

The problem of selection bias in studies of pre-mRNA splicing

Zachary W. Dwyer & Jeffrey A. Pleiss

 Check for updates

Here we demonstrate how selection bias in studies of pre-mRNA splicing can inadvertently lead to inaccurate biological conclusions which nevertheless appear statistically robust. We argue that this may be a pervasive problem with potentially significant consequences for the field.

The problem

Selection bias, also sometimes referred to as sample bias or sample selection bias, generally refers to a distortion (bias) of statistical testing that results from the way that samples are collected (selection). While this problem has been well understood in clinical and social sciences for quite some time^{1–3}, its significance in molecular biology is less widely appreciated. Work from Oshlack's group^{4,5} provides perhaps the most compelling demonstration of the issue of selection bias in molecular biology, wherein they demonstrate its impact on RNA-seq experiments where analyses of differential gene expression are coupled to analyses of GO-term enrichment. Oshlack's group highlights a major pitfall in studies like this which is that not all transcripts are measured with the same statistical power in a typical RNA-seq experiment: because longer and more highly expressed transcripts are sampled more frequently than are shorter and lower expressed transcripts, there is more statistical power to identify long and/or highly expressed transcripts as being differentially expressed. The result of this is a distortion in the statistics that measure enrichment: the set of transcripts identified as differentially expressed is dependent not only on their biological behavior but on distinct properties that enhance their capacity for detection in the experiment. Importantly, because splicing-informative reads are rare within standard RNA-seq datasets⁶, the problem of selection bias can be particularly problematic in studies involving pre-mRNA splicing. We therefore aim to raise awareness of the problem of selection bias in analyses of next-generation sequencing studies designed to understand quantitative changes in pre-mRNA splicing.

Causes and consequences of selection bias

To demonstrate both the problem of selection bias and the deleterious consequences of this bias in studies of pre-mRNA splicing, we designed a simple experiment that examines changes in splicing in the background of a well-characterized genetic variant of a canonical spliceosomal component: the RNA helicase Prp2. Work from several groups has established a role for Prp2 in rearranging the spliceosome prior to the first catalytic step^{7–10}, and as such a reasonable expectation is that loss of Prp2 function would result in defective splicing for all (or nearly all) expressed transcripts. Using a targeted sequencing approach termed Multiplexed Primer Extension Sequencing, or MPE-seq^{6,11},

which massively enriches for splicing informative reads, we generated rich datasets, equivalent to ~ an entire lane of NextSeq550 sequencing for each of triplicate samples from a budding yeast strain harboring the conditional *prp2-1* allele and a matched wild-type strain. To demonstrate the effect of sequencing depth on experimental outcome, we then computationally downsampled this large experiment to generate three smaller subsets of data, equivalent in an RNA-seq setting to what could be considered high, medium, and low sized experiments, or ~ 80, 40, and 20 million reads per replicate, respectively.

To analyze these datasets, for each intron-containing gene in the genome we calculated the fold change in abundance of reads corresponding to both premature and mature isoforms and then assessed these for differential expression using DESeq2 (Ref. 12). Of the 272 splicing events profiled in the full dataset, 261 demonstrated statistically significant differential splicing in the mutant relative to wildtype (Fig. 1A). For the majority of these (211), both the premature and mature versions of the transcript were detected as differentially expressed, whereas a smaller number of transcripts displayed differential expression of only one of the two isoforms; presumably reflecting different intrinsic properties of the rates of synthesis or degradation of these transcripts¹³. Importantly, the absence of evidence for differential expression of either splicing isoform for the remaining 11 transcripts in this experiment cannot be interpreted as evidence of an absence of an impact of the *prp2-1* variant on these transcripts. While such a biological conclusion might be true, it could also be that the design of this experiment was flawed—from a biological standpoint—perhaps because these transcripts were not actively expressed under the chosen conditions. Equally plausible, however, is the possibility that the experiment was technically flawed because even at this high sequencing depth it lacked sufficient statistical power to detect real changes in the splicing efficiency of these transcripts.

