ConnectivityMap

GEO CMap LINCS User Guide v2.1

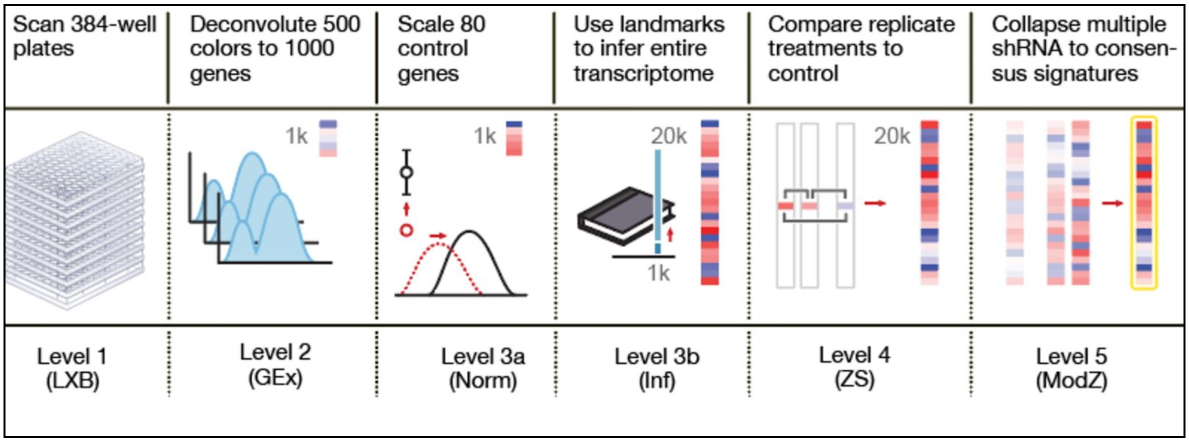
Preamble

1. Much of the work described in this guide is being written up for publication. We will do our best to address queries on the data and analytical methods, but kindly note that we have limited resources (most of which are directed towards making data and discoveries, rather than user support), so we appreciate your understanding and patience as we develop this resource. Email questions to clue@broadinstitute.org.

2. The L1000 datasets are deposited into GEO in the same spirit as initial human genome sequencing data. That is, while we do have elaborate QC/QA measures, that doesn't in itself mean that the data is free of errors or artifacts.
3. No part of this document may be reproduced in any form whatsoever without permission.
4. You are allowed to redistribute the data, but please do the field a favor by (a) referring back to the appropriate GEO project as the data source (b) describing differences, if any, between your pre-processing of the data and ours.

Update November 2017: The **L1000-based paper for Connectivity Map** has been published in Cell: http://www.cell.com/cell/abstract/S0092-8674(17)31309-0

Overview of signature generation process


Scan 384-well plates
Deconvolute 500 colors to 1000 genes
Scale 80 control genes
Use landmarks to infer entire transcriptome
Compare replicate treatments to control
Collapse multiple shRNA to consensus signatures
1k
1k
20k
1k
20k
Level 1 (LXB)
Level 2 (GEx)
Level 3a (Norm)
Level 3b (Inf)
Level 4 (ZS)
Level 5 (ModZ)

Deposited data is posted at all levels. (Unless you are interested in pre-processing details, our suggestion is for you to use Level 5.)

Note: We release data before before full analysis and before publication, with the expectation that the data will be useful to others even in its early form. Therefore, some best practices to keep in mind for using released data:

- check back here or contact us before you publish to see if any details have been updated
- if you notice something odd / inconsistent / unexpected in the data, please email us with details

Data files attached to the series

GSE70138 (*aka* **LINCS Phase II L1000 dataset)**

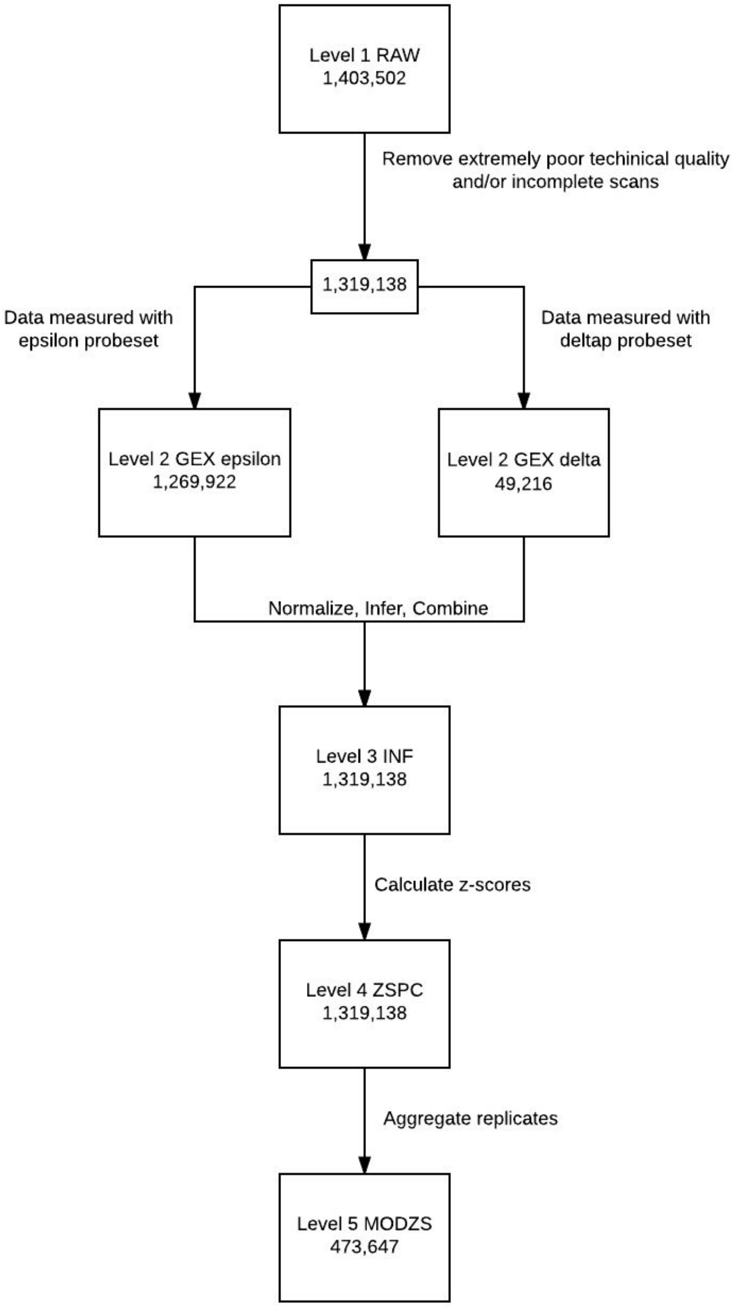
Filename	description	File type
GSE70138_Broad_LINCS_Level1_LXB_n345976.tar.gz	Level 1 data (raw fluorescence intensity measurements / LXB)	Gzipped tar of directory containing binary lxb files
GSE70138_Broad_LINCS_Level2_GEX_n345976x978.gctx.gz	Level 2 data (raw gene expression / GEX)	GCTX
GSE70138_Broad_LINCS_Level3_INF_mlr12k_n78980x22268_2015-06-30.gct.gz	Level 3 data (normalized & inferred / INF)	GCTX
GSE70138_Broad_LINCS_Level4_ZSPCINF_mlr12k_n345976x12328.gctx.gz	Latest Level 4 data (robust z-scores / ZSPC)	GCTX
GSE70138_Broad_LINCS_Level5_COMPZ_n118050x12328.gctx.gz	Latest Level 5 data (signatures from aggregating replicates)	GCTX
	Experimental metadata	
GSE70138_Broad_LINCS_inst_info.txt.gz	Metadata for individual experiments (levels 1-4)	Gzipped tab-delimited text
GSE70138_Broad_LINCS_sig_info.txt.gz	Metadata for signatures of aggregated replicates (level 5)	Gzipped tab-delimited text
GSE70138_Broad_LINCS_gene_info.txt.gz	Metadata for rows / genes of matrices	Gzipped tab-delimited text
GSE70138_Broad_LINCS_gene_info_delaprime.txt.gz	Metadata for rows / genes of matrices, applies to all profiles generated using the delta prime probe pool. See FAQ for details	Gzipped tab-delimited text
GSE70138_SHA512SUMS	Text file containing checksums calculated for each of the most recent above files, for use in verifying integrity of downloaded files	Gzipped text

GSE92742 (*aka* LINCS Phase I L1000 dataset)

Please note: an earlier version of this release incorrectly contained ~5k profiles that it should not have, they have subsequently been removed.

