Developing a Variant Calling Pipeline for Ribosomal DNA

SUPPLEMENTARY FIGURES

Table of contents

Figure S1. GC content per rDNA region with ascending window values. 2
Figure S2. IGV images of the uncalled variants in 30x and 100x sets. 2
Figure S3. Percentage of retained reads during the preprocessing. 3
Figure S4. Simulated variants. 4
Figure S5. DNA-like regions included in the set of sequences to generate reads. 4
Figure S6. Distribution of the parameters of the final output VCF for the first version of the pipeline 5
Figure S7. Summary statistics of the results' performance for the first version of the pipeline. 5
Figure S8. Distribution of the parameters of the final output VCF for the second version of the pipeline. 6
Figure S9. Summary statistics of the results' performance for the second version of the pipeline. 7
Figure S10. Manual Join-genotyping for the second version of the pipeline. 7
Figure S11. Evaluation of the distributions of the different evaluated positions in the final version. 8
Figure S12. GQ distributions for all variants called. 8
Figure S13. Heatmap of the chrR, 24 unique copies of CHM13, 14 maternal, 23 paternal copies sequences. 9

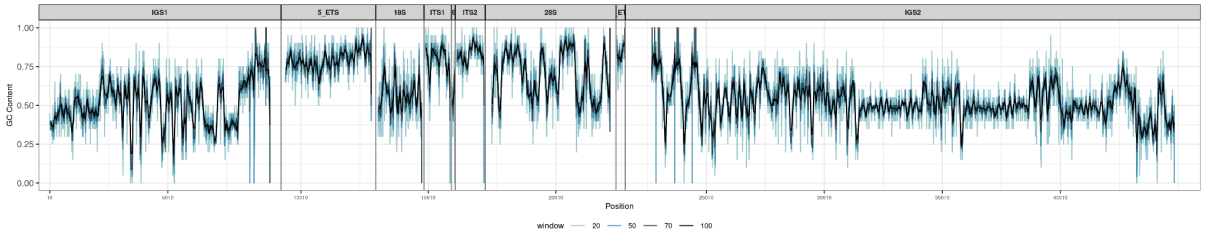

Figure S1. GC content per rDNA region with ascending window values.

