Check for updates

SOFTWARE TOOL ARTICLE

SeqPlots - Interactive software for exploratory data analyses, pattern discovery and visualization in genomics [version 1; referees: 2 approved, 1 approved with reservations]

Przemyslaw Stempor, Julie Ahringer

The Gurdon Institute and the Department of Genetics, University of Cambridge, Cambridge, CB2 1QN, UK

v1 **First published:** 15 Nov 2016, **1**:14 (doi: 10.12688/wellcomeopenres.10004.1)
Latest published: 15 Nov 2016, **1**:14 (doi: 10.12688/wellcomeopenres.10004.1)

Abstract

Experiments involving high-throughput sequencing are widely used for analyses of chromatin function and gene expression. Common examples are the use of chromatin immunoprecipitation for the analysis of chromatin modifications or factor binding, enzymatic digestions for chromatin structure assays, and RNA sequencing to assess gene expression changes after biological perturbations. To investigate the pattern and abundance of coverage signals across regions of interest, data are often visualized as profile plots of average signal or stacked rows of signal in the form of heatmaps. We found that available plotting software was either slow and laborious or difficult to use by investigators with little computational training, which inhibited wide data exploration. To address this need, we developed SeqPlots, a user-friendly exploratory data analysis (EDA) and visualization software for genomics. After choosing groups of signal and feature files and defining plotting parameters, users can generate profile plots of average signal or heatmaps clustered using different algorithms in a matter of seconds through the graphical user interface (GUI) controls. SeqPlots accepts all major genomic file formats as input and can also generate and plot user defined motif densities. Profile plots and heatmaps are highly configurable and batch operations can be used to generate a large number of plots at once. SeqPlots is available as a GUI application for Mac or Windows and Linux, or as an R/Bioconductor package. It can also be deployed on a server for remote and collaborative usage. The analysis features and ease of use of SeqPlots encourages wide data exploration, which should aid the discovery of novel genomic associations.

Open Peer Review

Referee Status: ? ✓ ✓

	Invited Referees		
	1	**2**	**3**
version 1 published 15 Nov 2016	? report	✓ report	✓ report

1 **Simon J van Heeringen**, Radboud University Medical Center Netherlands

2 **Boris Lenhard**, Imperial College London UK, **Malcolm Perry**, Imperial College London UK

3 **Jean-Christophe Andrau**, Institute of Molecular Genetics of Montpellier (IGMM) - UMR5535 France, **Anne Coleno**, Institute of Molecular Genetics of Montpellier (IGMM) - UMR5535 France

Discuss this article

Comments (0)

Corresponding authors: Przemyslaw Stempor (p.stempor@gurdon.cam.ac.uk), Julie Ahringer (ja219@cam.ac.uk)

How to cite this article: Stempor P and Ahringer J. **SeqPlots - Interactive software for exploratory data analyses, pattern discovery and visualization in genomics [version 1; referees: 2 approved, 1 approved with reservations]** Wellcome Open Research 2016, **1**:14 (doi: 10.12688/wellcomeopenres.10004.1)

Grant information: This work was supported by the Wellcome Trust [101863].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: No competing interests were disclosed.

First published: 15 Nov 2016, **1**:14 (doi: 10.12688/wellcomeopenres.10004.1)

Introduction

Sequencing based techniques such as ChIP-seq and RNA-seq are widespread experimental tools that generate vast amounts of data for downstream analyses such as uncovering global patterns of genomic activity. After aligning sequence reads to the reference genome, read coverage is calculated. Visualizing coverage tracks using genome browsers is the simplest way to inspect the results. Nevertheless, calculating and plotting signals across groups of selected genomic locations is essential for genome-wide hypothesis testing and quantitative comparisons.

