Journal of Genetics and Genomics 45 (2018) 415–417

Contents lists available at ScienceDirect

Journal of Genetics and Genomics

Journal homepage: www.journals.elsevier.com/journal-of-genetics-and-genomics/

Letter to the editor

WocEA: The visualization of functional enrichment results in word clouds

The integration, analysis and visualization of the big omics data are critical for addressing a broad spectrum of biological questions. One of the most frequently conducted procedures is enrichment analysis, which statistically tests whether individual functional annotations of Gene Ontology (GO) or Kyoto Encyclopedia of Genes and Genomes (KEGG) are significantly over- or under-represented in an "interesting" gene or protein list against the reference set (Tavazoie et al., 1999). We went through five leading journals including *Nature, Cell, Cancer Cell*, *Genome Research* and *Genome Biology*, and found that ~27.8% (563 out of 2027) of all research papers published in 2015 contained at least one enrichment analysis (Table S1). In most of these studies, tens of or even hundreds of significant annotation categories would be detected, and the enrichment results were usually present in tables or histograms, which were far from intuitive or informative for readers. To obtain an accurate, concise and compact artwork, it was proposed that enrichment results could be illustrated in word clouds, in which the font size is proportionally correlated with the statistical significance (*p*-value) (Supek et al., 2011; Tabas-Madrid et al., 2012; Kolde and Vilo, 2015). However, two major problems must be resolved for a much broader usage of such a display. First, the word cloud was originally designed for visualizing short words or phrases (Viegas et al., 2009), whereas GO or KEGG descriptions are relatively long. Thus, the layout algorithms need to be significantly optimized to meet aesthetic criteria for the publication. Second, a typical enrichment analysis using the hypergeometric distribution calculates an enrichment ratio (E-ratio) and a *p*-value for each category, and both measurements should be considered in the word-cloud visualization.

Here, we present a novel software package of WocEA (Word cloud for the Enrichment Analysis) for the visualization of enrichment analysis results in word clouds. Because words can be illustrated in different font sizes and colors, we proposed that *p*-value can be linearly shown by font size and E-ratio can be represented by font color, or *vice versa*. In default, the font size of an annotation term was correlated with *p*-value and determined as below:

$$FS = \left(\frac{NP - Min}{Max - Min}\right)^{r} \times (LFS - SFS) + SFS \quad (1)$$

While

- FS = the font size
- NP = -log$_a$ (*p*-value) (default a = 10)
- Max = the maximum of all NPs
- Min = the minimum of all NPs
- r = the coefficient (ranges from 0.1 to 10, default = 1)
- LFS = the largest font size (default = 62)
- SFS = the smallest font size (default = 9)

So the font color can be decided as below:

$$FC = \left(\frac{E - Min}{Max - Min}\right)^{r} \times 256 \times 3 \quad (2)$$

While

- FC = the font color
- E = E-ratio
- Max = the maximum of all E-ratios
- Min = the minimum of all E-ratios
- r = the coefficient (ranges from 0.1 to 10, default = 1)

In WocEA, users can reversely present *p*-value and E-ratio for their own purposes. The layout was adapted from previous studies with a great improvement by exhaustively testing (Viegas et al., 2009; Baroukh et al., 2011; Oesper et al., 2011; Supek et al., 2011; Tabas-Madrid et al., 2012; Kolde and Vilo, 2015). For the usage, the enrichment results can be prepared either in Microsoft Excel spreadsheet (.xls and.xlsx) or tab-separated value (TSV) formats, and the automatically generated word cloud can be easily resized, re-shaped and re-colored in a customized manner. The final artworks can be exported into publication-quality figures in multiple image formats. Moreover, all information of a word cloud can be saved as a project file in XML (The Extensible Markup Language) format, enabling a further reuse. We also implemented an option of the enrichment analysis for several model organisms, by downloading GO annotations from the UniProt database (http://www.uniprot.org/). A typical procedure for the manipulation of WocEA was shown in a 2′39″ video (http://wocea.biocuckoo.org/faq.php).

Besides the visualization of enrichment results, WocEA also supported the classical word frequency-based analysis. For example, in 2017, *Journal of Genetics and Genomics* (*JGG*) published a special issue on Database, containing up to five research articles (Xue and Wang, 2017). To get a quick snapshot of biological themes mentioned in this issue, we directly pasted the full texts of all articles into WocEA for a word-cloud illustration (Fig. 1A). Clearly, the special issue was mainly focused on "gene", "database" and "data", since these words frequently occurred throughout all papers (Fig. 1A). Previously, we developed an integrative database of CGDB, containing ~73,000 circadian genes in 148 eukaryotes (Li et al., 2017). Using 1889 human circadian genes, we carried out both GO- and KEGG-based enrichment analyses for a better understanding the biological functions of circadian clocks (Li et al., 2017).