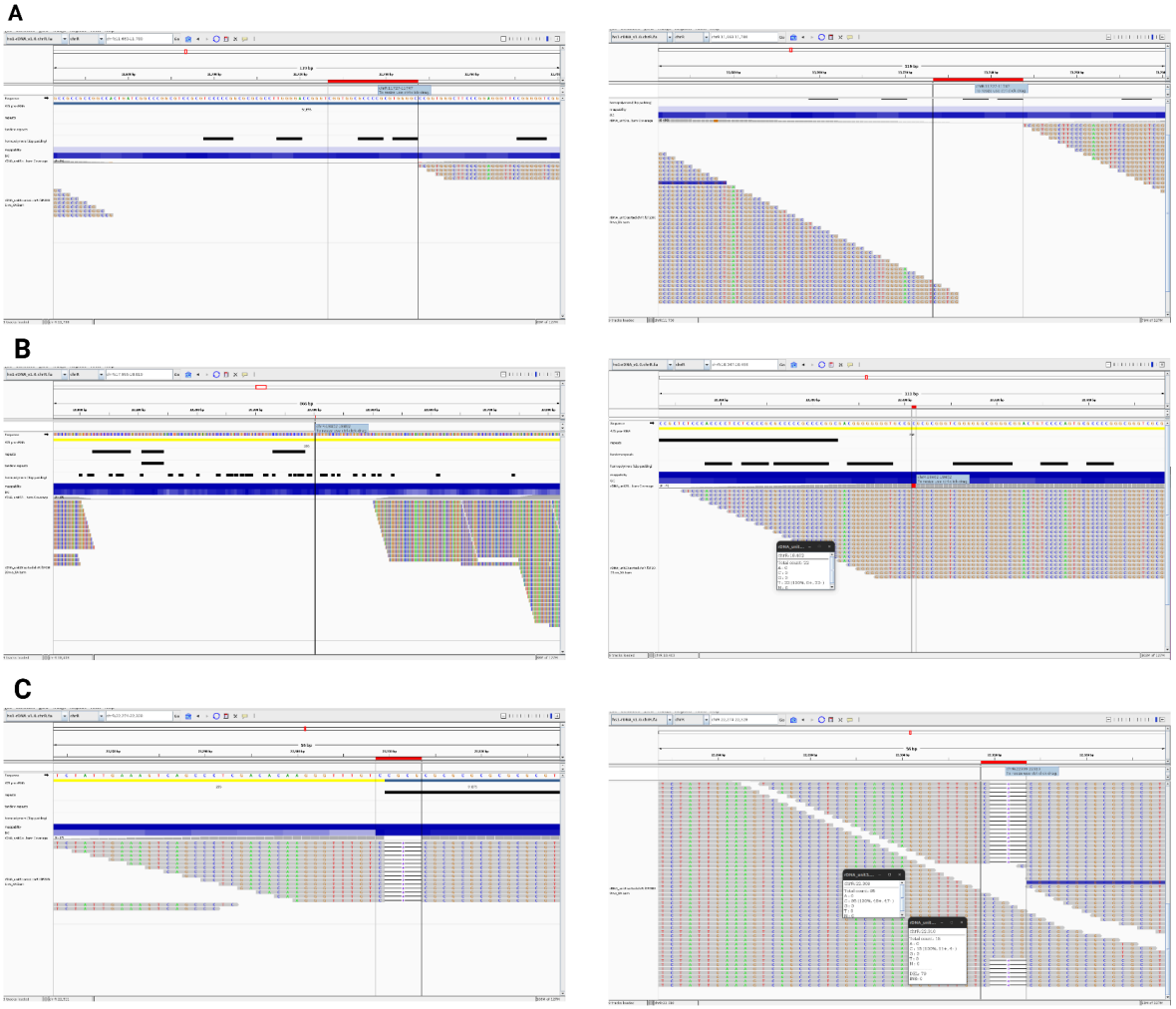

Figure S2. IGV images of the uncalled variants in 30x and 100x sets.

IGV images of the uncalled variants for positions **(A)** 11727, **(B)** 18402, and **(C)** 22309 in 30x (right) and 100x (left) sets.

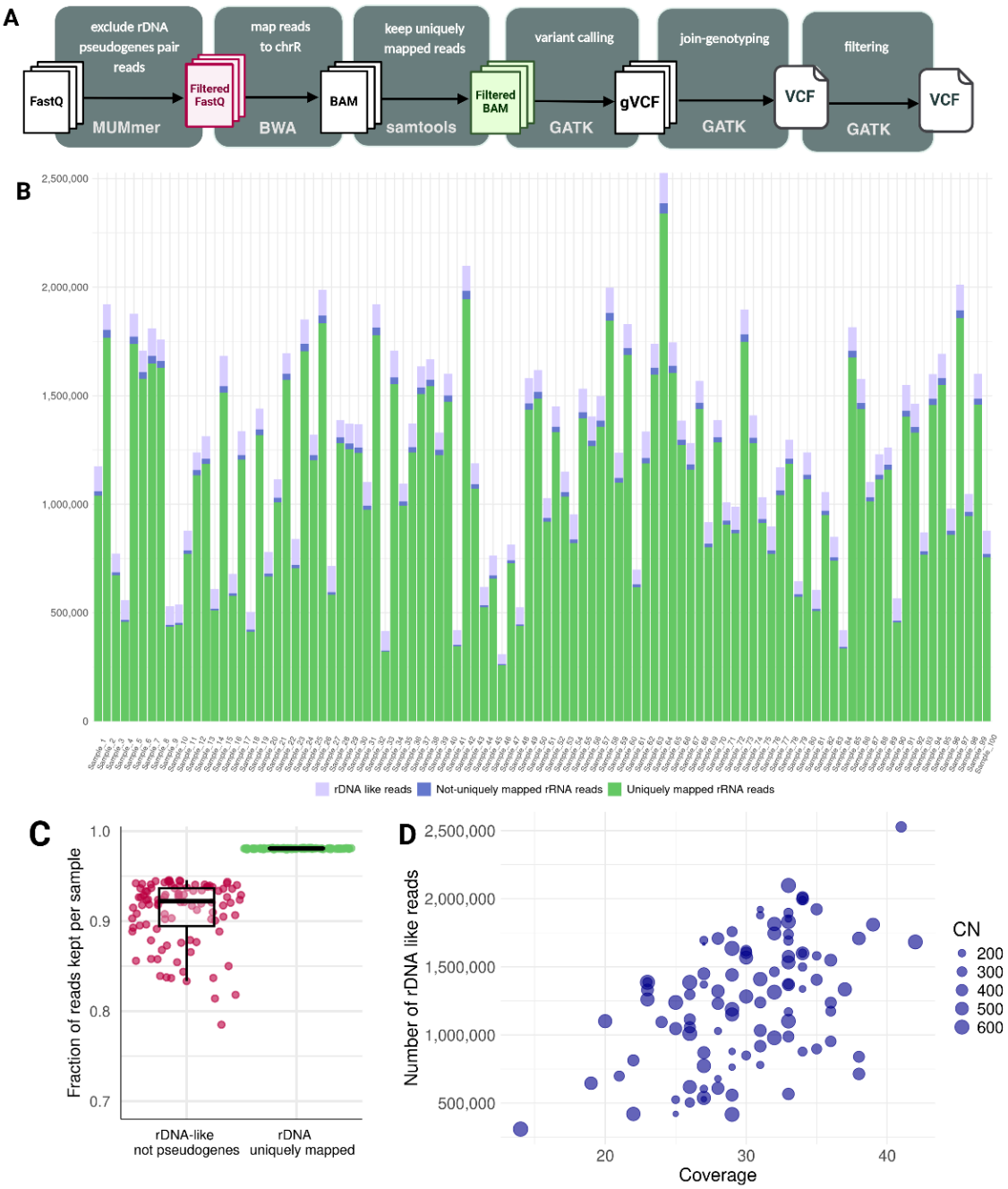

Figure S3. Percentage of retained reads during the preprocessing.

