<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>GWIPS-viz: 2018 update</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Michel, Audrey M.; Kiniry, Stephen J.; O'Connor, Patrick B. F.; Mullan, James P.; Baranov, Pavel V.</td>
</tr>
<tr>
<td><strong>Publication date</strong></td>
<td>2017-09-05</td>
</tr>
<tr>
<td><strong>Type of publication</strong></td>
<td>Article (peer-reviewed)</td>
</tr>
<tr>
<td><strong>Link to publisher's version</strong></td>
<td><a href="http://dx.doi.org/10.1093/nar/gkx790">http://dx.doi.org/10.1093/nar/gkx790</a></td>
</tr>
<tr>
<td><strong>Rights</strong></td>
<td>© The Authors 2017. Published by Oxford University Press on behalf of Nucleic Acids Research. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<a href="http://creativecommons.org/licenses/by-nc/4.0/">http://creativecommons.org/licenses/by-nc/4.0/</a>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact <a href="mailto:journals.permissions@oup.com">journals.permissions@oup.com</a> <a href="http://creativecommons.org/licenses/by-nc/4.0/">http://creativecommons.org/licenses/by-nc/4.0/</a></td>
</tr>
<tr>
<td><strong>Item downloaded from</strong></td>
<td><a href="http://hdl.handle.net/10468/4844">http://hdl.handle.net/10468/4844</a></td>
</tr>
</tbody>
</table>

Downloaded on 2018-06-29T02:36:52Z
GWIPS-viz: 2018 update

Audrey M. Michel, Stephen J. Kiniry, Patrick B. F. O’Connor, James P. Mullan and Pavel V. Baranov*

School of Biochemistry and Cell Biology, University College Cork, Cork, Ireland

Received August 08, 2017; Editorial Decision August 28, 2017; Accepted August 29, 2017

ABSTRACT

The GWIPS-viz browser (http://gwips.ucc.ie/) is an on-line genome browser which is tailored for exploring ribosome profiling (Ribo-seq) data. Since its publication in 2014, GWIPS-viz provides Ribo-seq data for an additional 14 genomes bringing the current total to 23. The integration of new Ribo-seq data has been automated thereby increasing the number of available tracks to 1792, a 10-fold increase in the last three years. The increase is particularly substantial for data derived from human sources. Following user requests, we added the functionality to download these tracks in bigWig format. We also incorporated new types of data (e.g. TCP-seq) as well as auxiliary tracks from other sources that help with the interpretation of Ribo-seq data. Improvements in the visualization of the data have been carried out particularly for bacterial genomes where the Ribo-seq data are now shown in a strand specific manner. For higher eukaryotic datasets, we provide characteristics of individual datasets using the RUST program which includes the triplet periodicity, sequencing biases and relative inferred A-site dwell times. This information can be used for assessing the quality of Ribo-seq datasets. To improve the power of the signal, we aggregate Ribo-seq data from several studies into Global aggregate tracks for each genome.

INTRODUCTION

Ribosome profiling (Ribo-seq) is a biochemical technique that utilizes high throughput sequencing that captures the mRNA fragments that are protected by actively translating ribosomes (1) thereby providing Genome-Wide Information on Protein Synthesis (GWIPS) (2). Ribo-seq was first carried out in Saccharomyces cerevisiae (1) and has since been used in many organisms resulting in a substantial growth in the number of published datasets. The numerous applications of the ribosome profiling technique as well as its limitations are described in details elsewhere (3–14). While the majority of Ribo-seq datasets represent footprints of elongating ribosomes, a number of studies have used protocols for enriching footprints deriving from initiating ribosomes and more recently a modification of the ribosome profiling protocol allowed footprinting of scanning ribosomes (15).

To account for differences in mRNA abundance, most Ribo-seq studies also generate parallel datasets where total mRNA (or total RNA) is randomly degraded and subsequently sequenced. Here we refer to such datasets as mRNA-seq. To date, the majority of published Ribo-seq/mRNA-seq raw sequencing data have been deposited in NCBI’s Sequence Read Archive (SRA) (16).

The GWIPS-viz browser (http://gwips.ucc.ie/) uses the functionality of the UCSC Genome Browser (17) to provide visualizations of Ribo-seq coupled with mRNA-seq controls so that users can freely explore pre-populated Ribo-seq/mRNA-seq tracks without the need to download, preprocess and align raw sequencing data to the corresponding genomes. Since its original publication (18), we have striven to expand the repertoire of Ribo-seq/mRNA-seq data hosted on GWIPS-viz. We have also incorporated additional tracks as well as improved visualizations to help users better interpret the Ribo-seq/mRNA-seq data.

