Supplementary Information for “L1000FWD: Fireworks visualization of drug-induced transcriptomic signatures”

Zichen Wang\textsuperscript{1}, Alexander Lachmann\textsuperscript{1}, Alexandra B. Keenan\textsuperscript{1}, Avi Ma’ayan\textsuperscript{1,*}

\textsuperscript{1}BD2K-LINCS Data Coordination and Integration Center; Mount Sinai Center for Bioinformatics; Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place Box 1603, New York, NY 10029 USA

*To whom correspondence should be addressed.

1 Supplementary Methods

Preparation of signatures

Drug/small-molecule perturbation profiles from the quantile normalized LINCS L1000 data (Level 3) \cite{1} were accessed through a gcxt file provided to us by the Broad LINCS Center for Transcriptomics. Drug-induced gene expression signatures are defined as gene expression changes computed by comparing gene expression data collected before and after treating a specific cell line with a specific drug, applied in a specific concentration, and where gene expression was measured at a specific time point. Therefore, each signature contains the four attributes: drug, cell line, dosage, and time point. The method proposed by \cite{2} was adopted to adjust for batch effects. Briefly, the L1000 gene expression profiles were first mean centered on the gene axis. Then, the expression signatures in the space of landmark genes and all genes were computed using drug treated samples versus samples treated with all other drugs, instead of comparing to biological controls. Drug/small-molecule perturbation signatures were computed using the Characteristic Direction (CD) method \cite{3}, which assigns CD coefficients to each gene to estimate the magnitude and direction of differential expression in a multivariate fashion.

To evaluate the quality of the signatures, we adopted a method for computing Signature Consistency Score (SCS), which is derived from an empirical permutation test with the null hypothesis that the observed CD signatures are not significantly different from a pool of random CD signatures generated using random expression profiles \cite{4}. Hence, we shuffled the samples used for computing the observed CD signatures to compute random CD signatures, then calculated the average cosine distance among the random CD signatures. We repeated this process 10,000 times to obtain an empirical distribution of average cosine distances. We then computed the one-tail empirical p-value to measure the consistency among the drug treated replicates.

To convert CD signatures into differentially expressed gene sets, we performed z-tests for each CD signatures (vectors of CD coefficients) in the full gene expression space (landmark genes and imputed genes) to identify genes with significantly (p-value <0.05) positive or negative CD coefficients, as up regulated and down regulated genes, respectively. Those differentially expressed gene sets are used for the signature similarity search functionality.
Benchmarking signature connectivity

Signature connectivity is a way to measure the similarity between a pair of gene expression signatures. To evaluate different signature connectivity methods as well as to evaluate the quality of signatures before and after batch effect removal, we used both intrinsic and extrinsic benchmarks. We evaluated three different connectivity methods including Gene Set Enrichment Analysis (GSEA) [5] computed using the most up and down regulated 50 genes (ES50), cosine similarity, and Pearson’s correlation coefficient among the 1000 landmark genes. The intrinsic benchmark for connectivity examines the distribution of signature connectivity for the following four groups: 1) Same drug on the same batch; 2) Same drug on different batches; 3) Different drugs on the same batch; 4) Different drugs on different batches. Batch effects can be revealed if signatures from the same batch have higher connectivity than signatures from different batches for the same drug (Fig. S1). The extrinsic benchmark for connectivity examines the ability of the drug-induced gene expression signature to recover MOA of drugs. To implement this benchmark, we used drug MOA curated by [6] as the gold standard and evaluated the rank ordered signature pairs by different methods (Fig. S2).

Construction of a signature similarity network

The LINCS L1000 dataset used covers over 20,000 drugs and compounds and over 127,000 signatures. We kept the 48,098 significant signatures with empirical p-value < 0.05, which covers 4,944 drugs and 71 cell lines. Few HDAC inhibitors were used as positive control in the L1000 dataset, yielding hundreds of significant signatures for the same few drugs. To ameliorate the dominant effect of those positive control compounds, we kept the top 20 most significant signatures for each drug/compound, resulting in 34,502 signatures. Next, we computed the adjacency matrix for the pairwise signatures using cosine similarity. To select an optimal threshold to construct a network from the adjacency matrix, we performed unsupervised clustering of the signatures using DBSCAN [7]. The resultant cluster labels were evaluated against the MOA of the drugs using Adjusted Mutual Information (AMI), which measures the concordance between the clustering of the signatures and the signatures grouped by MOA. We then determined the optimal threshold to be the 99.95% percentile (cosine similarity of 0.61). After removing edges below this threshold, the network contained 18,082 nodes and 595,177 edges. We further removed connected components with less than 10 nodes to derive the final network for the fireworks visualization made of 16,848 nodes and 594,372 edges, which covers 68 cells, 132 dosages, and 3,237 drugs/compounds.