(A) Pipeline's schema. From left to right, the user input and the output files are generated and used as input in each step. The colored files had been filtered for rDNA-like reads that are not pseudogenes (red) and rDNA uniquely mapped (green). The same color schema is used across figures. **(B)** Barplot of the removed reads per sample. **(C)** Boxplots with the total fraction kept of rDNA-like read that are not pseudogenes (red) and rDNA uniquely mapped (green) per sample. **(D)** Distribution of each sample's coverage and copy number of DNA-like reads.

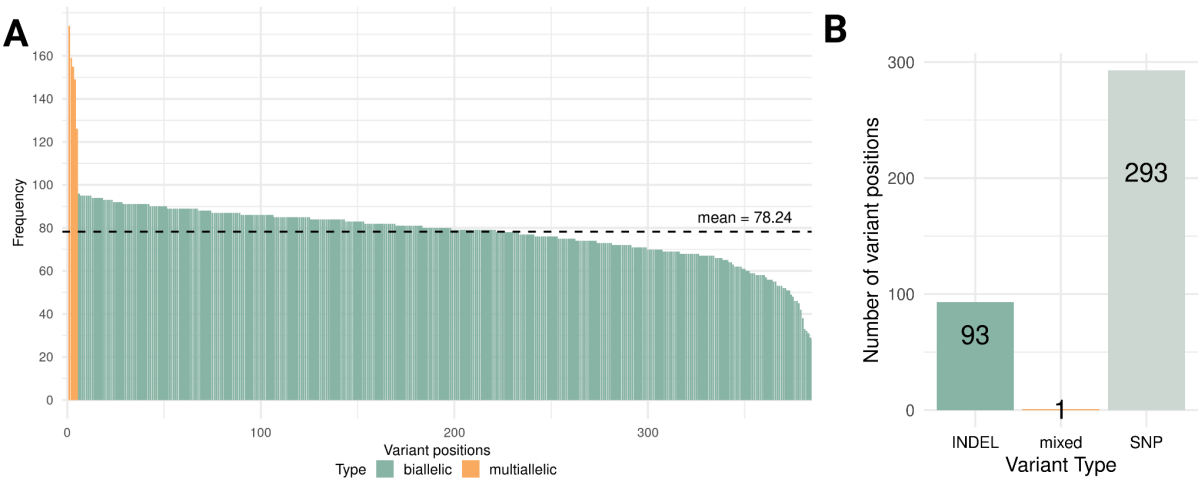

Figure S4. Simulated variants

A) Number of samples in which each artificial variant was included. B) Number of variants included in the samples per category.

