



# CRISPR v1.0

## Data release README

---

July 2018

# Data access

# Using the data

1. Please review caveats in using the data
2. The data is released pre-publication.
3. Our team is using these data to address Specific Aims of the LINCS project. We welcome your input and if your work overlaps with these aims, please consider reaching out to us to discuss.
4. You may freely download and use the data in your work.

# Build Files

File name	Description
CRISPR_Broad_LINCS_Level1_LXB.tar.gz	LXB - raw fluorescent intensity (FI) values measured for every bead detected by Luminex scanners.
CRISPR_Broad_LINCS_Level2_GEX_n978x55822.gctx.gz	Gene expression levels for the 978 landmark genes, deconvoluted from the measured fluorescent intensity values.
CRISPR_Broad_LINCS_Level3_INF_n55065x12328.gctx.gz	Gene expression (GEX, Level 2) are normalized to invariant gene set curves and quantile normalized across each plate.
CRISPR_Broad_LINCS_Level4_ZSPCINF_n55065x12328.gctx.gz	Z-scores for each gene based on Level 3 with respect to the entire plate population.
CRISPR_Broad_LINCS_Level5_MODZS_n18618x12328.gctx.gz	Replicate-collapsed z-score vectors based on Level 4.
Touchstone connectivity results (see “Touchstone Connectivities” section of the release page on clue.io)	<a href="#">Connectivity scores</a> of Level 5 z-score signatures to the CMap <a href="#">Touchstone dataset</a> .

More information on the data levels can be found [here](#).  
All files can be downloaded from the “Download” section of [the release page on clue](#).

# Important caveats in using these data

---

- This is an early access dataset released by the Broad Institute LINCS center
- Data generation and analysis is ongoing
  - Consequently, included samples and analysis techniques may change
- Once reviewed more thoroughly these data will be integrated with the remainder of the LINCS Phase II dataset at NCBI GEO [GSE70138](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE70138)

Please contact [ana@broadinstitute.org](mailto:ana@broadinstitute.org) if you have specific inquiries or analyses to discuss.

# Specific aims of LINCS

Aim 1. Compare shRNAs (RNAi) to sgRNAs (CRISPRs) from the perspective of generating transcriptional signatures of genetic loss of function

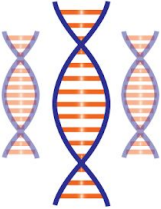

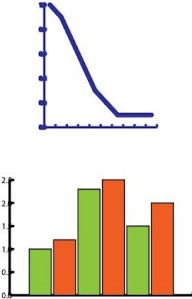
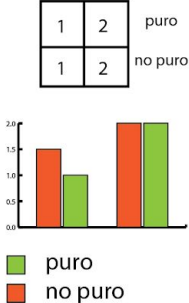
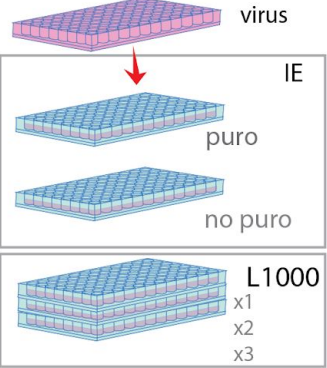
Aim 2. Develop computational approaches to determine off and on target effects of CRISPRs on the transcriptome

Aim 3. Develop computational approaches to integrate transcriptional signatures with cell viability and other relevant data; use these approaches to improve the ability to assign function to genes

Aim 4. Integrate small-molecule and genetic signatures for MoA and target elucidation

# Experimental Design

# Overview of data generation workflow

GENES LIST	CELL LINES	OPTIMIZATION	PRESCREEN	SCREEN, COLLECT LYSATE
	 <p>A735 A549 HA1E HT29 ... ...</p>			
Pick genes and order virus	Choose relevant cell lines	Determine optimal puromycin doses, amount of virus and amount of cell per well.	Test experimental conditions chosen	Infect cells; include 3 replicates for L1000 analysis

