Citation: CPT Pharmacometrics Syst. Pharmacol. (2016) 5, 588–598; doi:10.1002/psp4.12107

ORIGINAL ARTICLE

Systematic Quality Control Analysis of LINCS Data

L Cheng[1,2] and L Li[1,2]*

The Library of Integrated Cellular Signatures (LINCS) project provides comprehensive transcriptome profiling of human cell lines before and after chemical and genetic perturbations. Its L1000 platform utilizes 978 landmark genes to infer the transcript levels of 14,292 genes computationally. Here we conducted the L1000 data quality control analysis by using MCF7, PC3, and A375 cell lines as representative examples. Before perturbations, a promising 80% correlation in transcriptome was observed between L1000- and Affymetrix HU133A-platforms. After library-based shRNA perturbations, a moderate 30% of differentially expressed genes overlapped between any two selected controls viral vectors using the L1000 platform. The mitogen-activated protein kinase, vascular endothelial growth factor, and T-cell receptor pathways were identified as the most significantly shared pathways between chemical and genetic perturbations in cancer cells. In conclusion, L1000 platform is reliable in assessing transcriptome before perturbation. Its response to perturbagens needs to be interpreted with caution. A quality control analysis pipeline of L1000 is recommended before addressing biological questions.
CPT Pharmacometrics Syst. Pharmacol. (2016) **5**, 588–598; doi:10.1002/psp4.12107; published online 31 October 2016.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

☑ The Library of Integrated Cellular Signatures (LINCS) project provides comprehensive transcriptome profiling of human cell lines before and after chemical and genetic perturbations. Its L1000 platform utilizes 978 landmark genes to computationally infer to other 14,292 genes expression. However, there is no quality control data analysis on the reproducibility of this L100 gene expression platform.

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ For the first time, this study conducted quality control analysis on the LINCS L1000 gene expression platform.

WHAT THIS STUDY ADDS TO OUR KNOWLEDGE

☑ It shows a promising 80% correlation in transcriptome between L1000 and Affymetrix HU133A for MCF7 breast cancer, A357 melanoma, and PC3 prostate cancer cells. L1000 reproducibility analyses show that a moderate 30% of differentially expressed genes overlapped between any two selected controls viral vectors in the genetic perturbation screening. The MAPK, VEGF, and T-cell receptor pathways are pointed out for the most significantly connected breast cancer cell in chemical and genetic perturbations. A quality control pipeline of L1000 data quality is recommended before addressing biological questions.

HOW MIGHT THIS CHANGE DRUG DISCOVERY, DEVELOPMENT, AND/OR THERAPEUTICS?

☑ LINCS data provide the ability to establish a system pharmacology view of drug effect on the transcriptome. This landmark database provides a molecular basis for further study of single drug and/or drug combinatory effect at the cell level. LINCS data will eventually lead to more rational drug selection for patients.

The combination of "omics" technologies and cell-based drug screening tools have enabled us to evaluate cellular responses to drug perturbations and explore drug targets and their mechanisms. Large-scale drug screens for anticancer projects, such as the National Cancer Institute 60 (NCI60) human tumor cell line panel,[1] Connectivity Map (CMAP),[2,3] Cancer Cell Line Encyclopedia (CCLE),[4] Genomics of Drug Sensitivity in Cancer (GDSC),[5] and the cancer therapeutics response portal (CTRP)[6] all used cancer cell baseline genomes and/or transcriptomes to predict drug cell responses. Cell-based pooled short hairpin RNA (shRNA) screening is another important strategy for systematically identifying essential genes, and, eventually, therapeutic drug targets. The Achilles Project and DPSC-Cancer shRNAs interference detection[7,8] provided genome-wide shRNA dropout signature profiles for identifying cell vulnerabilities associated with genetic alterations. These databases attempted to address various aspects of the associations between molecular profiles, genetic and chemical perturbations (perturbagens), and cell responses to the perturbagens.[9] However, using these data sources it is difficult to form an integrated picture between cancer cell molecular profiles and their responses to perturbagens, because these data were generated through different experimental platforms.[11] By combining the chemical compounds and RNAis, the Library of Integrated Network-based Cellular Signatures (LINCS) project uses a novel transcriptome platform, L1000, to assess cell responses to perturbagens (lincs.hms.harvard.ed).[9,12] It allows us to integrate transcriptomes, perturbagens, and cell responses to drugs at the same time.[12–14]

[1]Centers for Computational Biology and Bioinformatics, School of Medicine, Indiana University, Indianapolis, Indiana, USA; [2]Department of Medical and Molecular Genetics, School of Medicine, Indiana University, Indianapolis, Indiana, USA. *Correspondence to: L Li (lali@iu.edu)
Received 28 June 2016; accepted 27 July 2016; published online on 31 October 2016. doi:10.1002/psp4.12107

Table 1 The samples number for L1000 data quality control and platform comparison between the Affymetrix HT-HG-U133A and L1000 in 22268 probe sets

				Cell lines		
Platform	**Timepoints**	**Experiment**	**Types**	**MCF7**	**PC3**	**A375**
L1000	H6	Compounds	#Gene expression profiles	43862	39605	27428
			#Chemical compounds	5434	4737	3083
	H24	Compounds	#Gene expression profiles	57475	57380	28601
			#Chemical compounds	6976	5845	2169
	H96	Knockdown	#Gene expression profiles	36023	41414	40640
			#Genes	3472	3824	3827
		Overexpression	#Gene expression profiles	9220	10271	10109
			#Genes	2160	2281	2281
	H144	Knockdown	#Gene expression profiles	20204	20414	/
			#Genes	1838	1726	/
L1000		Base Line	#Gene expression profiles	2922	27	24
Affymetrix HT-HG-U133A		Base Line	#Gene expression profiles	56	8	16

A cost-effective bead-based assay, the L1000 platform, is the LINCS primary technology that measures transcriptomic responses to perturbagens.[10] The L1000 platform directly measures transcripts of 978 landmark genes, from which the transcript levels of 14,292 genes are computationally inferred according to Gene Expression Omnibus (GEO) genes expression.[10,12] The relatively low cost of the L1000 allows the LINCS project to assess transcriptomic responses to 20,413 small molecule compounds and 22,119 genetic interference perturbagens, under more than four million different conditions (100-fold larger than other existing screening studies) (support.lincscloud.org). However, the enormous impact of the L1000 technology on chemical and genetic perturbagen screening depends heavily on its data quality.

The existing quality control analysis of the L1000 platform[10] focused only on 90 differentially expressed "landmark" genes in a validation experiment. It showed a high correlation for the 90 landmark genes, 0.92, between their gene expressions in the L1000 platform and their reverse-transcription polymerase chain reaction (RT-PCR) validations. However, the quality control analysis did not include the 14,292 L1000-inferred genes. Hence, a valuable and ideal evaluation would be a whole-genome gene expression comparison between the L1000 and other established whole-genome microarray platforms. Many cell-based molecular profiling datasets in public domain databases (e.g., the GEO), allows us to compare L1000 data quality to the other gene expression platforms. Similar to all inhibitory RNA (shRNA) libraries, L1000 RNA interference studies are complicated by the sequence design of perturbation shRNAs, as they affect silencing efficiency.[15,31] Thus, the effectiveness of genetic perturbagens on non-landmark genes should be further investigated too. In L1000, a number of perturbagens and a number of experimental conditions are used to explore perturbagen-induced change in transcriptome. It includes different cellular backgrounds, multiple chemical dosing concentrations, multiple timepoints,[12] as well as different empty control vectors (i.e., GFP, RFP, Luciferase, lacZ, and PGW) for single-gene knockdown or overexpression. These experiment conditions provide us an opportunity to broadly explore perturbagen effects on cells. In this study, we are particularly interested in whether the selection of controls will influence the shRNAs effect and gene overexpression effect on the cells and their transcriptome response.

