Supporting Information

Towards performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling

Mathias J. Wawer[a], Kejie Li[a], Sigrun M. Gustafsdottir[a], Vebjorn Ljosa[b], Nicole E. Bodycombe[a], Melissa A. Marton[c], Katherine L. Sokolnicki[b], Mark-Anthony Bray[b], Melissa M. Kemp[a], Ellen Winchester[c], Bradley Taylor[c], George B. Grant[c], C. Suk-Yee Hon[a], Jeremy R. Duvall[d], J. Anthony Wilson[a], Joshua A. Bittker[d], Vlado Dančík[a,e], Rajiv Narayan[f], Aravind Subramanian[f], Wendy Winckler[c], Todd R. Golub[f], Anne E. Carpenter[b], Alykhan F. Shamji[a], Stuart L. Schreiber[a]*, Paul A. Clemons[a]*

[a]Center for the Science of Therapeutics, [b]Imaging Platform, Broad Institute, [c]Genomics Platform, Broad Institute, [d]Center for the Development of Therapeutics, and [f]Cancer Program, Broad Institute, 7 Cambridge Center, Cambridge, MA 02142, USA; and [e]Mathematical Institute of the Slovak Academy of Sciences, 04001 Košice, Slovak Republic

***Corresponding authors**

Stuart L. Schreiber, stuart_schreiber@harvard.edu
Paul A. Clemons , pclemons@broadinstitute.org

Supporting Materials and Methods

Compound selection

The DOS compound collection represents a structurally diverse subset of 19,637 compounds selected from 23 DOS libraries. (1-8) These libraries were synthesized using a build-couple-pair strategy (9) to combine simple chiral building blocks into diverse and complex compounds. (10) Each library is built around a common chiral core by varying side chains and the configurations of core stereocenters. For most compounds, all stereoisomers were synthesized.

For each library, we determined the set of unique stereochemical "parent" structures, *i.e.*, structures with unspecified stereochemistry. We exposed these sets to a maximum dissimilarity selection algorithm using Tanimoto similarity (11) on ECFP4 fingerprints (12) for each chiral core separately. In collaboration with the Broad Institute Discovery Chemistry and Compound Management teams, we determined the desired proportion of compounds from each library, and for those stereochemical parents selected, included all stereoisomers with physical samples available.

The BIO collection comprised three different compound sets. First, we included 2,222 drugs, natural products, and small-molecule probes that are part of the Broad Institute known bioactive compound collection. The collection contains structurally diverse compounds across a wide range of biological activities with known targets for many compounds. Second, we extended this set by selecting 274 hits or structural analogs from various probe-development projects sponsored by the Molecular Libraries Program (MLP). Third, we selected 10,162 compounds from the Molecular Libraries Small Molecule Repository (MLSMR). Assay activity data from Molecular Libraries Probe Production Centers (MLPCN) screening centers reported as percent activity were retrieved from PubChem (http://pubchem.ncbi.nlm.nih.gov/). To include bioactive but not promiscuous compounds, we kept only compounds that had been tested in >50% of all assays (unique AIDs) performed on the MLSMR collection and that were hits in at least 2 but fewer than 10% of assays tested. Compounds were selected to cover many different chemical structures and biological activities.

Differences between the total compound numbers reported here and those in Table S1 are due to quality control filters on GE and MC experimental data.

Gene-expression assay

We followed the protocol published by Peck et al. (13) Briefly, we seeded 3,500 U-2 OS cells (ATCC, cat. no. HTB-96) per well in 384-well plates. After 24h of incubation at 37°C, compounds were added to the cells, followed by another 6h of incubation. Treatments were carried out in triplicates. The cells were then lysed and lysates transferred to oligo-dT plates to capture mRNAs. We reverse-transcribed mRNA and amplified cDNA via polymerase chain reaction (PCR). We annealed the cDNA to a mix of upstream/downstream probe pairs, each of which was designed to be specific for one of 977 transcripts. Each upstream probe consisted of a universal 20-nucleotide (nt) primer site

(complementary to T7 primer), a unique 24-nt barcode sequence, and a 24-nt sequence designed to bind to the 3'-end of one specific transcript. Downstream probes were designed to anneal contiguous to their corresponding upstream probe on the transcript. They consist of a 5'-phosphorylated 20-nt transcript-specific sequence and a 20-nt universal primer site (T3). Any unbound probes were removed after the annealing step. Only upstream and downstream probes that bound next to each other on a transcript cDNA molecule were ligated in the next step and then amplified by PCR using T3 and 5'-biotinylated T7 primers. We added these amplicons to a mix of color-coded Luminex microspheres, each of which carried capture probes complementary to one of the barcode sequences in the amplicons. Streptavidin-phycoerythrin was added to add fluorescent markers to the biotinylated amplicons. The number of captured amplicons was then quantified by flow cytometry measuring phycoerythrin fluorescence. Transcript identity was identified by microsphere color.

Gene-expression: cell plating and compound treatment

1. U-2 OS cells (ATCC, cat. no. HTB-96)
2. 384 well plates (Corning, cat. no. 3712)
3. culture medium
    - DMEM (Fisher Scientific, cat.no. MT10017CV)
    - 10%FBS (Life technologies, cat.no. 10437028)
    - 1% penicillin-streptomycin (Fisher Scientific, cat. no. MT30002CI)
4. TCL buffer (Quiagen, cat. no. 1031576)
5. cold-storage adhesive sealing foil (VWR, cat. no. 89049-034)

3500 U-2 OS cells per well were plated in 384-well plates with 50 μL culture medium. After 24 h of incubation at 37°C, compounds were added. Cells were treated for 6 h at 37°C before 40 μL of medium was removed and 30 μL TCL buffer added to lyse the cells. Plates were sealed with sterile sealing foil and, after 30 min incubation at room temperature (RT), stored at -80°C.

Gene-expression: sample preparation and measurement

We closely followed the protocol published by Peck et al. (13) Differences from the published materials and methods are summarized below. A detailed protocol follows.

1. Luminex microspheres: we used MagPlex instead of COOH
2. bead coupling preparation: we used a different volume of beads and wash solutions
3. probe hybridization: we introduced a ramp-down of annealing temperatures
4. polymerase: we used HotStarTaq Plus instead of HotStarTaq
5. PCR cycling conditions were different (29 cycles, 1min step times)
6. to measure all transcripts, we used 2 different bead mixes per sample during detection; furthermore, a different volume of each mix was used
7. bead/amplicon hybridization time: increased from 1 h to 16 h – 20 h

8. we introduced bead washes before and after streptavidin-phycoerythrin addition during detection, including new wash solutions
9. Luminex detection instrument: we used FlexMap 3D instead of Luminex 100

General Notes

All liquid transfers were automated on the Agilent Bravo Liquid Handling Platform. Between each step of the LMA protocol, unreacted products were removed by inverting the reaction plate onto a laboratory towel and centrifuging at 1000 g for 1 min. All non-room-temperature incubations took place on a Thermo Electron MBS 384 Satellite Thermal Cycler.

