.

Refining Feulgen: low-cost and accurate genome size measurements for everyone

Mohammed M.Tawfeeq[1,2*], Ursula Swaelus[1], Florence Rodriguez Gaudray[1], Jenny Arrensdorff[1], Felipe Ennes Silva[1,3], Laurent Grumiau[1], Thomas Verdebout[4] and Jean-François Flot[1,2,5*]

[1]Department of Organismal Biology, Université libre de Bruxelles (ULB), Av. F.D. Roosevelt, CP 160/12, Brussels, 1050, Belgium.
[2]Interuniversity Institute of Bioinformatics in Brussels – (IB)[2], Belgium.
[3]Department of Anthropology, New York University, USA.
[4]Department of Mathematics, Université libre de Bruxelles (ULB), Campus La Plaine, Bd du Triomphe, CP 210, Bruxelles, 1050, Belgium.
[5]Brussels Laboratory of the Universe - BLU, Belgium.

*Corresponding author. Email: mohammed.tawfeeq@ulb.be; jean-francois.flot@ulb.be;
Contributing authors: ursula.swaelus@skynet.be; florence.rodriguez@ulb.be; jenny.arrensdorff@gmail.com; felipe.ennes.silva@nyu.edu; laurent.grumiau@ulb.be; thomas.verdebout@ulb.be

Abstract

Feulgen image analysis densitometry (FIAD) has been extensively used for decades to measure genome sizes, but is usually considered less precise and less reproducible compared to other methods such as flow cytometry (FCM). However, FCM requires fresh tissue samples, which is a major impediment to using it on specimens collected in the field far from the lab. By contrast, FIAD works on ethanol-preserved samples kept at room temperature, which is a huge advantage for biodiversity genomics. In this study, we present refinements on the Feulgen protocol

and downstream analyses that produce measurements at least as reliable and precise as those obtained using whole-genome sequencing. We illustrate it by measuring for the first time the genome size of the red bald uakari monkey *Cacajao rubicundus* using the American cockroach *Periplaneta americana* and the black garden ant *Lasius niger* as standards.

Keywords: Feulgen Image Analysis Densitometry, C-value, DNA content, genome size

1 Introduction

Measuring genome size is a critical prerequisite for modern biological research, underpinning everything from strategic planning of genome sequencing projects to fundamental investigations of genome evolution. Genome sizes in eukaryotes are commonly reported as C-values, defined as half of the DNA content of a diploid somatic nucleus [1] and expressed in (pg) of DNA or in (Mb) or (Gb) of bases (1 pg = 978 Mb). C-values have profound biological implications: for instance, a larger genome size has been found to be positively correlated with an increased likelihood of extinction in both plants [2, 3] and vertebrates [4]. This has spurred the creation of comprehensive online databases to systematically catalog C-value data across various species [5].

The current state-of-the-art for measuring genome size is flow cytometry [6]. This technique is widely adopted for its speed and precision, providing reliable measurements with clear estimates of confidence. However, its utility is significantly hampered by practical constraints. The method is costly in terms of both equipment and reagents, and it requires large numbers of intact nuclei [7], which in many cases can only be obtained from fresh tissues. These requirements pose major logistical challenges for large-scale studies, particularly those involving samples collected from remote field locations. Other frequently used techniques, such as qPCR [8, 9], k-mer-based approaches [10] and genome assembly, also present drawbacks, such as the need for species-specific optimization for qPCR, lack of precision of k-mer based approaches, and assembler-specific variations in behavior when dealing with heterozygous or otherwise polymorphic regions [11].

Historically, one of the foundational methods for DNA quantification has been the Feulgen reaction, which was first proposed to stain DNA in cells and tissues by Robert Feulgen and Heinrich over a century ago [12] and later used to measure genome sizes [13]. The key advantage of Feulgen image analysis densitometry (FIAD) for genome size measurement is that it can be used with ethanol-preserved specimens, making it ideal for surveying specimens from field collections. However, FIAD is often considered as less precise and less quantitative than flow cytometry [14].

In the present study, we refined the classic FIAD approach to transform it into a cost-effective and accurate method, well-suited for surveying genome

sizes in large-scale studies. We used our optimized protocol to produce a robust genome size estimate for the red-faced uakari monkey (*Cacajao rubicundus*, a primate species endemic to the Brazilian Amazon Rainforest, with a distribution restricted to the middle Solimões River basin [15]), using the American cockroach *Periplaneta americana*(C-value = 3.41 pg ∼ 3,334 Mb) [16, 17] and the black garden ant *Lasius niger* (C-value = 0.30 pg ∼ 293.4 Mb [18]) as standards of known genome size. The C-value obtained from *C. rubicundus* using our method was subsequently validated using whole-genome sequencing and assembly.

2 Results

Our Feulgen protocol yielded for each of our three species (our *C. rubicundus* sample as well as our *L. niger* and *P. americana* standards) a large number of nuclei suitable for downstream analyses. Only few G2 nuclei were detected and discarded for *C. rubicundus* and *P. americana*, whereas they were far more numerous for *L. niger* (presumably before we used near-complete bodies for the ant vs. muscle and brain tissues for the primate and cockrach, respectively). The resulting Feulgen data passed all quality controls (Figs. 1 and 2). and yielded a C-value for *C. rubicunduns* of 2.67 pg (Fig. 3), corresponding to 2.61 Gb (with a CV of 8%). By comparison, k-mer analysis of the Nanopore reads using KAT yielded a genome size estimate of 2.66 Gb, whereas the size of the *de novo* assembly of the same sequence data using Flye was 2.69 Gb. Details of the genome assembly in terms of N50, BUSCO score, and k-mer completeness are presented in Table 1 and Figure 4 respectively.

Measure	**All scaffolds**
Flye assembly size (bp)	2,693,375,772
No. of scaffolds	3165
N50 (bp)	11,957,981
GC (%)	40.6%
Heterozygosity	0.12%
BUSCO Complete	99.5%
BUSCO Complete and Single	98.8%
BUSCO Complete and Duplicated	0.7%
BUSCO Fragmented	0.3%
BUSCO Missing	0.2%

Table 1. Assembly statistics for the *C. rubicundus* genome.