Typically, users plot the abundance of signal (e.g., read coverage) across a set of genomic regions (e.g., transcription start sites) either as a profile plot of average signal or as stacked rows of individual signals visualized as a heatmap. Such plots are usually generated using online or command line tools such as Galaxy/Cistrome, ngs.plot, and deeptools, or using custom scripts combined with plotting software such as Gnuplot[1–5]. We found that these methods were either laborious, as each plot needed to be set up individually, or were difficult to use by those with little computational training. These factors inhibited users from generating a large number of plots for data exploration.

To address this, we developed SeqPlots, a highly configurable, graphical user interface (GUI) operated application that rapidly generates publication quality average profile plots or heatmaps that can be clustered using different algorithms to uncover patterns within the data. A key feature of SeqPlots is the ability to select a set of features and signals, then rapidly plot them in any combination, facilitating wide data exploration.

Methods

SeqPlots can plot signals from any experimental or *in silico* data (e.g. ChIP-seq or RNA-seq read coverage, density of sequence motifs, mappability, nucleosome occupancy) over one or multiple sets of genomic features, (e.g. TSSs, gene bodies, peak calls). Users first add signal tracks and genomic feature files to an integrated SeqPlots database (see Table 1 for accepted file formats). Then any combination of signal and feature files in the database, together with any user entered sequence motifs, can be analyzed. Plots can be anchored at either end of a feature, at both ends, or at centers, and users can define which lengths of upstream and downstream sequence to plot. Additionally, three different methods can be used to cluster heatmaps: k-means, hierarchical clustering, and self organizing maps (unsupervised neural networks); heatmap rows can also be sorted by signal strength.

Table 1. File formats accepted by SeqPlots.

Genomic feature formats	
File formats	**Recognized extensions**
General Feature Format	gff
Browser Extensible Data	bed
General Transfer Format	gtf
Signal track formats	
File formats	**Recognized extensions**
bigWig Track Format[a]	bw
Wiggle Track Format[b]	wig
BedGraph Track Format[b]	bdg or bedGraph
Binary Sequence Alignment/Map[c]	bam

