Supplemental file - A curated resource for phosphosite-specific signature analysis

Supplemental Figure 1: Typical PTM-SEA workflow. The input to PTM-SEA is a *p* x *n* data matrix containing the abundance profiles of *p* phosphosites in *n* samples and can be obtained by processing raw MS data by computational tools such as Spectrum Mill or MaxQuant. The representation of phosphosites in these tables depends on the software used to derive the data tables as well as the protein database used to search MS/MS spectra (e.g. UniProt, RefSeq, Ensemble) and therefore a pre-processing step is typically required to re-format PTM site identifiers, to create single site-centric reports and to convert the data table into GCT v1.3 format. Pre-processing can be done in Excel or scripting languages like R, python, perl or similar. In order to support different protein sequence databases and to guarantee stable representation of PTM sites, PTMsigDB was assembled in three different formats for PTM site representation (**Table 1**). The UniProt-centric format can readily be used if MS/MS spectra were queried against a UniProt protein sequence database. For other sources of protein sequence databases we provide the flanking amino acid sequence (+/- 7 residues) around the modified residue (center position) as well as the PSP site group identifier, which represents a stable id across different database builds, within protein families (isoforms) and across species. We recommend to use the flanking sequence (+/- 7 AA) as identifier for PTM sites which are less susceptible to updates of protein accession numbers or updated annotation of protein N-termini which alter residue numbers. The gene cluster text format (GCT) in version 1.3 is a tab-delimited text file format for matrix-type datasets and allows metadata about an experiment to be stored alongside with the data. To conveniently convert a data table into a GCT file we recommend the cmapR R-package (https://github.com/cmap/cmapR) and the matrix manipulation tool Morpheus (https://software.broadinstitute.org/morpheus). PTM-SEA is implemented in R and can be run from command line or in RStudio (1). In addition, we made PTM-SEA available in GenePattern (2) enabling users to run PTM-SEA directly in a web browser. The code base of PTM-SEA is shared with the ssGSEA2.0 implementation (https://github.com/broadinstitute/ssGSEA2.0). If PTMsigDB is specified as database, the program will automatically perform PTM-SEA. The program will apply PTM-SEA to each sample column separately. The results are exported in GCT format and comprise signature enrichment scores, corresponding p-values and FDR-corrected p-values. These results can be directly imported into Morpheus for visualization purposes.

Supplemental Figure 2: Phosphorylation sites in PTMsigDB. A) Number of perturbagen signatures (y-axis) semi-automatically derived from PhosphoSitePlus (PSP) as a function of minimal signature length (x-axis). The line charts depict different values for the minimal number of studies (PMIDs) each site was required to be consistently reported in order to be part of a signature. Highlighted are the parameters used in this study. **B)** UpSet-plot (3) comparing phosphorylation sites in three signature categories in PTMsigDB. Horizontal bars depict the total number of sites in each category, vertical bars illustrate the number of intersecting sites as indicated in the set matrix below. **C)** Venn diagram comparing phosphorylation sites in PTMsigDB to regulatory phosphorylation sites of human origin derived from PhosphoSitePlus (PSP). **D)** Barchart depicting the distribution of serine, threonine and tyrosine sites in PTMsigDB.

Supplemental Figure 3: Comparison of kinase enrichment scores jointly detected by PTM-SEA and KSEA. Enrichment scores of kinase signatures calculated on the same set of phosphosites from the Sharma *et al.* dataset *(4)* by PTM-SEA (x-axis) and KSEA App (https://casecpb.shinyapps.io/ksea/) (5) (y-axis). The different panels correspond to replicate measurements of three experimental conditions: treatment with DMSO (red), EGF (blue) and nocodazole (green).

Supplemental Figure 4: Site-centric vs. gene-centric signature enrichment analysis. A) Clustering of site-centric normalized signature enrichment scores (NES) into three clusters partitioned samples into DMSO, EGF and nocodazole treatment while grouping replicate measurements together. Clustering was based on signatures detected in both approaches. **B)** Clustering of gene-centric NES resulted in a mixed cluster containing nocodazole and DMSO-treated samples. **C)** Box plots depicting normalized enrichment scores (NES) of the site-centric EGF signature set (PERT-PSP_EGF) across replicate measurements in DMSO, EGF and nocodazole treated samples. Numbers at the bottom indicate the median NES and FDR across replicate measurements. **D)** Box plots depicting NES of the gene-centric EGF signature set.

Supplemental Figure 5: Gene-centric-redundant signature enrichment analysis. A) Hierarchical clustering of normalized enrichment scores (NES) into three clusters resulted in a mixed cluster containing nocodazole and DMSO-treated samples, similar to the gene-centric approach. **B)** Silhouette analysis of the hierarchical clustering results. The bar chart depicts silhouette scores (x-axis) of each sample (y-axis) colored according to the assigned cluster. Average silhouette scores of each custer are depicted at the right side of the bar charts. **C)** Box plots depicting normalized enrichment scores (NES) of the gene-centric-redundant EGF signature set across replicate measurements in DMSO, EGF and nocodazole treated samples. Numbers indicate minimal achieved FDRs. **D)** Box plots depicting NES of the gene-centric-redundant nocodazole signature set. Numbers indicate median NES and FDR.

Supplemental Figure 6: Correlation between phosphosite ratios before and after normalization to protein expression. Log2-transformed phosphosite ratios (BYL719/DMSO) were normalized by subtracting the corresponding protein log2 ratio. The panels show correlation of phosphosite ratios before (x-axis) and after (y-axis) normalization in three replicate measurements in the 6h and 24 time point, respectively.

