KSEAPlus

User Guide

Version 0.1 (August 2019)

URL: http://ksea-plus.herokuapp.com

Table of Contents

1. Pre-Analysis Instructions

1.1 Dataset Requirements

- **Please ensure only .tsv files no larger than 5MB are uploaded.**

- Microsoft Excel can be used to generate .tsv files. To do so, open your dataset and choose **Edit → Select all.** Save the file by clicking **File → Save as…** and select **Tab-Delimited Text (.txt)** as the file format. Please ensure the generated file contains the .tsv extension. You can add this manually by renaming the file and adding the extension.

- Sample datasets are provided under the **_Getting Started_** tab of the KSEAPlus application.

- Each column must be titled and each row must have a value.

- As shown below, the first column of a dataset must refer to the detected phosphosites in the format of **gene(position);**. Multiple sites may be listed on the same line but they must be separated by semicolons. The remaining columns must refer to the corresponding log2(FC)/intensity values for each site. Different dataset formats are required depending on the algorithm chosen:

1.1.1 Z-Test/Kolmogorov-Smirnov

- These tests require columns **2** and onwards to contain pre-calculated log2 transformed fold-change (FC) values for each site. Before taking the log2 of the FC, it is generally advised to ensure the FC has been calculated in the format of _Treatment / Control_ and not the opposite.

	1	2	3	4
1	sh.index.sites	MCF7.GDC0941.fold	MCF7.AZD6482.fold	MCF7.CAL101.fold
2	KDM1A(T59);	0.028609794	0.084441611	-0.63413072
3	INCENP(S143);	0.400172957	-0.103155816	0.645288689
4	INCENP(S142);	0.270414436	-0.243416725	0.380882629
5	INCENP(T135);	0	0	0

1.1.2 KARP

- This test requires columns **2** and onwards to contain peptide intensity values for each identified phosphosite.

	1	2	3	4
1	gene;	AZD5438.3.PP2.1	AZD5438.4.PP2.2	DMSO.3
2	KDM1A(T59);	0.000216	0.000582	1.64017
3	INCENP(S148);	6.720432	7.867873	5.45268
4	INCENP(T145);	7.942432	9.098164	5.69186
5	INCENP(S143);	31.752216	37.470582	26.6093
6	INCENP(S142);	0.000324	0.000873	0.00051
7	INCENP(M136);	8.240108	21.65	14.4902
8	INCENP(S143);INCENP(M136);	1.52	2.31	0.8
9	INCENP(T135);	0.000216	0.000582	0.00034
10	EIF3J(S11);EIF3J(S13);	58.57	48.43	42.56
11	RPL4(353-364, no_mod);	0.000108	0.000291	0.00017
12	SAMD1(S161);	167.1	170.3	322.8

1.2 Analysis Parameters

- In order to upload a phosphoproteomic dataset, navigate to **Upload File** and select **Browse** to locate the .tsv file of interest.
- Click **Select kinase-substrate database** to choose the kinase-substrate relationships database for the dataset to be scanned against.
- Select a suitable KSEA algorithm from the **Select KSEA algorithm** dropdown list. Each algorithm can run against a single sample (e.g. one cell line) dataset or multiple samples (e.g. several cell lines) dataset. Single sample analysis returns interactive and downloadable result tables and a barplot of scores for each kinase. Multiple sample analysis returns a heatmap and downloadable .csv result files only.
- ***Generate a barplot/heatmap?*** allows to specify whether or not you would like to generate graphical kinase score results in the form of a barplot

(single sample analysis) or heatmap (multiple samples). Selecting **No** can speed up the analysis if the number of samples to analyse is large.

- ***Plot parameters*** - applicable to analyses with graphical output only (i.e. user has selected **Yes** to above). Minimum number of substrates determines which kinases to display on the barplot/heatmap. E.g. if 5 is selected, only kinases with 5 or more substrates will be displayed on the plots. KSEA results for the remaining kinases are calculated regardless, however they will not be shown on the plots.

2. KSEAPlus Results

2.1 Overview

The results section is divided into four main tabs:

- **Kinase Activity Data**: contains activity scores and related statistics for each identified kinase. This is presented as an interactive table for the single-sample analysis.
- **Kinase-Substrate Relationships**: contains additional information on identified unique kinase-substrate relationships, e.g. neighbouring residue sequences associated with a phosphosite, database source, and log2(FC)/intensity value for each phosphosite. This is presented in the form of an interactive table for single sample analysis.
- **Graphs:** if graphical output is selected, this is where a plot will appear.
- **Download**: contains downloadable kinase activity data (.csv), K-S Relationships data (.csv) and barplot/heatmap (.svg).

2.2 Kinase Activity Data

Each KSEAPlus algorithm returns kinase activity scores and associated information for each kinase identified in the database search. The results may differ depending on the algorithm chosen, as outlined below:

2.2.1 Z-Test

- **Kinase:** gene corresponding to the identified kinase.
- **Sub.Count:** the number of substrates from the experimental dataset found for the kinase of interest.
- **mnlog2(FC):** mean log2(FC) across all of the kinase's substrates.

