Published online 22 May 2019 *Nucleic Acids Research, 2019, Vol. 47, Web Server issue* ***W199–W205***
doi: 10.1093/nar/gkz401

WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs

Yuxing Liao [1], Jing Wang[1], Eric J. Jaehnig[1], Zhiao Shi[1] and Bing Zhang[1,2,*]

[1]Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, TX 77030, USA and [2]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA

Received March 04, 2019; Revised April 23, 2019; Editorial Decision May 01, 2019; Accepted May 01, 2019

ABSTRACT

WebGestalt is a popular tool for the interpretation of gene lists derived from large scale -omics studies. In the 2019 update, WebGestalt supports 12 organisms, 342 gene identifiers and 155 175 functional categories, as well as user-uploaded functional databases. To address the growing and unique need for phosphoproteomics data interpretation, we have implemented phosphosite set analysis to identify important kinases from phosphoproteomics data. We have completely redesigned result visualizations and user interfaces to improve user-friendliness and to provide multiple types of interactive and publication-ready figures. To facilitate comprehension of the enrichment results, we have implemented two methods to reduce redundancy between enriched gene sets. We introduced a web API for other applications to get data programmatically from the WebGestalt server or pass data to WebGestalt for analysis. We also wrapped the core computation into an R package called WebGestaltR for users to perform analysis locally or in third party workflows. WebGestalt can be freely accessed at http://www.webgestalt.org.

INTRODUCTION

Functional enrichment analysis plays a critical role in interpreting high-throughput experiment results, which frequently generate a list of interesting genes or proteins. WebGestalt (WEB-based GEne SeT AnaLysis Toolkit) is one of the most widely used gene set enrichment analysis tools that help users extract biological insights from genes of interest (1). Following updates in 2013 and 2017 (2,3), WebGestalt had 27 000 unique users per year according to Google Analytics and has been cited in 739 papers indexed in Scopus since 2017.

Since the last update in NAR, we continued to improve the usability and accessibility of WebGestalt. The functional annotations and ID mappings were updated with new additions, including support for the analysis of phosphoproteomics data. While no change has been made to the input user interface (UI), we have completely redesigned the output UI. Several new interactive plots were added to help users visualize the enrichment results, and the high-quality static images can be used directly in presentations or publications. In addition to improved visualization, we have implemented two computational methods to identify non-redundant, representative subsets of significant functional categories, which is particularly useful when the number of significant functional categories is overwhelming. To expand the audience of WebGestalt from biologists and clinical researchers to computational biologists and software developers, we have implemented core computation in an R package called WebGestaltR. WebGestaltR supports all three methods available on the WebGestalt server, namely Over Representation Analysis (ORA) (4), Gene Set EnrichmentAnalysis (GSEA) (5) and Network Topology-based Analysis (NTA) (6). It allows users to locally generate the same results as those from the WebGestalt server and enables batch runs and integration into other pipelines. A web application programming interface (API) is also provided to enable programmatic access to the underlying data and the job submission form.

DATA UPDATE AND SUPPORT FOR PHOSPHORYLATION SITE ENRICHMENT

WebGestalt now supports 342 ID types from 12 organisms. The total number of functional categories compiled from public databases and computational analyses is 155 175 (Table 1). Although the increase in numbers mainly results from updates to upstream data sources, a few new databases were added in WebGestalt 2019 (Table 1), including a cancer-related subset of WikiPathways (7), co-expression network modules derived from cancer proteomics datasets from CPTAC (Clinical Proteomic Tumor Analysis Consortium), the protein complex database CORUM (8), the disease phenotype database OMIM (9), and kinase target genes compiled from PhosphoSitePlus (10). Note that, by uploading custom functional annotations in GMT files, users can perform enrichment analysis on any organism and dataset in principal.

