

Refinebioc: A Bioconductor Package for Seamless Integration of refine.bio Transcriptome Data

Alexander Ho¹, BS; Sean Davis², MD, PhD

¹University of Colorado Anschutz School of Medicine, Research Track

² University of Colorado Anschutz School of Medicine, Department of Biomedical Informatics

Background

- Billions of dollars of data spanning millions of genomic samples from technologies like microarrays and sequencing are accessible through repositories such as Gene Expression Omnibus (GEO), ArrayExpress, & Sequence Read Archive (SRA).
- Refine.bio is one such repository that offers uniformly processed and normalized transcriptome data.
- There is a lack of tools to easily access these data from refine.bio.
- Therefore, we introduce the refinebioc software package that bridges refine.bio with Bioconductor, a large open-source community that develops R software to promote repeatability in analysis of biological data

Methods

- The refinebioc package offers the following functionality:
 - Data Overview & Retrieval
 - Download, Extract, & Load
 - Share and Access Webtools
- Data can be browsed and accessed directly in R, extracted as a SummarizedExperiment object, ensuring compatibility with many existing Bioconductor tools.
- Refinebioc allows for users to streamline their research workflows.

Results

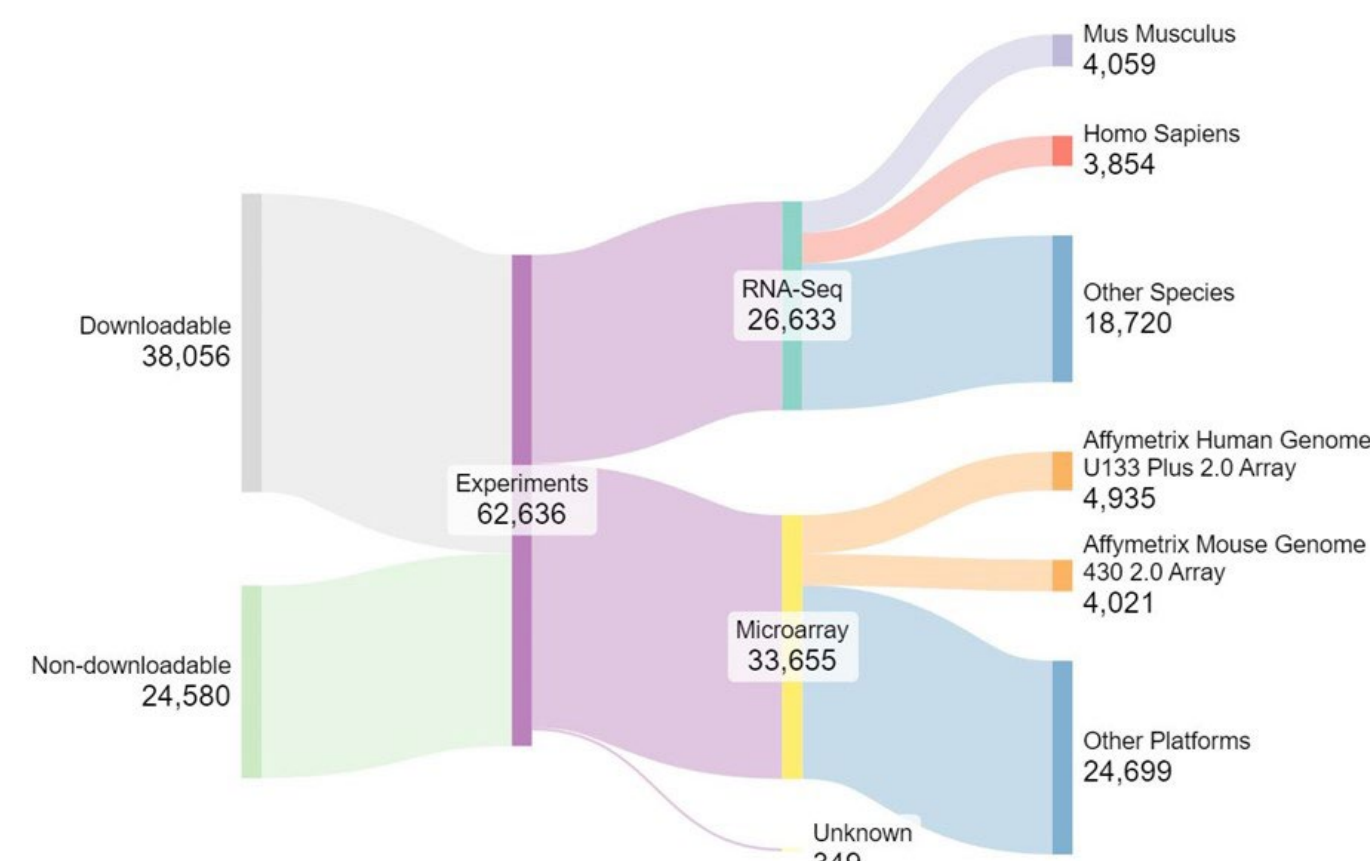


Figure 1. Flow diagram representing the number of experiments available within the refine.bio repository.

