splicekit: a comprehensive toolkit for splicing analysis from short-read RNA-seq

Gregor Rot* ✉, Arne Wehling*, Roland Schmucki, Nikolaos Berntenis, Jitao David Zhang, Martin Ebeling

Roche Pharmaceutical Research and Early Development, Roche Innovation Center Basel, Basel, Switzerland

* Equal contribution
✉ Corresponding author. Roche Pharmaceutical Research and Early Development, Roche Innovation Center Basel, Basel, Switzerland. E-mail: gregor.rot@roche.com

Abstract

Motivation: Analysis of alternative splicing using short-read RNA-seq data is a complex process that involves several steps: alignment of reads to the reference genome, identification of alternatively spliced features, motif discovery, analysis of RNA-protein binding near donor and acceptor splice sites, and exploratory data visualization.
Results: We introduce *splicekit*, a python package that provides a comprehensive set of tools for conducting splicing analysis.
Availability and implementation: https://github.com/bedapub/splicekit and over PyPI.

1 Introduction

Alternative splicing of RNA is a fundamental biological process that is critical for the generation of protein diversity. Dysregulation of splicing has been implicated in many human diseases such as cancer and neurological disorders (Scotti et al., 2016). Recent advances in splicing modulation using compounds, i.e. small molecules (Schneider-Poetsch et al., 2021), such as Risdiplam for the treatment of spinal muscular atrophy (Ratni et al., 2018), have renewed interest in developing new therapies targeting splicing.

Splicing analysis using short-read RNA-seq data is a multifaceted process that involves several steps and requires the integration of diverse software tools. For the analysis of differential splicing events, the community can potentially benefit from a comprehensive and efficient analysis toolbox.

To address this need, we introduce *splicekit*, a Python package that provides and integrates a set of existing and novel splicing analysis tools (Figure 1A). It offers functionalities to identify differentially expressed features (junctions, exons and genes), cluster samples, perform motif analysis to elucidate potential regulatory patterns, visualize changes of junction versus those of genes, and to identify RNA-protein binding in the vicinity of regulated features.

2 Splicing Analysis

The first step in *splicekit* is to identify regulated features in the comparison, for which *splicekit* runs edgeR (Robinson et al., 2009; Chen et al., 2016) with the diffSpliceDGE function to estimate differential splicing (on junction and exon counts within their respective gene context) and the edgeR glmQLFTest function to estimate differential gene expression.

splicekit then integrates diverse existing and novel analysis tools and methods to provide comprehensive differential splicing analysis. It introduces junction-DGE (juDGE) plots, a novel visualization technique to gauge the level of change in the splicing vs. gene context. It implements Donor Junction Analysis (DonJuAn) and motif analysis with DREME (Bailey et al., 2011) to elucidate potential regulatory patterns. In addition, it performs RNA-protein binding scanning (scanRBP) to identify RNA-protein binding in the vicinity of regulated donor and acceptor splice sites.

To visualize read coverage and alignments, *splicekit* provides an integrated JBrowse2 (Diesh et al., 2022) with a containerized web server. Exploring the data in JBrowse2 includes opening the web browser with the local JBrowse2 instance (Figure 1B).

We introduce and further describe novel analysis tools integrated with *splicekit* in sections 2.1 - 2.3.

2.1 junction-DGE (juDGE) and cluster logFC plots

To estimate if treated samples display mostly splicing or gene expression changes compared to control samples, *splicekit* produces juDGE plots (Figure 1C). By including genes and junctions and plotting gene log2 fold change (logFC) vs. logFC of junctions, we can estimate the level of alternative splicing in contrast to differential gene expression. A tall vertical plot with a low plot score, defined by the ratio of standard variance of x- and y-axis values, suggests detected changes are mostly on the splicing level, while a wider horizontal plot (high score) suggests there is extensive differential gene expression involved.

In case of multiple comparisons, *splicekit* also provides logFC clustering analysis at the level of junctions, exons and genes. In the Curtiss et al. (2022) dataset, the samples cluster by experimental group (time/cell type) rather than by treatment at all levels (Figure 1D). Such cluster analysis allows an additional overview of the diverse treatments in the context of splicing and gene expression.

2.2 Scanning for RNA-protein binding (scanRBP)

To investigate potential involvement of RNA-binding proteins (RBPs) in the mode of action of detected differential splicing events, we established an RBP analysis tool named scanRBP. It plots CLIP data cumulatively (van Nostrand et al., 2016; König et al., 2010) around a set of regulated features or use PWMs (112 RBPs from mCrossAtlas, Feng et al., 2019) to plot log-odds signals for diverse proteins.

