SOFTWARE TOOL ARTICLE

PEGS: An efficient tool for gene set enrichment within defined sets of genomic intervals [version 1; peer review: 2 approved]

Peter Briggs¹, A. Louise Hunter ², Shen-hsi Yang³, Andrew D. Sharrocks³, Mudassar Iqbal ⁴

¹Bioinformatics Core Facility, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, M13 9PL, UK

²Division of Diabetes, Endocrinology & Gastroenterology, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, M13 9PL, UK

³Division of Molecular & Cellular Function, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, M13 9PL, UK

⁴Division of Informatics, Imaging & Data Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, M13 9PL, UK

V1 First published: 15 Jul 2021, 10:570
<https://doi.org/10.12688/f1000research.53926.1>
 Latest published: 02 Nov 2021, 10:570
<https://doi.org/10.12688/f1000research.53926.2>

Abstract

Many biological studies of transcriptional control mechanisms produce lists of genes and non-coding genomic intervals from corresponding gene expression and epigenomic assays. In higher organisms, such as eukaryotes, genes may be regulated by distal elements, with these elements lying 10s–100s of kilobases away from a gene transcription start site. To gain insight into these distal regulatory mechanisms, it is important to determine comparative enrichment of genes of interest in relation to genomic regions of interest, and to be able to do so at a range of distances. Existing bioinformatics tools can annotate genomic regions to nearest known genes, or look for transcription factor binding sites in relation to gene transcription start sites. Here, we present PEGS (Peak set Enrichment in Gene Sets). This tool efficiently provides an exploratory analysis by calculating enrichment of multiple gene sets, associated with multiple non-coding elements (peak sets), at multiple genomic distances, and within topologically associated domains. We apply PEGS to gene sets derived from gene expression studies, and genomic intervals from corresponding ChIP-seq and ATAC-seq experiments to derive biologically meaningful results. We also demonstrate an extended application to tissue-specific gene sets and publicly available GWAS data, to find enrichment of sleep trait associated SNPs in relation to tissue-specific gene expression profiles.

Keywords

Genomic data integration, ChIP-seq, RNA-seq, gene set enrichment, genomic intervals

Open Peer Review

Approval Status

	1	2
version 2		
(revision)		
02 Nov 2021		
version 1		
15 Jul 2021		

1. **Nicolae Radu Zabet**, Queen Mary University of London, London, UK

2. **Aziz Khan** , Stanford University, Stanford, USA

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

This article is included in the **Bioinformatics** gateway.

Corresponding author: Mudassar Iqbal (mudassar.iqbal@manchester.ac.uk)

Author roles: **Briggs P:** Data Curation, Investigation, Resources, Software, Visualization; **Hunter AL:** Data Curation, Investigation, Resources, Validation, Writing – Review & Editing; **Yang Sh:** Resources, Validation, Writing – Review & Editing; **Sharrocks AD:** Resources, Supervision, Validation, Writing – Review & Editing; **Iqbal M:** Conceptualization, Investigation, Methodology, Project Administration, Software, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: MI was funded by the MRC (MR/M012174/1). ALH was funded by the MRC (MR/N021479/1). ADS and SHY were funded by the Wellcome Trust (103857/Z/14/Z)

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

Copyright: © 2021 Briggs P *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Briggs P, Hunter AL, Yang Sh *et al.* **PEGS: An efficient tool for gene set enrichment within defined sets of genomic intervals [version 1; peer review: 2 approved]** F1000Research 2021, 10:570 <https://doi.org/10.12688/f1000research.53926.1>

First published: 15 Jul 2021, 10:570 <https://doi.org/10.12688/f1000research.53926.1>

Introduction

Gene expression control in higher organisms is achieved through a complex hierarchical process involving opening of chromatin, histone modifications, and binding of transcription factors (TFs). Experimental approaches to understand transcriptional regulatory mechanisms in a biological context involve large-scale measurement of gene expression. Depending on the design of the experiment, these analyses produce differentially expressed gene sets or clusters for further analysis. These studies are often complemented by assays which map, on a genome-wide scale, TF binding sites (ChIP-seq) or regions of chromatin accessibility (DNase-seq, ATAC-seq). Analyses of these data produce a collection of genomic intervals (peak sets). An important computational task is then to integrate these data to produce meaningful results; i.e. to relate gene sets to peak sets to aid functional interpretation. Bearing in mind distal regulation, an important consideration here is to be able to calculate gene set enrichment at multiple genomic distances from peak sets, and to be able to do this efficiently within the same analysis.

