SYTO 14 nucleic acid stain	Nucleoli	Cyanine probe that binds to ssRNA
Wheat germ agglutinin (WGA) – AlexaFluor 555	Golgi Apparatus and Plasma Membrane	Lectin that selectively binds to sialic acid and N-acetylglucosaminyl residues enriched in the trans-Golgi network and plasma membrane
Phalloidin – AlexaFluor 568	F-actin (cytoskeleton)	Phallotoxin (bicyclic heptapeptide) that binds filamentous actin
MitoTracker Deep Red	Mitochondria	Accumulates in active mitochondria

Joshua Harrill

Example Chemicals



→ Mitochondrial compactness/texture

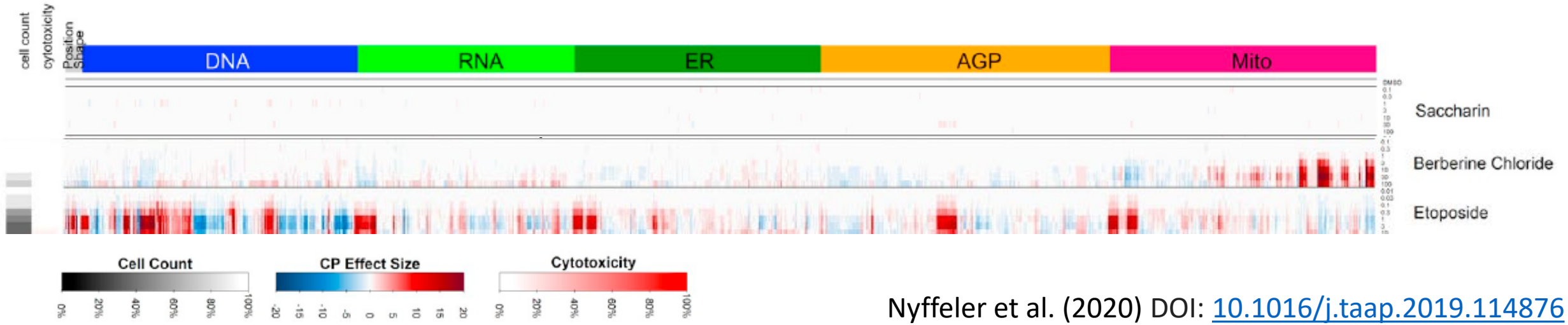


→ Cells are larger

Strong phenotypes are observable qualitatively and can be measured quantitatively using imaging processing software

Nyffeler et al. (2020) *Tox Appl Pharm* DOI: 10.1016/j.taap.2019.114876

Quantification of Cellular Features



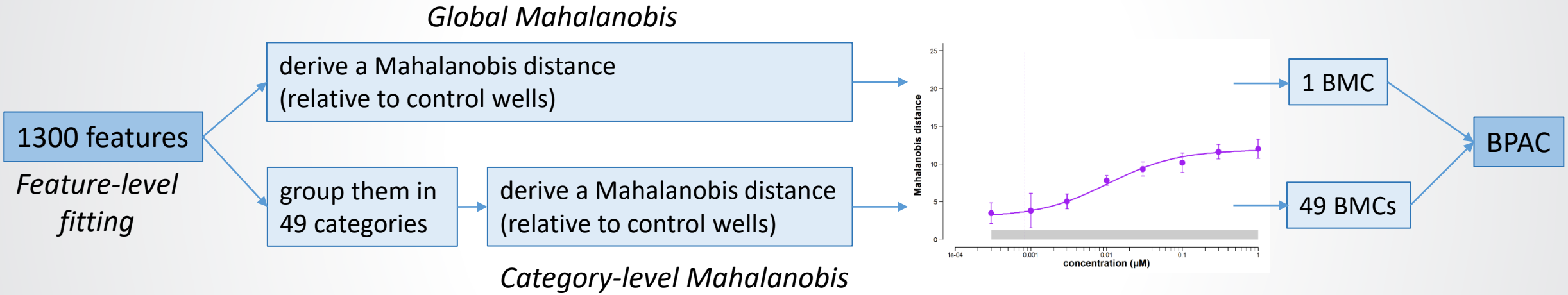
Nyffeler et al. (2020) DOI: 10.1016/j.taap.2019.114876

