medRxiv preprint doi: https://doi.org/10.1101/2024.09.18.24313945; this version posted September 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

ExonViz: A website and Python package to visualize transcripts and genetic variants

Redmar R. van den Berg[1,2,3], Marlen C. Lauffer[1,2], and Jeroen F.J. Laros[1,4]

[1] *Department of Human Genetics, Leiden Univerisity Medical Center, The Netherlands*
[2] *Dutch Center for RNA Therapeutics, Department of Human Genetics, Leiden Univerisity Medical Center, The Netherlands*
[3] *Department of Hematology, Leiden University Medical Center, The Netherlands*
[4] *Department of Bioinformatics and Computational Services, National Institute of Public Health and the Environment (RIVM), The Netherlands*

Abstract

Summary

ExonViz is an application and website that creates biologically accurate transcript figures, including features such as coding regions, genetic variants and exon reading frames. Transcript definitions are automatically retrieved from Ensembl and RefSeq. We illustrate the full functionality of ExonViz by generating a figure for all ClinVar variants reported in *CYLD*.

Alvailability and Implementation

ExonViz is available online via Dutch Center for RNA Therapeutics and can be installed locally via PyPI.

Contact

Redmar R. van den Berg

Supplementary information

Extensive documentation on ExonViz is available on Read the Docs.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.18.24313945; this version posted September 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

Introduction

The visualization of transcripts, including features like coding and non coding regions, reading frames and the mutational landscape is important within the field of clinical and human genetics (Walker et al., 2023). Especially when new genes or transcripts are discovered, illustrating the exon structure and how variants are distributed across the gene is common practice. These illustrations are also used to assess potential genetic treatment options (e.g., canonical exon skipping), in teaching settings, in diagnostics, to identify mutational hotspots and for genetic counseling.

To date, most researchers have to resort to manually drawing transcripts with tools like Illustrator, Photoshop or BioRender, or forgo illustrations altogether. Some R scripts have been made available that aid in drawing of transcripts like ggtranscript (Gustavsson et al., 2022) or wiggleplotr (Alasoo, 2017). Additionally, web tools like genepainter (Mühlhausen et al., 2015) are useful to compare different transcript isoforms. However, specifically for drawing reading frames and for obtaining illustrations in a timely manner, no proper tool is available. Furthermore, a tool must be quick and easy to use if it is to be utilized in clinical and day to day settings, rather than to create a bespoke figure for a manuscript or presentation. Moreover, knowledge on the exon frames aids in the assessment of pathogenicity of genetic variants using the ACMG-AMP guidelines (Richards et al., 2015) when evaluating exon spanning deletions, and when interpreting the effects of splice altering variants (Walker et al., 2023).

Here we present ExonViz, an application that automatically creates biologically accurate transcript visualizations, requiring minimal user input. By using arrows, notches and straight edges to indicate the reading frames of exons, it helps users identify which exons share common reading frames. This aids in genetic counseling, genetic therapy assessment and educational settings.

Availability

To facilitate ease of use, ExonViz is available as a web application on the website of the Dutch Center for RNA Therapeutics. The web interface features the most commonly used options with sensible default values. For more advanced uses, ExonViz is also available as a Python programming library accompanied by a command line interface, which can be installed via PyPI. Extensive documentation on ExonViz is available on Read the Docs. Figures generated by ExonViz are free to use under the Creative Commons BY license. The source code itself is available on Github under the AGLP-3.0 license.

The command line interface allows for more fine grained control over both the visualization parameters (such as exon and variant colors and names) as well as the transcript, allowing users to modify existing transcripts or to specify their own. This is particularly useful to show the effect of splice altering variants, to visualize novel transcripts and to show the distribution of variants across a transcript, as can be seen in the online documentation.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.18.24313945; this version posted September 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

Usage

The minimum input required is a gene name, a transcript identifier or an HGVS description. If a gene name is specified, ExonViz will automatically determine the corresponding MANE Select transcript (Morales et al., 2022), since this is usually the most relevant transcript (Pozo et al., 2022). In cases where multiple MANE Select transcripts have been defined for a gene, the first transcript will be used. ExonViz will automatically fetch annotations like exon structure and coding region for the selected transcript. By default, only the coding region of the transcript will be shown, but the non coding region can optionally be included. The non coding region will be drawn at half the `height` of the coding region, as can be seen in Figure 2.

The transcript layout can be modified using the `scale` and `width` parameters. The `scale` parameter determines how many pixels should be used per base pair (default is 1), while the `width` parameter determines the width of the page. Exons that are too large to fit on the page will be split over multiple rows. Exons can also be split if they are close to the end of the page (e.g. Exon 4 in Figure 3). Since the arrows and notches take up a certain amount of space (relative to the `height`), some exons have to be drawn at a scale larger than 1. If this is the case, the user will be notified of the minimum scale at which the transcript can be drawn.

