A Data Analysis Pipeline Accounting for Artifacts in Tox21 Quantitative High Throughput Screening Assays

Jui-Hua Hsieh, Alexander Sedykh, Ruili Huang, Menghang Xia, and Raymond R. Tice

Table of Contents

Supplemental Material, Methods

Pipeline components

Preview

The flowchart of pipeline with KNIME (http://knime.org, version 2.92) workflow names can be found in the Supplemental Material, Figure S2.

Data storage

The plate-level data processed by the pipeline are stored using RData format (R version: 3.01) in Odum Institute Dataverse Network (http://arc.irss.unc.edu/dvn/dv/curvepwauc). The files are named based on the rule: ${pathway}_${readout}.RData. The ${} notation represents a variable. The included pathways can be found in Table 2. The readout information is listed in the section of "Tox21 assays" in the main article. The plate-level column description can be found in Supplemental Material, Table S4. All the R scripts (R version: 3.01) and KNIME workflows in each component at plate-level, source-level, and compound-level are also uploaded to the Odum Institute Dataverse Network (http://arc.irss.unc.edu/dvn/dv/curvepwauc). The response pattern recognition algorithm in terms of carry-over and U-shape detection (Supplemental Material, p. 10 and 13) has been implemented into Curvep. The source code can be found in the GitHub (https://github.com/sedykh/curvep)

Data visualization

The plate-level concentration-response curves can be visualized in the public Rstudio Shiny server (http://spark.rstudio.com/moggces/plotting). The compound-level activity profiling results can be visualized in the http://spark.rstudio.com/moggces/profiling/. The source code can be

found in Github (https://github.com/moggces/ActivityProfilingGUI/archive/v1.0-beta.zip and https://github.com/moggces/CurveVisualizationGUI4Tox21/archive/v1.0-beta.zip)

Plate level

NCATS data standardization

The goal of this component is to standardize NCATS data with additional informative columns. The component includes an R script file (standardize_NCATS.r) implemented into KNIME workflow (standardize_NCATS.zip)

Input files

The tab-delimited text files (${basename}.txt, e.g., basename: tox21-p53-bla-p1_ch2) downloaded from NCATS Tox21 internal Gateway http://tripod.nih.gov/tox/

Output files

The standardized NCATS data files (${basename}_std.txt, e.g., tox21-p53-bla-p1_ch2_std.txt) with additional columns

Output columns

The pathway, readout, Cmpd_Library, Library, Library_seq, uniqueID, Tox21AgencyID, Chemical.ID, and Chemical.Name parameters are generated. For plate-level column description, please see Supplemental Material, Table S4.

Signal artifacts handling

The goal of this component is to apply Curvep with appropriate input parameters to clean the artifacts and with additional masking options (see Supplemental Material, Figure S4) for severe non-monotonic concentration-response data seen in the activation-type assays. The main R script (run_curvep.r) is implemented in the KNIME workflow (run_curvep.zip)

Dependent scripts/programs

io.r, get.r, Curvep

Input files

The standardized NCATS data files from the KNIME workflow of standardize_NCATS.zip or tab-delimited text file with appropriate format:

- pathway column: *pathway*, the assay identifier
- readout column: *readout*, the readout identifier
- unique ID column by pathway: *uniqueID*
- concentration columns: *conc[0-9]+* , (e.g., conc0, conc1, conc2, ... etc.), log10(M) transformation
- response columns: *resp[0-9]+* (e.g., resp0, resp1, resp2, ... etc.), -100% ~ 0% ~100%; 0% is the baseline

The words with italic font represent column names.

Input parameters

rename_text – the string to be appended to the original file name. Although any input is allowed, some controlled terms are already employed: u (setting assuming increasing signal); d (setting assuming decreasing signal); u0 (self-masking with setting assuming increasing signal); u1 (ch1-masking with setting assuming increasing signal)

pathway – the pathway name for filtering the files based on file names

readout – the readout name for filtering the files based on file names

direction – the string value to indicate whether increasing response (“up”) or decreasing response (“down”) should be treated by Curvep

thr – Curvep baseline noise threshold (*THR*)

thr_cro – Curvep carryover threshold (*CRO*). Set as “default” to use the default values. See Supplemental Material, Methods (Supplemental Material, p. 10) for more information about the *CRO*.

selfmask – whether self-masking should be conducted (default: no). If yes, the ${basename}_curvep_d.txt needs to be available

ch1mask – whether ch1-masking should be conducted (default: no). If yes, the ${basename%_*}_ch1_curvep_twoD.txt needs to be available

Output files

The tab-delimited text files (${basename}_${rename_text}.txt) with additional Curvep columns. If the input data files are processed by both self-masking and ch1-masking setting, the ch1-masking data are used only when the curve is flat in the self-masking setting.

Output columns

Curvep-related columns are appended. See Supplemental Material, Table S4

Quality control

The goal of this component is to check the reproducibility between plates for Tox21 data after Curvep. The R script is provided (qc.r).

Input files

The plate-level Rdata files (${basename}_${readout}.RData, e.g., ar-bla-antagonist_ratio.RData) or output after KNIME workflow (run_curvep.zip).

Input parameters

dmso_sd – the SD value of responses in DMSO control plates

curvep_thr – the Curvep baseline noise threshold (*THR*)

metric – a string value to indicate the type of signals (wauc or pod)

Output

T_1 & T_2 (assay-dependent wAUC thresholds. see "hit calling/ranking" in the main article)

The statistics are stored in an R list structure, including Pearson's r correlation between plates, intraclass correlation coefficient when using T_1 (or 0) and T_2 to categorize signals into: two groups (hit and others), or three groups (active, marginal active, and others), or five groups (strong increasing signal, weak increasing signal, inactive, weak decreasing signal, strong decreasing signal), and percentage of signals based on different categories.

Source level

wAUC-related data collapsing #1

The goal of this component is to 1) combine the two directions (increasing/decreasing) of signal if available and to 2) calculate source-level signal statistics. The R script (collapse_source_wauc.r) is implemented in the KNIME workflow (collapse_source_wauc.zip).

