L1000CDS2 User Manual

Summary

Description of the L1000 Data

Signatures Available for Download

Tutorial

***Input***

Search

Examples and Preloaded External Signatures

Configuration

Metadata

Recent Searches

***Result***

Table View

Header

Table of Combinations

Substructure Enrichment Analysis

API

Contact

**Summary**

L1000CDS² is a LINCS L1000 characteristic direction signature search engine. It enables users to find consensus L1000 small molecule signatures that match user input signatures. The underlying dataset for the search engine is a portion of the [LINCS L1000](http://www.lincscloud.org/) small molecule expression profiles generated at the Broad Institute by the [Connectivity Map team](http://lincscloud.org/about/cmap-team/). The differentially expressed (DE) genes of these profiles were calculated using the [characteristic direction](http://maayanlab.net/CD/) method. Depending on the user’s input, L1000CDS² uses either a gene-set method or cosine distance method to compare the input signatures to the L1000 signatures to perform the search. When up/down gene lists are submitted to L1000CDS², the search engine compares the input gene lists to the DE genes computed from the LINCS L1000 data and descriptive information of the top 50 matched signatures is returned. When a signature is submitted to L1000CDS² in the format of "gene symbol, expression value", the search engine calculates a cosine distance between the input signature and every characteristic direction signature in the underlying dataset, and the top 50 signatures of either the largest (reverse mode) or the smallest (mimic mode) cosine distances are returned. L1000CDS² leverages the efficiency of matrix operations to perform the search. The search finishes a query against more than 30,000 signatures in less than a second using the gene-set method, or less than three seconds using the cosine distance method. The L1000CDS2 application is developed by the [Ma'ayan Laboratory](http://icahn.mssm.edu/research/labs/maayan-laboratory" \t "_blank) at the [Icahn School of Medicine at Mount Sinai](http://icahn.mssm.edu/) for the [BD2K-LINCS-DCIC](http://www.lincs-dcic.org/) and the [KMC-IDG](http://commonfund.nih.gov/idg/fundedresearch) NIH funded projects.

**Description of the L1000 mRNA Expression Data**

The L1000 mRNA gene-expression dataset is generated as part of the [Library of Integrated Network-based Cellular Signatures (LINCS) project](http://www.lincsproject.org/), an NIH [Common Fund](https://commonfund.nih.gov/) program. LINCS aims to systematically profile the molecular and phenotypical outcomes of agent perturbed human cells. The variety of agents includes chemical compounds, different micro-environments, endogenous ligands, gene knockdown and overexpression. The L1000 dataset comprises of over a million gene expression profiles of chemically or genetically perturbed human cell-lines. The L1000 technology developed at the Broad Institute measures only ~1000 genes in each sample and estimates the expression of the rest of the transcriptome using an empirical linear model. In this way, the technology significantly reduces the cost and makes large-scale gene expression profiling possible. L1000CDS² currently includes a subset of the chemically perturbed gene expression profiles, specifically the profiles in the LJP, CPC and CPD peturbagen groups. After averaging replicates and removing insignificant signatures, we remained with 33,197 significant signatures served by the default search in L1000CDS².

