Received: 19 July 2021 | Revised: 29 November 2021 | Accepted: 30 November 2021

DOI: 10.1111/1755-0998.13570

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

ModEst: Accurate estimation of genome size from next generation sequencing data

Markus Pfenninger[1,2,3] | Philipp Schönnenbeck[3] | Tilman Schell[2]

[1]Senckenberg Biodiversity and Climate Research Centre, Frankfurt am Main, Germany

[2]LOEWE Centre for Translational Biodiversity Genomics, Senckenberg Biodiversity and Climate Research Centre, Frankfurt am Main, Germany

[3]Institute for Organismic and Molecular Evolution, Johannes Gutenberg University, Mainz, Germany

Correspondence
Markus Pfenninger, Senckenberg Biodiversity and Climate Research Centre, Frankfurt am Main, Germany.
Email: Markus.Pfenninger@senckenberg.de

Funding information
LOEWE Translational Biodiversity Genomics Centre

Abstract

Accurate estimates of genome sizes are important parameters for both theoretical and practical biodiversity genomics. Here we present a fast, easy-to-implement and accurate method to estimate genome size from the number of bases sequenced and the mean sequencing depth. To estimate the latter, we take advantage of the fact that an accurate estimation of the Poisson distribution parameter lambda is possible from truncated data, restricted to the part of the sequencing depth distribution representing the true underlying distribution. With simulations we show that reasonable genome size estimates can be gained even from low-coverage (10×), highly discontinuous genome drafts. Comparison of estimates from a wide range of taxa and sequencing strategies with flow cytometry estimates of the same individuals showed a very good fit and suggested that both methods yield comparable, interchangeable results.

KEYWORDS
Biodiversity genomics, comparative genomics, generation sequencing, genome trait, Next generation sequencing

1 | INTRODUCTION

Eukaryotic genomes vary tremendously in size (Bennett & Leitch, 2005; Carta et al., 2020; Kapusta et al., 2017; Oliver et al., 2007; Petrov, 2001), yet the underlying processes for this variability are not yet fully understood (Elliott & Gregory, 2015). To understand and study the mechanisms of genome size variation, such as proliferation of repetitive elements (Blommaert et al., 2019), effective population size (Lefébure et al., 2017; Lynch & Conery, 2003) or correlation to other traits (Gardner et al., 2020; Prokopowich et al., 2003), reliable estimates for the taxon under scrutiny are therefore mandatory. This is all the more important as substantial changes in genome size may even occur among closely related sister species, that is over relatively short evolutionary timescales (Agudo et al., 2019; Keyl, 1965; Vitales et al., 2020). An accurate estimation of genome size is also important for genomic projects. For example, in the assembly of genomes, the proportion of the true genome size covered by a given assembly draft is a quality criterion and limits the maximum size of the draft. In addition, resequencing projects requiring a certain sequencing depth (e.g., for genotyping) profit from a reliable genome size estimate (Fountain et al., 2016).

Flow cytometry is generally deemed to yield reliable estimates of genome size (Doležel & Greilhuber, 2010; Johnston et al., 2019). Yet, this method is not without caveats (Wang et al., 2015) and requires specialized laboratory skills and availability of the relatively expensive equipment. Moreover, the method depends on the availability of fresh or frozen tissue with largely intact cells, which narrows the range of taxa for which such analyses are practically feasible (Johnston et al., 2019).

Bioinformatical analysis of next generation sequencing (NGS) data provides an alternative for estimating genome size (Vurture et al., 2017). Besides the widely used k-mer-based methods (Li & Waterman, 2003; Lipovský et al., 2017), Schell et al. (2017) introduced a very simple method for genome size estimation, relying on mapping statistics

of NGS reads mapped back to a draft assembly. The approach assumes that the probability of sequencing a genome position is identical over the entire genome (i.e., that their true coverage is Poisson distributed). Even though there is a slight bias regarding the double strand breaking positions during DNA preparation for NGS sequencing, the impact on the resulting sequencing coverage distribution is negligible and thus this assumption is usually met (Poptsova et al., 2014). In a perfect assembly covering the entire genome, lambda as the parameter of the underlying Poisson distribution (as well as the mean) of the coverage distribution should therefore be identical to the true coverage. Dividing the number of sequenced, successfully back-mapped bases by the lambda of the observed coverage should yield a precise estimate of the true genome size. In most real draft genomes, however, repetitive regions are not resolved, which results in collapsed repeat regions, and in an assembly that is shorter than the true length (Treangen & Salzberg, 2012). These collapsed repeat regions are overproportionally covered, skewing the coverage distribution, and hence estimates of lambda upwards. A second source of systematic error in assemblies are relatively diverged heterozygous regions, such as from inversions that are not identified as homologous. These will result in a double representation of the respective region in the genome assembly, making it longer (Asalone et al., 2020). Consequently, the expected sequencing depth of these regions in the assembly will be half the true sequencing depth and skew the distribution and parameters estimated from it downwards. In real genome assemblies, both errors probably occur to various extents (Sohn & Nam, 2018), rendering a naïve use of parameters estimated from the observed coverage distribution misleading.

We show here how the observed sequencing depth distribution and an estimate of the number of bases sequenced from genome assembly drafts can be used to infer accurate estimates of genome size. We name the approach ModEst from Modal Estimation of genome size. We tested the methods with simulations, including various degrees of divergent heterozygous sites and a tetraploid genome, and compare genome size estimates from real data over a wide range of genome sizes with those derived from flow cytometry and k-mer-based methods.

2 | MATERIALS AND METHODS

2.1 | Theoretical background

Under the assumption that NGS sequencing methods sequence all bases in a genome with equal probability, dividing the number of bases sequenced (N) by the true length of the genome (L) yields the mean or expected sequencing depth (c) (Sims et al., 2014).

$$c = N/L$$

Since the sequencing depth distribution is discrete, it can be modelled by a Poisson distribution with parameter λ as c. As we are interested in L, we need to find reliable estimates for N and c from empirical data.

The number of bases used for the assembly of a particular genome is usually known. This number is, however, not necessarily identical to the number of bases sequenced from the target genome. Depending on the origin of the DNA, the data set may contain more or fewer reads originating from contaminations, the microbiome, and certainly reads from the mitochondrial or plastid genomes (Kumar et al., 2013). Even though several tools and pipelines exist to remove the bulk of such reads (Challis et al., 2020), this rarely succeeds completely. The number of bases after thorough cleaning, N_{clean}, therefore estimates rather the upper limit of N.

An alternative is the number of bases mapped back to the genome assembly draft N_{bm}. For this number to represent a good approximation of the number of bases sequenced from the corresponding genome, all genomic elements (telomers, centromers, repeats) must be represented in the assembly at least once without presence of contamination etc., and all reads must map back. This number is therefore a lower limit estimator of N.