- Image analysis software quantifies multiple features per cell & fluorescence channel (intensity, size, texture, etc.): CellProfiler *(open source)* Harmony *(commercial)*
- Cell-level features are summarized per well & normalized to controls
- Different chemicals induce distinct, dose-responsive profiles

Joshua Harrill

Computing Overall POD (Cell Painting)

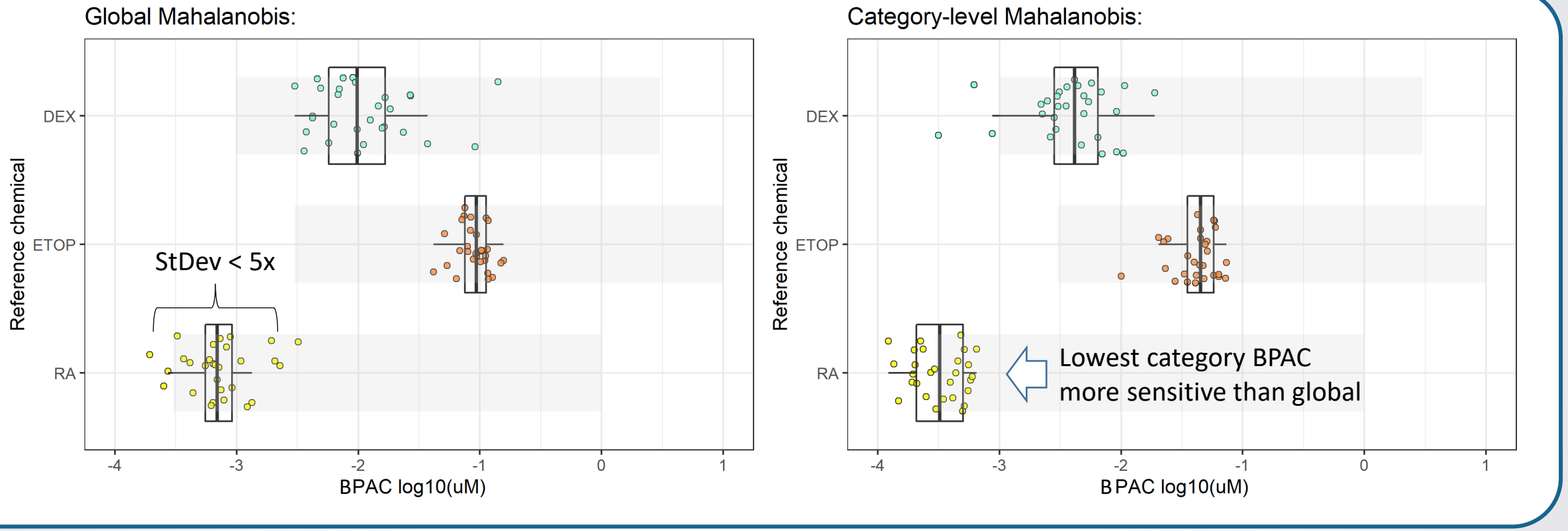
Mahalanobis Distance (D_M): A multivariate distance metric that measures the distance between a point (vector) and a distribution.



- Chemicals where a BMC can be determined using either the global or category D_M approach are considered active.
- The minimum of the global or most sensitive category BMC is the **Biological Phenotype Altering Concentration** (BPAC)