### L1000 CD signature download

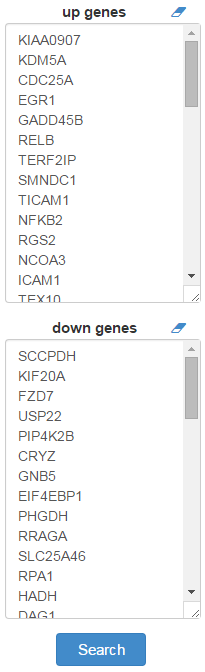
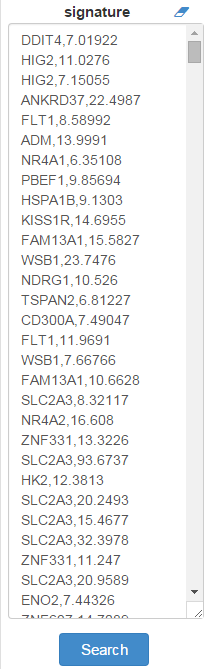
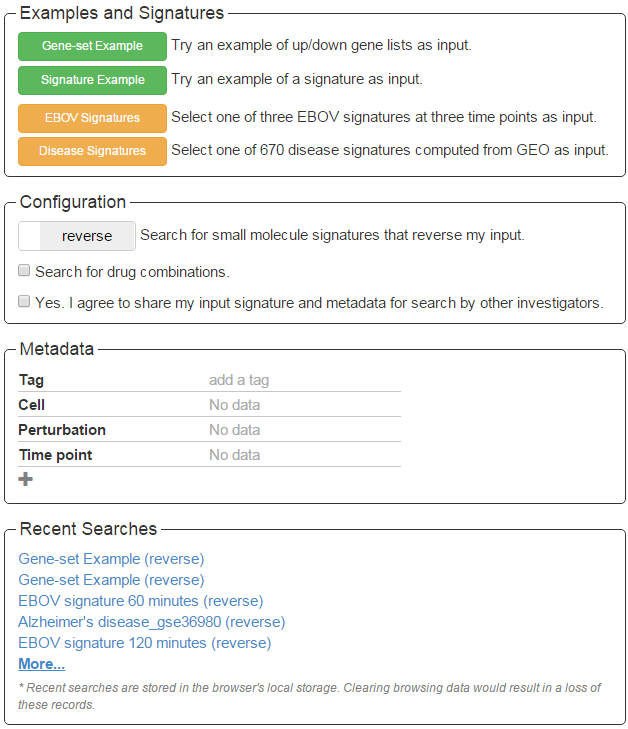
A large collection of L1000 gene expression signatures computed using the characteristic direction method, including signatures computed from the LJP, CPC and CPD collections can be downloaded from [here](http://amp.pharm.mssm.edu/public/L1000CDS_download/). The link provides instructions how to install a MongoDB database and then loading the data into this database for easy access.

**Tutorial**

*Input*

**Search**

The entry point into L1000CDS² is to paste up/down gene lists into the up/down gene text boxes (Fig. 1) or paste a signature (Fig. 2) into the up gene text box. A signature is a list of genes and their differential expression values separated by a comma. The search button will only become enabled when both up/down genes text boxes are filled by gene lists, or when the up gene textbox is filled with a properly formatted signature. Clicking the Search button, the information for the top 50 signatures will be displayed in a [table](http://localhost:8182/L1000CDS2/help/#table) in a new page.

*Fig3.* *Screenshot of the four input form sections: Examples and Signatures, Configuration, Metadata and Recent Searches*

*Fig2. Screenshot of the input up gene text box filled with a signature.*

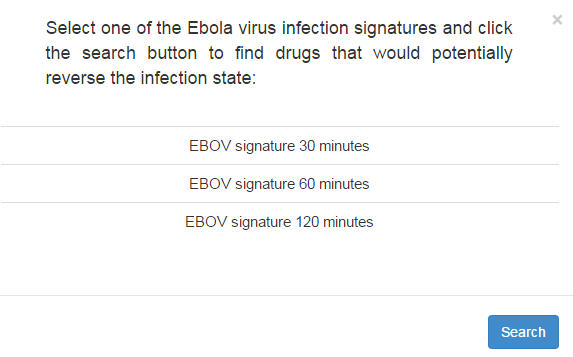
*Fig1. Screenshot of the input text boxes filled with up/down gene lists.*

