Estimating chromosome sizes from karyotype images enables validation of *de novo* assemblies.

Arne Ludwig[*,1,2], Alexandr Dibrov[*,1,2], Gene Myers[1,2], Martin Pippel[†,1,2]

May 22, 2022
Revision 1

*Equal contribution
†To whom correspondence should be addressed:

Martin Pippel
Max Planck Institute of Molecular Cell Biology and Genetics
Pfotenhauerstr. 108
01307 Dresden, Germany
Tel: +49 351 210-1972
Mail: pippel@mpi-cbg.de

Keywords: *de novo* assembly,karyotype,image processing

[1]Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstr. 108, 01307 Dresden, Germany
[2]Center for Systems Biology Dresden, Pfotenhauerstr. 108, 01307 Dresden, Germany

Abstract

Highly contiguous genome assemblies are essential for genomic research. Chromosome-scale assembly is feasible with the modern sequencing techniques in principle, but in practice, scaffolding errors frequently occur, leading to incorrect number and sizes of chromosomes. Relating the observed chromosome sizes from karyotype images to the generated assembly scaffolds offers a method for detecting these errors.

Here, we present KICS, a semi-automated approach for estimating relative chromosome sizes from karyotype images and their subsequent comparison to the corresponding assembly scaffolds. The method relies on threshold-based image segmentation and uses the computed areas of the chromosome-related connected components as a proxy for the actual chromosome size. We demonstrate the validity and practicality of our approach by applying it to karyotype images of humans and various amphibians, birds, fish, insects, mammals, and plants. We found a strong linear relationship between pixel counts and the DNA content of chromosomes. Averaging estimates from eight human karyotype images, KICS predicts most of the chromosome sizes within an error margin of just $6\,\mathrm{Mb}$.

Our method provides additional means of validating genome assemblies at low costs. An interactive implementation of KICS is available at `https://github.com/mpicbg-csbd/napari-kics`.

Introduction

Highly contiguous, complete, and accurate genome assemblies are fundamental to associating genotypes with phenotypes [19, 22]; genome-based evolution [19] and speciation studies [22]; analyzing repeat-organization and function [28]; population genetics [33]; and, ultimately, biomedical research [29, 47]. *De novo* assemblies can achieve chromosome-length scaffolds using third-generation long sequencing reads combined with additional sequencing data for scaffolding such as Bionano optical maps or Hi-C chromatin interaction maps [24, 40, 31].

However, the resulting assemblies often contain scaffolding errors that must be manually curated [18]. Knowing the true chromosome sizes in this step helps identify severe misjoins and incomplete scaffolds. An open-source tool to estimate chromosome sizes would assist in the *de novo* genome assemblies of unsequenced species like those targeted by the Vertebrate Genome Project [8], the Bird $10\,000$ Genomes Project [48], and the Earth BioGenome Project 2020 [27].

Karyotyping is a well-established technique in cytogenetics [13] using photomicrographs of complete chromosome sets. It has been practiced for more than a century [13]

producing karyotype images for hundreds of species. Such data distinctly renders the individual chromosomes' outlines and can thus be used to estimate their morphological properties. Commercial software packages like LUCIA Karyo [26] or Ikaros Karyotyping Platform [30] offer karyotype segmentation and subsequent analysis capabilities. We would argue that the scientific field would benefit from an open-source tool providing similar features.

Here, we present the karyotype image-based chromosome size estimator (KICS), a semi-automated method to estimate relative chromosome sizes from karyotype images. The open-source tool is implemented as a plugin for the general-purpose image viewer `napari` [9] and is available at `https://github.com/mpicbg-csbd/napari-kics`.

Results

Method Overview

The semi-automated method presented in this paper estimates relative chromosome sizes from karyotype images in four steps: (1) initial image segmentation by thresholding, (2) labeling of connected components, (3) manual curation of image labels, and (4) (semi-)automatic naming, ordering, and grouping of chromosomes. Optionally, the user may provide an estimate for the haploid genome size in order to derive absolute chromosome sizes. The results are available as an annotated image and in tabular format for further analysis. Figure 1 gives an overview of the process. We implemented the workflow as an extension to the general-purpose image viewer `napari` [9], which supplies functionality to load and annotate images.

Quality Analysis

To evaluate the accuracy and precision of KICS, we tested it on a set of eight human karyotypes, HUMAN8, which provides consistently high-quality images of 367 chromosomes and accurate reference chromosome sizes from the novel human telomere-to-telomere assembly [37]. We generated the estimates with a threshold of $\theta = 0.05$, blurring radius $\sigma_{\mathrm{B}} = 1$, and genome size $G = 3.1\,\mathrm{Gb}$. In manual curation, we only joined falsely separated pieces of chromosomes (fig. 1b) and removed noise-induced objects (fig. 1c). We named the chromosomes using the automatic method and renamed the sex chromosomes to X and Y, as indicated in the images. The dataset includes eight chromosomes with major translocations and deletions. These are ex-