https://doi.org/10.1016/j.jgg.2018.02.008

In this work, we re-performed the enrichment analysis of GO biological process terms for human circadian genes, and compared the illustration of WocEA (Fig. 1B) with several existing tools, such as GeneCodis3 (Tabas-Madrid et al., 2012) (Fig. S1A), Genes2-WordCloud (Baroukh et al., 2011) (Fig. S1B) and GOsummaries (Kolde and Vilo, 2015) (Fig. S1C). Besides the three tools, there were two additional word-cloud visualization programs developed for processing biological data, such as Cytoscape WordCloud (Oesper et al., 2011) and REVIGO (Supek et al., 2011). The former was specifically designed to visualize biological texts based on word frequencies for given sub-networks but not gene lists, whereas the visualization of REVIGO was not available. Thus, the two programs were not chosen for the comparison. It should be noted that although WocEA provided an option for GO-based enrichment analysis, it should be mainly regarded as a visualization tool, because using outdated GO annotations might generate misleading results (Wadi et al., 2016). Here, we recommend that users can obtain the latest version of annotation files for conducting the enrichment analysis. Obviously, the aesthetic presentation of WocEA in the horizontal and perpendicular layout was much better than other programs (Fig. 1B and Fig. S1).

Additional layouts were also supported in WocEA for the illustration of enrichment results. For example, Ronningen et al. (2015) conducted a comprehensive analysis of lamin A/C-associated domains in human adipose-derived stromal cells (ASCs) at different stages of adipogenesis, by using the chromatin immunoprecipitation-sequencing (ChIP-seq). They found lamin A/C-chromatin interactions to be dramatically reorganized and totally identified 2716 genes up-regulated within one day after adipogenic induction. The GO-based enrichment analysis revealed that a number of metabolic processes were statistically associated with the transition from preadipogenic to adipogenic stages (Ronningen et al., 2015). Using WocEA, we re-illustrated the enrichment results in a concentric circle layout (Fig. 1C). Moreover, Villar et al. (2015) profiled the genome-wide occurrence of two histone modifications including acetylated lysine 27 on histone H3 (H3K27ac) and trimethylated lysine 4 of histone H3 (H3K4me3) with ChIP-seq to detect enhancers and promoters in liver of 20 mammalian species. They defined “highly-conserved enhancers”, if the regulatory activity was steadily present in all ten of highest-quality genomes. By using genes near highly-conserved enhancers, they performed a GO-based enrichment analysis and observed that highly-conserved enhancers might regulate multiple fundamental processes, such as DNA replication and transcription (Villar et al., 2015). Here, WocEA was used to visualize the results in a compact layout (Fig. 1D).

In conclusion, we developed a highly useful software package of WocEA mainly for the word-cloud visualization of enrichment results, and the traditional word frequency-based illustration was also supported. We anticipated that such a display will be popular for the mainstream analysis of the multi-dimensional omics data. WocEA will be continuously maintained and refined upon users’ feedbacks. The software packages of WocEA 1.0 are freely available for academic research at: http://wocea.biocuckoo.org/. For convenience, we also designed an online service of WocEA at: http://wocea.biocuckoo.org/online.php.

