User Manual: KSEA App

The KSEA App computationally infers relative kinase activity from quantitative phosphoproteomics data.

Version: 1.0 (February 2017)

URL: https://casecpb.shinyapps.io/ksea/

Please cite the following papers if using this tool:

1. Wiredja D.D., Koyutürk M., Chance M.R. (2017) The KSEA App: a web-based tool for kinase activity inference from quantitative phosphoproteomics. *Submitted for review.*
2. Casado, P., Rodriguez-Prados, J.-C., Cosulich, S. C., Guichard, S., Vanhaesebroeck, B., Joel, S., and Cutillas, P. R. (2013) Kinase-substrate enrichment analysis provides insights into the heterogeneity of signaling pathway activation in leukemia cells. *Sci. Signal*. 6, rs6-rs6

Current specifications for the Kinase-Substrate Database:

- PhosphoSitePlus Database: July 2016 Release, restricted to human proteins
- NetworKIN predictions: pre-computed dataset from Human ENSEMBL version 59, restricted to predictions with scores 1 or greater

Version 1.0 only supports human protein analysis!

Thank you for using the KSEA App!

Any questions? Please email ddw34@case.edu.

Table of Contents

Instructions

1) Format the Data File

The KSEA App only accepts an experimental data file that adheres to the following specific format.

The data must be a comma delimited .csv file containing these columns with the ***exact*** header (in bold) and order:

- **Protein**: the Uniprot ID for the parent protein [optional]
- **Gene**: the HUGO gene symbol for the parent protein
- **Peptide**: the sequence of the peptide [optional]
- **Residue.Both**: the modified site (multiple sites are supported but must be separated by semicolons without spaces)
- **p**: the p-value of the differential phosphorylation between the two groups [optional]
- **FC**: the fold change between the two groups' peptide intensities (not log-transformed); only numeric values are allowed, and there can be no character values or blank cells

Sample dataset correctly formatted for input into the KSEA App interface:

	A	B	C	D	E	F
1	Protein	Gene	Peptide	Residue.Both	p	FC
2	P62263	RPS14	IEDVTPIPSDSTR	S139	2.84E-05	0.423607
3	Q9H4L7	SMARCAD1	KLSSSSEPYEEDEFNDDQSIK	Y217;S214	4.51E-05	0.591793
4	Q8N5I9	C12orf45	IEVLDSPASK	S178	0.000082	0.502393
5	P29375	KDM5A	VEPEVLSTDTQTSPEPGTR	S204	0.000222	0.476324
6	Q5UIP0	RIF1	SPQRPSDWSK	S782	0.000353	0.523072
7	Q86W50	METTL16	EGEAAAVEGPCPSQESLSQEENPEPTE	S455	0.000424	1.776738
8	O75400	PRPF40A	SDSPESDAEREK	S888	0.000439	0.718616
9	Q6PL18	ATAD2	NNSNTCNIENELEDSRK	S1243	0.000623	0.416637
10	P10588	NR2F6	AAEDDSASPPGAASDAEPGDEERPGL	S40	0.000793	0.407655
11	P31641	SLC6A6	SPGTRPEDEAEGKPPQR	T28	0.000829	0.326469

Important notes:

- The Gene, Residue.Both, and FC values are necessary for the calculations.
- The Protein, Peptide, and p values are optional, but the columns and headers must be present in the input file. You can leave the rest of the column as "NULL" but not blank.
- **DO NOT** leave cells blank or as **NA**, or else the entire row will be discarded!
- Rows with FC = 0 will be discarded since log2(0) cannot be calculated.
- FC values must use a period to indicate the decimal point, not a comma.
- *This current KSEA application only compares two groups at a time; therefore, if a dataset contains multiple conditions, the user must submit separate input files, each with a single fold change column, for every pairwise comparison.*

A sample data file input is provided at the top of the website's left panel.

2) Upload the Data

Click on the **Browse** button on the left panel.

This will allow you to select your formatted data file.

3) Select the Analysis and Plotting Parameters

Select kinase-substrate dataset

- Choose the kinase-substrate (K-S) dataset that the algorithm will use to extract substrates for each kinase
- **PhosphoSitePlus** alone is the more conservative choice since it only uses curated annotations.
- **PhosphoSitePlus + NetworKIN** allows users to include K-S predictions generated from NetworKIN; this option is recommended if you want to score more kinases or improve the coverage of experimental phosphosites used for KSEA calculations.