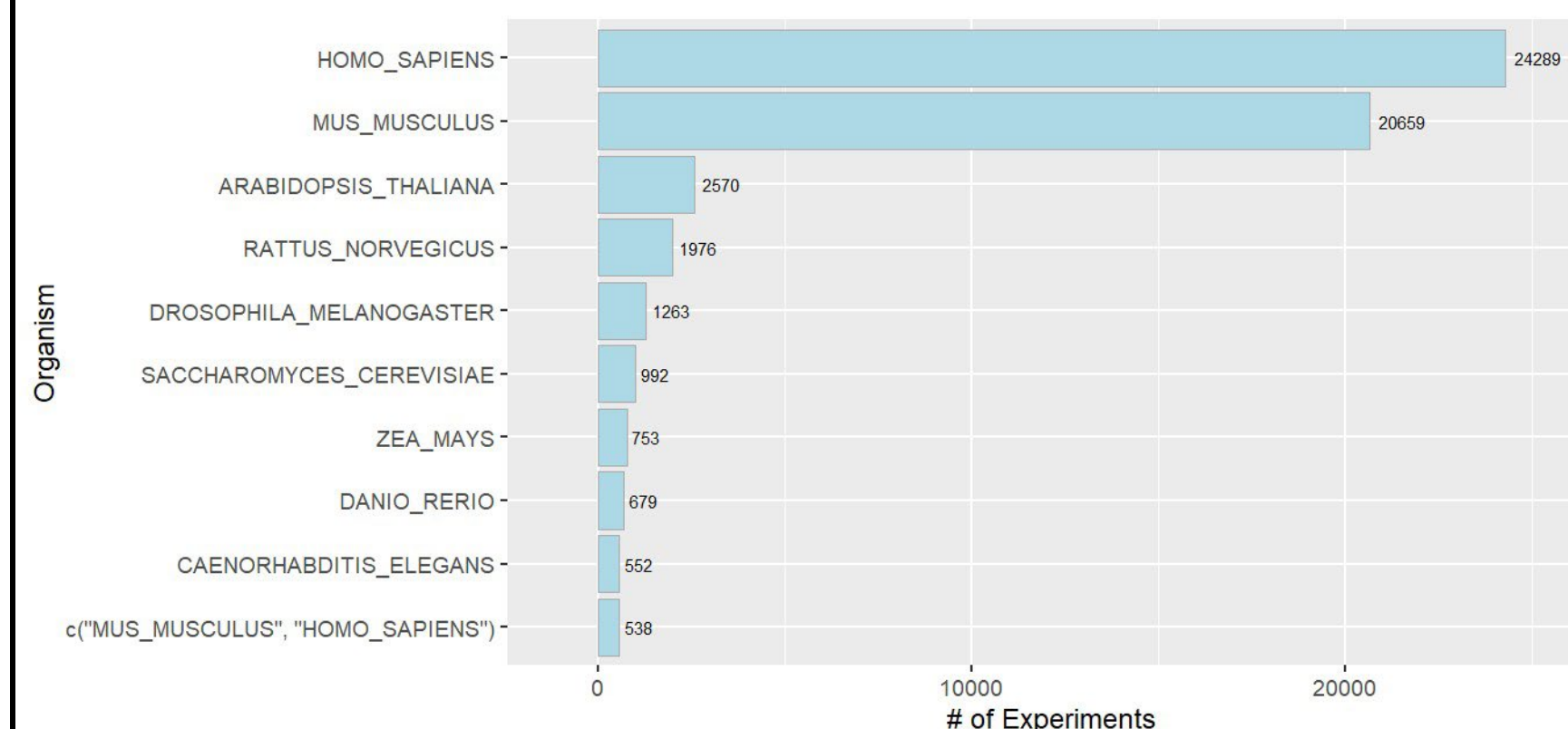


Figure 2. Top 10 Organisms within refine.bio.

