



SOFTWARE TOOL ARTICLE

**REVISED** Gene Annotation Easy Viewer (GAEV): Integrating KEGG’s Gene Function Annotations and Associated Molecular Pathways [version 3; peer review: 3 approved]

Trung Huynh , Sen Xu

Department of Biology, University of Texas at Arlington, Arlington, Texas, 76019, USA

**v3** **First published:** 29 Mar 2018, 7:416 (<https://doi.org/10.12688/f1000research.14012.1>)  
**Second version:** 28 Jun 2018, 7:416 (<https://doi.org/10.12688/f1000research.14012.2>)  
**Latest published:** 09 May 2019, 7:416 (<https://doi.org/10.12688/f1000research.14012.3>)

**Abstract**

We developed a Gene Annotation Easy Viewer (GAEV) that integrates the gene annotation data from the KEGG (Kyoto Encyclopedia of Genes and Genomes) Automatic Annotation Server. GAEV generates an easy-to-read table that summarizes the query gene name, the KO (KEGG Orthology) number, name of gene orthologs, functional definition of the ortholog, and the functional pathways that query gene has been mapped to. Via links to KEGG pathway maps, users can directly examine the interaction between gene products involved in the same molecular pathway. We provide a usage example by annotating the newly published freshwater microcrustacean *Daphnia pulex* genome. This gene-centered view of gene function and pathways will greatly facilitate the genome annotation of non-model species and metagenomics data. GAEV runs on a Windows or Linux system equipped with Python 3 and provides easy accessibility to users with no prior Unix command line experience.

**Keywords**

molecular pathway, *Daphnia*, genome annotation, visualization, homologous genes

**Open Peer Review**

**Reviewer Status**

	Invited Reviewers		
	1	2	3
<b>version 3</b> published 09 May 2019	<b>REVISED</b>	 report	 report
<b>version 2</b> published 28 Jun 2018	 report	↑	 report
<b>version 1</b> published 29 Mar 2018	 report	↑	 report

- Fragiskos N Kolisis**, National Technical University of Athens, Athens, Greece  
**Efthymios Ladoukakis** , National Technical University of Athens, Athens, Greece
- Tonia S. Schwartz** , Auburn University, Auburn, USA
- Simo V. Zhang** , Indiana University Bloomington, Bloomington, USA

Any reports and responses or comments on the article can be found at the end of the article.

**Corresponding author:** Sen Xu ([Sen.Xu@uta.edu](mailto:Sen.Xu@uta.edu))

**Author roles:** **Huynh T:** Investigation, Methodology, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Xu S:** Conceptualization, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** University of Texas at Arlington

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Copyright:** © 2019 Huynh T and Xu S. This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Huynh T and Xu S. **Gene Annotation Easy Viewer (GAEV): Integrating KEGG's Gene Function Annotations and Associated Molecular Pathways [version 3; peer review: 3 approved]** F1000Research 2019, 7:416 (<https://doi.org/10.12688/f1000research.14012.3>)

**First published:** 29 Mar 2018, 7:416 (<https://doi.org/10.12688/f1000research.14012.1>)

**REVISED** Amendments from Version 2

A section was added under use cases to further explain how to use the batch jobs feature, and various typos raised by reviewers were addressed.

See referee reports

## Introduction

Describing the biological function of computationally annotated genes in non-model assemblies and the molecular pathways formed by these genes' products is critical for identifying the genetic basis of the various unique biological attributes (e.g., physiology, life history, behavior) of the species in question. Computational search against DNA/protein databases, e.g., NCBI Blast (Boratyn *et al.*, 2013), UniProt (Bateman *et al.*, 2017), InterPro (Finn *et al.*, 2017), based on homology and protein domain information using computational tools, such as Blast (Camacho *et al.*, 2009), InterProScan (Jones *et al.*, 2014), and Hmmer (Mistry *et al.*, 2013), can make predictions for individual gene functions. In contrast, delineating the molecular pathways encoded by the entire suite of genes of a single species is a much more challenging task, especially for non-model species. To this extent, mapping genes to the molecular pathways derived from intensively studied model organisms provides an entry point for addressing this need.