Luminex microsphere (bead) preparation

Materials:

1. Luminex xMAP MagPlex microspheres
2. bead binding buffer
   - 0.1 M 2-(N-morpholino)ethansulfonic acid (pH 4.5)
3. EDC solution
   - 10 mg/mL 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride in water.
4. 0.02% Tween
5. 0.1% sodium dodecyl sulfate
6. TE (pH8)
   - 10 mM Tris-HCl (pH 8.0)
   - 1 mM EDTA

Luminex xMAP MagPlex microspheres were coupled to anti-barcode capture oligonucleotides. 490 distinct microsphere species were aliquoted into 800µL round-bottom deep well 96 plates, approximately 12.5 x 10^6 mirospheres per well. Beads were pulled down by centrifugation and magnetic separation, storage buffer was removed, and beads were re-suspended in 130 µL binding buffer. 500 pmol of capture probe was added to each microsphere well, such that each microsphere species well received a different probe. 5 µL of a freshly prepared EDC solution was added, and the reactions were incubated for 30 min. EDC addition was repeated a second time. Microspheres were then captured by magnetic pull-down and washed successively in 500 µL of Coupling 0.02% Tween, 0.1% sodium dodecyl sulfate, and TE (pH 8). Coupled microspheres were re-suspended in TE (pH 8) and pooled to a final concentration of 50,000 microspheres/µL. This pool constituted the first coupled microsphere set, referred to as dp52.

This process was repeated, such that the same 490 microsphere species were coupled to a different set of capture oligonucleotides, forming the second coupled microsphere set dp53.

Finally, 10 microsphere species were coupled to 80 different capture oligonucleotides, 8 oligonucleotides per microsphere species, to form a coupled microsphere pool assaying 10 invariant meta-genes.

Ligation Mediated Amplification

RNA extraction

Materials:

1. Turbocapture 384 mRNA kit (Qiagen, cat. no. 72271)

Frozen lysate was thawed for 1 h at room temperature. 20 µL of lysate was transferred to a 384-well oligo-dT capture plate, and incubated at room temperature for 1h. During this incubation, mRNA was immobilized to the plate via binding of the poly-A tail.

cDNA Generation

Materials:

1. M-MLV Reverse Transcriptase kit (Promega, cat. no. M1705)
2. 100mM dNTP set (Invitrogen, cat. no. 10297018)

cDNA was generated from immobilized mRNA via reverse transcription. A 5 µL M-MLV reaction mix was added to the reaction plate and incubated at 37°C for 1.5 h.

Probe Hybridization

1. pool of custom up- and down-stream probes
2. Taq DNA Ligase Reaction Buffer (NEB, cat. no. B0208S), used to prepare probe-pool working dilution

Custom up- and down-stream probes were hybridized to the cDNA. Upstream probes contained 20 nt of gene-specific sequence, a 24 nt FlexMap barcode, and the T7 priming site. Down-stream probes contained 20 nt of gene-specific sequence (designed to bind adjacent to the upstream probe) and the T3 priming site. Hybridization consisted of a 5 min 95°C denature followed by an annealing ramp-down from 70°C to 40°C, 12 min per degree. The reaction was held at 4°C overnight.

Probe Ligation

Materials:

Taq DNA ligase kit (NEB, cat. no. M0208L)

Bound probes were ligated via a 5 µL Taq DNA ligase reaction, incubated at 45°C for 1 h, followed by 65°C for 10 min.

PCR Amplification

Materials:

1. HotStarTaq Plus DNA polymerase kit (Qiagen, cat. no. 203603)
2. 100mM dNTP set (Invitrogen, cat. no. 10297018)

Ligated probes form a template competent for PCR using T3 and T7 universal primers. The T7 primer was biotinylated. A 15 µL HotStarTaq Plus reaction mix was added to the plate, and the following PCR program performed:

a. initial denature: 15 min @ 92°C
b. 29 amplification cycles: 1 min @ 92°C – 1 min @ 60°C – 1 min @ 72°C
c. final extension: 5 min @ 72°C

Hybridization and Detection

Microsphere Hybridization

Materials:

1. dp52 coupled bead set
2. dp53 coupled bead set
3. invariant-gene bead set
4. 1.5x TMAC hybridization solution
    - 4.5 M tetramethylammonium-chloride
    - 0.15% N-lauryl sarcosine
    - 75 mM Tris-HCl (pH 8)
    - 6 mM EDTA

Dilutions of each coupled flexmap microsphere set dp52 and dp53 were prepared in 1.5x TMAC hybridization solution, such that each reaction ultimately contained about 200 microspheres of each species. Additionally, "invariant-gene" microspheres were added to each dilution at a similar concentration. Invariant genes were genes selected to show relatively stable expression levels (coefficient of variation < 10%) across a large number of distinct reference samples. Eight invariant genes were selected at each of 10 expression levels. For each expression level, all 8 invariant genes were combined on one bead species. 23 µL of diluted microspheres are plated in 384. 5 µL of amplified LMA product was added to a dp52 dilution well and another 5 µL to a dp53 dilution well. Samples were arranged such that the dp52 and dp53 wells were found on the same plate. The

resulting detection plate therefore contained 192 samples assayed on both bead sets. The detection plate was denatured at 95°C for 2 min, and incubated at 45°C for 16 h – 20 h overnight.

Microsphere Detection

Materials:

1. 1x TMAC hybridization solution
   - 3 M tetramethylammonium-chloride
   - 0.1% N-lauryl sarcosine
   - 50 mM Tris-HCl (pH 8)
   - 4 mM EDTA
2. low-stringency wash buffer
   - 6x SSPE
   - 0.01% Tween-20
3. high-stringency wash buffer
   - 0.1x MES
   - 25 mM NaCl
   - 0.01% Tween-20
4. reporter mix
   - 3% streptavidin-phycoerythrin (Invitrogen, cat. no. S866) in 1x TMAC hybridization solution

Microspheres were captured in the detection plate by centrifugation at 1000 rpm for 1 min followed by magnetic pull-down. Microspheres were washed successively in low-stringency wash buffer and high-stringency wash buffer. 10 µL of reporter mix were added to each sample. Samples were incubated at 45°C for 10 min to allow streptavidin-phycoerythrin to bind the biotinylated amplicons. Samples were then centrifuged, magnetically pulled down, and washed with low-stringency wash buffer and 3 times with 1X TMAC wash solution. Labeled and washed microspheres were analyzed using the Luminex FlexMap 3D detector.

