OXFORD GENETICS

GENETICS, 2024, iyae099
https://doi.org/10.1093/genetics/iyae099
Advance Access Publication Date: 13 June 2024
Investigation

Genome size and chromosome number are critical metrics for accurate genome assembly assessment in Eukaryota

Carl E. Hjelmen*

Department of Biology, Utah Valley University, 800 W. University Parkway, Orem, UT 84058, USA

*Corresponding author: Department of Biology, Utah Valley University, 800 W. University Parkway, Orem, UT 84058, USA. Email: Carl.Hjelmen@uvu.edu

The number of genome assemblies has rapidly increased in recent history, with NCBI databases reaching over 41,000 eukaryotic genome assemblies across about 2,300 species. Increases in read length and improvements in assembly algorithms have led to increased contiguity and larger genome assemblies. While this number of assemblies is impressive, only about a third of these assemblies have corresponding genome size estimations for their respective species on publicly available databases. In this paper, genome assemblies are assessed regarding their total size compared to their respective publicly available genome size estimations. These deviations in size are assessed related to genome size, kingdom, sequencing platform, and standard assembly metrics, such as N50 and BUSCO values. A large proportion of assemblies deviate from their estimated genome size by more than 10%, with increasing deviations in size with increased genome size, suggesting nonprotein coding and structural DNA may be to blame. Modest differences in performance of sequencing platforms are noted as well. While standard metrics of genome assessment are more likely to indicate an assembly approaching the estimated genome size, much of the variation in this deviation in size is not explained with these raw metrics. A new, proportional N50 metric is proposed, in which N50 values are made relative to the average chromosome size of each species. This new metric has a stronger relationship with complete genome assemblies and, due to its proportional nature, allows for a more direct comparison across assemblies for genomes with variation in sizes and architectures.

Keywords: genome size; genome assembly; sequencing; eukaryote; chromosome number; assessment; genomics

Introduction

Since the advent of genome sequencing in 1977 (Sanger *et al.* 1977), the technology has rapidly advanced both in terms of efficiency and effectiveness (Shendure *et al.* 2017). It could be argued that there is not a better example of these increases in technology than the sequencing of the human genome. While the original human genome project lasted more than a decade and cost 3 billion US dollars (Hood and Rowen 2013), there is now the newly announced $200 human genome from Illumina sequencing technology (Greatest impacts in years: A look back at Illumina and the evolution of genomics 2023). Additionally, third generation technologies are allowing us to sequence ultra-long reads with progressively less starting genetic material, allowing for reference quality de novo genome assemblies for nonmodel species (Etherington *et al.* 2020; Jaworski *et al.* 2020; Kress *et al.* 2022). These advancements in technology and improvements in sequencing quality and read length make the almost "Star Trek"-dream of sequencing every species a reality. This reality is becoming clear through projects such as the Earth BioGenome Project (Lewin *et al.* 2018, 2022; Coddington *et al.* 2019), the Darwin Tree of Life Project (The Darwin Tree of Life Project Consortium 2022), B10K (Zhang 2015), i5K (Consortium 2013), and more. While we have been able to establish incredibly cost- and material-efficient pipelines for pumping out highly contiguous genomes for large groups of species, such as the *Drosophila* genus (Kim *et al.* 2021, 2023), we still run into an issue with many of our assemblies: genome size and the C-value paradox/enigma (Gregory 2000).

Genome size and the C-value enigma

Genome size (C-value) varies extensively across the tree of life (Gregory 2000), with more than 60,000-fold variation across Eukaryota (Gregory 2000; Elliott and Gregory 2015). More specifically, there is nearly 7,000-fold variation within animals (Palazzo and Gregory 2014) and 2,400-fold variation in plants (Pellicer *et al.* 2018). Decades of research on the topic have continued to show that there is little-to-no relationship between the amount of genetic material and organismal complexity; very little of this variation in genome size is due to protein-coding sequence in eukaryotes (Gregory and Hebert 1999; Gregory 2000, 2001; Elliott and Gregory 2015). The variation we see in genome size is largely due to noncoding DNA (not protein coding), such as repetitive DNA, introns, and transposable elements (Kidwell 2002; Ågren and Wright 2011; Sessegolo *et al.* 2016). This "nongenic" DNA has affectionately, or not-so affectionately, taken on the nickname of "junk-DNA" (Comings 1972; Ohno 1972) or "the dark matter of the genome" (Blaxter 2010; Sedlazeck *et al.* 2018; Girardini *et al.* 2023), with some arguing that this additional DNA is simply nonfunctional and likely deleterious to maintain (Makalowski 2000, 2003; Biémont and Vieira 2006; Doolittle 2013; Palazzo and Gregory 2014). So, simply put, the bigger the genome, the larger proportion that is not protein coding and that is likely repetitive.

Received on 02 April 2024; accepted on 06 June 2024

While historically there has been a focus on protein-coding DNA, there is an abundance of evidence suggesting there is a lot of important information we might be missing in this nongenic DNA. Many have hypothesized that larger genomes are deleterious and are fixed by genetic drift (Lynch and Conery 2003; Knight *et al.* 2005; but see: Whitney and Garland 2010). However, studies also show adaptive benefits to variations in genome size. In some cases, larger genomes can improve reproductive fitness in seed beetles (Arnqvist *et al.* 2015), be related to clinal variation/climate in plants (Díez *et al.* 2013; Bilinski *et al.* 2018), and even song-attractiveness in grasshoppers (Schielzeth *et al.* 2014). More recently, there have been attempts to find consistent phylogenetic patterns for genome size variation and change, but the overwhelming result continues to be: "it depends on the group we're looking at" (Leitch *et al.* 1998; Arnqvist *et al.* 2015; Clark *et al.* 2016; Jeffery *et al.* 2016; Alfsnes *et al.* 2017; Lower *et al.* 2017; Hjelmen *et al.* 2019; Bainard *et al.* 2020; Liu *et al.* 2020; Yuan *et al.* 2021).

Even beyond the arguments of overall genome size and its impact on the fitness and adaptability, we must consider the impact of structural variation in the genome within and among species. There is an abundance of evidence suggesting that these structural genomic variants play important roles in evolutionary adaptations and success (reviewed in Mérot *et al.* 2020). In humans, structural variants are thought to have been involved in adaptation to different diets, infectious diseases, and even to our brains (reviewed in Hollox *et al.* 2022). At the level of copy number variation, there are strong ties to drug responses and diseases in humans (Gamazon and Stranger 2015), drought tolerance in *Hordeum spontaneum* (Kalendar *et al.* 2000), clinal variation *in Drosophila simulans* (Vieira *et al.* 1998; Adrion *et al.* 2019), and a large number of important adaptations in *Populus balsamifera* (Prunier *et al.* 2019). While there is such evidence, the impact of structural variants is very much context-dependent and complicated (Lucek *et al.* 2019), similar to conclusions in studies of phylogenetic patterns in genome size evolution (discussed above). We must continue to investigate the role of structural variants on evolutionary success and adaptation.

Assembling and assessing genome assemblies

Let's face it, large genomes are just harder to assemble, especially due to the additional repetitive DNA. For example, the migratory locust genome (6.5 Gbp assembly) has only recently been undertaken as a genome assembly project. Early drafts of this genome "boast" a staggering 1,000,000 + contigs, with a minute 9.26 Kb contig N50 (179,233 contigs) and 320.3 Kb scaffold N50 (3,440 scaffolds), which are miniscule compared to its genome size (Wang *et al.* 2014). Results are similar for the first draft of the desert locust, *Schistocerca gregaria* (8.8 Gbp, Verlinden *et al.* 2021). Even the world-changing human genome project, which "completed" the assembly for the ~3 billion base pair human genome in 2001, was still missing 8–10% of the genome. With a staggering amount of coverage and sequencing technologies, we finally completed the telomere to telomere assembly of the entire human genome 22 years after its "completion" (Nurk *et al.* 2022). The issue in the case of the human genome was overwhelmingly common and repetitive rDNA sequences in acrocentric chromosomes (Nurk *et al.* 2022). Due to our historic and continued focus on protein-coding sequences, it seems easier to settle for assemblies that are high quality in regard to protein-coding genes, and hope that the resulting assembly contig/scaffold number aligns with the haploid chromosome count. Over time, with this focus on protein-coding genes, we have also seen increases in contiguity. Much of this progress has been thanks to long-read technologies and other methodologies like Hi-C and optical mapping (Zhang *et al.* 2019; Giani *et al.* 2020; Kronenberg *et al.* 2021; Rhie *et al.* 2021; Kong *et al.* 2023; Wang, Yu *et al.* 2023).

Standard methods of assessing genome assemblies rely on contiguity, for example contig and scaffold N50 (Thrash *et al.* 2020). These values are indicative of the size of the smallest contig, or scaffold, which is needed to reach half of the genome size, when contigs are ordered from largest to smallest. What a "good" N50 value is continues to be ambiguous. Usually, we consider a bigger N50 value to be "better"; however, it is important to note that a larger N50 value could also be due to artificially and incorrectly joined contigs. We must also consider the size of the genome we are assessing when determining a satisfactory N50 value. Some genomes are much larger than others; an N50 of 15 Mbp may be excellent for a 200 Mbp *Drosophila* genome, but may be concerning for a massive plant genome.

Other standard methods of assessing assemblies rely on the assemblies inclusion of core gene sets, often assessed with BUSCO (Benchmarking Universal Single Copy Orthologs; Simão *et al.* 2015; Manni *et al.* 2021a). Here, a gene is considered part of the BUSCO requirement if it is found as a single copy in at least 90% of the genomes at that taxonomic rank. Due to this requirement, there are fewer BUSCO genes at broader taxonomic ranks than there are at narrower taxonomic ranks (e.g. 255 genes in Eukarya vs 3258 genes for Diptera in their respective BUSCO v5 databases; Manni *et al.* 2021b). BUSCO scores are listed as a percentage, with a higher percentage score indicating the presence of more complete and single copy orthologs in the genome. While BUSCO can be a useful metric, it is important to note that this metric is only assessing the presence of genes considered to be universal and present in 1 copy, and not necessarily all genes present in an organism.