Set NetworKIN score cutoff

- This option only appears if you select **PhosphoSitePlus + NetworKIN** in the previous parameter.
- Choose the minimum NetworKIN score to decide which predicted K-S relationships to discard.
- A higher number corresponds to a more confident K-S prediction.
- The minimum cutoff is 1.

-------The following two parameters are for annotating the bar plot output only------

[for plot] Set p-value cutoff

- Decide which kinases in the bar plot would get marked for being statistically significant.
- Kinases with significant negative scores will be marked blue, whereas those that have significant positive scores will be red.
- $p < 0.05$ is the default setting.

[for plot] Set substrate count cutoff

- Decide the minimum number of detected substrates a kinase must have to be included in the bar plot.
- A lower cutoff thus allows more kinases into the bar plot.
- Kinases that don't pass the cutoff will be excluded from the bar plot but will still be in the two table outputs.

4) Click GO!

Then click on the **Plot**, **KSEA Kinase Scores**, and **Kinase-Substrate Links** tabs to see the results.

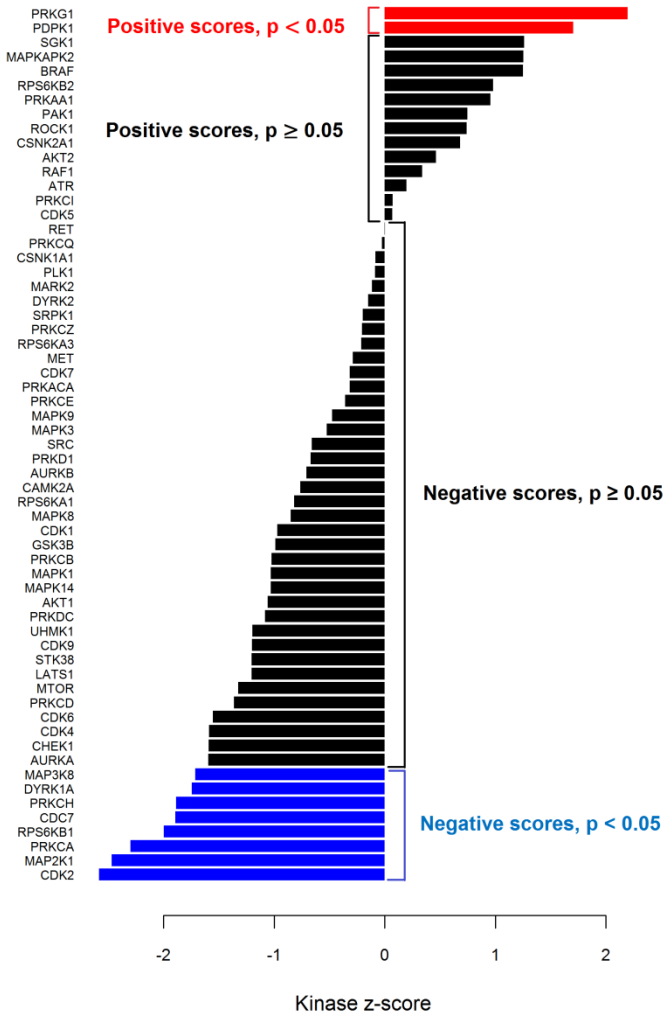
5) Download the Results

The KSEA results could be saved by downloading the individual outputs under the **Downloads** tab.

Plot

- This is the bar plot that summarizes the KSEA results.
- Note that not all kinases are included. The kinase substrate count cutoff, set in the left panel, decides which kinases to include in this plot.
- The p-value cutoff, set in the left panel, decides which kinases to color blue/red for visual annotation of kinases that reach statistical significance.
- Kinases with non-significant scores will be black.
- Do not worry if the preview bar plot under the **Plot** tab appears squished. The file that is available under the **Downloads** tab offers a clearer, well-proportioned .tif image.

Below is a bar plot, generated from phosphoproteomics data from Kim *et al., Mol. Cancer Res*, 2016.