3 Discussion

The genome size of *C. rubicundus* was determined as 2.61 Gb using our refined Feulgen protocol vs. 2.66 using KAT and 2.69 Gb using genome assembly. These values show a very high level of agreement, with the KAT k-mer-based estimate being just 2% higher than the Feulgen estimate, and the genome

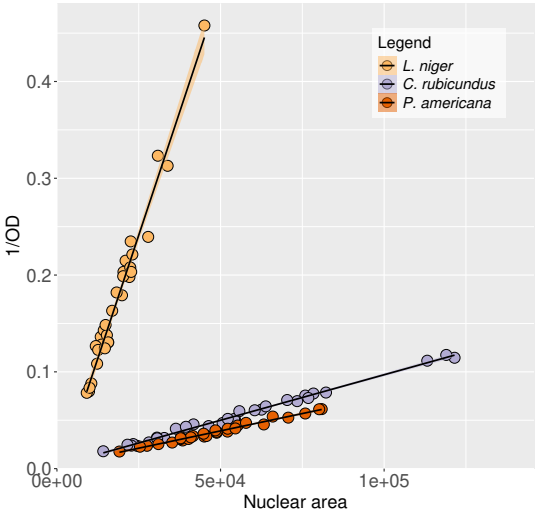

Fig. 1. Quality control confirming that the amount of coloration in the G1 nuclei of each species analysed is constant regardless of their area on the microscopy slides. The shaded bands around the regression lines indicate the 95% confidence intervals.

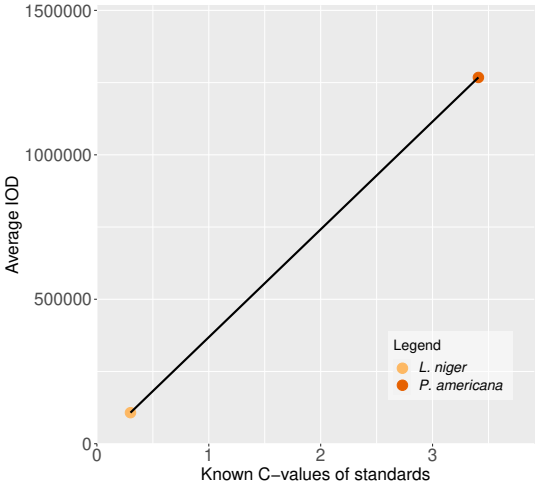

Fig. 2. Quality control confirming that the average amount of coloration of the nuclei for each standard species is proportional to their known C-value.

assembly size being 3% higher than the Feulgen estimate. These slightly higher values obtained from sequence-based analyses compared to Feulgen may reflect

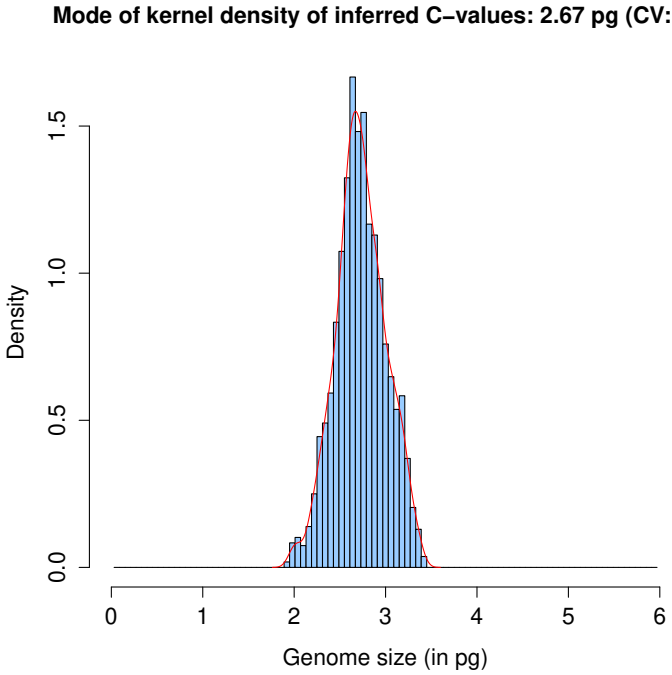

Fig. 3. Genome size histogram of *Cacajao rubicundus* with the corresponding Gaussian kernel density distribution (red curve).

some imprecision in the Feulgen method or might be caused by a small amount of contamination in the reads or (in the case of the assembly length) by a few residual uncollapsed haplotypes [19]. Future completion of the assembly of the *C. rubicundus* genome by scaffolding it to chromosome level using Hi-C [20] will allow us to detect and exclude both contaminating contigs and incompletely collapsed haplotypes, possibly yielding an assembly size even more congruent with our Feulgen-based estimate.

A key advantage of the Feulgen method is that it requires very few nuclei (we recommend 30, but a good estimate can already be obtained using as few as 15 nuclei). This is much less than flow cytometry, which typically requires 10^5 to 10^6 nuclei [21], hence the Feulgen method appears particularly suitable when working with small organisms or limited tissue samples. Furthermore, the Feulgen reaction is very suited for samples preserved in ethanol (in the case of fresh tissues, the first step in the protocol is actually to plunge them in ethanol).

Besides, flow cytometry is a technique that relies on sophisticated and expensive equipment not ubiquitously available in all laboratory settings. In contrast, the Feulgen method is extremely cost-effective, requiring only a standard microscope, a digital camera, and open-source software. This accessibility

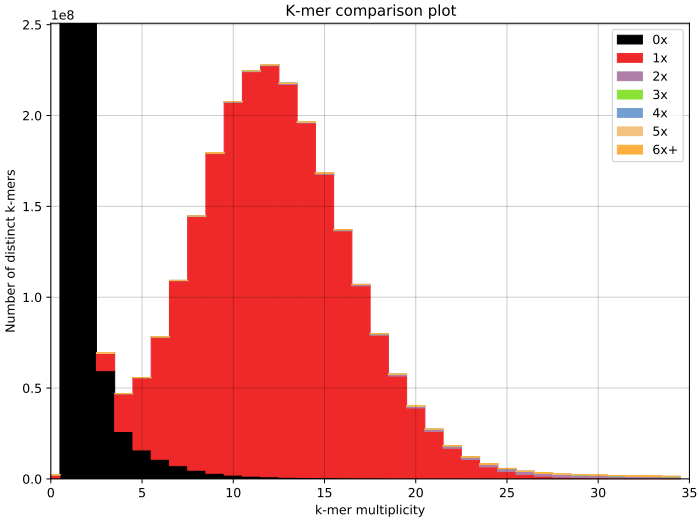

Fig. 4. KAT plot of *Cacajao rubicundus* comparing the k-mers in the Nanopore raw reads and in the assembly generated using Flye. Based on the k-mer histogram, KAT estimated the genome size as 2.66 Gb.

makes robust genome size estimation feasible for a broader range of research projects, particularly in laboratories with limited funding.