No assessment should be used alone to assess the quality of a genome, especially if we want to consider a genome as "complete" (Thrash *et al.* 2020; Jauhal and Newcomb 2021). N50 values should be considered as relative to the size of the genome, ideally the chromosome size, of the organism being assembled. BUSCO scores are based on what we know about what should be there in genomes based on what we have already sequenced, but does not inform us of the number of genes that should be there, regardless of copy number. There are a number of proposals for updated metrics for assessing genomes (Wang and Wang 2023), but we should really be thinking at the foundational and structural level of the genome. For instance, does our assembly match the estimated genome size for that species? Are our contigs/scaffolds similar in size to that of the chromosomes of this organism?

Investigating genome assembly vs genome size

The goal of this work is to assess available genome assemblies in regard to estimated genome sizes for their corresponding species. Additionally, information such as sequencing platform, taxonomic rank, and assessment metrics are taken into consideration when evaluating these genome assemblies. A new metric for assessing genome assemblies is proposed. In order to accomplish this goal, information for all eukaryotic genomes (41,358 genomes as of 2024 February 20) was downloaded from relevant NCBI databases (Assembly, BioSamples, SRA, and Taxonomy). The genome assembly information was analyzed and assessed when there was available information from the animal, plant, or fungi genome sizes databases (15,133 genomes with corresponding genome size estimates) as well as chromosome count information available for animals from karyotype databases.

Methods

Retrieving assembly information

A custom R script was developed to retrieve genome assembly information for Eukaryota (TaxID: 2759) from the National Center for Biotechnology (NCBI) databases using functions from the rentrez, XML, and httr packages in R v.4.2.1 in Rstudio 2022.07.1 (Lang and team 2012; R Core Team 2016; Winter 2017; Sayers *et al.* 2022; Wickham 2023b). A list of accession numbers associated with the Eukarota TaxID was generated on 2024 February 20 and consisted of 41,358 unique assemblies. Assembly information was downloaded from 2024 February 20–27.

Assembly statistics (Assembly length, Contig N50, Scaffold N50, BUSCO scores, etc.) were retrieved for each eukaryotic assembly accession from the NCBI Assembly Database where available using the rentrez_summary function. XML information was parsed with the xmlParse and xmlToList functions. Taxonomic information was retrieved from the Taxonomy database, while sequencing methodology information was retrieved from the Sequence Read Archive. The entrez_link function was utilized to ensure data retrieved from other databases (Taxonomy, BioSample, SRA) were correctly associated with the original assembly. It is important to note that some genome assemblies in the database are incomplete genome assemblies and will not be assembled near the genome size.

Retrieving genome size information

Genome size information for Metazoa was gathered from the Animal Genome Size database (https://genomesize.com/) as a downloaded CSV file on 2024 February 10 (Gregory 2024). Genome size records which included a range for single records were removed. This resulted in 8,416 records from 6,525 species. Genome size information for plants (12,273 records from 11,634 species) was retrieved from the Plant DNA C-values Database maintained by the Kew Royal Botanic Gardens (https://cvalues.science.kew.org/) on 2024 February 20 by copy and pasting all records into a CSV file (Pellicer and Leitch 2020). Fungal genome size information (2,412 records from 1,324 species) was gathered from the Fungal Genome Size database (http://www.zbi.ee/fungal-genomesize/) on 2024 February 20 by copy and pasting all records in a CSV file (Kullman *et al.* 2005). All genome size information will not be provided as supplemental information, as all records are freely available on these databases.

Additional genome size information was added for *Drosophila* (304 records from 152 species) and polyneopteran genome size records (60 records from 40 species) not available on genomesize.com at the time of this study [available as supplemental data from Hjelmen *et al.* (2020); Sylvester *et al.* (2020)]. All information was combined into 1 CSV document with corresponding taxonomic information. Genome sizes were converted into base pairs (bp) in order to correspond with assembly size values. Where records were reported in picograms (pg), bp was calculated by multiplying by 978 and then by 1 million. This effort resulted in 23,447 genome size estimation records across 19,577 species.

Retrieving karyotype information

Karyotype information for mammals, dipterans, coleopterans, amphibians, polyneopterans was downloaded from the karyotype database on 2024 February 20 (https://coleoguy.github.io/karyotypes/; Blackmon and Demuth 2015; Perkins *et al.* 2019; Sylvester *et al.* 2020; Morelli *et al.* 2022). All records (12,773) were combined in R with corresponding taxonomic information and unified to report the haploid chromosome number for each record.

Formatting data for analysis

An R script was used to pull out assembly records from aforementioned downloaded assembly information for those that had corresponding genome size estimation records. Records were considered a match if the Genus and species name of the assembly matched the genome size Genus and species record identically. Genome size differences were calculated by subtracting the total assembled size from the estimated genome size from the genome size databases. Species with multiple records for genome sizes had the estimate value averaged before subtracting the assembled sizes. Negative values indicate a genome was assembled larger than the estimated size, and positive values indicated the genome was assembled smaller than the estimated genome size. "Proportional differences from genome size" were calculated by dividing the above difference by the averaged genome size for each species. This allows the data to be relational and normalized to account for the vast variation in genome sizes across eukaryotic species.

Sequencing methodologies were simplified to 1 method per record. In cases where multiple platforms were used (Illumina, PacBio SMRT, Oxford Nanopore), the longer method was reported for the subsequent analyses. For example, if an assembly reported both Illumina and PacBio SMRT as sequencing platform, the method is reported here as PacBio.

Assessment of genome assemblies

Genome assemblies were assessed with the above described metric of "Proportional difference from genome size". A value of ±0.1 was selected arbitrarily as a metric of a "good" genome assembly. This value represents the assembly length as being within 10% of the estimated genome size. The distribution of the proportional difference from genome size was visualized across sequencing methodologies and the Kingdom level of taxonomy using ggplot, ggpubR, ggExtra, and rphylopic (Wickham 2016; Kassambara 2018; Attali and Baker 2023; Garnier *et al.* 2023; Gearty and Jones 2023). Data manipulation was performed with the dplyr, forcats, reshape, EnvStats packages in R (Wickham 2007, 2023a; Millard 2013; Wickham *et al.* 2023).

Comparison of distributions of "Proportional Difference" across sequencing methodologies and Kingdom were completed using Kruskal–Wallis tests with a post hoc Dunn's test with a Bonferroni correction in R using the "dunn.test" package (Dinno 2017). The relationship between Proportional Difference and genome size was tested using regression analysis with log(genome size) as the predictor variable and the absolute value of the "Proportion Difference" as the response variable.

The relationship between standard genome assembly assessment methods (Contig N50, Scaffold N50, and BUSCO) and the absolute value of "Proportion Difference" was assessed using a regression analysis. Additional analysis was performed assessing the relationship between "Proportion Difference" and Contig N50 and Scaffold N50 divided by the average chromosome size (Proportional N50, PN50). Average chromosome size was estimated by taking the genome size for a species divided by the haploid chromosome number from the karyotype databases. Plant karyotype information was not used as plants have a large number of instances of polyploidy events and hybridization, which complicate the analysis at this level.

Drosophila underreplication analysis

In order to identify if there was a relationship between the difference between assembly size and genome size and repetitive or heterochromatic DNA, we must utilize species in which there are estimations available for genome size and repetitive content. *Drosophila* provide a unique advantage in this regard, as they undergo 1 round of incomplete replication, or underreplication, in their thoracic tissues (Johnston *et al.* 2013, 2020). Estimation of this underreplication have been used to estimate the proportion of the genome which is late-replicating heterochromatin vs early replicating DNA. Estimations of underreplication in *Drosophila* were retrieved from previous work (Hjelmen *et al.* 2020). In order to calculate proportion of late-replicating heterochromatin, percent underreplication (how much is additionally replicated) reported in Hjelmen et al. was subtracted from 1. A regression analysis was completed in R in which Difference between genome size and assembly size was the response variable and proportion of late-replicating heterochromatin was used as the predictor variable.

R script information

Relevant R scripts can be found on a GitHub repository (https://github.com/cehjelmen/gs_assessment doi: 10.5281/zenodo.11506496).

Results

Assemblies and genome size information

The R script was identified and download information for 41,358 genome assemblies from NCBI. This assemblies corresponded to the 3 large Kingdoms (Metazoa, Viridiplantae, and Fungi), as well as other smaller classifications above the Phylum classification (referred to from here on as "Other"). More specifically, information was retrieved for 56 phyla, 204 classes, 731 orders, and 2,387 families. While 41,358 assemblies are available, only about 36.4% (15,133) of these assemblies have corresponding genome size records available on the online genome size databases. These records span all 3 major Kingdoms, as well as 28 phyla, 100 classes, and 336 orders. By Kingdom, there are 7,530 records across 480 species of Fungi, 4,938 records for 1,169 species of Metazoa, and 2,625 for 660 species of Viridiplantae (Table 1, Supplementary Fig. 1). Unsurprisingly, some model organisms had high representation of assemblies in the database. For example, of the 7,530 records in Fungi, 1,913 were of the model species *Saccharomyces cerevisiae*. In fact, 13 of the 20 top represented species of assemblies with genome size estimates were Fungi, whereas 3 were metazoans (*Homo sapiens*, *Drosophila melanogaster*, and *Mus musculus*) and 4 were from Viridiplantae (*Arabidopsis thaliana*, *Hordeum vulgare*, *Zea mays*, and *Oryza sativa*; Supplementary Fig. 2).

Table 1. Summary of number of assemblies with genome size estimates.