To address the challenges in LINCS data quality control, more than 6,975 chemicals and 3,827 single-gene knockdown, and 2,281 single-gene overexpression on three cancer cell lines, MCF7 (an estrogen receptor-positive luminal breast cancer cell[16]), A375 (human skin cell with malignant melanoma[17]), and PC3 (prostate metastatic cell), were investigated. L1000 data were analyzed from four different aspects for the first time: 1) cancer cell baseline transciptome (i.e., untreated) in L1000 platform was compared to their corresponding transcriptome in the Affymetrix HU133A platform; 2) transcriptomes were compared between chemical treated groups and controls at multiple timepoints for three cell lines; 3) RNAi experimental variation and its sensitivity to different control groups were explored; and 4) connectivity between genetic and chemical perturbations was investigated. Finally, a guidance on how to use the L1000 dataset is recommended.

METHODS

Materials general

In this study, level 3 data of the normalized profiles are used for the quality control data analysis. These data are described in great detail in the supplemental materials. L1000 adopts the practice of storing data annotations (metadata) and datasets separately.[12,13] The InstInfo file describes L1000 signature profiles under different experimental conditions (**Supplementary Figure 1** visualizes the experimental design). Each expression profile is assigned with a unique identifier, i.e., signature ID (or "distil_id"), and it connects the data with its metadata in InstInfo. **Table 1** lists the samples for L1000 data quality control analysis and platform comparison with the Affymetrix HU133A. **Supplementary Table 2** shows the gene expression profiles and chemical numbers before and after compound treatments in 6 hours (H6) and 24 hours (H24) in three cell lines. **Supplementary Table 4** lists these records and genes numbers for three types of cells at 96 hours (H96)

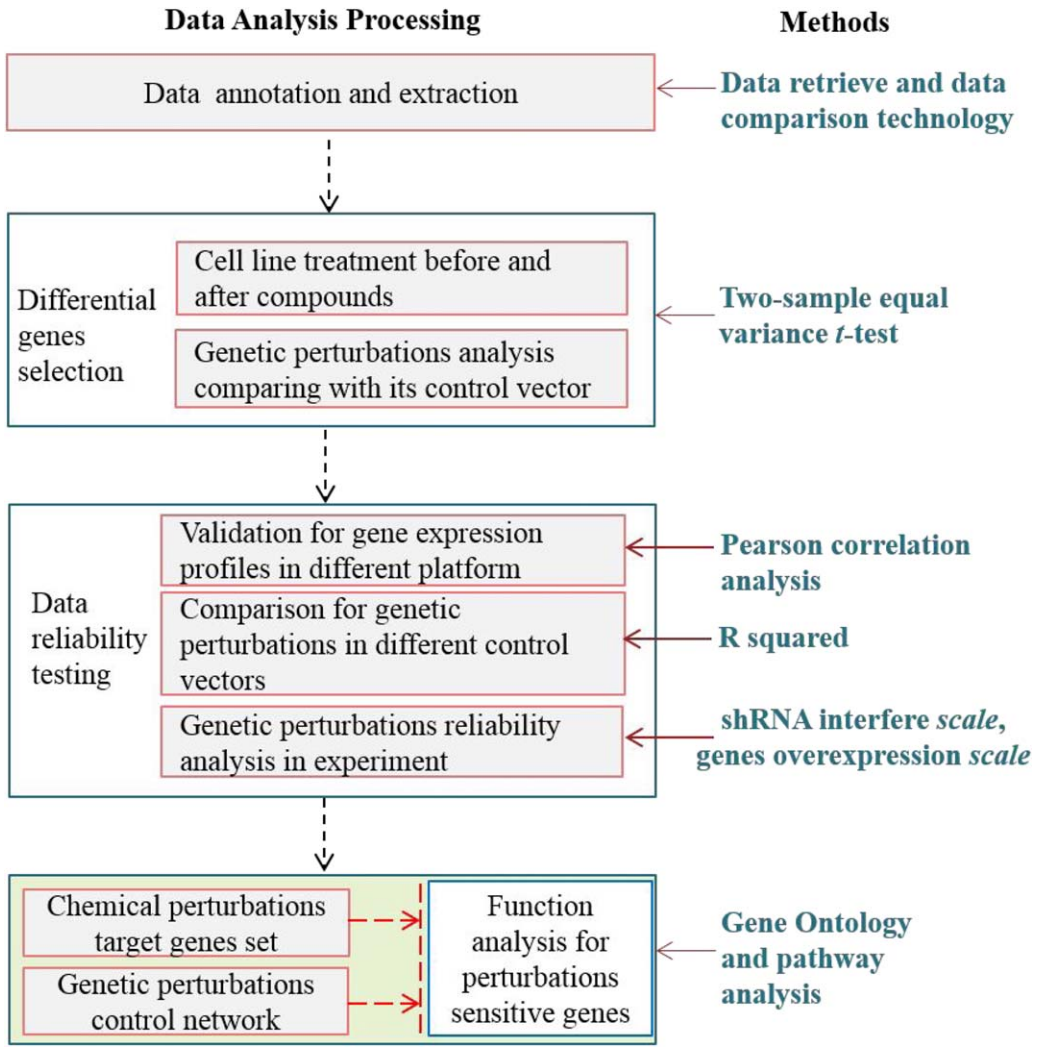

Figure 1 The overall L1000 quality control data analysis procedure. The figure shows the data analysis process and its associated methods in the study. The data analysis procedure can be divided into four steps. First, L1000 data is reannotated, retrieved, and extracted. Second, analysis of a differentially expressed gene (DEG) is performed before vs. after perturbation (chemical and genetic both type) through two-sample equal variance Student's *t*-test. Third, L1000 data quality is analyzed. It includes the mRNA correlation analysis between different platforms, the genetic perturbation comparisons using different control vectors through R-square; shRNA interfere scale and gene overexpression scale are calculated to evaluate the reliability of genetic perturbation experiments. Fourth, connectivity analysis is performed between chemical and genetic perturbations by GO enrichment and KEGG pathway analysis.

and 144 hours (H144) for knockdown and overexpression experiments against different control vectors.

Methods

The flow of data processing and analysis is shown in **Figure 1**.

Differentially expressed gene analysis

An unpaired two-tailed Student's *t*-test was used to evaluate the gene expression difference in data from two different groups, including the following: before and after chemical treatments (the gene profiles of chemical treatment vs. those incubated with dimethyl sulfoxide (DMSO) control) in MCF7, PC3, and A375 cells at 6 hours and 24 hours, respectively; before and after shRNAs/overexpression perturbations (the gene knockdown/overexpression group vs. its different control groups, such as empty vector GFP,[18] eGFP,[19] Luciferase, HcRed,[20] or lacZ,[21] respectively). The difference was considered significant if $P < 0.01$.