To conclude, the refined Feulgen method presented here is a robust, accurate, and economical protocol for estimating genome size. By avoiding the need for advanced equipment, large cell quantities, and exclusively fresh tissue, it represents a powerful and accessible alternative to flow cytometry. Another example of successful application of this method (where it yielded a result consistent both with k-mer spectrum genome size estimates and genome assembly size) was published recently [22]. While our methodology offers a significant cost advantage, its throughput is currently constrained by the time-consuming nature of manual image analysis. We are currently working on automating this process by developing a Python-based pipeline to streamline the entire workflow, from image processing through to the generation of genome size histograms.

4 Online Methods

Conducting Feulgen protocol in previous experiments presented inconsistencies ranging from the chemicals' preparation to the software and camera quality, leading to low reproducibility. To address these issues, we developed and optimized a robust protocol applicable to a diverse range of animal tissues. This refined procedure is presented below.

4.1 Sample Collection

We used a muscle tissue sample to estimate the genome size of a female red bald-headed uakari, *Cacajao rubicundus*. This sample was provided by the Mamirauá Institute for Sustainable Development (permit numbers: CITES 19BR033597/DF; SISBIO 55777).

Two species with well-characterized C-values were used as standards: the American cockroach, *Periplaneta americana* (C-value = 3.41 pg ∼ 3,334 Mb) [16, 17], and the black garden ant, *Lasius niger* (C-value = 0.30 pg ∼ 293.4 Mb [18]). Conveniently, these two species are reared at the Department of Organismal Biology of the Université libre de Bruxelles (ULB) for teaching and research purposes. We used the dissected brain of a female *P. americana* and three individual *L. niger* female worker ants (after removing their abdomens). The selection of female ant workers ensured the analysis of diploid somatic cells, as opposed to the haploid cells of male ants. Using brain tissues (for the cockroach) and avoiding abdomens (for the ant) was in order to avoid actively dividing tissues (such as the gut) that contain a mix of G1 and G2 cells. Tissues from all specimens were preserved in 96% ethanol after collection and until processing.

4.2 Reagents and Equipment

The essential laboratory equipment includes an incubator, a fume hood, and a digital balance. For image acquisition, a TOUPCAM™ FHD camera (TOUPTEK Photonics CO., LTD) was utilized, connected to a computer running ToupView software. Subsequent image analysis was performed using ImageJ, and statistical analyses were conducted in RStudio.

4.3 Slide preparation

A tissue sample, approximately 0.4 cm in length and 0.3 cm in width, was excised from the specimen using a sterilized razor blade or scissors. The excised tissue was placed directly onto a clean, dry microscope slide, which had been pre-cleaned with absolute alcohol.

Following placement on the slide, one to two drops of 40% acetic acid were added to the tissue. The tissue was then meticulously chopped into fine pieces directly on the slide with a sterilized razor blade. To prevent cell clumping and ensure a homogenous single-cell layer, additional drops of 40% acetic acid were applied, and the resulting cell suspension was carefully spread in a circular motion across the slide.

All prepared slides were left in a horizontal position at room temperature under the hood for 3 hours to allow for the complete evaporation of the acetic acid. Subsequently, the slides were fully immersed in 96% ethanol and allowed to air dry completely. Finally, the finished slides were stored in a dark, room-temperature environment for 48 hours before imaging and analysis to ensure optimal nuclei staining. The same detailed procedure was followed for the preparation of the standards.

4.4 Fixation

This step is crucial, as it dehydrates cells while preserving the tissue's structure in its original state, cross-linking nuclear proteins, and preventing DNA loss or degradation [14]. The fixation reagent was prepared by mixing 100% methanol, 37% formaldehyde and 90% acetic acid solutions in relative volumes (85:10:5). The prepared slides of standards and samples were placed in the slide rack then carefully immersed in a glass tank filled with the fixation reagent. The tank was sealed with parafilm and aluminum foil, then stored in a dark environment for 24 hours at 25 °C.

4.5 Washing with distilled water

The slide rack was carefully removed from the fixative tank and immersed in distilled water for 5 minutes. This was repeated a second time to remove residual chemicals.

4.6 Hydrolysis

To depurinate DNA and create free aldehyde groups [23], we immerged the slide rack in a fresh glass tank containing hydrochloric acid (HCl) at a 5M concentration. The container was sealed with parafilm, covered with aluminum foil, and stored in an incubator for 2 hours [24] at 25 °C. The rinsing procedure described in step 4.5 was then performed again.

4.7 Staining

Binding of the fuchsin on the free aldehyde groups formed on the DNA molecule during hydrolysis [23] was performed by placing the slide rack in a new glass tank containing Schiff reagent, then sealing the tank with parafilm and aluminum foil before incubating for 2 hours [14] in a dark setting at 25 °C.

4.8 Washing with sodium disulphite

A sodium disulphite solution was prepared to a volume of 200 ml by dissolving 1g of sodium disulphite in 10 ml of distilled water, adding 188 ml of distilled water, then finally adding 2 ml of HCl 5M. To remove and de-colorize unbound fuchsin stain, [14], the slides were placed in a tank filled with freshly prepared sodium disulphite buffer for 5 minutes, then a second time. Disulfite residues were then eliminated by performing two rinses in distilled water, as in section 4.5.

4.9 Dehydration

Finally, a few drops of 70% ethanol were added onto each processed slides and they left to dry for approximately 15 minutes until all the ethanol had evaporated. This was then repeated with 90% ethanol. The dehydrated slides were subsequently stored in a dark, dry environment till microscopy.

4.10 Microscopy

A few drops of immersion oil was applied onto each slide before covering them with a coverslip [14]. One or two drops of immersion oil were added on the coverslip itself, then a 100x oil-immersion objective was used to examine the slide on a KERN Optic microscope. The microscope settings, including maximum light intensity, the condenser's distance from the mechanical stage, and the iris diaphragm aperture, were maintained fixed to ensure consistent measurements across all standards and samples.

A 5-megapixel ToupCam digital color camera attached to the microscope was used to observe the slides; images were captured using the ToupView software (Windows version 4.11.19) at a resolution of 2560 × 1922 pixels.