Kingdom	Phylum	Count
Fungi ($n = 7,530$)	Ascomycota	6,416
	Basidiomycota	889
	Blastocladiomycota	2
	Chytridiomycota	8
	Cryptomycota	2
	Microsporidia	30
	Mucoromycota	178
	Zoopagomycota	5
Metazoa ($n = 4,938$)	Annelida	14
	Arthropoda	1,113
	Brachiopoda	2
	Chordata	3,427
	Cnidaria	25
	Ctenophora	1
	Echinodermata	23
	Mollusca	92
	Nematoda	175
	Onychophora	2
	Placozoa	1
	Platyhelminthes	28
	Porifera	11
	Rotifera	21
	Tardigrada	3
Viridiplantae ($n = 2,625$)	Chlorophyta	28
	Streptophyta	2,597
Other ($n = 40$)	Bacillariophyta	3
	No Phylum	8
	Rhodophyta	29

Counts of assemblies with corresponding values of genome size estimates for the species in an online genome size database. Counts are displayed for each Phylum and Kingdom.

Assembly assessment by Kingdom

Of the 15,133 genomes, 7,310 (48.3%) are within 10% of the genome size estimation for that species (between −0.1 and 0.1) and 6,538 (43.2%) were above 0.1 (Fig. 1). Positive values indicate assemblies smaller than the estimated genome size and negative values indicate assemblies larger than the estimated genome size.

When broken down by Kingdom, 18.17% of metazoan assemblies, 14.51% of assemblies within Viridiplantae, and 79.93% of assemblies of Fungi are within 10% of the estimate (Supplementary Fig. 3). A visualization of the distribution of "Proportional Difference" can be found in Fig. 1. When investigating the distribution of "Proportional Difference" across Kingdoms, the mean value in Fungi is −0.00066, whereas the mean values are 0.1609 and 0.3321 in Metazoa and Viridiplantae, respectively. This indicates that the average Fungi assembly size is at the genome size estimate, whereas assemblies in either Metazoa or Viridiplantae are likely to be smaller than the genome size estimate. It is important to note that Fungi, on average (31.63 Mbp average), have smaller genome sizes than species in either Metazoa (2,215 Mbp average) or Viridiplantae (2,587 Mbp average; Fig. 2). The distributions of genome sizes in these Kingdoms are found to be statistically significantly different from each other Kingdom according to a Kruskal–Wallis and Dunn's test with a Bonferroni correction ($P < 0.0001$)

A regression analysis finds that there is a significant positive relationship between log(genome size) and the absolute value of "Proportion Difference" (Adj. $R^2 = 0.08248$, $P < 0.0001$), indicating that smaller genomes are more likely to be assembled closer to the genome size estimate for that species and larger genomes are more likely to deviate (Fig. 2).

Assembly assessment by sequencing methodology

The data were subset to include information from assemblies which indicated they were performed by Illumina, Oxford Nanopore Technology, or PacBio SMRT sequencing platforms. Of the 15,133 assemblies with corresponding genome size estimates, 7,770 were performed with Illumina (51.7%), 609 were from Oxford Nanopore (4.02%), and 1,627 were from PacBio SMRT sequencing (10.75%). Assemblies from Illumina sequencing had a mean "Proportion Difference" of 0.124, while assemblies from Oxford Nanopore had mean values of 0.129 and those from PacBio

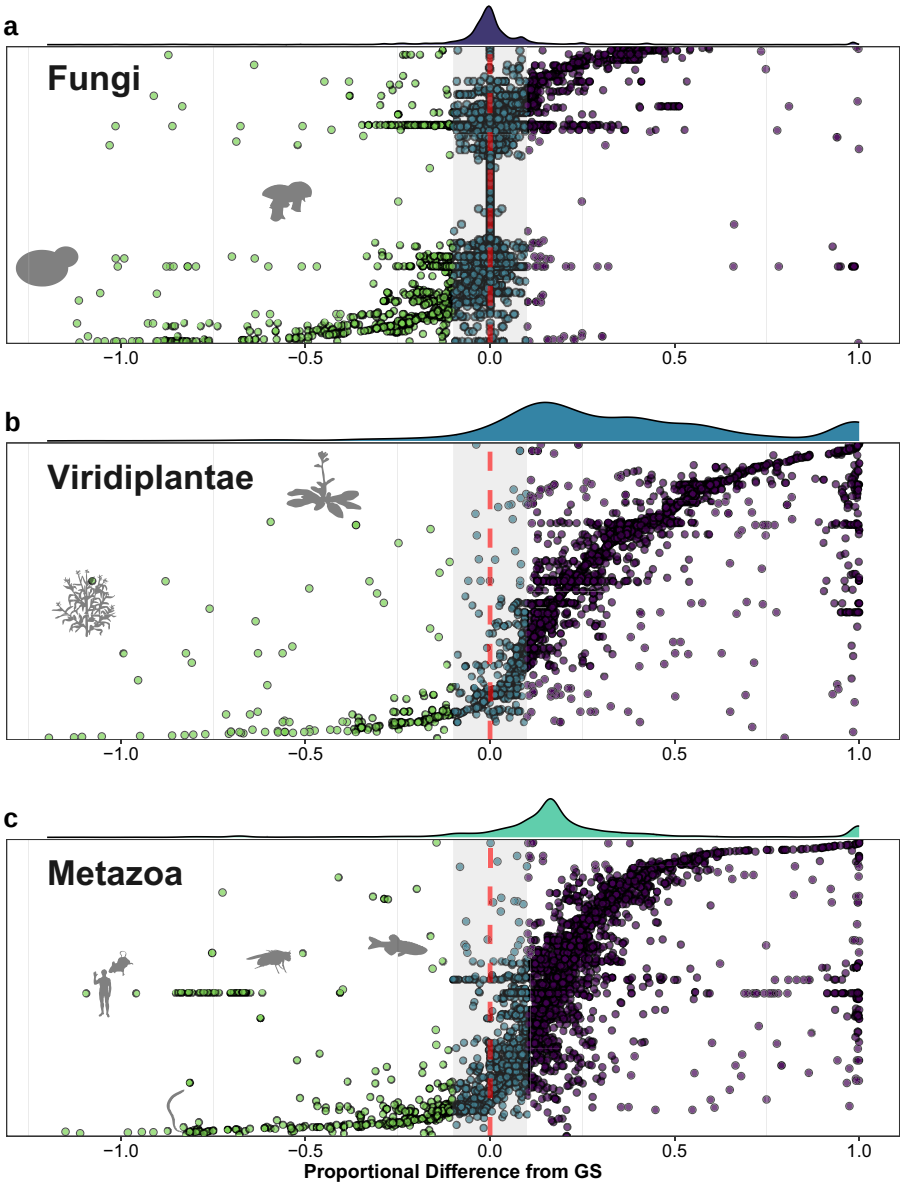

Fig. 1. Proportional difference in assembly size from genome size (GS) across 15,133 records. Each point represents a genome assembly, with the vertical axis representing each species across Fungi a), Viridiplantae b), and Metazoa c). Values are calculated as (genome size – assembly size)/(genome size) to account for proportional differences rather than absolute differences. This proportional difference allows a more relative comparison across a wide range of organisms and GSs. Negative values correspond to genomes that are assembled larger than the corresponding estimate for the species, while positive values indicate the assembled size is smaller than the estimated size. The gray shaded area represents a ±0.10 around 0, indicating a 10% variation above or below the size. This value was selected arbitrarily. Assemblies within this 10% range are within the shaded region and are colored blue, while genomes larger than the estimated size are in the left half of the plot and are colored green, and values under the 10% value are in the right portion of the plot and filled in purple. The x-axis limits were set to −1.2 to +1 to represent the vast majority of genomes (some assemblies were far lower than −1.2). Marginal distribution density plots are positioned above each panel to aid in visualization of the distribution of Proportional Difference scores for each Kingdom. PhyloPic silhouettes for model and well-studied organisms were placed near their representative assemblies (top to bottom: *Amanita* mushroom, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Zea mays*, *Caenorhabditis elegans*, *Mus musculus*, *Homo sapiens*, *Danio rerio*, and *Drosophila melanogaster*).

SMRT had mean values of 0.142. Of the assemblies using Illumina sequencing, 52.63% were within 10% of the genome size estimate, where Oxford Nanopore had 31.86% and PacBio SMRT had 28.33% within 10%.

When assessed with a Kruskal–Wallis and Dunn's post hoc test with Bonferroni correction, there were significant differences between distributions of "Proportion Difference" between assemblies from each sequencing platform, with Illumina's distribution of "Proportion Difference" aligning more closely with genome size estimates (Dunn's test, Illumina $P < 0.00001$ from either, and $P = 0.0245$ for Oxford Nanopore vs PacBio SMRT; Fig. 3a).