New genomes in GWIPS-viz

In 2014, GWIPS-viz provided Ribo-seq/mRNA-seq data for nine genomes: Homo sapiens (hg19), Mus musculus (mm10), Danio rerio (danRer7), Caenorhabditis elegans (ce10), S. cerevisiae (sacCer3), Escherichia coli K12 (ASM584v2), Bacillus subtilis (11/09/2009), human cytomegalovirus (HHV5 strain Merlin) and bacteriophage lambda (NC_001416). Today GWIPS-viz provides Ribo-seq/mRNA-seq data for an additional 14 genomes: Rattus norvegicus (rn6), Xenopus laevis (v6.0), Drosophila melanogaster (dm3), Trypanosoma brucei brucei (TriTrypDB TREU927 – v 5.1), Plasmodium falciparum (ASM276v1), Schizosaccharomyces pombe (ASM294v2), Neurospora crassa (or74a/GCF_000182925.2_NC12), Arabidopsis thaliana (Nov-2013), Zea Mays B73 (GCF_00005005.1_NC_024459.1), E. coli BW25113 (ASM75055v1), Caulobacter crescentus (ASM2200v1), Streptomyces coelicolor (ASM20383v1), Staphylococcus aureus USA300_FPR3757 (ASM1346v1), S. aureus NCTC
New tracks in GWIPS-viz

As well as the addition of new genomes to GWIPS-viz, the number of hosted tracks has grown by 10-fold. This is largely a result of our automated computational pipeline for the integration of new Ribo-seq and mRNA-seq data for genomes already in the browser, bringing the total number of tracks to 1792 tracks across the 23 genomes. The increase has been particularly substantial for Ribo-seq data generated for human as well as for mouse and S. cerevisiae. New data since the original GWIPS-viz publication include: H. sapiens hg38 assembly (19–49), H. sapiens hg19 assembly (43,50–54), M. musculus (51,52,55–75), R. novegicus (76–78), D. rerio (79–81), X. laevis (82), C. elegans (83–85), D. melanogaster (86–88), T. brucei brucei (89), P. falciparum (90), S. cerevisiae (15,64,82,85,91–109), S. pombe (82,110), N. crassa (111), A. thaliana (112–114), Z. Mays B73 (115), E. coli K12 (116–124), E. coli BW25113 (125), B. subtilis (126), C. crescentus (127), S. coelicolor (128), S. aureus USA300_FPR3757 (129), S. aureus NCTC 8325 (130). This expansion of datasets allows for improved cross-species comparison of orthologous genes while the availability of datasets from multiple research groups permits the assessment of technical reproducibility of the ribosome densities (131).

In addition to individual tracks reflecting Ribo-seq data generated under different conditions for each study, we aggregate each study’s data into an All track. We then aggregate the All tracks from each study into a Global Aggregate track for each genome (Figure 1A–D). This has the effect of improving the overall Ribo-seq signal by reducing the contribution of dataset specific biases and stochastic noise due to low coverage. The increased number of datasets is expected to yield higher sensitivity. The Global Aggregate tracks are set as the default for each genome and users can turn on/off each study’s data contribution to the aggregated data and then refine the visualizations by turning on/off individual tracks in each study. In addition, we provide Global Aggregate tracks through the UCSC Genome Browser for the human hg38 and hg19 assemblies.

We have also incorporated a new type of data into GWIPS-viz. Recently the Press group developed a technique called translation complex profile sequencing (TCP-seq), where ribosome subunits are cross-linked to mRNA, allowing footprinting of both elongating and scanning ribosomes (15). We now provide this data in the Small Ribosomal Subunits (Footprints) track group for S. cerevisiae. Given that TCP-seq is a powerful tool for studying translation initiation (132), we anticipate that the experimental protocol will be adapted for other species quite soon and we will strive to incorporate new TCP-seq data into GWIPS-viz as it becomes available.

For S. cerevisiae we also generated an additional gene annotation track from transcript isoform sequencing (TIF-seq) data (133). Given that Saccharomyces Genome Database (SGD) (134) and Ensembl (135) gene annotations for S. cerevisiae do not include UTR regions, and given that Ribo-seq has shown that extensive translation occurs in 5' leader regions (1) and to a lesser extent in 3' trailer regions (97), we thought it useful to integrate the 5' leader and 3' trailer gene annotations as interpreted by (133) from their TIF-seq data for S. cerevisiae.