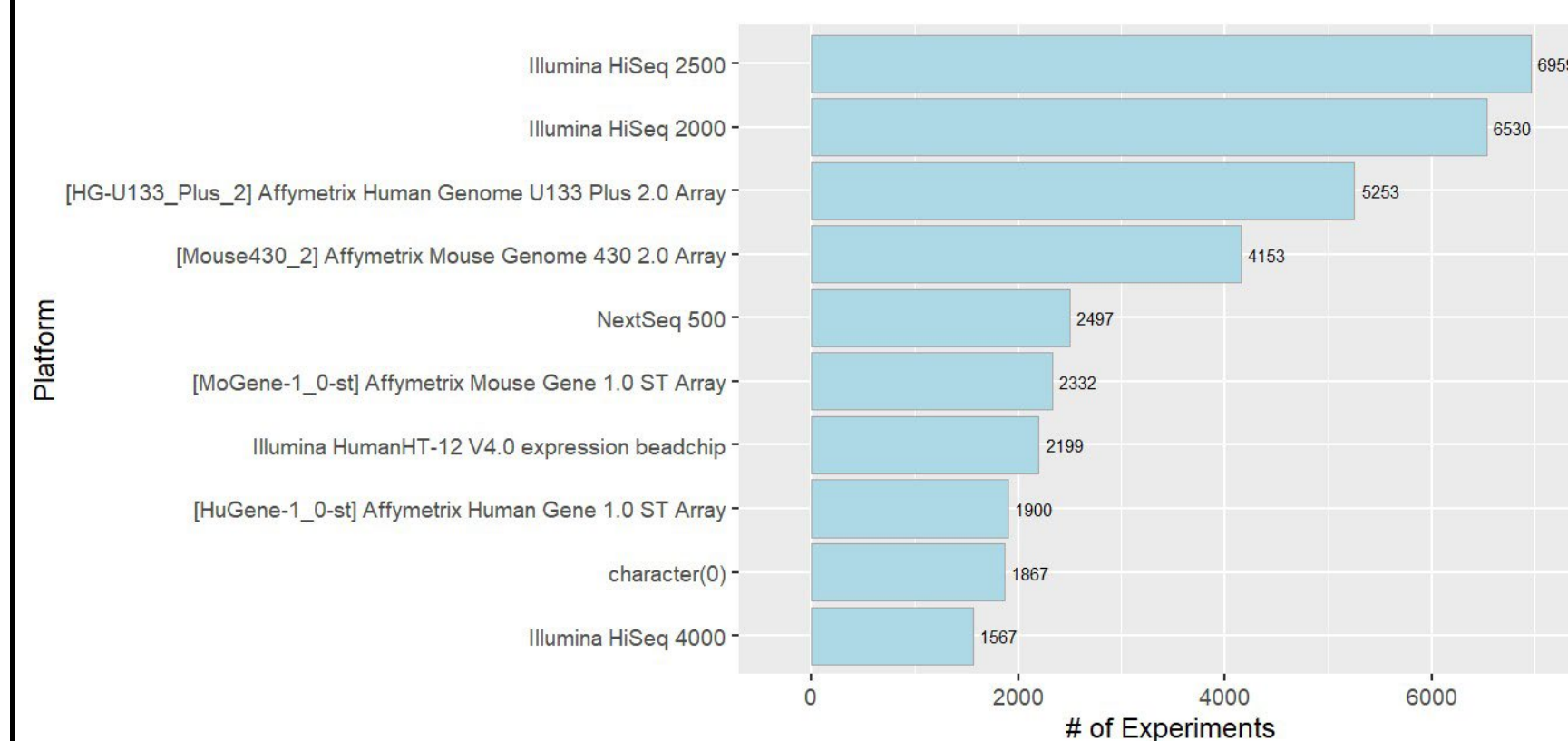


Figure 3. Top 10 Platforms within refine.bio.

