Precise estimation of genome size from NGS data

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

[1] Senckenberg Biodiversity and Climate Research Centre, Georg-Voigt-Str. 14-16, 60325 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

Abstract

Precise estimates of genome sizes are important parameters for both theoretical and practical biodiversity genomics. We present here a fast, easy-to-implement and precise method to estimate genome size from the number of bases sequenced and the mean sequence coverage. To estimate the latter, we take advantage on the fact that a precise estimation of the Poisson distribution parameter lambda is possible from truncated data, restricted to the part of the coverage distribution representing the true underlying distribution. With simulations we could show that reasonable genome size estimates can be gained even from low-covered, 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.

Introduction

Eukaryotic genomes vary tremendously in size (Oliver *et al.* 2007; Bennett & Leitch 2005; Petrov 2001), yet the underlying processes for this variability are not yet fully understood (Elliott & Gregory 2015). To understand and study 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, i.e. over relatively short evolutionary time scales (Keyl 1965). A precise 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. Also resequencing projects requiring a certain coverage 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 (Johnston *et al.* 2019; Doležel & Greilhuber 2010). Yet, this method is not without caveats (Wang *et al.* 2015) and requires specialised laboratory skills and availability of the relatively expensive equipment. Moreover, the method depends on 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 data provides an alternative for estimating genome size (Vurture *et al.* 2017). Besides the widely used k-mer based methods (Lipovský *et al.* 2017; Li & Waterman 2003), 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 is based on the assumption that the probability to sequence a genome position is identical over the entire genome, i.e. that their true coverage is Poisson distributed. In a perfect assembly covering the entire genome, the lambda as parameter of the underlying Poisson distribution (as well

as the mean and median) of the coverage distribution should therefore be identical to the true coverage. Dividing the number of sequenced, respectively 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 over proportionally covered, skewing the coverage distribution, and hence, estimates of lambda upwards. A second source of systematic error in assemblies are relatively diverged heterozygous regions, e.g. from inversions that are not identified as homologous. These will result in a double representation of the respective region in the genome, making it longer (Asalone *et al.* 2020). Consequently, the expected coverage of these regions in the assembly will be half of the true coverage and skew the coverage distribution and parameters estimated from it downwards. In real genome assemblies, both errors likely 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 coverage distribution and an estimate of the number of bases sequenced from genome assembly drafts can be used to infer precise estimates of genome size. We tested the methods with simulations and compare genome size estimates from real data with those derived from flow cytometry and k-mer based methods.

Material and Methods

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 coverage (c) (Sims *et al.* 2014).

$$c = N / L$$

Since the coverage 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 that genome. Depending on the origin of the DNA, the data set may contain more or less 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 (Chaliis et al. 2020), this rarely succeeds completely. The number of bases after thorough cleaning, N_{clean}, estimates therefore 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 coverage 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 the part of the distribution 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), i.e. 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 smaller than or equal to 1 and therefore becomes relatively less biased when λ is high (i.e. high mean coverage). Nevertheless, better methods for estimating λ from truncated Poisson distributions exist (Delignette-Muller & Dutang 2015; Nadarajah & Kotz 2006; Böhning & Schön 2005; David & Johnson 1952).

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. In particular, samtools stats and bedtools genomecov can be easily 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 starting point for a Maximum Likelihood 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 truncations limits and give recommendations below. The respective R-code can be found in the Supplement and a Perl wrapper-script, including all necessary dependencies can be found at https://github.com/schellt/backmap.

Simulations

To illustrate the influence of factors like sequencing depth, genome size, repeat content and -distribution on the different genome size estimation methods, we simulated five different genomes according to real examples. The latest 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/Simulate-Genome. 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.

Table 1: Simulated genomes and their characteristics, rr = repeat regions.