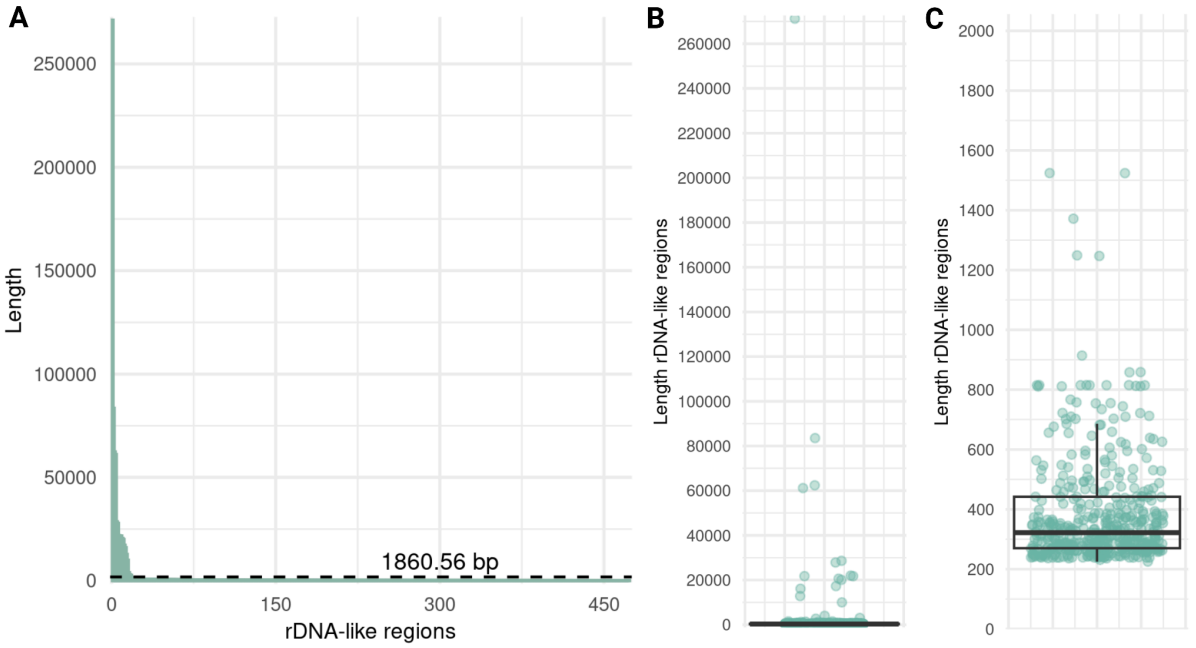

Figure S5. DNA-like regions included in the set of sequences to generate reads

A) Barplot with the rDNA-like regions sorted by length with an average of 1860.56 bp. **B)** Boxplot of the rDNA-like regions lengths. **C)** Zoomed boxplot (0 to 3000 bp range) to visualize better the distribution. It can be seen that most regions are around 300 bp long.

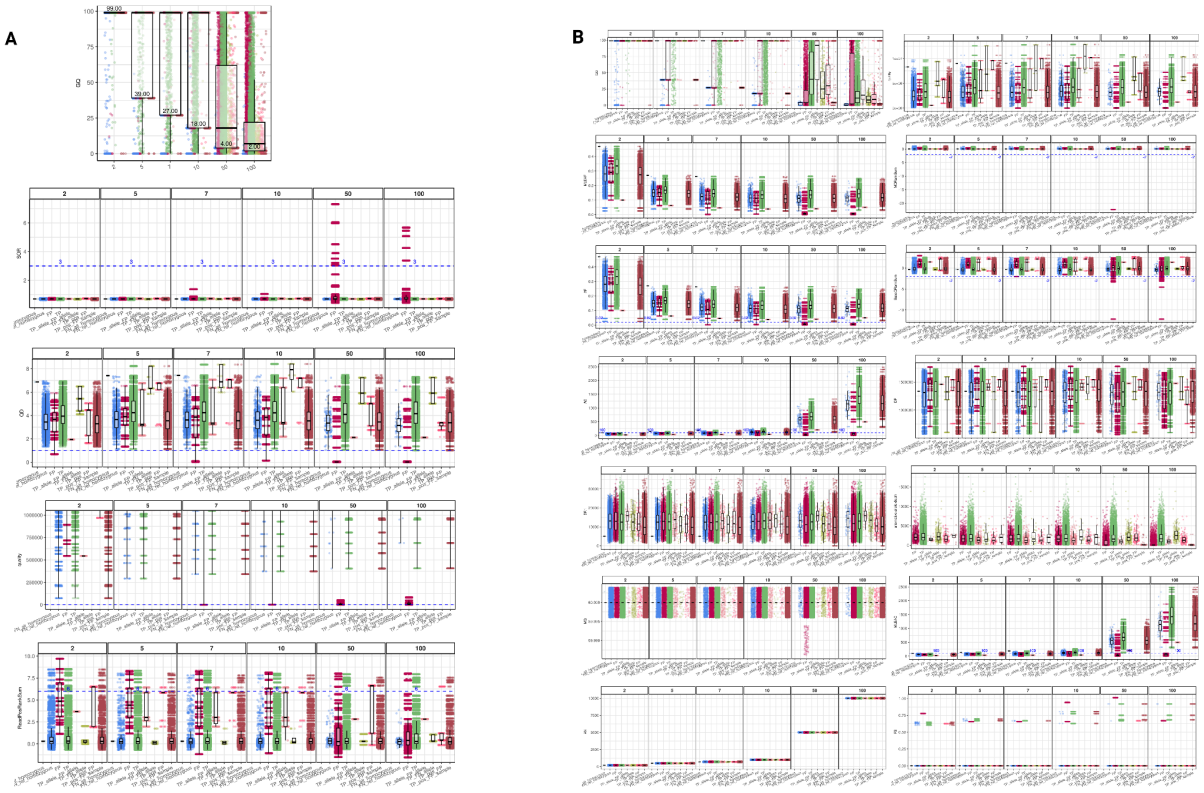

Figure S6. Distribution of the final VCF output file parameters in the first version of the pipeline
A) Parameters considered for the filtering step. **B)** Parameters discarded for the filtering step.