Gene-expression assay: data processing

Raw fluorescence intensity curves were processed by a peak-detection algorithm to yield expression values for each transcript in a sample. For each sample, the binary logarithms of expression values were normalized based on 80 pre-determined "invariant genes", *i.e.*, genes that show relatively stable expression levels (coefficient of variation < 10%) across a large number of distinct reference samples. Eight invariant genes were selected at each of 10 expression levels. A calibration curve was computed for each sample using the median expression of these invariant genes. Samples were then rescaled using a reference curve computed from a large collection of expression profiles and limited to the range [0, 15] (http://lincscloud.org/how-data-were-

prepared/). Detailed descriptions of the data-collection and data-processing pipeline will be published separately by the NIH LINCS project, and are summarized online (http://lincscloud.org/).

Gene-expression array: data normalization and correction

For each plate, the distributions of per-well gene-expression levels were quantile normalized. (14) Plate medians for all transcripts were subtracted from each well profile. Positional effects for each gene were corrected using GeneData Screener Assay Analyzer. (15) Robust Z-scores were calculated by dividing the resulting values by 1.4826 * plate median absolute deviation (MAD). We then calculated Stouffer's Z-score (16) to combine replicates into the final profiles for each compound.

Multiplexed cytological imaging assay

We followed the protocol published by Gustafsdottir et al. (17) Briefly, we seeded 1,500-2,000 U-2 OS cells (ATCC, cat. no. HTB-96) per well in 384-well clear-bottom imaging plates. After 24h of incubation at 37°C, compounds were added to the cells, followed by another 48h of incubation. Treatments were carried out in quadruplicates. We then stained six different cell compartments and organelles with fluorescent dyes: nucleus (Hoechst 33342), endoplasmic reticulum (concanavalin A/AlexaFluor488 conjugate), nucleoli (SYTO 14 green fluorescent nucleic acid stain), Golgi apparatus, and plasma membrane (wheat germ agglutinin/AlexaFluor594 conjugate, WGA), F-actin (phalloidin/AlexaFluor594 conjugate) and mitochondria (MitoTracker Deep Red). WGA and Mitotracker were added to living cells. The remaining stains were carried out after cell fixation with 16% paraformaldehyde. Images were captured in 5 fluorescent channels from 9 sites per well (20× magnification). We used the CellProfiler image-analysis software to calculate morphological features for each cell. (17)

Multiplex cytological imaging assay: materials

1. U-2 OS cells (ATCC, cat. no. HTB-96)
2. Aurora 384-well black/clear bottom plates, imaging quality (Brooks, cat. no. 1022-11330)
3. culture medium
    - DMEM (Fisher Scientific, cat.no. MT10017CV)
    - 10%FBS (Life Technologies, cat.no. 10437028)
    - 1% penicillin-streptomycin (Fisher Scientific, cat. no. MT30002CI)
4. Mitotracker Deep Red (Invitrogen, cat. no. M22426)
5. wheat germ agglutinin Alexa 594 conjugate (Invitrogen, cat.no. W11262)
6. paraformaldehyde 16%, methanol free (Electron Microscopy Sciences, cat. no. 15710-S)
7. Hank's Balanced Salt Solution, HBSS (Invitrogen, cat. no. 14065-056)
8. Triton X-100 (Sigma, cat. no. T8787)
9. phalloidin 594 (Invitrogen, cat. no. A12381)
10. concavalin A 488 (Invitrogen, cat. no. C11252)

11. Hoechst 33342 (Invitrogen, cat. no. H3570)
12. SYTO 14 green fluorescent nucleic acid stain (Invitrogen, cat.no. S7576)
13. sodium biocarbonate (HyClone, cat. no. SH30033.01)
14. methanol (BDH, cat. no. 67-56-1)
15. bovine serum albumin
16. REMP blue thermo heat seal (REMP/Nexus Biosystems, cat. no. 1800336)
17. ImageXpress Micro (Molecular Devices)

Multiplex cytological imaging assay: assay protocol

U-2 OS cells were plated at a density of 1500-2000 cells per well with 50 μL culture medium. After 24 h incubation at 37°C, compounds were added. Cells were treated for 48 h at 37°C. A 1 mM solution of Mitotracker in DMSO and a 1 mg/mL solution of wheat germ agglutinin (WGA) in distilled water were used to prepare a staining solution of 500nM Mitotracker and 60 μg/mL WGA in pre-warmed medium. After removal of 40 μL of media from the cells, 30 μL of the staining solution were added to each well and incubated for 30 min at 37°C. Cells were fixed for 20 min at RT with 10 μL paraformaldehyde and afterwards washed once with 70 μL HBSS. To permeabilize cells, 30 μL of a 0.1% solution of Triton X-100 in 1x HBSS were added, incubated for 10-20 min, and washed two times with 70 μL 1xHBSS. Concanavalin A was dissolved to 1 mg/mL in 0.1 M sodium bicarbonate solution. Phalloidin was dissolved in 1.5 mL methanol per vial. Staining mix was prepared from 0.025 μL phalloidin/μL, 100 μg/mL Concanavalin, 5 μg/mL Hoechst, and 3μM SYTO staining solution in 1x HBSS 1% BSA. Aliquots of 30 μL staining mix were added to each well and incubated for 30 min. After staining, cells were washed three times with 70 μL 1xHBSS without final aspiration. Plates were thermally sealed at 171°C (4 seconds).

Multiplex cytological imaging assay: image capture

We captured images on an ImageXpress Micro epifluorescent microscope. We recorded 9 sites per well at 20x magnification in 5 fluorescent channels, DAPI (387/447 nm), GFP (472/520 nm), Cy3 (531/593 nm), TexasRed (562/642 nm), Cy5 (628/692 nm). The first site of each well was used for laser-based auto-focus in the DAPI channel.

Multiplex cytological imaging assay: image analysis and data processing

CellProfiler (18) software version 2.0.9925 was used to locate and segment cells and measure morphological features for each cell. We used pipelines described and provided by Gustafsdottir *et al.* (17) to correct for uneven illumination and segment cells into nuclei and cytoplasm. Size, shape, texture, intensity statistics, and local density were measured for nuclei, cytoplasm, and entire cells. (17) Cell-morphology features were normalized by linearly scaling the 1st and 99th percentiles of

the DMSO-control distributions to 0 and 1, respectively. Plate medians were subtracted from each profile and positional effects corrected with GeneData Screener Assay Analyzer. (15)

HTS assay information

HTS assay results were assembled from an internal database at the Broad Institute. However, for the majority of assays, results have been deposited in public databases (ChemBank and PubChem/BARD; Datasets S1 and S2). We distinguished between screening projects, assays, and individual assay measurements with screening projects representing the highest level of organization in the respective database. For ChemBank, assays were defined as all experiments in a screening project that share the same detection method. Assay measurements were defined as all experiments in an assay that share the same experimental conditions and time point.