**Examples and Preloaded Signatures from External Resources**

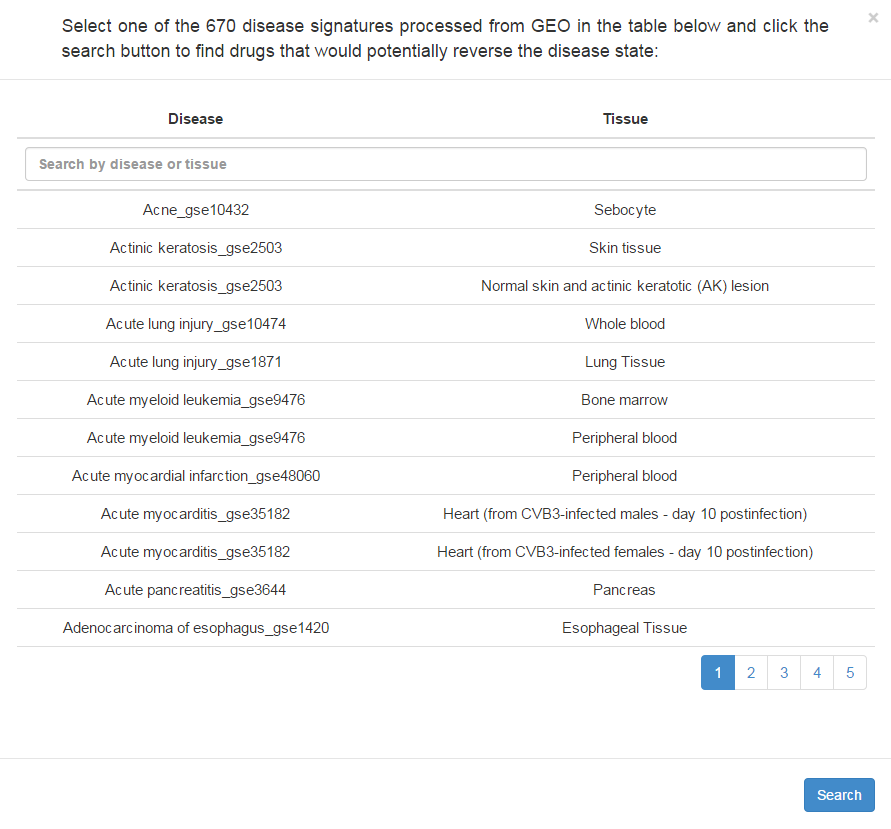
Clicking the Gene-set Example button will fill in an example of up/down gene lists in the text boxes for a demo search using the gene-set method. Clicking the Signature Example button will fill in an example of a signature in the up gene text box for a demo of searching with the cosine distance method.

Clicking the EBOV Signatures button will open a table with three EBOV signatures at three time points (Fig. 4). Selecting a signature by a single click and clicking the Search button can be used to perform a cosine distance search. The selected signature will be automatically filled in the up gene text box and the associated metadata will be filled in the [metadata](file:///C:\Users\Avi-Maayan\AppData\Local\Microsoft\Windows\Temporary%20Internet%20Files\Content.Outlook\6172SJPM\to%20metadata%20section) section.

Clicking on the Diseases Signature button will open a table of disease names and their tissue types including the GEO ID referring to the original study from which the disease signature was extracted from (Fig 5). The table is searchable by disease name or tissue type and can be sorted by either column. Each row represents a signature of a disease. This means it consists of the differentially expressed genes and their expression values in the disease compared with the relevant normal tissue. These signatures were calculated from gene expression data deposited in GEO. Clicking on a row will select that disease signature as an input to L1000CDS2. Clicking on the search button will fill in the signature of the selected disease in the up gene text boxes and the associated metadata in the metadata section and perform a search for small molecules at the same time.

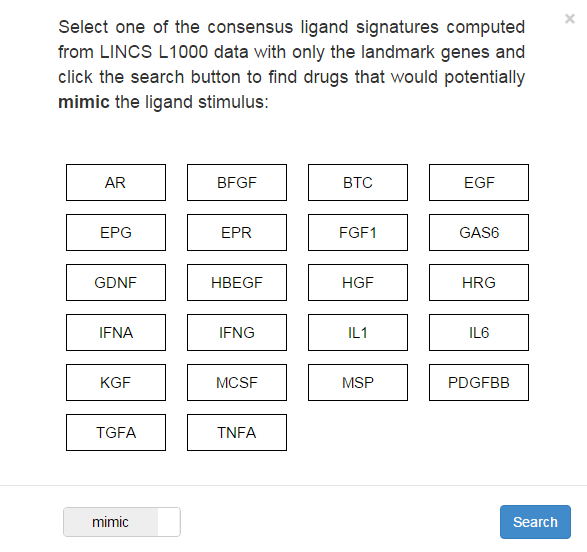


*Fig 4. Screenshot of the EBOV signature table*



*Fig 5. Screenshot of the disease signature table*

Clicking the Ligand Signatures button will open a grid of 22 consensus ligand signatures (Fig. 6). These signatures are characteristic direction signatures computed from the LJP LINCS L1000gene expression data using the **landmark genes only**. Selecting a signature by a single click, and then clicking on the Search button, L1000CDS2 performs a cosine distance search on a dedicated database that contains landmark genes CD signatures (cpcd-gse70138-lm-v1.0). The default mimic/reverse setting is mimic. Users can switch the mimic/reverse slider which can be found under the grid, to search in the reverse direction.