For mapping genes into known molecular pathways, the *Kyoto Encyclopedia of Genes and Genomes* (KEGG) provide comprehensive web services (Kanehisa *et al.*, 2017; Kanehisa & Goto, 2000; Kanehisa *et al.*, 2016a). KEGG is an integrated database for biological interpretation of genome sequences. The molecular function of genes is classified using ortholog groups, i.e., KEGG Orthology (KO). KEGG also contains KEGG pathways, BRITE hierarchies, and KEGG modules, all of which are networks of KO nodes. It is possible to annotate the molecular functions of a set of genes from complete/partial genome assembly or metagenomics dataset and their encoded molecular pathways using KEGG automatic annotation services that are provided through web servers BlastKOALA and GhostKOALA (Kanehisa *et al.*, 2016b). For a non-model species, we can use KAAS (KEGG Automatic Annotation Server) web services to annotate the complete or random set of genes to describe their molecular function and map them into identified molecular pathways. The annotation results consist of KO numbers for each gene, genes mapped to KEGG pathway database, and genes mapped to BRITE. Nonetheless, the resulting complete set of pathways and BRITE hierarchy can only be viewed through the temporary URL provided by KEGG, which are only available for several days after the analyses are completed. Although these results are organized through either curated KEGG pathways or BRITE hierarchy, KAAS does not provide an integrative gene-centered view of gene function and pathways, i.e., the complete summary of gene function and all associated molecular pathways for each gene.

As can be envisioned, integrating the gene function annotation based on KEGG orthology and KEGG pathways can provide an

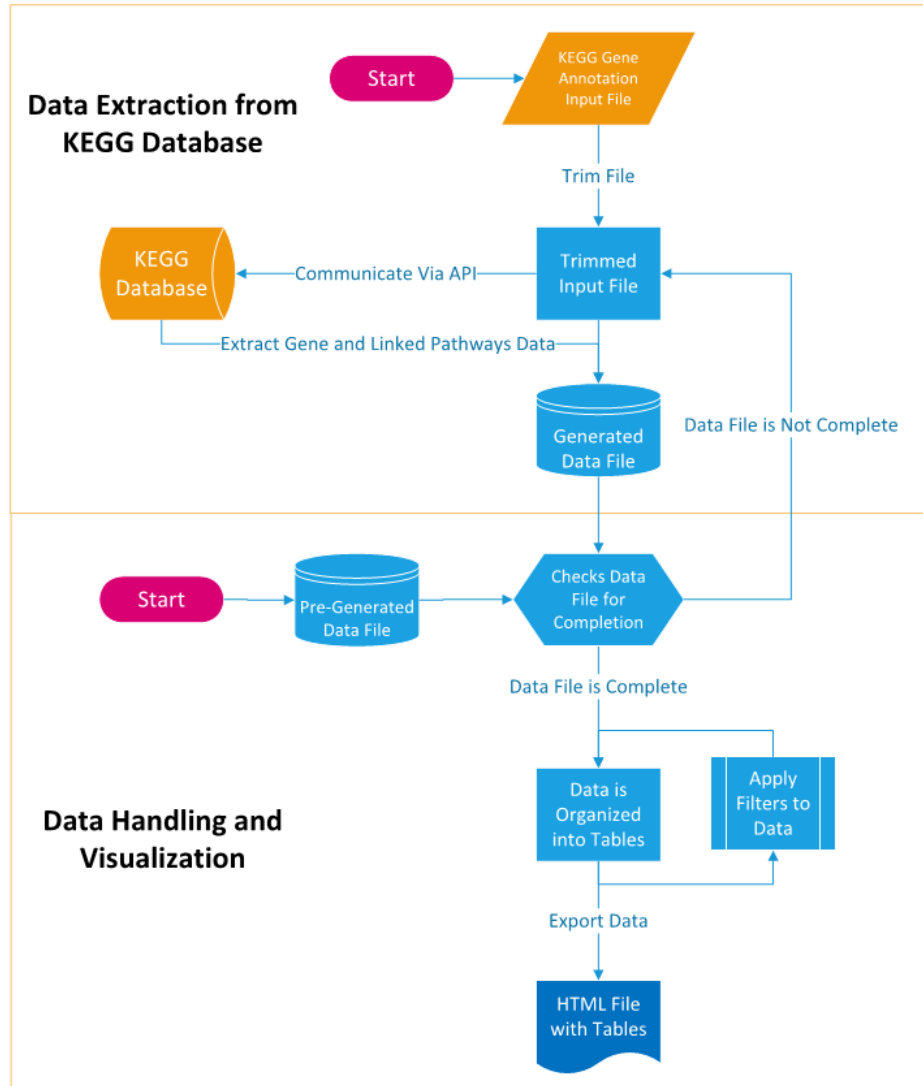
efficient way to characterize both the predicted genes and associated pathways for a newly assembled genome or metagenomics dataset. Despite numerous computational packages for retrieving KEGG pathways using the API interface provided by KEGG database (e.g., Moutselos *et al.*, 2009; Wrzodek *et al.*, 2011), none of these packages to our best knowledge allows us to reconstruct the complete set of molecular pathways contained in a newly assembled genome. To provide a means to utilizing the highly informative resources at KEGG for annotating genomic sequences and molecular pathways for non-model species, we have developed a Gene Annotation Easy Viewer (GAEV) for integrating results of KEGG orthology annotation and KEGG pathways mapping using KEGG API tools in both Windows and Linux environment. GAEV is aimed to provide a gene-centered view of gene function and pathways, i.e., the complete summary of gene function and all possibly associated molecular pathways for each gene. This is distinct from other KEGG-related software such as MEGAN (Huson *et al.*, 2016) and MinPath (Ye & Doak, 2009). MEGAN can achieve overall functional analysis of microbiome data with KEGG data (Huson *et al.*, 2016), whereas Minpath aims to provide a conservative and faithful estimation of the biological pathways for a query dataset (Ye & Doak, 2009). GAEV is implemented in Python 3 and can be used as an independent package.

