SOFTWARE TOOL ARTICLE

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

Peter Briggs[1], A. Louise Hunter[2], Shen-hsi Yang[3], Andrew D. Sharrocks[3], Mudassar Iqbal[4]

[1]Bioinformatics Core Facility, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, M13 9PL, UK
[2]Division of Diabetes, Endocrinology & Gastroenterology, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, M13 9PL, UK
[3]Division of Molecular & Cellular Function, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, M13 9PL, UK
[4]Division of Informatics, Imaging & Data Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, M13 9PL, UK

v2 **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.

Open Peer Review

Approval Status ✓ ✓

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

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.

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

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.

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 2; peer review: 2 approved]** F1000Research 2021, **10**:570 https://doi.org/10.12688/f1000research.53926.2

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

REVISED **Amendments from Version 1**

In this version, in response to reviewer's feedback, we have made following changes:

We have updated our software package to a newer version (0.6.3) with improved functionality in terms of command line options for input and output files. We have updated the documentation accordingly. We have also uploaded updated versions of all three figures and improved the text throughout the manuscript.

Any further responses from the reviewers can be found at the end of the article

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 using user-provided genomic distances 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, separately for each genomic distance and/or TADs. These tasks are performed in PEGS as follows:

1. Creating a gene interval file in BED (Browser Extensible Data) format for all TSSs 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 a specified genomic distance in both directions (and up to overlapping TAD boundaries, if provided). Intersection of these extended peaks with the gene intervals BED file from step 1 is calculated using BEDTools (RRID:SCR_006646)[2]. This leads to a gene set with TSSs overlapping with extended peaks.

3. Using the intersection of the input gene set, and unique genes from step 2 (thus removing genes with multiple TSSs), 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.

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

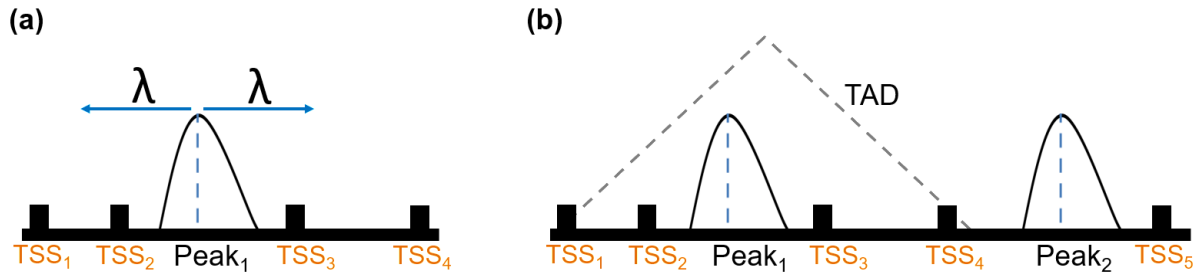

Figure 1. Cartoon showing peak expansion and overlapping TSSs in PEGS, with a specified genomic distance λ from the centre of the peak in both directions (**a**) where TSS_2 and TSS_3 are included, and a TAD overlapping with the left peak in (**b**) where all four TSSs within the TAD are included

Step 2 and 3 are repeated for every combination of gene cluster, peak set and genomic 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 BEDTools[2] 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. 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 the gene sets consisting of putative targets of glucocorticoid receptor (GR) in mouse liver. These are up- and down-regulated genes obtained by an RNA-seq study of liver samples from mice treated acutely with synthetic glucocorticoid dexamethasone or vehicle[4]. 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