However, as Fungi are noted to have significantly smaller genome sizes and "Proportion Difference" than species in Metazoa or Viridiplantae, an additional comparison between sequencing technologies was made excluding species within Fungi. When Fungi are excluded, the mean "Proportion Difference" in Illumina is 0.274, while Oxford Nanopore and PacBio SMRT have mean values of 0.172 and 0.213, respectively. When excluding Fungi, 14.47% of assemblies from Illumina sequencing were within 10% of the genome size estimate, while 19.51% of Oxford Nanopore assemblies and 20.7% of PacBio SMRT assemblies where within 10% of the genome size estimate. When assessed

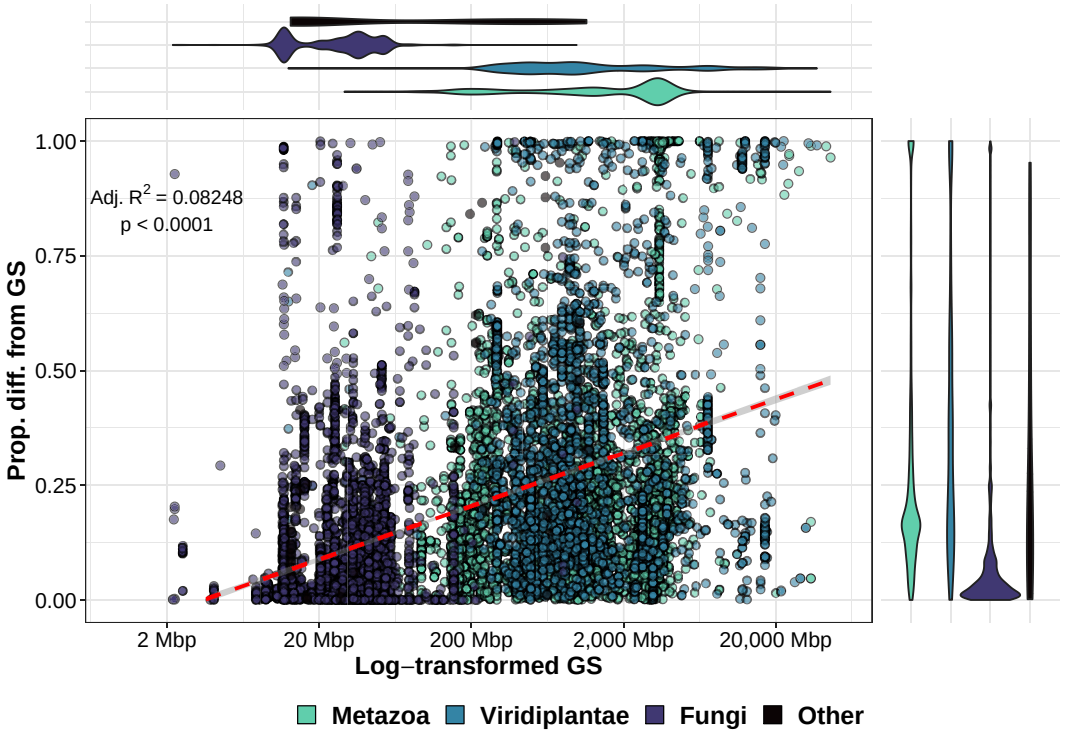

Fig. 2. Significant relationships between genome size (GS) and "Proportion Difference" as well as GS differences by Kingdom. There is a significant relationship between the log of GS and the absolute value of "Proportion Difference" in assembly size from estimated GS according to a linear regression model. GS is log-transformed, with the axis labels in true GS (Mbp). Colors of dots in plot indicate Kingdom for the assembly. There are significant differences in GSs between each of the Kingdoms of Eukarotya, in which Fungi are significantly smaller than Metazoa and Viridiplantae. These differences in GS are visualized by the violin plot above the scatterplot. The distribution of "Proportion Difference" is plotted in the violin plot positioned to the right of the scatterplot. Colors of violin plots indicate Kingdom.

with Kruskal–Wallis and Dunn's test, there are significant differences in the distribution of Illumina from either long-read technology (Dunn's test, $P < 0.0001$), but there is not a significant difference between the distribution of Oxford Nanopore or PacBio SMRT sequencing (Dunn's test, $P = 0.9399$; Fig. 3b).

Analysis of late-replicating heterochromatin

Drosophila species provide a unique opportunity to investigate the proportion of the genome which is considered to be late-replicating heterochromatin as they undergo 1 round of incomplete replication, or underreplication, in their thoracic tissues (Johnston *et al.* 2013, 2020). Data on underreplication estimates from *Drosophila* retrieved from previous work (Hjelmen *et al.* 2020) was subset to species of *Drosophila* which had estimates of underreplication proportion and genome size. In order to calculate proportion of late-replicating heterochromatin, percent underreplication (how much is additionally replicated) reported in Hjelmen et al. was subtracted from 1. When comparing the raw difference between estimated genome size and assembly size in these species (Fig. 4a) with the proportion of the genome which is estimated to be late-replicating heterochromatin, there was found to be a significant positive relationship (Adj. $R^2 = 0.333$, $P < 0.0001$; Fig. 4b). When a genome has a higher proportion of late-replicating heterochromatin, the difference between the estimated genome size and the assembly size is greater.

Analysis of standard genome assembly assessments

In order to assess the quality of standard genome assembly assessments, such as Contig N50, Scaffold N50, and BUSCO scores, regression analyses were completed between these metrics and the absolute value of the "Proportion Difference". In each of these regression analyses, "Proportion Difference" was the response variable, where the assembly assessment was the predictor variable. There is a moderately significant negative relationships between Contig N50 and the absolute value of Proportion Difference, but with very little explanatory value (Adj. $R^2 = 0.0003242$, $P = 0.034$; Table 2). No significant relationship was found between Scaffold N50 and "Proportion Difference". There was a significant negative relationship between complete BUSCO score and "Proportion Difference" (Adj. $R^2 = 0.016$, $P = 0.00423$; Table 2). Visualization of "Proportion Difference" vs each of these assessments can be found in Fig. 5a–c.

As the values of Contig and Scaffold N50 are in raw base sizes, it may be appropriate to make an assessment which is more specific to the genome architecture of the organism that is being sequenced, a metric proposed here as PN50, or "Proportional N50". In order to address this, Contig and Scaffold N50 values were divided by an estimated measure of the average chromosome size (genome size/haploid chromosome number). When this new measurement was placed as the predictor variable in a regression analysis, there were highly significant relationships with "Proportion Difference" ($P < 0.00001$) with higher Adj. R^2 values for either analysis (Table 2). Visualization of the performance of this PN50 assessment metric can be found in Fig. 5d, e, with direct visual comparisons to the standard Contig and Scaffold N50s above in 5a and 5b. Additionally, assembly assessment information, including PN50 values, from reference quality assemblies for *H. sapiens*, *Gorilla gorilla*, *Panthera tigris*, *Drosophila pseudoobscura*, *Anopheles gambiae*, *Z. mays*, and *O. sativa* can be found in Table 3.

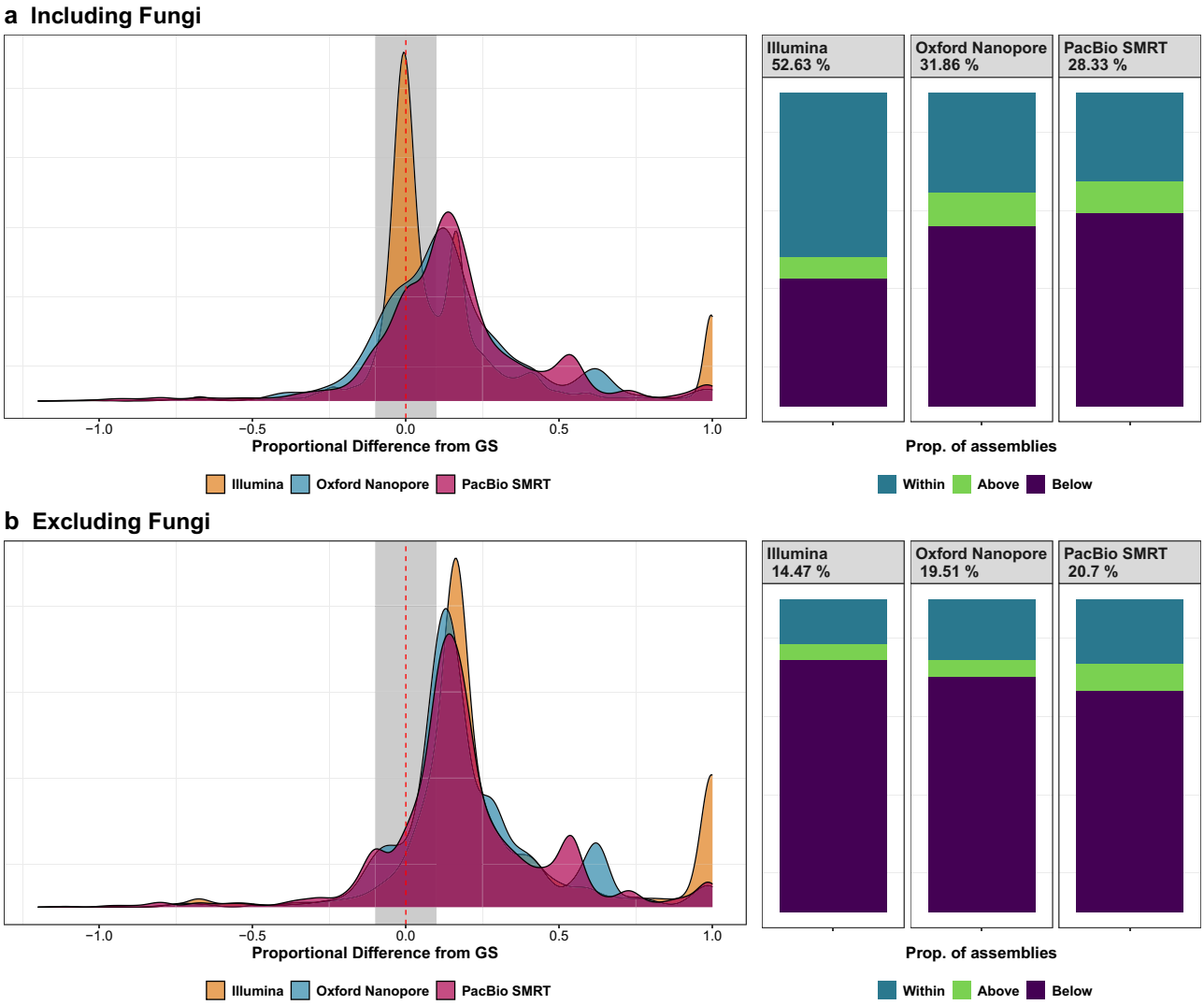

Fig. 3. Distributions of "Proportional Difference" of assemblies by sequencing platform. When all data are included for Eukaryota in which Illumina (orange), Oxford Nanopore Technology (blue), or PacBio SMRT (magenta) sequencing were used a), there is a significant difference in the distribution of Illumina genomes compared to Oxford Nanopore and PacBio genome assemblies. The bars on the right correspond to the proportion of genomes assembled from each sequencing platform as being within 10% of the genome size (GS) estimate (blue, top portion of stacked barplot), greater than GS estimate by over 10% (green, middle portion), or below the estimate by greater than 10% (purple, bottom portion). Percentages on the bars indicate the proportions which are within 10% of the GS estimate. When Fungi are removed from the analysis b), there is a significant difference between all platforms, but Illumina has a more positive value, indicating more assemblies lower than the GS estimate. This is reflected in the proportion of genome assemblies within 10% of the estimate (blue, top portion of stacked barplot) vs greater than (green, middle portion) and smaller than (purple, bottom portion) in the bar charts on the right.