*To whom correspondence should be addressed. Tel: +1 713 798 1443; Fax: +1 713 798 1693; Email: bing.zhang@bcm.edu

© The Author(s) 2019. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Table 1. Comparison of the functional categories supported by WebGestalt 2017 and 2019

Class	Database	Number of categories		Data source
		2017	2019	
Gene ontology	Biological process	19 239*	20 121	http://www.geneontology.org
	Cellular component	2317*	2537	
	Molecular function	5946*	6008	
Pathway	KEGG	2677	2815	https://www.kegg.jp/
	WikiPathways	1345	1466	https://www.wikipathways.org
	Reactome	11 804	15 019	https://www.reactome.org/
	PANTHER†	1336	1323	http://www.pantherdb.org/
Network	Hierarchical mRNA co-expression modules	30 852	30 852	Firehose (http://gdac.broadinstitute.org/)
	Hierarchical protein co-expression modules	NA	409	CPTAC data portal (https://cptac-data-portal.georgetown.edu/cptacPublic/)
	Hierarchical protein interaction modules	2979	3625	BioGrid (https://thebiogrid.org/)
	MicroRNA target	2210	2431	MSigDB (http://software.broadinstitute.org/gsea/msigdb/index.jsp)
	Transcript factor target	6150	6765	MSigDB (http://software.broadinstitute.org/gsea/msigdb/index.jsp)
	Kinase target genes	NA	363	PhosphoSitePlus (https://www.phosphosite.org)
	Kinase target phosphosites	NA	716	RegPhos (http://140.138.144.141/~RegPhos/)
	PTMsigDB	NA	692	https://github.com/broadinstitute/ssGSEA2.0
	CORUM	NA	3380	http://mips.helmholtz-muenchen.de/corum/
Disease	DisGeNET	7607	13 074	http://www.disgenet.org/
	GLAD4U	2997	3071	http://glad4u.zhang-lab.org
	OMIM	NA	5169	https://www.omim.org/
Drug	DrugBank	4831	5825	https://www.drugbank.ca/
	GLAD4U	2355	2454	http://glad4u.zhang-lab.org
Phenotype	Phenotype	16 531*	18 720	Human Phenotype Ontology (http://www.human-phenotype-ontology.org); Mammalian Phenotype Ontology (http://www.informatics.jax.org)
Chromosomal location	Cytogenetic band	6574	7149	Entrez Gene (http://www.ncbi.nlm.nih.gov/gene)
Others	User uploaded and community contributed datasets	NA	1191	
Total		127 750	155 175	

*The numbers have been revised compared with original publication. Specifically, categories without annotated genes were excluded in the revised counts.
†The decrease is due to updated curation in PANTHER.

One notable addition in the new version is support for the site-level analysis of phosphoproteomics data. Phosphorylation is one of the most studied post-translation modifications and plays a crucial role in most cellular processes. With technological advances in mass spectrometry, phosphoproteomic studies can now identify phosphorylation sites at a much larger scale (11), but tools and annotation datasets for analyzing phosphorylation sites (phosphosites) are still limited. Phosphosites usually have to be aggregated at the gene level and information for individual sites is inevitably lost (12). To evaluate kinase substrate enrichment within phosphosite level datasets, one approach has been to use sets of phosphosites that are known substrates for kinases for GSEA (13). As a unique new feature of WebGestalt 2019, we compiled datasets of kinases and their target phosphosites from RegPhos v2.0 database (14) for human, mouse and rat, which contains annotations for 377, 218 and 121 kinases for each organism, respectively. Recently, a database of post-translation modification (currently only phosphorylation) signatures of kinase substrates, perturbations and pathways, PTMsigDB, was published (12), and the GMT files from version 1.8.1 were processed and included. In total, the annotations include 7805 human, 1229 mouse and 702 rat phosphosites. Phosphosite enrichment analysis is performed essentially the same way as for gene-centric datasets, except for the use of site IDs rather than gene IDs in the input Gene List and the curated GMT files. Site ID annotation as the 15-mer sequence motif with the phosphosite in the middle or as the protein ID from Uniprot, Ensembl or RefSeq followed by the amino acid and its position (e.g. P12956_S51) is supported for the input for ORA and GSEA analysis. Different phosphorproteomic methods have different levels of sensitivity and biases and some peptides are hard to detect at all. This will not affect the GSEA