[a] preferred track format

[b] converted to bigWig upon upload

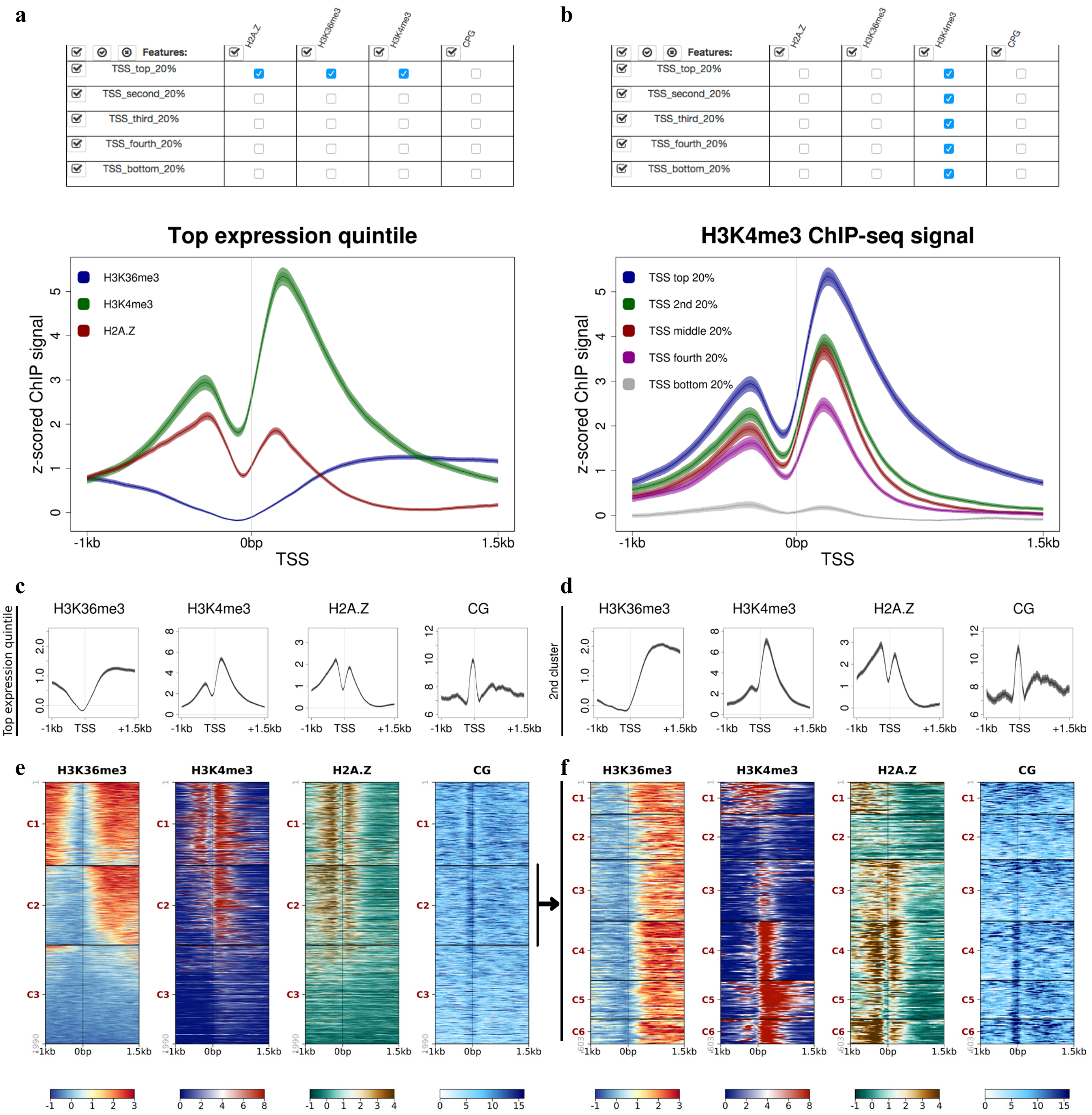
[c] coverage is calculated using all aligned reads

Implementation

SeqPlots utilizes indexing and the multi-layer summarization properties of bigWig files for rapid data acquisition[6], and precalculates and stores profiles for all combinations of selected signals and features. Users are presented with a clickable array of signal/feature pairs that can be plotted individually or in any combination in a matter of seconds. Average profile plots or heatmaps are immediately displayed as previews and can be downloaded as PDF files. Profile plots can display standard error and 95% confidence intervals. Spreadsheets with annotated heatmap clusters can be downloaded for downstream analyses such as additional clustering or gene enrichment analyses. Scaling, colors, axes, and titles are also easily configurable. Signal and feature files uploaded to the integrated SeqPlots database are available for use in later plot setups. Users can search and sort uploaded files, and annotate them with comments, user names and reference genome versions.

Use case

Figure 1 illustrates a typical use of SeqPlots. Five feature files in bed format containing genomic coordinates of protein coding genes in different expression bins were selected together with three bigWig signal files (normalized read coverage of H3K4me3, H2A.Z, and H3K36me3). In addition the dinucleotide motif CG was inputted and SeqPlots generated a CG density track for use in the analyses. A plot type anchored at the start position (the TSS) was then selected, and 1 kb upstream and 1.5 kb downstream of the TSS was specified. Following the setup and calculation, SeqPlots presented a clickable grid (top of Figure 1a,b). Selecting the desired combinations and plot type (average profile plot or heatmap) generates a plot. In Figure 1a, three signals (H3K36me3, H3K4me3, and H2A.Z) and one feature (top 20% TSSs) were selected for an average profile plot. For Figure 1b, these were deselected and a new combination was selected (H3K4me3 and all five TSS expression classes). For Figures 1c and d, single combinations of feature and signal were selected. A three-cluster heatmap was then generated using all four signal tracks, clustered using just the H3K36me3 signal. This simple clustering identified regions with bidirectional (C1), unidirectional (C2) or little (C3) H3K36me3. The unidirectional cluster (C2) was extracted from the cluster annotation spreadsheet and uploaded for further re-clustering. A self organizing map with 6 neurons was applied to the three other features – H3K4me3, H2A.Z and CpG, revealing clusters with different patterns of H3K4me3 and H2A.Z marking. For example, cluster C4 shows strong H3K4me3 downstream of the TSS and H2A.Z enrichment both upstream and downstream of the TSS whereas cluster C6 has a similar H3K4me3 pattern, but H2A.Z shows higher enrichment upstream of the TSS. Additionally, clusters C4–C6 have a stronger CpG signal at the TSS than clusters C1–C3. This simple example shows how SeqPlots can be used to



