CODECHECK Certificate 2022-018

10.5281/zenodo.7084333

Raniere Silva

September 27, 2022

Table 1: CODECHECK summary

Title	svaRetro and svaNUMT: Modular packages for annotation of retrotransposed transcripts and nuclear integration of mitochondrial DNA in genome sequencing data
Authors	Ruining Dong, Daniel Cameron, Justin Bedo, Anthony T Papenfuss
Reference	https://doi.org/10.46471/gigabyte.70
Summary	Only visualisation steps performed. All created figures match those in the article.
Repository	https://gitlab.com/cdchck/community-codechecks/2022-svaRetro-svaNUMT.git

Table 2: Summary of output files generated

Files	Comment
figure-2b.pdf	Figure 2(b) of the article
figure-3b.pdf	Figure 3(b) of the article
figure-4.pdf	Figure 4 of the article
figure-5.pdf	Figure 5 of the article
figure-6.pdf	Figure 6 of the article

Summary

The reproduction of the figures, from output data, in the article was straightforward given that the authors provided R Markdown (`.Rmd`) files. Figure 3 and Figure 2 have minor aesthetics differences. The whole pipeline was not reproduced!

CODECHECKER notes

Data and Code

As a repository was not provided by the author, codechecker made one. After creating an empty repository, `MANIFEST` was created. `Scripts.zip` were copied to `scripts`. Other data and code were downloaded from the supplemented material in Zenodo (Dong et al. 2022) by running

```
$ make download
```

Software Installation

The provided `.Rmd` files requires many packages from Bioconductor. To facilitate the installation of packages, the `bioconductor/bioconductor_docker` Docker image (version 3.15) was used. The Docker container can be started by running

```
$ docker compose up dev
```

Packages installation instructions are included in the `.Rmd` files.

Running the Script

`Figures2-4.Rmd` is the main script and was edited to include `sessionInfo()` at the end of the document, see Git history for details. To regenerate the figures, we simply render `Figures2-4.Rmd` using the RStudio Server included as part of the Bioconductor Docker image.

Figure 4 uses statistics from `sim_reads_rt.Rmd` which was not executed.

`gnomad.Rmd` is the script that renders Figure 5 and Figure 6 of the article and was edited to include `sessionInfo()` at the end of the document, see Git history for details. The block

```
#function from SVEnsemble
wkdir <- getwd()
gnomad.bnd.gr <- suppressWarnings(gnomadSV(paste0(wkdir, "/gnomad_v2.1_sv.sites.vcf.gz")))

gnomad.rt <- rtDetect(
  filter(gnomad.bnd.gr, FILTER=="PASS"),
  hg19.genes,
  maxgap = 1000,
  minscore = 0.4
)
```

takes a couple of hours to execute and might benefit to have the calculation cached. The pipeline

```
gnomad.insSite.pass.rmsk <- gnomad.insSite.pass %>% filter(rtFoundSum==T) %>%
  find_overlaps(., rmsk.gr, maxgap = 100) %>%
  unique() %>%
  as_tibble() %>%
  count(repClass) %>%
  bind_rows(tibble(repClass='non-repeats', n=n.rt-sum(.$n)))
```

failed with

```
Error in count(., repClass) : Argument 'x' is not a vector: list
```

that was resolved by replacing

```
count(repClass) %>%
```

with

```
dplyr::count(repClass) %>%
```

This error is due `svaRetro` be loaded after `dplyr`:

```
## Loading required package: matrixStats
## Attaching package: 'matrixStats'
## The following object is masked from 'package:dplyr':
##
##     count
```

The pipeline

```
gnomad.insSite.pass.rmsk %>%
  ggplot(., aes(x = reorder(repClass, n), y = n)) +
  geom_bar(stat = "identity", fill="lightblue3") + coord_flip() +
  labs(y="count", x="repClass") + ylim(0, 300) +
  geom_text(aes(label=n), hjust=-0.5, color="black",
            position = position_dodge(0.9), size=3.5) +
  #scale_fill_brewer(palette="Accent") +
  theme_minimal()
```

failed with

```
Error in ggplot(., aes(x = reorder(repClass, n), y = n)) :
  could not find function "ggplot"
```

that was resolved by adding

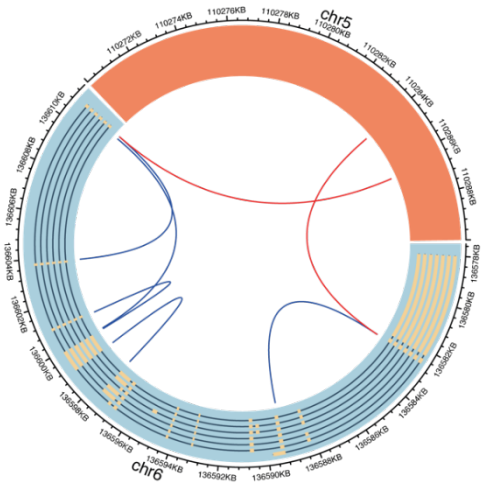
```
library(ggplot2)
```

at the begin of the code block.