With respect to 5' leader region delineation, we incorporated Riken 5' cap analysis gene expression (CAGE) data (136) as a permanent track in the Annotations Tracks and External Data track group for the human hg19 assembly. At the time of the incorporation, Riken CAGE data tracks for other GWIPS-viz assemblies were not available for permanent track integration. However, as GWIPS-viz also now includes the UCSC Genome Browser’s Track Hub functionality (137), we provide Riken’s FANTOM5 tracks for hg38, hg19, mm9 and rn6 as public track hubs. While these tracks are hosted and managed by the Riken group on their own server, a simple connection makes it easy to explore their CAGE data in conjunction with our Ribo-seq/mRNA-seq tracks in GWIPS-viz.

Initially we did not provide UCSC Genome Browser’s custom track feature (138) in GWIPS-viz. The custom track is only accessible to the user who uploads it, i.e. it is not a publicly available track. Many GWIPS-viz users, however, expressed an interest in the custom track feature as a means to explore their own Ribo-seq data in the context of published data and so we now include it. The custom track feature is also particularly useful for users of RiboGalaxy (139), a Galaxy based platform (140) that we have developed specifically for processing, mapping and analysing Ribo-seq data. Researchers can use the GWIPS-viz suite of tools in RiboGalaxy to generate Ribo-seq profiles that infer either the A-site (elongating ribosomes) or P-site (initiating ribosomes) from either the 5' end or the 3' end of Ribo-seq reads and the resulting profiles can be directly visualised as custom tracks in GWIPS-viz. The direct interface between GWIPS-viz and RiboGalaxy also allows data from GWIPS-viz to be retrieved into RiboGalaxy. We also provide a direct link to RiboGalaxy (http://ribogalaxy.ucc.ie/) from the GWIPS-viz homepage.

Improvements in data visualizations

Previously for bacterial genomes (E. coli K12, B. subtilis) our Ribo-seq and mRNA-seq profiles on GWIPS-viz did not provide strand orientation information. Our Ribo-seq density plots also used the center-weighted approach (141) to infer ribosome A-sites. Since then, several studies have shown that inferring the ribosome decoding center from the 3' ends of bacterial Ribo-seq data is more accurate (124,142,143). We decided to carry-out an overhaul of our bacterial tracks and now provide strand orientation information using the UCSC Genome Browser overlay functionality (144) in addition to A-site inference using a fixed offset from 3' footprint ends (Figure 1A,B). We have extended these improvements to the new bacterial genomes we now host in GWIPS-viz (E. coli BW25113, C. crescentus, S. coelicolor, S. aureus NCTC 8325, S. aureus USA300_FPR3757).

Recently we integrated the multi-region exon-only view (17), which is particularly useful for displaying Ribo-seq data for higher eukaryotes where exonic regions may be interrupted by long intronic regions (Figure 1C,D).
Figure 1. Exploring ribosome profiling data using GWIPS-viz. (A and B) Strand specific representation of the data for overlapping genes nudG and ynjH in the E. coli genome. In panel A, the Ribo-seq and mRNA-seq reads mapping to the forward strand (red) and to the reverse strand (blue) are both displayed. In panel B, only the reads mapping to the reverse strand are displayed. The profiles were generated using the Global aggregate tracks for E. coli in GWIPS-viz. (C and D) Aggregated human Ribo-seq data (red) at the SLC35A4 locus show that most of translation takes place at the uORF that spans the first three exons rather than the CDS (50,146,147). The exon-only view of the SLC35A4 locus improves the visualization of the translated uORF, the conservation of which is shown using the 100 vertebrates basewise conservation by PhyloP (148). (E) A RUST metafootprint profile that reveals the influence of mRNA codons on the relative read density in the vicinity of the ribosome is shown in grey in the top panel (145). The Kullback-Leibler divergence (blue for a single codon, green for adjacent codons) indicates the influence of each mRNA location on the frequency of ribosome footprint occurrence in the library. This is an example of a dataset with low sequencing biases, where the A-site codon influence is the highest. The lower left panel shows RUST estimates of relative codon decoding rates. The lower right panel shows the triplet periodicity signal (1,149) for individual read lengths. Panel E is taken from GWIPS-viz for study (20). (F) A screen-shot of the Downloads page that provides Ribo-seq and mRNA-seq read alignments for all tracks available in GWIPS-viz.
For higher eukaryotic datasets, we also now provide characteristics obtained with RUST (145). RUST utilizes Riboseq Unit Step Transformation to normalize ribosome profiling data. It further provides characteristics of ribosome profiling datasets among which is a metafootprint profile which shows the difference between observed (experimental) and expected (equiprobable) frequencies of specific sequences (commonly codons) in the vicinity of a ribosome footprint. The expectation is that the highest variation in codon frequencies should occur at the ribosome decoding center (A-site) (Figure 1E). A high variation at the end of footprints would occur due to sequencing biases. Thus, metafootprint profiles can be used for assessing the level of sequencing biases in individual datasets. Clicking on each study link in the GWIPS-viz genome page will open a new page with the link to the RUST quality plots which include the RUST metafootprint profile as well as a plot showing triplet periodicity for reads of different lengths. The RUST plots also include a panel that shows the relative inferred A-site dwell times for each amino acid.