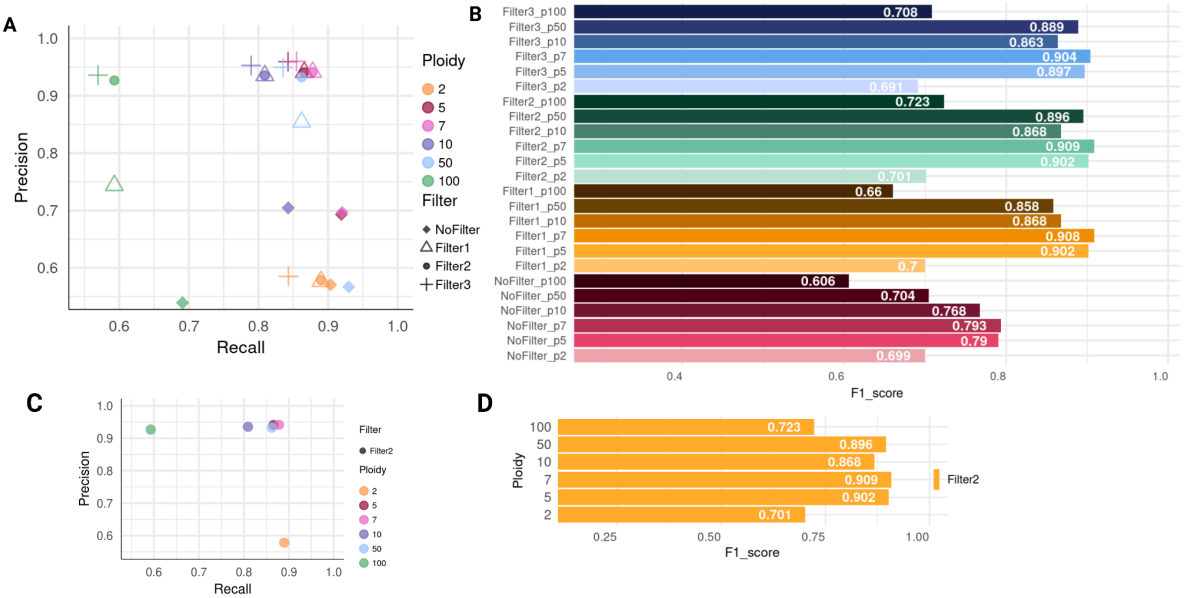

Figure S7. Summary statistics of the results' performance for the first version of the pipeline.
A) Precision against recall plot of all explored ploidy-filter combinations. **B)** Barplot of the F1 score of each explored ploidy-filter combinations. **C)** Precision against recall plot of the best-performing filter per ploidy. **D)** Barplot of the F1 score of the best-performing filter per ploidy.

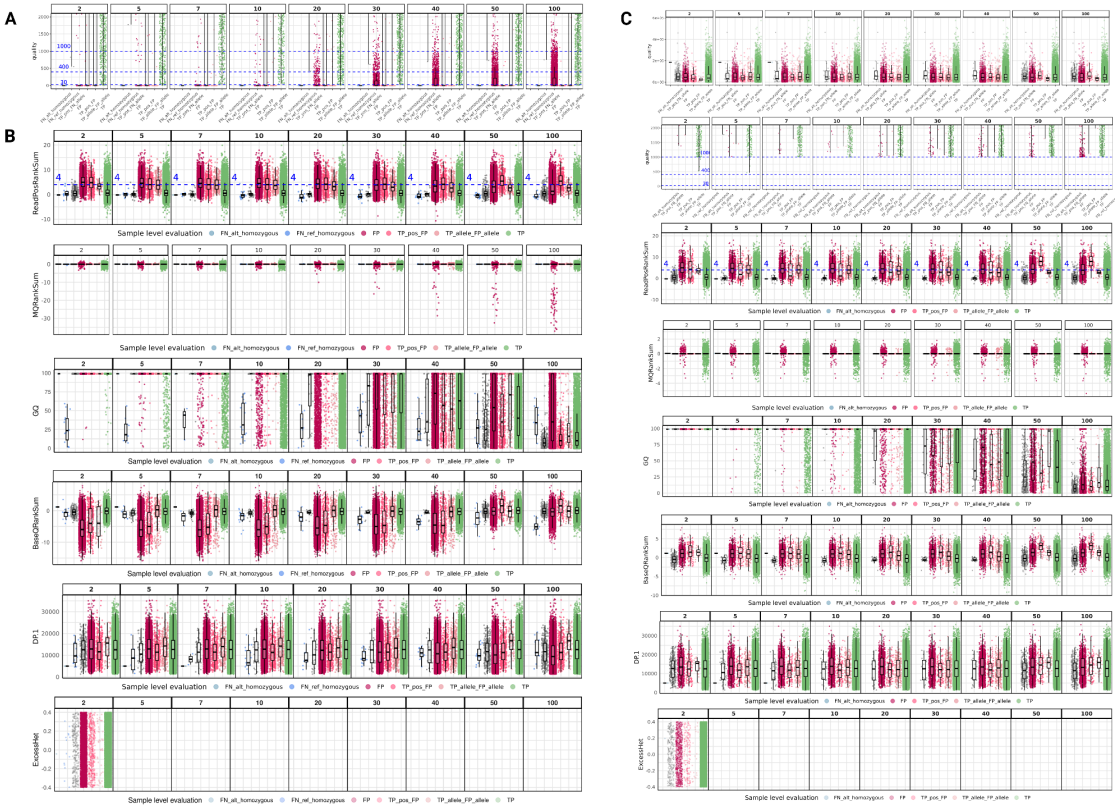

Figure S8. Distribution of the parameters of the final output VCF for the second version of the pipeline

A) Parameters considered for the filtering step. **B)** Parameters discarded for the filtering step.