## Methods

Assuming that the KEGG ortholog number is known for a single gene, the KO information can be retrieved from KEGG database by utilizing KEGG REST-style API. GAEV uses the 'get' operation of the KEGG API to extract data on the gene and linked pathways of every K number provided in the input file. The data extracted from KEGG database are stored in data files that can be loaded into GAEV to skip the data extraction step (Figure 1).

Some genes will not have a KO number associated with them. Before data extraction, GAEV will trim the input file to remove all genes that do not have a KO number or have a KO number that cannot be searched in the KEGG database. Once data extraction from KEGG's database is complete and the data file is generated, GAEV helps the user handle and visualize the data by exporting the data as a table in an HTML file. GAEV populates the table with the user defined gene ID provided in the input file and the associated K number provided in the input file, as well as the gene name, definition, and linked pathways that have been retrieved from the KEGG database. The linked pathway map URLs that highlight identified genes in the genome assembly are created using the following formula: `http://www.kegg.jp/kegg-bin/show_pathway?map=[mapno]&multi_query=%23ffffbf%0d%0a[k-num1]+%23ffffbf%0d%0a[k-num2]+...%23ffffbf%0d%0a[k-num_interest]+%23[node_color],%23[font_color]`.

In the above URL, [mapno] represents the pathway accession number. [k-num{1,2,3...}] represents the K number for each gene in the pathway that is present in the provided genome assembly, and [k-num\_interest] represents the K number of the focal



**Figure 1. Workflow of Gene Annotation Easy Viewer (GAEV).**

gene that will be highlighted with a unique color. **[node\_color]** and **[font\_color]** represent the desired color of the focal gene's node and font on the pathway map, respectively. By default, the node color of the focal gene is dark red, whereas the node color of other genes in the same pathway that are present in the genome assembly is light green.

## Use cases

### Installation

The most up-to-date version of this software can be downloaded at [https://github.com/UtaDaphniaLab/kegg\\_path\\_generator](https://github.com/UtaDaphniaLab/kegg_path_generator). This software requires Python 3 or newer to run. It is recommended that this software be used as a standalone program simply by double clicking on GAEV.py or by using the 'python 3 GAEV.py' command.

### Annotation

We analyze the newly published *Daphnia pulex* genome (Ye *et al.*, 2017) to demonstrate the usage of our package. The

required input file for our package contains two columns. The first column contains the gene names, whereas the second column represents the KO (KEGG orthology) numbers (Figure 2, Supplementary File 1). The KO numbers for the entire set of genes can be obtained through [KEGG Automatic Annotation Server](#). Briefly, users can provide the query protein sequences in a fasta file and use one of the provided search algorithms (e.g. Blast, GhostX, GhostZ) to assign KO numbers to each of the queried genes. The *Daphnia* protein fasta file can be found at [https://figshare.com/articles/PA42\\_3\\_0\\_protein\\_new\\_txt/6653297](https://figshare.com/articles/PA42_3_0_protein_new_txt/6653297). With a gff/gtf genome annotation file, users can also use tools such as gff2sequence (Camiolo & Porceddu, 2013) to extract DNA/protein sequences from genomic assemblies, which can be used as query sequences. Furthermore, researchers working with non-model organisms could use protein sequences extracted from an assembled transcriptome as input data. At the end of this analysis, the user will receive via email a link to the result page, where the query result can be downloaded. The downloaded query result can be directly used as input file for our package

gene1	
gene2	K12829
gene3	K14963
gene4	K20672
gene5	K12184
gene6	
gene7	K04958
gene8	
gene9	K09075
gene10	

**Figure 2. Example input file.**

even when some genes are not provided a KO number (which will be automatically excluded from further analysis).