Supplemental Figure 7: Reproducibility of normalized enrichment scores. A) Hierarchical clustering and **B)** pairwise correlation analysis of normalized enrichment scores (NES) demonstrated high reproducibility of NES between triplicate measurements in each time point. Boxes group replicate measurements of the same time point; numbers indicate pearson correlation coefficients.

Supplemental Figure 8: Phosphosite abundance levels of signatures affected by PI3K inhibition. The heatmaps show normalized log2 TMT ratios of individual phosphorylation sites that contribute to a signature set. T47D cells were treated with BYL719 for 6h and 24h and experiments were conducted in triplicate. Metadata tracks on top depict the time points, signature scores and FDR-corrected p-values of the corresponding signature. The 'signature.direction' data track at the right illustrate the annotated direction of change of each site in the signature. **A)** Kinase-substrate signature of CK2A1 kinase. **B)** Leptin pathway signature curated from NetPath. **C)** PI3K-Akt pathway obtained from WikiPathways. Heatmaps were created in Morpheus (`https://software.broadinstitute.org/morpheus/`).

References

1. RStudio Team (2016) RStudio: Integrated Development Environment for R.

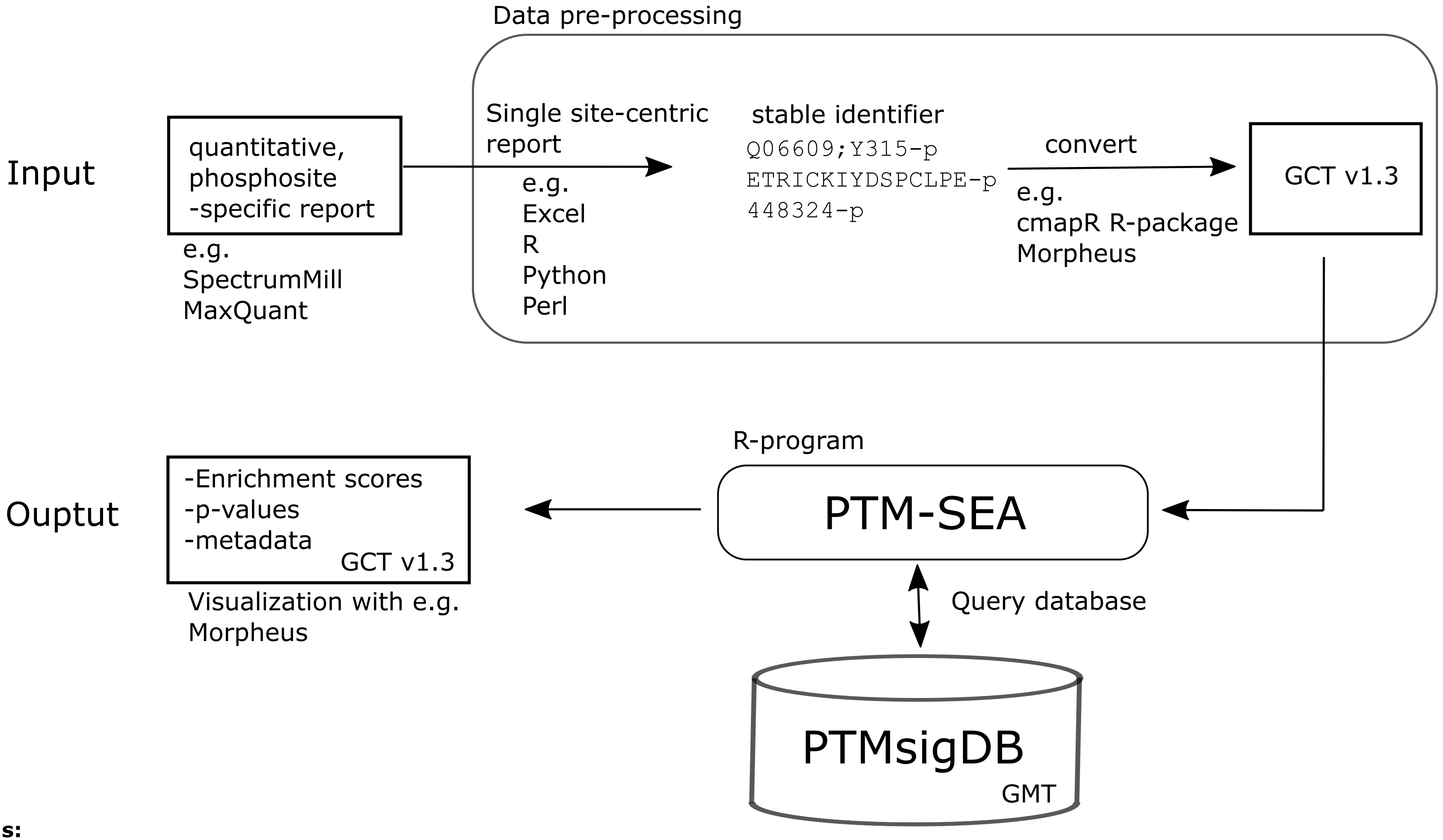
2. Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P., and Mesirov, J. P. (2006) GenePattern 2.0. *Nat. Genet.* 38, 500–501

3. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014) UpSet: Visualization of Intersecting Sets. *IEEE Trans. Vis. Comput. Graph.* 20, 1983–1992