For PubChem/BARD and internal screening projects, assays were defined as annotated by the experimenter who submitted the screen. Assay measurements were defined as all direct measurements and calculated values that convey information different from the direct measurements (e.g., a ratio of two direct measurements).

Compound activity for profiling assays and activity score

We used the multidimensional perturbation value (mp-value) as described by Hutz et al. (19) to determine compound activity in profiling screens. The profiles for all replicates of a compound within a batch were combined into a matrix with the profiles from the corresponding negative DMSO-control wells in the same batch such that rows represent wells and columns represent profiling features. The matrix, generated for each compound separately, was then standardized by first calculating a z-score across rows and then columns. Principal component analysis was performed on the standardized matrices and the first n principal components that sum up to a variance of 0.9 were retained. Each of these n principal components was weighted by the percentage of variance it explains (by multiplying the matrix with the vector of variances) to obtain the normalized matrix P.

P was split into treatment and control rows and for each of the parts a covariance matrix was calculated. Each of two covariance matrices (treatment and control) was weighted by the number of samples in each group. The sum of the resulting matrices was used to calculate the Mahalanobis distance between treatment and control samples.

To calculate a p-value based on the Mahalanobis distance, we performed an empirical test with 1000 permutations. Each time the treatment/control labels were randomly assigned and the distance recalculated to estimate a distribution of distances. A p-value was then calculated as the fraction of distance values that are equal to or larger than the real distance value. Compounds with a p-value lower than 0.05 were considered active.

We calculated a normalized "activity score" to use the calculated Mahalanobis distance as an additional constraint for activity (as suggested by Hutz et al. (19)). We scaled the distribution of distances for all compounds linearly such that the [0.2, 99.8]-percentile range mapped to [0, 1].

Promiscuity probability

The probability of a compound showing promiscuous HTS assay activity (or 'cross-reactivity') was calculated according to Dančík *et al.* (20) Based on past screening results, we calculated the mean (0.13) and standard deviation (0.012) of hit frequencies for all compounds. These values were used to determine parameters α and β of a beta-distribution:

$$\mu = \frac{\alpha}{\alpha + \beta},$$

$$\sigma^2 = \alpha\beta(\alpha + \beta)^{-2}(\alpha + \beta + 1)^{-1}$$

For each compound, we determined the number of assays N in which it was tested, and the number of assays n in which it scored as a hit and calculated the probability of having a hit frequency θ higher than $\theta_0 = 0.25$ using the MATLAB function `betainc`,

$$P(\theta > \theta_0) = betainc(\theta_0, n + \alpha, N - n + \beta, 'upper').$$

Supporting Figures

Figure S1. Annotated bioactive compounds clustered based on MC profiles form groups with similar biological effects.

We hierarchically clustered all compounds of the MC hit set for which a common name was available based on their imaging profiles. We used complete linkage applied to correlation distance (1-Pearson coefficient). Compounds that do not have a neighbor closer than 0.2 correlation distance are omitted for clarity. Compound names are reported next to the dendrogram. Where known, the compound's primary biological effect or use is reported. If groups of compounds with related biological effects co-cluster, their common effect is summarized (indicated by a larger font size).