With the obtained input file, the annotation analysis can be started by simply running GAEV.py and following the instructions of the menus. The first menu provides the option of using the obtained input file to extract data from KEGG or skipping the data extraction step by loading a pre-generated data file. Next, GAEV will prompt the user for the location of the input or data file. Both absolute and relative paths are accepted, but it is recommended that the GAEV.py file be placed in the same folder as the input or data file, so that the relative path can be easily used. After the data extraction from KEGG's servers is completed, a data file will be created, which can be repeatedly used for making different pathway tables. The next several menus guide the user through the process of customizing the output table. The user has the options to apply filters so that GAEV only outputs a table using genes with a specific keyword in its definition or linked pathways.

### Output file

The output file is an html file that can be opened in any internet browser (for example see [Supplementary File 2](#)). The results are organized in three different sections. The first section is the Genes and Linked Pathways, where for each query gene the molecular function based on KO and relevant pathways are listed. For each gene, its associated pathway(s) contains a link to the corresponding pathway page on KEGG website, where this specific gene is colored in red and all the identified genes from the genome assembly are colored in green. The other two sections contain a list of the pathways sorted by the number of identified genes and by alphabetic order, respectively. These two sections provide a pathway-centered view of the functions of the annotated genome.

### Batch jobs

The batch functions located in the first menu can be used when there are several sets of genes in different input files that the user

wants to annotate. The batch functions require a file with the relative/absolute paths of each input file on separate lines. Alternatively, entering 'all' will direct GAEV to run using every file with a txt extension as an input file if new data files need to be generated or every file with a dat extension if new tables need to be created from existing data files.

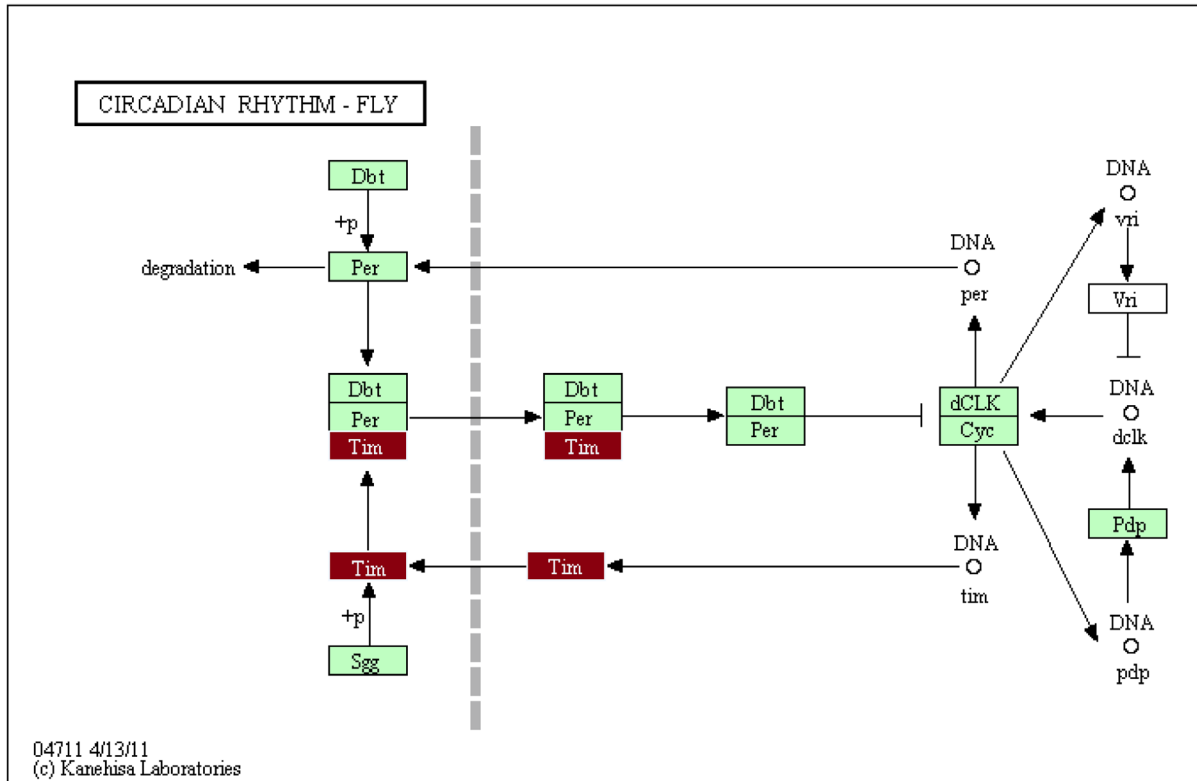
Filters can be applied to batch jobs and will apply to all sets of genes. An html output file like the one described above will be created for each set of genes in the same folder as its respective input and data file.