4. Sharma, K., D'Souza, R. C. C. J., Tyanova, S., Schaab, C., Wiśniewski, J. J. R., Cox, J., and Mann, M. (2014) Ultradeep Human Phosphoproteome Reveals a Distinct Regulatory Nature of Tyr and Ser/Thr-Based Signaling. *Cell Rep.* 8, 1583–1594

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

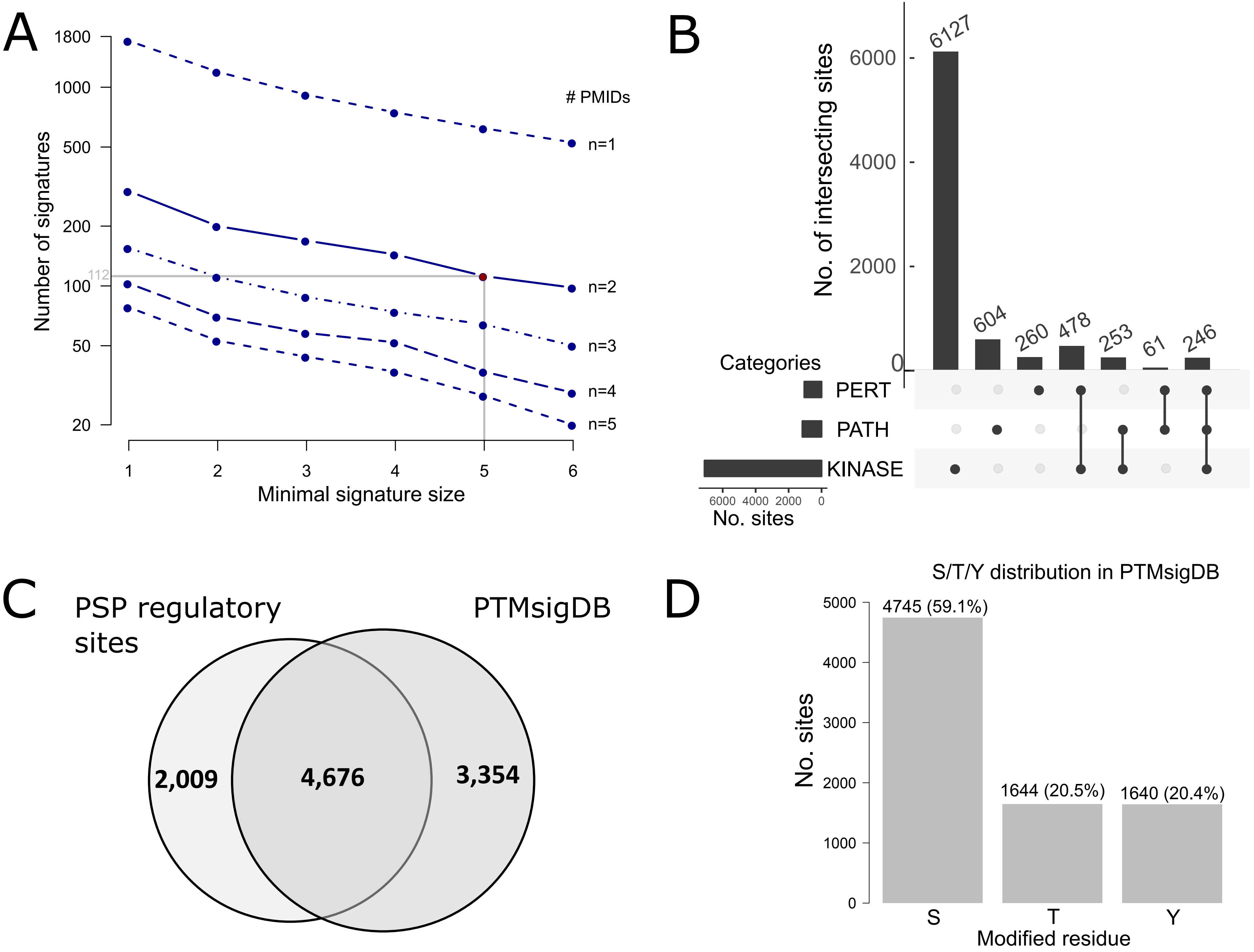
Supplemental Figure 1: PTM-Signature Enrichment Analysis