Joshua Harrill

Assay Performance / Reproducibility

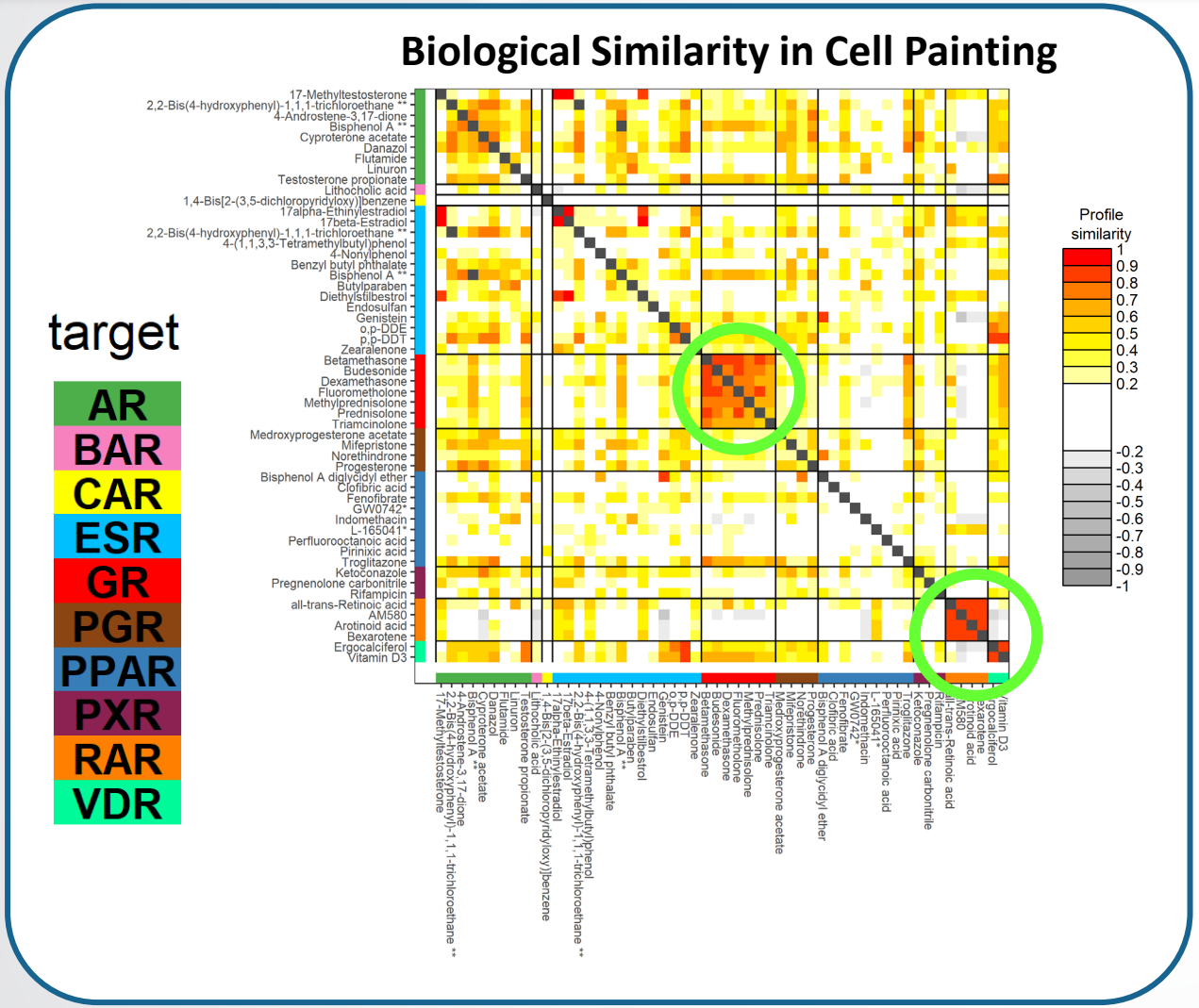


Used distance-based metrics (Mahalanobis) to derive the
Biological Phenotype Altering Concentration (BPAC)

BPAC	
Retinoic Acid:	~ 0.3 nM
Dexamethasone:	~ 3 nM
Etoposide:	~ 30 nM

Joshua Harrill

Profile Similarity Reflects Common Targets



- Cell Painting results for multiple chemicals in U-2 OS Cells
- Glucocorticoid & Retinoic Acid agonists display characteristic profiles
- Potential for similar applications to Connectivity Mapping in transcriptomics!

Joshua Harrill

Summary

- High-throughput/broad profiling technologies can assess many types of bioactivity at once
- No single best way to analyze the data – depends on technology, experimental design, and use-case!
- Transcriptomics & Cell Painting can be used for:
  - Mechanism-agnostic POD determination, e.g. transcriptional POD (tPOD)
  - Mechanistic/Mode of Action (MOA) inference
  - Guiding/prioritizing further targeted testing
  - Generalized Read Across

Questions?

everett.logan@epa.gov

Logan J. Everett, Ph.D.

Bioinformatics Scientist

Computational Toxicology & Bioinformatics Branch

Biomolecular & Computational Toxicology Division

Center for Computational Toxicology & Exposure

Office of Research and Development, U.S. EPA

Research Triangle Park, North Carolina. U.S.A.

Contributors

Joshua Harrill

Johanna Nyffeler

Richard Judson

Imran Shah

CCTE Leadership

Rusty Thomas

Sid Hunter

John Cowden

Kimberly Slentz-Kesler