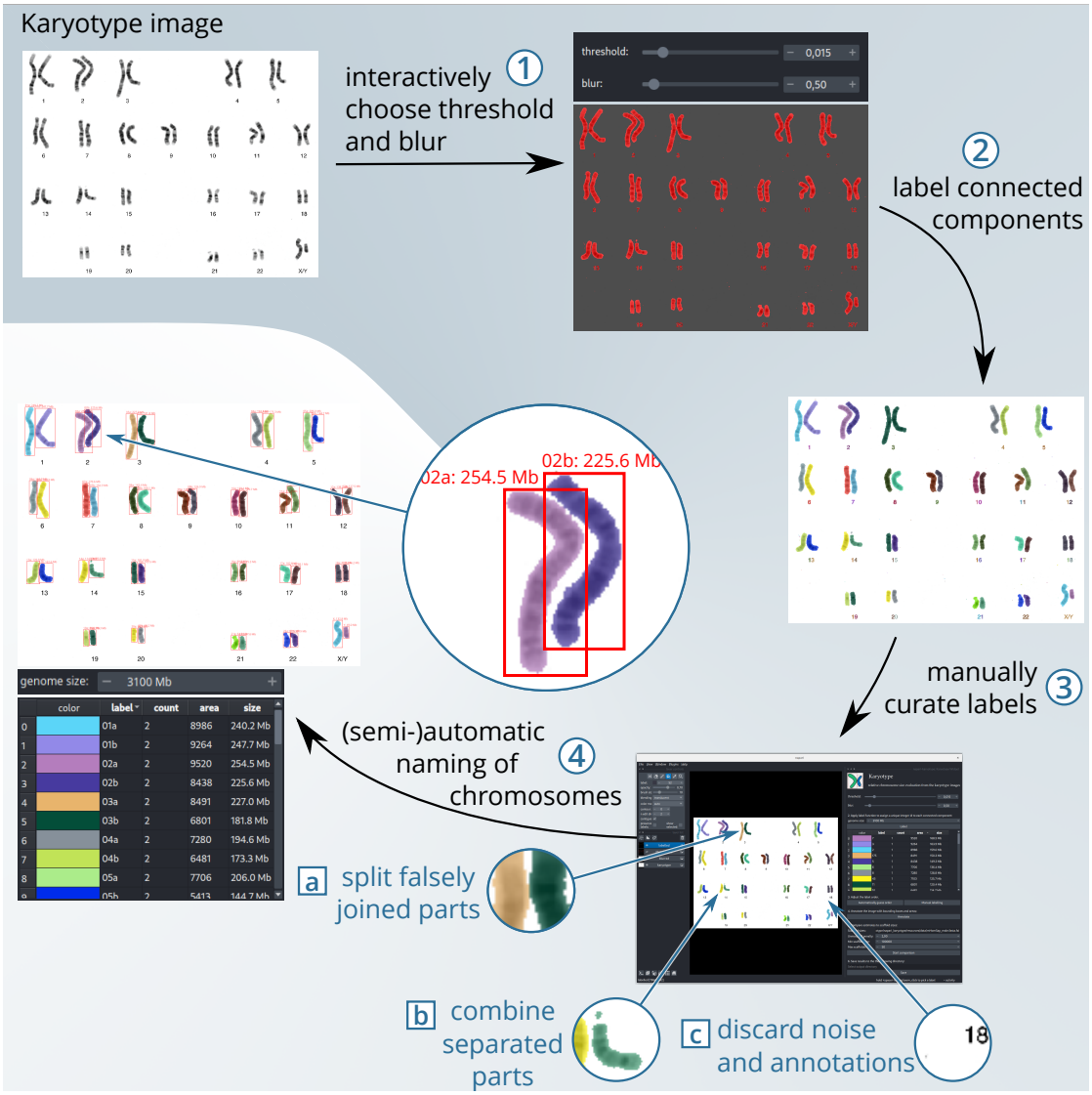

Figure 1: Overview of the workflow for estimating chromosome sizes from a karyotype image.

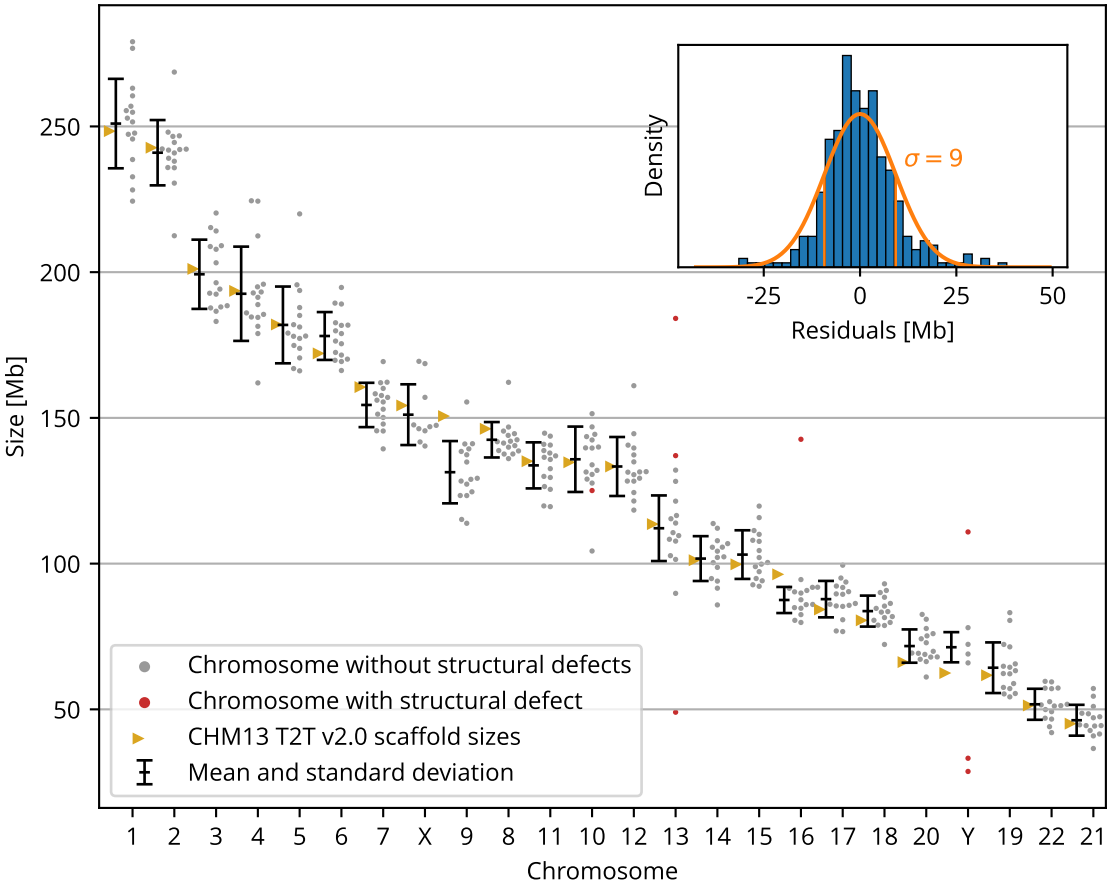

Figure 2: Estimated sizes of individual chromosomes with mean and standard deviation. Inset: distribution of residuals (histogram) and the inferred normal distribution (orange curve). Chromosomes with structural defects (red dots) are not included in the evaluation but highlight the potential of KICS to reveal unexpected deviations.

cluded from statistical analyses but highlighted in the plots demonstrating the potential of our method to reveal unexpected deviations. The resulting estimates are presented in figure 2.

In the following, we examine the accuracy and precision of KICS. While *accuracy* describes systematic errors in the estimates, *precision* is a measure of their statistical spread. We define the *accuracy* of the estimates as the absolute difference between their mean and the reference value; and their *precision* as their standard deviation.

KICS estimates chromosome sizes to a high degree of accuracy, mostly within $6\,\mathrm{Mb}$ of the reference size. On a per-chromosome basis, the method achieves an accuracy of $4.2\,\%$ or better relative to the true chromosome size. Exceptions of the above accuracy are chromosomes 9 ($-19\,\mathrm{Mb}/-13\,\%$), 16 ($-9\,\mathrm{Mb}/-9\,\%$), 20 ($-6\,\mathrm{Mb}/-8\,\%$), and Y ($+9\,\mathrm{Mb}/+14\,\%$). The variations of these estimates could be due to the satellite sequences in the centromeres [3]. Satellite sequence gets tightly packed and may therefore appear smaller in karyotype images. For example, chromosomes 9 and 16 are underestimated by KICS and contain about $20\,\%$ and $16\,\%$ of satellite sequence, respectively [3]. On the other hand, the length of chromosome Y is overestimated which might indicate missing sequence. Despite these exceptions, in general KICS is still able to estimate chromosome sizes from karyotype images with a high degree of accuracy.