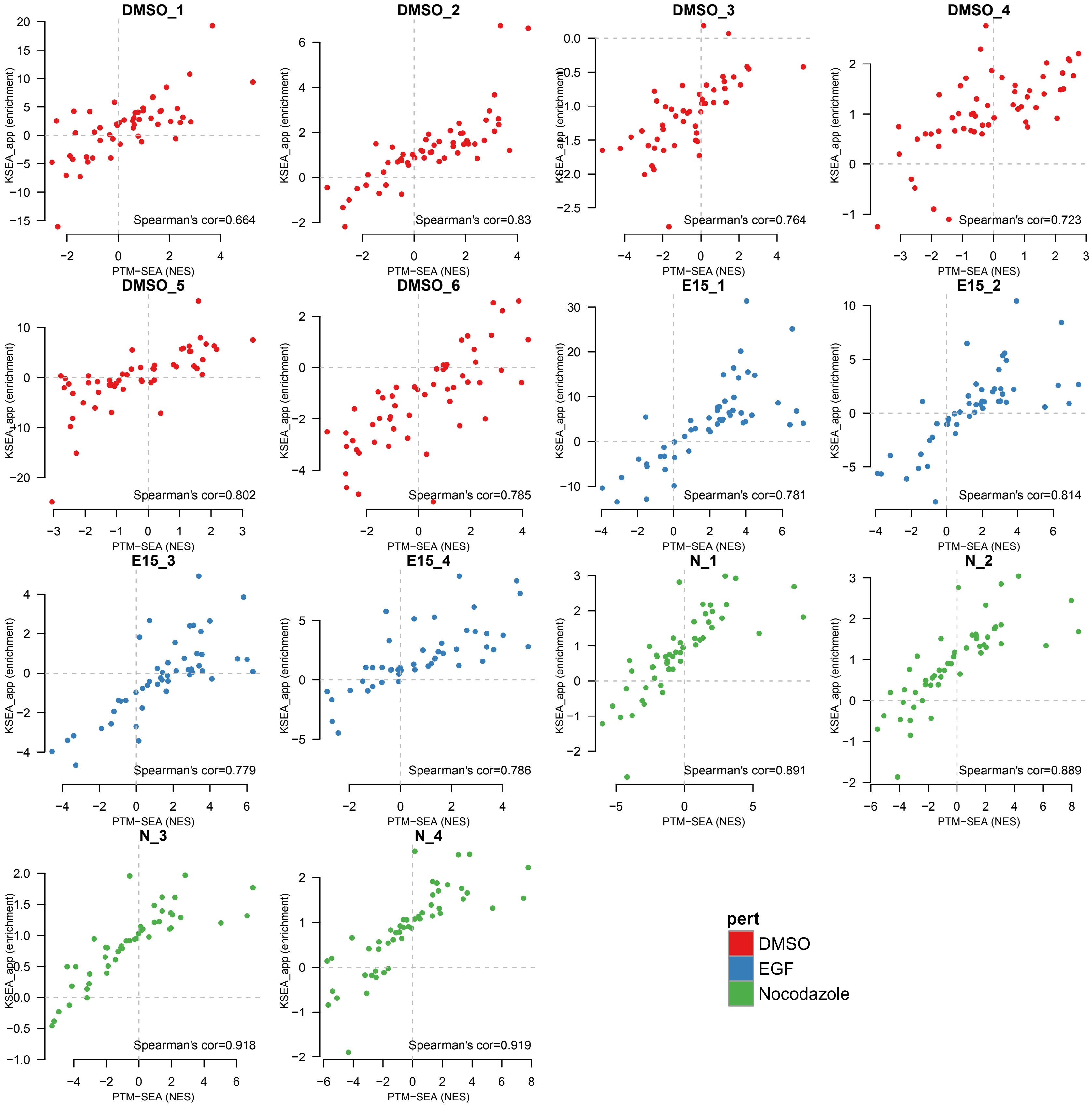
Tools:
Spectrum Mill (https://www.agilent.com/en/products/software-informatics/masshunter-suite/masshunter-for-life-science-research/spectrum-mill)
MaxQuant (http://www.biochem.mpg.de/5111795/maxquant)
cmapR (https://github.com/cmap/cmapR)
Morpheus (https://software.broadinstitute.org/morpheus/)

Data formats:
GCT - Gene Cluster Text (https://clue.io/connectopedia/gct_format)
GMT - Gene Matrix Transposed (https://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats)

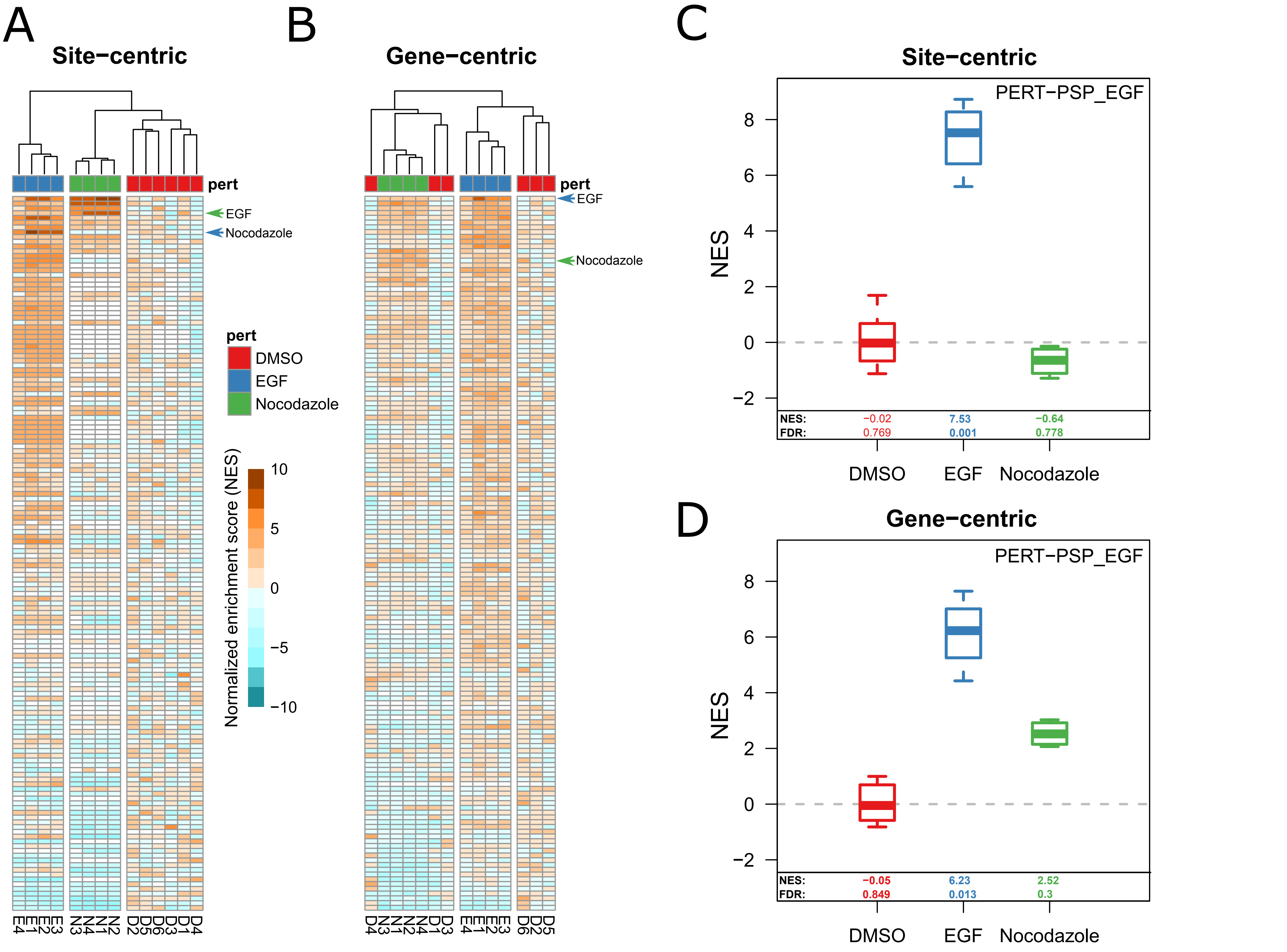
Supplemental Figure 2: Phosphorylation sites in PTMsigDB