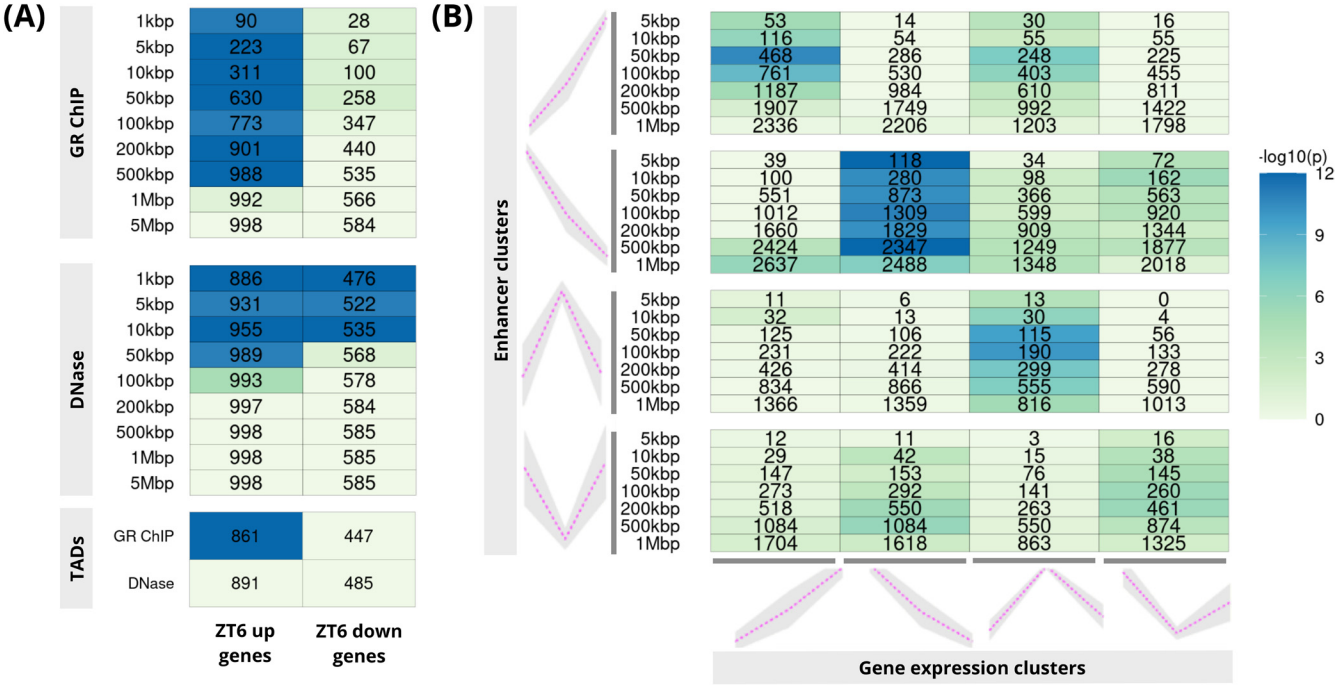

Figure 2. PEGS applications: (**A**) gene expression, ChIP-seq and DNase I data on mouse liver; upper two panels correspond to GR ChIP-seq and DNase peaks expanded to different genomic distances while the bottom panel shows both GR ChIP-seq and DNase peaks expanded to overlapping TAD boundaries (**B**) gene clusters derived from scRNA-seq and intergenic putative enhancer clusters from bulk ATAC-seq from 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).

sratoolkit 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. Using these GR ChIP-seq peaks and DHSs as peak sets, 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 (Figure 2A, top panel), 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 (Figure 2A, middle panel). On the other hand, PEGS analysis using TADs boundaries (Figure 2A, bottom panel) shows significant enrichment only in the case of up-regulated genes in GR ChIP-seq. This is suggestive of a direct role for GR in gene activation rather than repression.

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 time points (ESCs-embryonic stem cells, day1 EpiLCs - epiblast-like cells, day2 EpiLCs) during early embryonic stem cell differentiation[11]. Early embryonic development (naïve mouse ESCs to EpiLCs) involves large changes in the chromatin landscape through the action of many transcription factors and chromatin regulators leading to specific gene expression programs. Using our publicly available data[11], we defined our open chromatin peak sets as the intergenic regions with differential accessibility between any two time points. These 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 similarly clustered into four patterns across three time points. These constituted our gene sets for PEGS analysis. As shown in Figure 2B, application of PEGS to these data shows strong association 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 at corresponding time points. This application shows the utility of PEGS for integrating chromatin accessibility and gene expression data, leading to biologically meaningful association of enhancer and gene clusters.

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

Genome-wide association studies (GWAS) are commonly employed to study genotype-disease associations. Here, 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) which provides data, as well as analysis and visualisation resources for human genetic information regarding sleep and related traits. We extracted SNPs (single nucleotide polymorphisms) for certain sleep associated phenotypes (with genome wide p-value cutoff <=5 $e-8$). These SNPs constitute our input peak sets to PEGS, while we defined corresponding gene sets as tissue-specific genes from GTEx portal. These were created as following; we obtained median transcripts per million (TPM) data for different tissues in GTEx, and a gene list for a specific tissue was defined as genes with greater than 5x median TPM compared to the average in the remaining tissues.

In Figure 3, using PEGS, we show enrichment of SNPs from three sleep related phenotypes, namely chronotype, daytime sleepiness adjusted for BMI, and sleep duration. These enrichments are 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 tissue[12], 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 background population 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 analysis is limited to enhancer-genes associations based on genomic proximity. It builds on some aspects of, and is complementary to, GREAT[3], 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.ChIP[13], ranking of their target genes with Cistrome-GO[14] or BETA[15], or predicting which TFs might regulate differentially expressed gene sets with Lisa[16]).