Plotting parameters:
- [for plot] Set p-value cutoff: 0.05
- [for plot] Set substrate count cutoff: 4

KSEA Kinase Scores

- This is a complete table listing *ALL* the kinases, including those that are not featured in the bar plot, that have at least one identified substrate in the input dataset.
- Please refer to the original Casado *et al*. publication for detailed description of these columns and what they represent.
- **Kinase.Gene** indicates the gene name* for each kinase.
- **mS** represents the mean log2(fold change) of all the kinase's substrates.
- **Enrichment** is the background-adjusted value of the kinase's mS.
- **m** is the total amount of detected substrates from the experimental dataset for each kinase.
- **z.score** is the normalized score for each kinase, weighted by the number of identified substrates.
- **p.value** represents the statistical assessment for the z.score.
- **FDR** is the p-value adjusted for multiple hypothesis testing using the Benjamini & Hochberg method.

This table is useful if you want to 1) look at all the scored kinases, not just the filtered ones from the bar plot, and 2) generate custom graphs to supplement the KSEA App's default bar plot.

Note: the KSEA Kinase Scores table will NOT be altered if you change the bar plot's p-value cutoff and the kinase substrate count cutoff. However, this table will be altered if you adjust the NetworKIN score cutoff for the K-S dataset parameter.

Kinase-Substrate Links

- This is a complete table listing *ALL* the K-S relationships identified from the experimental dataset. This includes relationships for kinases that are not featured in the bar plot.
- For each kinase, every substrate identified from the dataset was used for the KSEA calculations (in other words, there was no filtering of the substrates).
- **Kinase.Gene** represents the gene name* for each kinase.
- **Substrate.Gene** indicates the gene name for each substrate linked to that kinase.
- **Substrate.Mod** is the substrate's specific amino acid residue that was modified.
- **Source** shows the database where the K-S annotation was derived from.
- **log2FC** is the log2(fold change) value of that particular substrate phosphosite from the experiment. If that same site was detected across multiple peptides that map to the same protein, the average log2FC is reported.

This table is useful if you want the specific substrates that contributed to each kinase's score.

Note: the Kinase-Substrate Links table will NOT be altered if you change the bar plot's p-value cutoff and the kinase substrate count cutoff. However, this table will be altered if you adjust the NetworKIN score cutoff for the K-S dataset parameter.

*__WARNING__: The *PDK1* gene name is often mistaken for the wrong protein. The KSEA App follows the assignments in this table, which are based on the PhosphoSitePlus and UniprotKB convention.

Kinase.Gene	Uniprot ID	Protein Name
PDK1	Q15118	pyruvate dehydrogenase kinase isoform 1
PDPK1	O15530	3-phosphoinositide-dependent protein kinase 1