Simulated genome	Size (Mbp)	average distance between rr	average length 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

From these simulated genomes, we generated synthetic next-generation sequencing short read sets of 10X, 30X and 60X coverage 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 datasets 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 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 coverage were back-mapped to the simulated genomes they were derived from. 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-divergent 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, 10X coverage), we evaluated the effect of different truncation limits on the precision of the λ estimation. For coverage 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 eleven around the estimated modal value. We calculated four different genome size estimates:

i) N_{clean}/λ, the number of "sequenced" bases after cleaning and trimming divided by the ML λ estimate derived from the empirical coverage distribution.
ii) N_{clean}/m, the number of "sequenced" bases after cleaning and trimming divided by the modal value of the empirical coverage distribution.
iii) N_{bm}/λ, the number of back-mapped bases divided by the ML λ estimate derived from the empirical coverage distribution.
iv) N_{bm}/m, the number of back-mapped bases divided by the modal value of the empirical coverage distribution.

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

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 also flow cytometry estimates gained from the very same individual/clone/population were available. The taxonomic range of genomes comprised plants and several animal taxa with a focus on insects (Table 2).

Table 2. Genomes used for empirical evaluation.

Species	Taxon	Flow-cytometry estimate [Mb]	Backmapping estimate [Mb]	k-mer based estimate [Mb]	Citation
Hydropsyche tenuis	Insecta	260.60	228.56	222.82	Heckenhauer *et al.* 2019
Plectrocnemia conspersa	Insecta	455.20	364.91	316.26	Heckenhauer *et al.* 2019
Agapetus fuscipens	Insecta	721.79	583.45	463.24	Heckenhauer et al. 2021
Odontocerum albicorne	Insecta	1616.01	1270.00	1103.37	Heckenhauer et al. 2021

Drusus annulatus	Insecta	840.20	684.34	592.32	Heckenhauer et al. 2021
Halesus radiatus	Insecta	1212.35	972.29	918.70	Heckenhauer et al. 2021
Micropterna sequax	Insecta	1434.71	1100.00	981.67	Heckenhauer et al. 2021
Micrasema longulum ML1	Insecta	663.60	707.65	650.70	Heckenhauer et al. 2021
Micrasema longulum ML3	Insecta	663.60	637.83	635.22	Heckenhauer et al. 2021
Micrasema minimum	Insecta	588.81	329.32	333.84	Heckenhauer et al. 2021
Rhyacophila evoluta Rss1	Insecta	651.30	581.81	518.83	Heckenhauer et al. 2021
Rhyacophila evoluta HR1	Insecta	651.30	565.46	514.41	Heckenhauer et al. 2021
*Glaux maritima (*also known as *Lysimachia maritima)*	Angiosperm plant	1270.00	1541.39	1221.29	Segers et al. unpublished
Radix auricularia	Mollusca	1575.00	1603.00	947.10	Schell *et al.* 2017
Crematogastor levior	Insecta	455.00	356.00	255.93	Hartke *et al.* 2019
Daphnia galeata	Crustacea	155.00	157.00	150.45	Nickel *et al.* 2021
Candidula unifasciata	Mollusca	1540.00	1420.00	977.55	Chueca *et al.* 2021c
Styela plicata	Tunicata	430.92	468.60	338.81	Galià-Camps et al. unpublished
Callionymus lyra	Teleostei	645.00	653.17	562.00	Winter *et al.* 2020
Pimpla turbinella	Insecta	300.00	298.00	206.00	Reumont
Fagus sylvatica	Angiosperm plant	582.40	542.00	541.00	Mishra *et al.* 2021
Aedes japonicus	Insecta	857.00	836.25	699.00	Reuss et al. unpublished
Nyctereutes procynoides	Mammalia	3100.00	3230.00	-	Chueca *et al.* 2021b
Microthlaspi erraticum	Angiosperm plant	194.50	211.00	211.00	-	Mishra *et al.* 2020

If not stated otherwise in the citations, genome size estimates from flow cytometry was estimated following a protocol with propidium iodide-stained nuclei described in (Hare & 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 propidium iodide (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 for all genomes. The modal value was chosen, because the coverage exceeded 50X in most cases. For comparison, we performed or used published k-mer based estimates. First the 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 and histo. 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.