Filename	description	File type
GSE92742_Broad_LINCS_Level1_LXB_n1403502.tar.gz	Level 1 data (raw fluorescence measurements)	Gzipped tar
GSE92742_Broad_LINCS_Level2_GEX_delta_n49216x978.gctx.gz	Level 2 data for delta probes/features (raw gene expression / GEX)	Gzipped gctx
GSE92742_Broad_LINCS_Level2_GEX_epsilon_n1269922x978.gctx.gz	Level 2 data for epsilon probes/features (raw gene expression / GEX)	Gzipped gctx
GSE92742_Broad_LINCS_Level3_Q2NORM_n1319138x12328.gctx.gz	Level 3 data (normalized & inferred / INF)	Gzipped gctx
GSE92742_Broad_LINCS_Level4_ZSPC_n1319138x12328.gctx.gz	Level 4 data, plate-control normalized (robust z-scores / ZSPC)	Gzipped gctx
GSE92742_Broad_LINCS_Level5_MODZS_n473647x12328.gctx.gz	Level 5 data (moderated z-scores / MODZS)	Gzipped gctx
Experimental metadata		
GSE92742_Broad_LINCS_cell_info.txt.gz	Metadata for each cell line that was used in the experiments	Gzipped tab-delimited text
GSE92742_Broad_LINCS_gene_info.txt.gz	Metadata for each measured feature / gene (metadata for rows of the data matrices)	Gzipped tab-delimited text
GSE92742_Broad_LINCS_gene_info_including_delta.txt.gz	Metadata for each measured feature / gene (metadata for rows of the data matrices) including the delta probes that were used in a small number of initial experiments	Gzipped tab-delimited text
GSE92742_Broad_LINCS_inst_info.txt.gz	Metadata for each experiment in the Levels 3-4 matrices (metadata for the columns in the Levels 3-4 data matrices)	Gzipped tab-delimited text
GSE92742_Broad_LINCS_pert_info.txt.gz	Metadata for each perturbagen that was used in the experiments	Gzipped tab-delimited text
GSE92742_Broad_LINCS_pert_metrics.txt.gz	Calculated / derived / inferred metrics and annotations associated with each perturbagen that was used in the experiments.	Gzipped tab-delimited text
GSE92742_Broad_LINCS_sig_info.txt.gz	Metadata for each signature in the Level 5 matrix (metadata for the columns in the Level 5 data matrix)	Gzipped tab-delimited text

GSE92742_Broad_LINCS_sig_metrics.txt.gz	Calculated / derived / inferred metrics and annotations associated with each signature in the Level 5 matrix	Gzipped tab-delimited text
Supplemental data		
GSE92742_Broad_LINCS_auxiliary_datasets.tar.gz	Auxiliary datasets used for supplementary calculations. Please see the table below for a description of the individual files included in the archive.	Gzipped tar of set of gctx
Inventory / hash codes		
GSE92742_SHA512SUMS.txt.gz	Text file containing checksums calculated for each of the above files, for use in verifying integrity of downloaded files	Gzipped text



Overview of the GSE92742 dataset. Level 1 data are the raw mean fluorescent intensity (MFI) values that come directly from the Luminex scanner. Expression levels of the 978 landmark genes and controls are associated with the appropriate genes to form Level 2 data. Note that most genes have been measured with the epsilon probeset, but an earlier probeset, delta prime, was used in a small number of experiments. Level 2 data undergoes scaling and normalization steps and is then used to algorithmically infer expression of an additional 11,350 genes, forming Level 3 data. A robust z-scoring procedure is used to generate differential expression values from the normalized profiles (Level 4 data). Finally, we apply a moderated Z-scoring procedure to replicate samples of each experiment (generally 3 replicates are made) to compute a weighted average signature. For several types of genetic perturbagens (see pert_type table below; for example, where multiple hairpins were used to target the same gene) we further collapse level 5 data to create a single consensus gene signature (CGS) that represents that perturbation.

Supplemental Datasets
Contents of GSE92742_Broad_LINCS_auxiliary_datasets.tar.gz

Name	Brief description / use case
DS_GEO_n12031x22268.gctx	Collection of publicly available gene expression profiles on Affymetrix HGU133A arrays that was used to select landmark genes and train the inference model. *rows: 22,268 features* *columns: 12,301 samples*
DS_GEO_OLS_WEIGHTS_n979x21290.gctx	The matrix of weights learned by training the L1000 inference algorithm, ordinary least squares (OLS) linear regression, on DS_{GEO}. *rows: 21,290 inferred features* *columns: 978 landmark genes + intercept = 979*
DS_CMAP_AFFXB01_n566x22268.gctx	Pilot connectivity Map dataset of 455 gene expression profiles of human cell lines treated with 164 small molecule compounds that were used in simulations to determine the optimal number of landmarks. *rows: 22,268 features* *columns: 455 samples*
DS_REPRODUCIBILITY_n216x978.gctx	Samples of purified total RNA from six human cancer cell lines, purchased from Life Technologies, were subjected to L1000 profiling. L1000 expression profiles were generated consisting of 12 technical replicates for each of the six cell lines, all done in three consecutive, independent LMA batches, yielding 36 replicate profiles per cell line and a total of 216 total profiles (6 cell line x 12 replicates x 3 batches). These profiles were used to assess the technical reproducibility of the L1000 assay. *rows: 978 landmark genes* *columns: 216 samples*
DS_KDLM_n955x978.gctx	We evaluated if probes designed against individual target landmark genes worked in the multiplexed gene assay format. To assess the specificity of L1000 landmark probe measurements, we procured shRNAs from The RNAi Consortium (TRC) that targeted landmark genes, treated MCF7 and PC3 cell lines with these shRNAs. The resulting dataset contains, as columns, an individual shRNA targeting a landmark gene performed in either MCF7 or the PC3 cancer cell line. Rows are z-scores of all measured landmark genes. For each gene in each sample, we computed differential expression values (z-scores) by comparing the gene's expression value in the given sample to that same gene's expression values in all other samples. *rows: 978 landmark genes*

	columns: 955 samples
DS_GTEX_RNASEQ_n8555x12320.gctx	Compendium of 8,555 RNA-seq samples obtained from the GTEx consortium (version 6). A subset of 3,176 of these samples were also profiled on L1000 and were used to compare the two platforms. *rows: 12,320 genes* *columns: 8,555 samples*
DS_GTEX_L1000_n3176x12320.gctx	From the GTEx version 6 collection, a subset of 3,176 samples were generously donated by the GTEx consortium for profiling in the L1000 assay. *rows: 970 landmark + 11,350 inferred = 12,320 genes* *columns: 3,176 samples*
DS_GTEX_RNASEQ_LMONLY_n8555x970.gctx	For convenience of use during assay validation and inference testing, we separated out from the overall GTEx RNA-seq dataset ($\text{DS}_{\text{GTEx-RNA-seq}}$), the subset of landmark genes and refer to it as $\text{DS}_{\text{GTEx-RNA-seq-lmonly}}$ in the methods. Note that the GTEx pre-processing provided mappings for 970 of the landmark genes. *rows: 970 landmark genes* *columns: 8,555 samples*
DS_GTEX_RNASEQ_INF_n8555x12320.gctx	Inferred version of $\text{DS}_{\text{GTEx-RNA-seq}}$ dataset generated by applying $\text{DS}_{\text{GEO-OLS}}$ to $\text{DS}_{\text{GTEx-RNA-seq-lmonly.}}$ *rows: 970 landmark + 11,350 inferred = 12,320 genes* *columns: 8,555 samples*