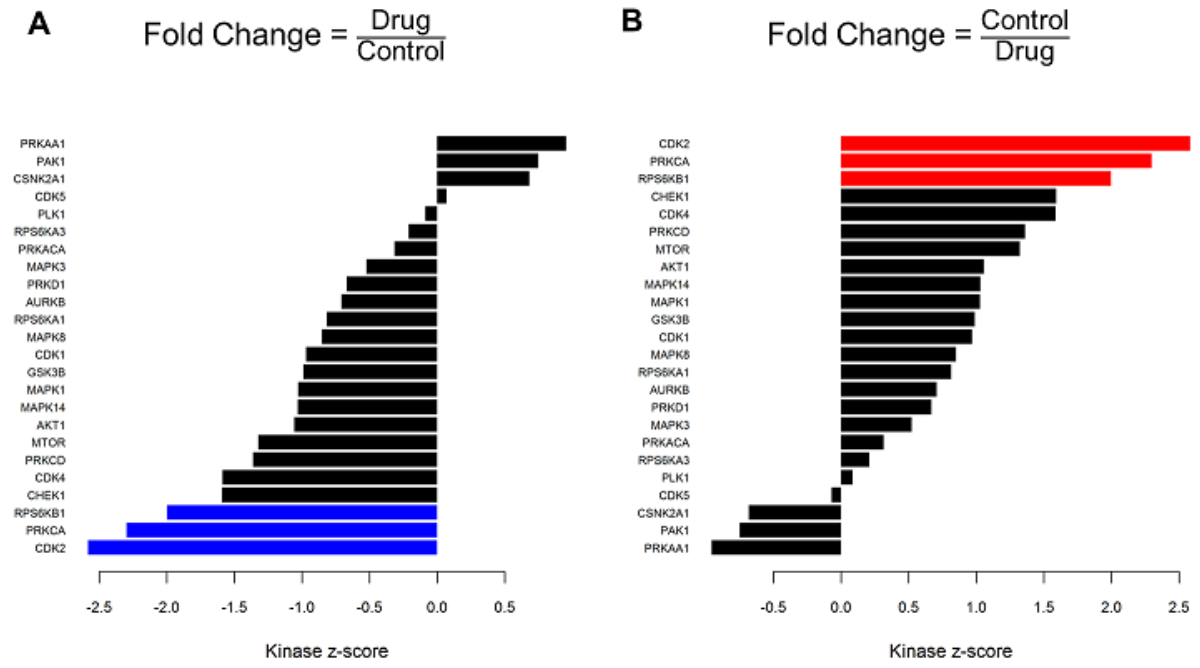
How to Interpret the Results

Interpretation of the kinase scores depends on how the fold change values were calculated!

We will demonstrate using a sample case study:

Take, for instance, a phosphoproteomics experiment looking at two groups: Control vs. Drug-Treated.

The following plots are generated by KSEA using the same dataset, but the fold change values were calculated differently.



A. Bar plot of the kinase scores calculated from the dataset in which the fold change was determined by **Drug/Control**. In this case, phosphosites with decreased phosphorylation in the drug-treated group will have a fold change ratio that is **less than one**, leading to a log2(fold change) that is **negative**. Thus, if a kinase (in this case, like CDK2, PRKCA, or RPS6KB1) has substrates that are collectively dephosphorylated with drug, its normalized score will be **negative** in value. *In turn, this kinase is deemed downregulated with drug because its signaling output is decreased in that group relative to control.*

B. Bar plot of the kinase scores calculated from the same experimental dataset, but this time the fold change was calculated as **Control/Drug**. The ratio is basically the inverse of A. Note how the kinases from **A** have the same magnitude of the kinase z-scores as in **B**, but the directionality is opposite (negative score in **A** becomes positive in **B**). In this case, phosphosites

with decreased phosphorylation in the drug-treated group will have a fold change ratio that is **greater than one**, leading to a log2(fold change) that is **positive**. Thus, if a kinase has a high proportion of substrates dephosphorylated with drug, its normalized score will be **positive** in value. Regardless, the final interpretation is the same as in **A**: *this kinase is deemed downregulated with drug because its signaling output is decreased in that group relative to control*.

Take-away message: the fold change setup determines whether a downregulated kinase has a negative or positive score!

We recommend keeping the reference/control group at the denominator of the fold change ratio. This makes interpretation more logical, given the bar plot's built-in coloring scheme.

Troubleshooting

An incorrectly-formatted dataset input is the most common cause for errors.

Below are examples of <u>unacceptable</u> inputs:

	A	B	C	D	E	F
1	Gene	Protein	Peptide	Residue.both	p	fc
2	YWHAE	P62258	AAFDDAIAELDTLSEESYK	S210	0.0005	1.19
3	SFN	P31947	DNLTLWTADNAGEEGGEAPQEPQS	S248	0.0017	3.89
4	YWHAQ	P27348	DNLTLWTSDSAGEECDAAEGAEN	S232	0.0019	1.83
5	YWHAZ	P63104	TAFDEAIAELDTLSEESYK	S207	0.0026	0.996
6	YWHAZ	P63104	NLLSVAYK	S45	0.0028	1.07
7	YWHAZ	P63104	DNLTLWTSDTQGDEAEAGEGGEN	T232	0.0029	2.75
8	TNKS1BP1	Q9C0C2	RFSEGVLQSPSQDQEK	S429	0.003	0.805
9	TNKS1BP1	Q9C0C2	YESQEPLAGQESPLPLATR	S601	0.0031	1.27
10	TNKS1BP1	Q9C0C2	SFGTRPLSSGFSPEEAQQQDEEFEK	S984	0.0032	0.815
11	TNKS1BP1	Q9C0C2	VSGAGFSPSSK	S1138	0.0033	0.872

A. The columns are in the incorrect order; it must follow the exact order as specified in p. 3.

B. The column titles are case-sensitive; the highlighted headers should be “Residue.Both” rather than “Residue.both” and “FC” instead of “fc”.