Outputs

Figure 1: Comparison of Figure 2(b) of the article.

(a) From pre-print.



(b) Re-run of `scripts/Figures2-4.Rmd`.

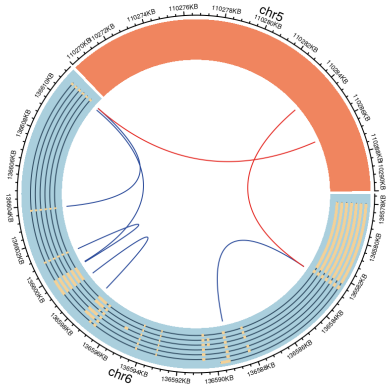
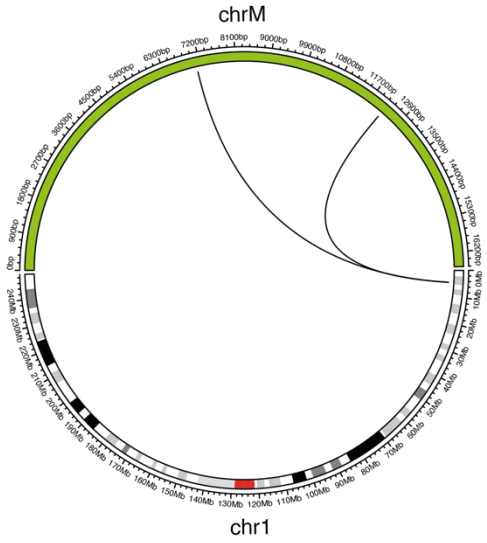


Figure 2: Comparison of Figure 3(b) of the article.

(a) From pre-print.



(b) Re-run of `scripts/Figures2-4.Rmd`.

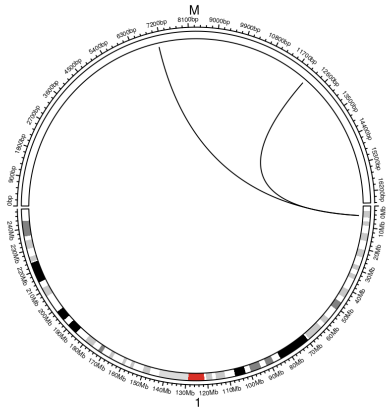
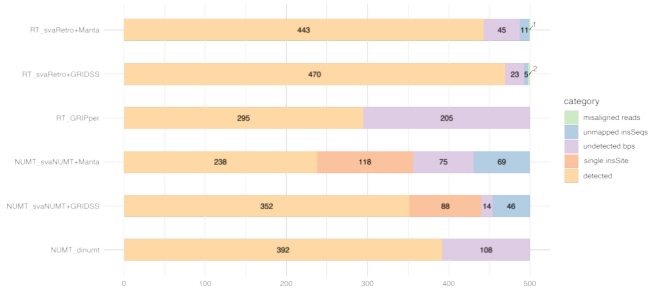


Figure 3: Comparison of Figure 4 of the article.

(a) From pre-print.



(b) Re-run of `scripts/Figures2-4.Rmd`.