Discussion

Number of records with genome size estimates

While we have made incredible progress in our efforts to sequence as many species as possible, with 41,358 eukaryotic assemblies on NCBI's Assembly database as of 2024 February 20, it is astounding how few of these assemblies have corresponding genome size estimations available on the genome size databases. Only 36.4% of assemblies have a corresponding species record on the animal, plant, or fungi genome size databases (Supplementary Fig. 1). And while 15,133 assemblies are quite a few, much of this quantity is due to high numbers of assemblies for model organisms (*S. cerevisiae*, *M. musculus*, *A. thaliana*, *D. melanogaster*, etc.). These 15,133 assemblies with genome size estimations only represent 2,309 unique species. It is possible that some records have strain information in their taxonomic entry, which will not match identically with the species name in the respective genome size databases. Conversely, it is possible that we have genome size estimations for a species that were not used to inform an assembly. Regardless, we must do better at reporting genome sizes with our genome assemblies, both to inform future studies, but also to ensure we have appropriately assembled genomes available to the public.

It is important to note that not all genome size estimates have been uploaded to these databases, and that it is possible that some of these 63.6% of genome assemblies had corresponding estimates in their works. Just as it is important to upload raw reads to databases such as NCBI's SRA, we should also make efforts to upload these data to the corresponding genome size databases as well. This study makes it clear that there is an abundance of genome size estimation records across these databases, with 23,447 records across 19,577 species. While this study has emphasized that many assemblies do not have corresponding genome size estimations, the number of species with estimates exceeds the number of species with assemblies with genome size

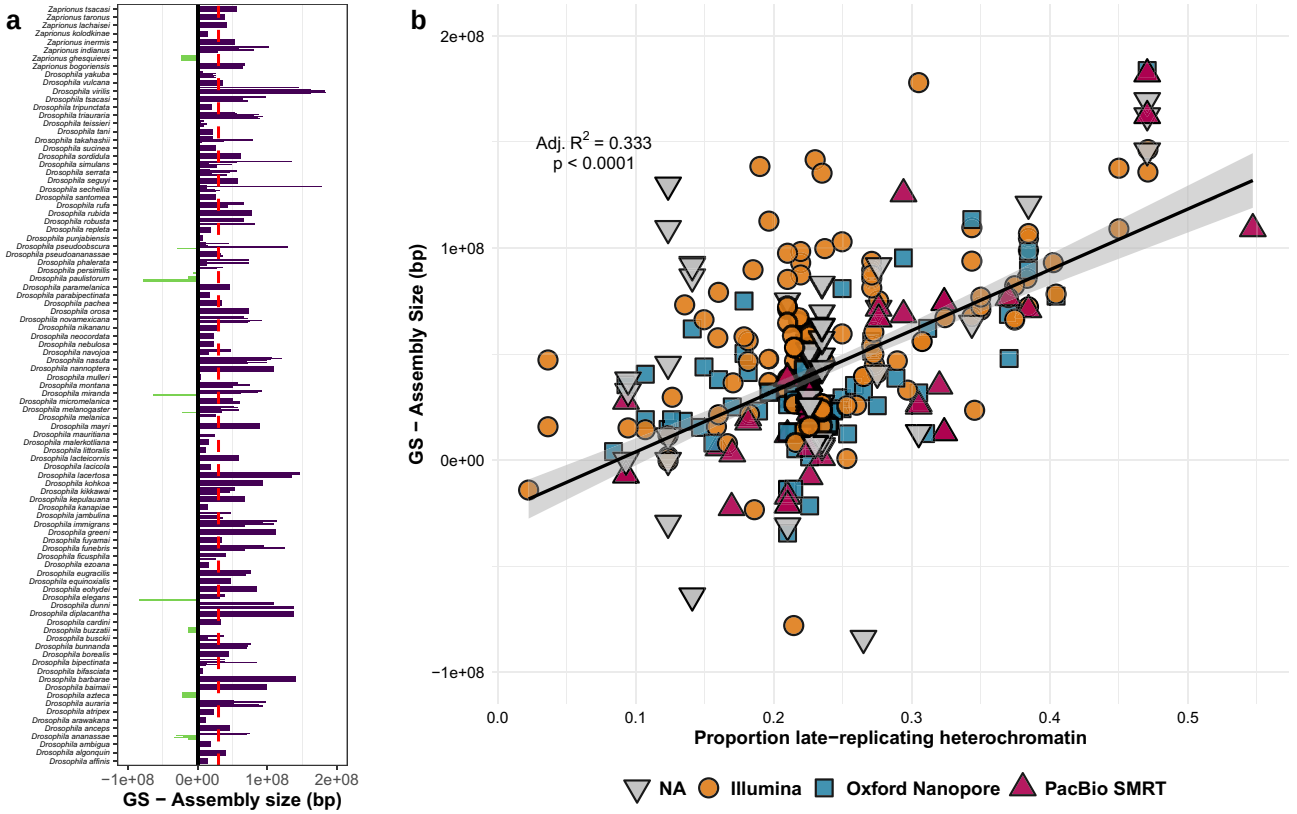

Fig. 4. *Drosophila* genomes suggest a strong relationship between proportion late-replicating heterochromatin and difference in assembly size from genome size (GS). a) Genome size – assembly size (bp) for *Drosophila* species. Purple bars on the right represents genome smaller than estimates; green bars on the left represents genome assemblies larger than estimates. The vertical, dashed, red line indicates 30 Mbp under the GS estimate. b) There is a strong relationship between percent late-replicating heterochromatin and raw difference in GS in *Drosophila* according to a linear regression model in which the difference in assembly size from GS is the response variable.

Table 2. Results of linear regression analyses for genome assembly assessments.

Assessment	Coefficient	*F*-statistic	df	Adj. R^2 value	*P*-value
Contig N50	−5.137E-10	4.25	10004	0.0003242	0.0394
Scaffold N50	−6.708E-13	0.00	10004	0.000	0.9882
BUSCO	−0.791	8.27	456	0.016	0.00423
Contig PN50	−0.311	67.48	2092	0.031	<0.00001
Scaffold PN50	−0.302	190.40	2092	0.083	<0.00001

The absolute value of "Proportion Difference" in assembly size from genome size was used as the response variable in each analysis, in which standard assessments of genome assembly (Contig N50, Scaffold N50, and BUSCO) were used as predictor variables. Additionally, new suggested metrics of Contig and Scaffold PN50 (N50/(genome size/haploid chromosome number)) were used as predictor variables in linear regression analyses.

estimations. This indicates that studies on genome size variation and evolution often include many species which are not selected for genome sequencing. We must also keep in mind that these databases are often maintained by a small number of individuals and are often not funded. Therefore, we must do our best to cite these databases appropriately when their data are used.

Assembly size vs estimated genome size

A value of 10% was arbitrarily selected as an acceptable proportion difference from the estimated genome size for this study. This value of 10% was selected, as was the use of proportions rather than absolute differences, to serve as a point of comparison across assemblies. When reported assembly sizes are compared to genome size estimates for each species, it is not common that an assembly comes within 10% of the estimated genome size (48.3% of total assemblies, Fig. 1). This value is biased by the number of assemblies that are within 10% of the estimate in Fungi (79.93%), compared to Metazoa (18.17%), and Viridiplantae (14.51%). As there is a significant negative positive relationship between genome size and "Proportion Difference" in assembly size from genome size (Fig. 2), the higher level of assemblies that are within 10% of the estimate for Fungi is likely due to the smaller genome size of Fungi. It should also be considered that *S. cerevisiae* was the first complete eukaryotic genome (reviewed in Engel et al. 2014), and that nearly 2,000 of the 7,530 assemblies for Fungi are from this model species (Table 1, Supplementary Fig. 1). Additionally, in part to their small genome sizes and medical importance, Fungi are highly represented in terms of number of genome assemblies (Table 1) but are not particularly speciose in terms of the distribution of the genome sequenced for that group. It is important to note that most assemblies that deviate from the estimate genome size are smaller than the estimated size,