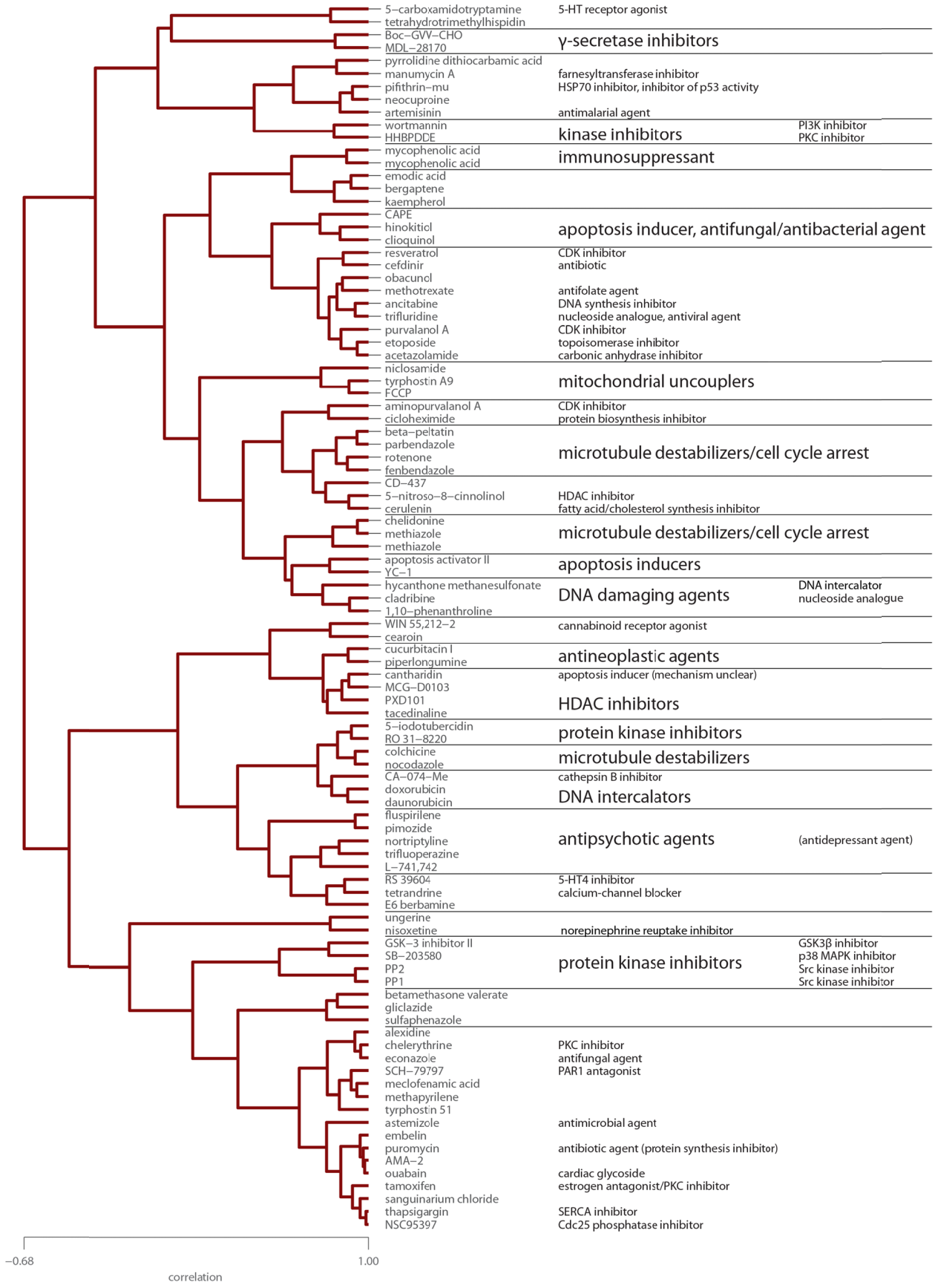
5–carboxamidotryptamine
tetrahydrotrimethylhispidin
5-HT receptor agonist
Boc–GVV–CHO
MDL–28170
γ-secretase inhibitors
pyrrolidine dithiocarbamic acid
manumycin A
farnesyltransferase inhibitor
pifithrin–mu
HSP70 inhibitor, inhibitor of p53 activity
neocuproine
artemisinin
antimalarial agent
wortmannin
HHBPDDE
kinase inhibitors
PI3K inhibitor
PKC inhibitor
mycophenolic acid
mycophenolic acid
immunosuppressant
emodic acid
bergaptene
kaempherol
CAPE
hinokitiol
clioquinol
apoptosis inducer, antifungal/antibacterial agent
resveratrol
cefdinir
CDK inhibitor
antibiotic
obacunol
methotrexate
ancitabine
trifluridine
purvalanol A
etoposide
acetazolamide
antifolate agent
DNA synthesis inhibitor
nucleoside analogue, antiviral agent
CDK inhibitor
topoisomerase inhibitor
carbonic anhydrase inhibitor
niclosamide
tyrphostin A9
FCCP
mitochondrial uncouplers
aminopurvalanol A
cycloheximide
CDK inhibitor
protein biosynthesis inhibitor
beta–peltatin
parbendazole
rotenone
fenbendazole
microtubule destabilizers/cell cycle arrest
CD–437
5–nitroso–8–cinnolinol
cerulenin
HDAC inhibitor
fatty acid/cholesterol synthesis inhibitor
chelidonine
methiazole
methiazole
microtubule destabilizers/cell cycle arrest
apoptosis activator II
YC–1
apoptosis inducers
hycanthone methanesulfonate
cladribine
1,10–phenanthroline
DNA damaging agents
DNA intercalator
nucleoside analogue
WIN 55,212–2
cearoin
cannabinoid receptor agonist
cucurbitacin I
piperlongumine
antineoplastic agents
cantharidin
apoptosis inducer (mechanism unclear)
MCG–D0103
PXD101
tacedinaline
HDAC inhibitors
5–iodotubercidin
RO 31–8220
protein kinase inhibitors
colchicine
nocodazole
microtubule destabilizers
CA–074–Me
cathepsin B inhibitor
doxorubicin
daunorubicin
DNA intercalators
fluspirilene
pimozide
nortriptyline
trifluoperazine
L–741,742
antipsychotic agents
(antidepressant agent)
RS 39604
tetrandrine
E6 berbamine
5-HT4 inhibitor
calcium-channel blocker
ungerine
nisoxetine
norepinephrine reuptake inhibitor
GSK–3 inhibitor II
SB–203580
PP2
PP1
protein kinase inhibitors
GSK3β inhibitor
p38 MAPK inhibitor
Src kinase inhibitor
Src kinase inhibitor
betamethasone valerate
gliclazide
sulfaphenazole
alexidine
chelerythrine
econazole
SCH–79797
meclofenamic acid
methapyrilene
tyrphostin 51
PKC inhibitor
antifungal agent
PAR1 antagonist
astemizole
embelin
puromycin
AMA–2
antimicrobial agent
antibiotic agent (protein synthesis inhibitor)
ouabain
tamoxifen
sanguinarium chloride
cardiac glycoside
estrogen antagonist/PKC inhibitor
thapsigargin
NSC95397
SERCA inhibitor
Cdc25 phosphatase inhibitor
–0.68
1.00
correlation

Figure S2. MC and GE profile diversity leads to HTS outcome diversity.

We compared the HTS outcome diversity for subsets of compounds selected from (a) the MC test collection and (b) the GE test collection. The subsets were selected to have diverse MC profiles, GE profiles, or chemical structures (CS), as indicated by the label for each curve. We compared the HTS performance diversity of sets selected based on MC or GE profile diversity or CS diversity to randomly selected compound sets of the same size (RND).

Diverse subset selection from the hit-sets was performed iteratively using a maximum dissimilarity strategy. The first compound in this process was selected randomly. We then iteratively added the compound to the growing list that had the most dissimilar MC profile (GE profile; CS) to the already selected ones. We monitored the change in HTS performance diversity throughout this selection process. HTS performance diversity was measured as the "true diversity" (see Methods). This measure penalizes redundancy and over-representation of individual profiles in a compound set. Therefore, adding a novel HTS profile to the set will increase diversity whereas adding a profile already contained in the set will decrease it. The maximum diversity (100%) is reached when an equal number of compounds represent each distinct assay profile in the respective hit set. (21) Effectively, this procedure provides a ranking of compounds by their contribution to the diversity of the entire set. Monitoring the HTS performance diversity provides a measure for the diversity of the top N diverse compounds, where N ranges from 1 to the size of the respective test collection. Because the first compound was selected randomly, the entire selection was repeated 500 times, each time with a different starting point (resulting in 500 distinct ranked lists). The average HTS performance diversity over these 500 rankings (+/- standard deviation, sd) is plotted for the top N compounds. The vertical line indicates the set size that achieved the maximum diversity across all of the selection methods. This set size was used for to generate the bar charts in Fig. 4b and 4c in the manuscript as well as Fig. S3.

(a) The diversity curves show that selection using MC profile diversity led to higher HTS outcome diversity than random selection and selection based on chemical structure diversity. Each selection method (MC, CS, RND) picked distinct profiles at first, leading to a steep increase in diversity near the top of the list. However, after many selections, random selection will choose profiles that are already in the set, indicated by a flattening of the diversity curve. Since there is no differential selection possible for the full set, all curves converge when the diversity of the full compound set is reached (23.9%; right end of the plot). However, selection based on the MC diversity rankings leads to a prolonged increase in HTS outcome diversity and hence reaches a higher level than random compound selection. This result indicates that MC profile diversity can inform the selection of a performance-diverse compound set for cell-based screens. Selection on chemical structure diversity performed worse than MC and even led to lower performance diversity than random selection. (b) Similar results are observed for the GE test collection, where diversity selection based on GE profiles outperforms RND and CS, leading to higher HTS performance diversity. Likely due to the lower numbers of compounds in the GE study, GE-profile-based selection led to a lower increase over random than MC.