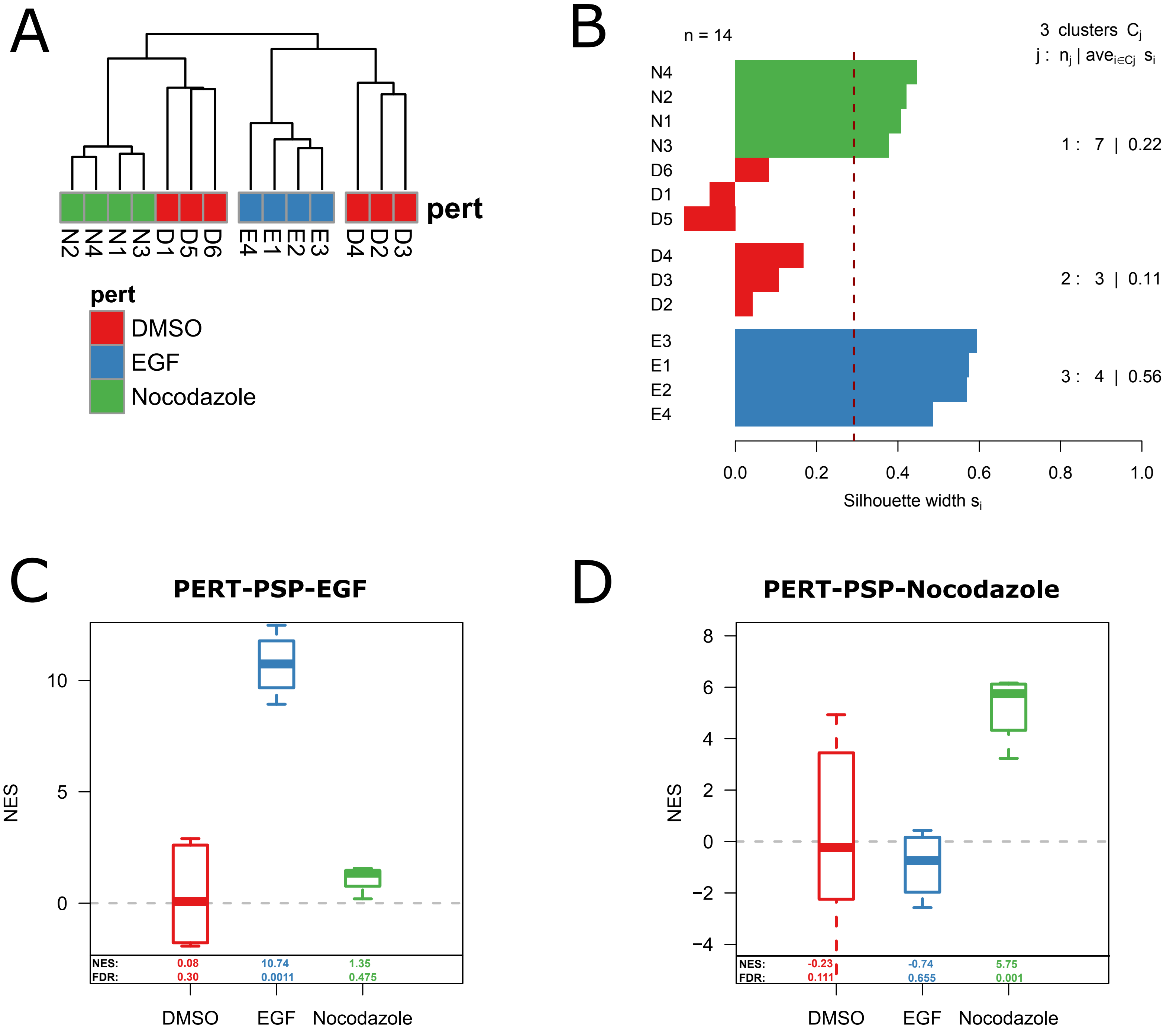
Supplemental Figure 3: Comparison of kinase enrichment scores jointly detected by PTM-SEA and KSEA