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 difference (Cohen *et al.* 2013).

Results

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 *S. saccharomyces*-like, the *C. elegans*-like and the *S. maximus*-like genomes, the total lengths of the assemblies were above 90% of the true size, for the remaining two below 80%. This was reflected in the back-mapping rates that were highly correlated to the relative assembly length ($r = 0.995$, $p < 0.001$, Table 3).

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	assembly size [bp]	proportion of true length	number of contigs	mean contig length	bp "sequenced"	bp mapped	proportion of bases mapped	estimated lambda	modal coverage
*Saccharomyces_cerevisae*_like	10	1.208E+07	1.124E+07	0.930	2,021	5,561	1.208E+08	1.130E+08	0.931	9.998362699	9
	30	1.208E+07	1.125E+07	0.931	1,823	6,170	3.624E+08	3.391E+08	0.932	29.99218743	29
	60	1.208E+07	1.125E+07	0.932	1,906	5,905	7.248E+08	6.777E+08	0.931	59.99714024	59
*Caenorhabditis_elegans*_like	10	1.000E+08	9.277E+07	0.927	105,104	883	1.000E+09	9.337E+08	0.904	9.995972787	9
	30	1.000E+08	9.280E+07	0.928	98,341	944	3.001E+09	2.807E+09	0.910	29.9914132	29
	60	1.000E+08	9.295E+07	0.929	99,957	930	6.002E+09	5.598E+09	0.904	59.99311067	59
*Arabidopsis_thaliana*_like	10	1.201E+08	9.283E+07	0.773	66,881	1,388	1.201E+09	9.566E+08	0.780	9.995704219	9
	30	1.201E+08	9.303E+07	0.775	63,695	1,461	3.602E+09	2.881E+09	0.785	29.99534966	29
	60	1.201E+08	9.275E+07	0.772	61,615	1,505	7.205E+09	5.759E+09	0.784	59.98860759	59
*Drosophila_melanogaster*_like	10	1.441E+08	1.078E+08	0.748	104,002	1,037	1.441E+09	1.118E+09	0.755	9.99480137	9
	30	1.441E+08	1.077E+08	0.747	95,701	1,125	4.322E+09	3.382E+09	0.763	29.98949853	29
	60	1.441E+08	1.076E+08	0.747	95,523	1,127	8.643E+09	6.725E+09	0.756	59.99197104	59
*Scophthalmus_maximus*_like	10	5.241E+08	5.234E+08	0.999	76,507	6,842	5.241E+09	5.220E+09	0.994	9.992722402	9
	30	5.241E+08	5.241E+08	1.000	63,360	8,261	1.572E+10	1.567E+10	0.995	29.99741619	29
	60	5.241E+08	5.241E+08	1.000	63,260	8,274	3.144E+10	3.134E+10	0.995	59.99244322	59

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). Closer analysis revealed that performance depended on the degree of successful reconstruction of the genome: if more than 90% of the genome were represented in the assembly, deviations were below 7% (Figure 1, Table 3), independent of estimator. There was also 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 contiguity among genomes assemblies reconstructed form different coverages, this factor had only a minor role for the precision of the genome size estimate (Table 3, Figure 1).