Data batch effects due to the plate are removed by the quartile normalization. In the LINCS L1000 experimental design, each plate will have its own control samples (i.e., DMSO) and perturbagen-treated samples. Our two-sample *t*-tests use perturbagen-treated and control samples from the same plate to analyze differentially expressed genes (DEGs). In order to let the statistical *t*-test be less sensitive to outliers or small variance, a minimum sample size of three was required for each group.

shRNAs interference and gene overexpression scale calculation

shRNA knockdown *scale* quantifies the gene knockdown accuracy. At first, a gene expression is normalized by the

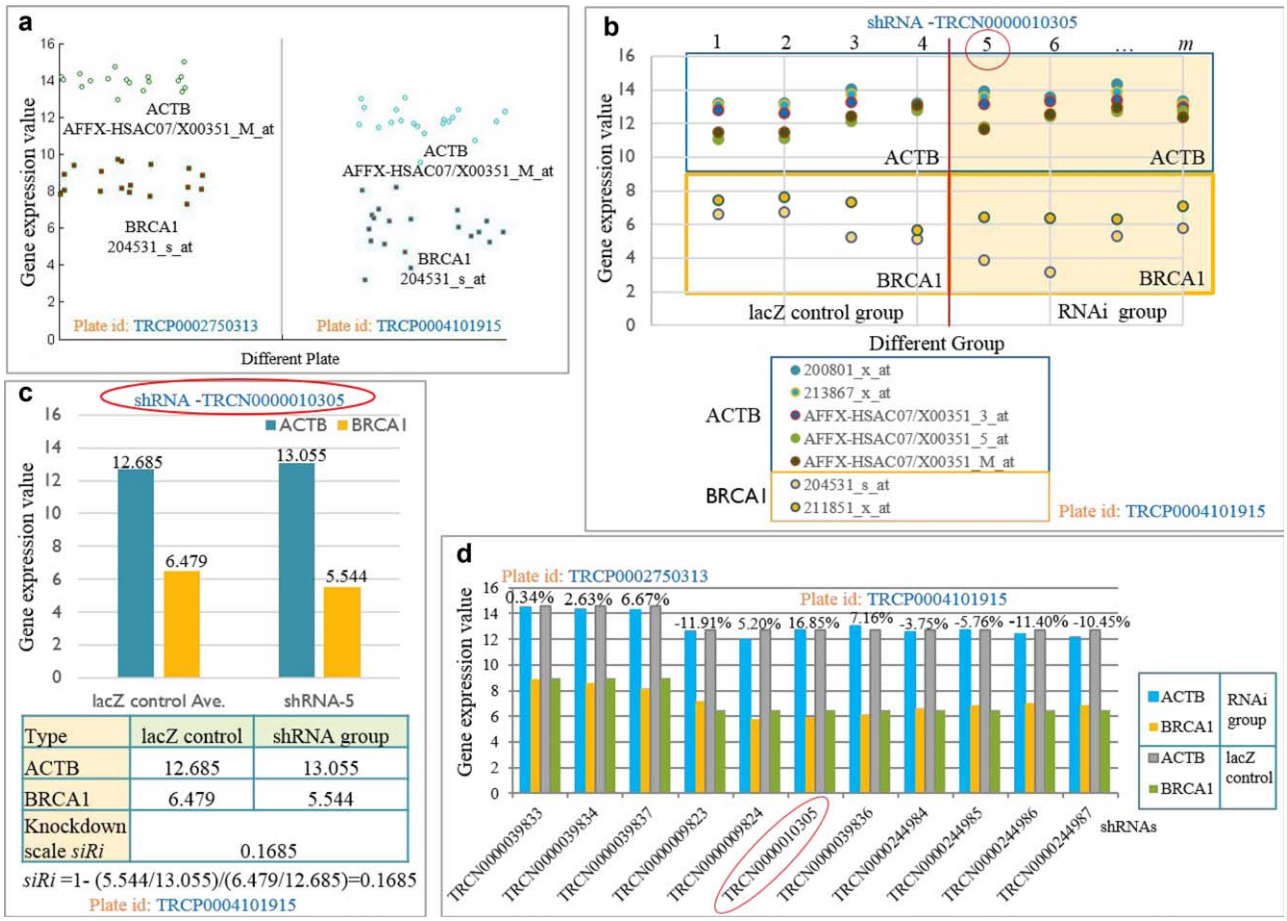

Figure 2 The BRCA1 knockdown scale calculation scheme under the lacZ control vector. (**a**) The scatterplot of two gene expressions (BRCA1 and housekeeping gene ACTB) under various conditions in two separate plates. (**b**) The same two gene expressions organized by different conditions, such as LacZ control, gene probes, and shRNAs targeting BRAC1. x-axis denotes different shRNAs, and y-axis denotes the gene expression value. (**c**) BRCA1 relative interfere scale calculation to control lacZ for the shRNA TRCN0000010305. First, the average expression of these probe sets for ACTB and BRCA1 was used to calculate siRi. Then the BRCA1 shRNA interference scale, with respect to ACTB, was siRi = 1-(Gene_exp/Gene_ctr) = 1-(5.544/13.055)/(6.479/12.685) = 0.1685. The shRNA TRCN0000010305 (the fifth shRNA targeting BRCA1) knocks down BRCA1 gene expression by 16.85% against its control vector lacZ. (**d**) BRCA1 shRNA interference scale relative to ACTB among different shRNAs. The x-axis is the shRNA clone, and the y-axis is the gene expression value. The shRNAs interference scales of 11 shRNA clones for BRCA1 are shown at 96 hours in MCF7.

housekeeping gene expression. Then, the knockdown *scale* quantifies an interfered gene expression change related to its uninterferred gene (i.e., control) in a cell. Denoting the control group as *ctr*, and the shRNA group as *exp*, the knockdown *scale* calculation is:

Gene_ctr = shRNA gene expression in control group/housekeeping gene expression in control group;

Gene_exp = shRNA gene expression in shRNA experimental group/housekeeping gene expression in the shRNA experimental group;

shRNA Knockdown *Scale*: *siRi* = 1- (*Gene_exp*/*Gene_ctr*);

In the article, if *siRi* > *0*, the experiment is defined as a success; otherwise it is a failure.

To calculate the gene overexpression *scale*, the formula is similar to the shRNAs interference *scale*:

Gene_ctr = Gene overexpression in control group/housekeeping gene expression in control group

Gene_exp = Gene overexpression in overexpression experiment group/housekeeping gene expression in overexpression experiment group

Gene Overexpression *scale*: *oeRi* = (*Gene_exp*/*Gene_ctr*)−1

When *oeRi* > *0*, the experiment is deemed successful; otherwise the experiment is deemed failure.

According to the housekeeping gene list in the references,[22,23] 13 housekeeping genes are selected as control genes. These genes have corresponding probe sets in Affymetrix HU133A Chip: ACTB (β-actin), CHMP2A, EEF1A1, EMC7, GAPDH, GPI, PSMB2, PSMB4, RAB7A, REEP5, SNRPD3, TUBA1A (α/β-tubulin), and VCP. **Figure 2** describes an example for shRNA knockdown *scale* calculation for gene BRCA1.