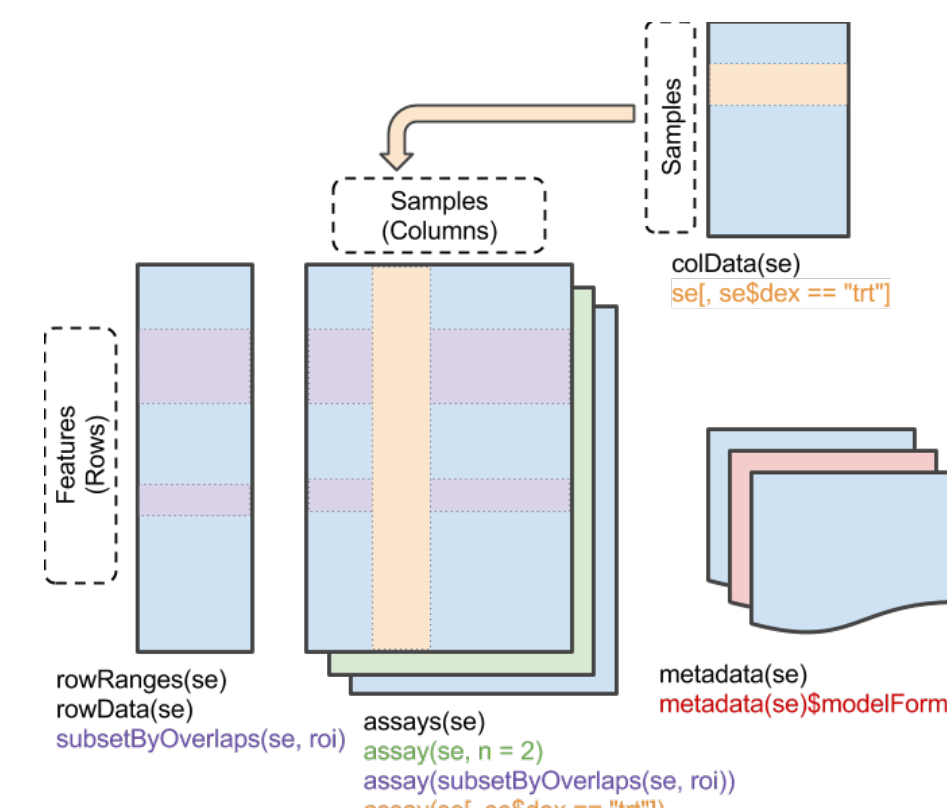


Figure 4. Depiction of the SummarizedExperiment object, compatible with Bioconductor.