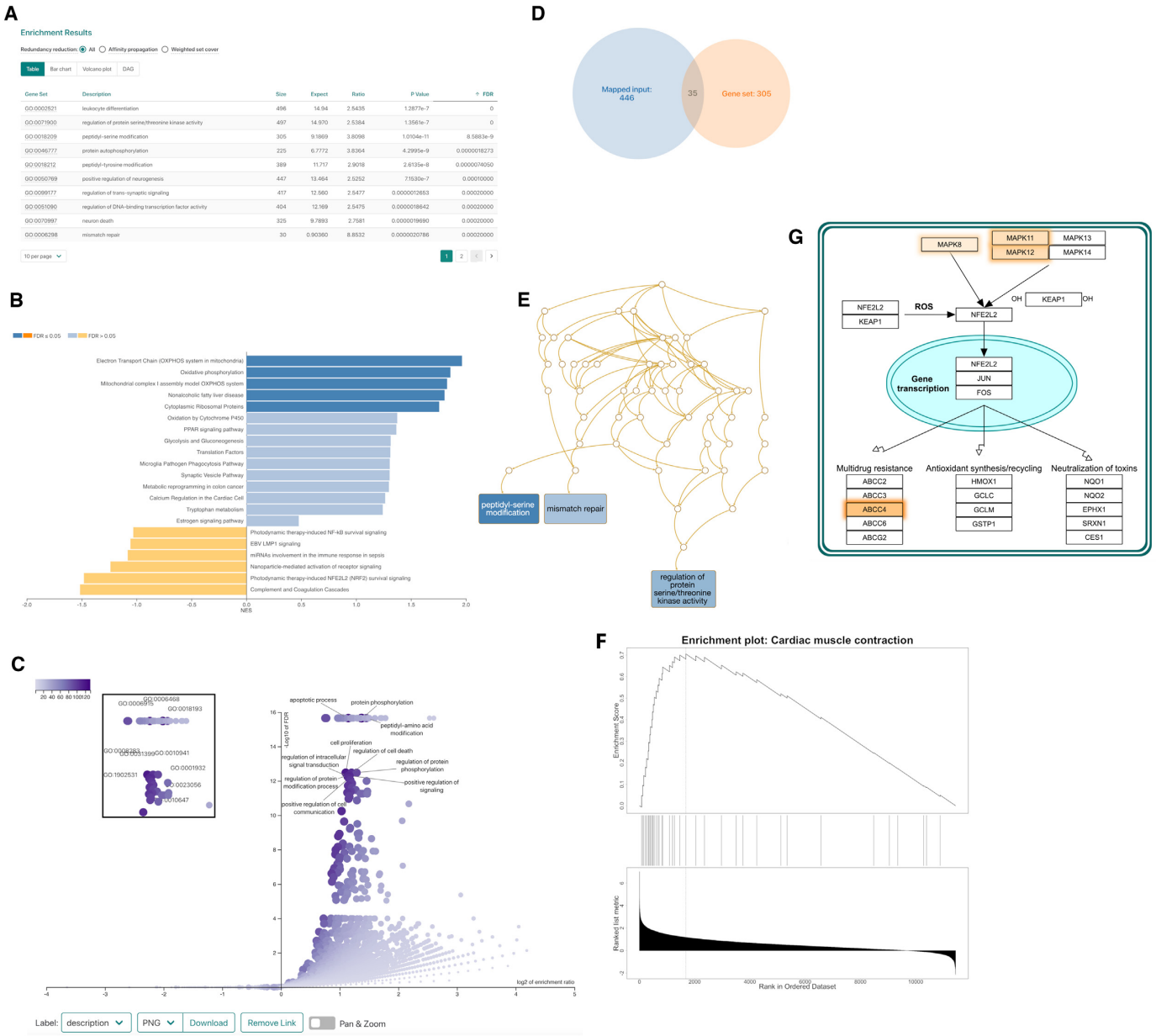

Figure 1. New and improved visualizations in the result page of WebGestalt 2019. (**A**) Table summary of significant results. The options of choosing reduced subsets are also shown. (**B**) Bar chart shows enrichment ratio or NES of results with direction. (**C**) Customizable volcano plot. Inset shows initial layout for comparison. (**D**) New implementation of DAG. Now it is more compact highlighting enriched nodes. (**E**) Venn diagram shows the overlap between the gene set in the input and in the reference. (**F**) GSEA enrichment plot. (**G**) Pathway view of WikiPathways highlighting leading edge genes based on score.

method because only quantifiable sites can be included in the input rank list. For ORA analysis, it is important to include only quantifiable sites as the background data set.