Input files

The files generated from KNIME workflow run_curvep.zip. For ${basename}_curvep_u.txt (data are cleaned using the setting for increasing signal), the ${basename}_curvep_d.txt file may be used to merge the data if "istwoD" parameter is set to TRUE.

Input parameters

istwoD – shows whether the curvep output file (${basename}_curvep_d.txt) should be used in the merging. When merging, the results from the preferred direction (determined from pathway/readout fields) have higher priority.

by_type – the type of chemical ID in the source level file (either Tox21.ID or Tox21AgencyID) used for collapsing.

Output files

If both signals from two directions are available, the merged tab-delimited text file (${baseman}_curvep_twoD.txt). This merged file is further converted to the reported RData file (${pathway}_${readout}.RData). Otherwise, the original input file will be renamed as (${baseman}_curvep_oneD.txt.

Also, source-level signal data files are generated. The files are named as ${readout}_${pathway}.wauc.source.txt if the readout is not *ratio*, *luc*, or *via*. Otherwise, ${readout}_${pathway}_main.wauc.source.txt is used.

Output columns

The descriptions of source-level columns can be found in Supplemental Material, Table S5.

Compound-dependent artifacts flagging

The goal of this component is to integrate various data sources to flag the potential artifacts (KNIME workflow flag_artifacts.zip).

Input files

Source-level auto-fluorescence data are needed for β-lactamase reporter gene assays as well as the readouts from either β-lactamase reporter gene assay or luciferase reporter gene assay.

Input parameters

pathway (name string) and pod_thr (max.difference in log-conc units between PODs of cytotoxic and inhibitory signals, e.g., $POD_{viability} - POD_{inhibition} < -0.5$)

Output files

The pathway-based, tab-delimited text files ${pathway}_main.wauc.source.txt (e.g., p53-bla_main.wauc.source.txt)

Output columns

Tox21.ID, Library, Cmpd_Library, curvep_wauc.[ratio|ch2|ch1|via|luc], curvep_pod.[ratio|ch2|ch1|via|luc], med_curvep_wauc[ratio|ch2|ch1|via|luc], med_curvep_pod.[ratio|ch2|ch1|via|luc], max_curvep_wauc.[blue|red|green], Comment, pvalue, mean_pod_diff, and med_pod_diff.

The compound ID column information can be found in Supplemental Material, Table S4. The source-level column descriptions can be found in Supplemental Material, Table S5.

Compound level

wAUC-related data collapsing #2

The goal of this component is to collapse the source-level activity data to the compound-level activity data, to generate hit calls, and to normalize wAUC [0-1]. The R script (collapse_cmpd_wauc.r) is implemented in the KNIME workflow (collapse_cmpd_wauc.zip).

Input file

The pathway-based, tab-delimited text files: ${pathway}_main.wauc.source.txt (e.g., p53-bla_main.wauc.source.txt) or the readout-based, tab-delimited text files: ${readout}_${pathway}_main.wauc.source.txt.

Input parameters

assay_id – the pathway (e.g., p53-bla)

T_1 & T_2 – see the "Quality control" (Supplemental Material, p.6)

data_type – the value (either activation or inhibition) to indicate whether the assay is activation-like or inhibition-like assay

Output files

Pathway-based, tab-delimited text file with compound-level activity information (${pathway}_main.wauc.cas.txt) or readout-based, tab-delimited text file with compound-level signal information (${readout}_${pathway}.wauc.cas.txt)

Output columns

The descriptions of compound-level columns can be found in Supplemental Material, Table S6.

Response pattern recognition

The examples of curves can be found in Table 1 and Supplemental Material, Figure S3.

Carry-over and baseline shift

Compound carry-over (Figure S3f) is caused by compounds that resist the pin-tool cleaning steps between plate runs and result in the activity due to the compound from the previous set of plates (at the highest concentration) being transferred to the next set of plates (which start at the lowest concentration for a different compound) (Table 1). It substantially affects wAUC. If the compound is inactive in the current plates, monotonically decreasing (increasing) responses could be observed in activation-type (inhibition-type) assays. These curves are corrected as baselines. However, if the compound is active in the current plates, the addition of transferred activity causes a baseline shift. Therefore, additional treatments are applied to adjust the part that caused by carry-over to baseline.

Pseudo code for carry-over detection (by Curvep)

The carry-over pattern identification is controlled by a carryover threshold (CRO) and maximum deviation (MXDV)[1]. The default value is set as 80% (CRO) and 5% (MXDV). For activation-type assays, a smaller CRO (e.g., 60%) is suggested for noisier assays. However, for stress response pathway assays, which have low noise (3SD < 10%), 30% could be used. The following code only applies to the carry-over in activation-type assays. For inhibition-type/cytotoxicity assays, the direction is simply reverse.

LOOP through the curves after noise filtering

Find the highest concentration (c) of the lowest response (m) (or the responses within MXDV)

Find the maximum responses before (rb) and after (ra) the c

IF c is the highest tested concentration

ra = m

IF (rb == 0)

normal curve, continue to next

IF rb < CRO

IF the curve is monotonically decreasing (rb – ra > MXDV)

carry-over

IF the curve is flat (abs(rb - ra) <= MXDV)

carry-over

IF the curve is monotonically increasing (ra – rb > MXDV)

baseline shift due to carry-over

IF rb >= CRO

IF the curve is monotonically decreasing (rb – ra > MXDV)

IF response type is inhibitory

carry-over

ELSE

undetermined (potential strong active)

IF the curve is flat (abs(rb - ra) <= MXDV)

undetermined (potential strong active)

END LOOP

To clarify whether the "undetermined" is a strong active (Figure S3e) or an artifact requires the incorporation of plate sequence information. If a signal from the previous plate is significant or is a carry-over resultant, the "undetermined" signal in current plate is highly probable to be an artifact – otherwise a strong active, where the response is saturated.