As detailed in the Introduction, the empirical sequencing depth distribution of back-mapped reads is usually biased by errors in the genome draft due to collapsed repeats and/or other assembly errors. However, commonly at least a substantial part of the back-mapped reads map to unique sequences in the genome draft and should consequently show a coverage distribution following the true underlying Poisson distribution. Estimating λ from this part of the distribution that we know is not biased by assembly errors should therefore yield a reliable estimator of c. In Schell et al. (2017), the modal value of the empirical coverage distribution (m), that is the most often observed coverage, was used as an estimator of c. The modal value is a fairly good approximation of λ because the difference is in all cases ≤1 and therefore becomes relatively less biased when λ is high (i.e., high mean sequencing depth). Nevertheless, better methods for estimating λ from truncated Poisson distributions exist (Böhning & Schön, 2005; David & Johnson, 1952; Delignette-Muller & Dutang, 2015; Nadarajah & Kotz, 2006).

As mentioned above, the sequencing depth distribution may show more than a single peak. One possibility to obtain a bimodal distribution arises from highly divergent heterozygous tracts in the respective genome. In the assembly process, such divergent tracts may not be identified as homologous by the algorithm and thus occur as separate regions. Consequently, the sequencing coverage in such areas is only half the true coverage. If a considerable proportion of the genome consists of such divergent heterozygous regions, a second peak may appear in the sequencing coverage histogram. It has its maximum usually at half the sequencing depth of the larger peak. In this case, the peak with the larger sequencing coverage represents the true sequencing coverage. Except for recent hybrid individuals, the latter peak should nevertheless always be the higher one.

Another possibility to obtain a multimodal coverage distribution arises from polypoid species. If the multiplied genomes diverged to an extent that both are completely represented in the assembly, the genome size estimation process is no different from a diploid species. The other extreme would be a multiplied genome that is so little

TABLE 1 Simulated genomes and their characteristics

Simulated genome	Size (Mbp)	Average count of bases between rr	Average count of bases of rr	% of rr
1 *Saccharomyces cerevisae*-like	12	1246.68	156.67	5.26
2 *Caenorhabditis elegans*-like	100	508.66	166.42	13.23
3 *Arabidopsis thaliana*-like	120	622.32	311.55	18.06
4 *Drosophila melanogaster*-like	144	372.42	242.48	23.39
5 *Scophthalmus maximus*-like	524	521.84	45.64	3.74

Abbreviation: rr = repeat regions.

diverged that only a single copy appears in the assembly. In an intermediate stage, some more diverged parts of the multiplied genomes may be resolved, while others are collapsed in the assembly. The collapsed parts are expected to be over-covered and therefore the lowest peak represents the true sequencing depth.

In general, the observation of a multimodal coverage distribution of the back-mapped reads is indicative of issues with the assembly. Genome size estimation with the proposed ModEst method should nevertheless be possible, given appropriate caution.

2.2 | Practical approach

All the figures needed to estimate the genome size according to the method described here are usually collected in the process of genome assembly or can be easily calculated with standard tools (e.g., García-Alcalde et al., 2012). In particular, *samtools stats* and *bedtools genomecov* (Danacek et al., 2021; Quinlan & Hall, 2010) can be used for this purpose. The output of *samtools stats* provides information on bases sequenced and mapped, while the output of *bedtools genomecov* provides the empirical coverage distribution. The latter can be used as input for R. After preparing the data, we first estimated the modal value of the empirical distribution. This modal value is used as a starting point for a maximum-likelihood (ML) method to estimate λ from a truncated Poisson distribution as implemented in the R-libraries *truncdist* and *fitdistrplus* (Delignette-Muller & Dutang, 2015; Nadarajah & Kotz, 2006). We empirically determined suitable upper and lower truncation limits and give recommendations below. The respective R-code can be found in the Supporting Information, and a Perl wrapper-script, including all necessary dependencies, can be found at https://github.com/schellt/backmap.

2.3 | Simulations

To illustrate the influence of factors such as sequencing depth, genome size, repeat content and repeat distribution on the different genome size estimation methods, we simulated five different genomes according to real examples. Publicly available genome assemblies and annotations of *Saccharomyces cerevisae*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Drosophila melanogaster* and *Scophthalmus maximus* were used to obtain distributions of size and distance between annotated repeat regions. Simulated genomes of the size of the five genome assemblies mentioned above were then created using a custom Python-tool, available at https://github.com/Croxa/SimulateGenome. Regions annotated as repeat regions (rr) were filled with random repeat units up to 10 bp length, high complexity regions with random nucleotides. For sake of ease, we simulated the genomes on a single chromosome. A mean GC content of 0.5 was applied to both categories. Characteristics of the simulated genomes can be found in Table 1.

From these simulated genomes, we generated synthetic next-generation sequencing short read sets of 10×, 30× and 60× sequencing depth using ART ILLUMINA 2.5.8 (Huang et al., 2012). This tool emulates the sequencing process with built-in, technology-specific read error models, base quality value profiles parameterized empirically for large sequencing data sets and even adds the sequencing adapters. The reads were simulated paired-end, length of 150 bp with a standard deviation of 10 and an insert size of 300 bp. The Illumina sequencing system profile was HiSeq 2500 (HS25).

The read sets were trimmed with TRIMMOMATIC 0.39 (Bolger et al., 2014). Trimmed were usual Illumina adapters (ILLUMINACLIP:adapter.fa:2:30:10), leading and trailing bases with a quality score lower than 5, sliding windows with the size of 20 and an average quality score below 5, and reads with a length of 50 or lower.

In a first set of experiments, the trimmed read sets of different sequencing depths were back-mapped to the simulated genomes from which they were derived. Mapping was executed within the wrapper script backmap.pl using BWA MEM 0.7.17 without changing default options from backmap.pl. BWA (Burrows–Wheeler Aligner) is a widely used algorithm for mapping low-divergence sequences against a large reference genome (Li, 2013).

To estimate the influence of genome assemblies of varying quality on the accuracy of the genome size estimate, we assembled each read set with SPADES, the St. Petersburg genome assembler. This algorithm is implemented in a toolkit containing various assembly pipelines (Bankevich et al., 2012). SPADES 3.13.0 was used to assemble both trimmed paired and unpaired reads in a one-pass assembly using default options. The respective read sets were back-mapped and analysed as described above. For one simulation (*A. thaliana*-like,

10× coverage), we evaluated the effect of different truncation limits on the precision of the λ estimation. For sequencing depth class windows ranging from 11 to 5, centred on the modal value, the deviation of the ML estimate decreased from 0.4% to 4%. We performed the λ calculations therefore with a window size of 11 around the estimated modal value.