4.11 Image analysis

Images were analyzed using ImageJ [25], an open-source tool for the analysis of scientific images developed by the National Institute of Health. We analyzed 30 nuclei per sample (clumped or overlapping nuclei were excluded from the analysis). Three important measures were taken: the area of the nucleus (A), the mean gray value of the nucleus (GVN), and the mean gray value of the background (GVB) (Fig. 5). From these we inferred the optical density (OD) of each nucleus as the difference between the mean gray value of the background and the mean gray value of the nucleus.

$$OD = GVB - GVN \tag{1}$$

This OD value can then be multiplied by the area of the nucleus (A) to obtain the integrated optical density (IOD), which is proportional to the genome size (C-value) of the specimen under investigation.

$$IOD = OD \times A \tag{2}$$

For the sake of consistency of analyses, we used only G1 nuclei for our downstream analyses and discarded G2 ones (that contain a double amount of DNA compared to G1 nuclei, resulting in an IOD approximately two times larger).

4.12 Data quality control

To control the quality of the measurements performed, we verified that they met the two basic assumptions of the Feulgen method. The first assumption is that nearly all nuclei of a given specimen contain the same amount of DNA, regardless of their area. This was tested by plotting 1/OD (the reverse of the optical density) vs. A (the nuclear area). The assumption is met for a given specimen if the resulting points fall on a line that passes nearby the origin of the plot, i.e., 1/OD is proportional to A, meaning that IOD = OD x A is constant for this specimen (Fig. 1).

The second assumption is that IOD is proportional to the genome size (C-value) of the species, i.e. the Schiff reagent colored DNA in a stoechiometric

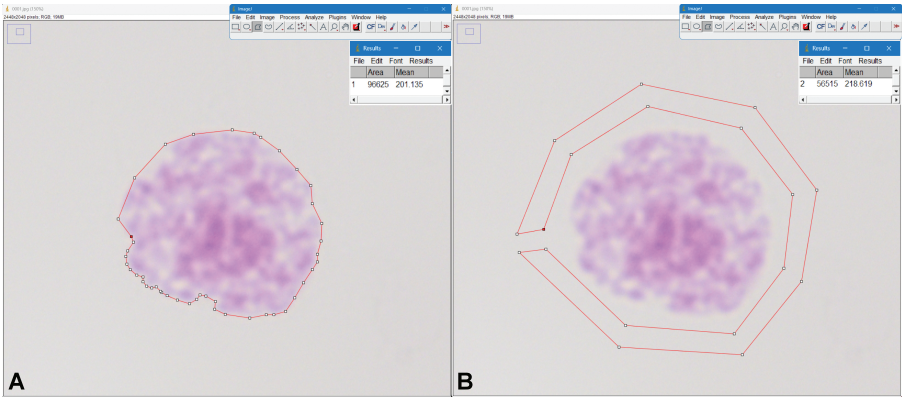

Fig. 5. Measurements of the nuclear area size and its light absorbance intensity from the same individual nucleus. **(A)** A nucleus boundary is outlined using ImageJ to extract the nucleus's area size and its mean gray value. **(B)** An external boundary of the background surrounding the nucleus is outlined to estimate the mean gray value of the background.

fashion. To verify this assumption, we checked that the IOD of the standards, when plotted against their known C-values, fell on a line that passed nearby the origin of the plot (Fig. 2).

4.13 Genome size calculation

Once verified that data passed both of the two quality controls detailed above, the C-value of the sample (C_{sample}) was estimated from its IOD (IOD_{sample}). Since the assumption is that IOD is proportional to C-value, we have

$$\frac{\text{Cvalue(sample)}}{\text{Cvalue(standard)}} = \frac{\text{IOD}_{\text{sample}}}{\text{IOD}_{\text{standard}}} \tag{3}$$

We calculated this C-value for each combination of one nucleus of the sample and one nucleus of the one of the standard: as there were 30 nuclei for our *C. rubicunduns* sample and 30 nuclei for each of our two standards, this resulted in a set of 1800 semi-independent estimates of the C-value of *C. rubicunduns*. These values were then visualized as a histogram superposed with the corresponding Gaussian kernel density, the mode of which yielded our final estimate of C-value of *C. rubicunduns* (expressed in pg of DNA) (Fig. 3). To estimate its standard deviation (σ), we measured the width of the Gaussian kernel density distribution at 0.882 time the height of the mode [26]. The coefficient of variation (CV_{sample}) was then calculated as the ratio of this standard deviation (σ_{sample}) to the modal C-value of the generated data matrix and expressed as a percentage (Fig. 3).

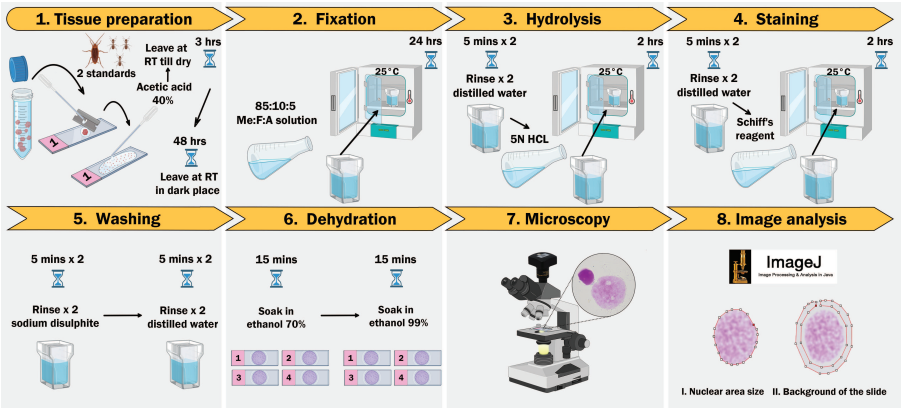

Fig. 6. The refined Feulgen Image Analysis Densitometry protocol. Created in BioRender. Flot, J. (2026) https://BioRender.com/oes4q0p.

4.14 DNA isolation, sequencing, and data processing

Genomic DNA was extracted from the muscle tissue of the same *C. rubicundus* specimen as for Feulgen, using Qiagen's MagAttract HMW DNA Kit (Ref: 67563). A ligation sequencing SQK-LSK114 Nanopore library was prepared according to the manufacturer's protocol and sequenced on a FLO-PRO114M flow cell with R10.4.1 pore proteins on a PromethION P2-solo machine (Oxford Nanopore Technologies, Oxford, UK). Sequencing was performed at the Evolutionary Biology and Ecology research unit, ULB (Brussels, Belgium). Nanopore base calling was performed using Guppy v3.2.10. The resulting reads were assembled de novo using Flye [27]. The quality and completeness of the assembly were assessed and calculated using K-mer Analysis Toolkit (KAT) v2.4.2 [28] and BUSCO v.5.8.3 with the odb12_primates dataset [29].