U-shape by masking

The U-shape curve violates the monotonicity assumption and represents mostly non-reproducible artifacts in qHTS assays. However, some of these responses could be real signals (judging by the occurrence of the same spikes in three runs) and can be handled by incorporating readout information as a mask (Supplemental Material, Figure S4). For example, in the self-masking component, the output of Curvep from the negative direction (response between 0% and -100%) is used to mask some of the responses in positive direction (response between 100% and 0%). Thus, the responses become monotonic (Figure S3b). A similar strategy can be applied (ch1-masking) by using ch1 (background readout in β-lactamase assay) to mask some of the responses in ch2 (gene expression readout) (Figure S3c).

U-shape by Curvep

Some of the U-shapes (real signals) cannot be handled by the readout masking but can be handled by Curvep itself (Figure S3d). The pseudo code handling U-shapes uses USHAPE parameter (set to 4; for low-noise assay, USHAPE=3 is applied) and is given below. Briefly, it detects an optimal pivot point (where change in monotonicity occurs), then masks the points on one of the shoulders (the other one will be retained).

LOOP through the curves

IF responses at lowest and highest test concentrations are different by more than THR

    normal curve, continue to next

IF standard deviation of responses < MXDV

    noisy flat curve, continue to next

Find the pivot point (c) with minimum number of corrections to restore monotonicity (e)

Find the number of points (p) in the plateau and the peak slopes around the pivot c

IF e > p AND response at c < CRO

    too many corrections, treat as flat noisy curve and continue to next

Find total number of spikes in the curve (x), i.e., number of changes in monotonicity

IF p < USHAPE OR x > p

    noisy curve with spikes, flatten and continue to next

Find the number of points in the retained shoulder of the U-shape curve (s)

IF 2*s < USHAPE OR c is lowest test conc.

    Not a true U-shape, treat as flat noisy curve and continue to next

U-shape curve is found. Correct e and treat as monotonic.

END LOOP

References

1. Sedykh, A.; Zhu, H.; Tang, H.; et al. Use of in Vitro HTS-Derived Concentration–Response Data as Biological Descriptors Improves the Accuracy of QSAR Models of in Vivo Toxicity. Environ. Health Perspect. 2011, 119, 364–370.

Supplemental Material, Tables

Supplemental Material, Table S1: the list of assays categorized by their signal type

assay	signal type	channel number	readout	cell type	SD (%)	T_1	T_2
ATAD5	activation	1	luminescence	HEK293	1.6	1.1	11.8
p53	activation	2	fluorescence	HCT116	2.8	1.2	25.6
AHR	activation	1	luminescence	HEPG2	3.0	1.5	23.5
AR (full)	activation	1	luminescence	MDAkb2	6.0	2.1	41.0
AR (partial)	activation	2	fluorescence	HEK293	7.4	3.1	28.7
ER (full)	activation	1	luminescence	BG1	7.7	3.3	21.7
ER (partial)	activation	2	fluorescence	HEK293	3.8	1.3	14.6
GR	activation	2	fluorescence	HeLa	3.6	1.3	24.8
TR	activation	1	luminescence	GH3	3.4	1.2	9.8
PPARγ	activation	1	fluorescence	HEK293	3.6	1.5	11.1
ATAD5	cytotoxicity	1	fluorescence	HEK293	6.2	3.6	23.4
p53	cytotoxicity	1	luminescence	HCT116	8.4	3.6	25.8
100 (mutant)	cytotoxicity	1	luminescence	DT40	7.8	3.3	35.5
653 (wild)	cytotoxicity	1	luminescence	DT40	7.8	3.3	33.0
657 (mutant)	cytotoxicity	1	luminescence	DT40	8.4	3.6	33.6
mitotox	cytotoxicity	1	luminescence	HepG2	4.5	2.3	21.4
AHR	cytotoxicity	1	fluorescence	HepG2	6.4	3.2	23.2
AR (full)	cytotoxicity	1	fluorescence	MDAkb2	4.8	2.4	25.4
AR (partial)	cytotoxicity	1	luminescence	HEK293	7.2	3.6	23.6
ER (full)	cytotoxicity	1	fluorescence	BG1	5.5	2.8	28.1
ER (partial)	cytotoxicity	1	luminescence	HEK293	9.6	4.1	24.8
GR	cytotoxicity	1	luminescence	HeLa	9.6	3.3	31.2
TR	cytotoxicity	1	fluorescence	GH3	9.6	4.1	33.2
aromatase	cytotoxicity	1	fluorescence	MCF7	6.1	3.0	26.6
mitotox	inhibition	2	fluorescence	HepG2	8.3	4.2	32.0
AR (full)	inhibition	1	luminescence	MDAkb2	8.4	3.6	27.5
AR (partial)	inhibition	2	fluorescence	HEK293	8.8	3.7	28.0
ER (full)	inhibition	1	luminescence	BG1	5.9	3.0	26.7
ER (partial)	inhibition	2	fluorescence	HEK293	6.4	3.2	24.7
GR	inhibition	2	fluorescence	HeLa	9.2	3.9	30.3
TR	inhibition	1	luminescence	GH3	7.0	3.5	34.0
aromatase	inhibition	1	luminescence	MCF7	7.1	3.0	30.4

Abbreviations: AHR: aryl hydrocarbon receptor: AR; androgen receptor; ER: estrogen receptor; GR: glucocorticoid receptor; mitotox: mitochondria toxicity; PPARγ: peroxisome proliferator-activated receptor gamma; TR: thyroid receptor; T_1 and T_2 are the two wAUC thresholds.