The results are RNA-maps that help suggest potential roles of RBPs in splicing changes due to treatment (Rot et al., 2017). Using data reported by Brown et al., 2022, we applied scanRBP to show how TDP-43 represses donor splice site usage (Figure 1E).

2.3 Donor Junction Analysis (DonJuAn)

Small molecular splicing modifiers are an arising therapeutic modality that targets specific junction donor sites to modify exon inclusion rates (Sivaramakrishnan et al., 2017). To identify sequence specific splicing effects, we implemented the Donor Junction Analysis (DonJuAn) module in *splicekit*. DonJuAn identifies the donor site sequences and exonic anchor regions of detected junctions for differential expression analysis. Exon inclusion produces positive junction/anchor logFC values, while exon skipping events give negative values which allows for filtering (Figure 1F, left).

As a demonstration, we analyzed public data (Ishigami et al., 2022) on Branaplam, an experimental drug to treat spinal muscular atrophy that binds to donor splice sites, surrounded by the sequence GAGTAAGT (Palacino et al., 2015). DonJuAn logFC junction stratification (Figure 1F, scatterplots) increases the signal for motif enrichment software such as DREME (Figure 1F, DREME).

3 Conclusion

We introduce *splicekit*, a comprehensive toolset for analyzing short-read RNA-seq datasets in the context of alternative splicing regulation. By integrating diverse analysis tools and methods, including external tools such as edgeR and DREME, as well as novel tools such as juDGE, DonJuAn, and scanRBP, *splicekit* provides a multifaceted approach to splicing analysis.

One of the key strengths of *splicekit* is its ability to interconnect basic feature analysis with motif search and RNA-protein binding analysis, allowing for a more in-depth understanding of splicing regulation. Additionally, *splicekit* provides exploratory tools for studying the mode of action of splicing in the context of different treatments and compounds.

splicekit can be run on a single computer or on a computer cluster, making it a versatile tool for researchers with varying computational resources. Overall, we believe that the scientific community may benefit from adopting, using, and further developing *splicekit*.

References

Bailey, T. L. (2011). DREME: Motif discovery in transcription factor ChIP-seq data. Bioinformatics, 27(12), 1653–1659. https://doi.org/10.1093/bioinformatics/btr261

Brown, A. L., Wilkins, O. G., Keuss, M. J., Hill, S. E., Zanovello, M., Lee, W. C., Bampton, A., Lee, F. C. Y., Masino, L., Qi, Y. A., Bryce-Smith, S., Gatt, A., Hallegger, M., Fagegaltier, D., Phatnani, H., Newcombe, J., Gustavsson, E. K., Seddighi, S., Reyes, J. F., Coon, S. L., Ramos, D., Schiavo, G., Fisher, E. M. C., Raj, T., Secrier, M., Lashley, T., Ule, J., Buratti, E., Humphrey, J., Ward, M. E., Fratta, P. (2022). TDP-43 loss and ALS-risk SNPs drive mis-splicing and depletion of UNC13A. Nature, 603(7899), 131–137. https://doi.org/10.1038/s41586-022-04436-3

Chen, Y., Lun, A. T. L., & Smyth, G. K. (2016). From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research*, *5*, 1438. https://doi.org/10.12688/f1000research.8987.1

Curtiss, B. M., VanCampen, J., Macaraeg, J., Kong, G. L., Taherinasab, A., Tsuchiya, M., Yashar, W. M., Tsang, Y. H., Horton, W., Coleman, D. J., Estabrook, J., Lusardi, T. A., Mills, G. B., Druker, B. J., Maxson, J. E., & Braun, T. P. (2022). PU.1 and MYC transcriptional network defines synergistic drug responses to KIT and LSD1 inhibition in acute myeloid leukemia. Leukemia, 36(7), 1781–1793. https://doi.org/10.1038/s41375-022-01594-1

Diesh, C., Stevens, G. J., Xie, P., de Jesus Martinez, T., Hershberg, E. A., Leung, A., Guo, E., Dider, S., Zhang, J., Bridge, C., Hogue, G., Duncan, A., Morgan, M., Flores, T., Bimber, B. N., Haw, R., Cain, S., Buels, R. M., Stein, L. D., & Holmes, I. H. (n.d.). JBrowse 2: A modular genome browser with views of synteny and structural variation. https://doi.org/10.1101/2022.07.28.501447