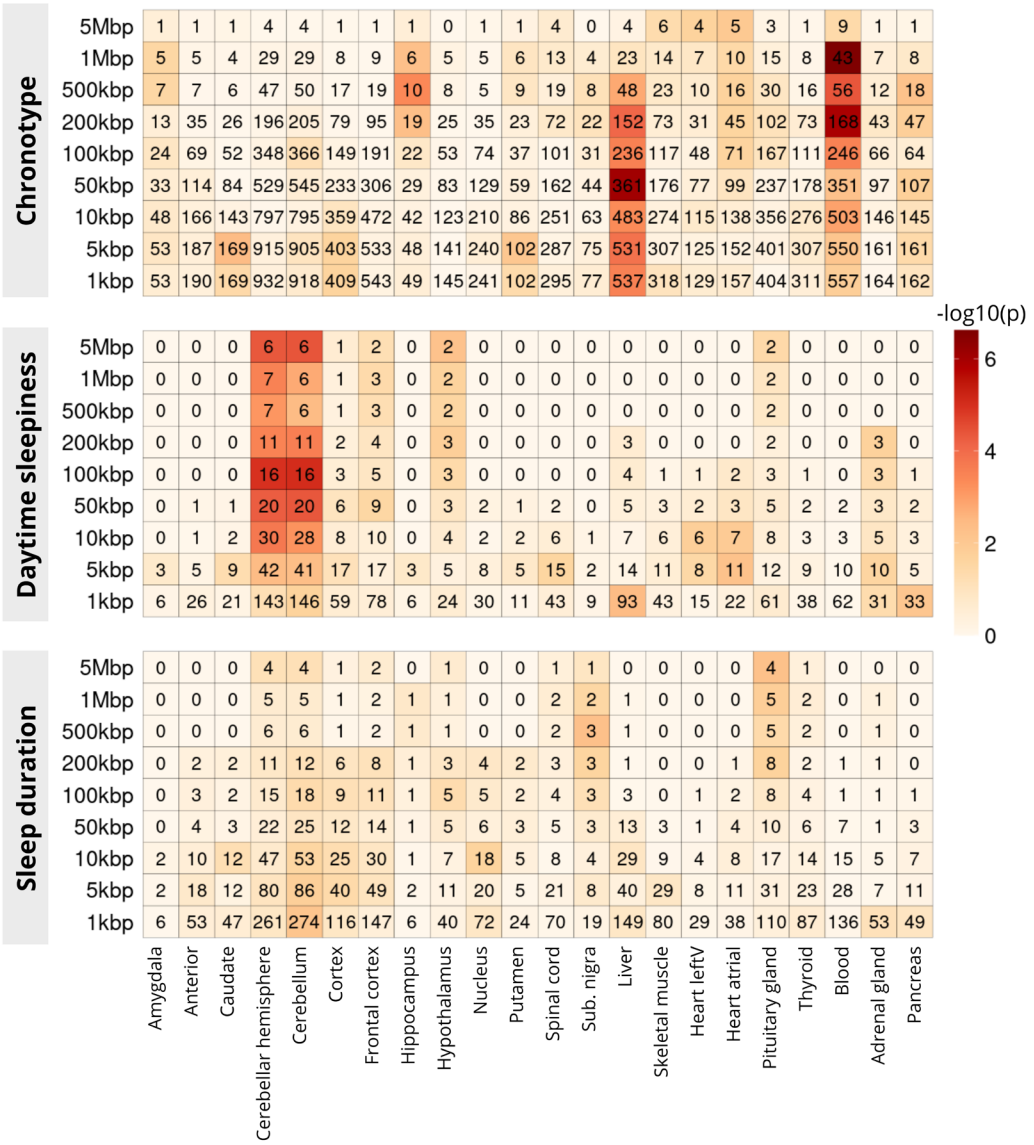

Figure 3. Enrichment of sleep traits SNPs in tissue-specific gene lists (GTEx). The x-axis shows different tissue-specific gene lists, and y-axis shows three sets of sleep related SNPs, expanded to multiple genomic distances. The colour of the cells show $-log_{10}$ of p-value of enrichment of corresponding 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)

Data availability

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

Software availability

Software is available from Zenodo: https://doi.org/10.5281/zenodo.5596224[1]. 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.5596224[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

1. Briggs P, mudassariqbal: **fls-bioinformatics-core/pegs: pegs-0.6.3 (0.6.3).** *Zenodo.* 2021.
http://www.doi.org/10.5281/zenodo.5596224
2. 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
3. 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
4. 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
5. 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
6. 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
7. Kim YH, Marhon SA, Zhang Y, *et al.*: **Rev-erbα dynamically modulates chromatin looping to control circadian gene transcription.** *Science.* 2018; **359**(6381): 1274–1277.
PubMed Abstract | Publisher Full Text | Free Full Text
8. 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
9. 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
10. 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
11. 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
12. 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
13. 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
14. 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
15. 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
16. 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 2

Reviewer Report 25 February 2022

https://doi.org/10.5256/f1000research.78711.r98741

✔ **Aziz Khan**
Stanford Cancer Institute, Stanford University, Stanford, CA, USA

The authors addressed all of my concerns/suggestions in the revised version.

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.

Reviewer Report 05 November 2021

https://doi.org/10.5256/f1000research.78711.r98740

✔ **Nicolae Radu Zabet**
Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, E1 2AT, UK

The authors have addressed all my comments.

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.

Version 1

Reviewer Report 26 August 2021

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

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, University of Manchester, Manchester, UK

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

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, University of Manchester, Manchester, UK

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.