Supplemental Material, Table S2: the comparison between the wAUC and AC_{50} from high-quality curves

assay name	activation-type		inhibition-type	
	correlation	frequency*	correlation	frequency**
ahr	0.43	4737	NA	
atad5	0.66	1004	NA	
aromatase	NA		0.90	4267
bg1er	0.85	3803	0.84	2764
gr	0.89	1610	0.83	3123
hek293ar	0.80	1639	0.77	5660
hek293er	0.88	2401	0.77	2843
mdakb2ar	0.76	1431	0.85	3650
mitotox	NA		0.85	6767
p53	0.51	2843	NA	
pparg	0.63	1191	NA	
tr	0.79	182	0.90	6902

*the AC_{50} from Curve Class = 1.1, 1.2, 2.1, 2.2

**the AC_{50} from Curve Class = -1.1, -1.2, -2.1, -2.2

Abbreviations: ahr: aryl hydrocarbon receptor; atad5: ATAD5 protein; bg1er: estrogen receptor in BG1 cell; gr: glucocorticoid receptor; hek293ar: androgen receptor in Hek293 cell; hek293er: estrogen receptor in Hek293 cell; mitotox: mitochondria toxicity; pparg: peroxisome proliferator-activated receptor gamma; tr: thyroid receptor; NA: not applicable

Supplemental Material, Table S3: the percentage of signal groups (n=8306)

Pathway	strong(+) (%)	weak(+) (%)	Inactive (%)	weak(-) (%)	strong(-) (%)
ahr_agonism	6.50	10.84	80.64	1.91	0.11
ar_agonism(hek293)	4.13	3.12	87.14	4.49	1.12
ar_agonism(mdakb2)	3.48	2.55	89.32	4.47	0.18
ar_antagonism(hek293)	0.61	1.95	76.85	10.96	9.63
ar_antagonism(mdakb2)	4.25	3.92	77.34	8.78	5.71
aromatase_antagonism	4.94	5.89	71.23	10.62	7.33
atad5	2.28	2.68	86.96	6.53	1.55
dna_damage(dsb)	0.77	7.36	77.76	12.05	2.06
dna_damage(srf)	0.85	8.86	77.39	11.08	1.82
er_agonism(bg1)	7.95	8.52	73.13	6.62	3.78
er_agonism(hek293)	4.00	6.73	89.09	0.17	0.01
er_antagonism(bg1)	4.70	3.66	77.71	9.15	4.78
er_antagonism(hek293)	1.81	4.15	80.42	8.22	5.39
gr_agonism	3.68	5.43	85.28	5.19	0.42
gr_antagonism	5.02	4.43	77.21	8.20	5.14
Mitotox	3.91	5.51	70.32	9.78	10.47
p53	5.29	9.28	85.20	0.18	0.05
pparg_agonism	3.14	3.19	81.50	8.15	4.02
tr_agonism	0.57	1.79	81.71	11.73	4.20
tr_antagonism	0.31	1.85	70.05	14.98	12.81

Abbreviations: ahr: aryl hydrocarbon receptor; ar: androgen receptor; dsb: double strand break; er: estrogen receptor; gr: glucocorticoid receptor; mitotox: mitochondria toxicity; pparg: peroxisome proliferator-activated receptor gamma; srf: stalled replication fork; tr: thyroid receptor; +/- represents the direction of the signal (increasing/decreasing)

Supplemental Material, Table S4: plate-level column descriptions

concentration-response data		
column names	column descriptions	potential value range or categories
pathway	e.g., p53	see Table 2
readout	e.g., ratio	[ch2|ch1|via|ratio|luc].batch#
conc[0-9]+	concentrations, log10(M)	-∞
resp[0-9]+	responses normalized by plate positive controls and negative controls	-∞ ~ ∞
curvep_r[0-9]+	responses after run_curvep.zip	-∞ ~ ∞
curvep-related parameters		
curvep_wauc	wAUC	-∞ ~ ∞
curvep_pod	POD, log10(M)	-∞
curvep_remark	warnings	OK, CHECK, CARRY_OVER, PART_U?, INVERSE, BLIP, U_SHAPE, BASE_SHIFT, SINGLE_POINT_ACT, NOISY
curvep_n_corrections	# of points corrected by Curvep	integer
curvep_mask	the mask that curvep employed to report the results	binary data string (1: the masked point)
input_mask	initial mask as the input for Curvep	binary data string (1: the masked point)
Chemical Identifiers		
Tox21.ID	Tox21 chemical ID	Tox21_[0-9]{6}_[0-9]
CAS	Chemical Abstract Service register number from EPA DSSTox	
Chemical.Name	from EPA DSSTox database	
Chemical.ID.GSID	GSID in EPA DSSTox database	
Cmpd_Library	compound libraries	[NTP|EPA|FDA]_[A-C]
Library	plate information	[NTP|EPA|FDA]_[A-C]_[1-3]
Library_seq	the sequence # of plate screening	integer
Row	row on the plate	integer
Column	column on the plate	integer
Tox21AgencyID	Tox21.ID @ Cmpd_Library	
Supplier	company which provides the chemical	

uniqueID	the base element used as input for Curvep at plate-level	N[0-9]+

Supplemental Material, Table S5: columns unique in source-level data

column names	column descriptions	potential value range or categories
med_curvep_[wauc|pod]	median collapsing of wAUC or POD data between plates	-∞ ~ ∞
max_curvep_wauc	maximum wAUC value in the auto-fluorescence data channels across cell types	0 ~ ∞
Comment	detailed flag information at the source level	AberrantRatio (contradictory readout), AutoBlue, AutoGreen_major (blue or green auto-fluorescence), NotPrimaryResp (cytotoxicity)
pvalue	Student's t-test (one-tailed)	
[mean|med]_pod_diff	POD difference between primary pathway readout and the viability readout if available	-∞ ~ ∞
cv.wauc	the absolute value of correlation of variation of wAUC between plates	0 ~ ∞ , NA (inactive)
sd.pod	the SD of PODs between plates	0 ~ ∞