The influence of different amounts of diverged heterozygous genome stretches on size estimation was evaluated using the *S. cervisiae*-like genome. We simulated the genome with X, Y and Z% heterozygous stretches. To ensure that these stretches were not collapsed in the assembly process, we chose a sequence divergence of 10%. Likewise, we inferred the effect of polyploidy on genome size estimation with our method. We doubled the *S. cervisiae*-like genome and randomly changed bases in the complex part of one of the genomes. We simulated divergences of 0.5%, 1% and 5% among the two genomes. Both sets of simulations were performed as described above with 30× sequencing depth.

For all simulations, we calculated four different genome size estimates:

1. N_{clean}/λ, the number of "sequenced" bases after cleaning and trimming divided by the truncated Poisson ML λ estimate derived from the empirical sequencing depth distribution.
2. N_{clean}/m, the number of "sequenced" bases after cleaning and trimming divided by the modal value of the empirical sequencing depth distribution.
3. N_{bm}/λ, the number of back-mapped bases divided by the ML λ estimate derived from the empirical sequencing depth distribution.
4. N_{bm}/m, the number of back-mapped bases divided by the modal value of the sequencing depth distribution.

For each estimate, we calculated the relative deviation from the true known genome size.

2.4 | Empirical data

We used data from *de novo* genome assemblies that were sequenced in the last few years at the LOEWE Translational Biodiversity Genomics Centre and for which flow cytometry estimates from the same individual/clone/population were available. The taxonomic range of genomes comprised plants and several animal taxa with a focus on insects (Table 2).

If not stated otherwise in the citations, genome size estimates from flow cytometry were estimated following a protocol with propidium iodide (PI)-stained nuclei described in Hare and Johnston (2012). Tissue of the organism was chopped with a razor blade in a petri dish containing 2 ml of ice-cold Galbraith buffer. The suspension was filtered through a 42-µm nylon mesh and stained with the intercalating fluorochrome PI (Thermo Fisher Scientific) and treated with RNase II A (Sigma-Aldrich), each with a final concentration of 25 µg/ml. The mean red PI fluorescence signal of stained nuclei was quantified using a Beckman-Coulter CytoFLEX flow cytometer with a solid-state laser emitting at 488 nm. Fluorescence intensities of 5000 nuclei per sample were recorded. We used the software CYTEXPERT 2.3 for histogram analyses. The total quantity of DNA in the sample was calculated as the ratio of the mean red fluorescence signal of the 2C peak of the stained nuclei of the target organism divided by the mean fluorescence signal of the 2C peak of the reference standard times the 1C amount of DNA in the standard reference. Six replicates were measured on six different days to minimize possible random instrumental errors. We report the mean value of these measurements.

For each of the genomes, we calculated N_{bm}/m since we could not reconstruct the exact state of taxonomic read cleaning *i.e., removal of contamination reads from other taxa for all genomes). The modal value was chosen because the sequencing depth exceeded 50× in most cases. For comparison, we performed or used published k-mer-based estimates as far as available. First, a k-mer profile was generated from Illumina reads using JELLYFISH 2.3.0 tools (Marçais & Kingsford, 2011) count with a length of $k = 21$ and counting k-mers on both strands. Subsequently, the generation histogram was used as input for the GenomeScope webserver (Vurture et al., 2017) together with the above-mentioned length of k and read length. For some organisms, the approach could find no appropriate model. In addition, it is not suitable for long read technologies.

2.5 | Statistical analysis

The performance of the two bioinformatic genome size estimation methods was evaluated by their linear regression fit with the respective flow cytometry estimates. We compared the two slopes of the regression for statistical differences (Cohen et al., 2013).

3 | RESULTS

3.1 | Simulations

The single-pass assemblies derived from the simulated short reads were highly fragmented with thousands of short scaffolds, almost independent of simulated coverage (Table 3). For the *Saccharomyces cerevisae*-like, the *Caenorhabditis elegans*-like and the *Scophthalmus maximus*-like genomes, the total lengths of the assemblies were above 90% of the true size, while for the remaining two they were below 80%. This was reflected in the back-mapping rates that were highly correlated to the relative assembly length ($r = .995$, $p < .001$, Table 3).

The least relative deviation from the true genome size overall was found for the N_{clean}/λ estimator (mean deviation 0.00017, range 0.00003–0.00056), followed by N_{clean}/m (0.054, 0.0169–0.111), N_{bm}/m (0.094, 0.014–0.209) and N_{bm}/λ (0.112, 0.003–0.224, Figure 1). There was a tendency for the method to perform better with higher coverage, mainly due to the smaller relative deviation of m from λ at higher coverage. Given the rather minor differences in

TABLE 2 Genomes used for empirical evaluation

Species	Taxon	Flow cytometry estimate (Mb)	Back-mapping estimate (Mb)	k-mer-based estimate (Mb)	Sequencing technique	Citation
Hydropsyche tenuis	Insecta	260.6	228.6	222.8	Short read	Heckenhauer et al. (2019)
Plectrocnemia conspersa	Insecta	455.2	364.9	316.3	Short read	Heckenhauer et al. (2019)
Agapetus fuscipens	Insecta	721.8	583.5	463.2	Short read	Heckenhauer et al. (2021)
Odontocerum albicorne	Insecta	1616.0	1270.0	1103.4	Short read	Heckenhauer et al. (2021)
Drusus annulatus	Insecta	840.2	684.3	592.3	Short read	Heckenhauer et al. (2021)
Halesus radiatus	Insecta	1212.4	972.3	918.7	Short read	Heckenhauer et al. (2021)
Micropterna sequax	Insecta	1434.7	1100.0	981.7	Short read	Heckenhauer et al. (2021)
Micrasema longulum ML1	Insecta	663.6	707.7	650.7	Short read	Heckenhauer et al. (2021)
Micrasema longulum ML3	Insecta	663.6	637.8	635.2	Short read	Heckenhauer et al. (2021)
Micrasema minimum	Insecta	588.8	329.3	333.8	Short read	Heckenhauer et al. (2021)
Rhyacophila evoluta Rss1	Insecta	651.3	581.8	518.8	Short read	Heckenhauer et al. (2021)
Rhyacophila evoluta HR1	Insecta	651.3	565.5	514.4	Short read	Heckenhauer et al. (2021)
Glaux maritima (also known as *Lysimachia maritima*)	Angiosperm plant	1270.0	1541.4	1221.3	Short read	Segers et al. unpublished data
Radix auricularia	Mollusca	1575.0	1603.0	947.1	Short read	Schell et al. (2017)
Crematogastor levior	Insecta	455.0	356.0	255.9	Short read	Hartke et al. (2019)
Daphnia galeata	Crustacaea	155.0	157.0	150.5	Short read	Nickel et al. (2021)
Candidula unifasciata	Mollusca	1540.0	1420.0	977.6	Short read	Chueca, Kochmann, et al. (2021)
Styela plicata	Tunicata	430.9	468.6	338.8	Short read	Galià-Camps et al. unpublished data
Callionymus lyra	Teleostei	645.0	653.2	562.0	Short read	Winter et al. (2020)
Pimpla turbinella	Insecta	300.0	298.0	206.0	Short read	Reumont et al. unpublished data
Fagus sylvatica	Angiosperm plant	582.4	542.0	541.0	Short read	Mishra et al. (2021)
Aedes japonicus	Insecta	857.0	836.3	699.0	Short read	Reuss et al. unpublished data
Nyctereutes procynoides	Mammalia	3100.0	3230.0	—	Long read	Chueca et al. (2021)
Microthlaspi erraticum	Angiosperm plant	194.5	211.0	—	Short read	Mishra et al. (2020)
Crematogastor levior, species B	Insecta	390.0	406.7	—	Long read	Feldmeyer et al. unpublished data
Camponotus femoratus	Insecta	330.0	340.0	—	Long read	Feldmeyer et al. unpublished data
Astacus astacus	Crustacaea	16,891.0	16,750.0	—	Short read	Theissinger et al. unpublished data
Lamprophis fuliginosis	Squamata	1480.0	1617.0	—	Long read	Hiller et al. unpublished data
Desmodus rotundus	Mammalia	2337	2089	—	Long read	Hiller et al. unpublished data