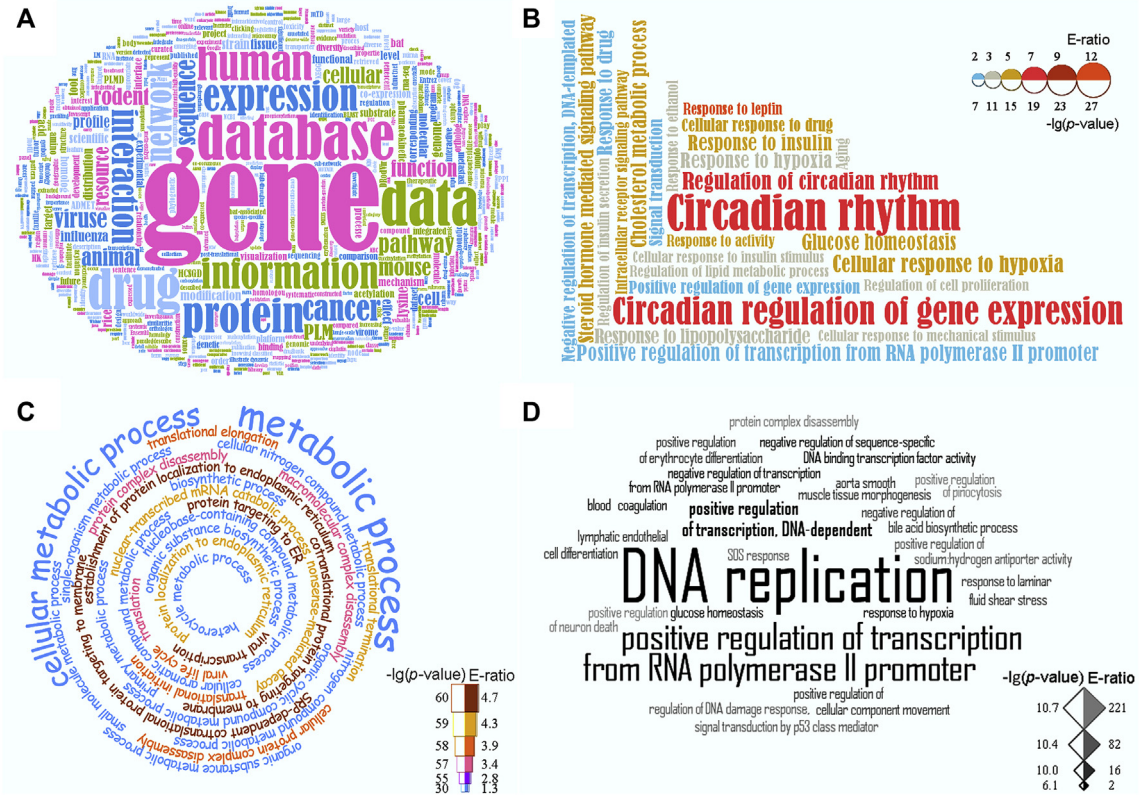

Fig. 1. The aesthetic illustrations of WocEA. **A**: The word-frequency display of the *JGG* special issue on Database. The larger font size represents a higher frequency of the word mentioned in the texts. **B**: The horizontal and perpendicular visualization of enrichment results of human circadian genes. The whole human proteome set was used as the background, whereas the hypergeometric distribution was adopted to test the statistical significance of individual GO biological process terms (p-value < 10^{-7}). In default, the font size and color were linearly correlated with -lg (p-value) and E-ratio, respectively. **C**: The concentric circle layout for enriched GO terms potentially associated with the adipogenic transition (Ronningen et al., 2015). **D**: The compact layout for over-represented GO biological processes potentially regulated by highly-conserved enhancers in mammalians (Villar et al., 2015).

Acknowledgments

This study was inspired from a ScienceNet blog post written by Dr. Wei Li (Netbase Solutions). This work was supported by the Special Project on Precision Medicine under the National Key R&D Program (2017YFC0906600), and the Natural Science Foundation of China (No. 31671360).

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jgg.2018.02.008.

References

Baroukh, C., Jenkins, S.L., Dannenfelser, R., Ma'ayan, A., 2011. Genes2WordCloud: a quick way to identify biological themes from gene lists and free text. Source Code Biol. Med. 6, 15.

Kolde, R., Vilo, J., 2015. GOsummaries: an R package for visual functional annotation of experimental data. F1000 Res. 4, 574.

Li, S., Shui, K., Zhang, Y., Lv, Y., Deng, W., Ullah, S., Zhang, L., Xue, Y., 2017. CGDB: a database of circadian genes in eukaryotes. Nucleic Acids Res. 45, D397–D403.

Oesper, L., Merico, D., Isserlin, R., Bader, G.D., 2011. WordCloud: a Cytoscape plugin to create a visual semantic summary of networks. Source Code Biol. Med. 6, 7.

Ronningen, T., Shah, A., Oldenburg, A.R., Vekterud, K., Delbarre, E., Moskaug, J.O., Collas, P., 2015. Prepatterning of differentiation-driven nuclear lamin A/C-associated chromatin domains by GlcNAcylated histone H2B. Genome Res. 25, 1825–1835.

Supek, F., Bosnjak, M., Skunca, N., Smuc, T., 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One 6, e21800.

Tabas-Madrid, D., Nogales-Cadenas, R., Pascual-Montano, A., 2012. GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. Nucleic Acids Res. 40, W478–W483.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., Church, G.M., 1999. Systematic determination of genetic network architecture. Nat. Genet. 22, 281–285.

Viegas, F.B., Wattenberg, M., Feinberg, J., 2009. Participatory visualization with wordle. IEEE Trans. Visual. Comput. Graph. 15, 1137–1144.

Villar, D., Berthelot, C., Aldridge, S., Rayner, T.F., Lukk, M., Pignatelli, M., Park, T.J., Deaville, R., Erichsen, J.T., Jasinska, A.J., Turner, J.M., Bertelsen, M.F., Murchison, E.P., Flicek, P., Odom, D.T., 2015. Enhancer evolution across 20 mammalian species. Cell 160, 554–566.

Wadi, L., Meyer, M., Weiser, J., Stein, L.D., Reimand, J., 2016. Impact of outdated gene annotations on pathway enrichment analysis. Br. J. Pharmacol. 13, 705–706.

Xue, Y., Wang, X., 2017. Bioinformaticians wrestling with the big biomedical data. J. Genet. Genomics 44, 223–225.

Wanshan Ning, Shaofeng Lin, Jiaqi Zhou, Yaping Guo, Ying Zhang, Di Peng, Wankun Deng*, Yu Xue*

Key Laboratory of Molecular Biophysics of Ministry of Education, College of Life Science and Technology and the Collaborative Innovation Center for Biomedical Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

* Corresponding authors.
E-mail addresses: dengwankun@hust.edu.cn (W. Deng), xueyu@hust.edu.cn (Y. Xue).

22 August 2017
Available online 13 April 2018