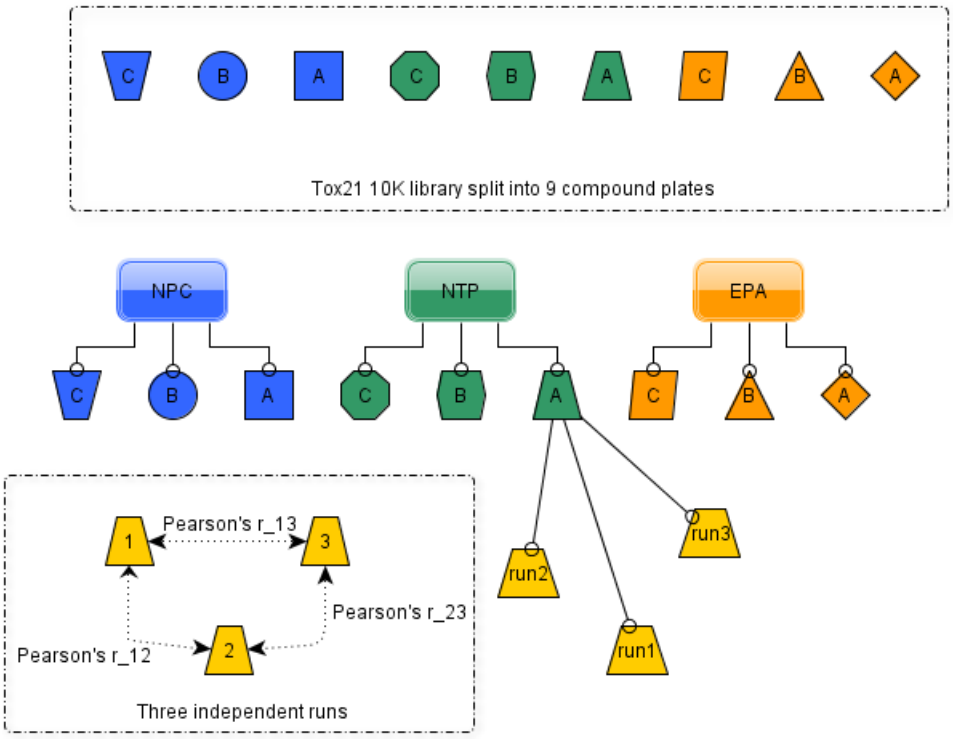
Supplemental Material, Table S6: columns unique to compound-level data

Property abbreviation	Property name	Property value range	Property value meaning
Properties related to activity			
nwauc.logit	overall activity (wAUC) based on logit normalization	-1 ~ 1	1. higher absolute number -> higher activity 2. positive value -> activity of assay interest 3. value > 0.05 -> active 4. negative value and 1E-4 -> inconclusive (assay interference)
hitcall	activity call	1, 0.5, 0, -0.5, -1, blank	1. value = 1 -> active 2. value = 0.5 -> marginal active 3. value = 0 -> inactive 4. negative value and blank -> inconclusive (assay interference)
npod	assay-dependent point-of-departure of activity	-∞ ~ ∞ (log10(M)), blank	1. higher absolute number -> more potent 2. positive value -> potency of assay interest 3. negative value and blank -> inconclusive (assay interference)
nac50	half maximal effect concentration (data collapsed from NCATS fitting results)	-∞ ~ ∞ (log10(M)), blank	same as npod
emax	maximal effect (data collapsed from NCATS fitting results)	- ∞ ~ ∞ (%)	similar to npod
cv.wauc	absolute value of	∞ , blank	1. higher value ->

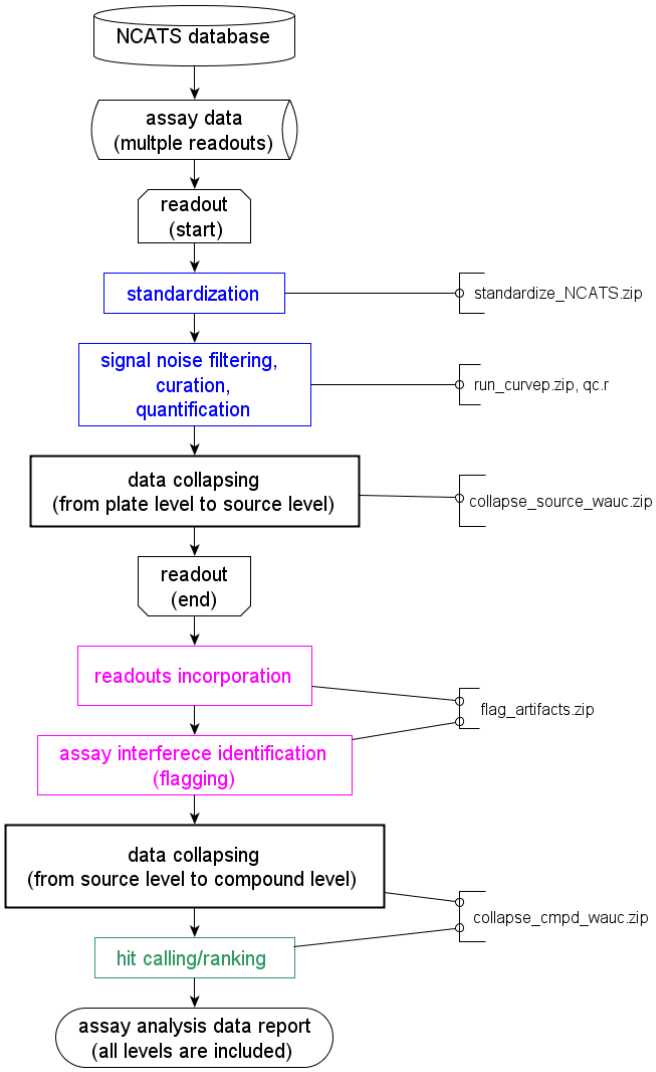
	correlation of variation between sources		higher variation 2. value > 1.4 -> at least one source is inactive 3. blank -> only one source available
pod_med_diff	log10 pod difference between primary signal and cytotoxicity signal	- ∞ ~ ∞ (log10 unit), blank	1. higher absolute number -> larger difference between these two signals 2. negative value -> primary signal is more potent (of assay interest) 3. blank -> no cytotoxicity
label	activity label	a_normal, b_autofluor, c_contradict, d_cytotoxic	a_normal -> normal b_autofluor -> auto-fluorescence c_contradict -> ch2 vs ratio issue d_cytotoxic -> cytotoxicity
a_normal	fraction of sources with “normal” label	0 ~ 100%, blank	1. higher value -> higher fraction of sources with normal label 2. blank -> only one source available
nwauc	unbounded overall activity (wAUC), raw	- ∞ ~ ∞, blank	1. higher absolute number -> higher activity 2. positive value -> activity of assay interest 3. negative value and blank -> inconclusive (assay interference)
n_collapsed	# of sources used to generate the provided activity data	positive integer	the denominator of the a_normal property
n_source	# of total sources screened	positive integer	
properties related to signal			