	A	B	C	D	E	F
1	Protein	Gene	Peptide	Residue.Both	p	FC
2	P62258	YWHAE	AAFDDAIAELDTLSEESYK	S210	NA	1.19
3	P31947	SFN	DNLTLWTADNAGEEGGEAPQEPQS	S248	NA	3.89
4	P27348	YWHAQ	DNLTLWTSDSAGEECDAAEGAEN	S232	NA	1.83
5	P63104	YWHAZ	TAFDEAIAELDTLSEESYK	S207	NA	0.996
6	P63104	YWHAZ	NLLSVAYK	S45	NA	1.07
7	P63104	YWHAZ	DNLTLWTSDTQGDEAEAGEGGEN	T232	NA	2.75
8	Q9C0C2	HGNC:19081	VPSSDEEVVEEPQSR	S1620;S1621	NA	1.28; 3.12
9	Q9C0C2	HGNC:19081	VPSSDEEVVEEPQSR	S1620	NA	1.72
10	Q9C0C2	HGNC:19081	YSSQDADEQDWEFQK	S919	NA	1.51
11	Q9C0C2	HGNC:19081	RESAASGLGGLLEEEGAGAGAAQEEVL	S1473;S1476	NA	1.81

C. These gene identifiers are not supported, so these rows will be missed when the algorithm searches against the K-S dataset. The application only accepts the HUGO Gene Symbol format.

D. These cells contain **NA**, which will lead to deletion of these rows in the analysis. Instead, write, “NULL**”** in these cells; <u>do not</u> remove the entire column or leave the cells empty. Remember: only the **Gene**, **Residue.Both**, and **FC** values are absolutely required for KSEA calculations, but the other three columns must still be present.

E. The fold change (FC) column incorrectly contains multiple values. There can only be one. See example K for another variation of this problem.

	A	B	C	D	E	F
1	Protein	Gene	Peptide	Residue.Both	p	FC
2	Q15019	2-Sep	IYHLPDAESDEDEDFK	S218	0.4766	0.163499
3	Q16181	7-Sep	ILEQQNSSRTLEK	T426	0.4768	-0.29513
4	Q9UHD8	9-Sep	LGDSSGPALKR	S22	0.4774	-0.30579
5	Q9UHD8	9-Sep	RLGDSSGPALK	S22	0.4775	0.042644
6	Q9NRG9	AAAS	FSPVLGR	S495	0.0315	0.250962
7	Q13685	AAMP	GTEGHQGPLTCVAANQDGSLILTGSVD	T290,T291	0.0362	0.124328
8	Q9NY61	AATF	YLVDGTKPNAGSEEISSEDDELVEEK	S316,S320,S321	0.3946	0.536053
9	Q9NY61	AATF	YLVDGTKPNAGSEEISSEDDELVEEK	S320,S321	0.3947	0.464668
10	Q9BZC7	ABCA2	ALVADEPEDLDTEDEGLISFEEER	T2412	0.0509	0.163499
11	P33527	ABCC1	HHNSTAELQK	S930	0.2906	-2.34373
12	O15439	ABCC4	DNEESEQPPVPGTPTLR	T646	0.2907	-0.33825

F

G

H

F. Be careful! If this datasheet was opened in MS Excel, some gene names that resemble months will be automatically converted to calendar dates. These genes will be missed when the algorithm searches against the K-S database. To avoid this, import the file into Excel by going to **Data** > **From Text**. After choosing the file to upload, pick **Delimited** under the file type, select **Comma** as the delimiter, and click **Next**. In the final step, click on the column with the gene names, and select **Text** under "Column data format." Click **Finish**.

G. The FC values are log-transformed, which will introduce errors. The correct FC values must be positive ratios ranging from 0 to infinity. However, the algorithm removes FC = 0 entries to avoid errors when these ratios are log-transformed in downstream analyses.

H. These residues are separated by commas, so these entries will be missed during the K-S search. While the application accepts multiple modification entries under **Residue.Both**, they must be separated by semicolons.