### Conclusions

The integrative annotation approach implemented in our package GAEV draws upon resources available at KEGG and provides an efficient way to explore the molecular pathways embodied in a draft genome. The integration of the generated html file with KEGG web services provides an intuitive interface to explore specific molecular pathways, with all the identified KEGG homologs highlighted in the pathway map. This type of information is essential to initial exploration of non-model organisms' genomes to understand the conservation of specific pathways compared to established model systems. For example, if we examine the circadian rhythm pathway in the *Daphnia* genome (by clicking on the link to the circadian rhythm pathway in the generated html file), we see strong conservation between *Daphnia* and *Drosophila*, with only 1 gene (i.e., *Vri*) in this pathway missing an identified homolog in the *Daphnia* assembly ([Figure 3](#)). Further efforts can be dedicated to verifying the absence of *Vri* gene in *Daphnia* genome. The strong conservation of the circadian pathway can greatly aid future efforts in using the freshwater microcrustacean *Daphnia* to understand the internal clock of aquatic organisms in response to aquatic environments.

In principle, GAEV can be used for visualizing functions and pathways for gene sets of any scale, ranging from genome-wide data to subsets of genes in a genome. For example, we can use GAEV to visualize the pathways that differentially expressed genes are involved in. Often the large number of differentially expressed genes from RNA-seq experiments prevents clear cataloguing of these genes and molecular pathways. Analyzing the genes of interest using our package can provide a quick, integrative view of the genes and affected pathways.

In summary, with a user-friendly design (e.g., no requirement of UNIX command line experience) in mind, we have developed GAEV to provide a fast, easily accessible summary for KEGG gene annotation results. We expect that GAEV will find its use in many bioinformatic analyses, especially those involving non-model species.



**Figure 3.** The circadian rhythm pathway in *Daphnia pulex* showing gene of interest (Tim) in red and other identified genes in green.

### Data and software availability

Software source code available from: [https://github.com/UtaDaphniaLab/Gene\\_Annotation\\_Easy\\_Viewer](https://github.com/UtaDaphniaLab/Gene_Annotation_Easy_Viewer)

Archived source code as at time of publication: <https://zenodo.org/record/2549592> (Trung, 2019)

License: This software is licensed under the MIT license

### Grant information

This work is supported by start-up funds from University of Texas at Arlington to SX.

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

### Acknowledgements

The authors thank M. Snyman for testing the software.

### Supplementary files

Supplementary File 1. Example input file [https://raw.githubusercontent.com/UtaDaphniaLab/Gene\\_Annotation\\_Easy\\_Viewer/master/gene\\_annotation\\_easy\\_viewer/example\\_input.txt](https://raw.githubusercontent.com/UtaDaphniaLab/Gene_Annotation_Easy_Viewer/master/gene_annotation_easy_viewer/example_input.txt)

Supplementary File 2. Example output file [https://raw.githubusercontent.com/UtaDaphniaLab/Gene\\_Annotation\\_Easy\\_Viewer/master/gene\\_annotation\\_easy\\_viewer/Example\\_Output.html](https://raw.githubusercontent.com/UtaDaphniaLab/Gene_Annotation_Easy_Viewer/master/gene_annotation_easy_viewer/Example_Output.html)