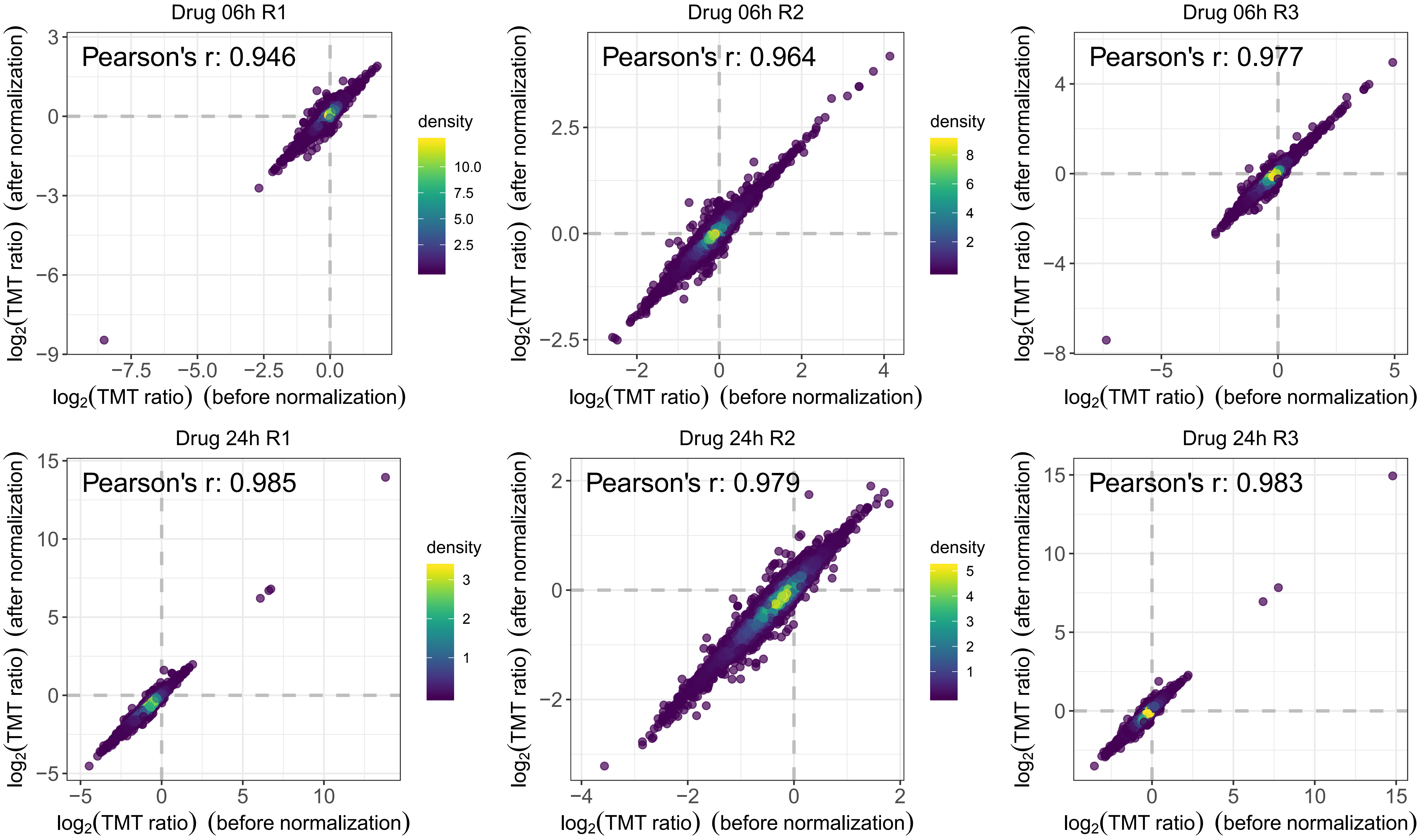
Supplemental Figure 4: Site-centric vs. gene-centric signature set enrichment analysis