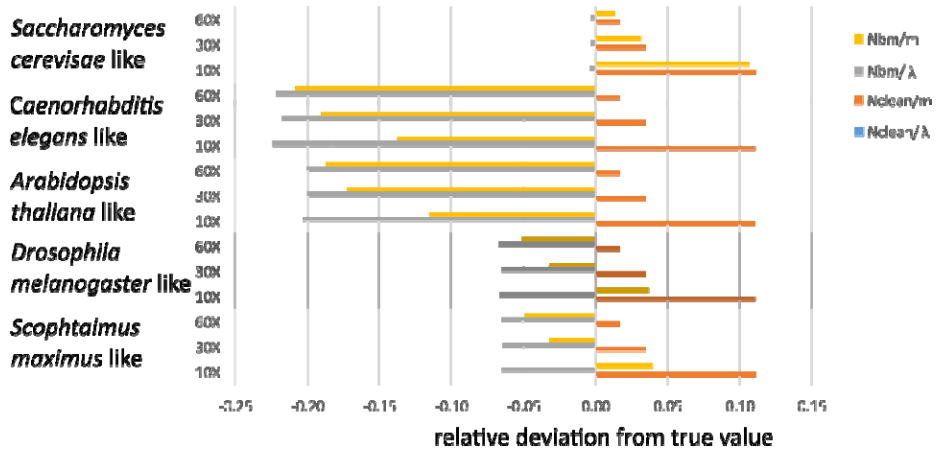

Figure 1. Relative deviations of genome size estimators from true values for different simulated genomes and simulated coverages. The deviations from N_{clean}/λ (blue) are so small that they are not visible on the scale.

Empirical data

Ordinary Least Squares Regression of 1C flow-cytometry estimates against the estimates derived from the coverage approach yielded a very good fit ($r^2 = 0.957$, $p = 3.9 \times 10^{-15}$). The estimated slope was with 1.01+/- 0.047 (s.e.) close to unity. The fit of the respective k-mer based estimates to the flow cytometry data also good ($r^2 = 0.879$, $p = 3.7 \times 10^{-10}$), however, the slope of 0.657 +/- 0.056 (s.e.) suggested a systematically lower k-mer estimate (Figure 2). The estimated slopes were significantly different from each other ($t = 4.75$, d.f. = 39, $p = 2 \times 10^{-5}$).

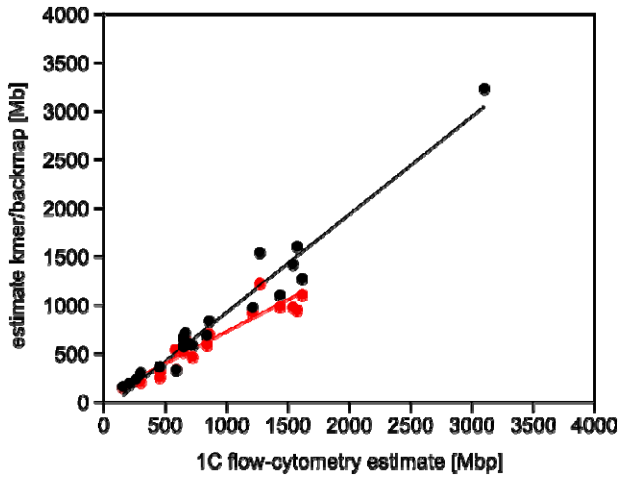

Figure 2. Ordinary least square regression for N_{bm}/m (black) and k-mer based (red) genome size estimates on 1C flow-cytometry estimates derived from the same individuals, respectively. Both regressions are highly significant ($p < 0.0001$). The N_{bm}/m estimates fitted better ($r^2 = 0.957$) than

their k-mer based counterparts ($r^2 = 0.879$). The slopes (1.01 for N_{bm}/m and 0.65 for k-mer based) are significantly different.

Discussion

As long as reliable whole chromosome sequencing is technically not yet 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 presented here a fast, easy-to-use and precise method for genome size estimation from NGS sequencing data. The method could become standard part of the genome assembly process, because it relies on data that is routinely collected. Albeit our method is not the first to propose the use of sequencing, respectively mapping statistics (Pflug *et al.* 2020; Pucker 2019), it requires less assumptions and much less bioinformatic effort than previously suggested approaches.

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 could show that the precision of the estimate is largely independent from 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 makes the method particularly suitable to obtain a reliable genome size estimate early in the assembly process. But also genome skimming projects (Dodsworth 2015) with low coverages could profit from the proposed method.