The variants specified in the HGVS description will be visualized at their correct location, as long as the *first* position of each variant is located inside an exon. Only variants that start inside an exon will be included in the figure, otherwise the variant is discarded and a warning will be shown to the user. This means that a big deletion that starts in an exon and includes part of an intron will be shown, whereas a big deletion that starts in an intron and includes part of an exon will be removed. If any variants are drawn on a transcript, a legend is added at the bottom of the figure to indicate the color and name for each variant.

Materials and Methods

ExonViz is written in Python 3, using Flask for the web interface. It uses the public Mutalyzer API (Lefter et al., 2021) to fetch transcript information. This gives ExonViz access to all transcripts defined in the RefSeq (O'Leary et al., 2016) and Ensembl (Harrison et al., 2023) databases across many species, ranging from human and mouse to fruit fly and coelacanth. Transcripts and annotations defined on the reverse strand are reversed on the fly, so ExonViz always visualizes transcripts in their forward orientation.

The start- and end phases of an exon refer to the alignment between the exon boundary and the reading frame. If the first base of an exon is also the first base of a codon, the phase is 0. If an exon starts at the second base of a codon, the start phase of the exon is 1, etc. The same holds for the end phases.

To visualize this, an exon ending in phase 1 is drawn with an arrow, and the

medRxiv preprint doi: https://doi.org/10.1101/2024.09.18.24313945; this version posted September 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

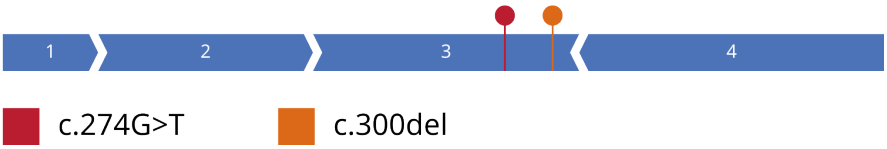

Figure 1: Coding region of *SDHD*(NM_003002.4), two variants in exon 3 are indicated using pins

start of the next exon is drawn with a notch, to signify the exons fit together. Conversely, a phase 2 exon ends with a notch, and starts with an arrow, while phase 0 is shown as two straight edges. If there are conflicting phases between exons, this is clearly visible from the figure.

The output of ExonViz is an SVG figure generated using the svg-py library, which can be used directly or modified using modern graphical editing programs. It is also possible to output the transcript and variants in TSV format, edit the transcript using any text editor or spreadsheet program, and draw the modified transcript using ExonViz. The online documentation has a number of examples of custom transcripts that can be visualized this way.

To increase the visibility of exon start- and -end phases for transcripts, we have reached out to the UCSC Genome Browser (Nassar et al., 2023) with the request to add this information to the gene tracks. This functionality has since been added to the Genome Browser. Simply hover your mouse over any exon in the Genome Browser and a pop-up will indicate the start- and end phase of the current exon, and whether the exon is in frame. Currently, the USCS Genome Browser is not able to visualize arrows and notches as used in ExonViz.

Results

ExonViz is an application that allows users to automatically draw biologically accurate transcript, including features like coding and non coding regions, exon reading frames and genetic variants. ExonViz can be employed to draw transcripts from any species where the transcript information can be retrieved (Lefter et al., 2021), requiring only a gene, transcript or specific variant(s) as input.

Figure 1 shows the default behaviour when a transcript (with variants) is visualized with ExonViz. It will render the coding regions and add the exon number to each exon. Any variants specified will be indicated using a colored pin, using different colors to distinguish the variants. The input to generate this figure was $NM_003002.4 : c.[274G > T; 300del]$.

Settings can be manually modified as illustrated in Figure 2, which shows the same transcript as Figure 1. Firstly, the figure now includes the non coding regions (drawn thinner) and the numbering of the exons has been removed. Notice how the last exon is split over two rows, since it is too large to be drawn at the specified `width`. The shape of the variants has been changed to a less

medRxiv preprint doi: https://doi.org/10.1101/2024.09.18.24313945; this version posted September 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

Figure 2: Full transcript for *SDHD*(NM_003002.4), including non coding regions. The same two variants as in Figure 1 are shown, this time annotated on RNA

intrusive `bar` instead of the default `pin`, and the `color` of the exons has also been changed. Finally, the `height` of the exons has been increased, which also influences the size of the arrows and notches. Since the figure has been scaled to fit on the page, increasing the `height` gives the transcript a more stocky appearance. Finally, notice that the variants are annotated on the RNA, not the DNA (as can be seen from the *r.* instead of *c.* in the variant description).