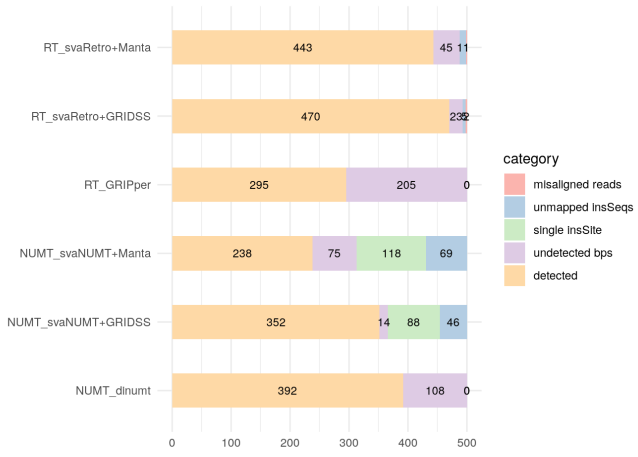
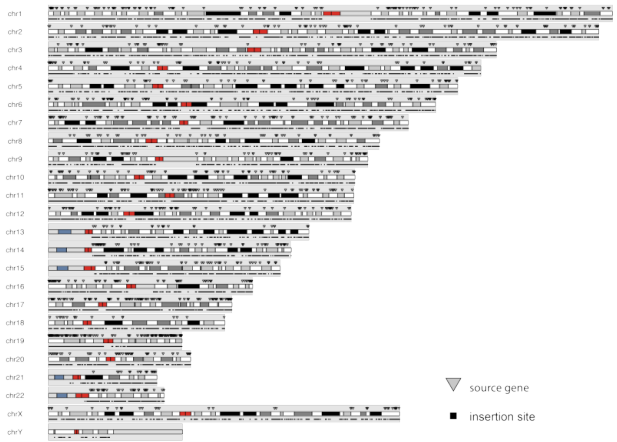


Figure 4: Comparison of Figure 5 of the article.

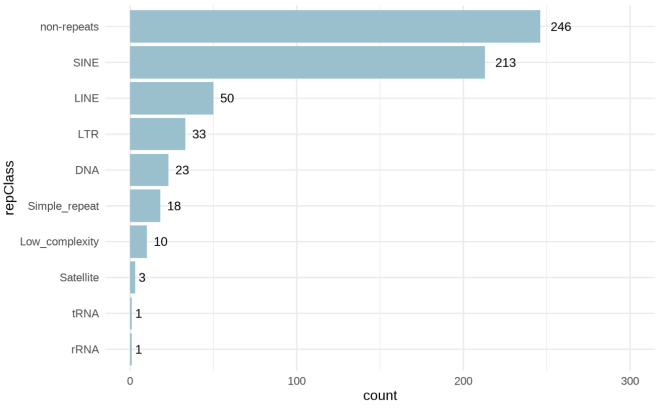
(a) From pre-print.



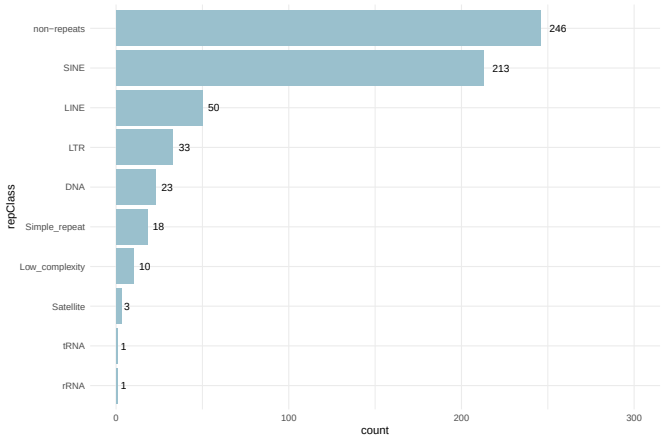
(b) Re-run of `scripts/gnomad.Rmd`.

Figure 5: Comparison of Figure 6 of the article.

(a) From pre-print.



(b) Re-run of `scripts/gnomad.Rmd`.



References

Dong, Ruining, Daniel Cameron, Justin Bedo, and Anthony T Papenfuss. 2022. "Data and Scripts for the Manuscript of svaRetro and svaNUMT: Modular Packages for Annotating Retrotransposed Transcripts and Nuclear Integration of Mitochondrial DNA in Genome Sequencing Data." Zenodo. https://doi.org/10.5281/ZENODO.7006177.

Colophon

This document was built with Quarto.

Session Info

```
sessionInfo()
```

```
R version 4.2.1 (2022-06-23)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 22.04.1 LTS

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.10.0
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.10.0

locale:
 [1] LC_CTYPE=en_GB.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_GB.UTF-8        LC_COLLATE=en_GB.UTF-8
 [5] LC_MONETARY=en_GB.UTF-8    LC_MESSAGES=en_GB.UTF-8
 [7] LC_PAPER=en_GB.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

loaded via a namespace (and not attached):
 [1] digest_0.6.29   jsonlite_1.8.0  magrittr_2.0.3  evaluate_0.16
 [5] highr_0.9       rlang_1.0.5     stringi_1.7.8   cli_3.4.0
 [9] rstudioapi_0.14 rmarkdown_2.16  tools_4.2.1     stringr_1.4.1
[13] xfun_0.33       yaml_2.3.5      fastmap_1.1.0   compiler_4.2.1
[17] htmltools_0.5.3 knitr_1.40
```