**Time-point: 96H**

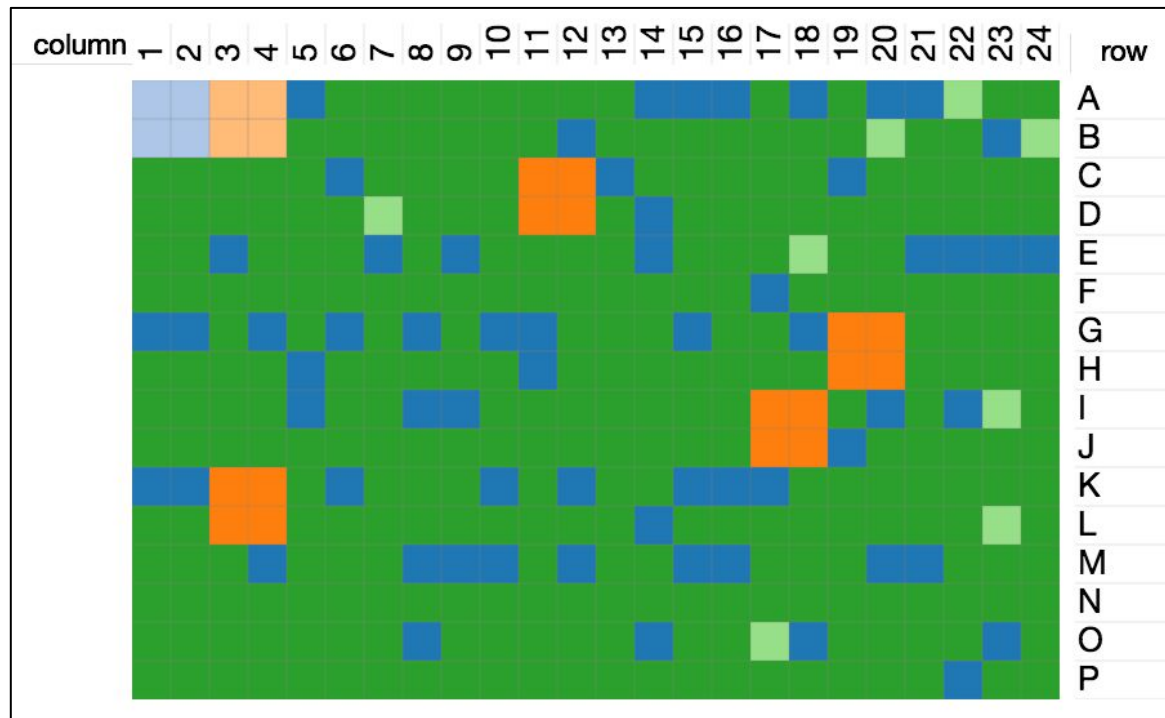


# Perturbagen types and controls

---

- Controls
  - Negative controls
    - Untreated wells (on average 8 per plate)
    - Non-targeting guides (on average 16 per plate)
      - These are different guides designed to not target any region of the human genome
  - Positive controls
    - Reference guides (4 per plate, in a fixed position)
- Targeting guides
  - Guides targeting directly measured (“landmark”) genes
  - Guides targeting non-directly measured genes

# Sample plate layout



Assay controls:

L1000 assay control

Negative controls:

Untreated well

Non-targeting guide

Positive controls:

Reference guide

Test guides:

Guide targeting directly  
measured gene

Guide targeting non-directly  
measured gene

# Data included

	Cell Line	<b>A375</b>	<b>A549</b>	<b>AGS</b>	<b>BICR6</b>	<b>ES2</b>	<b>HT29</b>	<b>MCF7</b>	<b>PC3</b>	<b>U251MG</b>	<b>YAPC</b>
	Cell line lineage	skin	lung	stomach	upper aero-digestive tract	ovary	large intestine	breast	prostate	central nervous system	pancreas
Virus plate 1	<b>XPR009</b>	3	3	3	3	3	3	3	3	3	3
Virus plate 2	<b>XPR016</b>	3	3	3	3	3	3	3	3	3	3
Virus plate 3	<b>XPR019</b>	3	3	3	3	3	3	3	3	3	3
Virus plate 4	<b>XPR025</b>	3	3	3	3	3	3	3	3	3	3
Virus plate 5	<b>XPR027</b>	3	3	3	3	3	3	3	3	3	3

Number inside squares corresponds to number of replicates included

# Data included

---

- Five 384-well virus plates of targeting guides x 10 cell lines
  - 3 replicates per plate
  - Wells with poor technical quality as failures/false in metadata files under column “QC\_pass”
- 1,884 landmark (directly measured) guide signatures, corresponding to 96 unique genes
- 15,428 non-landmark (not directly measured) guide signatures, corresponding to 789 unique genes