contiguity among genome assemblies reconstructed from different coverages, this factor had only a minor role on the precision of the genome size estimates (Table 3; Figure 1).

The genome size estimates from simulated genomes with varying proportions of divergent heterozygous sites all yielded the same estimates (Table S1). As can be seen in the respective coverage distributions, the only difference between the simulations was a second, lower peak at about half the expected coverage that grew with increasing amounts of heterozygous regions. The position of the true peak remained unaffected (Figure 2a).

TABLE 3 Characteristics of simulated genomes, their assemblies, back-mapping and estimation of the parameter of the underlying Poisson distribution

Simulated genome	Simulated coverage	True size (Mbp)	Assembly size (Mbp)	Proportion of true length	Number of contigs	Mean contig length (bp)	Mbp "sequenced" = N_{clean}	bp mapped = N_m	Proportion of bases mapped
Saccharomyces_cerevisae-like	10	12.08	11.24	0.930	2021	5561	120.8	113.0	0.931
	30	12.08	11.24	0.931	1823	6170	362.4	339.1	0.932
	60	12.08	11.24	0.932	1906	5905	724.8E+08	677.7	0.931
Caenorhabditis elegans-like	10	100.0	92.77	0.927	105,104	883	100.0E+09	933.7	0.904
	30	100.0	9280	0.928	98,341	944	300.1E+09	2807	0.910
	60	100.0	92.25	0.929	99,957	930	600.2E+09	5598	0.904
Arabidopsis thaliana-like	10	120.1	92.83	0.773	66,881	1388	1201	956.6	0.780
	30	120.1	93.03	0.775	63,695	1461	3602	2881	0.785
	60	120.1	92.75	0.772	61,615	1505	7205	5759	0.784
Drosophila melanogaster-like	10	144.1	107.8	0.748	104,002	1037	1441	1118	0.755
	30	144.1	107.7	0.747	95,701	1125	4322	3382	0.763
	60	144.1	107.6	0.747	95,523	1127	8643	6725	0.756
Scophthalmus maximus-like	10	524.1	523.4	0.999	76,507	6842	5241	5220	0.994
	30	524.1	524.1	1.000	63,360	8261	15720	15670	0.995
	60	524.1	425.1	1.000	63,260	8274	31440	31340	0.995
Saccharomyces cerevisae-like 1% divergent heterozygous regions	30	12.08	11.78	0.975	1152	10,226	356.1	352.2	0.989
5% divergent heterozygous regions	30	12.08	12.18	1.008	3020	4032	354.0	350.5	0.990
10% divergent heterozygous regions	30	12.08	12.68	1.049	5435	2333	351.4	347.1	0.988
10% divergent heterozygous regions	30	12.08	13.68	1.133	10,163	1346	346.1	341.9	0.988
Tetraploid *Saccharomyces cerevisae*-like 0.5% divergence among duplicated genomes	30	24.16	11.75	0.486	1197	9818	712.9	700.4	0.982
1% divergence	30	24.16	13.09	0.542	6699	1954	713.0	696.8	0.977
5% divergence	30	24.16	22.92	0.949	4719	4857	714.4	700.3	0.980

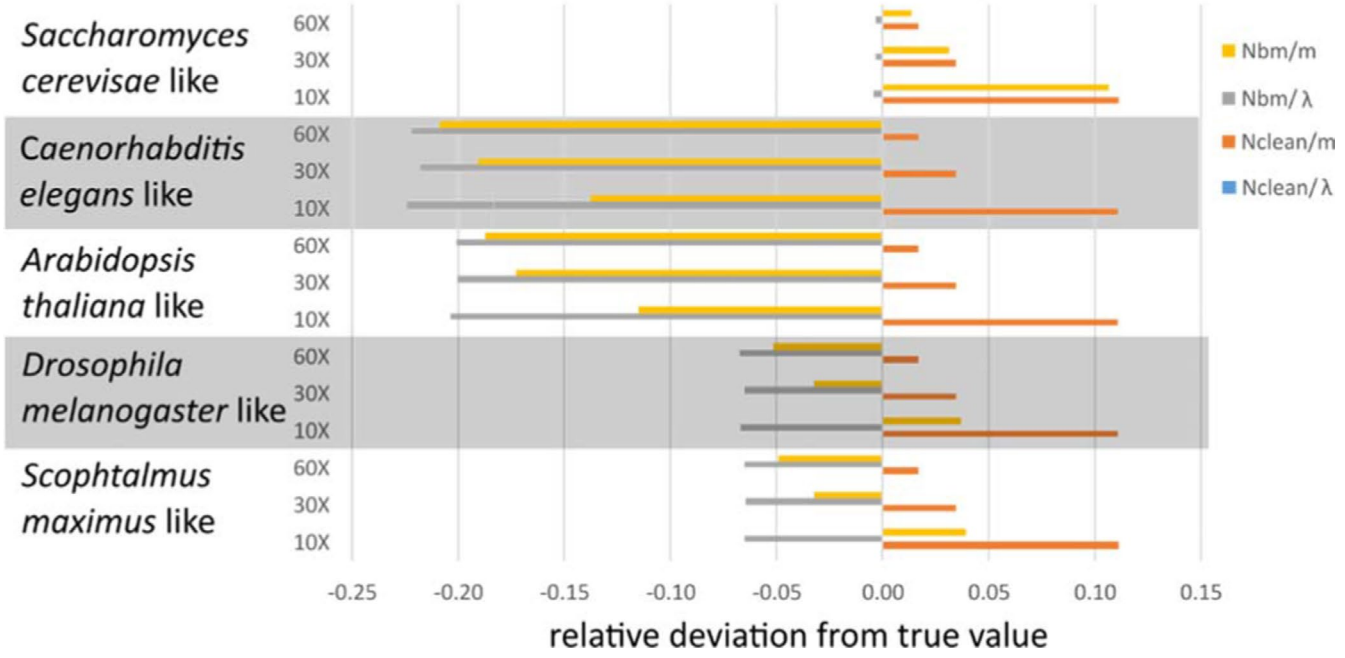

FIGURE 1 Relative deviations of genome size estimators from true values for different simulated genomes and simulated coverages. The deviations of N_{clean}/λ (blue) from the true value are so small that they are not visible on the scale. The raw data table to this figure can be found in Table S1