"Pseudo" R-squared calculation

Assuming x' is a variable, and y' is another control variable. x' and y' is normalized by $x=log_{10}(0.5+x')$ and

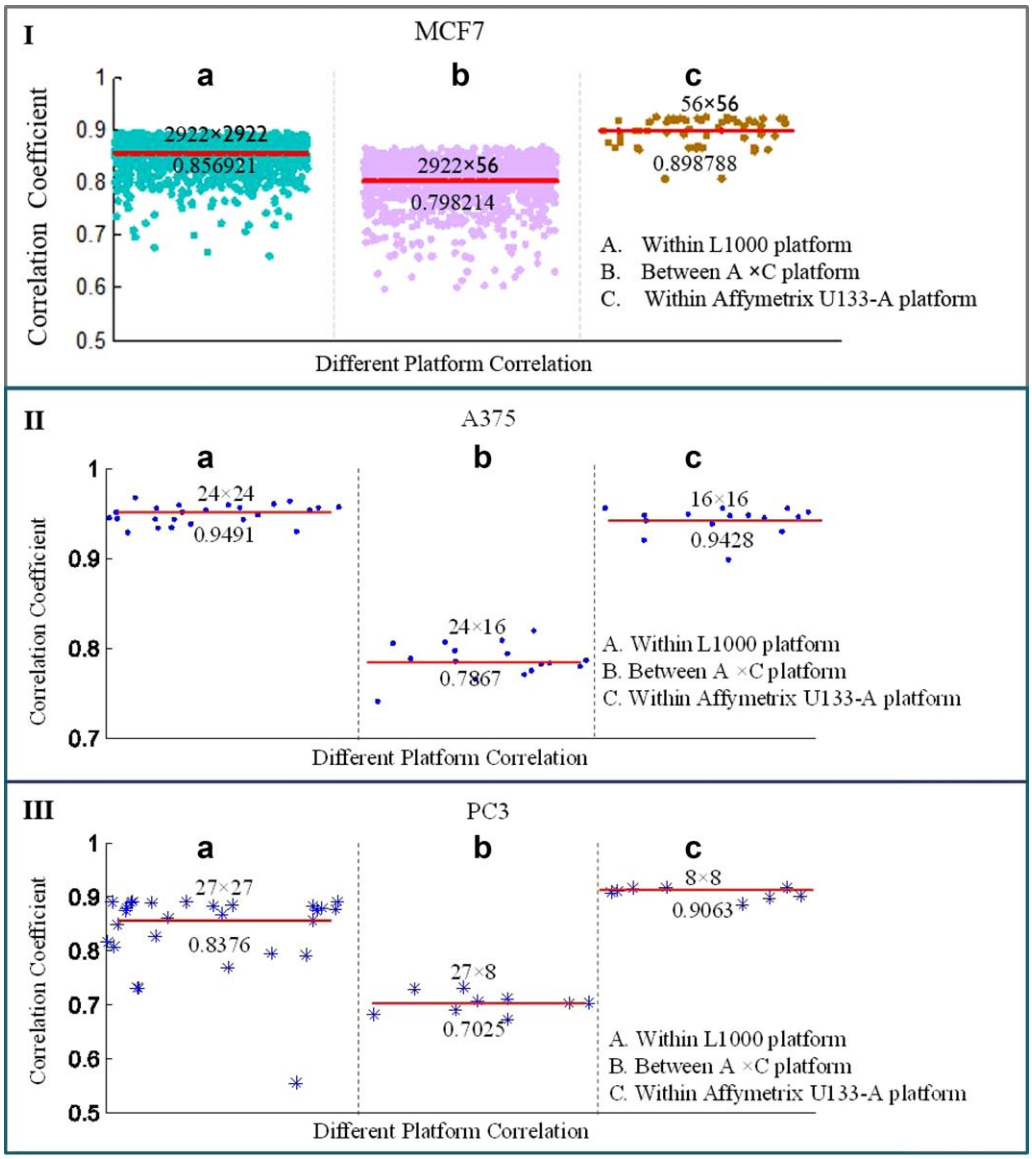

Figure 3 Transcriptome Pearson correlation analysis of within- and between-platforms: L1000 and Affymetrix HU133A. Baseline transcriptomes of L1000 and Affymetrix HU133A are compared for MCF7, A735, and PC3 cells, in I, II, and III, respectively. Subfigures titled in **"a"** present the within L1000 platform correlations; **"b"** shows cross-platform correlations; and **"c"** shows within-Affymetrix platform correlation.

$y = log_{10}(0.5 + y')$, respectively. According to the following formula to calculate R^2:

$$R^2 = 1 - \frac{SS(x-y)}{(SS(x)+SS(y))/2}$$

where $SS(x) = \sum_{i=1}^{n} x_i^2$, $SS(y) = \sum_{i=1}^{n} y_i^2$, $SS(x-y) = \sum_{i=1}^{n} (x_i - y_i)^2$, is the sum square of error that x is not explained by y, and vice versa. $(SS(x)+SS(y))/2$ is the average sum of square between x and y. Their ratio represents the percent of variation in x or y that are not explained by each other. Our defined R-square, which is one minus this ratio, represents the variations in x or y that can be explained by each other.

Software

MatLab (MathWorks, Natick, MA) was used to analyze the data and advanced graphics and visualization. The parse_gctx function was used to read .gct and .gctx files format and extract data according to its annotation in MatLab; for all the code source, see **Supplementary Code**.

Database for Annotation, Visualization, and Integrated Discovery (DAVID, https://david-d.ncifcrf.gov/) is an online analysis resource. It provides a comprehensive set of functional annotation tools for researchers to identify biological function by gene lists.[32] Gene Ontology (Go) function and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses are conducted on the selected differential genes. These analyses are performed in DAVID v. 6.8.

Table 2 The average interference and overexpression ratio in shRNAs knockdown and overexpression experiment by 13 housekeeping genes as the control reference respectively in MCF7 96 hours (H96) and 144 hours (H144) in L1000