IMPROVEMENT OF OUTPUT REPORT AND VISUALIZATION

Significant effort was devoted to improving the results page since last update. Now the output for ORA and GSEA is divided into two major sections: summary and enrichment results. While the summary section keeps the job parameters and the GO Sim summary of the input genes from the previous version, this section is now collapsed by default to highlight the enrichment results section. Links to downloading the ZIP file with the HTML report and text files of all results and the functional annotation database in GMT format are available.

In the enrichment results section, different visualizations of results are arranged in tabs, and detailed information of the functional category selected in the view is listed below the visualization. We introduced a few more visualizations to help users understand the enrichment results. Most plots are implemented in the browser with JavaScript and can be downloaded in SVG and high-resolution PNG formats. We designed the visualizations with the hope that they would be readily adopted for publication. In the 2017 version de-

scribed in the last update paper, the significant gene sets in the enrichment results were only summarized in a table or shown in a directed acyclic graph (DAG) when the functional categories in the search database have a hierarchical structure. In the 2019 version, this table has been enhanced with more information and can be sorted by scores and statistics by clicking on the headers (Figure 1A). The number of rows in the table per page can be also adjusted. Like in the other visualizations, clicking the gene set name will populate the section below with detailed information about the category. In addition, we added bar charts and volcano plots as new components and improved the DAG.

The bar chart plots the enrichment results vertically with the bar width equal to enrichment ratio in ORA or to the normalized enrichment score (NES) in GSEA (Figure 1B). The direction and color of the bars also indicated the direction of the association of the enriched category from GSEA. When top results are chosen to be returned and the false discovery rate (FDR) for the categories is $\leq$0.05, the colors of the bars are in darker shade than when the FDR exceeds 0.05. The height of the chart can adjust to include all significantly enriched sets. Clicking on the bar will show detailed information for the enriched category in the section below.

The volcano plot is implemented with interactivity and customization in mind. It shows the log of the FDR versus the enrichment ratio or NES for all the functional categories in the database, highlighting the degree by which the significant categories stand out from the background (Figure 1C). The size and color of the dot is proportional to the number of overlapping (for ORA) or leading edge genes (for GSEA) of the category. Hovering over a dot will show more information about the category, and clicking on it will update the detailed information section. Users can pan and zoom the plot to adjust for a better view or focus in on a region. The significantly enriched categories are labeled, and the labels are positioned automatically by a force field-based algorithm at startup. Users can further adjust the positions manually by dragging the label around with mouse. The label can be switched between gene set names and IDs and a line linking the label and its data point can be drawn. Users can save the plot in SVG or PNG format after fine tuning in the bottom toolbox. The inset in Figure 1C shows an example of initial label layout compared to the layout of the same plot after manual adjustment.

If the functional database contains an inherent DAG structure of functional categories, such as GO terms, the relationship of the enriched gene sets can be visualized in another tab (Figure 1D). We have changed the implementation of the DAG from the obsolete Cytoscape Web library to the Cytoscape.js library (15), removing the need for the Flash plugin. The new version utilizes the full width of the browser window compared with the half in the previous version, since the DAG tends to be much wider when a loose threshold is specified. We also minimized the style of the nodes connecting the significant categories to the root to make the DAG more compact by default. When the DAG is still too large to perform well in the browser, users could export the layout in Cytoscape JSON format and open it in the desktop version of Cytoscape. Options and functions for downloading and resizing can be found in the right-click menu. Clicking on the nodes of significant categories will select the category to view the detailed information.

The detailed information section first tallies score statistics and has links to external databases and for downloading the gene table. The section can be updated to show a category of interest either by clicking on corresponding elements in the plots, directly searching or selecting through the dropdown box. The gene table lists the overlapping or leading edge genes with their gene symbols, names and links to NCBI and can be sorted by clicking headers. The number of rows per page can be adjusted. For ORA, a Venn diagram illustrates the overlap between the genes in the input and in the geneset (Figure 1E). For GSEA, it is replaced by an enrichment plot showing the rank distribution, the running sum and the position of the peak of the running sum, similar to the plot made by the original GSEA Java program (Figure 1F).