More advanced figures can be generated by using the command line version of ExonViz, as illustrated in Figure 3. This figure depicts all ClinVar variants in the coding region of transcript NM_001378743.1 of *CYLD*, grouped into three categories. The figure includes only the coding region for *CYLD*, which starts in exon three. The fourth exon has been split over two rows, since it would not fit on a single row. This transcript has to be drawn at a scale of at least 1.2, since exon six is too small to be drawn at a smaller scale.

Figure 3 clearly shows that the first part of exon sixteen contains a large number of pathogenic variants, indicating a mutational hotspot. Finally, the figure illustrates how ExonViz collapses the legend if multiple variants have the same name and color.

Figure 4 shows a section of transcript ENST00000357033.9 of the *DMD* gene, the largest human gene on the genome, containing 79 exons. The region that is shown covers exons 42 to 54, which is a region that is targeted by various exon skipping therapies. Here, the scale has been reduced to 0.3 (each pixel represents three base pairs), to condense the figure while still showing the reading frames and the exon sizes to scale. In addition, the gap between the exons has been increased for clarity.

If the start phase is the same as the end phase for an exon (in other words, if the length of the coding region is divisible by three), an exon can be skipped without disrupting the reading frame. This can be seen in Figure 4 for exons 47, 48 and 49. Note that although all in frame exons in Figure 4 are in phase 0, any exon with a coding length divisible by three is in frame.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.18.24313945; this version posted September 20, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

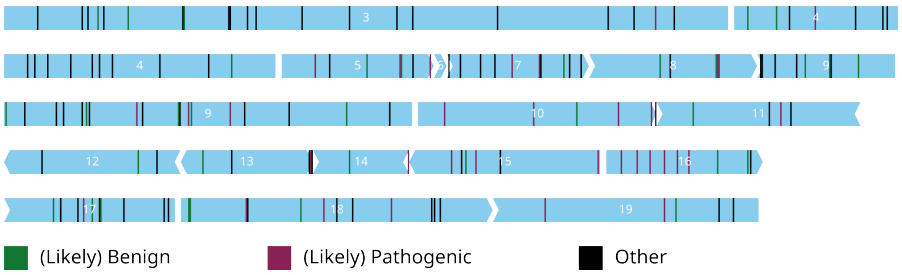

Figure 3: All ClinVar listed variants identified in the coding region of *CYLD*(NM_001378743.1). Colors were chosen to be distinguishable by color blind people. Likely benign and benign variants are depicted in dark green, likely pathogenic and pathogenic variants are depicted in purple and any other variant is shown in black.

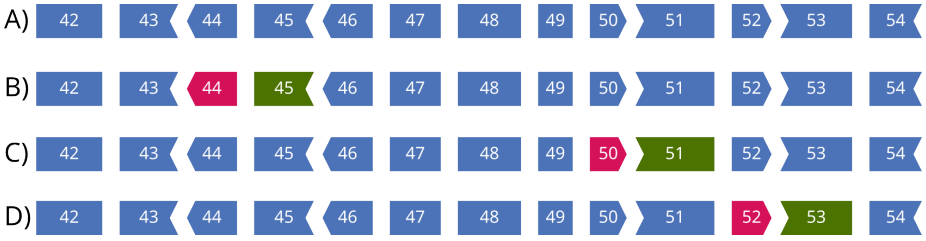

Figure 4: Partial coding region of *DMD*(ENST00000357033.9). Deleted exons are indicated in magenta, skipped exons are indicated in darkgreen. A) Wild-type transcript; B) Deletion of exon 44, which can be rescued by Casimersen skipping exon 45; C) Deletion of exon 50, which can be rescued by Eteplirsen skipping exons 51; D) Deletion of exon 52, which can be rescued by Viltolarsen skipping exon 53.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.18.24313945; this version posted September 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

Casimersen is an exon skipping therapy for Duchenne Muscular Dystrophy which skips exon 45 (Wagner et al., 2021) of the *DMD* transcript. As can be seen from Figure 4, this would cause a frame shift in a healthy version of the *DMD* gene. However, if a patient has a deletion of exon 44, which in itself is a frame shift mutation, skipping exon 45 restores the reading frame, since exon 43 and 46 fit together. Eteplirsen induces skipping of exon 51, restoring the reading frame for patients lacking exon 50 (Lim et al., 2017). Viltolarsen (Dhillon, 2020) and Golodirsen (Anwar & Yokota, 2020) skip exon 53, which can restore the reading frame for patients lacking exon 52. As can be seen from Figure 4, in theory these can also be used to treat deletions of 43-52, 45-52, 47-52, 48-52, 49-52 and 50-52 (Anwar & Yokota, 2020).