Group	Control vector	Housekeeping genes													
		ACTB	CHMP2A	EEF1A1	EMC7	GAPDH	GPI	PSMB2	PSMB4	RAB7A	REEP5	SNRPD3	TUBA1A	VCP	Average
H96 shRNA	lacZ	0.5901	0.6086	0.6027	0.6216	0.5892	0.5896	0.5976	0.5894	0.5908	0.5867	0.5597	0.5552	0.5758	0.5890
	Luciferase	0.6598	0.6639	0.6588	0.6622	0.6603	0.6608	0.6674	0.6743	0.7092	0.6794	0.6863	0.6995	0.6719	0.6734
	GFP	0.5886	0.5901	0.6097	0.6152	0.6124	0.6101	0.6052	0.6078	0.6175	0.6019	0.6052	0.5741	0.6312	0.6053
	RFP	0.6500	0.6610	0.6574	0.6500	0.6544	0.6540	0.6517	0.6576	0.6952	0.6852	0.6900	0.7017	0.6756	0.6680
	pgw	0.6272	0.6026	0.6542	0.6481	0.6698	0.6659	0.6970	0.6686	0.7201	0.7034	0.7110	0.7105	0.7027	0.6755
	Average	0.6231	0.6252	0.6366	0.6394	0.6372	0.6361	0.6438	0.6396	0.6666	0.6513	0.6504	0.6482	0.6514	0.6422
H144 shRNA	lacZ	0.6912	0.7021	0.6943	0.7145	0.6916	0.6901	0.7048	0.6843	0.7162	0.7063	0.6894	0.6952	0.7057	0.7001
	Luciferase	0.7015	0.7020	0.6940	0.6960	0.6929	0.6894	0.6921	0.6976	0.7227	0.7001	0.7077	0.7104	0.7059	0.7018
	GFP	0.6863	0.6885	0.6914	0.7002	0.6867	0.6883	0.6813	0.6812	0.6868	0.6868	0.6887	0.6855	0.7114	0.6900
	RFP	0.6877	0.6918	0.6977	0.6945	0.6906	0.6882	0.6940	0.6829	0.7130	0.7175	0.7248	0.7348	0.7168	0.7043
	pgw	0.6756	0.6522	0.6891	0.6952	0.6907	0.6864	0.7272	0.6879	0.7468	0.7370	0.7353	0.7399	0.7408	0.7091
	Average	0.6885	0.6873	0.6933	0.7001	0.6905	0.6885	0.6999	0.6868	0.7171	0.7095	0.7092	0.7132	0.7161	0.7011
H96 Overexpression	lacZ	0.4958	0.4925	0.5247	0.5167	0.5336	0.5362	0.5298	0.5338	0.5436	0.5161	0.5401	0.5320	0.5318	0.5271
	Luciferase	0.4978	0.4715	0.5219	0.5299	0.5586	0.5505	0.5487	0.5550	0.5432	0.5673	0.5265	0.5167	0.5195	0.5375
	HcRed	0.5176	0.5060	0.5270	0.5308	0.5446	0.5436	0.5429	0.5414	0.5272	0.5429	0.5457	0.5338	0.5586	0.5392
	eGFP	0.4978	0.4715	0.5219	0.5299	0.5586	0.5505	0.5487	0.5550	0.5432	0.5673	0.5265	0.5167	0.5195	0.5375
	Average	0.5023	0.4854	0.5239	0.5268	0.5489	0.5452	0.5425	0.5463	0.5393	0.5484	0.5347	0.5248	0.5324	0.5353

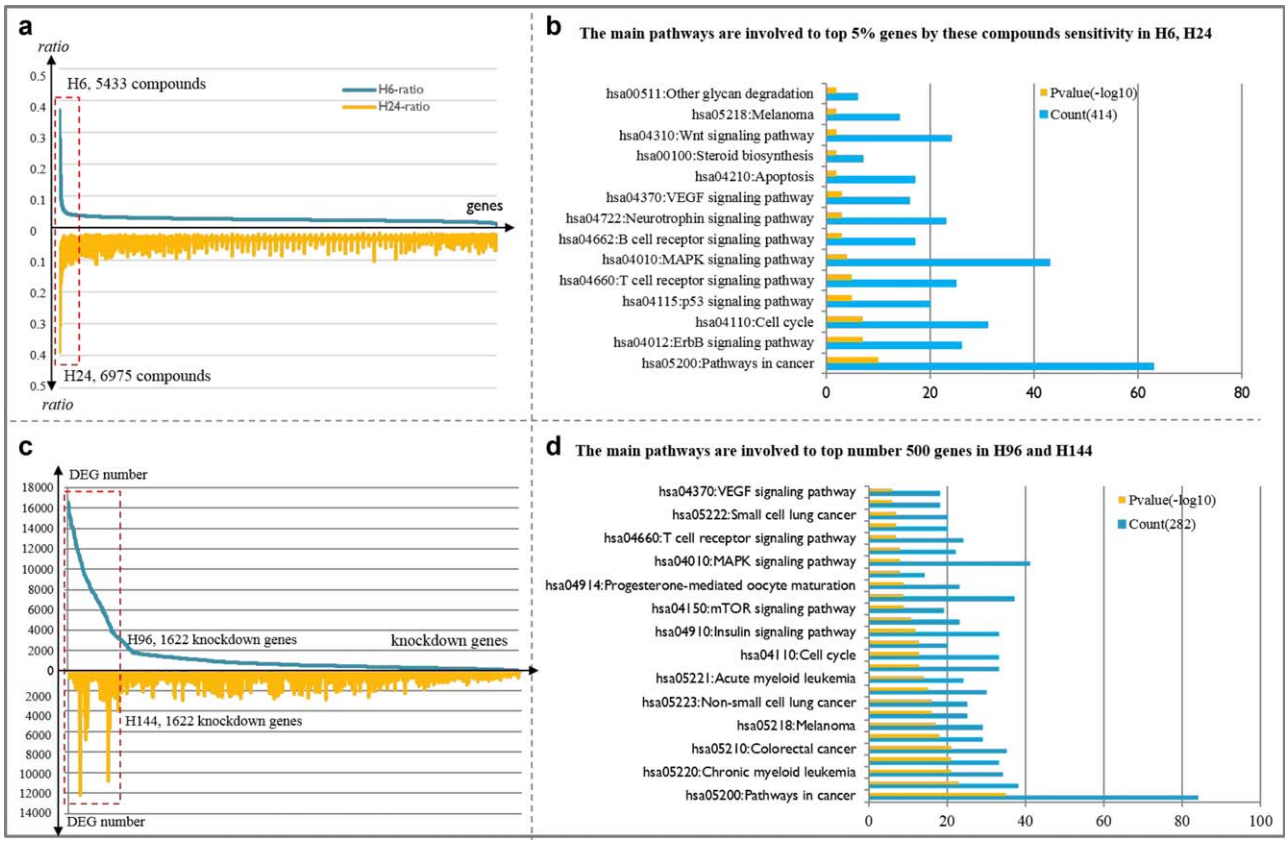

Figure 4 The genome analysis to chemical sensitivity and genetic perturbation sensitivity. (**a**) The chemical sensitivity to each of the genes. x-axis denotes 14,292 genes, while y-axis is the ratio of chemicals that perturbate the gene expressions number over the total number of chemicals. The upper panel of (**a**) ranks genes from the highest chemical sensitivity (left) to lowest sensitivity (right) after 6-hour chemical treatment of MCF7 cells (H6); and the lower panel of (**a**) displays the corresponding chemical sensitivity of at 24 hours (H24). (**b**) The pathway enrichment analysis for the top 5% genes with highest chemical sensitivity (a: red box with dot-line), using DAVID. (**c**) Genetic perturbation sensitivity to each of the genes in H96 and H144 for MCF7. The number of differentially expressed genes (DEGs) measures the impact of a genetic perturbation (y-axis). x-axis represents the common 1,622 knockdown genes between H96 and H144. The upper panel of (**c**) ranks gene knockdown from the highest impact (left) to lowest impact (right) at 96 hours (H96), while the lower panel shows the corresponding gene knockdown impact at 144 hours. (**d**) The pathway enrichment analysis for the top 500 genes (**c**) red box with dot-line).