`Figures2-4.Rmd`'s session info:

```
## R version 4.2.1 (2022-06-23)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.4 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
## LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/liblapack.so.3
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
```

```
##
## attached base packages:
## [1] stats4    stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] BSgenome.Hsapiens.UCSC.hg19_1.4.3
##  [2] BSgenome_1.64.0
##  [3] ggplot2_3.3.6
##  [4] readr_2.1.2
##  [5] plyranges_1.16.0
##  [6] svaNUMT_1.3.0
##  [7] svaRetro_1.2.0
##  [8] StructuralVariantAnnotation_1.12.0
##  [9] VariantAnnotation_1.42.1
## [10] Rsamtools_2.12.0
## [11] Biostrings_2.64.1
## [12] XVector_0.36.0
## [13] SummarizedExperiment_1.26.1
## [14] MatrixGenerics_1.8.1
## [15] matrixStats_0.62.0
## [16] rtracklayer_1.56.1
## [17] circlize_0.4.15
## [18] dplyr_1.0.10
## [19] TxDb.Hsapiens.UCSC.hg19.knownGene_3.2.2
## [20] GenomicFeatures_1.48.3
## [21] AnnotationDbi_1.58.0
## [22] Biobase_2.56.0
## [23] GenomicRanges_1.48.0
## [24] GenomeInfoDb_1.32.4
## [25] IRanges_2.30.1
## [26] S4Vectors_0.34.0
## [27] BiocGenerics_0.42.0
##
## loaded via a namespace (and not attached):
##  [1] bitops_1.0-7             bit64_4.0.5              RColorBrewer_1.1-3
##  [4] filelock_1.0.2           progress_1.2.2           httr_1.4.4
##  [7] tools_4.2.1              bslib_0.4.0              utf8_1.2.2
## [10] R6_2.5.1                 DBI_1.1.3                colorspace_2.0-3
## [13] withr_2.5.0              tidyselect_1.1.2         prettyunits_1.1.1
## [16] bit_4.0.4                curl_4.3.2               compiler_4.2.1
## [19] textshaping_0.3.6        cli_3.4.0                xml2_1.3.3
## [22] DelayedArray_0.22.0      labeling_0.4.2           sass_0.4.2
## [25] scales_1.2.1             rappdirs_0.3.3           systemfonts_1.0.4
## [28] stringr_1.4.1            digest_0.6.29            rmarkdown_2.16
## [31] pkgconfig_2.0.3          htmltools_0.5.3          highr_0.9
## [34] dbplyr_2.2.1             fastmap_1.1.0            rlang_1.0.5
## [37] GlobalOptions_0.1.2      rstudioapi_0.14          RSQLite_2.2.17
## [40] farver_2.1.1             shape_1.4.6              jquerylib_0.1.4
## [43] BiocIO_1.6.0             generics_0.1.3           jsonlite_1.8.0
## [46] BiocParallel_1.30.3      RCurl_1.98-1.8           magrittr_2.0.3
## [49] GenomeInfoDbData_1.2.8   Matrix_1.4-1             munsell_0.5.0
## [52] Rcpp_1.0.9               fansi_1.0.3              lifecycle_1.0.2
## [55] stringi_1.7.8            yaml_2.3.5               zlibbioc_1.42.0
## [58] BiocFileCache_2.4.0      grid_4.2.1               blob_1.2.3
## [61] parallel_4.2.1           crayon_1.5.1             lattice_0.20-45
```

```
## [64] hms_1.1.2                KEGGREST_1.36.3          knitr_1.40
## [67] pillar_1.8.1             rjson_0.2.21             codetools_0.2-18
## [70] biomaRt_2.52.0           XML_3.99-0.10            glue_1.6.2
## [73] evaluate_0.16            tzdb_0.3.0               png_0.1-7
## [76] vctrs_0.4.1              tidyr_1.2.1              gtable_0.3.1
## [79] purrr_0.3.4              assertthat_0.2.1         cachem_1.0.6
## [82] xfun_0.33                restfulr_0.0.15          ragg_1.2.2
## [85] tibble_3.1.8             GenomicAlignments_1.32.1 memoise_2.0.1
## [88] ellipsis_0.3.2
```