Acknowledgments

This study was supported by the Belgian Fonds de la Recherche Scientifique (F.R.S.-FNRS) via 'Projet de Recherches' grant T.0078.23 and "Action de Recherche Concertée" to JFF and via 'Chargé de Recherches' grant 40017464 to FES. We thank Clément Thijs and Stéphane Cherrier for providing the insects used as standards, and Olivier De Thier for useful discussions.

Table 1. Troubleshooting guidelines

Step	Problem	Possible reason	Solution
Microscopy	A pink smear of DNA is observed, without distinct nuclei.	The nuclei were degraded due to poor preservation, or the tissue was excessively chopped.	Use tissues recently preserved in 96% ethanol and keep in freezer till processing; do not chop the tissue too much (at most 5 mins) and avoid using a mortar and pestle.
Microscopy	Nuclei are overlapping or aggregated.	The cell suspension was not homogenized well.	Stir the suspension several times using the edge of the same razor blade that was used for chopping. Do this approximately every hour till the drop is completely dry.
Microscopy	Crystals and dirt overlap with the nuclei, reducing the quality of the pictures.	The slides were not clean enough.	Before using them, clean the slides with 96% ethanol using a laboratory wipe.
Image analysis	In the plot of 1/OD vs nuclear area size, the dots do not form a line passing through the origin.	The staining did not work well, or the nuclei are blurred (not in focus) or there is overlapping coloration from other nuclei.	Use only sharply defined nuclei well in focus and without overlapping coloration for measurements, or use only the largest nuclei (some camera may lack accuracy at measuring the optical density of very dense nuclei). If none of this solves the problem, consider repeating the Feulgen experiment.
Data analysis	The R codes fail with an error message complaining about missing columns.	Data were not exported correctly into the tab-delimited txt format expected by the R codes. For example, digits separated by dots in ImageJ become commas when pasted into Excel, which leads to value alteration.	Investigate the exported text files to make sure it is tab-delimited before using it in R.

Continued on next page

Table 1. Troubleshooting guidelines continued.

Step	Problem	Possible reason	Solution
Genome size calculation	The R code for genome size calculation returns a value several orders of magnitude larger than expected.	Differences in local settings between languages using a decimal point vs. a decimal comma.	Change the Excel settings to decimal dot instead of decimal comma and check the exported text files before running them in R.
Genome size calculation	The genome size is much lower than expected.	DNA degradation resulted in poor coloration.	Make sure the tissue samples were preserved in 96° ethanol and kept in a freezer; avoid using samples that are older than 5 years.
Genome size calculation	The genome size histogram displays several peaks corresponding to genome sizes that are multiple of each other.	G2 nuclei and/or polyploid nuclei resulting from endoduplication are present in the dataset.	Retain the peak with the smallest genome size, which is most likely the G1 non-polyploid nuclei. If ensure, redo the experiment avoiding tissues likely to contain dividing nuclei (e.g. gut cells) or polyploid nuclei (e.g. heart, and liver in mammals).
Genome size calculation	The genome size histogram displays several peaks corresponding to genome sizes that are not multiple of each other.	The inputted genome size of one of standards is not correct, or there are nuclei with different genome sizes in the sample.	Check whether the extra peaks disappear when using one standard at time. If so, then one standard has a problem - most probably, the known genome size that was used for calculation is not the correct not. If not, then the multiple peaks correspond to different populations of nuclei (e.g. host and symbionts), which ought better be distinguished by eye on the slides and analyzed separately.
Genome size calculation	Genome size is not accurate (when compared with genome size estimates obtained e.g. from k-mer spectrum analyses).	The standards and samples were not stained or photographed in a consistent way.	Because the Feulgen stain fades through time, standards and samples should be made in the same reagent bath and photographed at the same time.

Table 2. Estimated costs for key reagents and resources of this research.

Materials	Quantity	Purity	Source	Reference	Cost(€)	Website
Ethanol	1 L	99.50%	VWR®	8.18760.1000	64.20	Link-1
Methanol	2.5 L	99.80%	VWR®	20837.320	41.30	Link-2
Acetic acid	1 L	90%	VWR®	20109295	22.90	Link-3
Formaldehyde 37%	2.5 L	37%	Thermo Scientific™	A16163.0F	43.44	Link-4
Hydrochloric 5 N	2.5 L	100%	VWR®	30018320	46.10	Link-5
Schiff reagent	2.5 L	NA	Sigma-Aldrich®	1090332500	827.00	Link-6
Sodium disulphite	1 kg	NA	VWR®	27920295	43.90	Link-7
Slides	1×50	NA	VWR®	6311556	7.30	Link-8
Cover glasses	1×2000	NA	Corning®	CLS284518-2000EA	157.00	Link-9
Immersion oil	500 ml	NA	VWR®	A3494.0500	87.60	Link-10
Razor blades	1×10	NA	Swann-Morton®	9942	39.00	Link-11
Staining tray	2	NA	VWR®	BR474305	38.40	Link-12
Staining tank	1	NA	Marienfeld	6618005	97.00	Link-13
Tweezers	4	NA	Dutscher	53309	16.00	Link-14
Watchmaker forceps	1	NA	S Murray	12311379	23.85	Link-15
Becher 250 ml	1	NA	Glas-shop	102597	4.50	Link-16
Scissors	1	NA	S Murray & Co™	12368099	6.90	Link-17
Parafilm	100mm × 38m	NA	Merck	HS234526B-1EA	52.80	Link-18
Aluminum foil	1	NA	NA	NA	3.00	NA
				Total	**1479.39**	