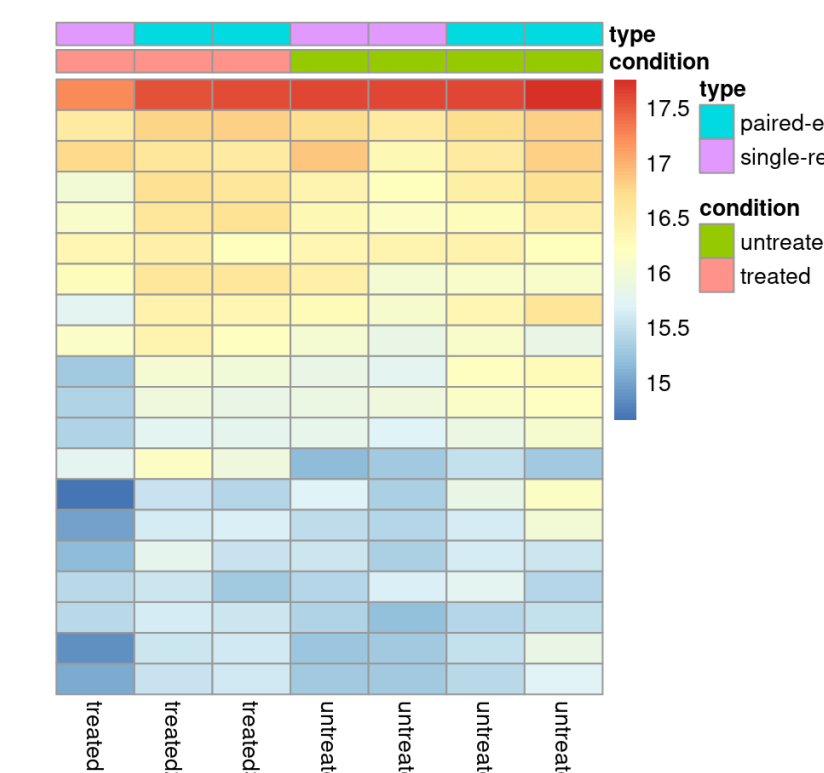


Figure 5. Heatmap of a count matrix utilizing the DESeq2 package from Bioconductor