GSE92743 (*aka* CMap-HBS Contest)

Filename	description	File type
GSE92743_Broad_Affymetrix_training_Level3_Q2NORM_n100000x12320.gctx.gz	Affymetrix data for 100,000 samples used by contestants for building their models	Gzipped gctx
GSE92743_Broad_GTEx_L1000_Level3_Q2NORM_n3176x12320.gctx.gz	All of the level 3 data (normalized & inferred / INF) of L1000 measurements on GTEx samples	Gzipped gctx
GSE92743_Broad_GTEx_L1000_Holdout_Level3_Q2NORM_n1000x12320.gctx.gz	Just the holdout level 3 data (normalized & inferred / INF) of L1000 measurements on GTEx samples	Gzipped gctx
GSE92743_Broad_GTEx_L1000_NotUsed_Level3_Q2NORM_n1526x12320.gctx.gz	Just the unused level 3 data (normalized & inferred / INF) of L1000 measurements on GTEx samples	Gzipped gctx
GSE92743_Broad_GTEx_L1000_Test_Level3_Q2NORM_n650x12320.gctx.gz	Just the test level 3 data (normalized & inferred / INF) of L1000 measurements on GTEx samples	Gzipped gctx
GSE92743_Broad_GTEx_RNAseq_Log2RPKM_q2norm_n3176x12320.gctx.gz	Level 3 data (normalized) of RNA-seq measurements on GTEx samples	Gzipped gctx
GSE92743_Broad_GTEx_gene_info.txt	Metadata for each measured feature / gene (metadata for rows of the data matrices)	tab-delimited text
GSE92743_Broad_GTEx_inst_info.txt	Metadata for each experiment in the Levels 3-4 matrices (metadata for the columns in the Levels 3-4 data matrices)	Gzipped tab-delimited text
GSE92743_Broad_OLS_WEIGHTS_n979x11350.gctx.gz	Matrix of weights used in current CMap L1000 inference model	Gzipped gctx
GSE92743_SHA512SUMS	Text file containing checksums calculated for each of the above files, for use in verifying integrity of downloaded files	text

GSE106127 (*aka* RNAi And CRISPR datasets)

Note that the data provided in this series are not new. They are subsets of the data, corresponding to genetic perturbational signatures of shRNAs and CRISPR reagents, that exist in GEO series GSE70138 and

GSE92742. Those seeking data for additional perturbation types will want to visit those repositories for the full datasets.

Filename	description	File type
GSE106127_level_4_zspc_n341336x978.gctx.gz	Level 4 data, plate-control normalized (robust z-scores / ZSPC)	Gzipped gctx
GSE106127_level_5_modz_n119013x978.gctx.gz	Level 5 data (moderated z-scores / MODZS)	Gzipped gctx
GSE106127_CGS_n33839x978.gctx.gz	Consensus gene signatures generated by combining level 5 signatures of individual shRNAs	Gzipped gctx
GSE106127_level_4_PRIME_zspc_n341336x978.gctx.gz	Level 4 data with the global first principal component (PC1) removed	Gzipped gctx
GSE106127_level_5_PRIME_modz_n119013x978.gctx.gz	Level 5 data with the global first principal component (PC1) removed	Gzipped gctx
GSE106127_CGS_PRIME_n33839x978.gctx.gz	CGS data with the global first principal component (PC1) removed	Gzipped gctx
GSE106127_pc_coeff_global_n978x978.gctx.gz	The principal component loadings derived by running PCA on the entire CMap level 5 data matrix (~470k signatures)	Gzipped gctx
GSE106127_inst_info.txt.gz	Metadata for each experiment in the Level 4 matrices (metadata for the columns in the Level 4 data matrices)	Gzipped tab-delimited text
GSE106127_sig_info.txt.gz	Metadata for each signature in the Level 5 matrices (metadata for the columns in the Level 5 data matrices)	Gzipped tab-delimited text
GSE106127_sig_metrics.txt.gz	Calculated / derived / inferred metrics and annotations associated with each signature in the Level 5 matrices	Gzipped tab-delimited text
GSE106127_CGS_meta.txt.gz	Metadata for each CGS signature in the CGS matrix (metadata for the columns in the CGS data matrix)	Gzipped tab-delimited text
GSE106127_CGS_PRIME_meta.txt.gz	Metadata for each CGS signature in the CGS PRIME matrix (metadata for the columns in the CGS PRIME data matrix)	Gzipped tab-delimited text
GSE106127_gene_info.txt.gz	Metadata for each measured feature / gene (metadata for rows of the data matrices)	Gzipped tab-delimited text

GSE106127_SHA512SUMS.txt.gz	Text file containing checksums calculated for each of the above files, for use in verifying integrity of downloaded files	Gzipped tab-delimited text

FAQ on Broad Institute LINCS Center Data Deposition

Q: Where do I get all of the LINCS CMap L1000 data?

A: All LINCS-funded CMap L1000 data is deposited into GEO.

- LINCS Phase 1 data is in GEO Series GSE92742. As this represents an earlier phase of LINCS, it will not be updated except for bug fixes, if any.
- LINCS Phase 2 data is in GEO series GSE70138. This series will be updated every 6 months as more L1000 data is produced and QC'ed over the duration of the LINCS program (Starting in 2016 through 2020). The same data will also become available through the LINCS DCIC portal. (http://lincsportal.ccs.miami.edu/dcic-portal/).
- Additionally, CMap / LINCS organizes datasets into discrete bundles to address particular questions in the form of contests, which are archived at GSE92743. The goal of this is to both engage the wider computational community in the improvement of analytics as well as to provide to LINCS users datasets along with benchmarks that are well organized for easier use.

Q: Who generated these data?

A: The Connectivity Map (CMap) group at the Broad Institute

Q: What is the relationship between CMap and LINCS?

A: The NIH established LINCS as a common fund project, inspired by the success of systematic approaches to functional discovery such as those developed by researchers using the NCI-60 cell lines for cytotoxicity profiling, studies with yeast knock-out mutants, and gene expression studies in human cells like those demonstrated in the Connectivity Map.

Researchers at the Broad, funded in part by LINCS, developed the L1000 assay as an alternative, cost effective, high throughput assay for use in scaling up the Connectivity Map dataset. Using the L1000 assay, the LINCS Center for Transcriptomics at the Broad Institute has grown the CMap resource to encompass > 1M profiles.

Q: What is the difference between data from GEO, LINCS portal (lincsproject.org, lincs-dcic.org), and CLUE?