a

☑ ⊘ ⊗ Features:	☑ H2A.Z	☑ H3K36me3	☑ H3K4me3	☑ CPG
☑ TSS_top_20%	☑	☑	☑	□
☑ TSS_second_20%	□	□	□	□
☑ TSS_third_20%	□	□	□	□
☑ TSS_fourth_20%	□	□	□	□
☑ TSS_bottom_20%	□	□	□	□

b

☑ ⊘ ⊗ Features:	☑ H2A.Z	☑ H3K36me3	☑ H3K4me3	☑ CPG
☑ TSS_top_20%	□	□	☑	□
☑ TSS_second_20%	□	□	☑	□
☑ TSS_third_20%	□	□	☑	□
☑ TSS_fourth_20%	□	□	☑	□
☑ TSS_bottom_20%	□	□	☑	□

Figure 1. An example of SeqPlots workflow to analyze H2A.Z, H3K36me3, H3K4me3 and CpG density across *C. elegans* protein coding TSSs separated by expression quintiles. (**a**,**b**) Top, GUI interface showing clickable grid of signal/feature combinations. Bottom, plots resulting from the clicked selections. (**c**) Plots of individual signals across genes in top expression quintile anchored at TSSs, plotting 1 kb upstream and 1.5 kb downstream of TSSs. (**d**) Heatmaps generated using k-means clustering (3 clusters) of TSSs in top expression quintile, using H3K36me3 signal for clustering. (**e**) Average signal profiles and (**f**) heatmaps generated from cluster 2 (C2) in (**d**) made by downloading full cluster data and uploading file with cluster 2 regions. Heatmaps were clustered using H3K4me3, H2A.Z and CpG signals. Data used to generate this figure are available from GEO (H3K4me3: GSE28770 - https://www.be-md.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE28770; H3K36me3: GSE62833 - https://www.be-md.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62833; H2A.Z/HTZ-1: GSE49717 - https://www.be-md.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE49717). TSS annotations are from 7,8, or Wormbase/Ensembl 81 if a gene had no TSS annotation in either dataset (available from https://gist.github.com/Przemol/c5114067cc2dd236ed1dbcaf41003472). Genes were divided into expression bins using DCPM values from Gerstein *et al.*[9].

find relationships between genomic features and signals. The rapid plotting capability and ease of use of SeqPlots should facilitate wide exploration of high-throughput sequencing data, leading to the discovery of novel biological associations.

Software availability

SeqPlots is distributed as user-friendly stand-alone applications for Mac and Windows or Linux, and is available as an R programming language package from the Bioconductor repository. SeqPlots can be also deployed as a server application, which is useful for data sharing within laboratories, collaborative usage and remote work. SeqPlots is an open source and open development project: source code wiki, bug tracker and pull requests are available via GitHub.