- **zSc:** Z-score denoting the activity level of the kinase. If the raw FC has been calculated in the format of Experiment / Control, a positive Z-score signifies an upregulated kinase activity in relation to the control. The opposite becomes true for negative Z-scores.
- **pVal:** p-value calculated based on the z-score. It denotes the statistical significance of the obtained result.

2.2.2 Kolmogorov-Smirnov (KS)

- **Kinase:** gene corresponding to the identified kinase.
- **Sub.Count:** the number of substrates from the experimental dataset found for the kinase of interest.
- **mnlog2(FC):** mean log2(FC) across all of the kinase's substrates.
- **(+/-) KS:** KS test statistic. It is signed based on the mean log2(FC) of the kinase.
- **pVal:** p-value to denote the statistical significance.
- **(+/-) -log10(pVal):** -log10 of the p-value that is then signed based on the mean log2(FC) of the kinase.

2.2.3 KARP

- **Kinase:** gene corresponding to the identified kinase.
- **Sub.Count:** the number of substrates from the experimental dataset found for the kinase of interest.
- **Total.Sub.Count:** the total number of substrates found for the kinase in the database of choice.
- **Sum.Ints:** the sum of all intensities across the substrate group.
- **kSc:** K-score denoting the activity level of the kinase.

2.3 Kinase-Substrate Relationships

- **Kinase:** gene associated with the identified kinase.
- **Site:** phosphosite phosphorylated by the corresponding kinase.
- ***Site.Seq (+/- 7AA):** A sequence of seven amino acids on either side of the affected phosphosite. Reported for KSEAPlus analyses against PhosphoSitePlus database only.
- **Source:** database source of the reported information.
- ***log2(FC):** log2(FC) value from the dataset associated with the identified phosphosite. Applicable only to Z-test and KS analyses.
- ***Ints.:** raw peptide intensity value from the dataset associated with the identified phosphosite. Applicable only to KARP analyses.

NOTE: if multiple samples analysis is selected, the column name specified by the user for each sample is appended to the relevant result columns in a repeating manner, e.g.:

| Kinase	Sub.Count	mnlog2(FC).CAL101	zSc.CAL101	pVal.CAL101	mnlog2(FC).AZD6482	zSc.AZD6482	pVal.AZD6482
PLK1	1	0.645620068	0.118089438	0.452998396	-0.155576994	0.017348215	0.493079411
CDK1	4	-0.179403447	-0.074167351	0.470438612	-0.449554994	-0.085009667	0.466126853
CDK2	2	0.221989856	0.054322327	0.478339181	-0.216898798	0.006877713	0.497256211
CDK5	1	2.376822021	0.443699547	0.328629911	-2.625218176	-0.485463405	0.313673833
MTOR	1	1.596297096	0.296895874	0.383273003	0.129651354	0.075419861	0.469940329
DYRK1A	1	0.208895522	0.035948861	0.485661568	-0.261739751	-0.004266207	0.498298035
MAPK3	1	-3.173241944	-0.600174584	0.274194945	-0.902877684	-0.134799989	0.446385007

2.4 Graphs

All kinase scores displayed on the plots are in the form of the Z-score, K-score and -log10(p-value) for the Z-test, KARP and KS test, respectively.

2.4.1 Single-Sample Analysis

Single-sample analysis returns a horizontal barplot containing scores corresponding to each identified kinase.

- For the Z-test and KS test, the bars are coloured differently depending on the statistical significance of the scores, i.e. bright red denotes p-values smaller than 0.01, pink indicates p-values greater than or equal to 0.01 and grey indicates p-values greater than or equal to 0.05.

An example barplot for a single-sample Z-test is shown below:
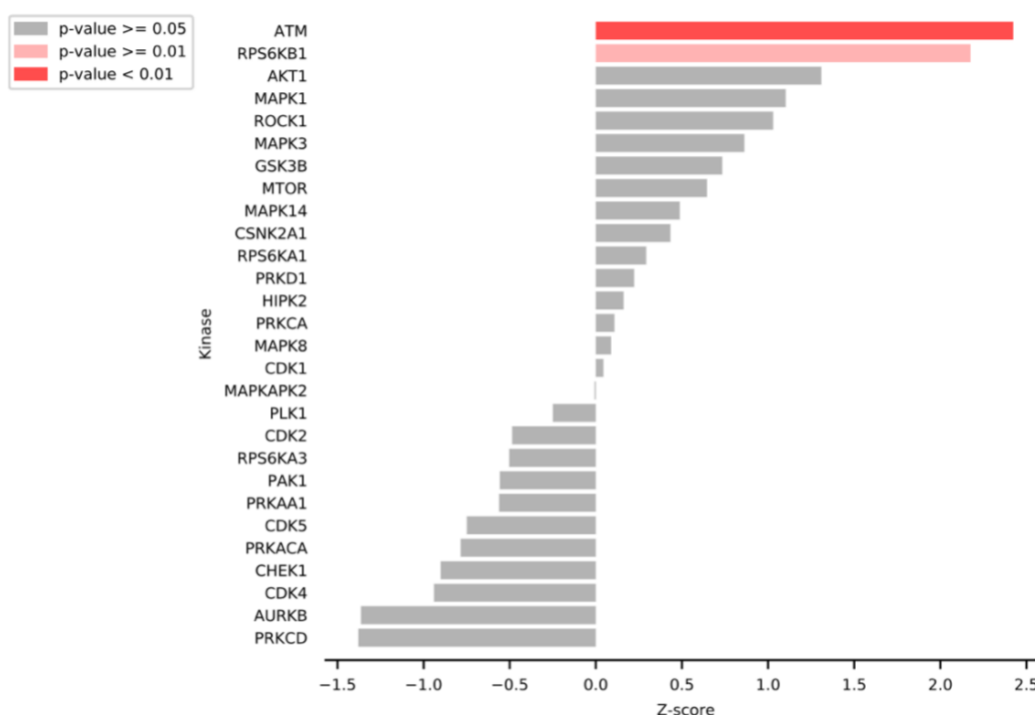
2.4.2 Multiple-Samples Analysis