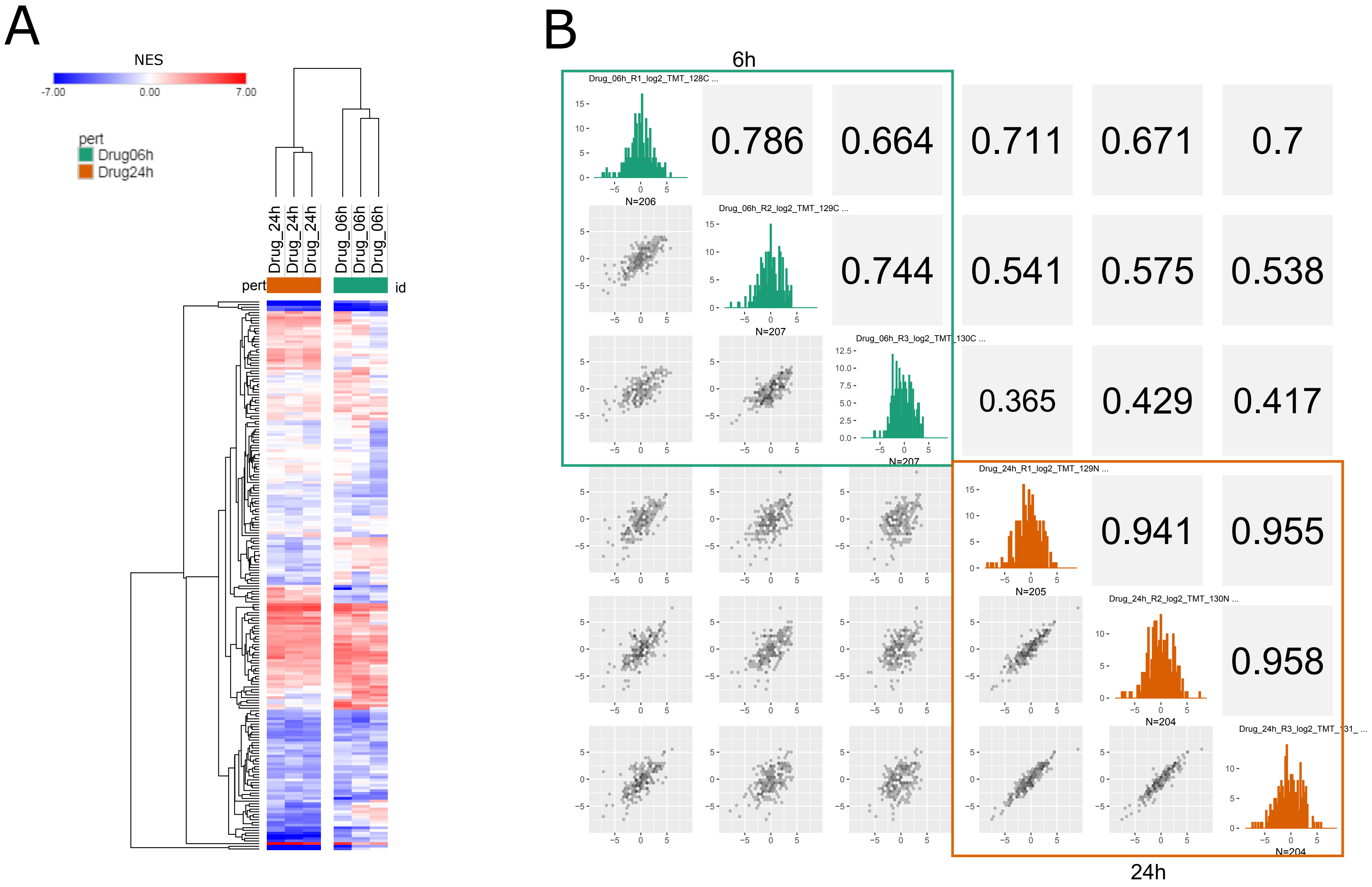
Supplemental Figure 5: Gene-centric-redundant signature enrichment analysis



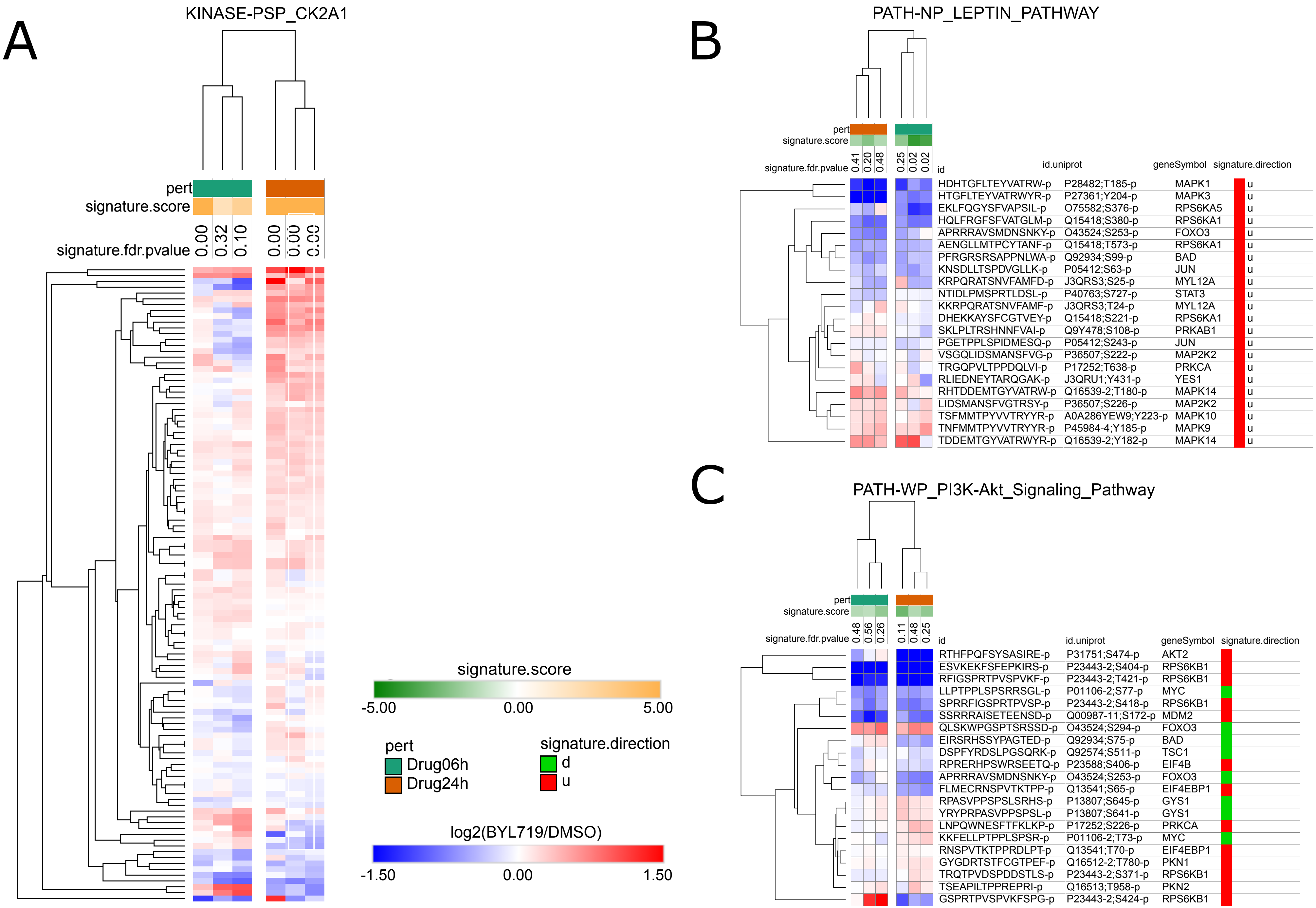
Supplemental Figure 6: Correlation between phosphosite ratios before and after normalization to protein expression