The consequences of decreased statistical power become readily apparent when considering our analyses of the downsampled datasets (Fig. 1B) wherein ever decreasing numbers of events were detected as differentially expressed with statistical significance as the read depth decreased. Whereas analysis of the full dataset demonstrated with statistical significance the widespread impact of *prp2-1* on the genome-wide splicing outcome, in standard sized experiments many of these splicing events lack the statistical power to be 'selected' within the class of transcripts considered impacted by *prp2-1*: the selection bias problem. Importantly, as Oshlack's group previously demonstrated for standard gene expression studies, loss of statistical power does not occur evenly across the complement of genome-wide events being monitored, but rather occurs as a function of intrinsic properties of those targets which may or may not be important in the context of the biological problem being examined. For example, Fig. 1C shows a comparison of the lengths of the introns that were identified as impacted or not at each of the different experimental depths. Whereas

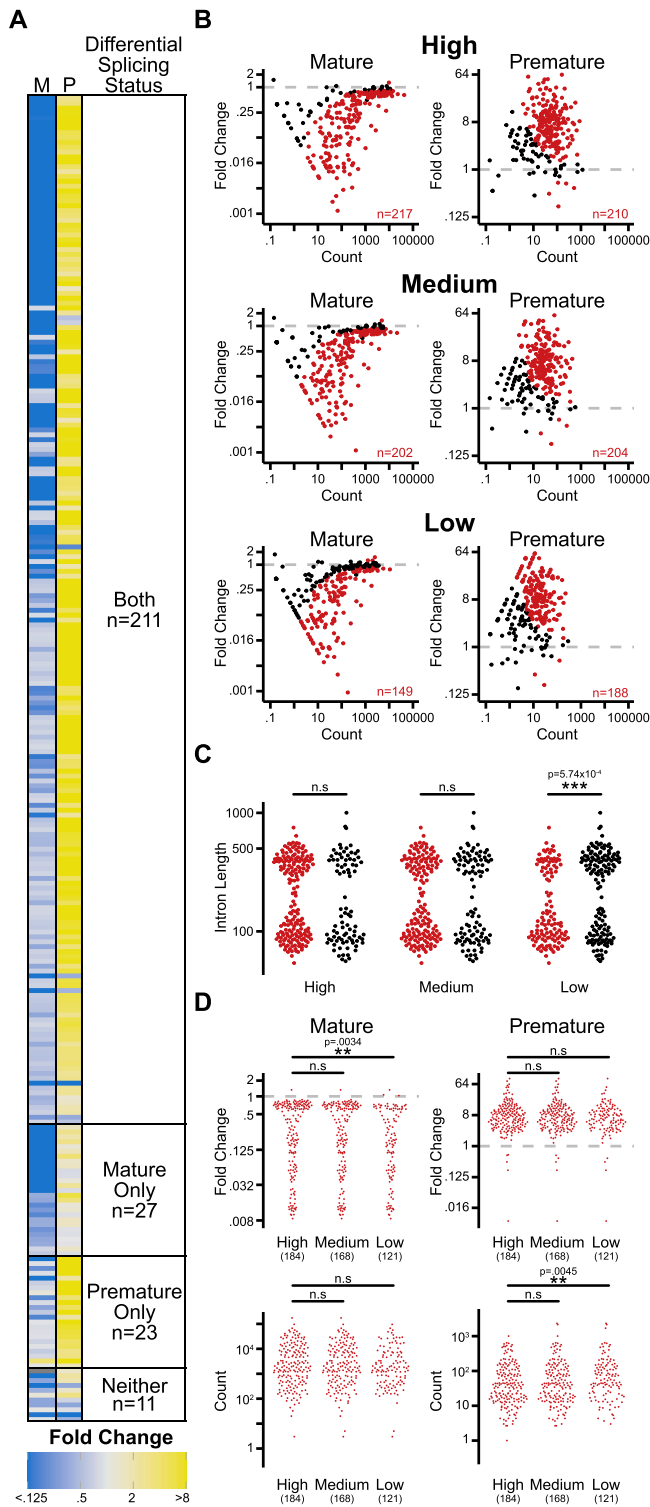


Fig. 1 | Selection bias introduces false correlations at insufficient read depth. Comparison of genome-wide splicing status in a *prp2-1* harboring strain relative to a matched wild-type strain after a ten min. shift to the non-permissive temperature (37 °C). **A** Heat map of fold change broken into categories of introns where both mature and premature, only mature, only premature, and neither mature nor premature supporting reads are significantly different between the Prp2 mutant and wild type as measured by DESeq2 at a multiple hypothesis corrected value of 0.05. **B** Fold change in number of mature or premature supporting reads as a function of expression between Prp2 mutant and wild type after downsampling to library sizes of 800,000 (High) MPE-seq reads to 400,000 (Medium) and 200,000 (Low) reads as measured by DESeq2. Red points are statistically significantly different at a multiple hypothesis corrected value of 0.05. **C** Length of introns that have (red) or do not have (black) significant difference in both premature and mature supporting reads as a function of read depth. One-sided Mann-Whitney test was performed to determine significance. **D** Fold change (upper) and count total (lower) for mature and premature supporting reads of introns that are significantly different in both mature and premature supporting counts at each downsampling. Two-sided Mann-Whitney test was performed to determine significance.