	A	B	C	D	E	F
1	Protein	Gene	Residue.Both	p	FC	FC
2	Q7Z7B0	FILIP1	T1205	0.1755	0.00297	0.869
3	Q64303	Pak2	S206	0.2641	0.00789	1.98
4	Q68D20	PMS2CL	S50	0.4215	0.00831	0.794
5	Q3B726	TWISTNB	S316	0.1312	0.0174	0.687
6	Q96I15	SCLY	S129	0.3415	0.0245	0.669
7	P05455	SSB	T362	0.2643	0.0331	3.11
8	Q02040	AKAP17A	S633	0.0305	0.0419	1.26
9	O14683	TP53I11	S14	0.5635	0.0677	1.2
10	Q8IVP5	FUNDC1	S13	0.1802	NULL	0.94
11	Q8N3D4	EHBP1L1	S1273	0.1472	0.0873	0.978

I

J

K

L

I. The **Peptide** column is missing completely. The header must be present in the file; just leave the subsequent cells as "NULL." *DO NOT* include **NA** or leave the cells empty/blank.

J. This protein/gene is from rats. The KSEA App currently supports only human protein analysis.

K. There are two **FC** columns where there can only be one. The current application only does two-group comparisons at a time. If the experimental dataset has 3+ conditions, you must create a new data file for each pairwise comparison.

L. This cell contains "NULL," which is a character value. This will create errors. Only positive numeric values are allowed under the **FC** column, so peptides/phosphosites with no FC measurements must be entirely discarded from the dataset.

Below are examples of <u>acceptable</u> inputs:

	A	B	C	D	E	F
1	Protein	Gene	Peptide	Residue.Both	p	FC
2	P62263	RPS14	IEDVTPIPSDSTR	S139	2.84E-05	0.423607
3	Q9H4L7	SMARCAD1	KLSSSSEPYEEDEFNDDQSIK	Y217;S214	4.51E-05	0.591793
4	Q8N5I9	C12orf45	IEVLDSPASK	S178	0.000082	0.502393
5	P29375	KDM5A	VEPEVLSTDTQTSPEPGTR	S204	0.000222	0.476324
6	Q5UIP0	RIF1	SPQRPSDWSK	S782	0.000353	0.523072
7	Q86W50	METTL16	EGEAAAVEGPCPSQESLSQEENPEPTE	S455	0.000424	1.776738
8	O75400	PRPF40A	SDSPESDAEREK	S888	0.000439	0.718616
9	Q6PL18	ATAD2	NNSNTCNIENELEDSRK	S1243	0.000623	0.416637
10	P10588	NR2F6	AAEDDSASPPGAASDAEPGDEERPGL	S40	0.000793	0.407655
11	P31641	SLC6A6	SPGTRPEDEAEGKPPQR	T28	0.000829	0.326469

	A	B	C	D	E	F
1	Protein	Gene	Peptide	Residue.Both	p	FC
2	NULL	E2F8	NULL	S71	NULL	1.7076724
3	NULL	ESYT2	NULL	S691	NULL	1.42234314
4	NULL	ESYT2	NULL	S755	NULL	1.62779682
5	NULL	ESYT2	NULL	S758	NULL	1.45135416
6	NULL	ESYT2	NULL	S761	NULL	1.3213071
7	NULL	MED19	NULL	S226	NULL	1.01828943
8	NULL	KIAA1598	NULL	S101	NULL	2.29095562
9	NULL	KIAA1598	NULL	S249	NULL	2.05053696
10	NULL	KIAA1598	NULL	S375	NULL	0.82156758
11	NULL	KIAA1598	NULL	S506	NULL	0.0807036

Warnings!

1. The KSEA App does not store any of the data once the site is reloaded or disconnected. Please download all the results before exiting!

2. To conserve server run time, the site disconnects automatically after 10 minutes of inactivity.

3. The KSEA App is on a public website hosted by RStudio's shinyapps.io online service (the "RStudio Service"). Please read their Terms of Use: https://www.rstudio.com/about/shinyapps-terms-use/.

   DO NOT upload confidential files, such as identifiable patient data:

   > RSTUDIO IS NOT RESPONSIBLE FOR THE CONFIDENTIALITY, AVAILABILITY, SECURITY, LOSS, MISUSE OR MISAPPROPRIATION OF ANY DATA YOU SUBMIT TO THE RSTUDIO SERVICE OR ANY APPLICATION MADE AVAILABLE VIA THE RSTUDIO SERVICE.

   If you wish to run KSEA on sensitive files, consider downloading the raw R code (supplemental file available through the journal) and running it locally on your computer.