## References

- Bateman A, Martin MJ, O'Donovan C, *et al.*: **UniProt: the universal protein knowledgebase.** *Nucleic Acids Res.* 2017; **45**(D1): D158–D169.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Boratyn GM, Camacho C, Cooper PS, *et al.*: **BLAST: a more efficient report with usability improvements.** *Nucleic Acids Res.* 2013; **41**(Web Server issue): W29–W33.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Camacho C, Coulouris G, Avagyan V, *et al.*: **BLAST+: architecture and applications.** *BMC Bioinformatics.* 2009; **10**: 421.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Camiolo S, Porceddu A: **gff2sequence, a new user friendly tool for the generation of genomic sequences.** *BioData Min.* 2013; **6**(1): 15.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Finn RD, Attwood TK, Babbitt PC, *et al.*: **InterPro in 2017-beyond protein family and domain annotations.** *Nucleic Acids Res.* 2017; **45**(D1): D190–D199.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Huson DH, Beier S, Flade I, *et al.*: **MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data.** *PLoS Comp Biol.* 2016; **12**(6): e1004957.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Jones P, Binns D, Chang HY, *et al.*: **InterProScan 5: genome-scale protein function classification.** *Bioinformatics.* 2014; **30**(9): 1236–1240.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kanehisa M, Furumichi M, Tanabe M, *et al.*: **KEGG: new perspectives on genomes, pathways, diseases and drugs.** *Nucleic Acids Res.* 2017; **45**(D1): D353–D361.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res.* 2000; **28**(1): 27–30.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kanehisa M, Sato Y, Kawashima M, *et al.*: **KEGG as a reference resource for gene and protein annotation.** *Nucleic Acids Res.* 2016a; **44**(D1): D457–D462.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kanehisa M, Sato Y, Morishima K: **BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences.** *J Mol Biol.* 2016b; **428**(4): 726–731.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Mistry J, Finn RD, Eddy SR, *et al.*: **Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions.** *Nucleic Acids Res.* 2013; **41**(12): e121.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Moutselos K, Kanaris I, Chatziioannou A, *et al.*: **KEGGconverter: a tool for the in-silico modelling of metabolic networks of the KEGG Pathways database.** *BMC Bioinformatics.* 2009; **10**: 324.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Trung: **UtaDaphniaLab/Gene\_Annotation\_Easy\_Viewer: Gene Annotation Easy Viewer v1.1.2 (Version v1.1.2).** *Zenodo.* 2019.  
<http://www.doi.org/10.5281/zenodo.2549592>
- Wrzodek C, Dräger A, Zell A: **KEGGtranslator: visualizing and converting the KEGG PATHWAY database to various formats.** *Bioinformatics.* 2011; **27**(16): 2314–2315.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ye YZ, Doak TG: **A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes.** *PLoS Comp Biol.* 2009; **5**(8): e1000465.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ye Z, Xu S, Spitz K, *et al.*: **A New Reference Genome Assembly for the Microcrustacean *Daphnia pulex*.** *G3 (Bethesda).* 2017; **7**(5): 1405–1416.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

# Open Peer Review

Current Peer Review Status:   

---

## Version 3

Reviewer Report 21 May 2019

<https://doi.org/10.5256/f1000research.20342.r48234>

© 2019 Schwartz T. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Tonia S. Schwartz** 

Department of Biological Sciences, Auburn University, Auburn, AL, USA

The authors have addressed my comments. I have retested the script and it now works just as described. I think this manuscript provides a nice addition to the bioinformatic method amenable to non-model organisms.

**Competing Interests:** No competing interests were disclosed.


**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Reviewer Report 16 May 2019

<https://doi.org/10.5256/f1000research.20342.r48232>

© 2019 Zhang S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Simo V. Zhang** 

Department of Computer Science, Indiana University Bloomington, Bloomington, IN, USA

I thank the authors for addressing all of my previous points. I have no additional comments and I recommend this manuscript to be indexed by the journal.

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Bioinformatics, genomics, genetics.



**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

---

**Version 2**

Reviewer Report 21 January 2019

<https://doi.org/10.5256/f1000research.16445.r42892>

© 2019 Zhang S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Simo V. Zhang**

Department of Computer Science, Indiana University Bloomington, Bloomington, IN, USA

Huynh and Xu present a new approach, GAEV, to annotate gene models of a given non-model genome assembly. It maps genes to their associated pathway counterpart of model species. GAEV specializes on improving the conventional KEGG analyses and focuses on providing a gene centric view of gene function and pathways. Overall, I think GAEV can be a good addition to the current bioinformatics research community. I recommend indexing of this manuscript once the authors could address my following questions/suggestions.

1. The first sentence of the manuscript confuses me about what problem the GAEV tries to solve. It provides a new and better way to annotate a given genome with pathways, but I don't think it tries to help *de novo* assembly per se. I can see a possible role that GAEV plays in evaluating a non-model assembly by looking at how many conservative pathways and genes are completed (for example, using *Drosophila*'s circadian rhythm pathway to evaluate the newly assembled *Daphnia* genome). So it would be better if the authors could be clearer what problems GAEV tries to solve.
2. I have tested GAEV using the example data provided by the authors. It worked without any errors. However, I found two typos in the first two options of the menu popping up from typing "python GAEV.py". Is it supposed to be "Create and generate a new data file and table from a new dataset of KO numbers"? Please fix it if it is the case.
3. GAEV provides an interactive interface to walk the users through the entire procedure. Such a design gives the users certain degree of flexibility. However, if some users need to try GAEV using different parameters or filters (which is usually true when one tries to annotate their genome/metagenome), such an interactive design can be overwhelming. I am wondering if the authors have plans to have an end-to-end automatic version of GAEV? By providing enough options from command line, I think GAEV can be both flexible and efficient. The likely problem with relative/absolute path to input files can also be naturally solved, and the requirement of putting the input files together with GAEV.py is not necessary as well. I am not asking the authors to have an implementation for this publication, but it would be nice to know the authors' thoughts on this.
4. GAEV implements a Trimming method. However, I cannot find some descriptions of this step. Such information is very important to users to understand what has been done internally and how it

has been done.