Assembly of a tetraploid *Saccharomyces cerevisae*-like genome with the two lowest divergences between the duplicated genomes (0.5% and 1%) resulted in the reconstruction of approximately a single haploid genome, respectively (assemblies of lengths 1.18 and 1.31 Mb, Table S1). Therefore, the highest observed coverages for these simulations were both 59 and the λ estimates close to 60 (Table S1; Figure 2b). Consequently, the genome size estimates were close to the haploid length. However, with divergence of 1%, a second peak with maximum 28 and λ of 28.9 emerged (Figure 2b; Table S1). Using this peak yielded estimates that were much closer to the truth (relative deviations between 0.005 and 0.06, depending on the estimator). With 5% divergence, the duplicated genomes were almost fully resolved in the assembly and, hence, the peak at the true coverage and therefore the genome size estimates were not further than 0.03 from the truth (Figure 2b; Table S1).

3.2 | Empirical data

Ordinary Least Squares Regression of 1C flow cytometry estimates against the estimates derived from our sequencing depth approach yielded an excellent fit ($r^2 = .998$, $p = 2.2 \times 10^{-34}$). Removing the outlier estimate for the crayfish genome did not change the result markedly ($r^2 = .958$, $p = 2.1 \times 10^{-17}$). The estimated slope was with 0.996 ± 0.043 (*SE*), very close to unity. The fit of the respective k-mer-based estimates to the flow cytometry data was equally good ($r^2 = .996$, $p = 1.1 \times 10^{-26}$), but the slope of 0.585 ± 0.007 (*SE*) suggested a systematically lower k-mer estimate (Figure 3). The estimated slopes were significantly different from each other ($t = 9.43$, $df = 44$, $p < 1 \times 10^{-6}$).

4 | DISCUSSION

As long as reliable whole chromosome sequencing remains technically not feasible and thus the true genome size is not known, genome size estimation of de novo sequenced genomes will be a necessary and important part of biodiversity genomics. We present here with ModEst a fast, easy-to-use and precise method for genome size estimation from NGS data. We have shown that the method works for a wide range of genome sizes. The method could become standard part of the genome assembly process, because it relies on data that are routinely collected. Although our method is not the first to propose the use of sequencing, and respectively mapping statistics (Pflug et al., 2020; Pucker, 2019), it requires fewer assumptions and much less bioinformatic effort than previously suggested approaches. The method does, admittedly, not solve the problem of how much sequence information should be produced in the first place if there is absolutely no a priori information on the expected genome size of the target organism. However, very low modal sequencing depths obtained with the method indicate that sequencing efforts should be increased.

To evaluate the performance of our method and the factors influencing it, we performed a simulation study. We simulated five different genomes with the characteristics and genome sizes typical for various eukaryotic taxa. We were able to show that the precision of the estimate is largely independent of the contiguity and quality of the underlying genome assembly as long as most sequence elements in the genome are represented in the assembly draft. This finding was confirmed with the empirical samples, where, for example, the size estimate for the very large genome of the crayfish *Astacus astacus* was gained from a preliminary, highly discontinuous assembly

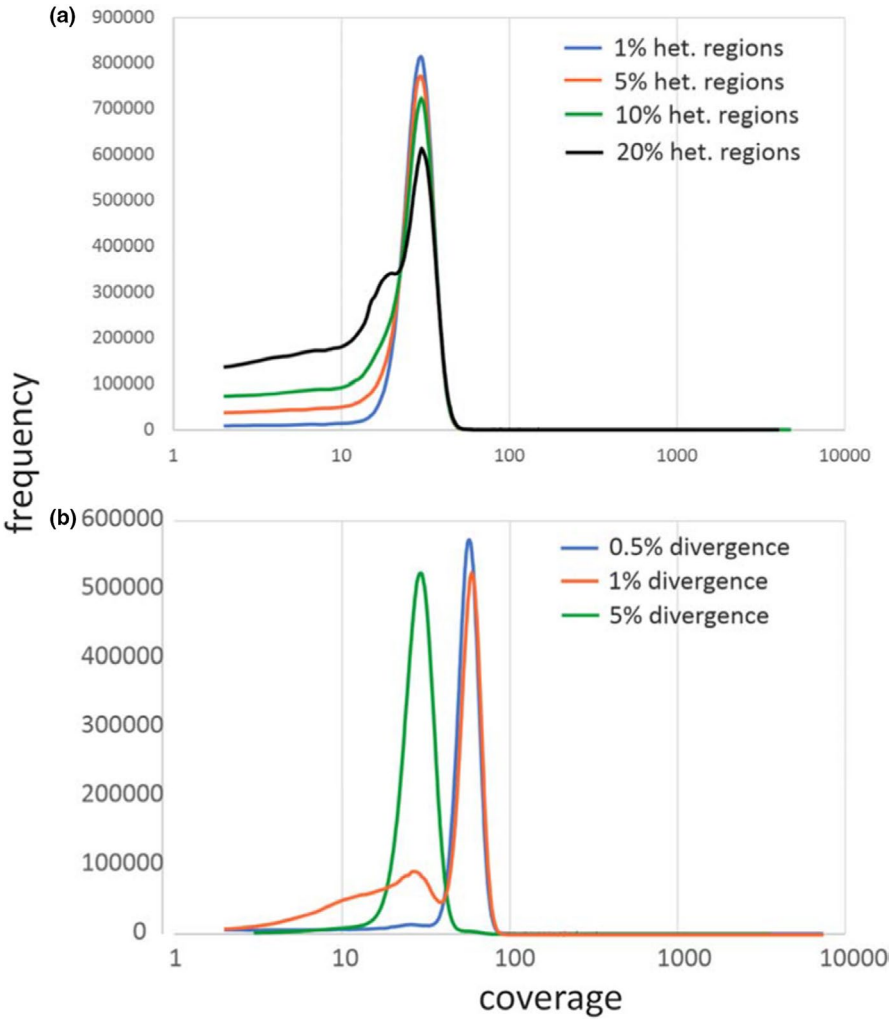

FIGURE 2 Coverage distributions for divergent heterozygous and tetraploid genomes. All distributions shown are based on the *Saccharomyces cerevisae*-like genome. (a) Coverage distributions for 0%, 5%, 10% and 20% of divergent heterozygous regions. (b) Coverage distributions for tetrapoid genomes with 0.5%, 1% and 5% divergence among the duplicated genomes. Note the logarithmic scale of the *x*-axes