*Fig 6. Screenshot of the ligand signature table*

**Configuration**

***Mimic/Reverse****:* Clicking the mimic/reverse slider can be used to switch between reverse and mimic modes. In the mimic mode for the gene-set search, the input up genes are intersected with the up genes of the gene expression profiles in the L1000CDS² database, and the input down genes are intersected with the down genes for each entry in the database. In the reverse mode for the gene-set search, the input up-genes are intersected with the down-genes, and the input down-genes are intersected with the up-genes. When a cosine distance search is performed, the top 50 signatures will be those of the smallest cosine distances from the input signature in the mimic mode, or those with the largest cosine distances in the reverse mode. **The default mode is reverse.**

***Database Version:* The DB version select chooses the version of database for search. Currently, there are two versions: cpcd-v1.0 and cpcd-gse70138-v1.0. cpcd-v1.0 covers only the CPC and CPD subset while cpcd-gse70138-v1.0 also includes the newest gse70138 subset. The latest option in the DB version select always corresponds to the most recent and comprehensive database which at present is cpcd-gse70138-v1.0. There is a third hidden DB named cpcd-gse70138-lm-v1.0. The signatures in this DB are basically the same as those in cpcd-gse70138-v1but were computed using only the 978 landmark genes. This DB is dedicated to the consensus ligand signature search for the reason that the ligand signatures are also computed using only the 978 landmark genes. This option will show up if user performs a ligand signature search.**

***Drug Combinations:* L1000CDS2 also provides the function to search for drug combinations. To enable this feature the user need to check the “Search for drug combinations” checkbox. When searching for combinations, L1000CDS2 compares every possible pair among the top 50 signatures and computes the potential synergy for each pair. With the gene-set search, the synergy is calculated as the combined overlap of the DE genes of two signatures with the input DE genes. In a cosine distance search, the synergy is calculated as the** orthogonality between two CD signatures. The rational for this is that if two perturbations are orthogonal, they may impart their overall effect through two independent pathways.

***Including more small molecules in the signature search:* By default L1000CDS2 searches only the significant collection of signatures as determined by the characteristic direction method. If the option to “*include less significant signatures”*  is selected, L1000CDS2 will search all signatures irrespective of their significance. This option allows users to increase their search breadth but run the risk of detecting and prioritizing small molecules that are less likely to produce the desire effect on global expression.**

***Share:*** Users can share their input signatures and metadata so other users can query the signatures and gene sets they submitted. To make submitted input gene sets and signatures “public” for research purposes, a checkbox is available (Fig. 3). The default is set to “No” such that users’ input is made private. Users can still share their input lists, signatures, metadata and results using the share icon on the result page [(Fig. 7).](http://10.91.53.50/L1000CDS2/help/#tableFig) Clicking on the share icon produced a permanent URL that can be shared through e-mail, publications and other documentation.