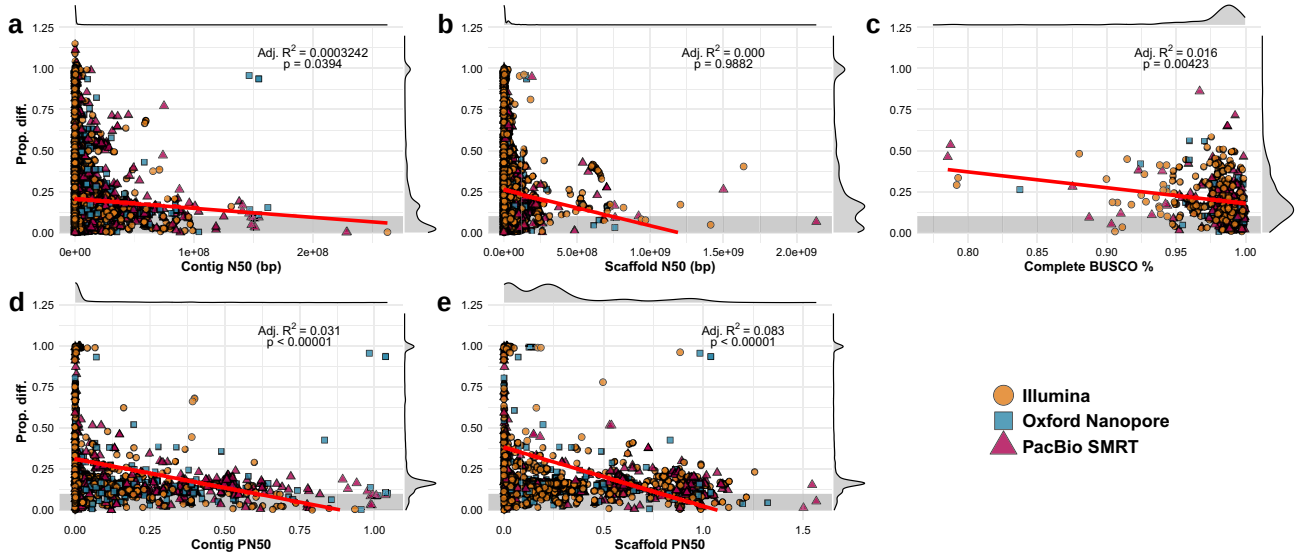

Fig. 5. Standard assessment metrics for genome assembly compared to "Proportion Difference" in assembly size from genome size when sequenced by Illumina Oxford Nanopore or PacBio SMRT sequencing platforms. The absolute value of "Proportion Difference" in assembly size from genome size was used as the response variable in linear regression analyses, in which standard assessments of genome assembly (Contig N50 a), Scaffold N50 b), and BUSCO c)) were used as predictor variables. Additionally, new suggested metrics of Contig PN50 d) and Scaffold PN50 e) [N50 values divided by average chromosome size (genome size/haploid chromosome number)] were used as predictor variables in linear regression analyses. *P*-values and R^2 values for each analysis are annotated on the scatterplots. Colors and shape of points indicate sequencing platform for the corresponding assembly. The gray shaded area indicates values of genomes within 10% above or below the available genome size estimate for each species. Marginal distribution density plots positioned above and to the right of each scatterplot to aid in visualization of the distribution of plots for the assessment metric (above scatterplot) and "Proportion Difference" (right of scatterplot). It is important to note slight differences in the distributions of Proportion Difference between N50 and PN50 plots due to fewer assemblies with corresponding chromosome numbers needed for the PN50 calculation (10,006 N50 values vs 2,092 PN50 values).

Table 3. Examples of PN50 scores for reference quality assemblies.

Species	Accession number	Avg GS (Mbp)	Haploid number	Contig N50 (Mbp)	Scaffold N50 (Mbp)	Contig PN50	Scaffold PN50	References
Homo sapiens	GCA_000001405.15	3423	23	57.90	67.80	0.389	0.456	Schneider *et al.* (2017)
Homo sapiens	GCA_009914755.3	3423	23	154.26	154.26	1.037	1.037	Nurk *et al.* (2022)
Gorilla gorilla	GCA_029281585.2	3668	24	150.80	150.80	0.987	0.987	Makova *et al.* (2024)
Panthera tigris	GCA_024034525.1	2650	19	0.15	145.23	0.001	1.041	Armstrong *et al.* (2022)
Drosophila pseudoobscura	GCA_009870125.1	164	5	30.71	32.42	0.935	0.987	Liao *et al.* (2021)
Anopheles gambiae	GCA_943734675.1	264	3	10.63	99.15	0.121	1.126	Habtewold *et al.* (2024)
Zea mays	GCA_026229685.1	2400	10	0.65	223.22	0.003	0.930	Wang, Hou *et al.* (2023)
Oryza sativa	GCA_001623345.3	430	12	32.12	32.12	0.896	0.896	Song *et al.* (2021)

Assemblies from a select few reference quality assemblies across Metazoa and Viridiplantae, including their NCBI assembly accession numbers, average genome size estimates for each species, haploid number, Contig N50, Scaffold N50, as well as calculated PN50 values. For *Homo sapiens*, the first listed assembly is the GRCh38 reference while the second listed assembly is the recently published telomere-to-telomere assembly (T2T-CHM13). There is a notable increase in the Scaffold PN50 with the updated assembly.

indicating that assemblies are systematically assembled to be smaller, or that genome size estimations are overestimating the size of genomes in a way that scales with genome size. However, it seems to be more likely that assemblies are missing information, for the reasons discussed below.

It is not surprising to see that larger genomes are more likely to deviate more strongly in assembly size from the estimated genome size (Fig. 2), as we know that larger eukaryotic genomes tend to have larger proportions of repetitive content and "junk DNA" (Makalowski 2003; Biémont and Vieira 2006; Doolittle 2013; Palazzo and Gregory 2014). It stands to reason that this proportion of the genome is more difficult to assemble due to its repetitive nature. However, repetitive DNA is not just restricted to "large" genomes. For example, if we are to look at the *nannoptera* group of *Drosophila*, we have recently reported assembly sizes of 134.5 Mbp *for D. nannoptera*, 127.52 Mbp for *D. pachea*, and 130.4 Mbp for *D. acanthoptera* (Kim *et al.* 2023). However, genome sizes for females of these species determined through flow cytometry are reported to be 245.9 Mbp, 176.5 Mbp, and 155.8 Mbp for *D. nannoptera*, *D. pachea*, and *D. acanthoptera*, respectively (Hjelmen *et al.* 2019). Unsurprisingly, early karyotypic studies of this group noted them to have highly heterochromatic chromosomes (Ward and Heed 1970), and follow-up work using estimations of underreplication of heterochromatin found that they did not have large

genomes, but uniquely heterochromatic genomes (Hjelmen *et al.* 2020). When expanded to all species of the *Drosophila* genus which have genome size estimates, assemblies, and estimations of thoracic underreplication, there is a statistically significant positive relationship between the proportion of the genome which is estimated to be late-replicating heterochromatin and the proportion difference between assembly size and estimated genome size (Fig. 4). When there is more heterochromatin in the genome, there is a higher chance of DNA missing in the assembly. If we are to use information on structural variation in the genome for informing our studies on populations, adaptation, evolution, and even health, we must be sure to prioritize including this often lost information.

Assembly performance by sequencing platform

Assembly performance, in terms of proportion difference of assembly size from genome size, was compared across the 3 most used sequencing technologies (Illumina, PacBio SMRT, and Oxford Nanopore Technologies). Upon a first glance, it seems as though Illumina does exceptionally well in terms of assemblies of genomes within 10% of their genome size estimate, with 52.63% of Illumina assemblies falling within 10% above or below the genome size estimate (Fig. 3a). The difference in performance between sequencing platforms is significant according to a Kruskal–Wallis test and a Dunn's post hoc test ($P < 0.0001$). Short, accurate read sequencing, such as Illumina, are likely used more often for reference-based assemblies, which could increase the contiguity of the assemblies compared to a de novo assembly with Illumina. Given that thought, it is possible that an abundance of assemblies for a few well-studied organisms could bias the results of Illumina due to higher instances of reference-based assemblies (Supplementary Figs. 2 and 3). As we are working toward sequencing and assembling genomes for all of life, it would be best to prioritize de novo assemblies (Consortium 2013; Zhang 2015; Lewin *et al.* 2018, 2022; Coddington *et al.* 2019; The Darwin Tree of Life Project Consortium 2022).

We must consider the strong relationship between genome size and proportion difference in assembly size. Smaller genomes are easier to assemble. A large proportion of the assemblies provided, especially for Illumina, are from Fungi (Supplementary Fig. 4), which have significantly smaller genome sizes on average compared to the 2 other large Kingdoms within Eukaryota (Kruskal–Wallis and Dunn's test, $P < 0.0001$, Fig. 2). When fungal assemblies are removed, Illumina's performance decreases and underperforms both PacBio SMRT and Oxford Nanopore Technology, with only 14.47% of assemblies within 10% above or below the estimated size, compared to 19.51% for Oxford Nanopore Technology and 20.7% for PacBio SMRT sequencing (Fig. 3b). Illumina's distribution is found to be significantly different from PacBio SMRT and Oxford Nanopore Technology, but there was no significant difference between the distributions of PacBio and Oxford Nanopore. Given this, it seems that long-read technologies tend to assemble proportionally more genomes within 10% of the genome size estimate. A larger sample size, both in terms of genome size estimations and assemblies, may start to tease apart slight differences in performance for Oxford Nanopore Technology and PacBio SMRT sequencing. For example, it was shown in a recent study that prioritizing high-long reads (PacBio HiFI) was more beneficial for assembly than ultra-long reads (PacBio CLR and Oxford Nanopore Technology; Hotaling *et al.* 2023). For now, it seems that long-read platforms have a higher percentage of assemblies within a 10% difference of their estimated genome size.

Assembly assessments

As discussed above, we often quantify the quality of our genome, in a broad sense, with metrics of contiguity (N50 values) and completeness of core sets of genes (BUSCO; Baker 2012; Simão *et al.* 2015; Thrash *et al.* 2020; Manni *et al.* 2021b). Hopefully it is clear by this point that we should also be considering the fundamental structure of our genome in these metrics, both in terms of total genome size, and in terms of our chromosome number. We have made leaps and bounds of progress in terms of chromosome-level assemblies, thanks to long-read sequencing technologies and other assistance, such as Hi-C and optical mapping; however, it is not clear how much of our data are missing on these chromosomes (Zhang *et al.* 2019; Giani *et al.* 2020; Kronenberg *et al.* 2021; Rhie *et al.* 2021; Kong *et al.* 2023; Wang, Yu *et al.* 2023). As discussed in the introduction, even our most "complete" genomes, such as the human genome, were missing nearly 10% of its sequence (mostly due to repetitive rDNA on acrocentric chromosomes; Nurk *et al.* 2022). We must come up with a way to assess our genomes at this fundamental level.