To analyze the precision of the estimates, we consider the distribution of residuals across all estimates. The resulting histogram visually agrees with a normal distribution (fig. 2). Thus, for practical purposes, we derive a normal-distributed error with mean $\mu = 0$ and standard deviation $\sigma = 9\,\mathrm{Mb}$. Considering the precision of individual chromosomes, we found a pronounced correlation between chromosome size and precision (supp. fig. 2C), indicating a multiplicative error model that we discuss in depth below.

Because chromosome condensation generally differs between species [23, 10, 5], our method may not be equally valid for all organisms. Therefore, we evaluated it on a wide range of species covering amphibians (*A. mexicanum*, *X. laevis*), birds (*G. gallus*), insects (*D. melanogaster*), plants (*A. thaliana*, *Z. mays*), fish (*D. rerio*), and mammals (*R. norvegicus*, *H. sapiens*), by correlating estimated with reference chromosome sizes (fig. 3). For individual species, we found Pearson correlations from slightly significant (*A. thaliana*, $\rho = 0.77$) to highly significant (*A. mexicanum*, *D. melanogaster*, *G. gallus*, *H. sapiens*, *R. norvegicus*; $0.97 \leq \rho \leq 0.98$). We observed that the correlation coefficient ρ is mainly influenced by the quality of the karyotype images, which we go into more deeply in the discussion. Overall, the estimates and reference sizes agree across species.

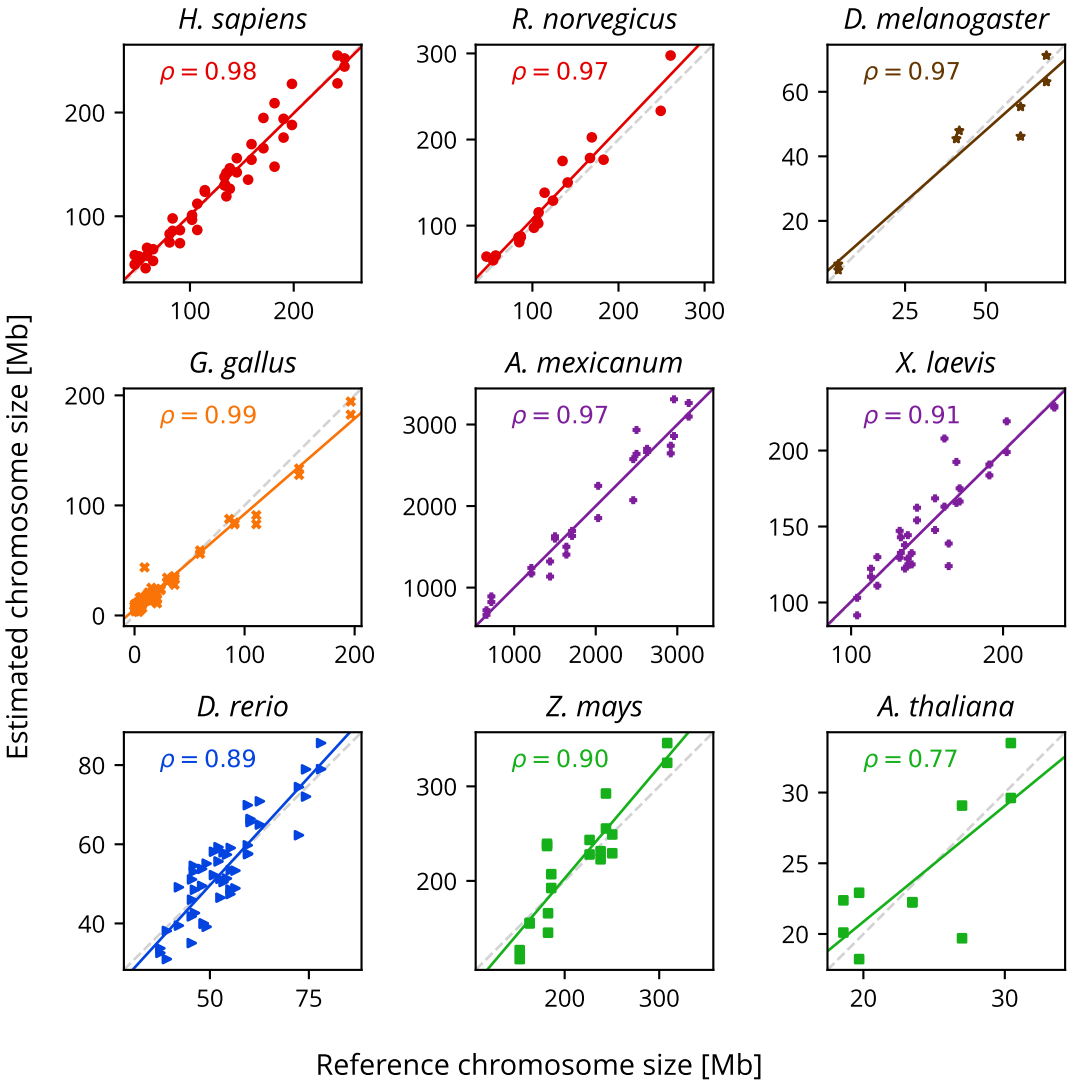

Figure 3: Reference versus estimated chromosome size for mammals (red circles), insects (brown stars), birds (orange crosses), amphibians (purple pluses), fish (blue triangles), and plants (green squares). Linear regression is depicted by a solid line and the identity function by a dashed line for reference. ρ shows the Pearson correlation coefficient.

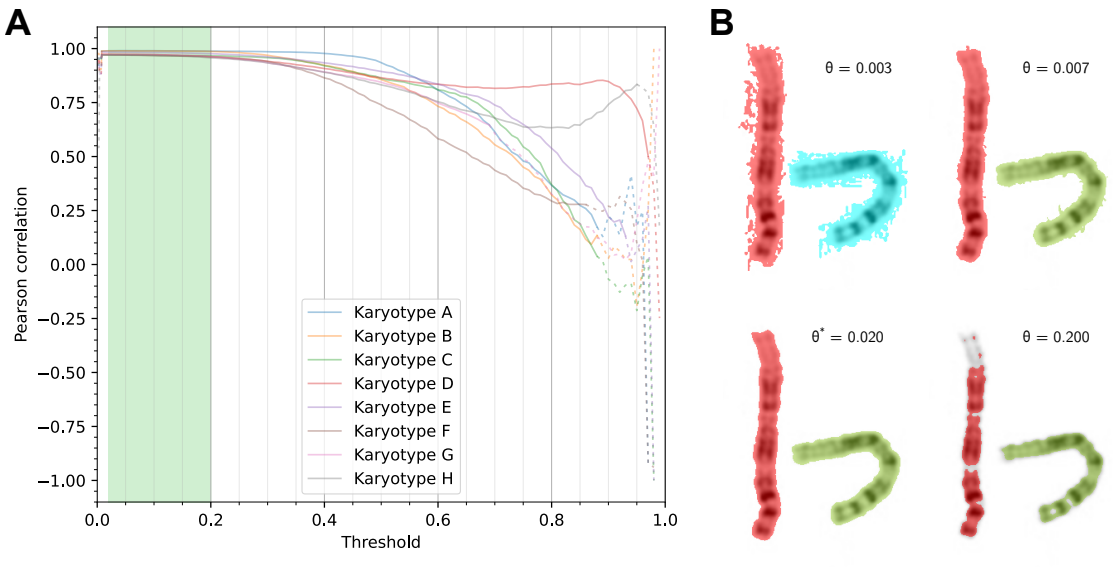

Figure 4: **A)** Pearson correlation of estimated and reference chromosome sizes at varying values of the threshold parameter θ for the HUMAN8 dataset. Dashed lines show segmentation into incorrect numbers of chromosomes. Highlighted in green is the near-optimal parameter space. **B)** Segmentation results of chromosome pair 1, karyotype H at varying thresholds θ.