**Metadata**

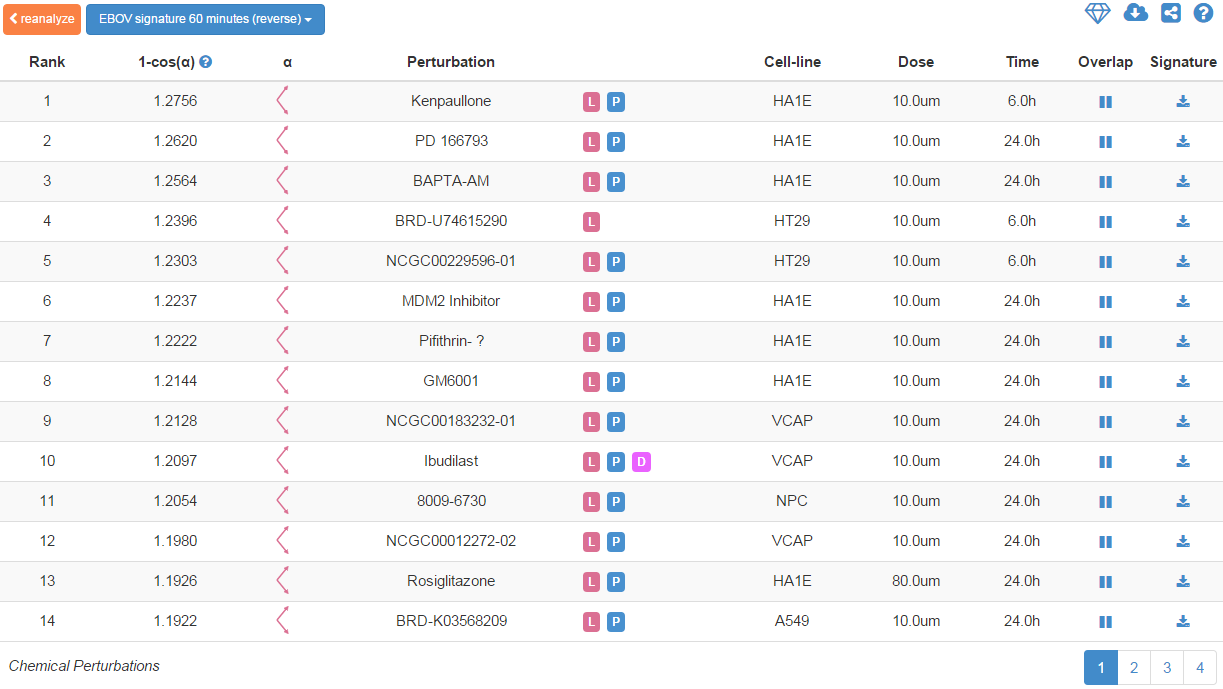
Any metadata associated with the input signature can be entered in the metadata section. By default, the section provides four input fields for metadata: Tag, Cell, Perturbation and Time Point (Fig. 3). Users can add new input fields for additional types of metadata by clicking the plus sign at the bottom or remove one by clicking the minus sign on the right of each row. The minus sign will only appear when the mouse cursor is hovering over a row. The tag field is used to enter few words which are most descriptive of the input signature.

**Recent Searches**

Recent search history will be displayed in this section as links (Fig 3). Clicking a link will show the results for that search. Recent searches are stored in the browser's local storage buffer. Clearing browsing data would result in a loss of these records. A maximum of 50 recent searches are stored for each user’s browser.

*Result*

**Table**

****

*Fig.7 Screenshot of the paginated results table and the metadata header*

The search results are rendered as a paginated table with 14 entries per page (Fig 7). Each entry provides seven pieces of information about the signature: rank, metric, perturbation, cell-line, dose, time point and overlap with the input.

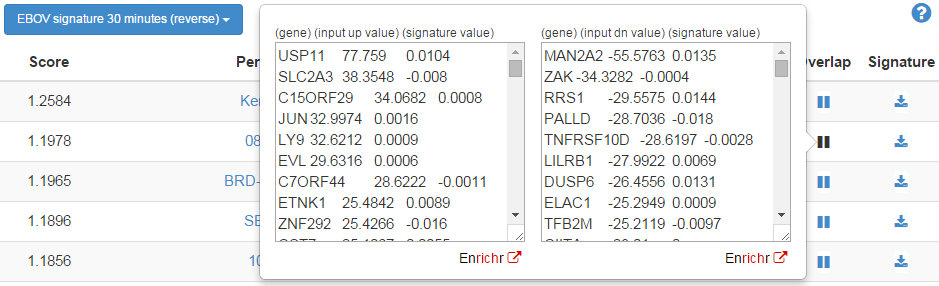
***Rank***: The rank of a signature is based on the search score.

***Search score:***  For gene-set search, the search score is the overlap between the input DE genes and the signature DE genes divided by the effective input. The effective input is the length of the intersection between the input genes and the extrapolated L1000 genes since some input lists contain genes that are not present in the L1000 extrapolated dataset. This includes all ~22,000 L1000 genes, not just the measured ~1000. This metric is visualized as a venn diagram in the venn column. For the cosine distance search, the metric is the cosine distance between the input signature and the L1000 characteristic direction signatures. The consensus signatures are sorted by their scores in descending order in reverse mode and in ascending order in mimic mode. The cosine value is subtracted by 1 such that a metric of 0 is for zero degree between the vectors, or fully aligned vectors 1 = cos(0) -> metric 0. Whereas a metric of 2 means that the two vectors point in the opposite direction with an angle of 180 degrees: -1=cos(180) -> metric 2. This metric is visualized as an angle in the α column.