Appendix A R codes

R script for generating Figure 1

```
# 1. Load libraries
library(ggplot2)
library(dplyr)
library(tcltk)

#2. Load data from separate files
data11 <- read.table(file.choose(), header = TRUE, sep = "\t")
data22 <- read.table(file.choose(), header = TRUE, sep = "\t")
data33 <- read.table(file.choose(), header = TRUE, sep = "\t")

#3. Label the data
data1 <- data11 %>% mutate(TrendSource = "Trend 1")
data2 <- data22 %>% mutate(TrendSource = "Trend 2")
data3 <- data33 %>% mutate(TrendSource = "Trend 3")
# Combine the data
combined_data <- bind_rows(data1, data2, data3)
combined_data$TrendSource <- as.factor(combined_data$TrendSource)

# 4. Plot with a custom legend
if (exists("combined_data") && all(c("Area", "OD_div", "TrendSource") %
    in% names(combined_data))) {
  base_text_size <- 11
 # Custom colors
  trend_colors <- c("Trend 1" = "#000000", "Trend 2" = "#000000", "
    Trend 3" = "#000000")
  trend_fills  <- c("Trend 1" = "#fdb863",  # Blue
                    "Trend 2" = "#B2ABD2",  # purple
                    "Trend 3" = "#e66101")  # red-orange

 # Species labels
  legend_labels <- c("Trend 1" = "L. niger",
                     "Trend 2" = "C. rubicundus",
                     "Trend 3" = "P. americana")
  gg_combined <- ggplot(combined_data, aes(x = Area, y = OD_div)) +

 # Data points shapes
    geom_point(data = subset(combined_data, TrendSource == "Trend 1"),
               aes(color = TrendSource, fill = TrendSource), size =
    4.5, shape = 21) +
    geom_point(data = subset(combined_data, TrendSource == "Trend 2"),
               aes(color = TrendSource, fill = TrendSource), size =
    4.5, shape = 21) +
    geom_point(data = subset(combined_data, TrendSource == "Trend 3"),
               aes(color = TrendSource, fill = TrendSource), size =
    4.5, shape = 21) +

```

```
  # Regression lines with confidence intervals
     geom_smooth(aes(group = TrendSource, color = TrendSource, fill = TrendSource),
                 method = "lm", formula = y ~ x,
                 linewidth = 0.75, alpha = 0.50, se = TRUE) +

  # Scales and labels
     scale_color_manual(values = trend_colors, labels = legend_labels) +
     scale_fill_manual(values = trend_fills, labels = legend_labels) +
     labs(
       x = "Nuclear area",
       y = "1/OD",
       title = "",
       color = "Legend",
       fill = "Legend"
     ) +
     coord_cartesian(
       xlim = c(0, max(combined_data$Area, na.rm = TRUE) * 1.2),
       ylim = c(0, max(combined_data$OD_div, na.rm = TRUE) * 1.05),
       expand = FALSE
     ) +
     theme_grey() +
     theme(
       axis.title = element_text(size = base_text_size * 1.5),
       axis.text = element_text(size = base_text_size * 1.5),
       legend.text = element_text(size = base_text_size * 1.3, face = "italic"),
       legend.title = element_text(size = base_text_size * 1.4),
       legend.position = c(0.80, 0.85),
       legend.background = element_rect(fill = alpha("white", 0.7), color = NA),
       legend.key = element_blank()
     ) +

  # Custom legend
     guides(
       color = guide_legend(
         override.aes = list(
           shape = 21,
           size = 4,
           fill = trend_fills,
           color = trend_colors)),
       fill = "none"
     )
   print(gg_combined)
} else {
   message("Error: 'combined_data' not found or is missing required columns (Area, OD_div, TrendSource). Please check data loading and preparation steps.")}
```

R code for generating Figure 2

```
#1. Load libraries
library(ggplot2)
library(dplyr)
library(tcltk)

#2. Load data
message("Select the TAB-separated file containing knowCv, IODaver, and Species columns")
Start_data <- read.table(tk_choose.files(caption = "Select TAB-separated data file"),
                         header = TRUE, sep = "\t", check.names = FALSE)

#3. Check, convert, and prepare data

required_cols <- c("knowCv", "IODaver", "Species") # Assuming these column names exist
if (!all(required_cols %in% names(Start_data))) {
  stop("Error: Input file must contain columns named 'knowCv', 'IODaver', and 'Species'. Please check file.")
}

# Convert relevant columns to numeric
Start_data$knowCv   <- suppressWarnings(as.numeric(as.character(Start_data$knowCv)))
Start_data$IODaver <- suppressWarnings(as.numeric(as.character(Start_data$IODaver)))

# Convert the species column to a factor
Start_data$Species <- as.factor(Start_data$Species)

# Remove rows with NA in plotting columns
plot_data <- Start_data %>% filter(!is.na(knowCv), !is.na(IODaver), !is.na(Species))

if(nrow(plot_data) == 0) {
  stop("Error: No valid data found in 'knowCv', 'IODaver', and 'Species' columns after conversion/filtering.")}

#4. Prepare plot parameters
min_x_val <- 0
max_x_val <- max(plot_data$knowCv, na.rm = TRUE)
min_y_val <- 0
max_y_val <- max(plot_data$IODaver, na.rm = TRUE)
base_text_size <- 11

#5. Create the ggplot
# Ensure data exists and has the required columns
```

```
if(exists("plot_data") && nrow(plot_data) > 0 && all(c("knowCv", "IODaver", "Species") %in% names(plot_data))) {

  gg_plot <- ggplot(plot_data, aes(x = knowCv, y = IODaver)) +

    # Add points, colored by the Species factor
    geom_point(aes(color = Species), shape = 19, size = 4.5) +

    # Add black regression line
    geom_smooth(method = "lm", formula = y ~ x, color = "black", se = FALSE, linewidth = 1) +

    # Define custom colors and labels
    scale_color_manual(
      name = "Legend",

      # Values to map colors to factor levels (ensure these match your data levels "1", "2")
      values = c("1" = "#fdb863",  # Color for P. americana (Level "1")
                 "2" = "#e66101"), # Color for L. niger (Level "2")

      # Legend labels
      labels = c("1" = expression(italic("L. niger")),
                 "2" = expression(italic("P. americana")))) +

    # Set labs and title
    labs(
      x = "Known C-values of standards",
      y = "Average IOD",
      title = "",
#IOD versus Known C-values
    color = "Species" # Define legend title (using scale now)) +

    # Set axis limits to start from zero
    coord_cartesian(
      xlim = c(min_x_val, max_x_val + 0.5),
      ylim = c(min_y_val, max_y_val + 250000),
      expand = FALSE
      ) +
    theme_grey() +
    theme(
        axis.title = element_text(size = base_text_size * 1.5),
        axis.text = element_text(size = base_text_size * 1.5),
        plot.title = element_text(size = base_text_size * 1.5, hjust =0.5),
        legend.title = element_text(size = base_text_size * 1.3),

        # Italic labels in the legend text
        legend.text = element_text(size = base_text_size * 1.3),
        legend.text.align = 0,
```

```

        #Position of the legend inside the plot
        legend.position = c(0.95, 0.1), # x=0.95 (near right), y=0.1 (near bottom) - Adjust
        legend.justification = c("right", "bottom"),
        legend.background = element_rect(fill = alpha("white", 0.5)) # Optional: semi-transparent background)

  # --- 6. Print the Plot ---
  print(gg_plot)}
  else {
   message("Error: 'plot_data' not found or has issues. Check previous steps.")}
```