Influence of Threshold

The thresholding value θ determines the segmentation and, consequently, may distort the chromosome size estimates. Therefore, we investigated the ease of choosing a near-optimal value and its robustness against perturbations. We used the Pearson correlation between the estimated and reference chromosome sizes as a measure of fit and evaluated it for different values of $0 \leq \theta < 1$ and $\sigma_{\mathrm{B}} = 1$ on the HUMAN8 dataset (fig. 4A). These results suggest an optimal threshold around $\theta^* = 0.02$ for all karyotypes and that any value $0.01 \leq \theta \leq 0.20$ (fig. 4) produces similarly good results. Without prior knowledge of this analysis, the manual choice $\theta = 0.05$ was close to the designated optimum θ^*. However, these results also suggest that the threshold is volatile towards zero. This agrees with the visual effects of low thresholds (fig. 4B), which the user easily detects. On the high end of the near-optimal interval, a threshold of $\theta = 0.20$ yields drastically impaired segmentation results and would be discarded by the user. Overall, the choice of the threshold is straightforward and robust.

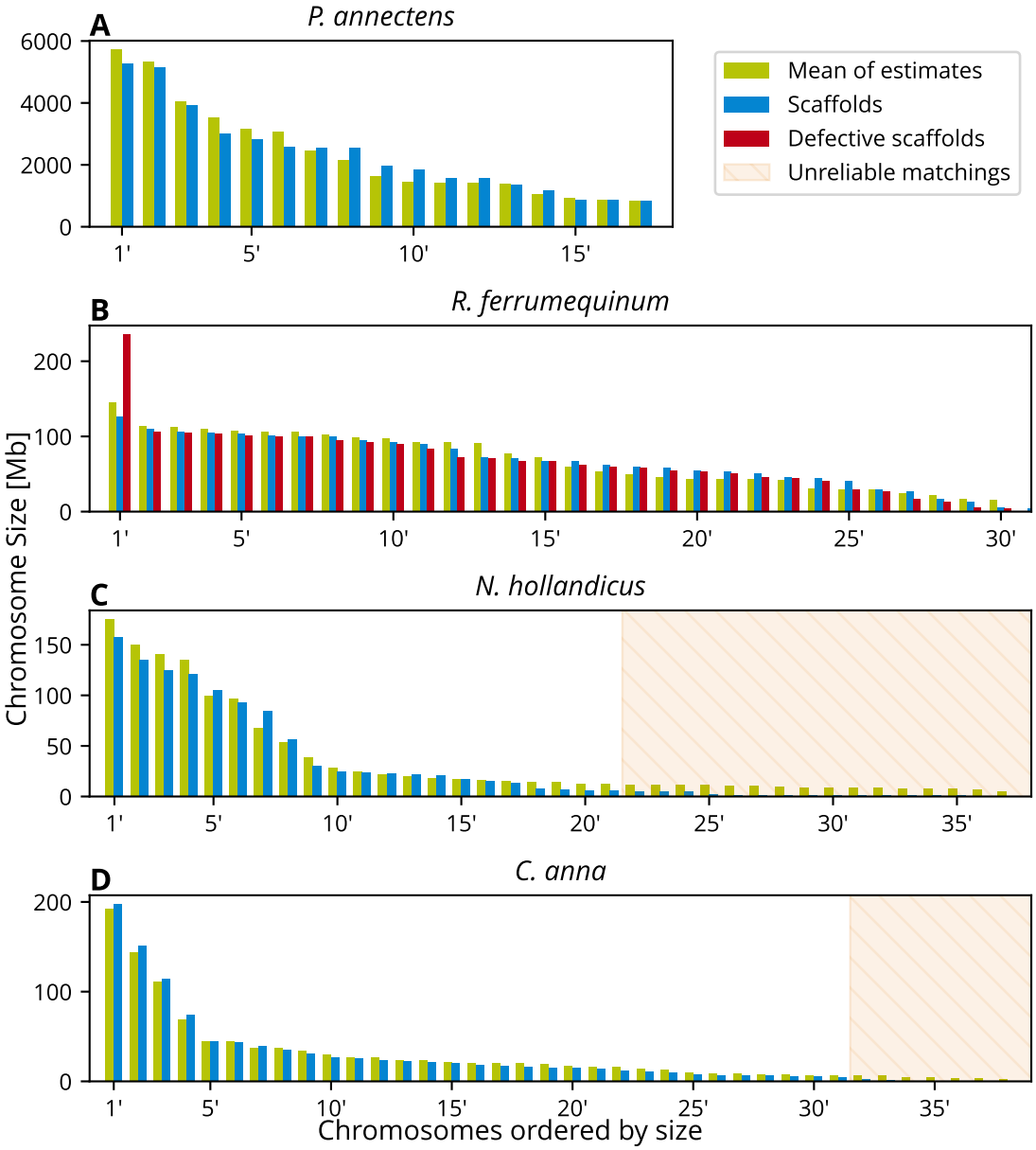

Figure 5: Comparison of chromosome size estimates and *de novo* assembly scaffold sizes without knowledge of correct matching. Both estimates and scaffold sizes are sorted independently by size before plotting. The hatched area highlights matches that are deemed unreliable, i. e. where the scaffolds have less than half the size of the corresponding estimate.

Examples

We wished to test KICS in the context of *de novo* genome assembly, where chromosome sizes are unknown *ab initio*. We calculated the estimates from single karyotype images and compared with manually curated chromosome-scale assemblies for four species – one fish, one bat and two birds. Because the true matching is unknown, we order the estimates and scaffold sizes independently by size from largest to smallest and match the numbers by rank.

The first example is the recent assembly of the West African lungfish (*Protopterus annectens*) which has an exceptionally large genome size of $40.5\,\mathrm{Gb}$ [46]. As can be seen in figure 5A, both number series have similar characteristics: (1) the two largest sizes are distinctly larger than the rest, (2) the three smallest sizes are very similar, and (3) the remaining twelve sizes have a roughly linear slope. Because there is no severe discrepancy between the number series, we cannot invalidate the scaffold sizes. In other words, the chromosome size estimates support the scaffold sizes.