For analysis using pathway databases, external links lead to database's viewer for the pathway. For the WikiPathways (7) and KEGG (16) databases, the overlapping/leading edge genes are highlighted, and the input scores for the genes are also used to specify a color gradient for the leading edge genes when GSEA is run against the WikiPathways database (Figure 1G).

The NTA results page was similarly refreshed with Cytoscape.js. The two side-by-side graphs show the retrieved or expanded network on the left and the DAG of GO Biological Process enrichment results on the right. Below are several tables of seeds and information about enriched GO terms. The enriched GO terms in the table and in the DAG are linked, and selecting one will highlight corresponding genes in the network.

REDUNDANCY REDUCTION OF ENRICHED GENE SETS

Depending on the internal redundancy between the functional categories in a given database and the significance threshold of the job, the number of enriched functional categories in the results can sometimes be overwhelming, even with the improved interface. Thus, we included a post-processing step in WebGestalt to identify the most representative significant gene sets for visualization with two methods of redundancy reduction, affinity propagation (17) and weighted set cover (18). In short, affinity propagation clusters gene sets using the Jaccard index as a similarity measurement and automatically identifies an 'exemplar' or representative for each cluster with priority for sets with significant *P*-values. Weighted set cover finds a minimum subset of gene sets that can cover all the genes from the enriched sets, while the weight or cost of adding a set is associated with its *P*-value. Weighted set cover may stop before convergence when the input parameter limiting the expected set number is reached.

To illustrate how much redundancy reduction can help with analyses using common databases, we used the example gene list on the WebGestalt website to perform ORA against several databases with FDR thresholds of 0.1, 0.05 and 0.01 and compared the numbers of enriched results under different thresholds and with redundancy reduction in Figure 2. There are many fewer gene sets after redundancy