***Perturbation:*** The perturbation column shows the names of the chemical perturbations.  The perturbation column shows the names of the chemical perturbations. User can click on the three **L** **P** **D** icons to look up perturbations in [LIFE](http://life.ccs.miami.edu/life/), [PubChem](https://pubchem.ncbi.nlm.nih.gov/" \t "_blank) and [DrugBank](http://www.drugbank.ca/" \t "_blank) that catalog detailed information about the chemical compound. Not every perturbation is available in all the three resources. There is also a signature column in the table.

***Cell-line, Dose and Time***: The cell-line, dose and time point used for generating the signature.

***Overlap***: Clicking the overlap button will show the overlapping genes (and their values) in two text boxes (Fig. 8). If the user input type is up/down gene lists, the first box will show the overlap genes between the input up genes and the signature up (down) genes and the second will show the overlap between the input down and the signature down (up) in mimic (reverse) mode. If the input is a signature, the first box will show genes with a positive value in input and their values in the signature; the second box will show genes with a negative value in input and their values in the signature. The signature values and input values in both boxes are expected to be mostly in the same sign in mimic mode and in the opposite sign in reverse mode. The Enrichr button under each text box will send the genes to Enrichr for enrichment analysis.



*Fig. 8 Screenshot of the overlap between the input signature and a small molecule signature*

***Download***: Clicking the download button in that column will download all the information about a signature as a JavaScript Object Notation (JSON) file.

***Target prediction***: L1000CDS2 contain a feature that is used to predict the target of small-molecules and drugs. This feature uses an independent external collection of gene expression studies of single gene perturbation signatures. The goal is to predict potential targets for all the small molecules profiled in the LINCS L1000 data. The single gene perturbation signatures are manually extracted from the Gene Expression Omnibus (GEO) through a crowdsourcing project. Students from the Coursera MOOC course Network Analysis in Systems Biology were asked to participate in a crowdsourcing project where they had to identify studies that perturbed a single gene in mammalian cells and where gene expression was measured before and after such perturbation. In total, 2,476 GEO signatures were generated and are used by L1000CDS2 for target prediction. The GEO signatures consist of perturbations that both up-regulate or down-regulate the target gene. To predict targets, we compare all the L1000 characteristic direction (CD) signatures to all the GEO single gene perturbation studies signatures. We aim to find the GEO signatures that are most similar to the L1000 CD signatures, assuming that the small molecule from the L1000 collection target the genes from the GEO studies. The top 20 GEO signatures ranked by cosine distance are listed as potential targets. Clicking on the target icon in each row will reveal the predicted targets as a table (Fig. X). The table displays the cosine distance, the perturbed gene, the direction of regulation and the GEO ID from which the signature was extracted. Clicking on the gene symbol displays detailed information about that gene as listed on the Harmonizome web portal we independently developed. Clicking on the GEO ID, the user is directed to the GEO page from which the gene expression of the single gene perturbation was extracted.

**Header**

On top of the table is a header bar that provides various functions (Fig. 7):

***Reanalyze:*** Clicking this button redirects the user back to the input page with input lists or signatures preloaded in the input textboxes. Users can then reanalyze their input using different configurations, or modify the associated metadata. This function also has a bearing on sharing results with others. It provides a way for users to reanalyze their input with different settings and obtain a permanent URL for each analysis.

***Tag:*** This button displays the tag and search mode. Clicking on the button shows the input metadata.

***Diamond*:** By clicking on this button, L1000CDS2 performs enrichment analysis on the substructures of the top ranked small-molecules. Refer to the Substructure Enrichment section for more detail about understanding the results from this feature.

***Cloud download:*** Clicking on this button downloads the table as a .csv file.

***Share:***  Clicking on the share icon produced a permanent URL that can be shared through e-mail, publications and other documentation.