RESULTS

Highly correlated transcriptomes between the L1000 and the Affymetrix HU133A platforms

The baseline samples for platform comparison between the Affymetrix HU133A and L1000 are described in **Table 1.** The detailed GEO datasets for Affymetrix HU133A platform are provided in **Supplementary Table 1**. A Pearson correlation is calculated for 22,268 probes between an L1000 sample or an Affymetrix sample. **Figure 3** depicts the transcriptome profile correlation analysis. It reveals that the L1000 platform shows the mean of within-platform correlation is 0.86 among 2,922 MCF7 samples (**Figure 3l.a**); the mean of cross-platform correlation between the L1000 and Affymetrix HU133A is 0.80 (**Figure 3l.b**); and the mean of within-platform correlation for the Affymetrix HU133A is 0.90 (**Figure 3l.c**). These correlations demonstrate the reproducibility of the L1000 related to the Affymetrix platform. Considering that the L1000 transcriptome is predicted from 978 landmark genes and the mean within-L1000 platform correlation is 0.86, the between-platform correlation of 0.80 clearly shows strong correlation between the two platforms. Similar results were observed in both PC3 and A375 cells (**Figure 3ll–3lll)**.

High concordance of differentially expressed gene numbers, but with moderate overlaps, in genetic perturbations using different controls

We examined five lentivirus control constructs (GFP, eGFP, Luciferase, RFP, HcRed and lacZ) to evaluate the reproducibility among them using L1000 in MCF7, PC3, and A375 cells. A transcriptome signature of a genetic shRNA perturbagen is composed of DEGs calculated from two-sample *t*-tests ($P < 0.01$) against each of the control vectors. Each shRNA has its own set of perturbated DEGs. Among shRNAs targeting the same gene, their perturbated DEG sets are merged together. This merged DEG set represents the targeted gene's shRNA overall impact on the transcriptome. In our analysis of lentivirus control constructs, each selected control generates a collection of shRNA perturbated gene sets. In order to compare two lentivirus controls, their

corresponding shRNA perturbated gene sets were first compared using the R-square method. This R-square calculates the relative amount of variance of shRNA perturbated gene numbers under one control against that under another control. An R-square close to one indicates a strong concordance between two different controls. Pairwise analysis shows that the concordance for shRNA-induced DEGs among five controls is high after either 96- or 144-hour infection times, where the R-square ranges from 0.94 to 0.98 (**Supplementary Figure 2**). Similarly, the concordance for the overexpressed DEGs among different controls is even higher, ranging from 0.97 to 0.99.

Between two control groups, the genome-wide DEG overlapping analysis focuses on not only statistically significant DEGs ($P < 0.01$), but also DEGs sharing the same direction of effect (i.e., upregulation or downregulation). The overlapping DEG scale is defined as the ratio of the common DEGs between two controls over the union of DEGs from two controls. The overlapping DEG scale is then calculated between any pair of control vectors. **Supplementary Table 4** shows average overlap scales for shRNA knockdown among any control pairs. The average is 0.36 at 96 hours and 0.26 at 144 hours, and the average overlapping scale of cDNA overexpression is 0.36 at 96 hours in the MCF7 cell. Moderately overlapped DEGs among different controls indicate potential differences between control vectors.

The whole-genome scale shRNA DEG numbers change with the time of cell after infection. **Supplementary Figure 4** compares the DEG numbers between 96- and 144-hour infection under each of the five controls in MCF7, and their average effects. We observe that shRNA infection leads to more downstream signaling changes (i.e., more DEGs) in 96 hours than 144 hours. This trend is consistent among five control groups (**Figure 3**). In addition, we observe that the DEG number in the lacZ control vector group is always less than those using the other controls, while DEGs in the PGW and Luciferase controls are always higher than those of the other groups. The pattern of DEGs is generally in agreement between 96 hours and 144 hours.

In particular, some genes, when knocking down and affecting a large number of other genes under one control, they also show the same impact under another control or in a differential timepoint. For instance, in **Supplementary Figure 3G,H**, knocking down ABL1 leads to 15,000 DEGs at 96 hours and around 5,000 DEGs at 144 hours. **Supplementary Figure 4** shows the top perturbagen sensitivity genes for cell MCF7 in 6 hours, 24 hours, and 144 hours.

Housekeeping gene selection does not affect the calculation of the effectiveness of genetic perturbations

The effectiveness of genetic perturbations is usually compared to the levels of shRNA knockdown target or overexpressed genes, relative to the expression of stable housekeeping genes. The shRNA interference scale approximates the ratio of gene expression before and after shRNA interference, normalized to the ratio of housekeeping gene expression before and after treatment. An shRNA interference scale larger than 0 means a successful knockdown experiment. A gene overexpression scale is defined similarly as a shRNA interference scale. To investigate the sensitivity of the scale calculation, 13 housekeeping genes were selected. **Table 2** shows shRNA interference and overexpression scales after 96 hours and 144 hours MCF7 cell infection. This suggests that there is very small variation in shRNA interference scales and overexpression scales among 13 housekeeping genes. The average shRNA interference scale is 0.64 at 96 hours (H96) and 0.70 at 144 hours (H144). The average gene overexpression scale is 0.53 for 96 hours (H96). In addition, these scales are consistent when different controls are used. The transfection efficiency and the shRNA knockdown scale of H144 are higher than that of H96 by one-sided Student's *t*-test for scales ($P < 0.05$). This suggests that a longer transfection time enhances transfection efficiency. In addition, according to the housekeeping gene recommendation,[23] a suitable housekeeping gene should show minimal variability under various experimental conditions. In LINCS, all 13 housekeeping genes generate fairly consistent inference scales. Thus, genetic perturbation quantification is not sensitive to the housekeeping gene selection (**Table 2**).

Connectivity analysis between MC7 cell chemical perturbation-sensitive and genetic perturbation-sensitive genes

Transcriptomic responses to perturbations can identify genetic biomarkers and their pharmacological mechanisms of chemical compounds in killing cancer cells, and detect essential genes and their signaling pathways associated with genetic perturbations. The impact of a chemical perturbation to a cell can be illustrated by a list of DEGs before and after perturbation. Large-scale chemical perturbation is helpful to understand genes that are sensitive to various perturbations. The gene sensitivity is calculated with its DEG frequency among all chemical perturbations ($P < 0.01$), i.e., relative DEG frequency (*y*-axis in **Figure 4a**). For instance, there were 5,433 chemicals tested at 6 hours to MCF7, 2,188 chemicals caused ATFI gene expression change. Hence, ATFI sensitive ratio to chemicals is $2{,}118/5{,}433 = 0.39$ at 6 hours. Similarly, the ATFI sensitive ratio is $2{,}580/6{,}975 = 0.37$ at 24 hours. **Figure 4a** shows gene sensitivity to all chemicals at 6 hours (*H6*) and 24 hours (H24), in which *x*-axis denotes 14,292 genes ranked by their sensitivity ratios (*y*-axis). The upper *y* axis is for timepoint 6 hours data, while the lower *y* axis is 24 hours. The Pearson correlation between H6 and H24 gene sensitivities is 0.712 ($P < 0.01$). The overall trend of the ratios is consistent between 6 hours and 24 hours. The red dotted box of **Figure 4a** displays the high sensitivity for the top 5% genes across all the chemicals in both H6 and H24. The top 24 sensitive genes in both H6 and H24 in MCF7 are displayed in **Supplementary Figure 4**; and the top 50 sensitive genes after 6- and 24-hour treatments are ranked in **Supplementary Table 5**. **Figure 4b** provides the pathway analysis for the top 5% sensitive genes (1,113 genes) by the DAVID tool (https://david.ncifcrf.gov/). The similarity of shRNA interfered DEGs is observed in MCF7 cells between 96 and 144 hours (**Figure 4c**). **Figure 4d** shows the