Standard genome assembly metrics, such as Contig N50 (but not Scaffold N50) and BUSCO do have significant negative relationships with our "Proportion Difference" in assembly size from genome size; however, they only explain a minute proportion of the variation in the data (Fig. 5, Table 2). Broadly speaking, there is a relationship between a larger N50 value and a higher BUSCO score and a lower deviation from our estimated genome size in our assembly, but much of the variation is still amorphous (Fig. 5a–c). We can then pose the question of "what is a good N50 value?". This could be up for debate, as discussed in the introduction, as genomes are extremely variable in size, and therefore 15 Mbp maybe excellent for some smaller genomes, but a cause for concern in large genomes. When comparing BUSCO values of our assemblies to the "Proportion Difference" values, there is evidence for a negative relationship; but the variation around the trendline suggests that it may not be entirely informative. It must also be noted that only assemblies with BUSCO values reported in the Assembly metadata from NCBI were included, which dramatically reduced the sample size for that metric (458 genomes for BUSCO vs 10,006 assemblies for contig N50). As discussed in the introduction, we should be wary of some of the limitations of BUSCO, as it is only assessing the presence of genes present in single copies that are present in 90% of genomes at that taxonomic rank. There is a lot of room for missing information. Overall, it seems our current metrics do indicate some support for a more complete genome assembly, but not necessarily overwhelmingly so.

If we consider the fundamental structure of our genome: our total size is split between a set number of chromosomes, we can increase and refine our ability to assess our genome assemblies utilizing a "Proportional" N50 value (PN50; equation 1, Fig. 5d, e).

$$\mathrm{PN50} = \frac{\mathrm{N50}}{\left(\frac{\mathrm{GS}}{\mathrm{HapNum}}\right)} \quad (1)$$

In order to calculate this metric, the N50 value (scaffold or contig) is simply divided by the "average chromosome size", which is calculated by dividing the haploid genome size by the haploid chromosome number. This calculation will convert the N50 values to proportional values, which have the added benefit of being more universal and comparable across organisms than a raw metric is. Additionally, the N50 metric is now being assessed

at the level of its size related to a functional unit of the genome (an estimated chromosome size) rather than an arbitrary length. Values are now proportional, with higher proportions of our assemblies within 10% of our estimated size when we have values of 0.5–1 (Fig. 5d, e). When this assessment metric is compared to the genome assemblies here, there is a strong significant, negative relationship between this metric and the "Proportion Difference" between genome size and assembly size. Not only is the *P*-value more significant in PN50 than the raw N50 values, but the corresponding adjusted R^2 value increases with the PN50 metric as well (Table 2). It seems that accounting for this fundamental structural unit can help us tease out genome assemblies which are contiguous and approaching the reported genome sizes. When this metric is applied to recent reference quality assemblies, we see Scaffold PN50 values approaching one, suggesting their Scaffold N50 values are approximately the average chromosome size for that organism (Table 3). It is interesting to compare the PN50 assessments for the 2 included assemblies for *H. sapiens*, the first being the commonly used GRCh38 reference (Schneider *et al.* 2017) and the second being the more recently published telomere-to-telomere assembly (T2T-CHM13; Nurk *et al.* 2022). There was a large increase in the Scaffold PN50 score from the GRCh38 assembly to the T2T-CHM13 assembly, from 0.456 to 1.037, suggesting the new T2T-CHM13 assembly has a scaffold N50 value that is approximately the size of the average human chromosome (Table 3).

Estimating genome size

If genome size is such an important metric, how do we go about getting proper estimates for our species of interest? Firstly, the recommendation would not be to utilize information from a closely related species; there is far too much variation among species, even with taxonomic ranks such as genus (Supplementary Fig. 5). In the *Drosophila* genus, there are genomes as small as 134.7 Mbp in *D. busckii* but as large as 342.7 Mbp in *D. suzukii* and 330.9 Mbp in *D. pallidipennis* (Hjelmen *et al.* 2019). In one of the largest studies of genome size variation at the genus level, it was found that there was nearly 4-fold variation at the genus level of *Synalpheus* shrimp (Jeffery *et al.* 2016). We clearly should not rely on estimates from the same genus or taxonomic ranks higher than genus.

If we cannot use previously estimated values for genus level or higher, we might be able to use previous values for the species. However, we know there are significant differences even at the species level. For example, in the *Drosophila* Genetic Reference Panel, 205 isolines were developed from a single collection event of *D. melanogaster* in Raleigh, NC, USA (Ellis *et al.* 2014; Huang *et al.* 2014). These lines varied in genome size from 169.7 to 192.8 Mbp, with a mean of these strains close to the standard estimate for *D. melanogaster* (175.6 Mbp). Interestingly, a there was a significant relationship between the variation in genome size in these strains and inversions, a common structural variant we might want to investigate across populations in assembled genomes (Huang *et al.* 2014). Outside of *Drosophila*, there is evidence for intraspecific variation in genome size for *Z. mays* (Bilinski *et al.* 2018), eyebrights (Becher *et al.* 2021), copepods (Leinaas *et al.* 2016), snapping shrimp (Jeffery *et al.* 2016), *Callosobruchus* seed beetles (Arnqvist *et al.* 2015), and countless others. This evidence suggests that we should also consider getting new estimates of genome size for any individual we want to sequence.

If we are to get new genome size estimates whenever possible, how do we go about doing this? Clearly, we cannot yet rely on the algorithms in genome assembly software. There have also been notable issues with accuracy of estimates utilizing k-mer methods of genome size estimation (Liu *et al.* 2013; Pflug *et al.* 2020, but see Hesse 2023). There are a few tried- and true methods for genome size estimations, but 2 are consistently and commonly used and considered the gold standard: flow cytometry and Feulgen densitometry. Detailed methods for flow cytometric genome size estimations in plants can be found at Pellicer and Leitch (2014), and at Johnston *et al.* (2019) for insects and other animals. Methods for Feulgen estimation are detailed in Jeffery and Gregory (2014) and Hardie *et al.* (2002). While these methods are not considered to be too problematic or difficult, they often require expensive specialized equipment and expert personnel for fine-tuned estimates. Often the best route for genome size estimations is to reach out to those who publish large quantities of estimates and/or methods chapters to inquire who may provide genome size estimations as a service for a modest cost or as part of a collaboration. These collaborations should include consideration for funding for estimating genome size of samples, which is relatively cheap per sample, and will provide the benefit of higher confidence in the final assembly of your organism's genome.

Conclusions

From this study, it is clear that we still need to work to gather and incorporate foundational architectural information about genomes we want to assemble. While sequencing technologies and assembly algorithms are improving, we cannot fully rely on them just yet for the most accurate assemblies. We must gather information on genome size, as well as chromosome number, in order to make the most informed genome assemblies. This study therefore proposes the use of the new metric "PN50", or proportional N50 value, a metric which is more universal and comparable across organisms than raw N50 values. This information, in conjunction with other assembly metrics, can allow us to be more confident in the end results, especially when it comes to noncoding and structural components of our genome assemblies.

Data availability

All data are publicly available through NCBI databases, the cited genome size databases, and the cited karyotype databases. All R scripts are available on a GitHub repository (https://github.com/cehjelmen/gs_assessment DOI: 10.5281/zenodo.11506496).

Supplemental material available at GENETICS online.

Acknowledgments

I would like to acknowledge the immensely helpful comments from my colleagues Geoffrey Zahn at Utah Valley University, Amely Bauer at University of Florida, and 2 of my closest mentors, J. Spencer Johnston and Heath Blackmon. I would also like to acknowledge my Fall 2022 and 2023 undergraduate genomics courses. Without their questions, discussions, and excitement for learning, I may not have been inspired to pursue this study. I would also like to thank the anonymous reviewers for their helpful comments on this manuscript and the UVU College of Science and Department of Biology for their continued support.

Funding

No funding was received for this work.

Conflicts of interest

The author(s) declare no conflict of interest.

Literature cited

25 Greatest impacts in 25 years: a look back at Illumina and the evolution of genomics. 2023. [accessed 2024 Mar 3]. https://www.illumina.com/company/news-center/feature-articles/25-greatest-impacts-in-25-years–a-look-back-at-illumina-and-the.html.

Adrion JR, Begun DJ, Hahn MW. 2019. Patterns of transposable element variation and clinality in *Drosophila*. Mol Ecol. 28(6): 1523–1536. doi:10.1111/mec.14961.

Ågren JA, Wright SI. 2011. Co-evolution between transposable elements and their hosts: a major factor in genome size evolution? Chromosome Res. 19(6):777–786. doi:10.1007/s10577-011-9229-0.

Alfsnes K, Leinaas HP, Hessen DO. 2017. Genome size in arthropods; different roles of phylogeny, habitat and life history in insects and crustaceans. Ecol Evol. 7(15):5939–5947. doi:10.1002/ece3.3163.

Armstrong EE, Campana MG, Solari KA, Morgan SR, Ryder OA, Naude VN, Samelius G, Sharma K, Hadly EA, Petrov DA. 2022. Genome report: chromosome-level draft assemblies of the snow leopard, African leopard, and tiger (*Panthera uncia, Panthera pardus pardus*, and *Panthera tigris*). G3 (Bethesda). 12(12):jkac277. doi:10.1093/g3journal/jkac277.

Arnqvist G, Sayadi A, Immonen E, Hotzy C, Rankin D, Tuda M, Hjelmen CE, Johnston JS. 2015. Genome size correlates with reproductive fitness in seed beetles. Proc Biol Sci. 282(1815): 20151421. doi:10.1098/rspb.2015.1421.