Submitting a dataset containing multiple samples for analysis returns a heatmap of kinase scores corresponding to each kinase per each sample.

- **Asterisk annotation:** heatmap scores are further annotated for the Z-test and KS test if they are statistically significant (significance level 0.05). A

single asterisk (*) denotes a p-value smaller than 0.05 and a double asterisk (**) indicates a p-value smaller than 0.01. No annotation is available for the KARP test.

- **Colour bar**: is shown on the right hand-side of the heatmap and indicates the magnitude of the scores. Red-like colours are associated with higher scores and blue-like colours with lower scores.

An example heatmap for a multiple samples Z-test analysis is shown below:

	MCF7.AZD6482.fold	MCF7.CAL101.fold	MCF7.GDC0941.fold	MCF7.HS173.fold	MCF7.PIK294.fold	MCF7.TGX221.fold	MCF7.CX4945.fold	MCF7.Go6983.fold	MCF7.KN93.fold	MCF7.MK2206.fold
PLK1	-0.2	0.7	-0.1	0.4	-0.2	0.9	1.7*	1.3	0.8	2.0*
CDK1	0.0	1.8*	1.2	2.1*	1.3	1.3	-2.3*	-1.1	-1.7*	-2.4**
CDK2	-0.5	0.9	0.6	0.2	0.8	0.5	-0.5	-0.2	-0.4	-1.1
CDK5	-0.7	0.4	0.9	0.1	0.2	0.3	-1.6	-1.4	-0.6	-1.7*
MTOR	0.6	1.0	-2.5**	-0.8	0.9	0.9	0.5	0.4	0.2	-1.4
MAPK3	0.9	0.9	0.3	1.5	0.9	1.1	-1.7*	-1.1	-0.3	0.1
CSNK2A1	0.4	-0.9	1.2	1.2	-0.5	0.2	5.7**	4.6**	4.8**	6.2**
MAPK1	1.1	1.0	0.2	1.4	1.1	1.0	-1.3	-1.2	-0.6	-0.9

3. Warnings

- **File size limit:** current version of KSEAPlus accepts .tsv files up to 5 MB in size.
- **App idling:** KSEAPlus is running on a free Heroku dyno which become idle when the website has not received any traffic for 30 minutes. This tends to cause slow app loading times on the first request and, rarely, on the subsequent request.
- **Request timeout:** by default Heroku will terminate requests that have not received a response within 30 seconds. Frequently getting the request timeout error message may indicate that your network connection speed is slow or that the application is experiencing high web traffic or connectivity issues.
- **Downloads/session timeout:** for each user session, all of the generated results data is stored temporarily in a 25 MB in-memory server-side database (Heroku-Redis). To preserve space, each session expires within 5 minutes of obtaining results. Please ensure you download your results before your session expires. In addition, data from your previous session will be overwritten each time a dataset is analysed.

4. Useful References

Hernandez-Armenta, C., Ochoa, D., Gonçalves, E., Saez-Rodriguez, J. and Beltrao, P. (2017). Benchmarking substrate-based kinase activity inference using phosphoproteomic data. *Bioinformatics*, [online] 33(12), pp.1845-1851. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5870625/ [Accessed 5 Aug. 2019].

Hornbeck, P., Zhang, B., Murray, B., Kornhauser, J., Latham V., Skrzypek E. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res. 2015 43:D512-20.

Wiredja, D., Koyutürk, M. and Chance, M. (2017). The KSEA App: a web-based tool for kinase activity inference from quantitative phosphoproteomics. *Bioinformatics,* 33(21), pp.3489-3491.

Casado, P., Rodriguez-Prados, J., Cosulich, S., Guichard, S., Vanhaesebroeck, B., Joel, S. and Cutillas, P. (2013). Kinase-Substrate Enrichment Analysis Provides Insights into the Heterogeneity of Signaling Pathway Activation in Leukemia Cells. *Science Signaling*, 6(268).

Wilkes, E., Casado, P., Rajeeve, V. and Cutillas, P. (2017). Kinase activity ranking using phosphoproteomics data (KARP) quantifies the contribution of protein kinases to the regulation of cell viability. *Molecular & Cellular Proteomics*, 16(9).