In our second example, we show a defect that we encountered assembling the genome of the greater horseshoe bat (*Rhinolophus ferrumequinum*) published in [19]: two of the largest chromosomes were falsely joined at the telomeres by the Hi-C scaffolding procedure. In figure 5B, the defective scaffold sizes are shown in red, while the corrected sizes are shown in green. The size of the falsely joined scaffold distinctly stands out because it has roughly double the size of the largest chromosome estimate. After manual curation of the scaffolds, both number series neatly align except for the last pair, possibly indicating missing sequence in one of the smallest chromosomes.

The last two examples (fig. 5D and C) highlight challenges associated with microchromosomes. Microchromosomes alongside high diploid numbers of $2n \approx 80$ chromosomes are commonly encountered in birds [12], such as the cockatiel (*Nymphicus hollandicus*, $2n = 72$) and Anna's hummingbird (*Calypte anna*, $2n = 64$). The assembly of *N. hollandicus* is an internal work in progress, and about half of the microchromosomes appear to be incomplete or missing. On closer examination, the seventh chromosome estimate is noticeably larger than the corresponding scaffold size. This may have at least two reasons which cannot be distinguished without further data: (1) a false join between one of the macrochromosomes and one of the microchromosomes, or (2) underestimated chromosome size (see human chromosomes 9 and 16, fig. 2). This illustrates an intrinsic limitation of this method: minor errors are virtually undetectable. Generally, though, the sizes of microchromosomes are estimated without notable bias, as demonstrated by the last example, *C. anna*, where all but seven microchromosomes were successfully assembled.

Discussion

We have presented KICS, a novel semi-automated method for estimating relative chromosome sizes from a karyotype image. The method was developed with the primary goal of providing additional means of validating *de novo* genome assemblies. Our analysis based on eight karyotype images of human individuals shows that KICS accurately estimates most chromosome sizes within an error margin of just $6\,\mathrm{Mb}$ and precision of $9\,\mathrm{Mb}$. Further, we demonstrated that KICS performs well on a wide variety of species by applying it to karyotypes of amphibians, birds, fish, insects, mammals, and plants. We provided evidence that the methods' main parameter, the threshold, is straightforward to determine and robust against perturbations. Finally, we presented four practical examples demonstrating the power and limitations of KICS.

The quality of the karyotype image is the most influential variable of KICS (supp. fig. 1A). High image contrast and uniform background are vital to a good threshold-based segmentation. Low-quality images may still be used but require increased efforts for manual segmentation. Another important consideration is the copy number of the chromosomes in the karyotype image and the scaffolds; they should, e. g., contain the same set of sex chromosomes. For some species, this may be even more complex, e. g., in the freshwater planarian (*Schmidtea mediterranea*) a large translocation between two chromosomes determines whether they reproduce sexually or asexually [36]. Microchromosomes are especially hard to estimate because their size is closer to the image resolution and they easily get out of focus [20].

An alternative approach to estimating chromosome sizes from existing cytogenetic analyses is to use length estimates from the literature directly. For example, we found appropriate estimates for the chicken (*G. gallus*) in [21], table 1. We computed estimates for the chromosome size from the mean relative lengths and converted them to base pairs analogous to our method (supp. fig. 1D). We found that both methods yield similar results (supp. fig. 1C). Thus, if available, chromosome length estimates from previous studies may provide a viable alternative to our method.

In our experiments, we found a significant linear correlation between the apparent chromosome area and the DNA content. Seemingly, this stands in contrast to other results about chromosome scaling laws [23, 10, 5] which find non-linear power laws for the area. However, we analyze chromosome scaling in a single karyotype, whereas the above-mentioned studies examine chromosome scaling between different species and karyotypes. It is a classical result that chromosomes in a single metaphase spread have similar widths [13, 2, 4]. Assuming that the volume scales linearly with the DNA content, we also get a linear relationship between area and DNA content.

We proposed a simple additive error model $\tilde{x} = x + \delta$ for the estimates $\tilde{x}$ where δ is normal-distributed with mean $\mu = 0$. This model is certainly fit for practical purposes but does not accurately describe the observed error distribution. The first observation is that the errors are, in fact, *dependent* because the estimates are always relative sizes meaning that an error in one direction in one chromosome affects all other chromosomes in the opposite direction. Secondly, we observed that a multiplicative error model $\tilde{x} = \varepsilon x$ where ε is log-normal-distributed with expected value $\mathrm{E}[\varepsilon] = 1$ is more accurate than the additive error model. This is equivalent to an additive normal-distributed error model in logarithmic space. Calculating the Shapiro-Wilk test statistic for normality [44, 39], we get a better fit for this model ($W = 0.99$) compared to the additive error model ($W = 0.96$). In particular, while both models do not fully describe the error distribution, i. e. p is close to zero, the multiplicative error model has a higher significance $p = 3 \times 10^{-3}$ compared to $p = 1 \times 10^{-7}$ in the case of the additive model. Comparing panels A and B in supplementary figure 2 reveals a slightly left-skewed, i. e. right-leaning, error distribution in linear space, while it appears more symmetric in logarithmic space. However, looking at the resulting standard deviations per chromosome (supp. fig. 2C), we observe a tendency for bigger errors in smaller chromosomes, indicating a mixed additive-multiplicative error model $\tilde{x} = \varepsilon x + \delta$. We did not investigate this model analytically because it requires non-standard statistical methods to estimate the involved distribution parameters. However, this analysis still provides valuable insights: for relatively large chromosomes, the multiplicative error term dominates the overall error, whereas the additive error term dominates the errors for relatively small chromosomes. Presumably, the multiplicative error originates from subtle difference in condensation between chromosomes and the additive error from inaccuracies of the segmentation.

Materials and Methods

Data Sources

KICS generally requires at least a karyotype image as input and produces estimated relative chromosome sizes as output. To acquire karyotype images, we used web searches for the species name with additional keywords like "karyotype", "cytogenetics", "chromosomes", or "G-stained". Absolute chromosome sizes can be computed using a conversion factor derived from an estimate of the genome size such as the assembly size or estimates from a genome size database like the Animal Genome Size Database [15], the Fungal Genome Size Database [25], the Plant DNA C-values Database [38], or the Bird Chromosome Database [12]. Because these databases contain literature ref-

Group	Species	Karyotype image	Chromosome Sizes
Amphibians	*Ambystoma mexicanum*	[6, fig. 1]*	[45, fig. 1]
	Xenopus laevis	[42, fig. 1c]	GCA_017654675.1
Birds	*Gallus gallus*	[20, fig. 3]	GCA_016699485.1
Fish	*Danio rerio*	[11, fig. 1A]	GCA_000002035.4
Insects	*Drosophila melanogaster*	[17, fig. 1]	[1, fig. 1]
Mammals	*Homo sapiens*	[35]	GCA_000001405.28
	Rattus norvegicus	[16, fig. 1]	GCA_015227675.2
Plants	*Arabidopsis thaliana*	[14, fig. 1]	GCA_000001735.2
	Zea mays	[41, fig. 1A]	GCA_902167145.1

Table 1: Listing of data sources for the DIVERSITY dataset. Chromosome sizes were acquired from the assemblies referenced by their GenBank accession number if given. *Chromosomes were arranged according to [45, fig. 1].