sensitized to loss of Prp2 function, the underlying data are more consistent with this being a biologically meaningless result of the loss of statistical power. As with most approaches for statistical testing, DESeq2 considers two important properties of the data in determining significance: the effect size, or difference in expression between the experimental and control samples; and the variance associated with the underlying measurements. For the relatively highly sampled mature isoforms (Fig. 1d, left), the subset of introns identified as differentially expressed are not characterized by higher read counts, but rather by larger fold-changes. Small fold-changes in the expression of the mature mRNA only surpass the significance threshold in the highly sampled dataset where overall variance is decreased. By contrast, for the relatively rare premature isoforms, the subset identified as differentially expressed is biased towards those that are highly sampled: even large fold-changes in differential expression fail to be deemed statistically significant if read depth is low (where variance is naturally higher).

Practical implications of selection bias

The above data demonstrate how selection bias has the potential to impact splicing studies, and we argue that this problem is likely pervasive in the field. To illustrate this, we consider here the data from one study which examined the role of the splicing factor HTATSF1, the ortholog of yeast Cus2¹⁴. As a core component of the U2 snRNP, and building off of significant prior work demonstrating a role for Cus2 in stabilizing a core structure of the U2 snRNA¹⁵⁻¹⁷, a reasonable expectation is that loss of HTATSF1/Cus2 activity would lead to decreased splicing efficiency across the complement of genome-wide substrates, akin to our expectations and observations for loss of Prp2 function as presented above (Fig. 1). In contrast, based on a knock-down experiment in mouse embryonic stem cells, it was reported that HTATSF1 appeared to function as a regulator of intron retention specifically in ribosomal proteins. While compelling statistical support was provided for intron retention within 45 different transcripts (many of which are involved in ribosome biogenesis and assembly), we wondered whether these transcripts were indeed uniquely impacted by loss of HTATSF1, or whether these were among the subset of transcripts for which there was sufficient statistical power in the experiment to detect a change in splicing efficiency. We therefore asked if the underlying data

no length difference was apparent between these classes at the High and Middle sizes, in the Low dataset a strong and statistically significant difference in the intron lengths was observed between the classes. While this result might suggest that short introns are more

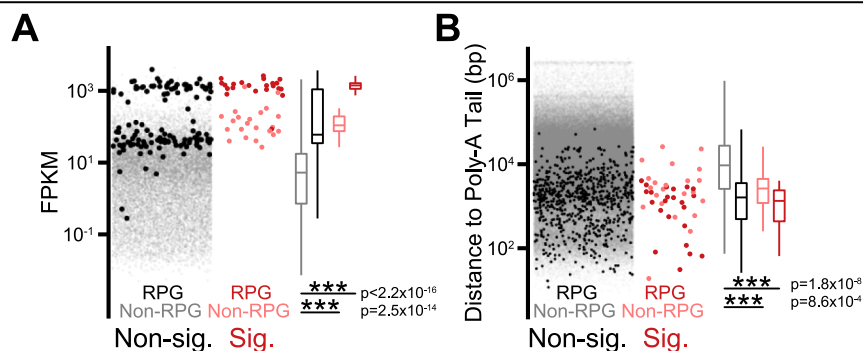


Fig. 2 | Selection bias in identified intron retention events. A Expression (measured in Fragments per Kilobase per Million Reads) of all introns versus those identified as retained broken out by ribosomal protein genes (RPG) and non-ribosomal protein genes (Non-RPG). **B** Distance from intron to poly-A Tail for all