The simulation study relied on simulated short reads as obtained e.g. by the widespread Illumina-platform, however, empirical examples (Chueca *et al.* 2021a) suggested that estimating the bases sequenced from the target genome from PacBio long reads work 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 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, in practice largely independent of coverage or assembly quality. However, since we gained the reads from simulated genomes, they were free of contaminations, i.e. reads from other organisms, respectively genomes. Whether $N_{clean,}$ the amount of bases sequenced, is reasonable for empirical estimations depends thus on the confidence with regard to the amount of residual contamination in the data set.

For the alternative, using the number 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 coverage 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 relative decreases with increasing coverage. Most genome sequencing projects use coverages of several dozen X 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 from empirical data to flow cytometry data obtained from the same individual, respectively clone achieved very good agreement. The regression slope of close to 1 indicated that the estimates obtained with our method can be used interchangeably. This allows to gather reliable and comparable genome size estimates also for species, where fresh material is difficult or impossible to obtain or access to flow-cytometry equipment is lacking.

While the k-mer based estimates available were almost as accurate those obtained from sequencing coverage, they were not precise. The k-mer approach consistently underestimated the true size by about one third. By their very nature, k-mer approaches estimate rather the content of high complexity regions (Lipovský *et al.* 2017). It will be therefore interesting to see whether the observed taxon independent relation of approximately 2/3 complexity regions to 1/3 repeat regions will hold true for more genomes and, if so, which processes are driving this pattern.

Acknowledgements

TBG colleagues for giving early access to assembled genomes.

References

Asalone KC, Ryan KM, Yamadi M*, et al.* (2020) Regional sequence expansion or collapse in heterozygous genome assemblies. *PLoS Computational Biology* **16**, e1008104.

Bankevich A, Nurk S, Antipov D*, et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal Of Computational Biology* **19**, 455-477.

Bennett MD, Leitch IJ (2005) Genome size evolution in plants. In: *The evolution of the genome*, pp. 89-162. Elsevier.

Blommaert J, Riss S, Hecox-Lea B, *et al.* (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**, 1-12.

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 AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120.

Challis R, Richards E, Rajan J, *et al.* (2020) BlobToolKit – Interactive Quality Assessment of Genome Assemblies. *G3 Genes|Genomes|Genetics* **10**, 1361–1374.

Chueca L, Kochmann J, Schell T*, et al.* (2021) De novo Genome Assembly of the Raccoon Dog (Nyctereutes procyonoides). *Frontiers in Genetics* **12***: 658256. doi: 10.3389/fgene*.

Chueca LJ, Schell T, Pfenninger M (2021) De novo genome assembly of the land snail Candidula unifasciata (Mollusca: Gastropoda). *G3: Genes, Genomes, Genetics in press.*

Cohen J, Cohen P, West SG, Aiken LS (2013) *Applied multiple regression/correlation analysis for the behavioral sciences* Routledge.

Danecek P, Bonfield JK, Liddle J, *et al.* (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.

Delignette-Muller ML, 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.

Doležel J, Greilhuber J (2010) Nuclear genome size: are we getting closer? *Cytometry Part A* **77**, 635-642.

Elliott TA, Gregory TR (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.

Fountain ED, Pauli JN, Reid BN, *et al.* (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.

García-Alcalde F, Okonechnikov K, Carbonell J*, et al.* (2012) Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics* **28**, 2678-2679.

Gardner JD, Laurin M, Organ CL (2020) The relationship between genome size and metabolic rate in extant vertebrates. *Philosophical Transactions of the Royal Society B* **375**, 20190146.

Hare EE, Johnston JS (2012) Genome size determination using flow cytometry of propidium iodide-stained nuclei. In: *Molecular methods for evolutionary genetics* (pp. 3-12). Humana Press.

Hartke J, Schell T, Jongepier E*, et al.* (2019) Hybrid genome assembly of a neotropical mutualistic ant. *Genome Biology and Evolution* **11**, 2306-2311.