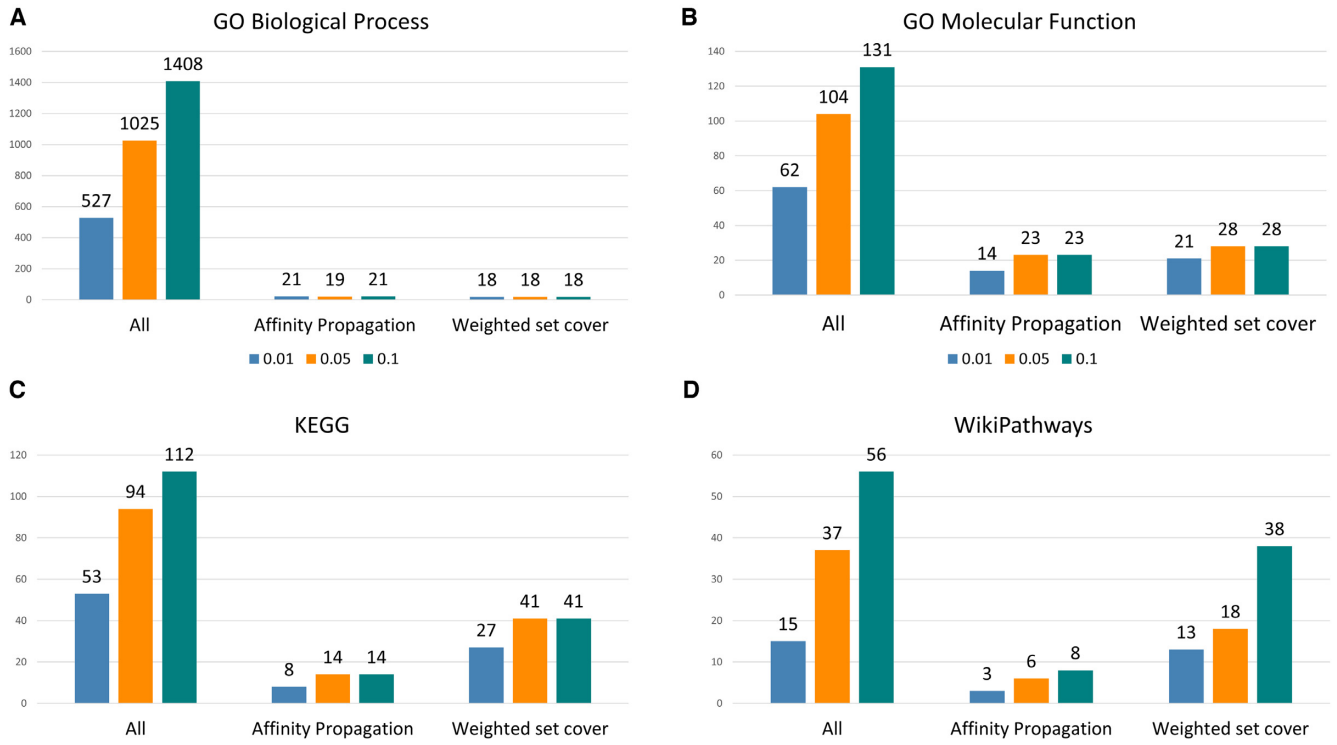

Figure 2. The number of all significant results are compared with the numbers with redundancy reduction regarding different thresholds and datasets including (A) Gene Ontology (GO) biological process, (B) GO molecular function, (C) KEGG pathway, and (D) WikiPathways.

reduction, making the results more manageable for users. For GO, the number of enriched gene sets are significantly reduced, due to known high redundancy between GO terms (Figure 2A and B). The KEGG (16) and WikiPathways (7) pathway databases also show a large degree of reduction (Figure 2C and D). Meanwhile, important biological themes are all covered with these selected gene sets. By selecting a representative subset in the HTML report, all the visualizations will be updated. Detailed information about the clustering can be found in the downloaded results, including the members of the affinity propagation cluster and the coverage of genes in the enrichment results by set cover.

WEBGESTALTR PACKAGE AND API

To further extend the usage of WebGestalt beyond the web server and reach users with programming skills, we have incorporated the core computation into an R package called WebGestaltR (Figure 3), which is published on CRAN (https://cran.r-project.org/package=WebGestaltR). WebGestaltR enables enrichment analysis in batch mode and integration with other computation pipelines, while still having access to the functional annotation and ID mapping data hosted on the WebGestalt server. In addition to using supported datasets and uploading gene lists or custom data to the server, WebGestaltR also supports directly passing R objects as input (Figure 3). In parallel with batch mode, we extended the enrichment analysis to be able to perform on multiple databases combination on the fly, such as all three GO categories or several pathway databases. The main function in the package returns the enrichment result as a data frame in R, writes all results to text files and generates an HTML report served by the WebGestalt server (Figure 3). The GSEA algorithm was implemented in R and Rcpp so that fine control over the output can be achieved. In this way, WebGestaltR blends the strengths of web- and R-based tools and fills the gap between these two categories of functional enrichment analysis tools. Users can still take advantage of the interactive and user-friendly visualizations in web browser while running large number of jobs locally.

By shifting the computation to WebGestaltR, the WebGestalt server now focuses on handling user submissions and serving the results. It has been enhanced with more functionality for interaction with the underlying data and the job submission form. RESTful APIs were designed to allow access to the data on the WebGestalt server, such as the functional categories and ID mapping, by WebGestaltR or any third-party programs (Figure 3). The web API endpoints accept HTTP POST requests with JSON data or GET requests with encoded parameters and may return JSON or plain text dependent on the endpoint (Table 2 and Supplementary File). The front page of WebGestalt was refreshed to provide better user input validation and support for ORA analysis on multiple databases and is able to prefill values in the submission form from parameters encoded in HTTP GET request or form data in POST request (Figure 3 and Table 2), which is useful for external websites to link to WebGestalt for enrichment analysis. Part of required data can be loaded, for example gene list, and let user to choose which database to run on or set advanced parameters. When all the parameters are prepared, the request can be posted to