C) Distribution of the parameters after applying the 4th version of the filter.

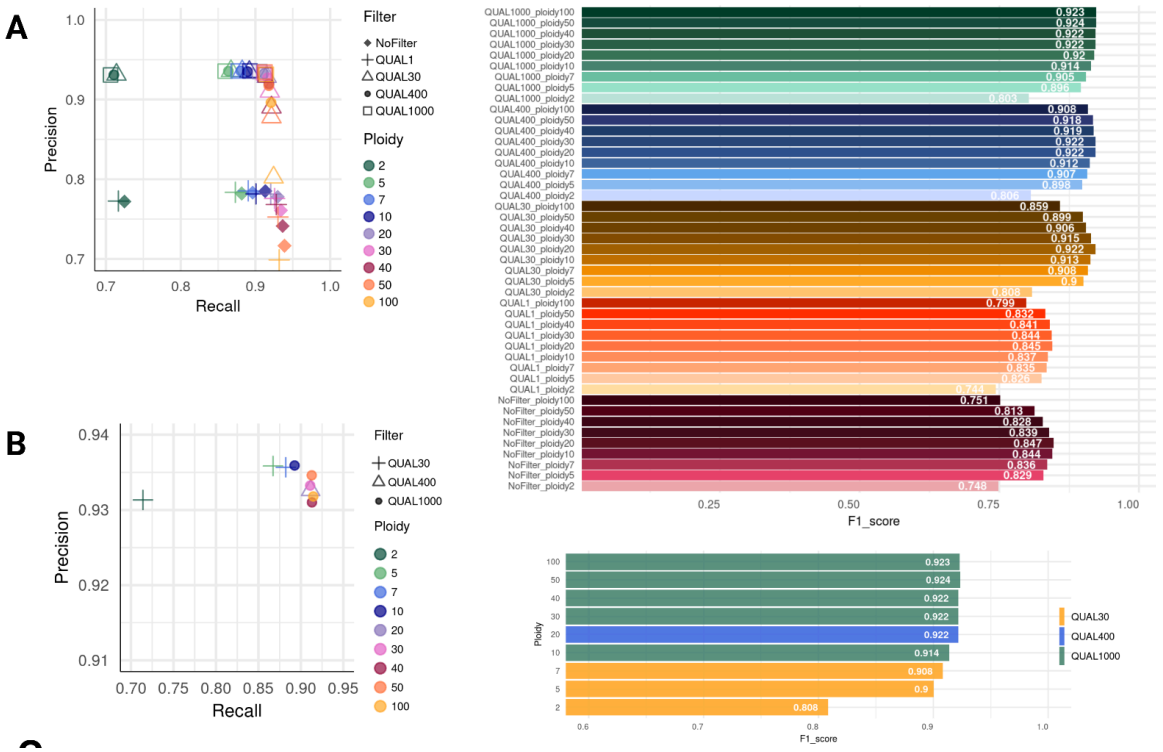

Figure S9. Summary statistics of the results' performance for the second version of the pipeline

A) Precision against recall plot and barplot of the F1 score of each explored ploidy-filter combination. **B)** Precision against recall plot and barplot of the F1 score of the best-performing filter per ploidy.