# Technical quality control measures assessed

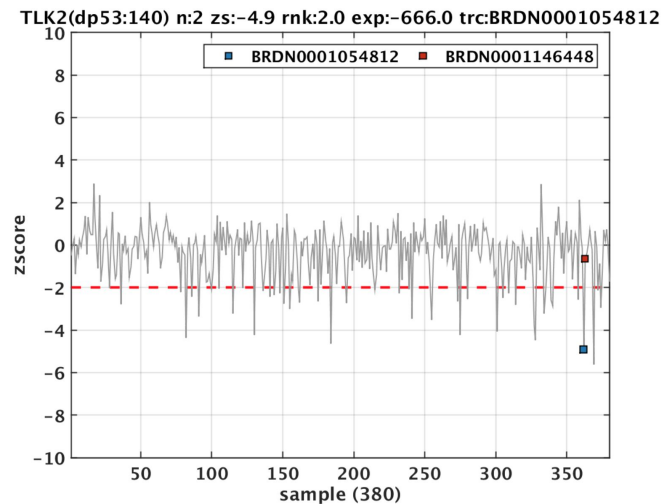
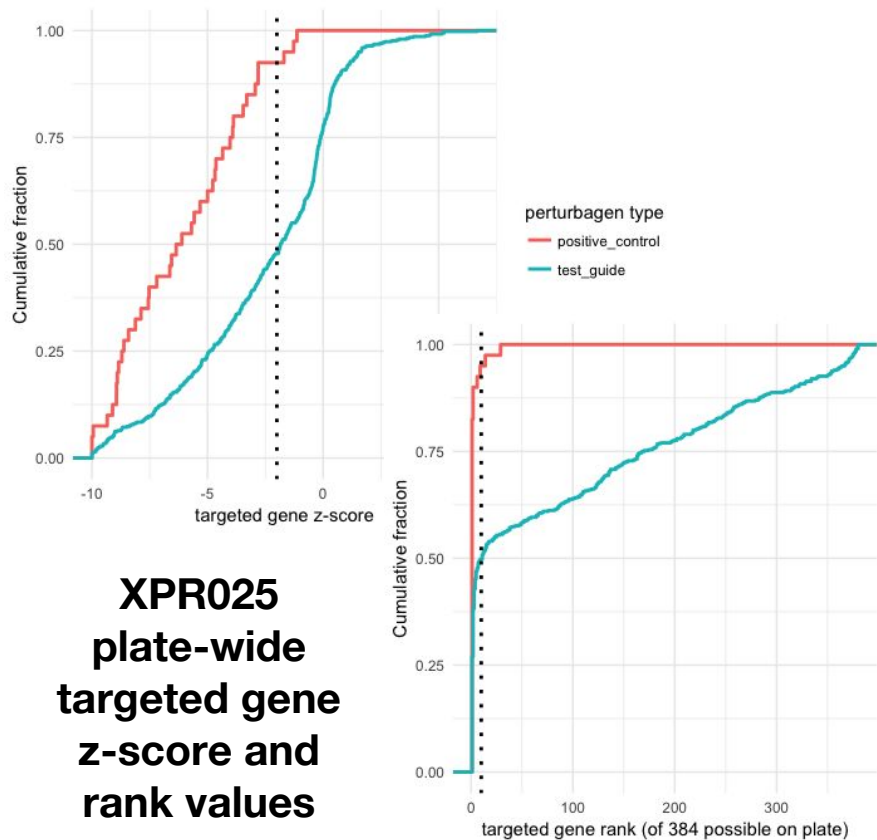
---

For an overview of general QC measures, please refer to our QC article [here](#); please also note that this article describes each of the measures listed below in more depth. Since these data correspond to a new perturbation type, QC was particularly stringent; instead of assessing QC by plate (as described in the link above), we looked at the technical measures below on an individual well basis and eliminated any signature in which any of the three replicates had any technical issues.