R script for generating Figure 3 (genome size histogram)

```
library(tcltk)

STANDARDS <- read.csv(tk_choose.files(caption = "Select a one-column
    file with DOI/C ratios of nuclei of one or several standards"))
SAMPLES <- read.csv(tk_choose.files(caption = "Select a one-column file
     with DOIs of nuclei of one sample for which you want to estimate
    the C-value"))

nb_standards=length(STANDARDS[,1])
nb_samples=length(SAMPLES[,1])

t=0
Cvalues=rep(0,nb_standards*nb_samples)
for (i in 1:nb_standards){
  for (j in 1:nb_samples){
    t=t+1
    Cvalues[t]=SAMPLES[j,1]/STANDARDS[i,1]
  }
}

if (max(Cvalues) < 0.1) max_histo<-0.1
if (max(Cvalues) > 0.1) max_histo<-1
if (max(Cvalues) > 1) max_histo<-10
if (max(Cvalues) > 10) max_histo<-100

density_estimate<-density(Cvalues)
mode_value <- density_estimate$x[which.max(density_estimate$y)]
mode3SD <-signif(mode_value,3)

library(svDialogs)
max_histo <- dlgInput("Please enter below the maximum number of X axis
    to plot (or click OK to use the proposed default value):", default
    = max_histo)$res

max_histo <- as.numeric(max_histo)
left_density <- list("x" = density_estimate$x[1:which.max(
    density_estimate$y)], "y" = density_estimate$y[1:which.max(
    density_estimate$y)])
right_density <- list("x" = density_estimate$x[which.max(
    density_estimate$y):which.max(density_estimate$x)], "y" =
    density_estimate$y[which.max(density_estimate$y):which.max(
    density_estimate$x)])

cutoff <- 0.882*max(density_estimate$y)
left_cutoff <- left_density$x[which.min((left_density$y-cutoff)^2)]
right_cutoff <- right_density$x[which.min((right_density$y-cutoff)^2)]
sigma <- right_cutoff - left_cutoff

```

```
CV <- sigma/mode_value

CVpercent <- paste0(round(CV*100,0),"%)")

breakvalues=seq(max_histo*0.005,max_histo*0.995,by=max_histo/100)

histogram1 <-hist(Cvalues[Cvalues >= max_histo*0.005 & Cvalues <
    max_histo*0.995], xlab="Genome size (in pg)",ylab="Density", main=
    paste("Mode of kernel density of inferred C-values:", prettyNum(
    mode3SD),"pg (CV:", CVpercent),xlim=c(0,max_histo), prob=TRUE,
    breaks=breakvalues, cex.lab=1.3, cex.axis=1.3, cex.main=1.3, cex.
    sub=1.3)
col="#99CCFF"

lines(density_estimate, col = "red")

#Used colors: col="#FF3300","#99CCFF", "#33FF00"
```

References

[1] Greilhuber, J., Doležel, J., Lysak, M.A., Bennett, M.D.: The origin, evolution and proposed stabilization of the terms 'genome size'and 'C-value'to describe nuclear DNA contents. Annals of Botany **95**(1), 255–260 (2005). https://doi.org/10.1093/aob/mci019

[2] Vinogradov, A.E.: Selfish DNA is maladaptive: evidence from the plant Red List. Trends in Genetics **19**(11), 609–614 (2003). https://doi.org/10.1016/j.tig.2003.09.010

[3] Soto Gomez, M., Brown, M.J.M., Pironon, S., Bureš, P., Verde Arregoitia, L.D., Veselý, P., Elliott, T.L., Zedek, F., Pellicer, J., Forest, F., Nic Lughadha, E., Leitch, I.J.: Genome size is positively correlated with extinction risk in herbaceous angiosperms. New Phytologist **243**(6), 2470–2485 (2024). https://doi.org/10.1111/nph.19947

[4] Vinogradov, A.E.: Genome size and extinction risk in vertebrates. Proceedings of the Royal Society of London. Series B: Biological Sciences **271**(1549), 1701–1705 (2004). https://doi.org/10.1098/rspb.2004.2776

[5] Gregory, T.R., Nicol, J.A., Tamm, H., Kullman, B., Kullman, K., Leitch, I.J., Murray, B.G., Kapraun, D.F., Greilhuber, J., Bennett, M.D.: Eukaryotic genome size databases. Nucleic Acids Research **35**(suppl_1), 332–338 (2007). https://doi.org/10.1093/nar/gkl828

[6] Galbraith, D.W., Harkins, K.R., Maddox, J.M., Ayres, N.M., Sharma, D.P., Firoozabady, E.: Rapid flow cytometric analysis of the cell cycle in intact plant tissues. Science **220**(4601), 1049–1051 (1983). https://doi.org/10.1126/science.220.4601.1049

[7] Michaelson, M.J., Price, H.J., Ellison, J.R., Johnston, J.S.: Comparison of plant DNA contents determined by Feulgen microspectrophotometry and laser flow cytometry. American Journal of Botany **78**(2), 183–188 (1991). https://doi.org/10.1002/j.1537-2197.1991.tb15745.x

[8] Mounsey, K.E., Willis, C., Burgess, S.T., Holt, D.C., McCarthy, J., Fischer, K.: Quantitative PCR-based genome size estimation of the astigmatid mites *Sarcoptes scabiei*, *Psoroptes ovis* and *Dermatophagoides pteronyssinus*. Parasites & Vectors **5**(1), 3 (2012). https://doi.org/10.1186/1756-3305-5-3

[9] Kadota, M., Tatsumi, K., Yamaguchi, K., Yamaguchi, A., Asahida, T., Sato, K., Sakamoto, T., Uno, Y., Kuraku, S.: Shark and ray genome size estimation: methodological optimization for inclusive and controllable biodiversity genomics. F1000Research **12**, 1204 (2023). https://doi.org/10.12688/f1000research.136385.1

[10] Moeckel, C., Mareboina, M., Konnaris, M.A., Chan, C.S.Y., Mouratidis, I., Montgomery, A., Chantzi, N., Pavlopoulos, G.A., Georgakopoulos-Soares, I.: A survey of k-mer methods and applications in bioinformatics. Computational and Structural Biotechnology Journal **23**, 2289–2303 (2024). https://doi.org/10.1016/j.csbj.2024.05.025