We present a new tool – PEGS (Peak set Enrichment in Gene Sets)¹ – which calculates mutual enrichment of multiple gene sets associated with multiple peak sets, simultaneously and efficiently. This can be at user-defined peak-to-TSS (transcription start site) distances, as well as constraining to topologically associated domains (TADs). Thus, PEGS quickly produces an overall picture of gene set enrichment in relation to peaks, and shows at what genomic distances this is most pronounced. It is applicable to gene sets derived from any source, and peak sets derived from different epigenomic assays, as well as single nucleotide polymorphisms (SNPs) from genome-wide association studies (GWAS).

Methods

Architecture and Implementation

In PEGS, input peaks are extended in both directions for a given distance or constrained within known TAD boundaries, if provided (Figure 1). Subsequently, the enrichment of the input gene set is calculated among the genes whose TSSs overlap with the extended peaks. These tasks are performed in PEGS as follows:

1. Creating a gene interval file in BED (Browser Extensible Data) format for all genes in the given genome

using *refGene* from UCSC Table Browser. This reference TSSs BED file only needs to be created once (human hg38 and mouse mm10 are provided with the tool; a utility is provided to create these for other genomes).

2. For a given peak set, peaks are extended to specified genomic distances in both directions, and up to overlapping TAD boundaries. Intersection of these extended peaks with the gene intervals BED file from step 1 is calculated using BEDTools (RRID:SCR_006646)², leading to a gene set whose TSSs overlap with extended peaks.
3. Using the intersection of the input gene set, and unique genes from step 2, a Hypergeometric test is performed to calculate the p-value using Equation 1, similar to GREAT (RRID:SCR_00580)³. Here, M is the total number of genes in the genome, N_c is the number of genes in the input cluster/set c , N_p is the number of unique genes overlapping the peaks for given distance and n_{pc} is the intersection of two gene sets.

$$p - value = \sum_{x=n_{pc}}^{\min(N_p, N_c)} \frac{\binom{N_p}{x} \binom{M - N_p}{N_c - x}}{\binom{M}{N_c}} \quad (1)$$

Step 2 and 3 are repeated for every combination of gene cluster, peak set and distance and/or TADs. The final combined heatmap shows $-\log_{10}$ of the resulting p-values.

PEGS is implemented in Python 3, where we have reused functions from existing Python packages included with Python distributions, or available from the Python Package Index (PyPI). We also make use of BEDTools² for working with genomic intervals. We provide online documentation (<https://pegs.readthedocs.io/en/latest/>), and an example analysis with input data at the PEGS GitHub repository.

Operation

PEGS works with Python ≥ 3.6 and, when installed through pip, automatically installs all the dependencies. These are listed in *requirements.txt* file in our PEGS GitHub repository.

Figure 1. Cartoon showing peak expansion and overlapping TSSs in PEGS, with a specified genomic distance λ from the centre of the peak in both directions (a) where TSS₂ and TSS₃ are included and a TAD overlapping with the left peak in (b) where all four TSSs within the TAD are included

We provide extensive documentation online at <https://pegs.readthedocs.io> which includes easy-to-follow instructions about:

- Installation and system requirements
- Format of input files, output files, and graphics
- PEGS commands for standard operations, as well as running PEGS with additional input options, e.g. TAD definition files
- Creating customised reference TSSs files for new genomes

Results
Use Cases

Here, we present three use cases where we apply PEGS to different publicly available data sets. The format of input files is the same for all use cases below. Gene clusters are provided as text files with one gene symbol on each line; genomic region coordinates are provided in standard BED format. These input files for Use Case 1, as well as example analysis reproducing Figure 2A, are provided in our GitHub repository (<https://github.com/fls-bioinformatics-core/pegs>).