introns versus those identified as retained, broken out by RPG and Non-RPG. Two-sided Mann-Whitney test was performed to determine significance. For all box-plots, the median value is represented by the center line, box limits represent the 25th and 75th quartiles, and whiskers are 1.5x the interquartile range.

suggested a bias in the subset of identified events by examining two parameters which are expected to influence pre-mRNA detection: expression level of the host transcript, and distance between the end of the affected intron and the polyadenylation site for that transcript. Our motivation for examining this second feature is that most RNA-seq protocols, including the one employed in Corsini et al.¹⁴, utilize a poly(A)⁺ enrichment step. Because splicing is coupled to transcription¹⁸, the likelihood of a retained intron being detected within a poly(A)⁺ pool of RNA is expected to be highest for those introns located closest to the polyadenylation site, as these would have the least amount of time for removal prior to polyadenylation. As shown in Fig. 2, the events identified in Corsini et al.¹⁴ appear biased for each of these properties such that they would be expected to have much greater statistical power for detecting differential expression than most of the other events in the genome. As such, while we do not question the role of coordinated control of ribosome biogenesis, our analysis suggests that the exact extent to which HTATSF1 controls splicing and intron retention specifically in ribosomal proteins could have been impacted by selection bias: an enhanced capacity to detect changes in splicing of these transcripts rather than a specific defect in their processing.

How can selection bias be mitigated?

Here we argue that selection bias not only can have deleterious impacts on RNA-seq-based studies of pre-mRNA splicing, but in fact has had and will continue to have such impacts unless and until knowledge of this problem and its potential solutions becomes widely appreciated. While this problem has been previously noted¹⁹, our hope here is to bring renewed attention to this issue. Importantly, while this problem has been carefully considered in the context of standard gene expression studies^{4,5}, a cursory examination of the literature suggests this problem nevertheless continues to pervade current work. Whereas Oshlack provided an elegant mathematical approach for mitigating the impacts of selection bias in GO-term enrichment analyses, we unfortunately offer no such solution here. While we hope that others might provide tools to mitigate this problem in the future, we suggest that the simplest solution to mitigate this problem is the use of approaches that either enrich for, or otherwise increase the number of, splicing-informative reads across the complement of genomic substrates^{11,20}, thereby reducing the differences in sampling across the datasets.

Importantly, while the work examined here involved short-read, Illumina-based sequencing, we note that this is not a problem unique to this platform but rather reflects a fundamental statistical challenge associated with analyzing datasets with small numbers of measured events. As such, we expect that this problem will be even greater in analyses of datasets from long-read sequencing platforms where the number of reads per experiment is typically much lower, and likewise in single-cell experiments where the number of reads per cell is dramatically reduced. Indeed, evidence of such bias in single-cell experiments has been recently demonstrated²¹. Similarly, while many software packages have been developed which enable more sensitive detection of splicing ‘hits’ within an experiment^{22–26}, the problem of selection bias is fundamentally an issue of how to handle those events which are not identified as hits: those wherein the absence of evidence cannot be interpreted as evidence of absence. We are unaware of any software packages that are widely available today that account for this problem but hope that such solutions will soon arise. In the meantime, we conclude by noting that as users of these technologies, whether that be as experimentalists generating and analyzing such data, or as consumers evaluating the work of others, it will be essential for all of us to consider the possibility that an apparently statistically significant conclusion may not reflect a meaningful biological property but instead may be the result of a statistical aberration.