[11] Kooij, P.W., Pellicer, J.: Genome size versus genome assemblies: are the genomes truly expanded in polyploid fungal symbionts? Genome Biology and Evolution **12**(12), 2384–2390 (2020). https://doi.org/10.1093/gbe/evaa217

[12] Feulgen, R., Rossenbeck, H.: Mikroskopisch-chemischer Nachweis einer Nucleinsäure vom Typus der Thymonucleinsäure und die-darauf beruhende elektive Färbung von Zellkernen in mikroskopischen Präparaten. **135**(5-6), 203–248 (1924). https://doi.org/10.1515/bchm2.1924.135.5-6.203

[13] Swift, H.: The constancy of desoxyribose nucleic acid in plant nuclei. Proceedings of the National Academy of Sciences **36**(11), 643–654 (1950). https://doi.org/10.1073/pnas.36.11.643

[14] Hardie, D.C., Gregory, T.R., Hebert, P.D.N.: From pixels to picograms: a beginners' guide to genome quantification by Feulgen image analysis densitometry. Journal of Histochemistry & Cytochemistry **50**(6), 735–749 (2002). https://doi.org/10.1177/002215540205000601

[15] Silva, F.E., Valsecchi do Amaral, J., Roos, C., Bowler, M., Röhe, F., Sampaio, R., Cora Janiak, M., Bertuol, F., Ismar Santana, M., de Souza Silva Júnior, J., Rylands, A.B., Gubili, C., Hrbek, T., McDevitt, A.D., Boubli, J.P.: Molecular phylogeny and systematics of bald uakaris, genus *Cacajao* Lesson, 1840 (Primates: Pitheciidae), with the description of a new species. Molecular Phylogenetics and Evolution **173**, 107509 (2022). https://doi.org/10.1016/j.ympev.2022.107509

[16] Hanrahan, S.J., Johnston, J.S.: New genome size estimates of 134 species of arthropods. Chromosome Research **19**(6), 809–823 (2011). https://doi.org/10.1007/s10577-011-9231-6

[17] Wang, L., Xiong, Q., Saelim, N., Wang, L., Nong, W., Wan, A.T.-Y., Shi, M., Liu, X., Cao, Q., Hui, J.H.L., Sookrung, N., Leung, T.-F., Tungtrongchitr, A., Tsui, S.K.W.: Genome assembly and annotation of *Periplaneta americana* reveal a comprehensive cockroach allergen profile. Allergy **78**(4), 1088–1103 (2023). https://doi.org/10.1111/all.15531

[18] Vizueta, J., Xiong, Z., Ding, G., Larsen, R.S., Ran, H., Gao, Q., Stiller, J., Dai, W., Jiang, W., Zhao, J., Guo, C., Zhang, X., Zuo, D., Zhong, W., Schiøtt, M., Liu, C., Zhang, H., Dai, X., Andreu, I., Shi, Y., Tretter, S.,

He, D., Gautam, S., Li, Z., Hickey, G., Ivens, A.B.F., Meurville, M.-P., Hita-Garcia, F., Kass, J.M., Guénard, B., Moreau, C., Paten, B., LeBoeuf, A.C., Economo, E.P., Chapuisat, M., Shik, J.Z., Ward, P.S., Heinze, J., Schultz, T.R., Li, Q., Dunn, R.R., Sanders, N.J., Liu, W., Schrader, L., Boomsma, J.J., Zhang, G.: Adaptive radiation and social evolution of the ant. Cell (1), 1–47 (2025). https://doi.org/10.1016/j.cell.2025.05.030

[19] Guiglielmoni, N., Houtain, A., Derzelle, A., Van Doninck, K., Flot, J.-F.: Overcoming uncollapsed haplotypes in long-read assemblies of non-model organisms. BMC Bioinformatics **22**(1), 303 (2021). https://doi.org/10.1186/s12859-021-04118-3

[20] Flot, J.-F., Marie-Nelly, H., Koszul, R.: Contact genomics: scaffolding and phasing (meta)genomes using chromosome 3D physical signatures. FEBS Letters **589**(20), 2966–2974 (2015). https://doi.org/10.1016/j.febslet.2015.04.034

[21] Givan, A.L.: Flow Cytometry: First Principles, 2nd edn., pp. 229–233 (2001). https://doi.org/10.1002/0471223948

[22] De Thier, O., Lebel, M., M.Tawfeeq, M., Faure, R., Dru, P., Blanchoud, S., Alié, A., Brown, F.D., Flot, J.-F., Tiozzo, S.: First chromosome-level genome assembly of the colonial chordate model *Botryllus schlosseri* (Tunicata). GigaScience **14**, 097 (2025). https://doi.org/10.1093/gigascience/giaf097

[23] Mello, M.L.S., Vidal, B.d.C.: The Feulgen reaction: a brief review and new perspectives. Acta Histochemica **119**(6), 603–609 (2017). https://doi.org/10.1016/j.acthis.2017.07.002

[24] Andersson, G.K.A., Kjellstrand, P.T.T.: Exposure and removal of stainable groups during Feulgen acid hydrolysis of fixed chromatin at different temperatures. Histochemie **27**(2), 165–172 (1971). https://doi.org/10.1007/BF00284958

[25] Schneider, C.A., Rasband, W.S., Eliceiri, K.W.: NIH Image to ImageJ: 25 years of image analysis. Nature Methods **9**(7), 671–675 (2012). https://doi.org/10.1038/nmeth.2089

[26] Bartlet, J.C., Smith, D.M.: The determination of the areas of resolved and partially resolved chromatography peaks. Canadian Journal of Chemistry **38**(11), 2057–2065 (1960). https://doi.org/10.1139/v60-280

[27] Kolmogorov, M., Yuan, J., Lin, Y., Pevzner, P.A.: Assembly of long, error-prone reads using repeat graphs **37**(5), 540–546 (2019). https://doi.org/10.1038/s41587-019-0072-8

[28] Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J., Clavijo, B.J.: KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies **33**(4), 574–576 (2016). https://doi.org/10.1093/bioinformatics/btw663

[29] Tegenfeldt, F., Kuznetsov, D., Manni, M., Berkeley, M., Zdobnov, E.M., Kriventseva, E.V.: OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. Nucleic Acids Research, 987 (2024). https://doi.org/10.1093/nar/gkae987