Visualization of cell specific maps

The cell specific maps of drug-induced signatures were visualized by two different methods: 1) the t-distributed Stochastic Neighbor Embedding (tSNE) [8] and 2) the fireworks display (FWD). We first selected the top 20 cell lines with the most significant signatures with empirical p-value < 0.05. Next, we computed the adjacency matrices for the pairwise signatures using cosine similarity for signatures grouped by the top 20 cell lines.

To visualize the adjacency matrix of signatures using tSNE, we initialized the embedding using principal component analysis (PCA), then performed tSNE algorithm implemented in scikit-learn library [9] with default parameters.

To visualize the cell specific signatures using the FWD, we constructed k-nearest neighbor (kNN) graphs from the adjacency matrices with k=5 to connect the top 5 most similar signatures with
each other. The layouts of the kNN graphs were then determined by executing the Allegro edge-repulsive strong clustering algorithm implemented in Cytoscape [10] with 10,000 iterations.

**Extraction of clinical information for prescribed drugs**

Clinical usage information for prescribed drugs was extracted from the Mount Sinai Health System (MSHS) Electronic Medical Records (EMR). Briefly, RxNorm was used to map the drug entities identified by brand names and/or generic names to active ingredients. A five-day sliding window was used for ~800,000 individuals in the EMR to count co-occurrences between drugs and diagnosis codes identified by the International Classification of Diseases (ICD)-9, as well as co-occurrences between pairs of drugs. To quantify the co-occurrences between two entities, Jaccard index was used to account for the total occurrences of both entities. The age at the time of prescription was recorded and aggregated to achieve a distribution of ages for the prescription of each drug.

**Similarity search for drug-induced gene expression signatures**

To quantify the similarity between an input pair of up/down gene sets, we used similarity score, which is the overlap between the input up/down genes and the signature up/down genes divided by the effective input [11]. The effective input is the length of the intersection between the input genes and all the L1000 genes since some input lists contain genes that are not present in the L1000 dataset. This includes all ~22,000 L1000 genes (landmark and inferred). We used Fisher exact test to assess the statistical significance of the overlap between the input up/down gene sets and the drug gene sets while penalizing the intersections between genes of opposite direction. We also simulated 10,000 pairs of random up/down gene sets as input to calculate the similarity scores for all drug signatures to get expected ranks. For a given observed rank, we calculate a Z-score quantifying the deviation of the observed rank from the expected rank. We also compute a combined score \( c \), which is an empirical scoring/ranking method previously developed for the gene list enrichment tool Enrichr [12]. Combined score is calculated by multiplying the logarithm of the p-value from the Fisher exact test and the Z-score as a composite index:

\[
c = z \cdot \log_{10}(p)
\]

We benchmarked those three scoring metrics (similarity score, Z-score, and combined score) using 257 drug-induced gene expression signatures from the Gene Expression Omnibus (GEO) [13] collected for a previous study [14] to rank the >40,000 drug-induced signatures from the L1000 data.
2 Supplementary Figures

Fig S1. **Intrinsic benchmark of signature connectivity.** The distribution of signature connectivity grouped by the four conditions indicated in the legend are plotted across the 6 panels, with (A-C) derived from signatures computed without adjusting for batch effects, and (D-F) derived from signatures with batch effect adjustment. ES50: Enrichment score computed using the most up and down regulated 50 genes; cosine: cosine similarity, correlation: Pearson’s correlation coefficient.
**Fig S2. Extrinsc benchmark of signature connectivity.** Partial receiver operating characteristic (ROC) curves are plotted for rank ordered signature pairs by cosine similarity (blue) and ES50 (red) with (solid lines) or without (dashed lines) batch effect adjustment in recovering MOA of drugs.

**Fig S3. Determining the optimal threshold for constructing the signature similarity network.** Adjusted mutual information (AMI) scores are plotted over different thresholds used for constructing the signature similarity network.
**Fig S4. Benchmarking of the signature similarity scoring methods.** The cumulative distribution of the scaled ranks (r) of matched drug-induced expression signatures between CREEDS and L1000. The deviation of the cumulative distribution of the scaled ranks from the uniform distribution (D(r) – r) is plotted against the scaled ranks. Area under the curve (auc) values are shown in the legend.
3 Supplementary References