Species	Karyotype image	Chromosome Sizes
Calypte Anna	[40, fig. 1g]	GCA_003957555.2
Nymphicus hollandicus	[34, fig. 4, all panels]	unpublished assembly
Protopterus annectens	[32, fig. 1c]	GCA_019279795.1
Rhinolophus ferrumequinum	[7, fig. 7]	GCA_004115265.3

Table 2: Listing of data sources for the examples. Chromosome sizes were acquired from the assemblies referenced by their GenBank accession number if given.

erences, they are also a good starting point for searching karyotype images.

To evaluate the performance of KICS, we compiled two datasets called HUMAN8 and DIVERSITY: The HUMAN8 dataset consists of 367 chromosomes from eight karyotype images (named karyotype A-H) of human individuals found in figure 1 of [43]. These karyotype images were acquired in clinical research and contain structural defects (translocations and deletions) in eight chromosomes. We highlighted these in our results for two reasons: first, they lead to outliers in the estimates that are independent of our method, and second, they present just the type of error that is detectable by our method. We used the scaffold sizes of the CHM13 T2T v2.0 assembly ([37], accession GCA_009914755.4) as reference chromosome sizes because it represents the human chromosomes from telomere to telomere. The DIVERSITY dataset covers species from amphibians, birds, fish, insects, mammals, and plants. It contains one karyotype image and one set of reference sizes for each species. Table 1 lists the species alongside the

data sources. The data sources for the examples presented in figure 5 are listed in table 2. The photomicrographs of the chromosomes of *A. mexicanum* shown in [45] could not be used for our method because they are compiled from several spreads.

Chromosome Size Estimation

Image Segmentation

The provided karyotype image is segmented in three phases (fig. 1.1). First, the image is converted to gray scale, where each pixel has a value in $[0, 1]$, where 0 is the darkest and 1 the brightest value. Second, the gray-scale image is blurred with a Gaussian kernel of user-adjustable size $\sigma_{\mathrm{B}} \geq 0$. Third, the blurred image $X = (x_{ij})$ is segmented by the threshold operation $x_{ij} < 1 - \theta$. We call θ the *threshold*. In our implementation, these operations are implicitly executed every time the user adjusts any of the parameters.

The user should start by selecting a threshold such that the segmentation agrees with the chromosomes. The blurring radius σ_{B} may be adjusted to improve the smoothness of the segmented areas, e. g., to compensate for jagged outlines. Areas that are segmented because of embedded annotations or other noise will be removed in the next step.

Labeling the Chromosomes

Once the basic segmentation is established, all 4-connected components are identified and labeled (fig 1.2). The background label is 0 and the foreground components are assigned labels $1, 2, 3, \ldots$. This label image is the basis for manual curation by the user. Typically, the user has to (1) assign different labels to falsely joined chromosomes (fig. 1A), (2) assign separate parts of chromosomes to the same label (fig. 1B), and (3) remove undesired labels, e. g. noise-induced segmentation or embedded labeling (fig. 1C). The core software `napari` provides "painting" tools (eraser, brush, fill bucket, pipette) for manipulating the label layer. Also, the user may easily remove small, noise-induced labels by selecting the n smallest labels in the table and pressing the backspace key.

Annotating the Chromosomes

With the image labels in place, the chromosomes can be meaningfully named (fig 1.4). Usually, chromosome names are mandatory because they carry information about the number of copies for each chromosome which is used to calculate their sizes correctly. Initial chromosomes names can be either generated automatically or by interactively striking them off in the desired order. The initial names can be manually curated using the interactive table.

The automatic labeling procedure tries to identify rows of chromosomes and in each row groups of chromosomes. The chromosome groups get numerical labels (01, 02, ...) starting in the top-left corner and proceeding in rows to the bottom-right corner. The chromosomes in each group are labeled from left to right with lowercase Latin letters (a, b, ...).

Annotated areas that overlap on the vertical image axis form the rows. Objects within a cutoff distance d^* grouped together, whereas further apart objects are placed in distinct groups. The cutoff distance d^* is determined for each row separately. Assume there are n objects in a row. Then the distances $d_1, \ldots, d_{n-1}$ are given as $d_i = l_{i+1} - r_i$ where l_i and r_i are the left- and right-most coordinates of object i, respectively. Without loss of generality, assume that d_i are sorted in non-descending order $d_1 \leq d_2 \leq \ldots \leq d_{n-1}$. Then the cutoff distance is given as $d^* = \max_{i=1,\ldots,n-2}\{d_i \mid 4d_i \leq d_{i+1}\}$ unless the set is empty, in which case $d^* = -\infty$, i. e., all objects are placed in separate groups. The set may be empty because either the row contains just a single object ($n = 1$) or none of the d_i satisfies the condition. The factor 4 in the above equation was determined empirically and will not work for every dataset. Also, the layout of karyotype images varies and may render this automatism useless.

In cases where the automatic naming fails, there is an interactive tool that determines the order and grouping by letting the user draw a path over each group of chromosomes. The chromosomes are then named according to the above naming scheme in chronological order and grouped if the same stroke marked them.

Finally, the names can be manually provided or altered in the interactive table. If the user enters a name in the same format as described above, it will also be interpreted in the same way. All other formats are interpreted as strings with no further meaning and chromosomes with the same names are taken to belong to the same group. This means that multiple occurrences of the same sex chromosome should get the same name, e. g. "X".

Estimation of Absolute Sizes

After correctly identifying all the chromosomes, the absolute size estimates can be computed. Suppose there are N distinct chromosomes $1, \ldots, N$ and each chromosome appears $c_i > 0$ times in the karyotype. Let A_{ij} be the area, i. e. the number of pixels, of the j-th annotated object ($j = 1, \ldots, c_i$) of chromosome i and G the user-provided estimate for the haploid genome size. Then we estimate the base pairs size of the annotated objects as $\tilde{x}_{ij} = \rho A_{ij}$, where $\rho = G / \sum_{ij} c_i^{-1} A_{ij}$ is the estimated DNA content per pixel.

Choosing the correct "genome size" may not always be straightforward. For example, the assembly in question may or may not contain both sex chromosomes, or chromosome names may be absent from a diploid karyotype impeding comparison to a haploid assembly. Generally, the "genome size" should be the number of base pairs present in the annotated image, taking into account that some chromosomes may occur multiple times. For example, a missing sex chromosome in the assembly can be compensated by skipping the annotation of that chromosome. An unannotated, diploid karyotype can be dealt with by artificially duplicating each scaffold.

Evaluation of Experimental Results

We evaluate the accuracy and precision of KICS by matching estimated sizes to known reference chromosome sizes from independent sources. While this makes a detailed per-estimate error analysis possible, the observed errors are always a mixture of estimation errors on one side and errors in the reference sizes on the other side. Hence, we do expect to observe differences between estimates and reference sizes.