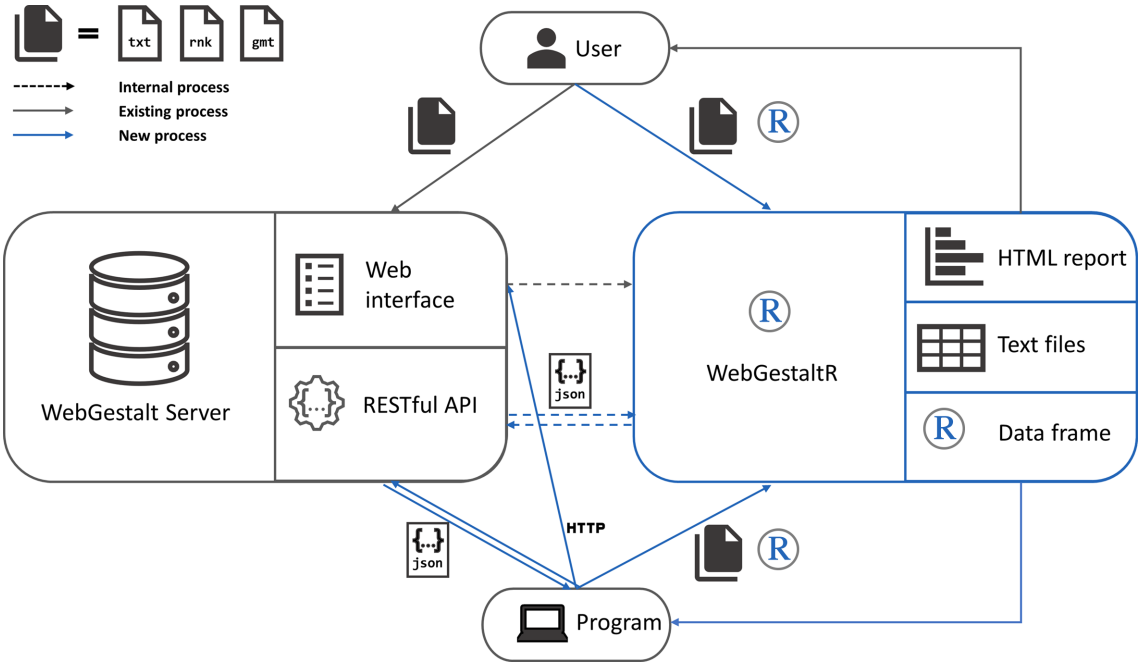

Figure 3. The redesigned system centered around the WebGestaltR package allows easy access of users and programs.

Table 2. List of API and programmatic access endpoints on the WebGestalt server

Endpoint	HTTP method	Description*
/api/summary/[type]	GET/POST	Get summary of available data types: 'idtype', 'geneset', 'referenceset', 'network'
/api/geneset	GET/POST	Get functional annotations e.g. GMT files
/api/reference	GET/POST	Get reference gene list used in ORA
/api/idmapping	GET/POST	Map gene IDs from one type to another
/	GET/POST	Prefill values in submission form
/process	POST	Submit job

*Detailed description about the parameters can be found in Supplementary File and WebGestalt website.

run the job directly as well (Table 2). The system is summarized in Figure 3 with existing processes of submitting jobs to WebGestalt server colored in grey and new processes colored in blue.

DISCUSSION

WebGestalt 2019 presents an update with expanded functionalities, revamped UIs, a new R package and web APIs. Since the other major web-based enrichment tools have not updated since the previous update of WebGestalt, the comparison from the WebGestalt 2017 paper remains valid for this update. With the WebGestaltR package and web APIs, we hope to expand the applicability of WebGestalt and serve users from different backgrounds. We expect the new visualizations and redundancy reduction in the report will be useful for interpreting the results and generating high-quality figures. The introduction of functional annotations for phosphosites in the database makes WebGestalt stand out from similar web servers as a unique -omics enrichment tool that addresses the emerging need for tools to analyze phosphoproteomics data. One limitation of the phosphosite annotation is that the coverage of kinases is low and skewed; most phosphosites do not have a known kinase or function and some kinases are much more studied and annotated than others (11). This can be potentially addressed by sequence-based prediction of kinase substrates; however, most kinases do not have sufficient numbers of annotated phosphosites for accurate computational modeling. We added the functionality to support the combination of multiple databases for ORA and GSEA in the WebGestaltR package and enabled it for ORA on the web server for now, since the running time of GSEA increases greatly with the total number of gene sets due to the permutation algorithm. We will continue to actively incorporate new gene set analysis algorithms and databases into WebGestalt in the future.