**Table of combinations**

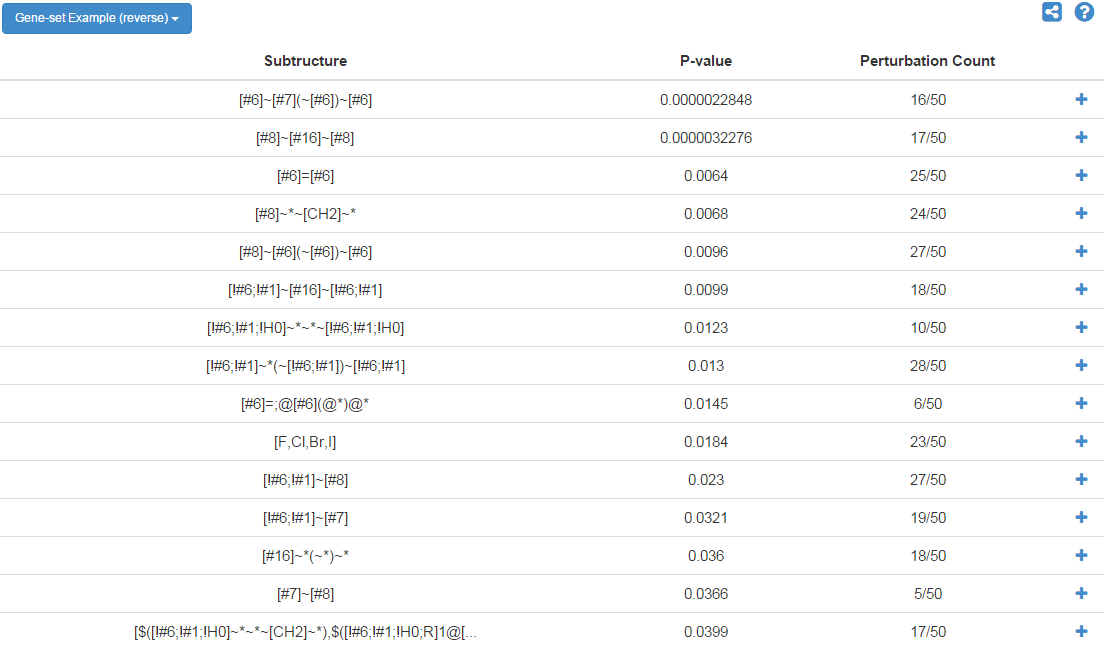
If the user chooses to search for drug combinations, a table of signature combinations will appear below the single perturbation result table. This table is also a paginated table with 14 entries per page (Fig 9). Each entry provides three pieces of information about the identified combinations: rank, synergy score and combinations. The synergy score has is described in the configuration section. The rank is based on the synergy score. The number before each chemical perturbation in the combinations column is the rank of that perturbation in the single signature result table. Clicking on a perturbation will highlight that perturbation in the single signature results table so the user can learn more detail about that perturbation. Clicking on the cloud download button on the upper right corner downloads the combination table as a .csv file.



**

Fig. 9 Screenshot of the drug combination table results

**Substructure Enrichment Analysis**

****

*Fig.10 Screenshot of the table of significantly enriched substructures.*

The results of the substructure enrichment analysis are displayed as a table where each row is a significantly enriched substructure (Fig. 10). Each row provides three pieces of information: substructure, p-value and perturbation count. The substructure is represented as a string in the SMARTS (add link) format. The p-value is computed using the Fisher’s exact test. The perturbation count shows the number of perturbations that have this substructure. Clicking on the share icon produces a permanent URL that can be share the substructure enrichment analysis results through e-mail, publications and other documentation.



*Fig.11 Screenshot of an expanded row*

Clicking on the plus sign shows a visualization of the substructure and a table of the top perturbations that contain the substructure (Fig. 11). The rank in the table is the rank of the perturbations in the top 50 signature table.

**API**

**[POST]** [**http://amp.pharm.mssm.edu/L1000CDS2/query**](http://amp.pharm.mssm.edu/L1000CDS2/query)

**Gene-set Search**

**PAYLOAD (content-type: application/json)**

|  |  |  |
| --- | --- | --- |
| data | Object | An object that saves the input up/down gene lists. |
| data.upGenes | [String] | An array of up-regulated genes. |
| input.data.dnGenes | [String] | An array of down-regulated genes. |
| config | Object | An object that saves the search configuration. |
| config.aggravate | Boolean | True to perform the search in aggravate mode and False in reverse mode. |
| config.seachMethod | String | “geneSet” |
| config.share | Boolean | True to agree to share input data and metadata. |
| config.combination | Boolean | True to search for drug combinations. |
| metadata | [Object] | An array of objects that saves the metadata of the input. |
| metadata[ ].key | String | The field name of a piece of metadata. |
| metadata[ ].value | String | The value of a piece of metadata. |