A: The Connectivity Map analysis platform (clue.io) contains L1000 data and other perturbagen datasets made from a variety of public, philanthropic, and industrially funded projects. NIH-supported LINCS data is deposited into GEO twice per year. That same data from GEO is also imported by the LINCS DCIC into the LINCS portal at lincsproject.org, so that it is inter-operable with other LINCS data. So, data in CLUE is a superset of public domain L1000 datasets.

Q: How do I get access to the other data at CLUE.io

A: Academic users can register and use the CLUE tools to query the L1000 data from across multiple funding sources. Groups from academic institutions and from companies who want to collaborate can email us at clue@broadinstitute.org.

Q: What are the kinds of files available for download from GEO?

A: Files provided contain data matrices and metadata annotation.

Type	Format	Notes
Matrix of numbers	GCTx	Binary format based on HDF5 that enables faster i/o than text. Code is available at cmap github. Use the cmapR, cmapPy and/or cmapM repositories.
Experimental Metadata	TXT	Information on perturbagens and cell lines that were profiled
Metrics	TXT	Statistics computed on signatures that reflect their characteristics, including reproducibility of profiles and the magnitude of gene expression changes

Q: Specifically, what are the metadata and metrics data files we provide?

A: This table shows the filenames for the metadata and metrics data.

File	Description
GSE***_Broad_LINCS_cell_info.txt.gz	Metadata describing cell lines used in perturbagen treatments
GSE***_Broad_LINCS_gene_info.txt.gz	Metadata describing measured and inferred genes
GSE***_Broad_LINCS_inst_info.txt.gz	Metadata pertaining to individual profiles (or instances, experiments)
GSE***_Broad_LINCS_pert_info.txt.gz	Metadata describing each perturbagen used in experiments
GSE***_Broad_LINCS_sig_info.txt.gz	Metadata for each signature in the Level 5 matrix
GSE***_Broad_LINCS_pert_metrics.txt.gz	Calculated / derived / inferred metrics and annotations associated with each perturbagen that was used in the experiments.
GSE***_Broad_LINCS_sig_metrics.txt.gz	Calculated / derived / inferred metrics and annotations associated with each signature in the Level 5 matrix

Q: Anything more about the metadata?

A: The meta data is accurate to the best of our knowledge at the time of deposition. However, given the large size of the data and the many people, organizations, and processes involved, there will inevitably be errors or holes. We will do our best to fix any errors, however accessing via the CLUE API is likely a more convenient and up-to-date mode for subscribers.

If you notice an error => Please email us at clue@broadinstitute.org

For CLUE users, the latest form of metadata is available via the CLUE API at clue.io/api

Q: How are new data releases made on GEO?

A: Data releases are appended to the GEO series object. This way you can see the earlier releases in addition to the current one. Each release (approximately once every 6 months) is a full package (i.e earlier releases are not overwritten). This enables researchers using the data for the first time to get the latest and greatest release, while at the same time providing a persistently addressable source for earlier datasets.

Q: Where do I learn more about the LINCS project?

A: For information on the LINCS consortium, visit http://lincsproject.org
For information on the Broad Institute Connectivity Map and LINCS efforts, visit https://clue.io

Q: What is GTEx?

A: GTEx stands for Genotype Tissue Expression, and the project is aimed at finding associations between genotype and gene expression. For more information about the project, including experimental and analytical methods, please consult the project website: http://www.gtexportal.org/home/documentationPage.

Q: Can you describe the inference contest briefly?

A: CMap utilizes a novel, high-throughput gene expression profiling technology called L1000 to generate gene expression profiles at scale. Essential to this approach is that, instead of measuring all ~20,000 genes in the human genome, CMap measures a select subset of approximately 1,000 genes and uses these “landmark” gene measurements to computationally infer a large portion of the remainder. The current algorithm is effective but imperfect, and improving the imputation methods will have an immediate impact on the quality of data and the biologically meaningful connections that can be discovered. With this in mind, we launched a crowdsourcing contest on the TopCoder platform to stimulate exploration of new and improved inference methods. The data used for that contest are available in this series. More details on the contest can be found here: https://clue.io/contest (Contest 1: Inference Challenge).

Q: What is the plan to profile non-cancer lines?

A: Most current data is from cancer cell lines as they are easier to work with in the lab. In LINCS Phase II, there is new emphasis on non-cancer cell lines and it is expected that the proportion of non-cancer cell line data will grow.

Q: Does CMap use Affymetrix chips anymore?

A: No, we use L1000.

Q: Are there plans for CMap to use RNA-Seq?

A: Not in production because (a) there is no concrete evidence yet that RNA-Seq improves connectivity analysis compared to L1000, and (b) reagent cost is ~$1.50 per sample - considerably less than the cost of RNA-Seq. Even at Broad, where there is considerable sequencing capacity, the cost of RNA-Seq is not yet competitive.

Q: Can I nominate a cell line / drug / gene for profiling?

A: Yes; visit https://clue.io/nominate.

Q: Are there control treatments in the dataset?

A: Yes - DMSO is the control for compound treatments. Empty vector and other forms of non-gene-coding inserts (e.g LacZ, GFP, etc) are controls for genetic perturbagens.

Q: How are differential signatures computed?

A: We take the difference between a treatment of interest and all other perturbagens on the same 384-well assay plate (this is referred to as a population control). In other forms of analyses we compare the treatment to a control such as DMSO or an empty vector. In our experience, use of a population control is a more rigorous form of signature generation because it is less sensitive to variations arising from inert perturbagens (which are seldom truly inert).

Q. What is the difference between a profile and a signature?

A: A profile (also termed an experiment or an instance) represents data points that are generated from a perturbagen used to treat a particular cell type at a specified treatment dosage and for a specified duration of treatment. The numbers in a profile represent either the raw fluorescent intensity values (level 1 or raw data) or these numbers post deconvolution (level 2) or post normalization (e.g., quantile normalization) which leads to level 3 data. Finally, profiles are compared to appropriate controls to generate a list of differentially expressed features (level 4). Usually we do each experiment with 3 replicates that are then robustly averaged into one differentially expressed vector, to create a signature (level 5).

Q: I have derived a hypothesis from these data and need reagents (compounds or genetic reagents like shRNAs or CRISPRs or follow-up). Any suggestions?

A: The Broad's compound management team manages access to compounds, and when possible they distribute compounds for a modest plating fee. Genetic reagents are procured or synthesized with the help of the Broad Genetic Perturbagen (GPP) platform. Please email us at clue@broadinstitute.org and we will facilitate coordination with compound management or GPP. Alternatively, you can contact these Broad groups directly.

Q: Can I publish my study that uses the GEO CMap LINCS Data?

A: Yes!

LINCS Phase I (GSE92742) you may use the data in whatever way you see fit.

LINCS Phase II (GSE70138) data has not been published yet. However, we are glad if you have found the data to be useful and would like to incorporate it into your paper. You do not need to wait for the Broad CMap group to first publish a paper; the only exception to this is a paper describing the overall contours of the LINCS dataset. This approach is similar to The Cancer Genome Atlas (TCGA) policy, which allows user to incorporate TCGA findings with respect to a gene of interest prior to the TCGA official paper, but does not allow users to publish on the entire landscape of the TCGA cancer dataset (until the lead center(s) publish).

If your paper needs a citation to our work on L1000 or LINCS, please contact us at clue@broadinstitute.org.

Please note that as the L1000 and LINCS datasets mature we will revisit these terms. For now, given the evolving nature of the project, we request that you contact us (or the LINCS program at NIH) directly if you have any questions on data access, or consult their release policy.