5. GAEV also supports “BATCH” mode. Could the authors provide some use cases and let the users know what they should expect in terms of the final HTML report?
6. One typo: “MEGAN can achieve overall functional analysisof microbiome ...” should be “MEGAN can achieve overall functional analysis of microbiome ...”.
7. In the URL formula, it seems that there are extra spaces after “www”. Please fix it if it is the case.

**Is the rationale for developing the new software tool clearly explained?**

Partly

**Is the description of the software tool technically sound?**

Yes

**Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?**

Partly

**Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?**

Yes

**Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Bioinformatics, genomics, genetics

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Reviewer Report 09 July 2018

<https://doi.org/10.5256/f1000research.16445.r35571>

© 2018 Kolisis F et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Fragiskos N Kolisis**

Laboratory of Biotechnology, School of Chemical Engineering, National Technical University of Athens, Athens, Greece

**Efthymios Ladoukakis** 

National Technical University of Athens, Athens, Greece

I have no further comments to make.

**Is the rationale for developing the new software tool clearly explained?**

Yes

**Is the description of the software tool technically sound?**

Yes

**Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?**

Yes

**Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?**

Yes

**Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?**

Yes

**Competing Interests:** No competing interests were disclosed.

**We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

---

### Version 1

Reviewer Report 02 May 2018

<https://doi.org/10.5256/f1000research.15230.r32642>

© 2018 Schwartz T. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Tonia S. Schwartz** 

Department of Biological Sciences, Auburn University, Auburn, AL, USA

This is a review of the manuscript “Gene Annotation Easy Viewer (GAEV): Integrating KEGG’s Gene Function Annotations and Associated Molecular Pathways” by Huynh and Xu. The authors present a method via a python script that can take KEGG IDs and output three tables of useful annotation information in the context of the molecular pathways.

The manuscript is well written and is easy to read. I envision that the tool developed will be useful, although see the comments below. With corrections I think this manuscript will be a good contribution to the resources available to researchers working in genomics/transcriptomics in non-model organisms.

#### Major Comments

Although I can start the python script running using python3, the example\_input.txt file cannot be found when I enter the relative path or the absolute path. This is despite the input file being in the same folder as the script. My bioinformatics technician also encountered the same problem. Thus it seems there is an error in the script that would need to be corrected prior to acceptance of the manuscript.

#### Minor Comments

This tool is presented as accessible and user-friendly, and the authors use the example from Ye et al. 2017 of “the query protein sequences in a fasta file”. But a protein sequence file is not provided with that paper (that I could find). Thus there is an assumed knowledge gap that your reader needs to be able to use this tool as you describe. I suggest you provide additional information on how users can go from a genome with predicted genes (i.e. a .gff or .gff3 file) to obtain those query sequences for input into the KEGG.

You may also want to mention that researchers working with non-model organisms could start with a de novo transcriptome or as an input file.

#### **Is the rationale for developing the new software tool clearly explained?**

Yes

#### **Is the description of the software tool technically sound?**

Yes

#### **Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?**

No

#### **Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?**

Yes

#### **Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?**

Partly

**Competing Interests:** No competing interests were disclosed.

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 23 May 2018

**Trung Huynh**, Department of Biology, University of Texas at Arlington, Arlington, USA

Dear Dr. Schwartz,

Thanks very much for your comments on our manuscript. Please see below for how we revised our manuscript to address your concerns.

*1. Although I can start the python script running using python3, the example\_input.txt file cannot be found when I enter the relative path or the absolute path. This is despite the input file being in the same folder as the script. My bioinformatics technician also encountered the same problem. Thus it seems there is an error in the script that would need to be corrected prior to acceptance of the manuscript.*

**Response:** We have found the bug in our code that prevented the script from finding files on Linux and Mac OS properly. This bug has been fixed as of version 1.1.1.