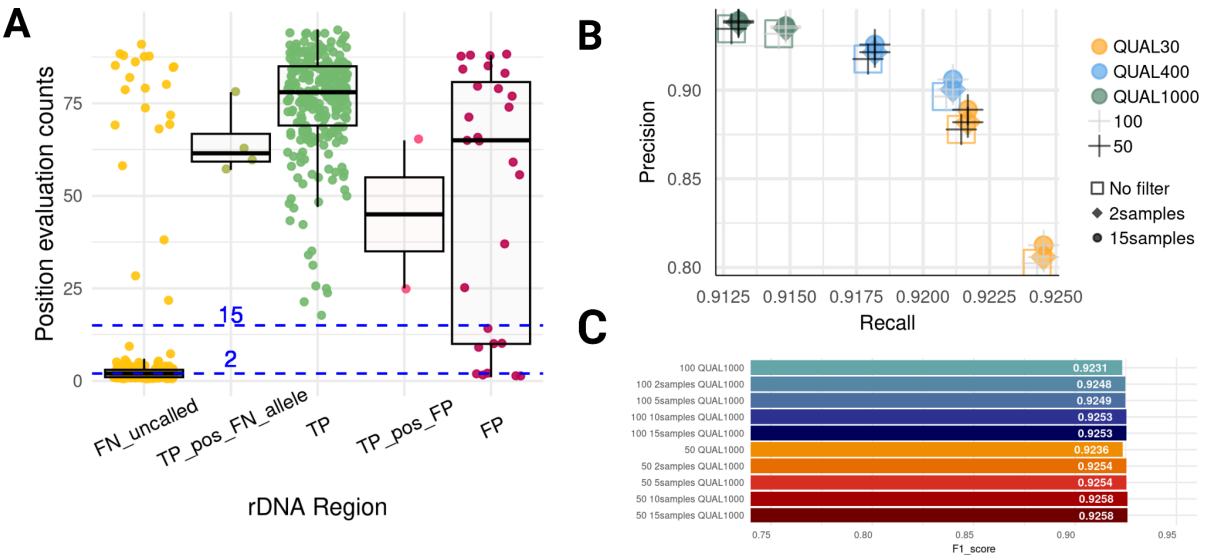

Figure S10. Manual Join-genotyping for the second version of the pipeline

A) Distribution of the number of times each position was classified in each evaluation. The thresholds in blue highlight the considered cut-off values. **B**) Precision by recall after removing those samples appearing in less than 2 or 15 samples. **C**) F1 score after removing those samples appearing in less than 2 or 15 samples.

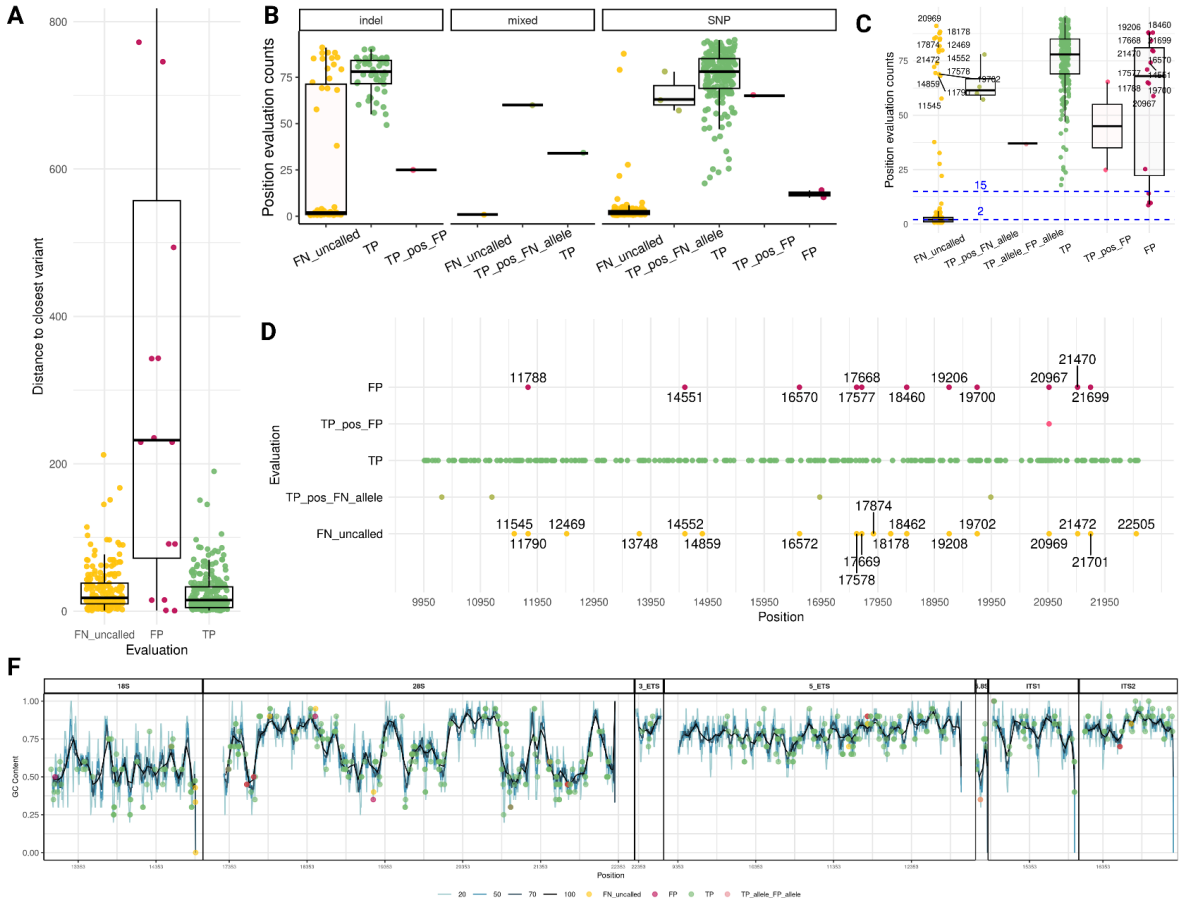

Figure S11. Distributions of the different evaluated positions after the final filter for version 2 of the pipeline (QUAL the 1000 and 2 samples)

A) Distance to the closest variant position distribution per evaluated type **B**) Distribution of the number of times each position was found in each category per evaluation category per variant type. **C**) Distribution of the number of times each position was found in each category per evaluation with the positions evaluated as FP or FN in more than 50 samples labeled. **D**) Distribution of the positions throughout the rDNA that were classified in a given evaluation category in more than 50 samples with the FPs and FNs labeled **F**) GC-content of the rDNA sequence with ascending widow size clustered per region, with the positions evaluated as true positives are located compared to those always missed (FNs) and always wrongly called (FPs).