Q: Is LINCS L1000 available to users from for-profit organizations?

A: Yes. All NIH funded L1000 data is freely accessible by all for download via GEO.

In addition, we offer subscriptions to clue.io as a portal to perturbational datasets that compliment the publicly funded data, and as an analysis environment that allows commercial users to combine proprietary and public data in a secure manner. See https://clue.io/subscribe for more information.

Analysis tools in clue.io are freely available for academic users.

Source code for algorithms is freely available through github.

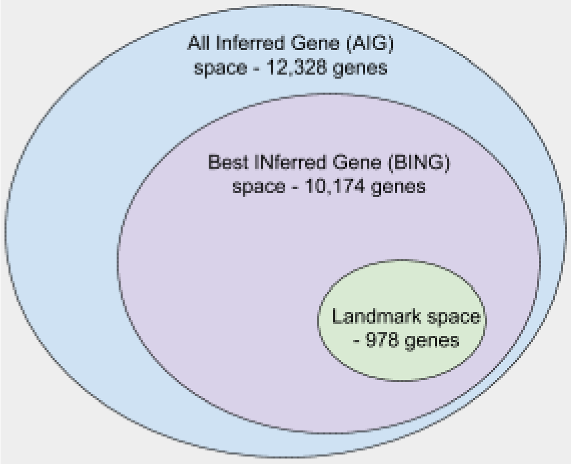
Q: What are features (rows) in the data matrix? Posed differently, what is the gene space accessible by L1000?

A: In L1000 datasets, features are genes and the matrix values correspond to their raw, normalized, or differential expression values, depending on which level of data is being used. L1000 reports on **12,328** unique genes; 978 of these are the landmark genes, which are directly measured. The remaining 11,350 are computationally inferred. 9,196 of these 11,350 genes are inferred with high fidelity, and together with the 978 landmarks comprise the Best INFerred Genes (BING) feature space, containing 10,174 genes total. We term the entire space of 12,328 genes as All Inferred Genes (AIG; see figure below).

The unique identifier for each row is the Entrez ID for the gene.

Note that earlier releases used Affymetrix-based identifiers. That is no longer necessary as data and inference is benchmarked against RNA-Seq datasets. Hence we use NCBI gene entrez gene identifiers (id and symbol).

Because we now map to gene symbols, the number of inferred features in the current matrices provided is 12,328 (unique genes) and not 22,268 (which used to be the count based on earlier Affymetrix probe sets).


All Inferred Gene (AIG) space - 12,328 genes
Best INferred Gene (BING) space - 10,174 genes
Landmark space - 978 genes

Q: For the inference contest, any further detail on the features (rows) in the data matrix?

In the files provided for the inference contest:

- For both RNA-seq and Affymetrix data, the features are all measured directly. The RNAseq data are quantile-normalized log2-scaled reads per kilobase-million (RPKM) values.

Q: What levels / types of data are available?

A: There are 5 levels of data available:

Level 1 - LXB - raw fluorescent count data generated by Luminex scanners. Fluorescent intensities measured for multiple beads of each of 500 different colors

Level 2 - GEX - Gene expression levels for the 978 landmark genes, deconvoluted from the measured fluorescent intensity values

Level 3 - INF_mlr12k - Gene expression (GEX, Level 2) levels that have been normalized to invariant gene set curves, and quantile normalized across each plate, and inferred values based on those normalized values

Level 4 - ZSVCINF_mlr12k - Z-scores for each gene based on INF_mlr12k / Level 3 with respect to the population of vehicle controls.
ZSPCINF_mlr12k - Z-scores for each gene based on INF_mlr12k / Level 3 with respect to the entire plate population

Level 5 - MODZ - replicate-collapsed z-score vectors based on Level 4.

For each of the levels except level 2, values are present for each of the 12,328 genes; in the case of the Level 2 GEX data, values are present for only the 978 Landmark features, since Level 2 GEX is determined prior to the inference step.

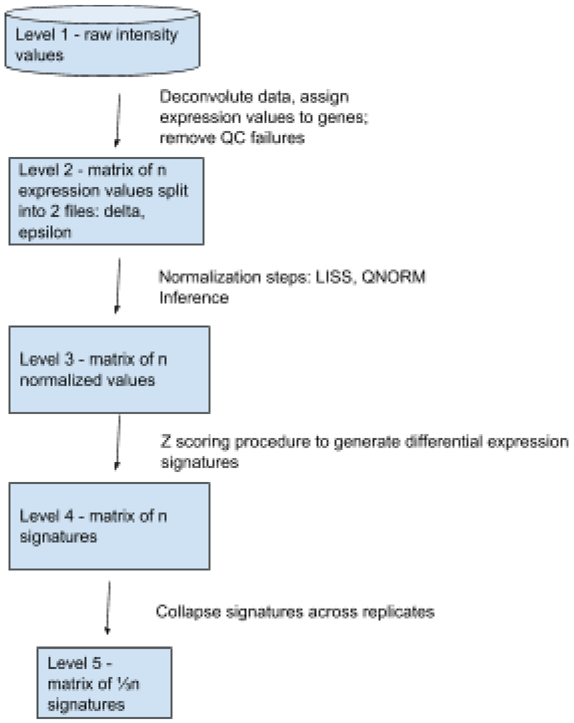
Q: How about levels for the Contest data?

There is one level of data available in series GSE92743: Level 3 - INF. This data is gene expression that has been normalized to invariant gene set curves, and quantile normalized across each plate. For L1000 data, it includes the inferred values calculated from those normalized values. For RNA-seq and Affymetrix, all features are measured directly.

Q: Why are there different numbers of columns between the different data levels?

A: For **GSE70138**, all data levels have the same number of columns (345,976) except for level 5 (MODZS) which has 118050 columns. The difference arises because the level 5 data is calculated by aggregating across individual replicates to generate a single signature for each group of replicates (generally 3 per experiment).

For **GSE92742**, there is more data at level 1 than the other levels because level 1 data includes samples that failed to pass QC. The level 2 data (GEX, which represents direct measurements) has been split into 2 files (“delta” and “epsilon”), based on which set of genes was measured directly; the total number of columns for these level 2 files is 49,216 + 1,278,882 = 1,328,098. That total (1,328,098) matches the number of columns for the level 3 (INF) and level 4 (ZSPC) data. As described above for GSE70138, the level 5 data (MODZS) is calculated by aggregating across the replicates within the level 4 data, and thus there are fewer columns - about ⅓ as many columns in the level 5 data as in the level 4 data.


Level 1 - raw intensity values
Deconvolute data, assign expression values to genes; remove QC failures
Level 2 - matrix of n expression values split into 2 files: delta, epsilon
Normalization steps: LISS, QNORM Inference
Level 3 - matrix of n normalized values
Z scoring procedure to generate differential expression signatures
Level 4 - matrix of n signatures
Collapse signatures across replicates
Level 5 - matrix of ⅓n signatures

Q: Ok, If i simply wanted to latest LINCS Phase II data, what are the specific file names (as of March 2017)

1. GSE70138_Broad_LINCS_**Level1**_LXB_n345976_2017-03-06.tar.gz
2. GSE70138_Broad_LINCS_**Level2**_GEX_n345976x978_2017-03-06.gctx.gz
3. GSE70138_Broad_LINCS_**Level3**_INF_mlr12k_n345976x12328_2017-03-06.gctx.gz
4. GSE70138_Broad_LINCS_**Level4**_ZSPCINF_mlr12k_n345976x12328_2017-03-06.gctx.gz
5. GSE70138_Broad_LINCS_**Level5**_COMPZ_n118050x12328_2017-03-06.gctx.gz

Q: Can I see the code that was used to calculate the different levels of data?