Use Case 1: Application of PEGS provides insight into glucocorticoid-mediated gene regulation in mouse liver
The first application (Figure 2A) uses up- and down-regulated glucocorticoid target genes obtained by an RNA-seq study of

liver samples from mice treated acutely with synthetic glucocorticoid dexamethasone or vehicle⁴. Corresponding GR ChIP-seq and chromatin accessibility data (DNase I hypersensitive (DHS) regions) were obtained from 5, and 6 respectively, whilst the mouse liver TAD boundaries were obtained from 7. Raw published datasets were downloaded from GEO Sequence Read Archive (RRID:SCR_005012) using *sra toolkit* v2.9.2 (<http://ncbi.github.io/sra-tools/>). Reads were aligned to the reference genome (mouse *mm10*), sorted and indexed using *Bowtie2* (v2.3.4.3, RRID:SCR_005476,⁸) and *SAMtools* (v1.9, RRID:SCR_002105,⁹). *MACS2* (v2.1.1.20160309, RRID:SCR_013291,¹⁰) was used to call peaks, using default settings. PEGS analysis shows strong association of dexamethasone up-regulated genes with dexamethasone-induced GR peaks at distances up to 500kbp from these peaks, but no enrichment of down-regulated genes, indicating distinct mechanisms of gene activation and repression by glucocorticoids. At the same time, there is promoter proximal enrichment for both up- and down-regulated genes in the DHS regions.

Use Case 2: PEGS demonstrates association of differential chromatin accessibility and gene expression during embryonic stem cell differentiation

Next, using PEGS, we calculated enrichment of gene clusters derived from single-cell RNA-seq and open chromatin regions defined by bulk ATAC-seq at three matching stages (ESCs-embryonic stem cells, day1 EpiLCs - epiblast-like cells, day2 EpiLCs) of early embryonic stem cell differentiation¹¹. The intergenic regions (peak sets) were defined as those with differential

Figure 2. PEGS applications: (A) gene expression, ChIP-seq and DNase I data on mouse liver, (B) gene clusters derived from scRNA-seq and intergenic putative enhancer clusters from bulk ATAC-seq at three matching early stem cell differentiation time-points. In both plots, numbers in the cells show common genes among the input genes (x-axis) and genes overlapping with expanded peaks (y-axis) and the colour shows $-\log_{10}$ of p-value (Hypergeometric test)

accessibility between any two time points and were clustered into four profiles based on z-score of tag densities, as described in 11. Similarly, differentially expressed genes were identified from pseudo-bulk gene expression data at each time point, and were clustered into four patterns. As shown in Figure 2B, strong association is observed between the matching gene expression (x-axis) and chromatin opening profiles (y-axis) at intergenic enhancers, reflecting correspondence between differential accessibility and gene expression changes.

Use Case 3: Extended application: PEGS detects enrichment of sleep trait SNPs in tissue-specific genes

Furthermore, we present an extended application of PEGS to GWAS data and find associations of SNPs for different sleep

phenotypes with sets of tissue-specific genes from the Genotype-Tissue Expression (GTEx) Portal, RRID:SCR_013042). For this purpose, we downloaded GWAS data from the Sleep Disorder Knowledge Portal (RRID:SCR_016611) for certain sleep associated phenotypes (with genome wide p-value cutoff $\leq 5e-8$) and calculated enrichment of tissue-specific gene lists defined using the GTEx Portal. Using median transcripts per million (TPM) data for different tissues in GTEx, a gene list for a tissue was defined as genes with 5x median TPM compared to the average in the remaining tissues.

In Figure 3, we show enrichment of SNPs from three sleep related phenotypes, namely chronotype, daytime sleepiness adjusted for BMI, and sleep duration. These enrichments are

Figure 3. Enrichment of sleep traits SNPs in tissue-specific gene lists (GTEx). The x-axis shows different tissue-specific gene lists, and y-axis shows three sets of sleep related SNPs, expanded to multiple genomic distances. The colour of the cells show $-\log_{10}$ of p-value of enrichment of corresponding gene list (x-axis) in the genes identified through overlap with expanded SNP intervals, the numbers in the cells show the common genes among the two (used in the calculation of Hypergeometric p-value)

calculated for tissue-specific genes lists created from GTEx for 22 tissues, the majority of them from the brain. Application of PEGS to these data reveals some strong associations, e.g. chronotype SNPs strongly enriched for genes expressed in liver and blood, while daytime sleepiness SNPs are enriched in gene sets for different brain tissues. Some of these associations are reported in the literature, e.g. daytime sleepiness SNPs in brain tissue¹², others may warrant further investigation.