Supplemental Figure 7: Reproducibility of normalized enrichment scores



Supplemental Figure 8: Phosphosite abundance levels of signatures affected by PI3K inhibition.


A
KINASE-PSP_CK2A1
pert
signature.score
signature.fdr.pvalue
signature.score
pert
Drug06h
Drug24h
signature.direction
d
u
log2(BYL719/DMSO)
B
PATH-NP_LEPTIN_PATHWAY
pert
signature.score
signature.fdr.pvalue
id
id.uniprot
geneSymbol
signature.direction
HDHTGFLTEYVATRW-p P28482;T185-p MAPK1 u
HTGFLTEYVATRWYR-p P27361;Y204-p MAPK3 u
EKLFQGYSFVAPSIL-p O75582;S376-p RPS6KA5 u
HQLFRGFSFVATGLM-p Q15418;S380-p RPS6KA1 u
APRRRAVSMDNSNKY-p O43524;S253-p FOXO3 u
AENGLLMTPCYTANF-p Q15418;T573-p RPS6KA1 u
PFRGRSRSAPPNLWA-p Q92934;S99-p BAD u
KNSDLLTSPDVGLLK-p P05412;S63-p JUN u
KRPQRATSNVFAMFD-p J3QRS3;S25-p MYL12A u
NTIDLPMSPRTLDSL-p P40763;S727-p STAT3 u
KKRPQRATSNVFAMF-p J3QRS3;T24-p MYL12A u
DHEKKAYSFCGTVEY-p Q15418;S221-p RPS6KA1 u
SKLPLTRSHNNFVAI-p Q9Y478;S108-p PRKAB1 u
PGETPPLSPIDMESQ-p P05412;S243-p JUN u
VSGQLIDSMANSFVG-p P36507;S222-p MAP2K2 u
TRGQPVLTPPDQLVI-p P17252;T638-p PRKCA u
RLIEDNEYTARQGAK-p J3QRU1;Y431-p YES1 u
RHTDDEMTGYVATRW-p Q16539-2;T180-p MAPK14 u
LIDSMANSFVGTRSY-p P36507;S226-p MAP2K2 u
TSFMMTPYVVTRYYR-p A0A286YEW9;Y223-p MAPK10 u
TNFMMTPYVVTRYYR-p P45984-4;Y185-p MAPK9 u
TDDEMTGYVATRWYR-p Q16539-2;Y182-p MAPK14 u
C
PATH-WP_PI3K-Akt_Signaling_Pathway
pert
signature.score
signature.fdr.pvalue
id
id.uniprot
geneSymbol
signature.direction
RTHFPQFSYSASIRE-p P31751;S474-p AKT2
ESVKEKFSFEPKIRS-p P23443-2;S404-p RPS6KB1
RFIGSPRTPVSPVKF-p P23443-2;T421-p RPS6KB1
LLPTPPLSPSRRSGL-p P01106-2;S77-p MYC
SPRRFIGSPRTPVSP-p P23443-2;S418-p RPS6KB1
SSRRRAISETEENSD-p Q00987-11;S172-p MDM2
QLSKWPGSPTSRSSD-p O43524;S294-p FOXO3
EIRSRHSSYPAGTED-p Q92934;S75-p BAD
DSPFYRDSLPGSQRK-p Q92574;S511-p TSC1
RPRERHPSWRSEETQ-p P23588;S406-p EIF4B
APRRRAVSMDNSNKY-p O43524;S253-p FOXO3
FLMECRNSPVTKTPP-p Q13541;S65-p EIF4EBP1
RPASVPPSPSLSRHS-p P13807;S645-p GYS1
YRYPRPASVPPSPSL-p P13807;S641-p GYS1
LNPQWNESFTFKLKP-p P17252;S226-p PRKCA
KKFELLPTPPLSPSR-p P01106-2;T73-p MYC
RNSPVTKTPPRDLPT-p Q13541;T70-p EIF4EBP1
GYGDRTSTFCGTPEF-p Q16512-2;T780-p PKN1
TRQTPVDSPDDSTLS-p P23443-2;S371-p RPS6KB1
TSEAPILTPPREPRI-p Q16513;T958-p PKN2
GSPRTPVSPVKFSPG-p P23443-2;S424-p RPS6KB1