Methods

Cell growth. Wild-type cells and those harboring the *prp2-1* allele were streaked from glycerol stocks onto solid rich media (YPD) and grown at the permissive temperature (25 °C) for three days. In triplicate, single colonies were inoculated into 5 mL of YPD and grown at 25 °C with shaking at 200 rpm overnight. Cultures were back-diluted into 20 mL of YPD to an OD₆₀₀ of 0.05 and incubated at 25 °C with shaking at 200 rpm. Upon reaching an OD₆₀₀ of approximately 0.75, cultures were transferred to a 37 °C shaking water bath (200 rpm) for 10 min. Cells were collected via vacuum filtration and pellets were immediately flash-frozen in liquid nitrogen and stored at –80 °C.

MPE-seq library preparation. RNA was purified and MPE-seq libraries were prepared as previously described⁶ with the exception that biotin-11-dUTP was used in place of aminoallyl-dUTP during reverse

transcription such that no separate biotin coupling step was necessary. Instead, following hydroxide treatment an elution volume of 50 μ L was used during the zymo column clean-up which went immediately into the first bead purification.

Downsampling. Using a custom script, each read was assigned a random number from 0 to 1 and sorted based on their random number. The top 800,000, 400,000, and 200,000 were included in the high, medium, and low datasets, respectively. Random number generation used a seed value of 1 to allow future reproducibility.

Alignment and quantification. The full and downsampled datasets were processed as follows: Reads were trimmed of sequencing adapters using fastp²⁷ with the following parameters: --adapter_sequence CTGTCTCTTATACACATCT --adapter_sequence_r2 CTGTCTCTTATACACATCT. Trimmed reads were aligned to the R64-2-1 genome release from SGD with hisat2²⁸ with the following parameters: --max-intronlen 2000 --no-unal and reads with MAPQ scores below 5 were removed with samtools²⁹. Unspliced and spliced counts were obtained with a custom script based on HTSeq³⁰. DESeq2¹² was used to assess differential splicing.

Data processing. FPKM values for host genes and identified retained introns were obtained from Corsini et al. (GEO GSM2535498 and Table S2 therein, respectively)¹⁴. All mm9 UCSC introns were broken into groups based on whether they were identified as significant and whether they are ribosomal protein genes. Distances from the end of each distinct intron (as determined by their chromosome, start, and stop positions) and the end of the host transcript were calculated from genomic coordinates. In the case that an intron existed within multiple transcript isoforms, the shortest isoform was considered.



Reporting summary. Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All newly generated sequencing data are available through NCBI's Gene Expression Omnibus (GEO) at accession number [GSE160046](#), and the data previously reported¹⁴ obtained under accession number [GSM2535498](#).

Code availability

Code for basic analysis steps is available on GitHub (<https://github.com/zdwyer/Problem-of-Selection-Bias>).

Zachary W. Dwyer  & Jeffrey A. Pleiss 

¹Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York, USA. ✉e-mail: jpleiss@cornell.edu

Received: 19 January 2021; Accepted: 24 March 2023;

Published online: 08 April 2023

References

1. Antman, K. et al. Selection bias in clinical trials. *J. Clin. Oncol.* **3**, 1142–1147 (1985).
2. Tripepi, G., Jager, K. J., Dekker, F. W. & Zoccali, C. Selection bias and information bias in clinical research. *NEC* **115**, c94–c99 (2010).