- High bead count: insufficient bead count results in unreliable expression measurement
- High calibration measures: high invariant 10 levels, high calibration curve slope & fit
- Correct gene measured: since we use 500 colors to detect about 1000 genes (more details [here](#)), it's important to assess that the gene measured is the one expected.
- High infection efficiency & virus titer

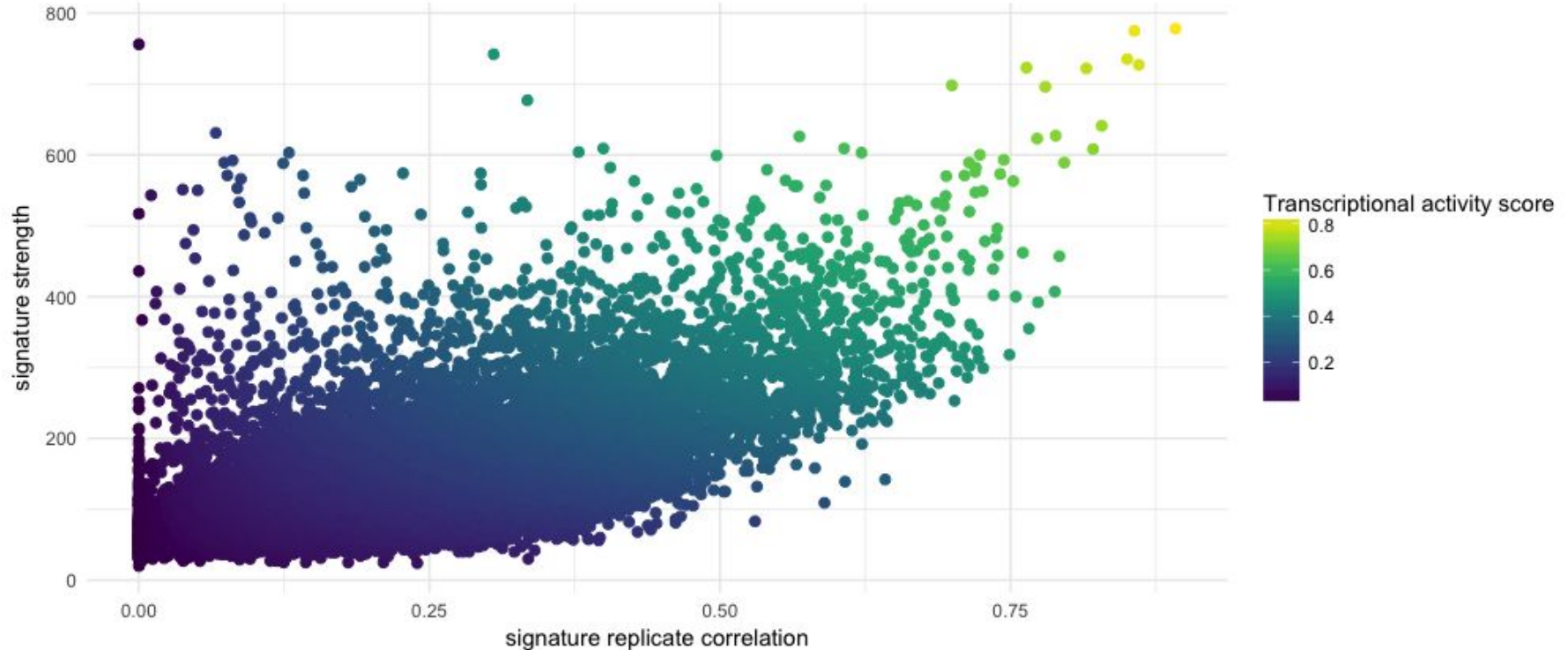
In the `siginfo.txt` and `instinfo.txt` metadata files made available with this dataset, there is a column called “QC\_pass” indicating whether the signature or replicate (respectively) passed these measures. Note that these are different from **functional** quality control measures, which involve a separate set of analyses.

# Example. Knockout results in decreased expression in one two guides targeting TLK2 in BICR6.311 (XPR025 plate)



**Example.** One of the guides targeting **TLK2** in BICR6.311 (**blue**) depletes expression of the gene; the z-score of TLK2 (blue dot) relative to its expression elsewhere on the plate (gray line) is low. The other guide targeting TLK2 (**red**) does not appear to have worked in this cell line.

# Signatures included have a range of correlation, strength, and TAS values



We commonly use signature strength and replicate reproducibility (combined into a transcriptional activity score) to assess signature quality. For more details on these measures please see [here](#).