Conclusions

Through the three different applications above, we demonstrate that PEGS is a versatile and highly efficient tool to integrate different genomic data, and is able to generate hypotheses for further analysis. The implementation of PEGS is highly efficient and as an example of computational efficiency, with pre-created reference TSS files, it only took 7.6 seconds to produce the output for **Figure 2A** on a laptop with Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz processor with 16GB RAM.

Furthermore, the user can adjust the background population and control for bias. For example, depending on the scientific question at hand, the reference gene interval file (TSSs BED file) could be limited to include only those genes known to be expressed in the tissue of interest. The efficiency of PEGS allows multiple gene and peak input files (e.g. with varying false discovery rate or fold-change cut-offs) to be tested quickly.

PEGS builds on some aspects of, and is complementary to, GREAT³, an existing tool, which performs functional enrichment of regulatory regions using annotations of nearby genes.

PEGS could also be used in conjunction with other tools to gain further mechanistic understanding (e.g. by finding enriched transcription factors with TFEA.ChIP¹³, ranking of their target genes with Cistrome-GO¹⁴ or BETA¹⁵, or predicting which TFs might regulate differentially expressed gene sets with Lisa¹⁶).

Data availability

All data underlying the results are available as part of the article or available publicly.

Software availability

Software available from Zenodo: <https://zenodo.org/record/5012058#.YNG7xmhKiUI>. It is easily installable through the Python Package Index (PyPI).

Source code available from Github: <https://github.com/fls-bioinformatics-core/pegs>.

Archived source code at the time of publication: <https://doi.org/10.5281/zenodo.5012058>¹

License: PEGS is distributed under BSD 3-Clause license.

Online manual: <https://pegs.readthedocs.io>

Acknowledgements

The authors thankfully acknowledge useful discussions with Magnus Rattray and Leo Zeef.

References

- Briggs P, Hunter AL, Yang S, et al.: **Pegs: An efficient tool for gene set enrichment within defined sets of genomic intervals.** Zenodo. 2021. <http://www.doi.org/10.5281/zenodo.5012058>
- Quinlan AR, Hall IM: **BEDTools: A flexible suite of utilities for comparing genomic features.** *Bioinformatics*. 2010; **26**(6): 841–842. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- McLean CY, Bristor D, Hiller M, et al.: **GREAT improves functional interpretation of cis-regulatory regions.** *Nat Biotechnol*. 2010; **28**(5): 495–501. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Caratti G, Iqbal M, Hunter L, et al.: **REVERBa couples the circadian clock to hepatic glucocorticoid action.** *J Clin Invest*. 2018; **128**(10): 4454–4471. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Grøntved L, John S, Baek S, et al.: **C/EBP maintains chromatin accessibility in liver and facilitates glucocorticoid receptor recruitment to steroid response elements.** *EMBO J*. 2013; **32**(11): 1568–1583. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Sobel JA, Krier I, Andersin T, et al.: **Transcriptional regulatory logic of the diurnal cycle in the mouse liver.** *PLoS Biol*. 2017; **15**(4): e2001069. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kim YH, Marhon SA, Zhang Y, et al.: **Rev-erba dynamically modulates chromatin looping to control circadian gene transcription.** *Science*. 2018; **359**(6381): 1274–1277. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Langmead B, Salzberg SL: **Fast gapped-read alignment with bowtie 2.** *Nat Methods*. 2012; **9**(4): 357–9. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Li H, Handsaker B, Wysoker A, et al.: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics*. 2009; **25**(16): 2078–9. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Zhang Y, Liu T, Meyer CA, et al.: **Model-based analysis of chip-seq (macs).** *Genome Biol*. 2008; **9**(9): R137. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Yang SH, Andrabi M, Biss R, et al.: **ZIC3 Controls the Transition from Naive to Primed Pluripotency.** *Cell Rep*. 2019; **27**(11): 3215–3227.e6. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Wang H, Lane JM, Jones SE, et al.: **Genome-wide association analysis of self-reported daytime sleepiness identifies 42 loci that suggest biological subtypes.** *Nat Commun*. 2019; **10**(1): 3503. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Puente-Santamaria L, Wasserman WW, Del Peso L: **TFEA.ChIP: A tool kit for transcription factor binding site enrichment analysis capitalizing on ChIP-seq datasets.** *Bioinformatics*. 2019; **35**(24): 5339–5340. [PubMed Abstract](#) | [Publisher Full Text](#)
- Li S, Wan C, Zheng R, et al.: **Cistrome-GO: A web server for functional enrichment analysis of transcription factor ChIP-seq peaks.** *Nucleic Acids Res*. 2019; **47**(W1): W206–W211. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Wang S, Sun H, Ma J, et al.: **Target analysis by integration of transcriptome and ChIP-seq data with BETA.** *Nat Protoc*. 2013; **8**(12): 2502–2515. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Qin Q, Fan J, Zheng R, et al.: **Lisa: Inferring transcriptional regulators through integrative modeling of public chromatin accessibility and ChIP-seq data.** *Genome Biol*. 2020; **21**(1): 32. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 26 August 2021