wauc	overall signal (wauc)	- ∞ ~ ∞	1. higher absolute number -> larger signal 2. positive value -> increasing signal 3. negative value -> decreasing signal
pod	assay-dependent point-of-departure	-∞ ~ ∞ (log10(M)), blank	1. higher absolute number -> higher potency of the signal 2. positive value -> increasing signal 3. negative value -> decreasing signal 4. blank -> no signal

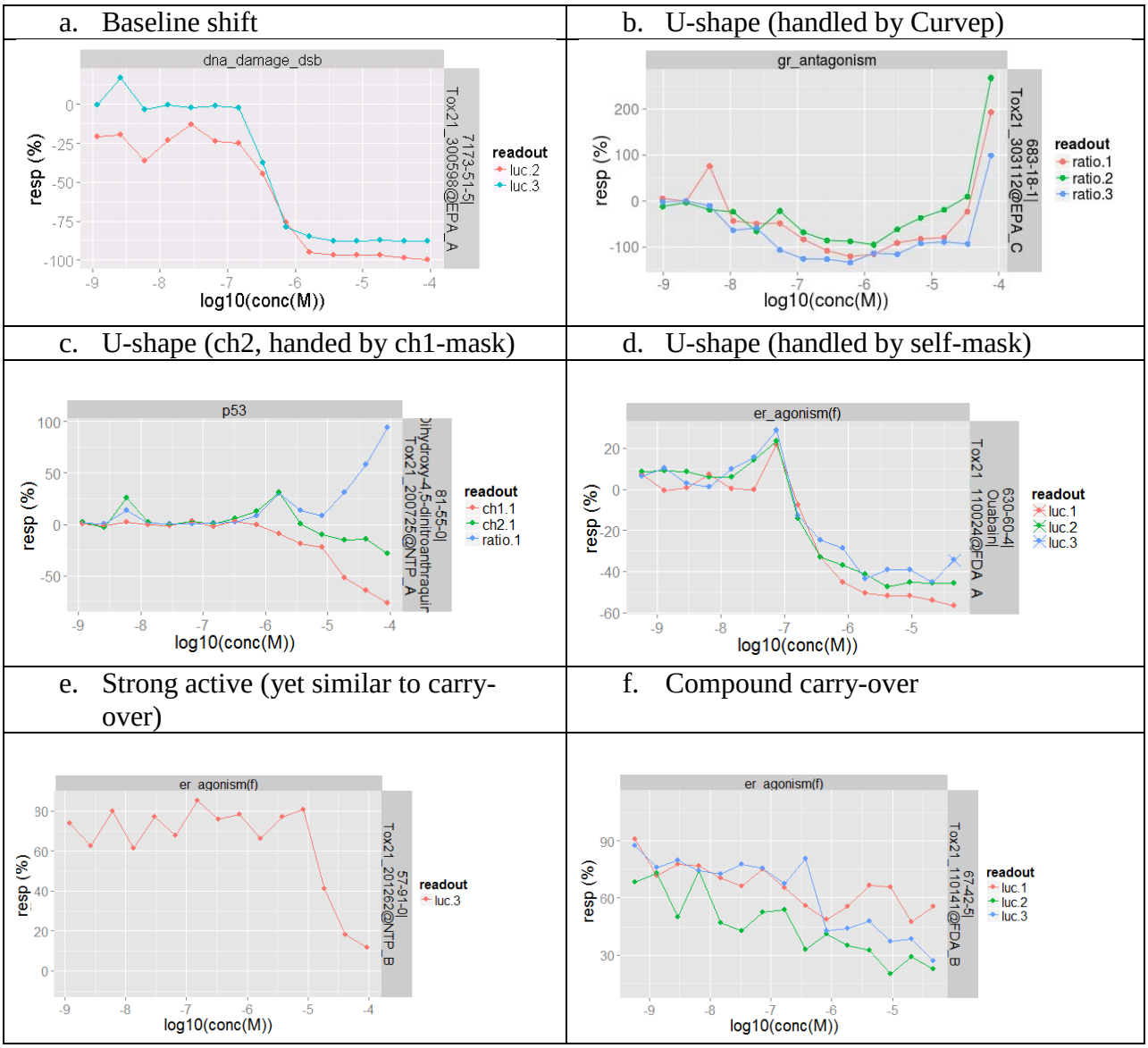
Supplemental Material, Figures



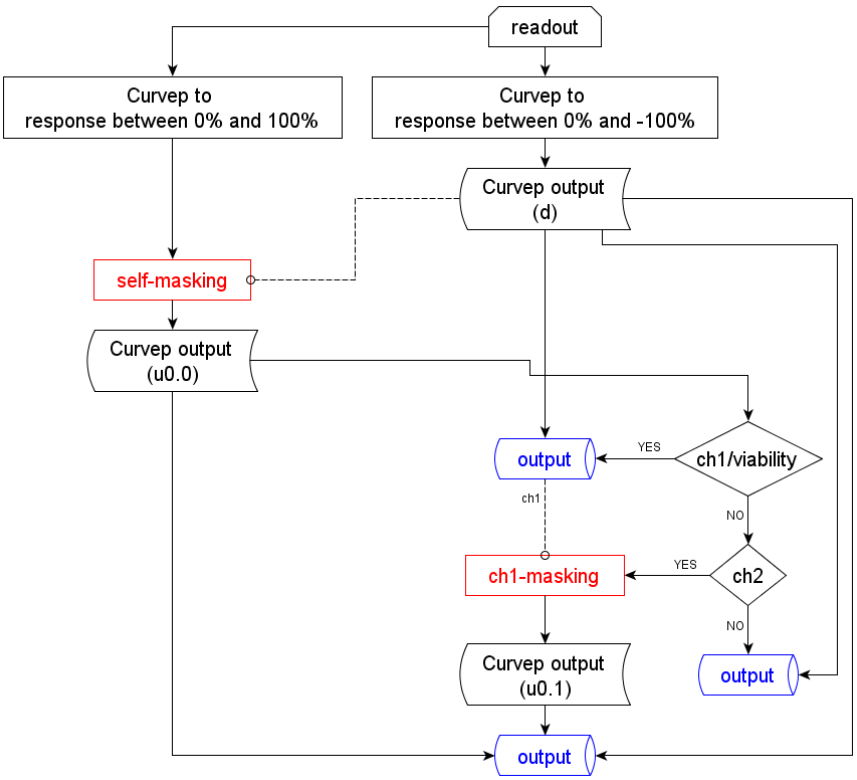
Supplemental Material, Figure S1: the Tox21 10K Library. The Tox21 library consists of nine different compound plates coming from three agencies: NCATS (NPC library), EPA, and NTP. Each compound plate is screened three times, each time on a separate day. The reproducibility (e.g., wAUC) between three different batches is compared using Pearson's r correlation coefficient (e.g., Pearson's r_13: the correlation coefficient between the batch 1 and batch 3). Thus, in total, there are 27 comparisons.



Supplemental Material, Figure S2: the pipeline components with KNIME workflow names (zip files). An R script for quality control of plate-level signals (qc.r) is also included.