In the HUMAN8 dataset, there are 13–16 estimates per non-defective chromosome (except for X with 10 and Y with 4 estimates). We report mean and standard deviation for each chromosome, assuming they have an additive and normal-distributed error. To assess the hypothesis of this error model, we evaluate the distribution of residuals by visually comparing it to the probability density function of a normal distribution with the parameters estimated from the samples.

Given a single karyotype image, we used the Pearson correlation as a measure of fit between the chromosome size estimates and corresponding reference sizes. We interpret values $\rho \leq 0.70$ as insignificant, $0.75 < \rho \leq 0.85$ as slightly significant, $0.85 < \rho \leq 0.90$ as moderately significant, $0.90 < \rho \leq 0.95$ as very significant, and $\rho > 0.95$ as highly significant. This interpretation is based on the observations we made in the course of this work.

Given estimates from a single karyotype image with unknown matching to the scaffold sizes, as typically encountered in applications of our method, we match the estimates sorted by size (largest first) with the largest scaffold sizes sorted equally but independently by size. We do not compute Pearson correlation for this data because the independent sorting introduces a strong correlation into every dataset. Instead, we evaluate the results visually by plotting the values side by side in a bar plot.

Declarations

Availability of Data and Materials

Karyotype images analyzed during this study are copyright protected and may be accessed through the original publications referenced in this published article. Remaining data generated or analyzed during this study are included in this published article and its supplementary information files.

Availability and Requirements

Project name:	KICS
Project home page:	`https://github.com/mpicbg-csbd/napari-kics`
Operating systems:	Platform independent
Programming language:	Python
Other requirements:	`napari` [9]
License:	MIT
RRID:	SCR_022224
biotools:	napari-kics

Competing Interests

The authors have no competing interests.

Funding

This work was supported by the Max Planck Society.

Abbreviations

KICS Karyotype image-based chromosome size estimator

Author Contributions

AL and AD implemented the software and analyzed the data. MP and GM conceived this study. MP applied and validated the software. GM supervised this study. AL wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgments

Not applicable.

References

[1] Mark D. Adams et al. "The Genome Sequence of Drosophila melanogaster." In: *Science* 287.5461 (Mar. 2000), pp. 2185–2195. DOI: 10.1126/science.287.5461.2185. eprint: https://www.science.org/doi/pdf/10.1126/science.287.5461.2185. URL: https://www.science.org/doi/abs/10.1126/science.287.5461.2185.

[2] S P Alexander and C L Rieder. "Chromosome motion during attachment to the vertebrate spindle: initial saltatory-like behavior of chromosomes and quantitative analysis of force production by nascent kinetochore fibers." In: *J Cell Biol* 113.4 (1991), pp. 805–15. DOI: 10.1083/jcb.113.4.805.

[3] Nicolas Altemose et al. "Complete genomic and epigenetic maps of human centromeres." In: *bioRxiv* (2021). DOI: 10.1101/2021.07.12.452052. eprint: https://www.biorxiv.org/content/early/2021/07/19/2021.07.12.452052.full.pdf. URL: https://www.biorxiv.org/content/early/2021/07/19/2021.07.12.452052.

[4] Daniel G Booth et al. "3D-CLEM Reveals that a Major Portion of Mitotic Chromosomes Is Not Chromatin." In: *Mol Cell* 64.4 (2016), pp. 790–802. DOI: 10.1016/j.molcel.2016.10.009.

[5] Sumitabha Brahmachari and John F Marko. "Chromosome disentanglement driven via optimal compaction of loop-extruded brush structures." In: *Proc Natl Acad Sci U S A* 116.50 (2019), pp. 24956–24965. DOI: 10.1073/pnas.1906355116.

[6] H G Callan. "Chromosomes and nucleoli of the axolotl, Ambystoma mexicanum." In: *J Cell Sci* 1.1 (1966), pp. 85–108. DOI: 10.1242/jcs.1.1.85.

[7] Ernesto Capanna and Maria Vittoria Civitelli. "Chromosomal Mechanisms in the Evolution of Chiropteran Karyotype Chromosomal Tables of Chiroptera." In: *Caryologia* 23.1 (1970), pp. 79–111. DOI: 10.1080/00087114.1970.10796365. eprint: https://doi.org/10.1080/00087114.1970.10796365. URL: https://doi.org/10.1080/00087114.1970.10796365.

[8] G10K Consortium. *Vertebrate Genomes Project*. URL: https://vertebrategenomesproject.org/ (visited on 01/31/2022).

[9] napari contributors. *napari: a multi-dimensional image viewer for python*. Version 0.4.14. 2019. DOI: 10.5281/zenodo.5975474.

[10] J R Daban. "Physical constraints in the condensation of eukaryotic chromosomes. Local concentration of DNA versus linear packing ratio in higher order chromatin structures." In: *Biochemistry* 39.14 (2000), pp. 3861–6. DOI: 10.1021/bi992628w.

[11] Rafael R. Daga, Guillermo Thode, and Angel Amores. "Chromosome complement, C-banding, Ag-NOR and replication banding in the zebrafishDanio reio." In: *Chromosome Research* 4.1 (1996), pp. 29–32. ISSN: 1573-6849. URL: https://doi.org/10.1007/BF02254941.

[12] Tiago M Degrandi et al. "Introducing the Bird Chromosome Database: An Overview of Cytogenetic Studies in Birds." In: *Cytogenet Genome Res* 160.4 (2020), pp. 199–205. DOI: 10.1159/000507768.

[13] L Doncaster. *An introduction to the study of cytology*. Ed. by University Press Cambridge. 1920. URL: https://archive.org/details/introductiontost00donc.

[14] Paul Fransz et al. "Cytogenetics for the model system Arabidopsis thaliana." In: *The Plant Journal* 13.6 (1998), pp. 867–876. DOI: https://doi.org/10.1046/j.1365-313X.1998.00086.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1365-313X.1998.00086.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-313X.1998.00086.x.

[15] T.R. Gregory. *Animal Genome Size Database*. 2022.

[16] M N Guilly et al. "Comparative karyotype of rat and mouse using bidirectional chromosome painting." In: *Chromosome Res* 7.3 (1999), pp. 213–21. DOI: 10.1023/a:1009251416856.

[17] Daryl S Henderson. "The chromosomes of Drosophila melanogaster." In: *Methods Mol Biol* 247 (2004), pp. 1–43. DOI: 10.1385/1-59259-665-7:1.

[18] Kerstin Howe et al. "Significantly improving the quality of genome assemblies through curation." In: *Gigascience* 10.1 (2021). DOI: 10.1093/gigascience/giaa153.

[19] David Jebb et al. "Six reference-quality genomes reveal evolution of bat adaptations." In: *Nature* 583.7817 (2020), pp. 578–584. DOI: 10.1038/s41586-020-2486-3.