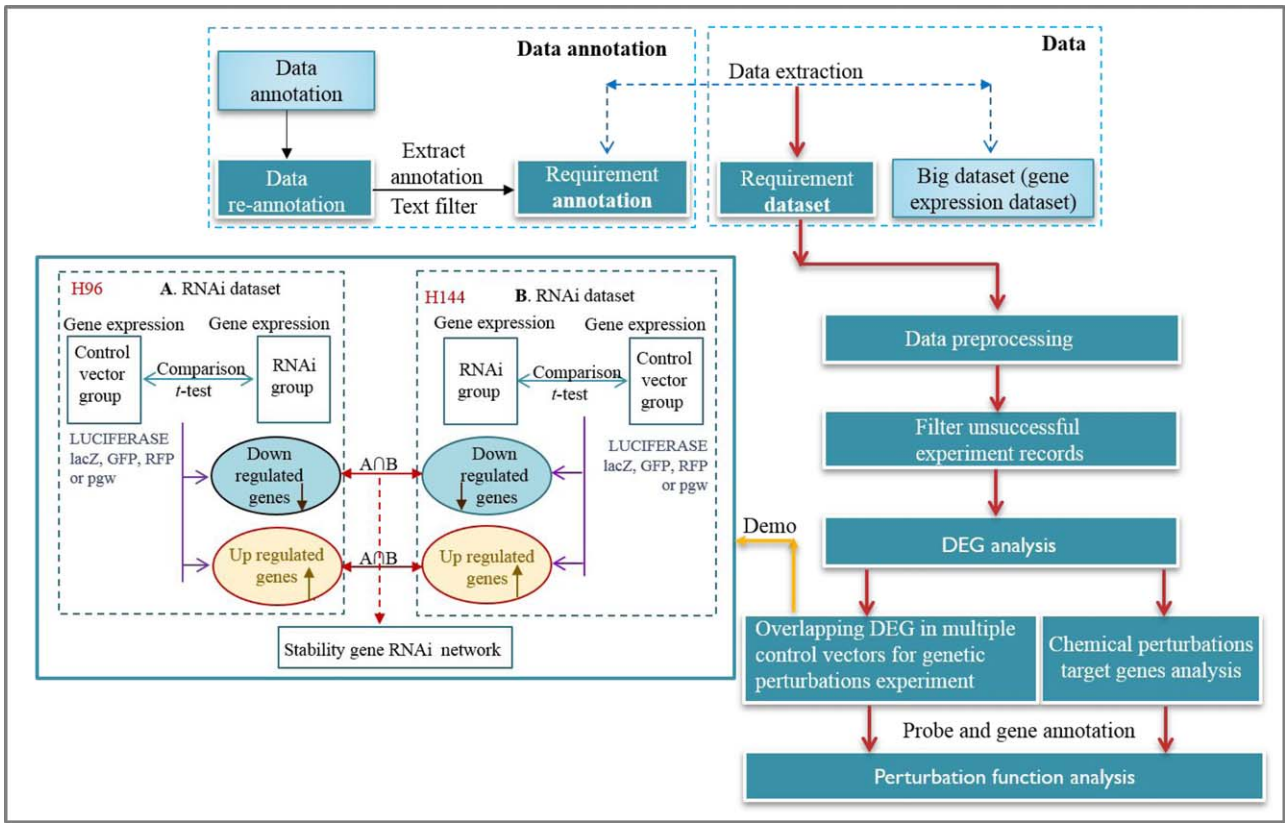

Figure 5 Recommended data processing schema for L1000 data analysis. L1000 data analysis starts from a data annotation retrieval and a data extraction. Then empty and batch effect is removed. Unsuccessful experiment records (i.e., failed shRNA knockdown or failed gene overexpression experiments) need to be filtered out before a DEG or other data analysis. Overlapped DEGs with respect to multiple controls are recommended for the genetic perturbation analysis.

pathway analysis for the top 500 interference genes by the DAVID tool. Comparing the sensitive gene pathways (**Figure 4b)** and shRNA interfere pathways (**Figure 4d**), several common pathways are identified between chemical and genetic perturbations. They include mitogen-activated protein kinase (MAPK), T-cell receptor, and vascular endothelial growth factor (VEGF). These shared cancer pathways reveal potential connections between chemical and genetic perturbations. Additional GeneOntology (GO) analyses further confirm the connectivity between chemical and genetic perturbations in **Supplementary Figures 5–7.**

DISCUSSION

The feasibility of using the L1000 platform for the LINCS project is due to its cost-effectiveness in measuring transcriptomic responses under millions of genetic and chemical perturbation conditions. Technically, L1000 measures only 978 landmark gene transcripts[10] upon which the expression of another 14,292 genes is computationally inferred.[9] Before using L1000 data to answer significant biological questions, it is critical to assess its quality. To that end, we compared L1000 and Affymetrix HU133A transcriptome measurements of MCF7, A375, and PC3 cancer cells. We found a mean within-platform transcriptome correlation of 0.86 for L1000 and 0.90 for Affymetrix, while the cross-platform transcriptome correlation was around 0.80. This demonstrates the remarkable reproducibility of L1000, considering that its transcriptome is predicted from only 978 landmark genes.

Control vectors play a critical role in gene silencing or overexpression experiments. Transcriptomic changes in cells treated with nonsilencing controls provide a baseline reference, which can guard against false positives in either molecular or cell responses of genetic perturbagens.[31] LINCS investigates a variety of "control hairpins" such as GFP, eGFP, Luciferase, RFP, HcRed, lacZ, and PGW.[24] We analyzed the number of DEGs from genetic perturbations using various control vector groups after cell lentivirus infection for 96 and 144 hours to PC3 and A375 cells. It demonstrated consistent numbers and patterns of DEGs. R-squares ranged from 0.95 to 0.98 for both shRNA and overexpression perturbations in MCF7. A more stringent analysis of overlapping DEGs (i.e., same identities, statistically different, and same directional change) shows less consistency between the five control vectors. The average DEG overlapping of any pair of control vectors is 0.38 and 0.26 in MCF7 cells before and after shRNA transduction for 96 and 144 hours, respectively; and 0.36 for 96 hours gene

overexpression experiments. Therefore, we feel the most reliable control selection is to use the overall overlapping DEGs among all the controls.

Through interference scale calculation, the average shRNAs interference scale is around 0.67 and the average gene's overexpression scale is 0.53, using the L1000 platform. These results illustrate the likely existence of a number of unsuccessful genetic perturbation experiments, due to off-target interference and small signal-to-noise scales. Consequently, we highly recommend removing these failed experimental records before the final analysis by its interference or overexpression scale calculation.

Both chemical and genetic perturbations influence cell viability. It is very interesting to investigate whether chemical and genetic perturbations stimulate similar signaling pathways. Through GO and KEGG pathway analyses of the two sets of gene signatures from chemical and genetic perturbations, a number of shared signaling pathways are identified, including MAPK, VEGF, and T-cell receptor pathways, all well-established signaling pathways in breast cancer.[25] Although the estrogen receptor alpha (ERα) signaling pathway is a primary carcinogenic mechanisms of MCF7 cells, it is not selected as a shared pathway between two perturbations. However, it is well established that ERα cross-talks with the MAPK cell proliferation pathway, as MAPK is a primary mediator of ERα activation.[26,27]

In order to improve the LINCS data quality, the following data processing and filtering steps are recommended, before it is used to answer biological questions.