<https://doi.org/10.5256/f1000research.57359.r89907>

© 2021 Khan A. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Aziz Khan

Stanford Cancer Institute, Stanford University, Stanford, CA, USA

In this paper, the authors presented a Python-based command-line tool, PEGS, for gene set enrichment in association with genomic regions. PEGS computes the enrichment of gene sets with proximity-based association with genomic region sets. These associations are further restricted within the Topologically Associated Domains (TADs), which is good. The manuscript is moderately written and it provides three use cases of the tool.

The tool itself is very useful but it lacks several key options to give users the flexibility to customize the input data and also the output heatmap.

I have the following comments for the authors to address:

1. It is useful to restrict peak-gene association within the TAD boundaries, but it is not the case that all the interactions, such as enhancer-gene interactions occur within the TAD boundaries. The enhancer-gene communication can also occur outside topological domains or in-between TADs. Do authors plan to provide an optional feature to integrate chromatin interaction data, such as HI-C?
2. The command-line tool can be further improved by providing additional options to improve user experience and its usage. Below are some recommendations.
 - Currently, the peaks sets and gene lists inputs arguments are positional and the tools can only scan files available in the provided folders. Instead of looking into provided folders for BEDs files and gene lists, the argument should also allow chaining a list of bed files with a path. This is because in real analysis scenarios BEDs can be spread across multiple folders or a single folder can have other visible/hidden files. For example, I was testing the tool on a Mac machine, and PEGS started processing peaks for `.DS_Store`, which is the default directory structure and a hidden file.
 - The tool arguments could be: `pegs --peaks peaks/*.bed --genes genes/*.txt` and also `pegs --peaks A.bed B.bed --genes A.txt B.txt`

- The output heatmap should also have an option to generate vector-based plots, such as PDF or SVG.
 - Authors may consider adding additional options to adjust the heatmap, such as setting labels, dimensions, colors, and gene/peak set names.
3. Figures can be further improved.
4. Please highlight the limitations of the tool such as the enhancer–gene associations are solely based on proximity.
5. Providing an installation option through Conda using the bioconda channel (<https://bioconda.github.io/>) will be useful and it will increase the usage/availability of the tool.

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.

Reviewer Expertise: bioinformatics, gene regulation, regulatory genomics, epigenomics

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

Author Response 22 Oct 2021

Mudassar Iqbal

We thank the reviewer for constructive review and useful suggestions. We have updated the software taking into account the reviewer's suggestions, improved the manuscript overall and added more text. Here we will address their individual comments.

1 - It is useful to restrict peak-gene association within the TAD boundaries, but it is not the case that all the interactions, such as enhancer-gene interactions occur within the TAD boundaries. The enhancer-gene communication can also occur outside topological domains or in-between TADs. Do authors plan to provide an optional feature to integrate chromatin interaction data, such as Hi-C?

We would like to emphasise that the genomic distances are not restricted to TAD boundaries. PEGS provides the user with the option of supplying any distances. In addition, we also provide the user with the option to restrict peak expansion to TADs boundaries (if available), as a separate analysis (Fig 2A). We have revised the relevant text and we hope this will address reviewer's main concerns and clarify any confusion. Integration of HiC data is beyond the scope of this work, but we will think of ways to incorporate that in future developments of PEGS.

2 - The command-line tool can be further improved by providing additional options to improve user experience and its usage. Below are some recommendations.