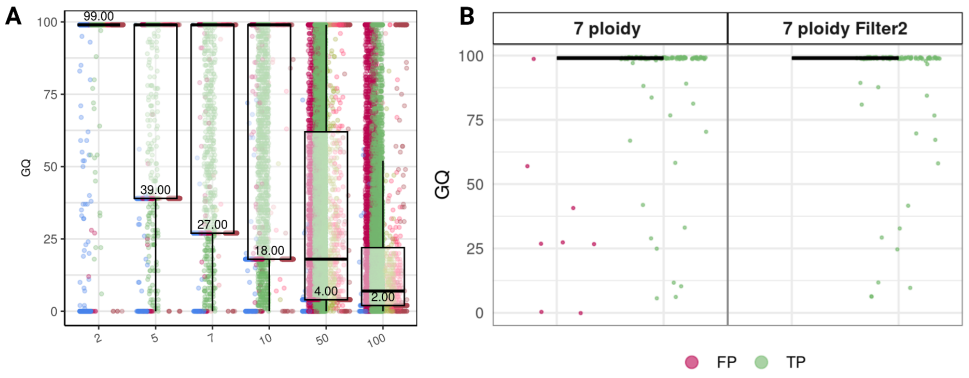

Figure S12. GQ distributions for all variants are called.

A) Simulated reads distribution. **B**) GQ distribution for the called variants of the T2T-CHM13 WGS data.

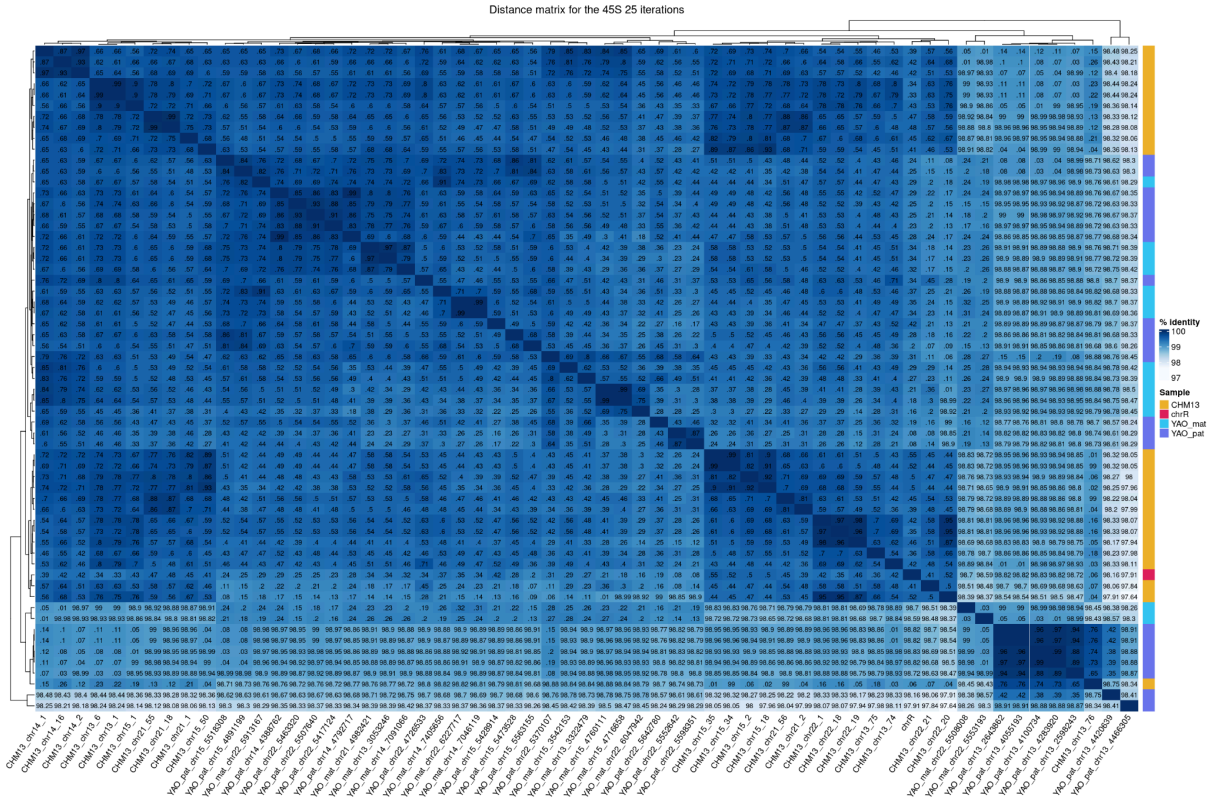

Figure S13. Heatmap of the chrR, 24 unique copies of CHM13, 14 maternal 24 paternal copies sequences.