A: Yes, it is available at https://github.com/cmap/cmapM. Note that analysis code is available in other languages; see clue.io/code for details.

Q: What is a sample (also called profile, instance or experiment)?

A: A sample (or experiment or instance) refers to the data measured and inferred from biological material from a single well of a plate (or other container) in which cells (present in the well) may have been treated with a perturbagen (at a given dose and for a specified duration). The columns of the various GCTX matrices from Level 2 to Level 4 correspond to samples; Level 5 contains replicates that have been robustly averaged into a signature.

Q: What experimental metadata is available? What are the meanings of the experimental metadata fields?

A: The metadata is stored in tab-delimited text files attached to the series. Metadata for individual experiments are in the "inst_info.txt" file attached to each series. Metadata for the rows / genes is present in the "gene_info.txt" file attached to each series. Please see the section below (Explanation of Metadata Column Headers) for information about the columns.

Q: What are Series, Super Series and subset Series?

A: A Series record links together a group of related Samples and provides a focal point and description of the entire study. Series records may also contain tables describing extracted data, summary conclusions, or analyses. Due to GEO technical limitations, for the CMap L1000 submission to GEO the samples have been broken up into multiple series (subset Series), which are linked together via a master or Super Series (GSE70138, http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE70138). The Super Series GSE70138 contains supplementary data files, one for each level of data; these contain the data and metadata for every sample in the series. The files are in GCTx format. See also "What levels / types of data are available?".

Most importantly, if you simply want the matrices, ignore the sub-series and sub-projects etc (see above "*Ok, If i simply wanted to latest LINCS Phase II data, what are the specific file names (as of March 2017*)).

Q: What are your data update plans?

A: Over the course of the LINCS initiative, L1000 data generated as part of LINCS will be released twice a year to GEO. L1000 data generated from other sources will be released into the public domain as funding sources permit.

Q: How do I cite L1000?

A: Please cite http://biorxiv.org/content/early/2017/05/10/136168 (we will update this once the manuscript has been published). If you use our code libraries, please cite Enache et al.

Q: How do I refer to the contest data?

A: Please cite GSE92743 when referring to the contest dataset.

Q: Are there tools available to analyze the data?

A: The Broad CMap project provides tools at: https://clue.io.

The LINCS DCIC provides tools to analyze L1000 and other LINCS data: http://lincs-dcic.org.

Q: Can I see the code that was used to calculate the normalized and inferred L1000 data?

A: The code is available at https://github.com/cmap/cmapM

Q: Where is the code to access GCTx files?

A: The GEO releases are in GCTx file format; the code in Python, R, and Matlab is available on GitHub; for more information see:
https://clue.io/code

Q: What is the GCTx file format?

A: GCTx is an HDF5-based file format that we developed to optimize i/o from files on local disk. As CMap datasets became larger and larger, we noticed that TXT based storage was slow. The Physics community had pioneered the use of HDF5 as a "data model, library, and file format for storing and managing data. It supports an unlimited variety of datatypes, and is designed for flexible and efficient I/O and for high volume and complex data". We adapted this so as to be able to store every profile we generate in a single unified file to which new signatures can be appended as they are generated. Each column in the GCTx matrix is a signature and each row is a gene in that signature. Other advantages of the GCTx / HDF5 file format include rapid random-access loading of the data. A preprint citing the GCTX format is available here; if you use these files or any of the Python, Matlab, R, or Java software packages to analyze these files, please cite the preprint. For more information see:
https://clue.io/code

Q: What is the source code policy?

A: All algorithm code is available in accordance with the BSD 3-Clause license. In accordance with this license, users can use this code as needed, as long as the BSD copyright and license notice are included in it.

For reference, the full license is provided here.

In your publications, please cite the code as Enache et al.

Q: Is the data "fully QC'd"

A: L1000 data production is a high throughput screening operation that uses several sophisticated pieces of automated instrumentation. We have implemented a large number of quality control and assurance procedures and routinely fail samples that seem sub-optimal. Yet the guiding principle of LINCS / CMap is to make data that has passed all technical QC measures available, even if parts of the data aren't perfect. Hence, we ask that you (a) pay attention to metrics included in metadata headers (rather than assuming all columns of data pass QC, (b) be aware that as in any HTS there could be unobserved issues (and if you observe any, email us at clue@broadinstitute.org), and (c) consider accessing data that we have vetted and organized for analysis in webapps at clue.io (as opposed to downloading data).

Q: Where do I go to learn more?

A: https://clue.io/

Q: Have you considered X or noticed Y about the data?

A: Maybe, but we would be glad to hear from you about it either way - see below, “How do I get other questions answered or send a comment to CMap?”

Q: How do I use the SHA512 sums / GSE70138_SHA512SUMS.txt.gz file?

This text files provides computed checksums for the files attached to the series. Specifically, the SHA512 version of the SHA-2 family of hash functions has been used. To determine if a file has been downloaded correctly, after downloading calculate the SHA512 hash of the downloaded file and then compare to the entry in the SHA512SUMS file; if there is a discrepancy then there was a problem with the download. Note that some operating systems will both do the calculation of the checksum on the downloaded file and compare to the expected checksum.

Q: Why are some compounds listed as “restricted”? For example, this appears in an entry:

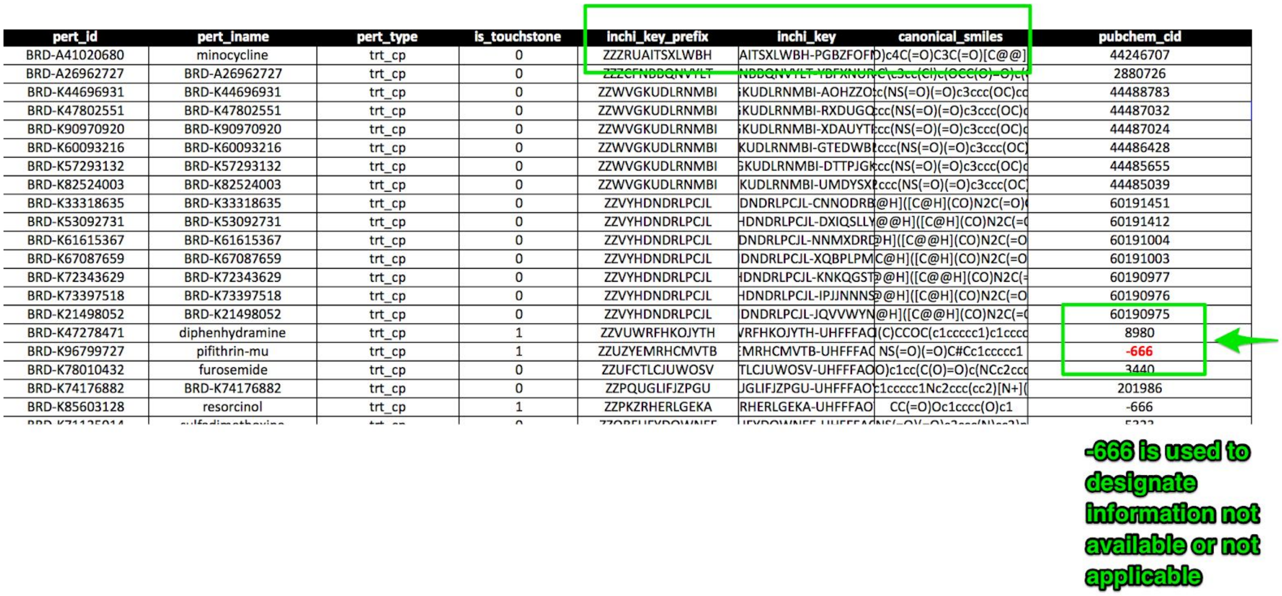
BRD-U97083655 <restricted>

A: Very few compounds are in this category; there are 41 and 10 unique restricted compounds in GSE92742 and GSE70138, respectively.