Response:

|  |  |  |
| --- | --- | --- |
| shareId | String | Unique ID for sharing the search results |
| combinations | [Object] | Drug combinations. Only available if config.combination is set to true. |
| combinations[ ].X1 | sig\_id | Unique identifier for a signature in the combination. |
| combinations[ ].X2 | sig\_id | Unique identifier for a signature in the combination. |
| combinations[ ].value | Number | Overlap metric |
| topMeta | [Object] | Descriptive information of the top 50 consensus signatures |
| topMeta[ ].score | Number | Score |
| topMeta[ ].cell\_id | String | Cell-line |
| topMeta[ ].pert\_desc | String | Perturbation name |
| topMeta[ ].pert\_id | String | Unique identifier for a perturbation |
| topMeta[ ].pert\_time | String | Time point |
| topMeta[ ].pert\_time\_unit | String | Time point unit |
| topMeta[ ].pert\_dose | String | Dose |
| topMeta[ ].pert\_dose\_unit | String | Dose unit |
| topMeta[ ].sig\_id | String | Unique identifier for a signature |
| topMeta[ ].overlap | Object | The overlapping genes between input DE genes and signature DE genes. |
| topMeta[ ].overlap.up/dn (reverse mode only) | [String] | The overlap between input up genes and signature down genes. |
| topMeta[ ].overlap.dn/up (reverse mode only) | [String] | The overlap between input down genes and signature up genes. |
| topMeta[ ].overlap.up/up (mimic mode only) | [String] | The overlap between input up genes and signature up genes. |
| topMeta[ ].overlap.dn/dn (mimic mode only) | [String] | The overlap between input down genes and signature down genes. |

**Cosine Distance Search**

**PAYLOAD (content-type: application/json)**

|  |  |  |
| --- | --- | --- |
| data | Object | An object that saves the input signature. |
| data.upGenes | [String] | An array of input genes. |
| input.data.dnGenes | [Number] | An array of input values that match to each input gene. |
| config | Object | An object that saves the search configuration |
| config.aggravate | Boolean | True to perform the search in aggravate mode and False in reverse mode. |
| config.seachMethod | String | “CD” |
| config.share | Boolean | True to agree to share input data and metadata. |
| config.combination | Boolean | True to search for drug combinations. |
| metadata | [Object] | An array of objects that saves the metadata of the input |
| metadata[ ].key | String | The field name of a piece of metadata. |
| metadata[ ].value | String | The value of a piece of metadata. |

Response:

|  |  |  |
| --- | --- | --- |
| shareId | String | Unique ID for sharing the search results |
| uniqInput | Object | Unique input genes overlapped with L1000 genome and their input values. |
| uniqInput.up | Object | Unique input up genes and their input values. |
| uniqInput.up.genes | [String] | Unique input up genes. |
| uniqInput.up.vals | [Number] | Values of unique input up genes. |
| uniqInput.dn | Object | Unique input down genes and their input values. |
| uniqInput.dn.genes | [String] | Unique input down genes. |
| uniqInput.dn.vals | [Number] | Values of unique input down genes. |
| combinations[ ].X1 | sig\_id | Unique identifier for a signature in the combination. |
| combinations[ ].X2 | sig\_id | Unique identifier for a signature in the combination. |
| combinations[ ].value | Number | Orthogonality |
| topMeta | [Object] | Descriptive information of the top 50 consensus signatures |
| topMeta[ ].score | Number | Score |
| topMeta[ ].cell\_id | String | Cell-line |
| topMeta[ ].pert\_desc | String | Perturbation name |
| topMeta[ ].pert\_id | String | Unique identifier for a perturbation |
| topMeta[ ].pert\_time | String | Time point |
| topMeta[ ].pert\_time\_unit | String | Time point unit |
| topMeta[ ].pert\_dose | String | Dose |
| topMeta[ ].pert\_dose\_unit | String | Dose unit |
| topMeta[ ].sig\_id | String | Unique identifier for a signature |
| topMeta[ ].overlap | Object | The signature values of unique input genes overlapped with L1000 genome. |
| topMeta[ ].overlap.up | Object | The signature values of unique input up genes. |
| topMeta[ ].overlap.dn | Object | The signature values of unique input down genes. |

Python Example:

….

Contact

Please contact Avi Ma’ayan and Qiaonan Duan for comments, suggestions, and support:[avi.maayan@mssm.edu](mailto:avi.maayan@mssm.edu), [qiaonan.duan@mssm.edu](mailto:qiaonan.duan@mssm.edu).