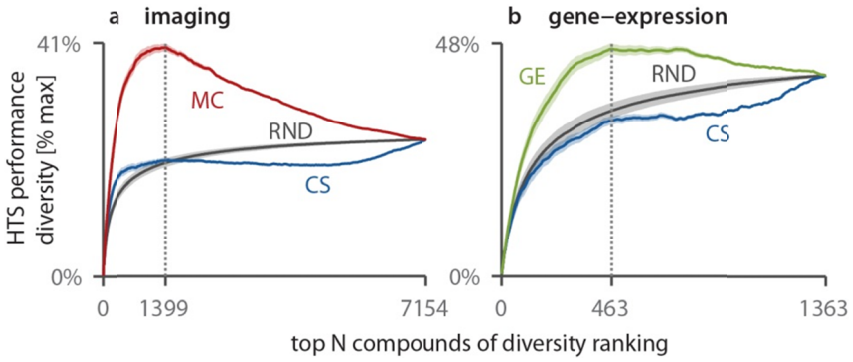
a imaging
b gene–expression
41%
0%
48%
0%
HTS performance
diversity [% max]
MC
RND
CS
GE
RND
CS
0
1399
7154
0
463
1363
top N compounds of diversity ranking

Figure S3. Assay-profile cluster coverage of different diversity selection methods.

We compared the number of distinct HTS assay profile clusters (groups of compounds with similar performance across HTS assays) that were covered by compound sets selected to have diverse MC profiles, GE profiles, or chemical structure (CS). We compared these to a baseline of randomly selected sets of the same size (RND). We clustered compounds based on similar HTS performance patterns using hierarchical clustering (see Methods for details). Diverse compound subsets were selected from (a) the MC test collection (1399 compounds; marked with a dashed line in Fig, S2a) and (b) the GE test collection (463 compounds; marked with a dashed line in Fig, S2b). The diversity subsets used here were identical to the ones used in Fig. 4 of the manuscript. Analogous to the results shown in Fig. 4b, 4c, and 5c of the manuscript, sets with diverse MC or GE profiles cover more HTS clusters than sets with diverse chemical structure or randomly selected sets. Likewise, selection of diverse GE profiles led to comparably smaller improvements of HTS cluster coverage over CS and RND (b) than selection of diverse MC profiles (a), likely due the lower number of compounds available for the GE study.

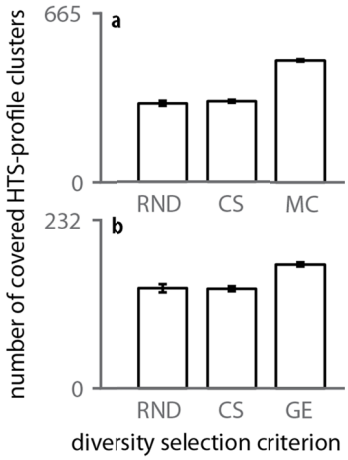

Figure S4. Annotated bioactive compounds clustered based on GE profiles form groups with similar biological effects.

We hierarchically clustered all compounds of the GE hit set for which a common name was available based on their gene-expression profiles. We used complete linkage applied to correlation distance (1-Pearson coefficient). Compounds that do not have a neighbor closer than 0.35 correlation distance are omitted for clarity. Compound names are reported next to the dendrogram. Where known, the compound's primary biological effect or use is reported. If groups of compounds with related biological effects co-cluster their common effect is summarized (indicated by a larger font size). Compounds that are present multiple times are positive controls included in multiple instances throughout the experiment.


valproic acid
15D–prostaglandin J2
tretinoin
tretinoin
tretinoin
tretinoin
tretinoin
tretinoin
tretinoin
tretinoin
tretinoin
RAR agonist
triazolothiadiazine, 5
beta–peltatin
microtubule destabilizer
geldanamycin
geldanamycin
HSP90 inhibitors
dihydrogambogic acid
cucurbitacin I
ZL3VS
protein synthesis inhibitor
puromycin
proteasome inhibitor
FCCP
tyrphostin A9
niclosamide
oxidative phosphorylation uncoupling agents
trifluoperazine
fluphenazine
fluphenazine
thioridazine
fluphenazine
thioridazine
thioridazine
trifluoperazine
fluphenazine
fluphenazine
SCH–79797
PAR1 inhibitor
fluphenazine
trifluoperazine
thioridazine
trifluoperazine
trifluoperazine
trifluoperazine
trifluoperazine
trifluoperazine
fluphenazine
fluphenazine
antipsychotic agents
9-aminoacridine
tilorone
interferon inducer
ouabain
proscillaridin
bufalin
convallatoxin
ouabain
sarmentogenin
digoxin
digoxigenin
cardiac glycosides
H7
BML–259
Cdk2/Cdk5 inhibitor
genistein
genistein
genistein
genistein
genistein
tyrosine kinase inhibitor, PPAR activator
geldanamycin
geldanamycin
tanespimycin
tanespimycin
geldanamycin
geldanamycin
geldanamycin
tanespimycin
tanespimycin
tanespimycin
tanespimycin
tanespimycin
HSP90 inhibitors
3–FLUOROBENZYLSPIPERONE
spiperone
antipsychotic agents
15D–prostaglandin J2
15D–prostaglandin J2
PPAR γ agonist
estradiol
estradiol
estrogen
troglitazone
troglitazone
PPAR γ agonist
LY–294002
LY–294002
LY–294002
LY–294002
LY–294002
LY–294002
LY–294002
LY–294002
PI3K inhibitor
sirolimus
sirolimus
sirolimus
sirolimus
sirolimus
sirolimus
sirolimus
sirolimus
sirolimus
immunosuppressant
mitoxantrone
mitoxantrone
mitoxantrone
mitoxantrone
mitoxantrone
mitoxantrone
mitoxantrone
mitoxantrone
mitoxantrone
topoisomerase II inhibitor
RO 31–8220
5–iodotubercidin
aminopurvalanol A
kinase inhibitors
PKC inhibitor
Adenosine kinase inhibitor
CDK inhibitor
triptolide
RNA synthesis inhibitor
daunorubicin
DNA intercalator
MCG–D0103
tacedinaline
PXD101
trichostatin A
trichostatin A
trichostatin A
trichostatin A
vorinostat
vorinostat
vorinostat
vorinostat
trichostatin A
vorinostat
trichostatin A
vorinostat
vorinostat
vorinostat
vorinostat
trichostatin A
HDAC inhibitors
−0.52
1.00
correlation

Figure S5. Compounds active for both MC and GE are more promiscuous than hits in either assay individually.

Cumulative distributions of promiscuity probability values (see "Promiscuity probability" in Supporting Methods for details) are shown for compounds active in the MC assay, the GE assay, or both assays. The right-shift of the red curve (hits in both MC and GE assays) compared to the black (hits in GE assay) and blue curves (hits in MC assay) indicate that the set of compounds that were hits in both assays was enriched for promiscuous compounds, *i.e.*, compounds that are active in many HTS assays, likely due to unspecific activity (*e.g.*, cytotoxicity).