with poor N50, which nevertheless yielded excellent concordance with the flow cytometry estimates (Table 2). This makes the method particularly suitable to obtain a reliable genome size estimate early in the assembly process and, if necessary, adjust the sequencing strategy. In addition, genome skimming projects (Dodsworth, 2015) with low sequencing depths could profit from the proposed method, as long as the obtained coverage is at least in the order of 2–5×. The simulations have further shown that divergent heterozygote stretches do not compromise the result of the genome size estimation.

The accuracy of genome size estimates of simulated tetraploid organisms depended strongly on the degree of divergence between the genome copies. When the divergence was low (0.5%), the assembly of the duplicated was almost completely collapsed and consequently the modal sequencing depth was twice as high as the true coverage. However, even with 1% sequence divergence between the duplicated genomes, an additional peak close to the true value of 30 was observed. For 5% sequence divergence and higher (not shown), the assembly more or less fully resolved the duplicated genomes and the highest peak was identical to the true coverage. This stressed that multimodal coverage distributions point to issues with the assembly and should always be carefully investigated. Nevertheless, if the ploidy of the organism is known, reliable estimates of the genome size can be gained with our method as well even for recent polyploidzation events.

The simulation study relied on simulated short reads as obtained, for example, by the widespread Illumina-platform. However, several included empirical examples (e.g. Chueca et al., 2021) suggested that estimating the bases sequenced from the target genome with PacBio long reads worked equally well. In principle, as long as the assumption of random sequencing of bases from the genome is fulfilled, every sequencing platform should yield reliable estimates. For mixed assemblies, however, it is advisable to use only one sort of data (preferably the one with the higher number of sequenced bases, see below), because the underlying coverage distributions are usually different.

We proposed four slightly different estimators of genome size. Simulations indicated that, as expected, the N_{clean}/λ estimator yielded by far the best results, largely independent of coverage or assembly quality. However, since we gained the reads from simulated genomes, they were by definition free of contaminations, that is reads from other organisms or other (e.g., organellar) genomes. Whether N_{clean}, the number of bases sequenced after cleaning and trimming, is reasonable for empirical estimations thus depends on

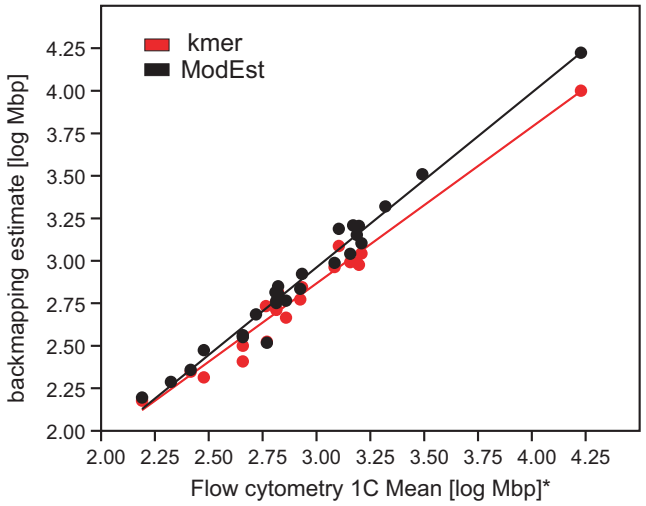

FIGURE 3 Ordinary least square regression for N_{bm}/m (black) and k-mer-based (red) genome size estimates on 1C flow cytometry estimates derived from the same individuals, respectively. For better graphical representation, estimates were log-transformed. Both regressions were highly significant ($p < .0001$). The N_{bm}/m estimates fit as well (r^2 =.998) as their k-mer-based counterparts (r^2 = .996). The slopes (0.995 for N_{bm}/m and 0.59 for k-mer-based) were significantly different

the confidence with regard to the amount of residual contamination in the data set.

For the alternative, using the number of back-mapped reads, N_{bm}, as an estimator of the bases sequenced, precision depended strongly on the completeness of the genome assembly in terms of presence of all sequence elements, regardless of their copy number. This seemed reasonable: if all repeat classes and complex regions are represented in the genome draft, all reads will find a place they can map to. If the confidence is high that N_{clean} is correct, the ratio N_{bm}/N_{clean} would be a good indicator of genome completeness in this sense.

We have shown that the λ parameter of the underlying true Poisson distribution of base coverage is readily and reliably found by ML estimation, if we truncate the data to a small window around the modal value of the sequencing depth distribution. Moreover, because the modal value of a Poisson distribution cannot deviate more than 1 from λ, the relative error from using m instead of λ decreases with increasing coverage. Most genome sequencing projects use coverages of several dozen × for at least one technique where the difference becomes marginal. Estimating genome size from low coverage (e.g., of genome-skimming projects), however, should entail proper estimation of λ.

Comparison of genome size estimates obtained with our sequencing coverage method to empirical data from flow cytometry obtained from the same individual achieved very good agreement, regardless of genome size. The regression slope of close to 1 indicated that the estimates obtained with our method can be used interchangeably with those from flow cytometry. This allows researchers to gather reliable and comparable genome size estimates for species where fresh material is difficult or impossible to obtain or where access to flow cytometry equipment is lacking.

While the k-mer-based estimates available were almost as consistent as those obtained from sequencing coverage, they were not as precise. The k-mer approach consistently underestimated the true size by more than one third. By their very nature, k-mer approaches estimate rather the content of high-complexity regions (Lipovský et al., 2017). It will therefore be interesting to see whether the observed taxon-independent relationship of approximately 2/3 complexity regions to 1/3 repeat regions as found here mainly for animal species will hold true for more genomes. The work of Novák et al. (2020) also showed an almost constant, albeit higher proportion of repetitive regions for plant genomes with sizes up to 10 Gb. Above this size, the relative proportion of repeats declined. Obtaining more reliable genome sizes from a broad taxon range will allow us to infer which processes are driving these patterns, to which the proposed ModEst method can contribute.

ACKNOWLEDGEMENTS

We thank our LOEWE-TBG colleagues for providing early access to their assembled genomes. Open access funding enabled and organized by ProjektDEAL.