Restricted compounds represent molecules synthesized by chemists associated with the LINCS effort at the Broad and Harvard Medical School who wanted to contribute the profiles to the public domain but didn't yet have permission to release structures. If these small number of compounds affect your work, email us and we can connect you to the data owners.

Q: In the metadata, what does “-666” mean?

A: Any missing value is represented by -666 (see table below), which indicates that the information is not available or not applicable.



pert_id	pert_iname	pert_type	is_touchstone	inchi_key_prefix	inchi_key	canonical_smiles	pubchem_cid
BRD-A41020680	minocycline	trt_cp	0	ZZZRUAITSXLWBH	AITSXLWBH-PGBZFOFI	O)c4C(=O)C3C(=O)[C@@]	44246707
BRD-A26962727	BRD-A26962727	trt_cp	0	ZZZCFNBDQNVYLT	NBDQNVYLT-YBFXNUR	C\c3cc(Cl)c(OCC(O)=O)c(	2880726
BRD-K44696931	BRD-K44696931	trt_cp	0	ZZWVGKUDLRNMBI	KUDLRNMBI-AOHZZO	c(NS(=O)(=O)c3ccc(OC)cc	44488783
BRD-K47802551	BRD-K47802551	trt_cp	0	ZZWVGKUDLRNMBI	KUDLRNMBI-RXDUGQ	ccc(NS(=O)(=O)c3ccc(OC)c	44487032
BRD-K90970920	BRD-K90970920	trt_cp	0	ZZWVGKUDLRNMBI	KUDLRNMBI-XDAUYT	ccc(NS(=O)(=O)c3ccc(OC)c	44487024
BRD-K60093216	BRD-K60093216	trt_cp	0	ZZWVGKUDLRNMBI	KUDLRNMBI-GTEDWB	ccc(NS(=O)(=O)c3ccc(OC)	44486428
BRD-K57293132	BRD-K57293132	trt_cp	0	ZZWVGKUDLRNMBI	GKUDLRNMBI-DTTPJGK	cc(NS(=O)(=O)c3ccc(OC)c	44485655
BRD-K82524003	BRD-K82524003	trt_cp	0	ZZWVGKUDLRNMBI	KUDLRNMBI-UMDYSX	ccc(NS(=O)(=O)c3ccc(OC)	44485039
BRD-K33318635	BRD-K33318635	trt_cp	0	ZZVYHDNDRLPCJL	DNDRLPCJL-CNNODRB	@H]([C@H](CO)N2C(=O)(	60191451
BRD-K53092731	BRD-K53092731	trt_cp	0	ZZVYHDNDRLPCJL	HDNDRLPCJL-DXIQSLLY	@@H]([C@H](CO)N2C(=(	60191412
BRD-K61615367	BRD-K61615367	trt_cp	0	ZZVYHDNDRLPCJL	DNDRLPCJL-NNMXDRD	@H]([C@@H](CO)N2C(=O	60191004
BRD-K67087659	BRD-K67087659	trt_cp	0	ZZVYHDNDRLPCJL	DNDRLPCJL-XQBPLPM	C@H]([C@H](CO)N2C(=O	60191003
BRD-K72343629	BRD-K72343629	trt_cp	0	ZZVYHDNDRLPCJL	HDNDRLPCJL-KNKQGST	@@H]([C@@H](CO)N2C(=	60190977
BRD-K73397518	BRD-K73397518	trt_cp	0	ZZVYHDNDRLPCJL	HDNDRLPCJL-IPJJNNNS	@@H]([C@H](CO)N2C(=O	60190976
BRD-K21498052	BRD-K21498052	trt_cp	0	ZZVYHDNDRLPCJL	DNDRLPCJL-JQVVWYN	@H]([C@@H](CO)N2C(=(	60190975
BRD-K47278471	diphenhydramine	trt_cp	1	ZZVUWRFHKOJYTH	VRFHKOJYTH-UHFFFAO	(C)CCOC(c1ccccc1)c1cccc	8980
BRD-K96799727	pifithrin-mu	trt_cp	1	ZZUZYEMRHCMVTB	EMRHCMVTB-UHFFFAO	NS(=O)(=O)C#Cc1ccccc1	-666
BRD-K78010432	furosemide	trt_cp	0	ZZUFCTLCJUWOSV	TLCJUWOSV-UHFFFAO	O)c1cc(C(O)=O)c(NCc2ccc	3440
BRD-K74176882	BRD-K74176882	trt_cp	0	ZZPQUGLIFJZPGU	UGLIFJZPGU-UHFFFAO	c1ccccc1Nc2ccc(cc2)[N+](	201986
BRD-K85603128	resorcinol	trt_cp	1	ZZPKZRHERLGEKA	RHERLGEKA-UHFFFAO	CC(=O)Oc1cccc(O)c1	-666
BRD-K71125014	sulfadimethoxine	trt_cp	0	ZZQBFUFYDOWNEF	UFYDOWNEF-UHFFFAO	NS(=O)(=O)c2ccc(N)cc2)n	5323

-666 is used to designate information not available or not applicable

Q: How were SMILES strings and InChIKey generated?

A: We rely on vendor-provided annotation coupled with literature curation to determine SMILES strings, and we generate InCHIKeys from those using chemaxon cheminformatics tools. SMILES strings and InCHIKeys are indicated in the figure above.

Q: What is the relationship between the older CMap build 02 website and clue.io?

A: CMap build 02 was assembled from approximately 7,000 profiles based on Affymetrix based scans, data from ~1,300 small molecules remains available (note that this data is in maintenance mode only; no fixes or updates are funded). The Build 02 site will be maintained; email us at clue@broadinstitute.org with any support issues. The CLUE.io site contains data from ~1 million profiles generated by L1000, covering approximately 50,000 unique perturbagens (as of March 2017). All newer data is being made available from this site.

Learning more / commenting / contacting

Q: Where do I go to learn more?

https://clue.io/

Q: How do I get other questions answered or send a comment to CMap?

A:

- Email us at: clue@broadinstitute.org
- Attend “office hours”. See Instructions
- Check for updates on Twitter @CMap_Broad

Explanation of Metadata and Metrics Data Column Headers

bead_batch: One instantiation of a complete set of beads, which have been coupled to probes at one time, under the same conditions.

bead_revision: The set of beads that applies to a particular collection of gene pairs for each bead color.

bead_set: A pair of barcodes for each gene pair used for a bead color; gene pairs are used for the tag duo procedure, where a bead color is coupled to two different genes.

cell_id:

cl_center_specific_id synonym for cell_id.

count_cv: The coefficient of variation of bead counts.

count_mean: The mean of per well-analyte bead counts.

det_mode: The detection mode used for acquiring L1000 data. Can be either DUO (two genes per analyte color) or UNI (one gene per analyte color).

det_plate: Detection plate, the plate of L1000 experiments that, at the end of the assay pipeline, is put through the Luminex scanners to detect the levels of landmark gene amplicons.

det_well: Detection well, which refers to each well of the detection plate in which an L1000 experiment is conducted.

distil_cc_q75: 75th quantile of pairwise spearman correlations in landmark space of replicate level 4 profiles.