*2. This tool is presented as accessible and user-friendly, and the authors use the example from Ye et al. 2017 of “the query protein sequences in a fasta file”. But a protein sequence file is not provided with that paper (that I could find). Thus there an assumed knowledge gap that your reader needs to be able to use this tool as you describe. I suggest you provide additional information on how users can go from a genome with predicted genes (i.e. a .gtf or .gff3 file) to obtain those query sequences for input into the KEGG.*

**Response:** We provide a fasta file of Daphnia protein sequence through the github page for GAEV. Moreover, in the manuscript we recommend users to use tools such as gff2sequence to create query sequences using information from gtf/gff files. See 2<sup>nd</sup> paragraph of Use Cases.

*3. You may also want to mention that researchers working with non-model organisms could start with a de novo transcriptome or as an input file.*

**Response:** Addressed accordingly. See 2<sup>nd</sup> paragraph of Use Cases.

**Competing Interests:** No competing interests were disclosed.

Reviewer Report 16 April 2018

<https://doi.org/10.5256/f1000research.15230.r32636>

© 2018 Kolisis F et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



### Fragiskos N Kolisis

Laboratory of Biotechnology, School of Chemical Engineering, National Technical University of Athens, Athens, Greece

Efthymios Ladoukakis

National Technical University of Athens, Athens, Greece

The authors of this article present a new tool, compiled as a single Python script, that aids in visualizing the gene annotation analysis results of the webserver-based KEGG services. The function of this tool is to generate an html-based view of the genes analyzed by KEGG tools and the appropriate pathways with

which they are linked. The html view provides also an overview of the total existing pathways per associated genes, a task which can be useful for whole genome and metagenome annotation queries.

The authors provide a thorough explanation about how the tool works and communicates with KEGG API by generating the appropriate links and exporting the generated data. The authors also provide some test datasets which can be used to generate the results mentioned in the manuscript.

Nevertheless this reviewer considers this endeavour to have already been covered by other bioinformatic tools with which a comparison would be necessary to underline the importance of the new tool. For example tools like MEGAN can provide a thorough investigation of the existing KEGG pathways in a genome/metagenome (although by using an older and not commercial version of the KEGG database). Furthermore MinPath can be also used in combination with KEGG generated datasets in order to provide a similar pathway reconstruction analysis. Maybe the authors could elaborate a little bit more about what makes their tool more suitable than these already published tools.

Moreover during the Conclusions section the authors do not seem to explain the methodology in order to examine the differences between *Daphnia* and *Drosophila* and how that (and similar analyses) can be achieved solely (or more intuitively) by exploiting this particular tool.

In general this tool seems like a good addition to a bioinformatic pipeline for genomic or metagenomic analysis but this reviewer thinks that the author must emphasize more on the differences and/or improvements regarding similar tools.

**Is the rationale for developing the new software tool clearly explained?**

Partly

**Is the description of the software tool technically sound?**

Yes

**Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?**

Yes

**Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?**

Yes

**Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?**

Partly

**Competing Interests:** No competing interests were disclosed.

**We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.**

Author Response 23 May 2018

**Trung Huynh**, Department of Biology, University of Texas at Arlington, Arlington, USA

Dear Dr. Kolisis,

Thanks very much for your comments on our manuscript. Please see below for how we revised our manuscript to address your concerns.

1. *Nevertheless this reviewer considers this endeavour to have already been covered by other bioinformatic tools with which a comparison would be necessary to underline the importance of the new tool. For example tools like MEGAN can provide a thorough investigation of the existing KEGG pathways in a genome/metagenome (although by using an older and not commercial version of the KEGG database). Furthermore MinPath can be also used in combination with KEGG generated datasets in order to provide a similar pathway reconstruction analysis.*

**Response:** Our goal with our new tool is to provide a gene-centric view of molecular pathways, where each gene is accompanied by all the pathways where this gene is predicted to play a role. This is different from the purposes of MEGAN and Minpath. We emphasized this idea in the last paragraph of Introduction and drew comparison with MEGAN and Minpath.

2. *Moreover during the Conclusions section the authors do not seem to explain the methodology in order to examine the differences between Daphnia and Drosophila and how that (and similar analyses) can be achieved solely (or more intuitively) by exploiting this particular tool.*

**Response:** This example of *Daphnia* circadian pathway is to demonstrate using this tool for initial exploration of non-model organisms' genomes to understand the conservation of specific pathways compared to established model systems (i.e., *Drosophila*). The *Drosophila* circadian pathway is provided through KEGG database. Users can directly examine their interested pathways from the results of GAEV (click on the link in the generated html file) and view the pathways and mapped genes on KEGG website. We provide a brief explanation of how to technically view the pathway on KEGG server in Discussion.

**Competing Interests:** No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact [research@f1000.com](mailto:research@f1000.com)

F1000Research