- *Currently, the peaks sets and gene lists inputs arguments are positional and the tools can only scan files available in the provided folders. Instead of looking into provided folders for BEDs files and gene lists, the argument should also allow chaining a list of bed files with a path. This is because in real analysis scenarios BEDs can be spread across multiple folders or a single folder can have other visible/hidden files. For example, I was testing the tool on a Mac machine, and PEGS started processing peaks for .DS_Store, which is the default directory structure and a hidden file.*
- *The tool arguments could be: `pegs --peaks peaks/*.bed --genes genes/*.txt` and also `pegs -peaks A.bed B.bed --genes A.txt B.txt`*
- *The output heatmap should also have an option to generate vector-based plots, such as PDF or SVG.*
- *Authors may consider adding additional options to adjust the heatmap, such as setting labels, dimensions, colors, and gene/peak set names.*

We thank the reviewer as these are very useful suggestions and we have updated PEGS to a new version (please see latest version 0.6.2), which includes all of the above command line options. We also output the heatmap data as an excel file, so the user can customise their heatmaps/plots according to their choice/requirements. We have updated the documentation accordingly.

3 - Figures can be further improved.

We have improved and updated all of the figures (please see new version of the manuscript)

4 - Please highlight the limitations of the tool such as the enhancer-gene associations are solely based on proximity.

We have updated the manuscript text making it clear that our enrichment calculations are based on genomic proximity, expanding peaks (in both directions) with given distances (and/or TADs), and obtaining genes whose TSSs overlap with the expanded peaks.

5 - Providing an installation option through Conda using the bioconda channel (<https://bioconda.github.io/>) will be useful and it will increase the usage/availability of the tool.

We thank the reviewer; our tool is now installable through Conda.

Competing Interests: No competing interests were disclosed.

Reviewer Report 02 August 2021

<https://doi.org/10.5256/f1000research.57359.r89906>

© 2021 Zabet N. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Nicolae Radu Zabet

Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, E1 2AT, UK

Briggs and co-authors present a new tool called PEGS to generate gene set enrichment for ChIP-seq and DNase-seq datasets. In fact, the tool can be applied to any set of genomic intervals, including SNPs datasets. Generation of gene set enrichment for genomic intervals is a very important task and the authors propose an interesting approach to address it. Particularly, I appreciate the use of TADs to limit the expansion of genomic intervals. They also provide three use cases with different datasets and prove the applicability of this tool.

I have the following comments:

1. Do you consider alternative TSS? Would a gene with multiple TSSs be overrepresented or not?
2. Do you think that distal loops connecting TSS with enhancers residing outside of TADs would affect your results?
3. While readable, the resolution of figure 2 is low. I would advise the authors to upload a higher resolution figure.
4. For case1, maybe I missed it, but I think it would be interesting to interpret the results with or without TADs. This would allow us to see the impact of TAD annotation on the analysis.
5. I think the authors should add more explanations in the text about the results of their three cases.

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.

Reviewer Expertise: Computational biology, bioinformatics, chromatin and epigenetics

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

Author Response 22 Oct 2021

Mudassar Iqbal

We thank the reviewer for constructive comments, here we will address their points one by one.

1 - Do you consider alternative TSS? Would a gene with multiple TSSs be overrepresented or not?

We consider all TSSs defined in the given genome build, which can include multiple TSSs for some genes. When calculating the enrichment for gene sets obtained through overlap of TSSs with expanded peaks, we remove duplicates. Therefore, genes with multiple TSSs are not over-represented.

2 - Do you think that distal loops connecting TSS with enhancers residing outside of TADs would affect your results?

We have two scenarios for enrichment calculations in PEGS. First, the user can provide genomic distances which are not constrained to TADs. Hence enhancers residing in a separate TAD to the TSS could be included. Secondly, we provide an option to the user to constrain the peak expansion to TAD/subTAD boundaries, if available. This will exclude

enhancers outside the TAD boundaries, but the distances option can still be utilised to test multiple distances within and beyond TAD boundaries.

3 - While readable, the resolution of figure 2 is low. I would advise the authors to upload a higher resolution figure.

We agree, and have added a high-resolution version of Fig. 2

4 - For case1, maybe I missed it, but I think it would be interesting to interpret the results with or without TADs. This would allow us to see the impact of TAD annotation on the analysis.

This is related to point 2. In Fig.2A, we do provide the analysis with and without TADs for case 1. Enrichment for multiple distances is at the top two panels, and enrichment calculations using TADs are at the bottom. We have improved Fig 2 to make this clear.

5 - I think the authors should add more explanations in the text about the results of their three cases.

We have added more text in the manuscript, further explaining the three cases.

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

F1000Research