distil_id: ID of an individual replicate profile, referred to as level 4 / z-score data, that is used in creating the signature from replicates assayed together on an L1000 plate. The signature is referred to as level 5 / aggregated z-score data.

distil_nsample: Number of individual replicate profiles (level 4 / z-score) that were used to create the signature (level 5 / aggregate z-score).

distil_ss: The number of significantly differentially expressed transcripts that arise from a particular perturbagen treatment.

ds_index: An arbitrary index used within a file; do not use this.

icc: Inter-cell connectivity (ICC). The similarity (aggregated WTCS) between signatures of a given perturbagen across cell lines. This number ranges between -1 and 1, and the higher the number, the more similar the signatures across cell lines. Only exemplar signatures are used in computing ICC. See **is_exemplar** for more details.

inf_model: Inference model designation.

inst_id: synonym for distil_id.

is_exemplar: A boolean indicating whether the given signature is an exemplar. Due to the redundancy of the CMap database, meaning that some perturbagens have many signatures even within the same cell line, it is convenient to identify a single 'exemplar' signature for each perturbagen in each cell line. These signatures are specifically designated for further analysis, such as ICC and aggregate TAS. Exemplar signatures were selected according to the following process. For each perturbagen in each cell line:

1. If possible, consider only signatures with between 2 and 6 replicates.

2. Within these signatures, select the one with highest transcriptional activity score (TAS). See **tas** for more details.
3. If there are no signatures that have between 2 and 6 replicates, simply select the one with highest TAS.

is_gold: A heuristic for assessing whether a signature is reproducible and distinct. Requirements include: distil_cc_q75 >= 0.2 and pct_self_rank_q25 <= 0.05.

is_touchstone: A boolean indicating whether the corresponding signature or perturbagen is a member of the Touchstone dataset. Touchstone is a term applied to the subset of CMap perturbagens that are well-annotated and that were systematically profiled across the majority of the core set of 9 cell lines at standardized conditions. Because of these properties Touchstone dataset well-suited as a reference compendium against with to compare external queries.

mfc_plate_dim: Manufacturer's stated dimensions of the pert plate.

mfc_plate_id: Manufacturer's designated plate id.

mfc_plate_name: Name of the plate as designated by the manufacturer.

mfc_plate_quad: Quadrant of the plate as designated by the manufacturer.

mfc_plate_well: Well of the pert plate as designated by the manufacturer.

ngenes_modulated_up_lm: The number of landmark genes that show increased expression in cells treated with perturbagen.

ngenes_modulated_dn_lm: The number of landmark genes that show decreased expression in cells treated with perturbagen.

pct_self_rank_q25: Self connectivity of replicates expressed as a percentage of total instances in a replicate set.

pert_desc: A brief summary of the biological function (for genetic perturbagens) or mechanism of action (for compound perturbagens).

pert_dose: Precise amount of compound used to treat cells.

pert_dose_unit: Unit (generally micromolar) applied to the dose of compound used to treat cells.

pert_id: A unique identifier for a perturbagen that refers to the perturbagen in general, not to any particular batch or sample.

pert_idose: The concatenation of pert_dose and pert_dose_unit to create a string containing the dose information. We use a standardized dose for a perturbagen treatment. For example, the less common dose of 10.04 is rounded to 10. This enables grouping of signatures by a common dose.

pert_iname: The internal (CMap-designated) name of a perturbagen. By convention, for genetic perturbations CMap uses the HUGO gene symbol.

pert_itime: The concatenation of pert_time and pert_time_unit to create a string containing the length of time that a perturbagen was applied to the cells. We use a standardized time for a perturbagen treatment. For example, if data is made by treating cells with a perturbagen for 5.5 hours, we round that time to the more common treatment time of 6 hours.

pert_mfc_id: A manufacturer's id for the perturbagen; by convention, for compounds registered with Broad Compound Management this is the full BRD containing both the compound ID and the batch ID.

pert_time: The length of time, expressed as a number, that a perturbagen was applied to the cells; does not include the unit.

pert_time_unit: The unit that applies to the pert_time numerical value.

pert_type: Abbreviated designation for perturbagen type, referring to compound or genetic perturbagens that are used in cell treatments to assess gene expression effects. The various pert_types used by CMap are listed in the table below.

Treatment	pert_type
Compound	trt_cp
Ligand	trt_lig
shRNA for LoF	trt_sh
Consensus signature from shRNAs for LoF	trt_sh.cgs
cDNA for overexpression of wild-type of gene	trt_oe
cDNA for overexpression of mutated form of gene	trt_oe.mut
CRISPR for LoF	trt_xpr
Controls - vehicle (e.g DMSO)	ctl_vehicle
Controls - vector (e.g empty vector)	ctl_vector
Controls - consensus signature from sister shRNAs	trt_sh.css
Controls - consensus signature of vehicles	ctl_vehicle.cns
Controls - consensus signature of vectors	ctl_vector.cns
Controls - consensus signature of untreated	ctl_untrt.cns
Untreated cells	ctl_untrt

pert_vehicle: The solvent or other vehicle used to deliver the perturbagen.

pool_id: Landmark probe pool used; generally the pool is epison (most users should **ignore** the older pools such as delta and deltaprime).

provenance_code: A shorthand code that tracks the different steps in data processing.

qc_f_logp: The -log10 of p-value (for f statistic), representing the goodness of fit of the power model used to convert raw fluorescence intensity values to log2 expression values during the LISS process.

qc_iqr: The interquartile range of normalized expression within a level 3 profile.

qc_slope: The line slope in degrees (arctan of slope) of the line of best fit through the observed invariant set expression levels and their expected expression ranks.

rn_target_gene_id: synonym for pert_univ_id.

rna_plate: Name of the plate as it was used throughout the assay prior to detection; the name includes all information except the bead_batch_id suffix, for example: LJP005_A375_24H_X1.

rna_well: Name of a well within an rna_plate.

seeds_seq_6mer: The 6-mer oligonucleotide seed sequence for the given shRNA.

seeds_seq_7mer: The 7-mer oligonucleotide seed sequence for the given shRNA.

sig_id: A CMap unique identification number assigned to each signature generated from L1000 data.

tas: Transcriptional activity score, a measure of the L1000 transcriptional response elicited by a perturbagen. TAS is computed as the geometric mean of the signature strength and the 75th quantile of pairwise replicate correlations for a given signature

tas_q75: Aggregated transcriptional activity score. For a given perturbagen, tas_q75 is computed as the 75th quantile of its TAS across cell lines. The higher the number, the more generally active the perturbagen. Only exemplar signatures are used for computing tas_q75. See **is_exemplar** for more details.

target_seq: The sequence within a gene that is targeted by a hairpin shRNA for knockdown, to abolish expression of that gene. This term applies to shRNA experiments only.

zmad_ref: The reference population used for Z scoring; generally population or vehicle, but it could be custom (see provenance code for what was used).

LINCS related identifiers (useful to synchronize the above with metadata you might find on the LINCS DCIC portal)

sm_center_compound_id: synonym for pert_mfc_id

sm_dose: synonym for pert_dose.

sm_dose_unit: synonym for pert_dose_unit.

sm_lincs_id: synonym for pert_id.

sm_name: synonym for pert_iname.

sm_pert_type: synonym for pert_type.

sm_time: synonym for pert_time.

sm_time_unit: synonym for pert_time_unit

Contest related metadata fields

gtex_id identifier for the sample as used within the GTEx project.

id other identifier used within CMap for the sample.

tissue name / description of the tissue from which the sample was obtained from.

inst_id the identifier for the sample exactly as it appears in the matrix files attached to the series.