The median promiscuity probability (pp) is significantly higher for the intersection of both assays [median(pp) = 0.33%] than for MC [median(pp) = 0.15%; $p = 7.6\times10^{-14}$] or GE alone [median(pp) = 0.15%; $p = 5.7\times10^{-7}$].

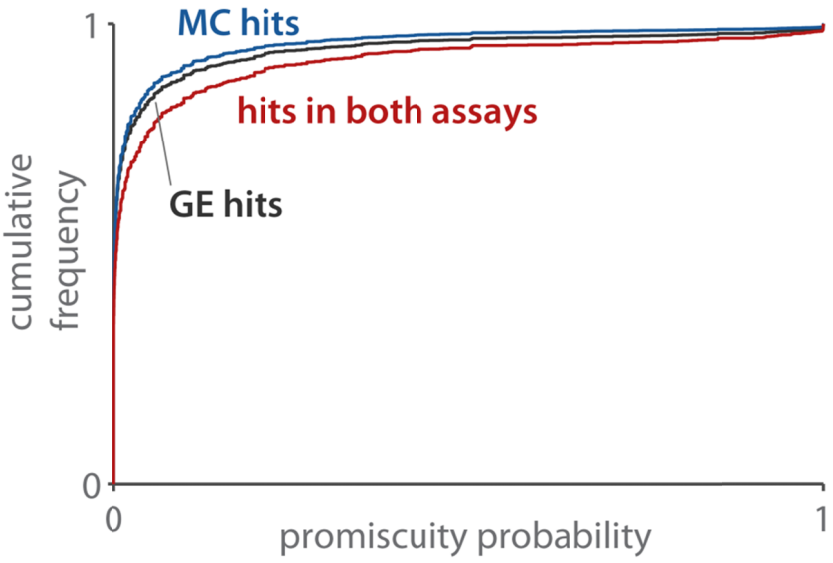

Figure S6. MC leads to slightly higher HTS performance diversity when compared to GE on the overlap between the MC and GE test collections. We applied maximum dissimilarity diversity selection to the overlap between the MC and GE test collections to directly compare the level of HTS performance diversity achieved by diversity selection based on MC and GE profiles on the same compound set (see Fig. S2 for a detailed description of maximum dissimilarity diversity selection).

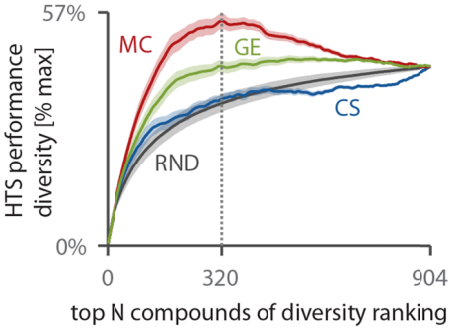

Figure S7. Assay-profile cluster coverage of different diversity-selection methods – calculated on the overlap of MC and GE test collections. The bar chart shows the number of HTS assay profile clusters covered by different selection methods (RND, random; CS, chemical structure; MC, imaging profiles; GE, gene-expression profiles) for a set size of 320 compounds, which achieved the overall highest performance diversity for the intersection of test collections (indicated by the dashed line in Fig. S6).

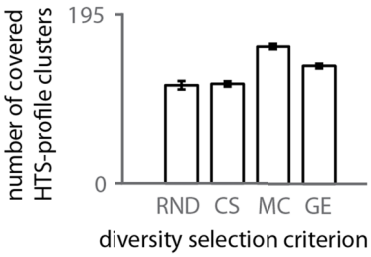

Table S1. Composition of profiling compound collection and hit sets.

	BIO			DOS			POS			all		
	N	**hits**	**hit rate**	**n**	**hits**	**hit rate**	**n**	**hits**	**hit rate**	**n**	**hits**	**hit rate**
MC	12431	8490	68.3	17805	6594	37.0	-	-	-	30236	15084	49.9
GE	4199	1639	39.0	17553	1924	11.0	29	28	96.6	21781	3591	16.5
union	12606	9009	71.5	19164	7606	39.7	-	-	-	31770	16615	52.3

Table S2. Detection method used in assay measurements.

detection method	frequency	frequency [%]
fluorescence	290	56.64%
luminescence	150	29.30%
cytoblot	28	5.47%
absorbtion	4	0.78%
qPCR	2	0.39%
alphaLISA	1	0.20%
other	37	7.23%

Table S3. Assay kits used in assay measurements.

assay kit	frequency	frequency [%]
CellTiter-Glo	100	19.53%
Resazurin	85	16.60%
Calcein	44	8.59%
JC-1	33	6.45%
Reporter	29	5.66%
Bromodeoxyuridine	20	3.91%
NileRed	9	1.76%
AmplexRed	7	1.37%
DAPI	6	1.17%
Caspase-Glo	6	1.17%
OilRedO	5	0.98%
MTT	4	0.78%
other	164	32.03%

Table S4. Cells used in assay measurements.

cell type	frequency	frequency [%]
primary	142	27.73%
A549	67	13.09%
HEK293	15	2.93%
NIH/3T3	14	2.73%
U2OS	14	2.73%
HeLa	14	2.73%
H1299	13	2.54%
PC9	13	2.54%
LN229	13	2.54%
hES	12	2.34%
MEF	7	1.37%
Alpha-TC-1;Beta-TC-3	6	1.17%
mES	5	0.98%
C2C12	5	0.98%
HTB-65	4	0.78%
786-O	4	0.78%
HUVEC	4	0.78%
A498	4	0.78%
MCH58	4	0.78%
HepG2	4	0.78%
worm	3	0.59%
DLD1	3	0.59%
MM1S	3	0.59%
RKO	3	0.59%
HT22	3	0.59%
RPMI8826	3	0.59%
Huh7	3	0.59%
SK-MEL-5	3	0.59%
HMLE	3	0.59%
MEF-1	2	0.39%
INS-1E	2	0.39%
J774A.1	2	0.39%
mPASMC	2	0.39%
U251	2	0.39%
BHK	2	0.39%
KoptK1	2	0.39%
Min6	2	0.39%
L6	2	0.39%
PC12	2	0.39%
H4	1	0.20%
D54	1	0.20%
BHK-21	1	0.20%
CEM21;HeLa	1	0.20%
BJ	1	0.20%
CRL-5865	1	0.20%
DKS8	1	0.20%
BG1	1	0.20%
LNCaP	1	0.20%
COS-7	1	0.20%
CCL-185	1	0.20%
BJAB	1	0.20%
Hct116	1	0.20%
HKE3	1	0.20%
other	87	16.99%

Table S5. MC and GE have significantly overlapping yet distinct hit sets.