Software is available from:

- http://przemol.github.io/seqplots (Mac, Windows, Linux, full documentation)
- http://bioconductor.org/packages/seqplots (R/Bioconductor)
- http://przemol.github.io/seqplots/#installation---server-deployment (server deployment)
- https://github.com/Przemol/seqplots (latest source code, open development tools, including wiki, bug tracker, and pull requests)

Archived source code as at the time of publication:

- DOI: 10.5281/zenodo.163638, https://zenodo.org/record/163638 (core R/Bioconductor package)[10]
- DOI: 10.5281/zenodo.163641, https://zenodo.org/record/163641 (Stand-alone GUI application for Mac, Windows and Linux)[11]

License: LGPL 2.1:
https://www.gnu.org/licenses/old-licenses/lgpl-2.1.html

Author contributions

PS conceived, designed, and wrote the software. JA contributed to the design and supervised the study. PS and JA wrote the manuscript.

Competing interests

No competing interests were disclosed.

Grant information

This work was supported by the Wellcome Trust [101863].

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

Acknowledgements

We are grateful to members of the Ahringer lab for feature suggestions and testing of SeqPlots.

References

1. http://www.gnuplot.info/.
2. Huber W, Carey VJ, Gentleman R, *et al.*: **Orchestrating high-throughput genomic analysis with Bioconductor.** *Nat Methods.* 2015; **12**(2): 115–121.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Liu T, Ortiz JA, Taing L, *et al.*: **Cistrome: an integrative platform for transcriptional regulation studies.** *Genome Biol.* 2011; **12**(8): R83.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Ramírez F, Dündar F, Diehl S, *et al.*: **deepTools: a flexible platform for exploring deep-sequencing data.** *Nucleic Acids Res.* 2014; **42**(Web Server issue): W187–191.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Shen L, Shao N, Liu X, *et al.*: **ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases.** *BMC Genomics.* 2014; **15**: 284.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Kent WJ, Zweig AS, Barber G, *et al.*: **BigWig and BigBed: enabling browsing of large distributed datasets.** *Bioinformatics.* 2010; **26**(17): 2204–2207.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Chen RA, Down TA, Stempor P, *et al.*: **The landscape of RNA polymerase II transcription initiation in *C. elegans* reveals promoter and enhancer architectures.** *Genome Res.* 2013; **23**(8): 1339–1347.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Kruesi WS, Core LJ, Waters CT, *et al.*: **Condensin controls recruitment of RNA polymerase II to achieve nematode X-chromosome dosage compensation.** *eLife.* 2013; **2**: e00808.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Gerstein MB, Rozowsky J, Yan KK, *et al.*: **Comparative analysis of the transcriptome across distant species.** *Nature.* 2014; **512**(7515): 445–448.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Stempor P, Ahringer J: **SeqPlots - core R/Bioconductor package.** *Zenodo.* 2016.
Data Source
11. Stempor P, Ahringer J: **SeqPlots - stand alone GUI application for Mac, Windows and Linux.** *Zenodo.* 2016.
Data Source

Open Peer Review

Current Referee Status: ? ✓ ✓

Version 1

Referee Report 30 November 2016

doi:10.21956/wellcomeopenres.10780.r17658

✓

Jean-Christophe Andrau, **Anne Coleno**

Institute of Molecular Genetics of Montpellier (IGMM) - UMR5535, Montpellier, France

Seqplots is an analysis and visualization software for genomic data sets, including ChIPseq, RNAseq and others. It allows investigating the pattern and abundance of signal coverage across genomic regions. Seqplots provides options for average profiling, performing heatmaps of normalized signal that can be clustered or not and further retrieved to explore the data. Seqplots is accessible as an application for all interfaces (Mac, Windows, Linux) or as R/Bioconductor package. The online documentation presents a dataset of a transcription factor and two epigenetic marks (H3K4me3 and H3K36me3) on the first chromosome of *C. elegans*.

Installation:
To challenge Seqplots, we tested our own data on the complete human genome. The latest version (3.0.12 MacOSX bundle version) of Seqplots was first downloaded on a Mac but yielded genome installation issues. We thus went on with an older version (1.7.16, MacOS bundle version 2.0.2) that contains two drosophila genome versions (dm3 and dm6), *C. elegans* genome ce10 and human hg19 and we proceeded with this one.