- Data annotation and extraction: L1000 uses a specific data storage style separate from its annotation. Several data annotation files exist, but they cannot be fully integrated because of inconsistent perturbagen annotations. Hence, the first step is annotation before extracting the corresponding LINCS dataset from the depository.
- Data preprocessing: Relatively sparse (e.g., less than 5%) empty data points among microarray probes in samples can be imputed based on the average value of several close-by probes. Otherwise, remove these empty data points from the dataset.
- Some unsuccessful genetic perturbation records should be filtered out. Only keep the record where the shRNA interference scale or overexpression scale is larger than zero.
- Within-plate DEG comparison of control vector and experimental groups, to reduce the technical variation between different plates or batches.
- Overlap analysis among different controls: To reduce the impact of off-target prediction in shRNA or overexpression experiments, overlapping DEGs in multiple control vectors is necessary (see **Figure 5**). The overlapped genes should keep the same up- or down-gene regulations under different control vectors and different timepoints in genetic perturbations. In chemical perturbation-sensitive gene analyses, the chemical compound names, dose, and time should be consistently annotated before and after treatment.
- Genetic perturbation-targeted gene annotations: Before a genetic perturbation function analysis, a consistent mapping annotation is necessary between probes and genes in Affymetrix U133A platform and shRNA or overexpressed gene names, as well as their affected probes.

In summary, we conclude that despite the above-mentioned considerations needed for LINCS data analysis, L1000 is reliable in assessing transcriptomic responses to a myriad of perturbagens. Several promising antineoplastic compounds have already been identified by the LINCS program.[28–30] We strongly believe that the additional considerations we present here will even further expedite the identification of therapeutic chemical/genetic agents.

Acknowledgments. This work was supported by National Institutes of Health DK102694, GM10448301, and LM011945.

Conflict of Interest. The authors declare no conflicts of interest.

Author Contributions. LC and LL acquired the data and performed the necessary computational analyses. Both authors contributed to preparation of the article.

1. Abaan, O.D. *et al.* The exomes of the NCI-60 panel: a genomic resource for cancer biology and systems pharmacology. *Cancer Res.* **73**, 4372–4382 (2013).
2. Lamb, J. The Connectivity Map: a new tool for biomedical research. *Nat. Rev. Cancer.* **7**, 54–60 (2007).
3. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science.* **313**, 1929–1935 (2006).
4. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature.* **483**, 603–607 (2012).
5. Garnett, M.J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature.* **483**, 570–575 (2012).
6. Basu, A. *et al.* An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell.* **154**, 1151–1161 (2013).
7. Cheung, H.W. *et al.* Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 12372–12377 (2011).
8. Marcotte, R. *et al.* Essential gene profiles in breast, pancreatic, and ovarian cancer cells. *Cancer Discov.* **2**, 172–189 (2012).
9. Duan, Q. *et al.* LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. *Nucleic Acids Res.* **42**, W449–460 (2014).
10. Peck, D. *et al.* A method for high-throughput gene expression signature analysis. *Genome Biol.* **7**, R61 (2006).
11. Benjamin, H.K. *et al.* Inconsistency in large pharmacogenomic studies. *Nature.* **504**, 389–393 (2013).
12. De Wolf, H., De, Bondt A., Turner, H., Göhlmann, H.W. Transcriptional characterization of compounds: lessons learned from the public LINCS data. *Assay Drug Dev. Technol.* **14**, 252–260 (2016).
13. Vempati, U.D. *et al.* Metadata standard and data exchange specifications to describe, model, and integrate complex and diverse high-throughput screening data from the Library of Integrated Network-based Cellular Signatures (LINCS). *J. Biomol. Screen.* **19**, 803–816 (2014).
14. Pritchard, J.R. *et al.* Defining principles of combination drug mechanisms of action. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E170–179 (2013).
15. Dorsett, Y. & Tuschl, T. siRNAs: applications in functional genomics and potential as therapeutics. *Nat. Rev. Drug Discov.* **3**, 318–329 (2004).
16. Levenson, A.S. & Jordan, V.C. MCF-7: the first hormone-responsive breast cancer cell line. *Cancer Res.* **57**, 3071–3078 (1997).
17. Fernandes, C. *et al.* New chiral derivatives of xanthones: synthesis and investigation of enantioselectivity as inhibitors of growth of human tumor cell lines. *Bioorg. Med. Chem.* **22**, 1049–1062 (2014).
18. Chalfie, M., Tu, Y., Euskirchen, G., Ward, W.W. & Prasher, D.C. Green fluorescent protein as a marker for gene expression. *Science.* **263**, 802–805 (1994).
19. Enoki, S., Saeki, K., Maki, K. & Kuwajima, K. Acid denaturation and refolding of green fluorescent protein. *Biochemistry.* **43**, 14238–14248 (2004).
20. Gurskaya, N.G. *et al.* GFP-like chromoproteins as a source of far-red fluorescent proteins. *FEBS Lett.* **507**, 16–20 (2001).
21. Jain, V.K. & Magrath, I.T. A chemiluminescent assay for quantitation of beta-galactosidase in the femtogram range: application to quantitation of beta-galactosidase in lacZ-transfected cells. *Anal. Biochem.* **199**, 119–124 (1991).
22. Eisenberg, E. & Levanon, E.Y. Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574 (2013).
23. Silver, N., Best, S., Jiang, J. & Thein, S.L. Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR. *BMC Mol. Biol.* **7**, 33 (2006).

24. Paddison, P.J., Caudy, A.A. & Hannon, G.J. Stable suppression of gene expression by RNAi in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 1443–1448 (2002).
25. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. *Nature.* **490**, 61–70 (2012).
26. Bjornstrom, L. & Sjoberg, M. Mechanisms of estrogen receptor signaling: convergence of genomic and nongenomic actions on target genes. *Mol. Endocrinol.* **19**, 833–842 (2005).
27. Thomas, R.S., Sarwar, N., Phoenix, F., Coombes, R.C. & Ali, S. Phosphorylation at serines 104 and 106 by Erk1/2 MAPK is important for estrogen receptor-alpha activity. *J. Mol. Endocrinol.* **40**, 173–184 (2008).
28. Guo, Y. *et al.* TSC1 involvement in bladder cancer: diverse effects and therapeutic implications. *J. Pathol.* **230**, 17–27 (2013).
29. Kim, H.G. *et al.* Discovery of a potent and selective DDR1 receptor tyrosine kinase inhibitor. *ACS Chem. Biol.* **8**, 2145–2150 (2013).
30. Weisberg, E. *et al.* Selective Akt inhibitors synergize with tyrosine kinase inhibitors and effectively override stroma-associated cytoprotection of mutant FLT3-positive AML cells. *PLoS One.* **8**, e56473 (2013).
31. Moore, C.B., Guthrie, E.H., Huang, M.T. & Taxman, D.J. Short hairpin RNA (shRNA): design, delivery, and assessment of gene knockdown. *Methods Mol. Biol.* **629**, 141–158 (2010).
32. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).

Supplementary information accompanies this paper on the *CPT: Pharmacometrics & Systems Pharmacology* website (http://www.wileyonlinelibrary.com/psp4)