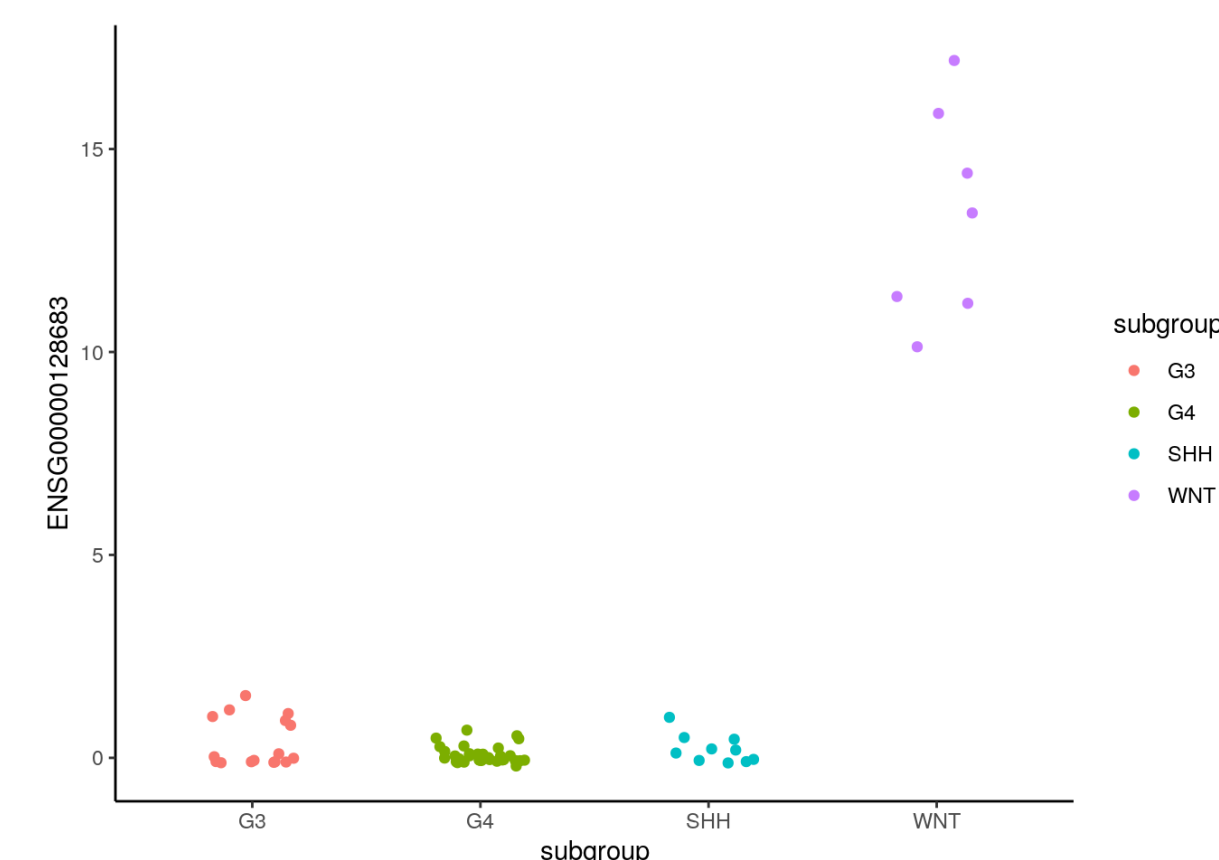


Figure 6. Differential Expression graph for a gene, utilizing the limma package from Bioconductor.

Conclusions

- Refinebioc connects refine.bio with the R/Bioconductor ecosystem, allowing seamless access to high-quality transcriptome data.
- Data can be analyzed in-depth by utilizing existing Bioconductor tools
- Research methods can be accelerated through utilizing the refinebioc package.
- As the volume and complexity of transcriptome data continue to grow, refinebioc can play an important role in facilitating data-driven discoveries and advancing understanding of biological systems.

Disclosures

- The authors do not have any conflicts of interest to disclose.

References

- Clough, Emily, and Tanya Barrett. 2016. "The Gene Expression Omnibus Database." Edited by Ewy Mathé and Sean Davis 1418: 93–110. https://doi.org/10.1007/978-1-4939-3578-9_5.
- Davis, Sean, and Paul S. Meltzer. 2007. "GEOQuery: A Bridge Between the Gene Expression Omnibus (GEO) and BioConductor." *Bioinformatics* 23 (14): 1846–47. <https://doi.org/10.1093/bioinformatics/btm254>.
- Huber, Wolfgang, Vincent J. Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S. Carvalho, Hector Corrada Bravo, et al. 2015. "Orchestrating High-Throughput Genomic Analysis with Bioconductor." *Nature Methods* 12 (2): 115–21. <https://doi.org/10.1038/nmeth.3252>.
- Kauffmann, Audrey, Tim F. Rayner, Helen Parkinson, Misha Kapushesky, Margus Lukk, Alvis Brazma, and Wolfgang Huber. 2009. "Importing ArrayExpress Datasets into R/Bioconductor." *Bioinformatics* 25 (16): 2092–94. <https://doi.org/10.1093/bioinformatics/btp354>.
- Leinonen, R., H. Sugawara, M. Shumway, and on behalf of the International Nucleotide Sequence Database Collaboration. 2011. "The Sequence Read Archive." *Nucleic Acids Research* 39 (Database): D19–21. <https://doi.org/10.1093/nar/gkq1019>.
- Parkinson, H., M. Kapushesky, M. Shojatalab, N. Abeygunawardena, R. Coulson, A. Farne, E. Holloway, et al. 2007. "ArrayExpress—a Public Database of Microarray Experiments and Gene Expression Profiles." *Nucleic Acids Research* 35 (Database): D747–50. <https://doi.org/10.1093/nar/gkl995>.
- "SummarizedExperiment." 2017. Bioconductor. <https://doi.org/10.18129/B9.BIOC.SUMMARIZEDEXPERIMENT>.
- Zhu, Yuelin, Robert M. Stephens, Paul S. Meltzer, and Sean R. Davis. 2013. "SRadb: Query and Use Public Next-Generation Sequencing Data from Within R." *BMC Bioinformatics* 14 (1): 19. <https://doi.org/10.1186/1471-2105-14-19>.