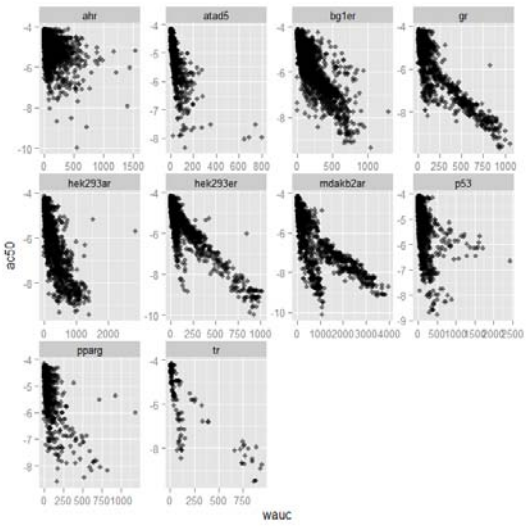


Supplemental Material, Figure S3: challenges in qHTS data analysis, continued. a) baseline shift, could be due to the compound activity carry-over. b) U-shape, handled by Curvep alone. c) U-shape of ch2, handled by using ch1 as mask. d) U-shape, handled by self-masking. e) strong active similar to carry-over. f) compound carry-over with monotonically decreasing signal. Note: Ratio" or "luc" represents the main readout in either β-lactamase assay or luciferase assay, respectively; "ch1" represents channel 1, the background in bla (β-lactamase assay) assay; "ch2" represents channel 2, the signal channel in bla assay; "via" represents cell viability. Numbers represent different batches.

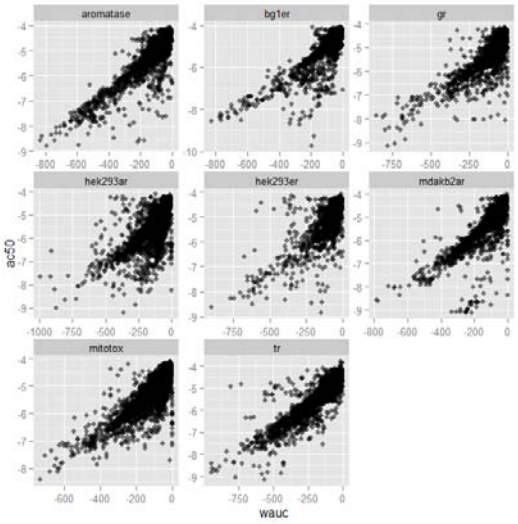


Supplemental Material, Figure S4: The masking flowchart for activation-type assays. The red components represent the two masking strategy: self-masking by responses between 0% and -100% after Curvep and ch1-masking particular for β-lactamase assays.