gnomad.Rmd's session info:

```
## R version 4.2.1 (2022-06-23)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.4 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
## LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/liblapack.so.3
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats4    stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] TxDb.Hsapiens.UCSC.hg19.knownGene_3.2.2
##  [2] GenomicFeatures_1.48.3
##  [3] AnnotationDbi_1.58.0
##  [4] svaRetro_1.2.0
##  [5] StructuralVariantAnnotation_1.12.0
##  [6] VariantAnnotation_1.42.1
##  [7] Rsamtools_2.12.0
##  [8] Biostrings_2.64.1
##  [9] XVector_0.36.0
## [10] SummarizedExperiment_1.26.1
## [11] Biobase_2.56.0
## [12] MatrixGenerics_1.8.1
## [13] matrixStats_0.62.0
## [14] rtracklayer_1.56.1
## [15] karyoploteR_1.22.0
## [16] regioneR_1.28.0
## [17] ggplot2_3.3.6
## [18] dplyr_1.0.10
## [19] plyranges_1.16.0
## [20] GenomicRanges_1.48.0
## [21] GenomeInfoDb_1.32.4
## [22] IRanges_2.30.1
## [23] S4Vectors_0.34.0
## [24] BiocGenerics_0.42.0
```

```
##
## loaded via a namespace (and not attached):
##   [1] colorspace_2.0-3         rjson_0.2.21             deldir_1.0-6
##   [4] ellipsis_0.3.2           biovizBase_1.44.0        htmlTable_2.4.1
##   [7] base64enc_0.1-3          dichromat_2.0-0.1        rstudioapi_0.14
##  [10] bit64_4.0.5              fansi_1.0.3              xml2_1.3.3
##  [13] codetools_0.2-18         splines_4.2.1            cachem_1.0.6
##  [16] knitr_1.40               Formula_1.2-4            jsonlite_1.8.0
##  [19] cluster_2.1.3            dbplyr_2.2.1             png_0.1-7
##  [22] compiler_4.2.1           httr_1.4.4               backports_1.4.1
##  [25] lazyeval_0.2.2           assertthat_0.2.1         Matrix_1.4-1
##  [28] fastmap_1.1.0            cli_3.4.0                htmltools_0.5.3
##  [31] prettyunits_1.1.1        tools_4.2.1              gtable_0.3.1
##  [34] glue_1.6.2               GenomeInfoDbData_1.2.8   rappdirs_0.3.3
##  [37] Rcpp_1.0.9               jquerylib_0.1.4          vctrs_0.4.1
##  [40] xfun_0.33                stringr_1.4.1            lifecycle_1.0.2
##  [43] ensembldb_2.20.2         restfulr_0.0.15          XML_3.99-0.10
##  [46] zlibbioc_1.42.0          scales_1.2.1             BSgenome_1.64.0
##  [49] ProtGenerics_1.28.0      hms_1.1.2                parallel_4.2.1
##  [52] AnnotationFilter_1.20.0  RColorBrewer_1.1-3       yaml_2.3.5
##  [55] curl_4.3.2               memoise_2.0.1            gridExtra_2.3
##  [58] sass_0.4.2               biomaRt_2.52.0           rpart_4.1.16
##  [61] latticeExtra_0.6-30      stringi_1.7.8            RSQLite_2.2.17
##  [64] BiocIO_1.6.0             checkmate_2.1.0          filelock_1.0.2
##  [67] BiocParallel_1.30.3      rlang_1.0.5              pkgconfig_2.0.3
##  [70] bitops_1.0-7             evaluate_0.16            lattice_0.20-45
##  [73] purrr_0.3.4              htmlwidgets_1.5.4        GenomicAlignments_1.32.1
##  [76] bit_4.0.4                tidyselect_1.1.2         magrittr_2.0.3
##  [79] R6_2.5.1                 generics_0.1.3           Hmisc_4.7-1
##  [82] DelayedArray_0.22.0      DBI_1.1.3                pillar_1.8.1
##  [85] foreign_0.8-82           withr_2.5.0              survival_3.3-1
##  [88] KEGGREST_1.36.3          RCurl_1.98-1.8           nnet_7.3-17
##  [91] tibble_3.1.8             crayon_1.5.1             interp_1.1-3
##  [94] utf8_1.2.2               BiocFileCache_2.4.0      rmarkdown_2.16
##  [97] bamsignals_1.28.0        jpeg_0.1-9               progress_1.2.2
## [100] grid_4.2.1               data.table_1.14.2        blob_1.2.3
## [103] digest_0.6.29            bezier_1.1.2             munsell_0.5.0
## [106] bslib_0.4.0
```