DATA AVAILABILITY STATEMENT

Data Accessibility Statement: All genomes simulated are available on Dryad: https://doi.org/10.5061/dryad.dr7sqvb0j.

Benefits Generated: Benefits from this research accrue from the sharing of our data and results on public databases as described above.

ORCID

Markus Pfenninger https://orcid.org/0000-0002-1547-7245
Tilman Schell https://orcid.org/0000-0002-6431-6018

REFERENCES

Agudo, A. B., Torices, R., Loureiro, J., Castro, S., Castro, M., & Álvarez, I. (2019). Genome size variation in a hybridizing diploid species complex in *Anacyclus* (Asteraceae: Anthemideae). *International Journal of Plant Sciences*, *180*, 374–385.

Asalone, K. C., Ryan, K. M., Yamadi, M., Cohen, A. L., Farmer, W. G., George, D. J., Joppert, C., Kim, K., Mughal, M. F., Said, R., Toksoz-Exley, M., Bisk, E., & Bracht, J. R. (2020). Regional sequence expansion or collapse in heterozygous genome assemblies. *PLoS Computational Biology*, *16*, e1008104. https://doi.org/10.1371/journal.pcbi.1008104

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, *19*, 455–477. https://doi.org/10.1089/cmb.2012.0021

Bennett, M. D., & Leitch, I. J. (2005). Genome size evolution in plants. In T. R. Gregory ed., *The evolution of the genome* (pp. 89–162). Elsevier.

Blommaert, J., Riss, S., Hecox-Lea, B., Mark Welch, D. B., & Stelzer, C. P. (2019). Small, but surprisingly repetitive genomes: transposon expansion and not polyploidy has driven a doubling in genome size in

a metazoan species complex. *BMC Genomics*, *20*, 466. https://doi.org/10.1186/s12864-019-5859-y

Böhning, D., & Schön, D. (2005). Nonparametric maximum likelihood estimation of population size based on the counting distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*, 721–737.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Carta, A., Bedini, G., & Peruzzi, L. (2020). A deep dive into the ancestral chromosome number and genome size of flowering plants. *New Phytologist*, *228*, 1097–1106. https://doi.org/10.1111/nph.16668

Challis, R., Richards, E., Rajan, J., Cochrane, G., & Blaxter, M. (2020). BlobToolKit – Interactive quality assessment of genome assemblies. *G3: Genes, Genomes, Genetics*, *10*, 1361–1374. https://doi.org/10.1534/g3.119.400908

Chueca, L., Kochmann, J., Schell, T. et al (2021). *De novo* genome assembly of the raccoon dog (*Nyctereutes procyonoides*). *Frontiers in Genetics*, *12*, 658256. https://doi.org/10.3389/fgene

Chueca, L. J., Schell, T., & Pfenninger, M. (2021). *De novo* genome assembly of the land snail *Candidula unifasciata* (Mollusca: Gastropoda). *G3: Genes, Genomes, Genetics*, *11*, jkab180.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*, giab008. https://doi.org/10.1093/gigascience/giab008

David, F., & Johnson, N. (1952). The truncated poisson. *Biometrics*, *8*, 275–285. https://doi.org/10.2307/3001863

Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, *64*, 1–34.

Dodsworth, S. (2015). Genome skimming for next-generation biodiversity analysis. *Trends in Plant Science*, *20*, 525–527. https://doi.org/10.1016/j.tplants.2015.06.012

Doležel, J., & Greilhuber, J. (2010). Nuclear genome size: are we getting closer? *Cytometry Part A*, *77*, 635–642. https://doi.org/10.1002/cyto.a.20915

Elliott, T. A., & Gregory, T. R. (2015). What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*, 20140331. https://doi.org/10.1098/rstb.2014.0331

Fountain, E. D., Pauli, J. N., Reid, B. N., Palsbøll, P. J., & Peery, M. Z. (2016). Finding the right coverage: The impact of coverage and sequence quality on single nucleotide polymorphism genotyping error rates. *Molecular Ecology Resources*, *16*, 966–978. https://doi.org/10.1111/1755-0998.12519

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Götz, S., Tarazona, S., Dopazo, J., Meyer, T. F., & Conesa, A. (2012). Qualimap: Evaluating next-generation sequencing alignment data. *Bioinformatics*, *28*, 2678–2679. https://doi.org/10.1093/bioinformatics/bts503

Gardner, J. D., Laurin, M., & Organ, C. L. (2020). The relationship between genome size and metabolic rate in extant vertebrates. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*, 20190146. https://doi.org/10.1098/rstb.2019.0146

Hare, E. E., & Johnston, J. S. (2012). Genome size determination using flow cytometry of propidium iodide-stained nuclei. In V. Orogozo & M. V. Rockmann eds., *Molecular methods for evolutionary genetics* (pp. 3–12). Humana Press.

Hartke, J., Schell, T., Jongepier, E., Schmidt, H., Sprenger, P. P., Paule, J., Bornberg-Bauer, E., Schmitt, T., Menzel, F., Pfenninger, M., & Feldmeyer, B. (2019). Hybrid genome assembly of a neotropical mutualistic ant. *Genome Biology and Evolution*, *11*, 2306–2311. https://doi.org/10.1093/gbe/evz159

Heckenhauer, J., Frandsen, P. B., Gupta, D. K., Paule, J., Prost, S., Schell, T., Schneider, J. V., Stewart, R. J., & Pauls, S. U. (2019). Annotated draft genomes of two caddisfly species *Plectrocnemia conspersa* CURTIS and *Hydropsyche tenuis* NAVAS (Insecta: Trichoptera). *Genome Biology and Evolution*, *11*, 3445–3451. https://doi.org/10.1093/gbe/evz264

Heckenhauer, J., Frandsen, P. B., Sproul, J. S. et al (2021). Genome size evolution in the diverse insect order Trichoptera. *bioRxiv*. https://doi.org/10.1101/2021.05.10.443368

Huang, W., Li, L., Myers, J. R., & Marth, G. T. (2012). ART: A next-generation sequencing read simulator. *Bioinformatics*, *28*, 593–594. https://doi.org/10.1093/bioinformatics/btr708

Johnston, J. S., Bernardini, A., & Hjelmen, C. E. (2019). Genome size estimation and quantitative cytogenetics in insects. In S. Brown & M. Pfrender eds., *Insect genomics* (pp. 15–26). Springer.