Conclusion

ExonViz is the first publicly accessible application that allows the user to draw transcripts, including additional features such as reading frames and the mutational landscape along the transcript. ExonViz can be used for illustrations within publications, assessment of treatment options, for teaching purposes or genetic counseling. Figures generated by ExonViz are free to use under the Creative Commons BY license. Furthermore, we allow the user to construct their own transcripts to incorporate features like poison or cryptic exons and alternative isoforms. ExonViz can be accessed as a web application via exonviz.rnatherapy.nl or installed via PyPI. The source code is available on Github.

Grants

RRvdB is supported by a ZonMW PSIDER grant. MCL is supported by a Walter Benjamin Fellowship from the DFG.

Acknowledgments

We would like to thank the members of the Dutch Center for RNA Therapeutics for their ideas and suggestions, and their feedback on earlier versions of ExonViz. We also thank Maximilian Haeussler and his colleagues at the UCSC for their efforts implementing exon frame information into the UCSC Genome Browser.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.18.24313945; this version posted September 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

References

Alasoo, K. (2017). Wiggleplotr: Make read coverage plots from bigwig files. *Bioconductor. doi*, *10*, B9.

Anwar, S., & Yokota, T. (2020). Golodirsen for duchenne muscular dystrophy. *Drugs of Today (Barcelona, Spain: 1998)*, *56*(8), 491–504.

Dhillon, S. (2020). Viltolarsen: First approval. *Drugs*, *80*(10), 1027–1031.

Gustavsson, E. K., Zhang, D., Reynolds, R. H., Garcia-Ruiz, S., & Ryten, M. (2022). Ggtranscript: An r package for the visualization and interpretation of transcript isoforms using ggplot2. *Bioinformatics*, *38*(15), 3844–3846.

Harrison, P. W., Amode, M. R., Austine-Orimoloye, O., Azov, A. G., Barba, M., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., Bhurji, S. K., Boddu, S., BrancoÂ Lins, P. R., Brooks, L., Ramaraju, S. B., Campbell, L. I., Martinez, M. C., Charkhchi, M., Chougule, K., … Yates, A. D. (2023). Ensembl 2024. *Nucleic Acids Research*, *52*(D1), D891–D899. https://doi.org/10.1093/nar/gkad1049

Lefter, M., Vis, J. K., Vermaat, M., den Dunnen, J. T., Taschner, P. E. M., & Laros, J. F. J. (2021). Mutalyzer 2: next generation HGVS nomenclature checker. *Bioinformatics*, *37*(18), 2811–2817. https://doi.org/10.1093/bioinformatics/btab051

Lim, K. R. Q., Maruyama, R., & Yokota, T. (2017). Eteplirsen in the treatment of duchenne muscular dystrophy. *Drug design, development and therapy*, 533–545.

Morales, J., Pujar, S., Loveland, J. E., Astashyn, A., Bennett, R., Berry, A., Cox, E., Davidson, C., Ermolaeva, O., Farrell, C. M., et al. (2022). A joint ncbi and embl-ebi transcript set for clinical genomics and research. *Nature*, *604*(7905), 310–315.

Mühlhausen, S., Hellkamp, M., & Kollmar, M. (2015). Genepainter v. 2.0 resolves the taxonomic distribution of intron positions. *Bioinformatics*, *31*(8), 1302–1304.

Nassar, L. R., Barber, G. P., Benet-Pagès, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J. N., Hinrichs, A. S., Lee, B. T., et al. (2023). The ucsc genome browser database: 2023 update. *Nucleic acids research*, *51*(D1), D1188–D1195.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (refseq) database at ncbi: Current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, *44*(D1), D733–D745.

Pozo, F., Rodriguez, J. M., Vázquez, J., & Tress, M. L. (2022). Clinical variant interpretation and biologically relevant reference transcripts. *NPJ Genomic Medicine*, *7*(1), 59.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: A joint con-

medRxiv preprint doi: https://doi.org/10.1101/2024.09.18.24313945; this version posted September 20, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY 4.0 International license .

sensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genetics in medicine*, *17*(5), 405–423.

Wagner, K. R., Kuntz, N. L., Koenig, E., East, L., Upadhyay, S., Han, B., & Shieh, P. B. (2021). Safety, tolerability, and pharmacokinetics of casimersen in patients with d uchenne muscular dystrophy amenable to exon 45 skipping: A randomized, double-blind, placebo-controlled, dose-titration trial. *Muscle & nerve*, *64*(3), 285–292.

Walker, L. C., de la Hoya, M., Wiggins, G. A., Lindy, A., Vincent, L. M., Parsons, M. T., Canson, D. M., Bis-Brewer, D., Cass, A., Tchourbanov, A., et al. (2023). Using the acmg/amp framework to capture evidence related to predicted and observed impact on splicing: Recommendations from the clingen svi splicing subgroup. *The American Journal of Human Genetics*, *110*(7), 1046–1067.