Genomes upload:
One can select the desired genome on the proposed list but the action of « install selected » stops half way after indicating « installing packages: BSgenome.Scerevisiae.U ». The blue progression line stops but no error message appears, leaving no option to correct any possibility of upload. There is no indication of genome format file needed to upload, whether it would be FASTA, indexed genome or anything else.

Data upload:
Data (bam, wig, bigwig) can be easily uploaded to Seqplots that automatically converts any format in bigwig to allow better handling. Data upload time is directly dependent on your data size. After upload, the file is directly converted to bigwig.

New Plot Set:
To generate a profile plot or heatmap, one has to chose two files: the file containing the signal intensity (wig, sam …) and the file containing the selected annotations of interest such as genes, protein coding genes, promoters. This second file can be at the bed, gtf or gff format.

Several files can be processed in a single analysis; the visualization tool uses results to plot individual or combined profiles as selected by the user.

Options to select:
- Bin size of the signal file has to be known in advance and it is important to adjust this parameter properly, ideally identical to the original data input bin size. If using a bin size below the actual data, it will introduce wholes in the plot. Conversely, using a larger bin size will smooth the data, which can be useful for clarity of the results but can also result in a loss of resolution. This point could be clarified in the documentation. We also note that the link to explanation is essentially inactive.

- The 'point view', 'end point' or 'midpoint' options are easy to understand as start, middle and end of the features. The 'anchored features' option is a bit less obviously accessible, and should be explained more clearly.

Profiles:
Profiles are easily generated and modified using indicated options, and overall functions well. That the user could directly change the labels below the plots could be a possible improvement.

Heatmaps:
This option also allows clustering of the data with a nice graphical interface, including nice color options. When the process is complete, this tool is very useful to identify classes of genes/features and to explore the data. However, we note that calculation time for heatmap generation can be limiting and varies quite a lot from one time to the other. Unfortunately, no error message appears to explain if something went wrong on the interface while generating the heatmap. Sometimes the message is blocked in "exporting results" but these are never exported.

General considerations and conclusion:
Seqplots can be used quite easily by non bio-informaticians with minimum training. We believe that this software fulfills many of the analysis options that biologists are looking for when dealing with high-throughput sequencing data sets, including ChIP-seq and RNA-seq. It has therefore a great potential of usage by the community of scientists interested in genomic science. We also found that improvements remain possible and suggest debugging more specifically the following points.

- The genome upload issue has to be fixed.
- Documentation should be developed for the sections 'genome upload',
- There is no possibility to run two jobs in parallel. Seqplots could allow the opening of two windows or more.
- Error messages have to be clearer and help the user to make a decision on what should be done.
- Heatmap generation jobs are often aborted during calculations.

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.

Competing Interests: No competing interests were disclosed.

Referee Report 28 November 2016

doi:10.21956/wellcomeopenres.10780.r17652

✓ **Boris Lenhard**, **Malcolm Perry**
Institute of Clinical Science, Imperial College London, London, UK

'SeqPlots' is clearly a mature package and it provides a graphical interface to make complex plots from within an web browser. I was able to get the package up and running, and generate plots from my own data very quickly after installation through the GUI, which was intuitively navigable. It is clear that a lot of work has gone into providing an impressive array of options to the user.

The manuscript itself gives an adequate, if brief, description of the purpose and typical output of the resource. It does not attempt to serve as user manual or tutorial.

A tutorial is provided as a separate, regularly updated document. The tutorial is very detailed, although a few relative links were broken. My one suggestion would be to provide easy access to the example data through the tutorial - I couldn't find it without digging into the package source. Adding an example line of `run(root='/path/to/ex/data')` in the tutorial would be a simple way to do this.

In the future, it may be worth considering developing a scriptable back-end to the package, so that users who are comfortable with R can automate their pipelines.