Kapusta, A., Suh, A., & Feschotte, C. (2017). Dynamics of genome size evolution in birds and mammals. *Proceedings of the National Academy of Sciences of the United States of America*, *114*, E1460–E1469. https://doi.org/10.1073/pnas.1616702114

Keyl, H.-G. (1965). A demonstrable local and geometric increase in the chromosomal DNA of *Chironomus*. *Experientia*, *21*, 191–193. https://doi.org/10.1007/BF02141878

Kumar, S., Jones, M., Koutsovoulos, G., Clarke, M., & Blaxter, M. (2013). Blobology: Exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Frontiers in Genetics*, *4*, 237. https://doi.org/10.3389/fgene.2013.00237

Lefébure, T., Morvan, C., Malard, F., François, C., Konecny-Dupré, L., Guéguen, L., Weiss-Gayet, M., Seguin-Orlando, A., Ermini, L., Sarkissian, C. D., Charrier, N. P., Eme, D., Mermillod-Blondin, F., Duret, L., Vieira, C., Orlando, L., & Douady, C. J. (2017). Less effective selection leads to larger genomes. *Genome Research*, *27*, 1016–1028. https://doi.org/10.1101/gr.212589.116

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

Li, X., & Waterman, M. S. (2003). Estimating the repeat structure and length of DNA sequences using ℓ-tuples. *Genome Research*, *13*, 1916–1922. https://doi.org/10.1101/gr.1251803

Lipovský, M., Vinar, T., & Brejova, B. (2017). Approximate abundance histograms and their use for genome size estimation. *ITAT*, *2017*, 27–34.

Lynch, M., & Conery, J. S. (2003). The origins of genome complexity. *Science*, *302*, 1401–1404. https://doi.org/10.1126/science.1089370

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*, 764–770. https://doi.org/10.1093/bioinformatics/btr011

Mishra, B., Ploch, S., Runge, F., Schmuker, A., Xia, X., Gupta, D. K., Sharma, R., & Thines, M. (2020). The genome of *Microthlaspi erraticum* (Brassicaceae) provides insights into the adaptation to highly calcareous soils. *Frontiers in Plant Science*, *11*, 943. https://doi.org/10.3389/fpls.2020.00943

Mishra, B., Ulaszewski, B., Meger, J. et al (2021). A chromosome-level genome assembly of the European Beech (*Fagus sylvatica*) reveals anomalies for organelle DNA integration, repeat content and distribution of SNPs. *bioRxiv*. https://doi.org/10.1101/2021.03.22.436437

Nadarajah, S., & Kotz, S. (2006). R programs for computing truncated distributions. *Journal of Statistical Software*, *16*, Code Snippet 2.

Nickel, J. H., Schell, T., Holtzem, T. et al (2021) Hybridization dynamics and extensive introgression in the Daphnia longispina species complex: New insights from a high-quality Daphnia galeata reference genome. *bioRxiv*. https://doi.org/10.1101/2021.02.01.429177

Novák, P., Guignard, M. S., Neumann, P., Kelly, L. J., Mlinarec, J., Koblížková, A., Dodsworth, S., Kovařík, A., Pellicer, J., Wang, W., Macas, J., Leitch, I. J., & Leitch, A. R. (2020). Repeat-sequence turnover shifts fundamentally in species with large genomes. *Nature*

Plants, *6*, 1325–1329. https://doi.org/10.1038/s41477-020-00785-x

Oliver, M. J., Petrov, D., Ackerly, D., Falkowski, P., & Schofield, O. M. (2007). The mode and tempo of genome size evolution in eukaryotes. *Genome Research*, *17*, 594–601. https://doi.org/10.1101/gr.6096207

Petrov, D. A. (2001). Evolution of genome size: new approaches to an old problem. *Trends in Genetics*, *17*, 23–28. https://doi.org/10.1016/S0168-9525(00)02157-0

Pflug, J. M., Holmes, V. R., Burrus, C. et al (2020). Measuring genome sizes using read-depth, k-mers, and flow cytometry: methodological comparisons in beetles (Coleoptera). *G3: Genes, Genomes, Genetics*, *10*, 3047–3060.

Poptsova, M. S., Il'Icheva, I. A., Nechipurenko, D. Y. et al (2014). Non-random DNA fragmentation in next-generation sequencing. *Scientific Reports*, *4*, 1–6.

Prokopowich, C. D., Gregory, T. R., & Crease, T. J. (2003). The correlation between rDNA copy number and genome size in eukaryotes. *Genome*, *46*, 48–50. https://doi.org/10.1139/g02-103

Pucker, B. (2019). Mapping-based genome size estimation. *bioRxiv*. https://doi.org/10.1101/607390

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033

Schell, T., Feldmeyer, B., Schmidt, H., Greshake, B., Tills, O., Truebano, M., Rundle, S. D., Paule, J., Ebersberger, I., & Pfenninger, M. (2017). An annotated draft genome for *Radix auricularia* (Gastropoda, Mollusca). *Genome Biology and Evolution*, *9*, 585–592. https://doi.org/10.1093/gbe/evx032

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: Key considerations in genomic analyses. *Nature Reviews Genetics*, *15*, 121–132. https://doi.org/10.1038/nrg3642

Sohn, J.-I., & Nam, J.-W. (2018). The present and future of de novo whole-genome assembly. *Briefings in Bioinformatics*, *19*, 23–40.

Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nature Reviews Genetics*, *13*, 36–46. https://doi.org/10.1038/nrg3117

Vitales, D., Álvarez, I., Garcia, S., Hidalgo, O., Nieto Feliner, G., Pellicer, J., Vallès, J., & Garnatje, T. (2020). Genome size variation at constant chromosome number is not correlated with repetitive DNA dynamism in *Anacyclus* (Asteraceae). *Annals of Botany*, *125*, 611–623. https://doi.org/10.1093/aob/mcz183

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics*, *33*, 2202–2204. https://doi.org/10.1093/bioinformatics/btx153

Wang, J., Liu, J., & Kang, M. (2015). Quantitative testing of the methodology for genome size estimation in plants using flow cytometry: A case study of the *Primulina* genus. *Frontiers in Plant Science*, *6*, 354. https://doi.org/10.3389/fpls.2015.00354

Winter, S., Prost, S., de Raad, J., Coimbra, R. T. F., Wolf, M., Nebenführ, M., Held, A., Kurzawe, M., Papapostolou, R., Tessien, J., Bludau, J., Kelch, A., Gronefeld, S., Schöneberg, Y., Zeitz, C., Zapf, K., Prochotta, D., Murphy, M., Sheffer, M. M., … Janke, A. (2020). Chromosome-level genome assembly of a benthic associated Syngnathiformes species: The common dragonet, *Callionymus lyra*. *Gigabyte*, *2020*, 1–10. https://doi.org/10.46471/gigabyte.6

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Pfenninger, M., Schönnenbeck, P., & Schell, T. (2022). ModEst: Accurate estimation of genome size from next generation sequencing data. *Molecular Ecology Resources*, 22, 1454–1464. https://doi.org/10.1111/1755-0998.13570