Feng, H., Bao, S., Rahman, M. A., Weyn-Vanhentenryck, S. M., Khan, A., Wong, J., Shah, A., Flynn, E. D., Krainer, A. R., & Zhang, C. (2019). Modeling RNA-Binding Protein Specificity In Vivo by Precisely Registering Protein-RNA Crosslink Sites. Molecular Cell, 74(6), 1189-1204.e6. https://doi.org/10.1016/j.molcel.2019.02.002

Ishigami, Y., Wong, M. S., Martí-Gómez, C., Ayaz, A., Kooshkbaghi, M., Hanson, S., McCandlish, D. M., Krainer, A. R., & Kinney, J. B. (n.d.). Title: Specificity, synergy, and mechanisms of splice-modifying drugs. https://doi.org/10.1101/2022.12.30.522303

König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M., & Ule, J. (2010). ICLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nature Structural and Molecular Biology, 17(7), 909–915. https://doi.org/10.1038/nsmb.1838

van Nostrand, E. L., Pratt, G. A., Shishkin, A. A., Gelboin-Burkhart, C., Fang, M. Y., Sundararaman, B., Blue, S. M., Nguyen, T. B., Surka, C., Elkins, K., Stanton, R., Rigo, F., Guttman, M., & Yeo, G. W. (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nature Methods, 13(6), 508–514. https://doi.org/10.1038/nmeth.3810

Palacino, J., Swalley, S. E., Song, C., Cheung, A. K., Shu, L., Zhang, X., van Hoosear, M., Shin, Y., Chin, D. N., Keller, C. G., Beibel, M., Renaud, N. A., Smith, T. M., Salcius, M., Shi, X., Hild, M., Servais, R., Jain, M., Deng, L., … Sivasankaran, R. (2015). SMN2 splice modulators enhance U1-pre-mRNA association and rescue SMA mice. Nature Chemical Biology, 11(7), 511–517. https://doi.org/10.1038/nchembio.1837

Ratni, H., Ebeling, M., Baird, J., Bendels, S., Bylund, J., Chen, K. S., Denk, N., Feng, Z., Green, L., Guerard, M., Jablonski, P., Jacobsen, B., Khwaja, O., Kletzl, H., Ko, C. P., Kustermann, S., Marquet, A., Metzger, F., Mueller, B., … Mueller, L. (2018). Discovery of Risdiplam, a Selective Survival of Motor Neuron-2 (SMN2) Gene Splicing Modifier for the Treatment of Spinal Muscular Atrophy (SMA). Journal of Medicinal Chemistry, 61(15), 6501–6517. https://doi.org/10.1021/acs.jmedchem.8b00741

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Rot, G., Wang, Z., Huppertz, I., Modic, M., Lenče, T., Hallegger, M., Haberman, N., Curk, T., von Mering, C., & Ule, J. (2017). High-Resolution RNA Maps Suggest Common Principles of Splicing and Polyadenylation Regulation by TDP-43. Cell Reports, 19(5), 1056–1067. https://doi.org/10.1016/j.celrep.2017.04.028

Schneider-Poetsch, T., Chhipi-Shrestha, J. K., & Yoshida, M. (2021). Splicing modulators: on the way from nature to clinic. In Journal of Antibiotics (Vol. 74, Issue 10, pp. 603–616). Springer Nature. https://doi.org/10.1038/s41429-021-00450-1

Scotti, M. M., & Swanson, M. S. (2016). RNA mis-splicing in disease. In Nature Reviews Genetics (Vol. 17, Issue 1, pp. 19–32). Nature Publishing Group. https://doi.org/10.1038/nrg.2015.3

Sivaramakrishnan, M., McCarthy, K. D., Campagne, S., Huber, S., Meier, S., Augustin, A., Heckel, T., Meistermann, H., Hug, M. N., Birrer, P., Moursy, A., Khawaja, S., Schmucki, R., Berntenis, N., Giroud, N., Golling, S., Tzouros, M., Banfai, B., Duran-Pacheco, G., … Metzger, F. (2017). Binding to SMN2 pre-mRNA-protein complex elicits specificity for small molecule splicing modifiers. Nature Communications, 8(1). https://doi.org/10.1038/s41467-017-01559-4