**Download Ribo-seq and mRNA-seq alignments**

Following user requests, we added the functionality to download our genomic alignments in bigWig format. While the Table Browser provides the option to download our Ribo-seq and mRNA-seq alignments in bedGraph format, many users requested our original alignment files. Hence, we built a separate Downloads page (Figure 1F) for this purpose. For each Ribo-seq study hosted on GWIPS-viz, users can download (1) ribosome profiles of elongating ribosomes (number of footprints whose inferred A-site match a specific coordinate), (2) Ribo-seq and (3) mRNA-seq coverage plots that provide the number of reads that map to each coordinate. Where available, data enriched with footprints of initiating ribosomes, represented as coordinates of inferred P-site codons, can also be downloaded. In addition, footprints of small ribosome subunits generated by TCP-seq (15) are available for download for *S. cerevisiae* as coverage plots.

**FUTURE PLANS**

The development of an automated computational data integration pipeline has greatly helped us to keep pace with the flux of new Ribo-seq data for genomes already existing in GWIPS-viz. We do, however, still have some backlog in terms of Ribo-seq data generated for genomes that we need to manually add to GWIPS-viz. We are examining ways in which we can improve our capacity to add new genomes particularly genomes that are not hosted on the UCSC Genome Browser.

We also aim to continue to improve the visualizations of Ribo-seq data. We wish to extend the overlay functionality with strand-specific display to all the data tracks in GWIPS-viz (currently provided for bacterial genomes only). Conversely, we want to provide RUST characteristics for Ribo-seq data generated for bacteria. This requires adapting the RUST programming code to use 3′ end offsetting for A-site inference which we plan to do soon. In addition, we want to use RUST parameters and other Ribo-seq specific parameters such as the triplet periodicity, to develop a quality scoring method and provide this information on GWIPS-viz for each dataset. This will allow users to compare Ribo-seq datasets in terms of their quality across studies. It will also help us and our user-base to determine what governs Ribo-seq data quality and how it can be maintained and improved in future Ribo-seq data.

We also wish to avail of new functionality as it becomes available on the UCSC Genome Browser. For this reason, we have carried out several upgrades of the GWIPS-viz browser to keep in-line with the UCSC Genome Browser. The recent exon-only view is one such example which is particularly beneficial for eukaryotic organisms that extensively use RNA splicing, permitting genomic alignments of Ribo-seq data to be explored without intervening introns.

**DATA AVAILABILITY**

GWIPS-viz is publicly and freely available at (http://gwips.ucc.ie/).

**ACKNOWLEDGEMENTS**

We wish to thank the following people who have helped at different stages with the continuous development of GWIPS-viz: Vimalkumar Velayudhan, Claire A Donohue, Svetlana Koroteeva, Kevin O’Riordan, Matthias Claes and Sandor Dedeyne. We also wish to thank the UCSC Genome Browser Development team who have always been so generous with their help.

Finally, we wish to thank all the world-wide users of GWIPS-viz as well as those users who have posted questions and suggestions on our GWIPS-viz forum (http://gwips.ucc.ie/Forum).

**FUNDING**

Science Foundation Ireland [12/IA/1335 to P.V.B]; A.M.M. and S.J.K. wish to acknowledge personal support from the Irish Research Council. Funding for open access charge: Science Foundation Ireland.

Conflict of interest statement. A.M.M. and P.V. B. are co-founders of Ribomaps Ltd., a company that offers ribosome profiling analysis that can be affected financially by the growing popularity of ribosome profiling.

**REFERENCES**