[20] Meng Ji et al. "Cultivation and Biological Characterization of Chicken Primordial Germ Cells." In: *Brazilian Archives of Biology and Technology* v. 59.e16150374 (2016). Accessed 2022-03-18. DOI: 10.1590/1678-4324-2016150374.

[21] Alberto Juan Solari. "Ultrastructure of the synaptic autosomes and the ZW bivalent in chicken oocytes." In: *Chromosoma* 64.2 (1977), pp. 155–165. ISSN: 1432-0886. DOI: 10.1007/BF00327055. URL: https://doi.org/10.1007/BF00327055.

[22] Andreas F. Kautt et al. "Contrasting signatures of genomic divergence during sympatric speciation." In: *Nature* 588.7836 (2020), pp. 106–111. ISSN: 1476-4687. URL: https://doi.org/10.1038/s41586-020-2845-0.

[23] Eric M Kramer, P A Tayjasanant, and Bethan Cordone. "Scaling Laws for Mitotic Chromosomes." In: *Front Cell Dev Biol* 9 (2021), p. 684278. DOI: 10.3389/fcell.2021.684278.

[24] Zev N. Kronenberg et al. "Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C." In: *Nat Commun* 12.1 (2021), p. 1935. DOI: 10.1038/s41467-020-20536-y.

[25] Bellis Kullman, Heidi Tamm, and Kaur Kullman. *Fungal Genome Size Database*. 2005. URL: http://www.zbi.ee/fungal-genomesize/.

[26] Laboratory Imaging s.r.o. *LUCIA Karyo*. Accessed 2022-02-22. URL: https://www.lucia.cz/en/products/lucia_karyo.

[27] Harris A Lewin et al. "The Earth BioGenome Project 2020: Starting the clock." In: *Proc Natl Acad Sci U S A* 119.4 (2022). DOI: 10.1073/pnas.2115635118.

[28] Tsung-Yu Lu, and Mark J P Chaisson. "Profiling variable-number tandem repeat variation across populations using repeat-pangenome graphs." In: *Nat Commun* 12.1 (2021), p. 4250. DOI: 10.1038/s41467-021-24378-0.

[29] Amanda D Melin et al. "Variation in predicted COVID-19 risk among lemurs and lorises." In: *Am J Primatol* 83.6 (2021), e23255. DOI: 10.1002/ajp.23255.

[30] MetaSystems. *Ikaros Karyotyping Platform*. Accessed 2022-03-17. URL: https://metasystems-international.com/en/products/ikaros/.

[31] Karen H Miga et al. "Telomere-to-telomere assembly of a complete human X chromosome." In: *Nature* 585.7823 (2020), pp. 79–84. DOI: 10.1038/s41586-020-2547-7.

[32] Maria A. Morescalchi, Lucia Rocco, and Vincenzo Stingo. "Cytogenetic and molecular studies in a lungfish, Protopterus annectens (Osteichthyes, Dipnoi)." In: *Gene* 295.2 (2002). Papers presented at the 3rd Anton Dohrn Workshop 'Fish Genomics: Structural and Functional Aspects', Ischia, 1-2 June 2001 Giorgio Bernardi, Giacomo Bernardi (Organizers), pp. 279–287. ISSN: 0378-1119. DOI: https://doi.org/10.1016/S0378-1119(02)00755-2. URL: https://www.sciencedirect.com/science/article/pii/S0378111902007552.

[33] Phillip A. Morin et al. "Reference genome and demographic history of the most endangered marine mammal, the vaquita." In: *Mol Ecol Resour* 21.4 (2021), pp. 1008–1020. DOI: 10.1111/1755-0998.13284.

[34] I Nanda et al. "Chromosome repatterning in three representative parrots (Psittaciformes) inferred from comparative chromosome painting." In: *Cytogenet Genome Res* 117.1-4 (2007), pp. 43–53. DOI: 10.1159/000103164.

[35] National Human Genome Research Institute. *Talking Glossary of Genetic Terms: Karyotype*. Accessed 2021-12-18. URL: https://www.genome.gov/genetics-glossary/Karyotype.

[36] Philip A Newmark and Alejandro Sánchez Alvarado. "Not your father's planarian: a classic model enters the era of functional genomics." In: *Nat Rev Genet* 3.3 (2002), pp. 210–9. DOI: 10.1038/nrg759.

[37] Sergey Nurk et al. "The complete sequence of a human genome." In: *bioRxiv* (2021). DOI: 10.1101/2021.05.26.445798. eprint: https://www.biorxiv.org/content/early/2021/05/27/2021.05.26.445798.full.pdf. URL: https://www.biorxiv.org/content/early/2021/05/27/2021.05.26.445798.

[38] Jaume Pellicer and Ilia J Leitch. "The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies." In: *New Phytol* 226.2 (2020), pp. 301–305. DOI: 10.1111/nph.16261.

[39] Nornadiah Mohd Razali and Yap Bee Wah. "Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests." In: *Journal of Statistical Modeling and Analytics* 2.1 (2011), pp. 21–33.

[40] Arang Rhie et al. "Towards complete and error-free genome assemblies of all vertebrate species." In: *Nature* 592.7856 (2021), pp. 737–746. DOI: 10.1038/s41586-021-03451-0.

[41] Mt Sadder and G. Weber. "Karyotype of maize (Zea mays L.) mitotic metaphase chromosomes as revealed by fluorescencein situ hybridization (FISH) with cytogenetic DNA markers." In: *Plant Molecular Biology Reporter* 19.2 (2001), pp. 117–123. ISSN: 1572-9818. URL: https://doi.org/10.1007/BF02772153.

[42] Adam M Session et al. "Genome evolution in the allotetraploid frog Xenopus laevis." In: *Nature* 538.7625 (2016), pp. 336–343. DOI: 10.1038/nature19840.

[43] Jing Sha et al. "Chromosomal abnormalities and Y chromosome microdeletions in infertile men with azoospermia and oligozoospermia in Eastern China." In: *J Int Med Res* 48.4 (2020), p. 300060519896712. DOI: 10.1177/0300060519896712. URL: https://doi.org/10.1177/0300060519896712.

[44] S. S. Shapiro and M. B. Wilk. "An Analysis of Variance Test for Normality (Complete Samples)." In: *Biometrika* 52.3/4 (1965), pp. 591–611. ISSN: 00063444. URL: http://www.jstor.org/stable/2333709.

[45] Jeramiah J Smith et al. "A chromosome-scale assembly of the axolotl genome." In: *Genome Res* 29.2 (2019), pp. 317–324. DOI: 10.1101/gr.241901.118.

[46] Kun Wang et al. "African lungfish genome sheds light on the vertebrate water-to-land transition." In: *Cell* 184.5 (2021), 1362–1376.e18. DOI: 10.1016/j.cell.2021.01.047.

[47] Wesley C Warren et al. "Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility." In: *Science* 370.6523 (2020). DOI: 10.1126/science.abc6617.

[48] Guojie Zhang et al. "Genomics: Bird sequencing project takes off." In: *Nature* 522.7554 (2015), p. 34. DOI: 10.1038/522034d.