Numbers represent compounds tested in both experiments. The null hypothesis of MC and GE hit sets being independent can be rejected for the set of all compounds and the DOS collection based on *p*-values calculated with Fisher's exact test. The BIO set does not show significant overlap because both MC and GE identify many of the 4053 compound as hits. Therefore, a large overlap is expected.

	n	**hits MC**	**hits GE**	**overlap**	***p*-value**
BIO	4053	3018	1557	1148	0.81
DOS	16194	6026	1837	912	1.13E-31
all	20247	9044	3394	2060	3.70E-94

Table S6. Compounds active in both MC and GE assays are often promiscuous

Shown are compounds active in both MC and GE with a promiscuity probability > 0.5 (see "Promiscuity probability" and Dančík *et al*. (20)) for which common names were available.

compound_name	PubChem_CID
3'-fluorobenzylspiperone	3248000
5-iodotubercidin	1830
5-nitroso-8-cinnolinol	44483284
AMA-2	160020
BADGE	2286
CD-437	135411
dihydroergocristine	11072143
FCCP	3330
GR 55562	128018
H7	3542
L-741,742	133008
LE-135	10410894
LY-294002	3973
MCG-D0103	9865515
NSC95397	262093
PXD101	6918638
R(+)-6-bromo-APB	10452020
R(-)-2,10,11-trihydroxy-N-propyl-noraporphine	6603798
RO 31-8220	5083
SB-415286	4210951
SCH-79797	4259181
ZL3VS	5497183
aminoacridine	7019
aminopurvalanol A	6604931
apomorphine hydrochloride	6005
bepridil	2351
beta-peltatin	92122
calcimycin	40486
cantharidin	2545
cicloheximide	6197
colchicine	6167
cucurbitacin I	44483311
curcumin	969516
daunorubicin	30323
emetine	10219
etoposide	36462
fenbendazole	3334
hinokitiol	3611
hycanthone methanesulfonate	3634
mefloquine	4046
metergoline	28693
methiazole	6604471
mycophenolic acid	446541

niclosamide	4477
nocodazole	4122
ouabain	439501
parbendazole	26596
phorbol myristate acetate	27924
pimozide	16362
puromycin	2724365
sanguinarium chloride	5154
suloctidil	657255
tacedinaline	2746
tetrandrine	73078
thapsigargin	446378
trifluoperazine	5566
trifluridine	6708818
tyrphostin A9	5614
tyrphostin AG 1296	2049
tyrphostin AG-1478	2051
tyrphostin AG879	6809654

Supporting Datasets

Dataset S1. Listing of assays and assay measurements published in ChemBank.
ChemBank experiment IDs and links are listed for each assay (provided as a separate Excel file).

Dataset S2. Listing of assay measurements published in PubChem and BARD.
PubChem Assay IDs (AIDs) and BARD Assay Definition IDs (ADIDs) are listed for each assay (provided as a separate Excel file).

Supporting References

1. Kelly AR, *et al.* (2009) Accessing skeletal diversity using catalyst control: formation of n and n + 1 macrocyclic triazole rings. *Org. Lett.* 11(11):2257-2260.
2. Marcaurelle LA, *et al.* (2010) An aldol-based build/couple/pair strategy for the synthesis of medium- and large-sized rings: discovery of macrocyclic histone deacetylase inhibitors. *J. Am. Chem. Soc.* 132(47):16962-16976.
3. Gerard B, *et al.* (2011) Synthesis of a stereochemically diverse library of medium-sized lactams and sultams via S(N)Ar cycloetherification. *ACS Comb. Sci.* 13(4):365-374.
4. Fitzgerald ME, *et al.* (2012) Build/couple/pair strategy for the synthesis of stereochemically diverse macrolactams via head-to-tail cyclization. *ACS Comb. Sci.* 14(2):89-96.
5. Comer E, *et al.* (2011) Fragment-based domain shuffling approach for the synthesis of pyran-based macrocycles. *Proc. Natl. Acad. Sci. U S A* 108(17):6751-6756.
6. Gerard B, *et al.* (2011) Large-scale synthesis of all stereoisomers of a 2,3-unsaturated C-glycoside scaffold. *J. Org. Chem.* 76(6):1898-1901.
7. Gerard B, *et al.* (2012) Application of a Catalytic Asymmetric Povarov Reaction using Chiral Ureas to the Synthesis of a Tetrahydroquinoline Library. *ACS Comb. Sci.*
8. Gerard B, *et al.* (2013) Synthesis of stereochemically and skeletally diverse fused ring systems from functionalized C-glycosides. *J. Org. Chem.* 78(11):5160-5171.
9. Nielsen TE & Schreiber SL (2008) Towards the optimal screening collection: a synthesis strategy. *Angew. Chem. Int. Ed. Engl.* 47(1):48-56.
10. Burke MD & Schreiber SL (2004) A planning strategy for diversity-oriented synthesis. *Angew. Chem. Int. Ed. Engl.* 43(1):46-58.
11. Rogers DJ & Tanimoto TT (1960) A Computer Program for Classifying Plants. *Science* 132(3434):1115-1118.
12. Rogers D & Hahn M (2010) Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50(5):742-754.
13. Peck D, *et al.* (2006) A method for high-throughput gene expression signature analysis. *Genome Biol.* 7(7):R61.
14. Bolstad BM, Irizarry RA, Astrand M, & Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19(2):185-193.
15. Genedata Screener Assay Analyzer (Genedata, Basel, Switzerland), 10 (2013).
16. Stouffer SA, Suchman EA, DeVinney LC, Star SA, & Williams RMJ (1949) *Studies in Social Psychology in World War II: The American Soldier* (Princeton University Press, Princeton).
17. Gustafsdottir SM, *et al.* (2013) Multiplex cytological profiling assay to measure diverse cellular states. *PLoS One* 8(12):e80999.
18. Kamentsky L, *et al.* (2011) Improved structure, function and compatibility for CellProfiler: modular high-throughput image analysis software. *Bioinformatics* 27(8):1179-1180.
19. Hutz JE, *et al.* (2013) The multidimensional perturbation value: a single metric to measure similarity and activity of treatments in high-throughput multidimensional screens. *J. Biomol. Screen.* 18(4):367-377.
20. Dancik V, *et al.* (2014) Connecting Small Molecules with Similar Assay Performance Profiles Leads to New Biological Hypotheses. *J. Biomol. Screen.* 19(5):771-781.
21. Clemons PA, *et al.* (2011) Quantifying structure and performance diversity for sets of small molecules comprising small-molecule screening collections. *Proc. Natl. Acad. Sci. U S A* 108(17):6817-6822.