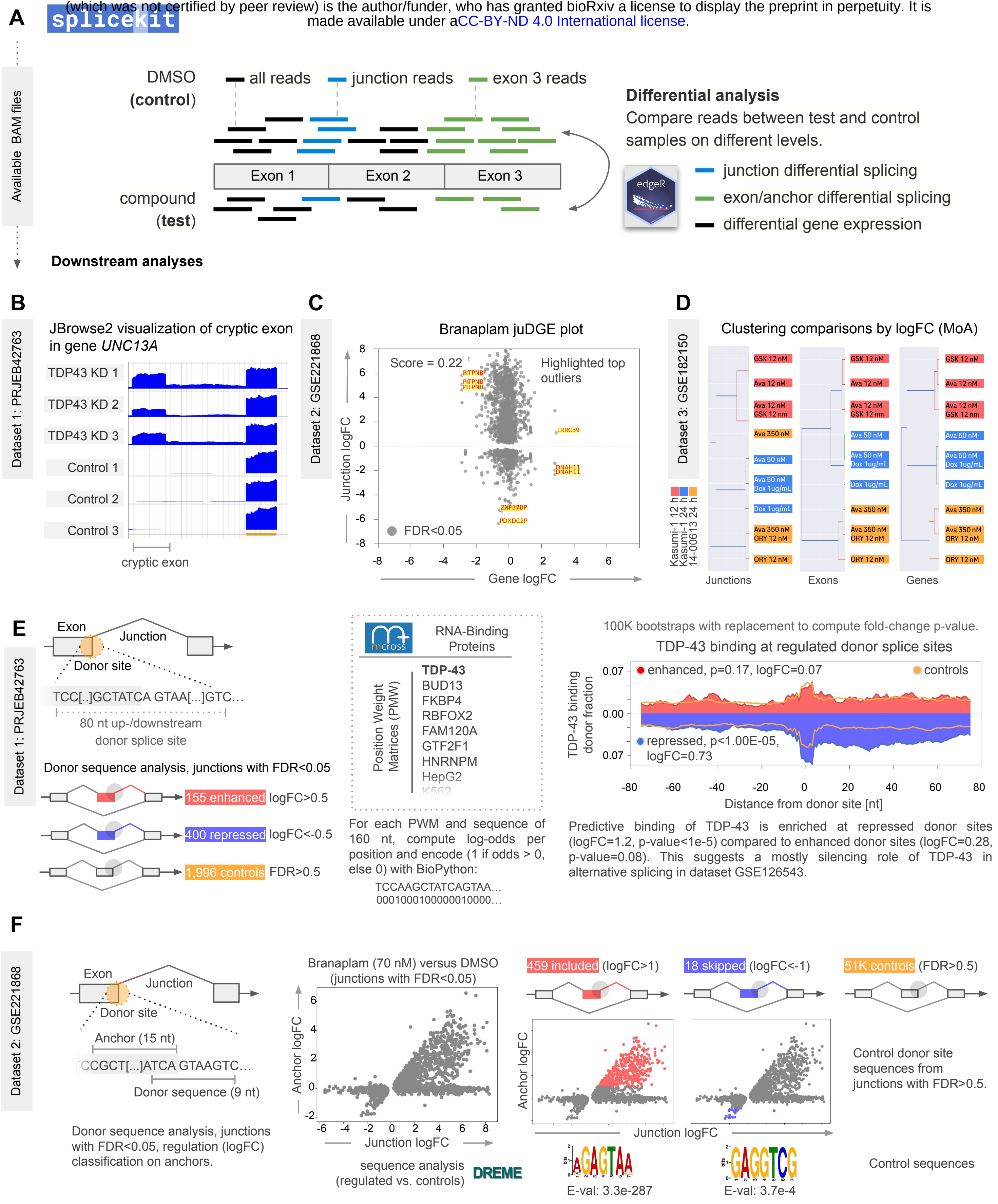

Figure 1. splicekit comprehensive toolkit for splicing analysis from short-read RNA-seq

(A) Initial input to splicekit is read alignments in BAM format. Comparisons are made between groups of test and control experiments. After initial differential calling on the level of splicing (junctions, exons) and genes, further downstream analysis include JBrowse2 visualizations, juDGE plots (logFC of genes and junctions), motif and RNA-protein binding enrichment analysis.

(B) JBrowse2 visualization of PRJEB42763 samples and *UNC13A* cryptic exon. The cryptic exon is reported by splicekit junction analysis.

(C) junction logFC vs. gene expression logFC (juDGE) plot in GSE221868 suggest Branaplam is a splicing modifying compound. A low juDGE score (stdev_x/stdev_y) suggest most regulation happens at the splicing (junction) level.

(D) Clustering of comparisons (test conditions vs. controls) at the junction, exon and gene levels. Only features with FDR<0.05 in at least one comparison.

(E) scanRBP analysis of TDP-43 RNA-protein binding in GSE126543 suggests enrichment of TDP-43 binding at repressed sites.

(F) Donor Junction Analysis (Don JuAn) on Branaplam versus DMSO in GSE221868 identifies relevant donor sites to detect known binding motif AGAGTAA (Palacino et al., 2015).