Heckenhauer J, Frandsen PB, Gupta DK*, et al.* (2019) Annotated draft genomes of two caddisfly species Plectrocnemia conspersa CURTIS and Hydropsyche tenuis NAVAS (Insecta: Trichoptera). *Genome Biology and Evolution* **11**, 3445-3451.

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

Huang W, Li L, Myers JR, Marth GT (2012) ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593-594.

Johnston JS, Bernardini A, Hjelmen CE (2019) Genome size estimation and quantitative cytogenetics in insects. In: *Insect genomics*, pp. 15-26. Springer.

Keyl H-G (1965) A demonstrable local and geometric increase in the chromosomal DNA of Chironomus. *Experientia* **21**, 191-193.

Kumar S, Jones M, Koutsovoulos G, *et al.* (2013) Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Frontiers in Genetics* **4**, 237.

Lefébure T, Morvan C, Malard F*, et al.* (2017) Less effective selection leads to larger genomes. *Genome Research* **27**, 1016-1028.

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

Li X, Waterman MS (2003) Estimating the repeat structure and length of DNA sequences using ℓ-tuples. *Genome Research* **13**, 1916-1922.

Lipovský M, Vinar T, Brejova B (2017) Approximate Abundance Histograms and Their Use for Genome Size Estimation, 27-34.

Lynch M, Conery JS (2003) The origins of genome complexity. *Science* **302**, 1401-1404.

Marçais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764-770.

Mishra B, Ploch S, Runge F, *et al.* (2020) The genome of Microthlaspi erraticum (Brassicaceae) provides insights into the adaptation to highly calcareous soils. *Frontiers in plant science* **11**, 943.

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*.

Nadarajah S, Kotz S (2006) R programs for computing truncated distributions. *Journal of.*

Nickel JH, 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*.

Oliver MJ, Petrov D, Ackerly D, *et al.* (2007) The mode and tempo of genome size evolution in eukaryotes. *Genome Research* **17**, 594-601.

Petrov DA (2001) Evolution of genome size: new approaches to an old problem. *Trends in Genetics* **17**, 23-28.

Pflug JM, Holmes VR, 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.

Prokopowich CD, Gregory TR, Crease TJ (2003) The correlation between rDNA copy number and genome size in eukaryotes. *Genome* **46**, 48-50.

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

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842.

Schell T, Feldmeyer B, Schmidt H, *et al.* (2017) An annotated draft genome for Radix auricularia (Gastropoda, Mollusca). *Genome Biology and Evolution* **9**, 00-00.

Sims D, Sudbery I, Ilott NE, Heger A, *et al.* (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics* **15**, 121-132.

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

Treangen TJ, Salzberg SL (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics* **13**, 36-46.

Vurture GW, Sedlazeck FJ, Nattestad M, *et al.* (2017) GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202-2204.

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.

Winter S, Prost S, De Raad J, *et al.* (2020) Chromosome-level genome assembly of a benthic associated Syngnathiformes species: the common dragonet, *Callionymus lyra. Gigabyte* **1**, *https://doi.org/10.46471/gigabyte.6*

Supplement

R-code for estimating lambda from a truncated Poisson distribution

```
library(fitdistrplus)
```

```
library(truncdist)
library(splitstackshape)
#transform Qualimap output to R-object
obj <- read.table("coverage_histogram.txt", header = TRUE)
obj <- expandRows(obj, "freq")
obj <- as.vector(obj$freq)
summary(obj)
#define function for mode
mode <- function(obj) {uniqv <- unique(obj) uniqv[which.max(tabulate(match(obj, uniqv)))]}
min <- mode – 5
max <- mode + 5
dtruncated_poisson <- function(x, lambda) {dtrunc(x, "pois", a=min, b=max, lambda=lambda)}
ptruncated_poisson <- function(q, lambda) {ptrunc(q, "pois", a=min, b=max, lambda=lambda)}
fitdist(obj, "pois", start = list(lambda = mode))
```