Overall, for those who want a graphical interface to make these plots this is a very useful resource.

Note: Malcolm Perry tested the software and wrote most of the above comments.

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.

Competing Interests: One of us (Malcolm Perry) is developing a R/Bioconductor package with some overlap in functionality to SeqPlots.

Referee Report 25 November 2016

doi:10.21956/wellcomeopenres.10780.r17648

? **Simon J van Heeringen**
Department of Molecular Developmental Biology, Radboud University Medical Center, Nijmegen, Netherlands

SeqPlots is a graphical, interactive tool for exploratory visualization of high-throughput sequencing data. To start with my major point, I regret to say that I can not recommend SeqPlots in its current form. The main issue I have is with installation. This is a common hurdle with bioinformatics software, however, it is an important factor for a software tool. I don't gain any satisfaction from pointing this out, as the author has clearly spent a lot of effort on creating installers for all three major operating systems. However, I was not able to install SeqPlots in any form. I do have to admit here that I am not experienced in R. However, I do use other R modules without problems and SeqPlots is positioned as a tool for those with little computational training.

I tried the following:

1. Linux AppImage (Ubuntu 14.04)

2. Linux zipped application (Ubuntu 14.04)

3. Linux R installation (two different computers running Gentoo Linux)

4. Windows MSI package (Windows 7)

5. Windows zipped application (Windows 7)

These all failed in different ways:
1. & 2. I get a “Loading, please wait” prompt, with “Seqplots is running: false” on the command line.
3. After the run() command I get a greyed out, non-responsive webpage.

4. “Installation has failed” (this was after I had to upgrade .NET to be able to run the installer).

5. There were so many files (700MB in 10,000+ files) that extracting/copying the application took too much time and broke it off.

I recommend that the authors simplify installation options and then make sure that these really work. For instance, for Windows, there are three different ways to install the package. It is a lot of work to test and support all these releases. In my case not one out of five options worked. Personally, I can highly recommend bioconda, however, other working options would also be fine.

Given that I could not use the application I have not been able to asses practical use of SeqPlots with my own data.

With that out of the way, I have to say that the application (as far as I could judge from the demo and the manuscript) seems to be a useful and interesting tool. Documentation is excellent at first glance, although I haven't practically used it. The software is available at Github and archived via Zenodo, perfect.

One thing that I wondered is how the authors address reproducibility issues. Are the parameters that were used for an analysis saved somewhere? As it is easy to generate many different plots, it would be good to have a way to keep track of this.

Some other minor remarks and points:

- The authors mention other tools that are “Slow and laborious” and “difficult to use by investigators with little computational training”. I would much prefer that the authors phrase this more positively. This publication will not be screened for novelty, so no reason to downplay other excellent, widely used tools. What are the key properties that are unique to SeqPlots?

- “Such plots are usually generated using online or command line tools such as Galaxy/Cistrome, ngs.plot, and deeptools, or using custom scripts combined with plotting software such as Gnuplot.” A bit pedantic, but I'm not sure anyone uses gnuplot in the various genomics fields. There are other methods that deserve mention, ggplot and R come to mind, and various other plotting tools that have more or less the same functionality.

- I am not a native English speaker, but I am unsure if “wide data exploration” is an existing concept.

- “Common examples are the use of chromatin immunoprecipitation for the analysis of chromatin modifications or factor binding, enzymatic digestions for chromatin structure assays, and RNA sequencing to assess gene expression changes after biological perturbations." Here, "factor binding" should probably be "transcription factor binding"?

- Are gzipped files accepted as input (for instance for bed and bigbed)?
- The distinction between the "Methods" and the "Implementation" section is a bit fuzzy to me.
- It would be good to add a LICENSE/COPYING file to the repository with the software license.
- Don't use the rainbow/jet palette as the default color scheme, it has too many shortcomings.

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.

Competing Interests: I am the author of a software tool with similar functionality.