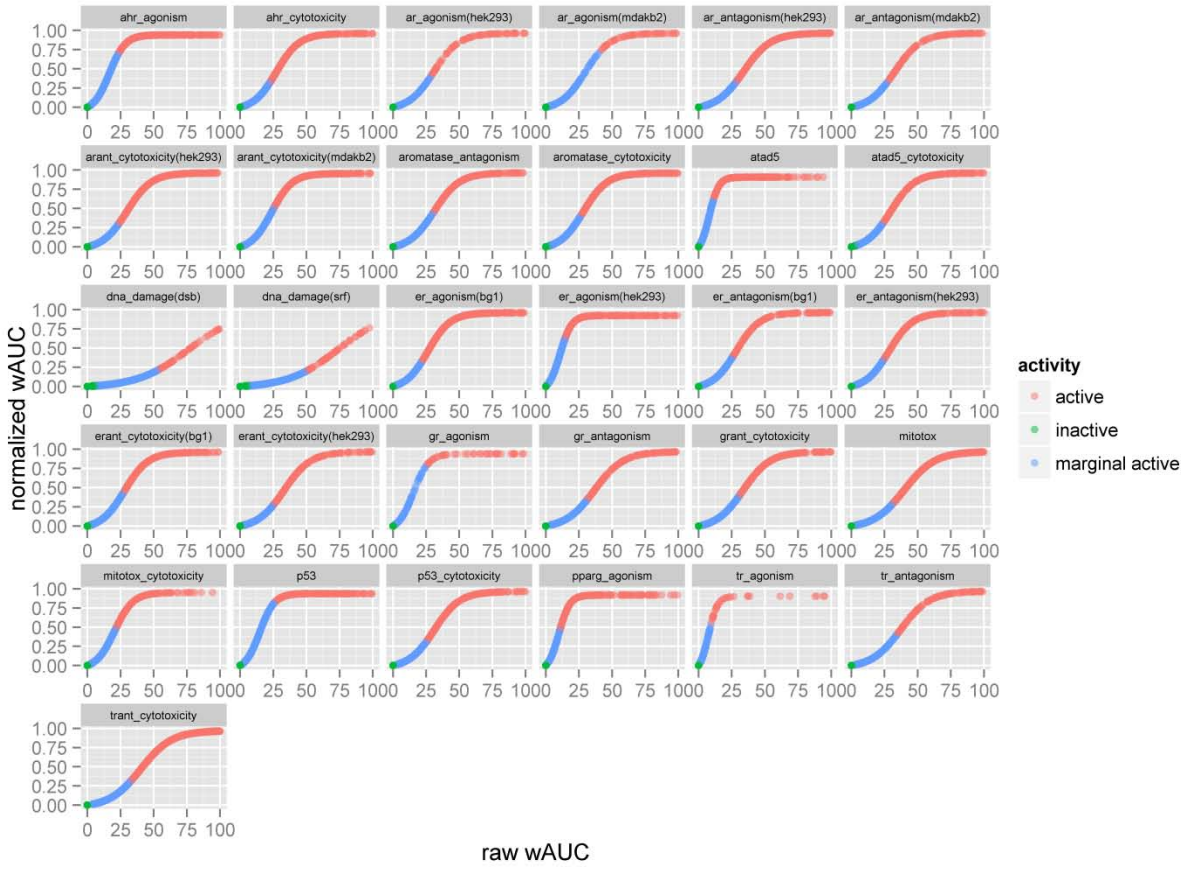
a.



b.

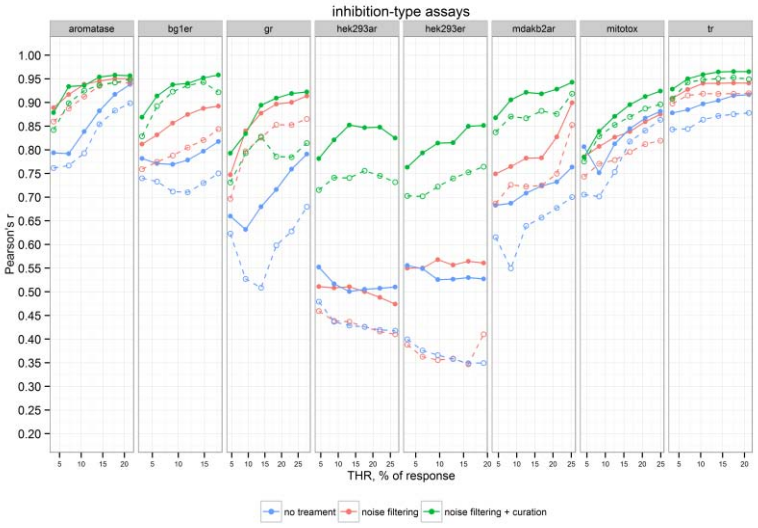


Supplemental Material, Figure S5: the relationship between the wAUC and AC_{50} from high-quality curves a. activation-type assays; b. inhibition-type assays

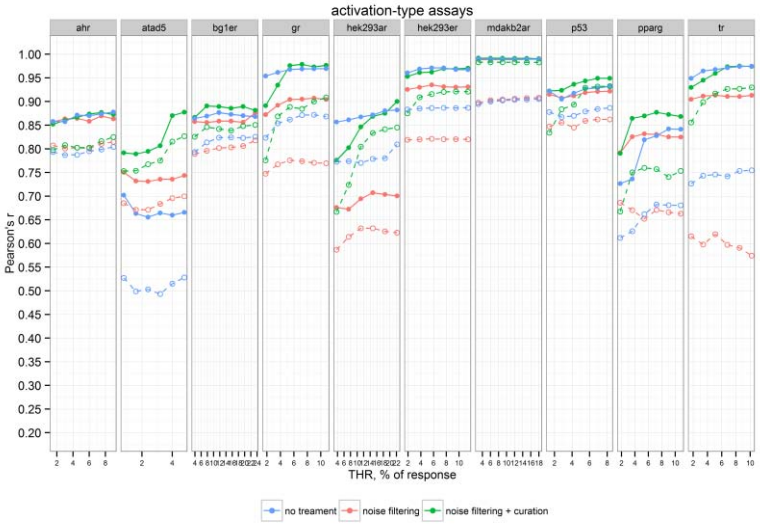


Supplemental Material, Figure S6: comparison between the raw wAUC values and the normalized wAUC values in assays. The color relates to the hit calls in assays. The T_1 and T_2 control the shape of curve. The curve shape is more flat in the inhibition-type assays and cytotoxicity assays than in the activation-type assays because of the higher background noise level in assays and the response satuaration feature (i.e., E_{max} can usually reach 100%) in curves; thus, in the inhibition-type assays and cytotoxicity assays, the signal needs to have higher raw wAUC value to achieve the same normalized wAUC value in the activation-type assays.

a



b



Supplemental Material, Figure S7: comparison of inhibition-type assay (a) and activation-type assay (b) reproducibility based on three signal processing protocols as a function of Curvep baseline noise threshold (THR). The solid/dashed line represents median or 25th percentile of Pearson's r values from the plate comparisons.