DATA AVAILABILITY

WebGestalt can be freely accessed at http://www.webgestalt.org. The WebGestaltR is published on CRAN (https://cran.r-project.org/package=WebGestaltR) and its source code is hosted on Github (https://github.com/bzhanglab/WebGestaltR).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Cancer Institute (NCI) CPTAC Award [U24 CA210954]; Cancer Prevention and Research Institutes of Texas (CPRIT) Award [RR160027]; McNair Medical Institute at The Robert and Janice McNair Foundation. Funding for open access charge: CPRIT Award [RR160027].
Conflict of interest statement. None declared.

REFERENCES

1. Zhang,B., Kirov,S. and Snoddy,J. (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res.*, **33**, W741–W748.
2. Wang,J., Duncan,D., Shi,Z. and Zhang,B. (2013) WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res.*, **41**, W77–W83.
3. Wang,J., Vasaikar,S., Shi,Z., Greer,M. and Zhang,B. (2017) WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res.*, **45**, W130–W137.
4. Khatri,P., Sirota,M. and Butte,A.J. (2012) Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Comput. Biol.*, **8**, e1002375.
5. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
6. Wang,J., Ma,Z., Carr,S.A., Mertins,P., Zhang,H., Zhang,Z., Chan,D.W., Ellis,M.J.C., Townsend,R.R., Smith,R.D. *et al.* (2017) Proteome profiling outperforms transcriptome profiling for coexpression based gene function prediction. *Mol. Cell Proteomics*, **16**, 121–134.
7. Slenter,D.N., Kutmon,M., Hanspers,K., Riutta,A., Windsor,J., Nunes,N., Mélius,J., Cirillo,E., Coort,S.L., Digles,D. *et al.* (2018) WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.*, **46**, D661–D667.
8. Giurgiu,M., Reinhard,J., Brauner,B., Dunger-Kaltenbach,I., Fobo,G., Frishman,G., Montrone,C. and Ruepp,A. (2019) CORUM: the comprehensive resource of mammalian protein complexes—2019. *Nucleic Acids Res.*, **47**, D559–D563.
9. Scott,A.F., Schiettecatte,F., Bocchini,C.A., Amberger,J.S. and Hamosh,A. (2014) OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.*, **43**, D789–D798.
10. Hornbeck,P. V., Zhang,B., Murray,B., Kornhauser,J.M., Latham,V. and Skrzypek,E. (2015) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.*, **43**, D512–D520.
11. Needham,E.J., Parker,B.L., Burykin,T., James,D.E. and Humphrey,S.J. (2019) Illuminating the dark phosphoproteome. *Sci. Signal.*, **12**, eaau8645.
12. Krug,K., Mertins,P., Zhang,B., Hornbeck,P., Raju,R., Ahmad,R., Szucs,M., Mundt,F., Forestier,D., Jane-Valbuena,J. *et al.* (2019) A curated resource for phosphosite-specific signature analysis. *Mol. Cell Proteomics*, **18**, 576–593.
13. Drake,J.M., Graham,N.A., Stoyanova,T., Sedghi,A., Goldstein,A.S., Cai,H., Smith,D.A., Zhang,H., Komisopoulou,E., Huang,J. *et al.* (2012) Oncogene-specific activation of tyrosine kinase networks during prostate cancer progression. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 1643–1648.
14. Huang,K.Y., Wu,H.Y., Chen,Y.J., Lu,C.T., Su,M.G., Hsieh,Y.C., Tsai,C.M., Lin,K.I., Huang,H. Da, Lee,T.Y. *et al.* (2014) RegPhos 2.0: An updated resource to explore protein kinase-substrate phosphorylation networks in mammals. *Database*, **2014**, bau034.
15. Franz,M., Lopes,C.T., Huck,G., Dong,Y., Sumer,O. and Bader,G.D. (2015) Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics*, **32**, 309–311.
16. Tanabe,M., Sato,Y., Morishima,K., Furumichi,M. and Kanehisa,M. (2016) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.
17. Frey,B.J. and Dueck,D. (2007) Clustering by passing messages between data points. *Science*, **315**, 972–976.
18. Golab,L., Korn,F., Li,F., Saha,B. and Srivastava,D. (2015) Size-constrained weighted set cover. In: *Proceedings - International Conference on Data Engineering*. IEEE, Vol. **2015**, pp. 879–890.