3. Westhoff, C. L. Epidemiologic studies: pitfalls in interpretation. *Dialogues Contracept.* **4**, 5–6 (1995).
4. Oshlack, A. & Wakefield, M. J. Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct* **4**, 14 (2009).
5. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* **11**, R14 (2010).
6. Gildea, M. A., Dwyer, Z. W. & Pleiss, J. A. Multiplexed primer extension sequencing: A targeted RNA-seq method that enables high-precision quantitation of mRNA splicing isoforms and rare pre-mRNA splicing intermediates. *Methods* <https://doi.org/10.1016/j.ymeth.2019.05.013> (2019).
7. Wlodaver, A. M. & Staley, J. P. The DEXD/H-box ATPase Prp2p destabilizes and proofreads the catalytic RNA core of the spliceosome. *RNA* **20**, 282–294 (2014).
8. Warkocki, Z. et al. The G-patch protein Spp2 couples the spliceosome-stimulated ATPase activity of the DEAH-box protein Prp2 to catalytic activation of the spliceosome. *Genes Dev.* **29**, 94–107 (2015).
9. Kim, S. H. & Lin, R. J. Spliceosome activation by PRP2 ATPase prior to the first transesterification reaction of pre-mRNA splicing. *Mol. Cell Biol.* **16**, 6810–6819 (1996).
10. Bai, R. et al. Mechanism of spliceosome remodeling by the ATPase/helicase Prp2 and its coactivator Spp2. *Science* **371**, ea8e8863 (2021).
11. Xu, H., Fair, B. J., Dwyer, Z. W., Gildea, M. & Pleiss, J. A. Detection of splice isoforms and rare intermediates using multiplexed primer extension sequencing. *Nat. Methods* **16**, 55–58 (2019).
12. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
13. Gildea, M. A., Dwyer, Z. W. & Pleiss, J. A. Transcript-specific determinants of pre-mRNA splicing revealed through in vivo kinetic analyses of the 1st and 2nd chemical steps. *Mol. Cell.* **82**, 2967–2981.e6 <https://doi.org/10.1016/j.molcel.2022.06.020> (2022).
14. Corsini, N. S. et al. Coordinated control of mRNA and rRNA processing controls embryonic stem cell pluripotency and differentiation. *Cell Stem Cell* **12**, 543–558.e12 (2018).
15. Yan, D. et al. CUS2, a yeast homolog of human Tat-SF1, rescues function of misfolded U2 through an unusual RNA recognition motif. *Mol. Cell Biol.* **18**, 5000–5009 (1998).
16. Rodgers, M. L. et al. Conformational dynamics of stem II of the U2 snRNA. *RNA* **22**, 225–236 (2016).
17. Zhang, Z. et al. Molecular architecture of the human 17S U2 snRNP. *Nature* **583**, 310–313 (2020).
18. Herzel, L., Ottoz, D. S. M., Alpert, T. & Neugebauer, K. M. Splicing and transcription touch base: co-transcriptional spliceosome assembly and function. *Nat. Rev. Mol. Cell Biol.* **18**, 637–650 (2017).
19. Nazarov, P. V. et al. RNA sequencing and transcriptome arrays analyses show opposing results for alternative splicing in patient derived samples. *BMC Genomics* **18**, 443 (2017).
20. Li, H., Qiu, J. & Fu, X.-D. RASL-seq for massively parallel and quantitative analysis of gene expression. *Curr Protoc Mol Biol* **Chapter 4**, Unit 4.13.1–9 (2012).
21. Buen Abad Najar, C. F., Yosef, N. & Lareau, L. F. Coverage-dependent bias creates the appearance of binary splicing in single cells. *Life* **9**, e54603 (2020).
22. Norton, S. S., Vaquero-Garcia, J., Lahens, N. F., Grant, G. R. & Barash, Y. Outlier detection for improved differential splicing quantification from RNA-Seq experiments with replicates. *Bioinformatics* **34**, 1488–1497 (2018).
23. Shen, S. et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci USA*. **111**, E5593–601 <https://doi.org/10.1073/pnas.1419161111> (2014).
24. Zhang, Z. et al. Deep-learning augmented RNA-seq analysis of transcript splicing. *Nat. Methods* **16**, 307–310 (2019).
25. Huang, Y. & Sanguinetti, G. BRIE: transcriptome-wide splicing quantification in single cells. *Genome Biol.* **18**, 123 (2017).
26. Katz, Y., Wang, E. T., Airoidi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7**, 1009–1015 (2010).
27. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
28. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
29. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
30. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).

Acknowledgements

The authors thank members of the Pleiss lab for critical feedback on this work. We thank P. Schweitzer and the BRC Genomics Facility at Cornell for outstanding technical support with Illumina sequencing. This work was funded by NIH grants R01GM098634 and R01GM140082 to J.A.P.

Author contributions

Z.W.D. performed the experiments, Z.W.D. and J.A.P. designed the experiments, analyzed the data, and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-023-37650-2>.

Correspondence and requests for materials should be addressed to Jeffrey A. Pleiss.

Peer review information *Nature Communications* thanks Yi Xing and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023