Ongoing analysis at the Broad LINCS team:

# **Methods for identifying high-quality signatures**

Note. These are likely to be adjusted as our research studies develop further



# Ongoing analyses

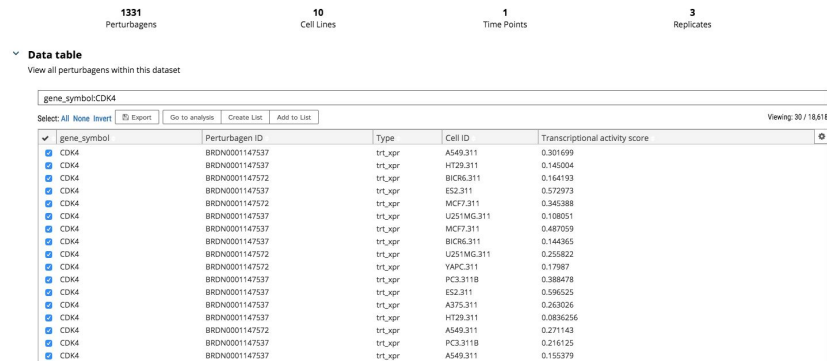
- Possible confounding factors
  - Presence/absence of spatial bias or technical effects
  - Effect of cas9 derivatization on gene expression signatures
  - Computational removal of generic responses to genetic perturbation
  - Presence/absence of [confounding cellular response due to gene copy number](#)
  - Evaluation of impact of off-target effects per guide
- Variation/calibration by cell line using expression patterns of [Hart core essential genes](#)
- Combining signatures across perturbagens (e.g., shRNA, sgRNA, cDNA) and other features to derive high-confidence gene knockout signatures

**Note.** Degree of impact of these factors is currently being assessed.  
Please contact [ana@broadinstitute.org](mailto:ana@broadinstitute.org) if you have specific inquiries or suggestions.

# Example. Connectivities of CRISPR-based signature targeting CDK4 enriched for CDK4 knockdown (via shRNA) and CDK inhibitor

Filter for guides targeting CDK4 from Data Table on clue.io Data Library page. Select all from Data table, and then click “View in ICV” to view connectivities.

## CRISPR\_Broad\_LINCS



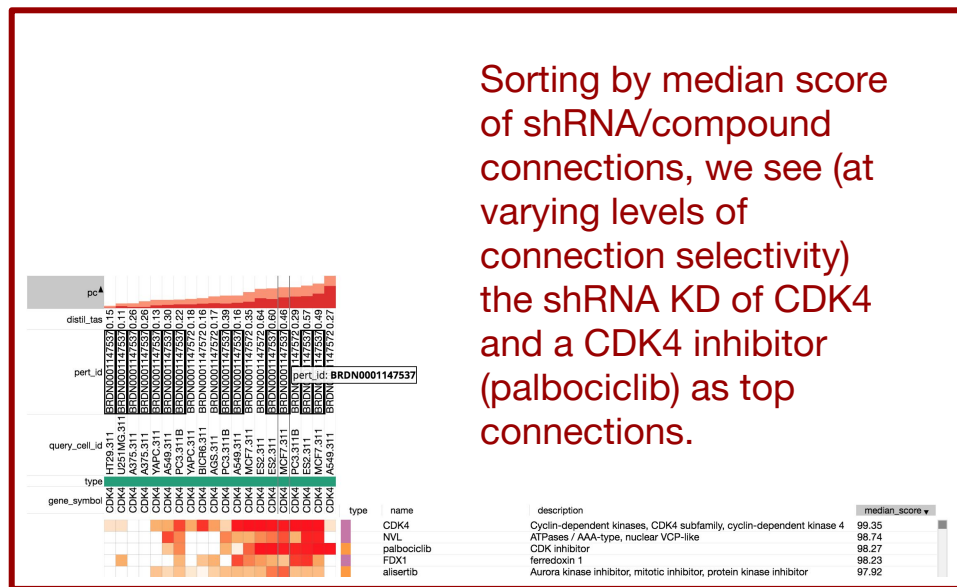
## Analysis in CLUE

### Touchstone connectivity

View connectivities between selected perturbagens and Touchstone perturbagens

Choose a subset: Selected from table (30)

VIEW IN ICV



Connectivities between 95 and 100 are coloured in orange -> red, where the darker the color the higher the connectivity score.

# **Additional resources**

# Contact information and useful resources

---

- Early release dataset & helpful applications
  - [available on clue.io](https://clue.io)
- Specific inquiries about this dataset
  - Email [oana@broadinstitute.org](mailto:oana@broadinstitute.org)
- General CMap analysis & data questions
  - Refer to [clue.io/connectopedia](https://clue.io/connectopedia)
- Open source code libraries: [clue.io/code](https://clue.io/code)
  - Available in Python, R, Matlab, Java
- In-person help (virtually or at the Broad Institute)
  - Attend our weekly [office hours](#)