Attali D, Baker C. 2023. ggExtra: add marginal histograms to "ggplot2", and more "ggplot2" enhancements. R package version 0100. [accessed 2024 Feb 27]. https://CRAN.R-project.org/package=ggExtra.

Bainard JD, Newmaster SG, Budke JM. 2020. Genome size and endopolyploidy evolution across the moss phylogeny. Ann Bot. 125(4):543–555. doi:10.1093/aob/mcz194.

Baker M. 2012. De novo genome assembly: what every biologist should know. Nat Methods. 9(4):333–337. doi:10.1038/nmeth.1935.

Becher H, Powell RF, Brown MR, Metherell C, Pellicer J, Leitch IJ, Twyford AD. 2021. The nature of intraspecific and interspecific genome size variation in taxonomically complex eyebrights. Ann Bot. 128(5):639–651. doi:10.1093/aob/mcab102.

Biémont C, Vieira C. 2006. Genetics: junk DNA as an evolutionary force. Nature. 443(7111):521–524. doi:10.1038/443521a.

Bilinski P, Albert PS, Berg JJ, Birchler JA, Grote MN, Lorant A, Quezada J, Swarts K, Yang J, Ross-Ibarra J. 2018. Parallel altitudinal clines reveal trends in adaptive evolution of genome size in *Zea mays*. PLoS Genet. 14(5):e1007162. doi:10.1371/journal.pgen.1007162.

Blackmon H, Demuth JP. 2015. Coleoptera karyotype database. Coleopt Bull. 69(1):174–175. doi:10.1649/0010-065X-69.1.174.

Blaxter M. 2010. Revealing the dark matter of the genome. Science. 330(6012):1758–1759. doi:10.1126/science.1200700.

Clark J, Hidalgo O, Pellicer J, Liu H, Marquardt J, Robert Y, Christenhusz M, Zhang S, Gibby M, Leitch IJ, *et al.* 2016. Genome evolution of ferns: evidence for relative stasis of genome size across the fern phylogeny. New Phytol. 210(3):1072–1082. doi:10.1111/nph.13833.

Coddington J, Lewin H, Robinson G, Kress WJ. 2019. The earth biogenome project. Biodivers Inf Sci Stand. 3:e37344. doi:10.3897/biss.3.37344.

Comings DE. 1972. The structure and function of chromatin. In: Harris H, Hirschhorn K, editors. Advances in Human Genetics. Boston (MA): Springer. p. 237–431.

Consortium i5K. 2013. The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. J Hered. 104(5):595–600. doi:10.1093/jhered/est050.

Díez CM, Gaut BS, Meca E, Scheinvar E, Montes-Hernandez S, Eguiarte LE, Tenaillon MI. 2013. Genome size variation in wild and cultivated maize along altitudinal gradients. New Phytol. 199(1):264–276. doi:10.1111/nph.12247.

Dinno A. 2017. dunn. test: Dunn's test of multiple comparisons using rank sums. R package version 1: 1. [accessed 2024 Feb 27]. https://CRAN.R-project.org/package=dunn.test.

Doolittle WF. 2013. Is junk DNA bunk? A critique of ENCODE. Proc Natl Acad Sci U S A. 110(14):5294–5300. doi:10.1073/pnas.1221376110.

Elliott TA, Gregory TR. 2015. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. Philos Trans R Soc Lond B Biol Sci. 370(1678):20140331. doi:10.1098/rstb.2014.0331.

Ellis LL, Huang W, Quinn AM, Ahuja A, Alfrejd B, Gomez FE, Hjelmen CE, Moore KL, Mackay TFC, Johnston JS, *et al.* 2014. Intrapopulation genome size in *D. melanogaster* reflects life history variation and plasticity. PLoS Genet. 10(7):e1004522. doi:10.1371/journal.pgen.1004522.

Engel SR, Dietrich FS, Fisk DG, Binkley G, Balakrishnan R, Costanzo MC, Dwight SS, Hitz BC, Karra K, Nash RS, *et al.* 2014. The reference genome sequence of *Saccharomyces cerevisiae*: then and now. G3 (Bethesda). 4(3):389–398. doi:10.1534/g3.113.008995.

Etherington GJ, Heavens D, Baker D, Lister A, McNelly R, Garcia G, Clavijo B, Macaulay I, Haerty W, Di Palma F. 2020. Sequencing smart: de novo sequencing and assembly approaches for a non-model mammal. GigaScience. 9(5):giaa045. doi:10.1093/gigascience/giaa045.

Gamazon ER, Stranger BE. 2015. The impact of human copy number variation on gene expression. Brief Funct Genomics. 14(5): 352–357. doi:10.1093/bfgp/elv017.

Garnier S, Ross N, Rudis B, Sciaini M, Camargo PA, Scherer C. 2023. viridisLite: colorblind-friendly color maps for R. doi:10.5281/zenodo.4679423.

Gearty W, Jones LA. 2023. Rphylopic: an R package for fetching, transforming, and visualising PhyloPic silhouettes. Methods Ecol Evol. 14(11):2700–2708. doi:10.1111/2041-210X.14221.

Giani AM, Gallo GR, Gianfranceschi L, Formenti G. 2020. Long walk to genomics: history and current approaches to genome sequencing and assembly. Comput Struct Biotechnol J. 18:9–19. doi:10.1016/j.csbj.2019.11.002.

Girardini KN, Olthof AM, Kanadia RN. 2023. Introns: the "dark matter" of the eukaryotic genome. Front Genet. 14:1150212. doi:10.3389/fgene.2023.1150212.

Gregory TR. 2000. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol Rev. 76(1):65–101. doi:10.1017/S1464793100005595.

Gregory TR. 2001. The bigger the C-value, the larger the cell: genome size and red blood cell size in vertebrates. Blood Cells. Mol Dis. 27(5):830–843. doi:10.1006/bcmd.2001.0457.

Gregory TR. 2024. Animal genome size database. [accessed 2024 Feb 27]. http://www.genomesize.com.

Gregory TR, Hebert PDN. 1999. The modulation of DNA content: proximate causes and ultimate consequences. Genome Res. 9(4):317–324. doi:10.1101/gr.9.4.317.

Habtewold T, Wagah M, Tambwe MM, Moore S, Windbichler N, Christophides G, Johnson H, Heaton H, Collins J, Krasheninnikova K, *et al.* 2024. A chromosomal reference genome sequence for the malaria mosquito, *Anopheles gambiae*, Giles, 1902, Ifakara strain. Wellcome Open Res. 8:74. doi:10.12688/wellcomeopenres.18854.1.

Hardie DC, Gregory TR, Hebert PDN. 2002. From pixels to picograms: a beginners' guide to genome quantification by Feulgen image analysis densitometry. J Histochem Cytochem. 50(6):735–749. doi:10.1177/002215540205000601.

Hesse U. 2023. K-Mer-based genome size estimation in theory and practice. In: Heitkam T, Garcia S, editors. Plant Cytogenetics

and Cytogenomics: Methods and Protocols. New York (NY): Springer. p. 79–113.
Hjelmen CE, Blackmon H, Holmes VR, Burrus CG, Johnston JS. 2019. Genome size evolution differs between *Drosophila* subgenera with striking differences in male and female genome size in *Sophophora*. G3 (Bethesda). 9(10):3167–3179. doi:10.1534/g3.119.400560.
Hjelmen CE, Holmes VR, Burrus CG, Piron E, Mynes M, Garrett MA, Blackmon H, Johnston JS. 2020. Thoracic underreplication in *Drosophila* species estimates a minimum genome size and the dynamics of added DNA. Evolution. 74(7):1423–1436. doi:10.1111/evo.14022.
Hollox EJ, Zuccherato LW, Tucci S. 2022. Genome structural variation in human evolution. Trends Genet. 38(1):45–58. doi:10.1016/j.tig.2021.06.015.
Hood L, Rowen L. 2013. The Human Genome Project: big science transforms biology and medicine. Genome Med. 5(9):79. doi:10.1186/gm483.
Hotaling S, Wilcox ER, Heckenhauer J, Stewart RJ, Frandsen PB. 2023. Highly accurate long reads are crucial for realizing the potential of biodiversity genomics. BMC Genomics. 24(1):117. doi:10.1186/s12864-023-09193-9.
Huang W, Massouras A, Inoue Y, Peiffer J, Ramia M, Tarone AM, Turlapati L, Zichner T, Zhu D, Lyman RF, *et al.* 2014. Natural variation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines. Genome Res. 24(7):1193–1208. doi:10.1101/gr.171546.113.
Jauhal AA, Newcomb RD. 2021. Assessing genome assembly quality prior to downstream analysis: N50 versus BUSCO. Mol Ecol Resour. 21(5):1416–1421. doi:10.1111/1755-0998.13364.
Jaworski CC, Allan CW, Matzkin LM. 2020. Chromosome-level hybrid de novo genome assemblies as an attainable option for nonmodel insects. Mol Ecol Resour. 20(5):1277–1293. doi:10.1111/1755-0998.13176.
Jeffery NW, Gregory TR. 2014. Genome size estimates for crustaceans using Feulgen image analysis densitometry of ethanol-preserved tissues. Cytometry A. 85(10):862–868. doi:10.1002/cyto.a.22516.
Jeffery NW, Hultgren K, Chak STC, Gregory TR, Rubenstein DR. 2016. Patterns of genome size variation in snapping shrimp. Genome. 59(6):393–402. doi:10.1139/gen-2015-0206.
Johnston JS, Bernardini A, Hjelmen CE. 2019. Genome size estimation and quantitative cytogenetics in insects. In: Brown S, Pfrender M, editors. Insect Genomics. New York (NY): Humana Press. p. 15–26.
Johnston JS, Schoener M, McMahon DP. 2013. DNA underreplication in the majority of nuclei in the *Drosophila melanogaster* thorax: evidence from Suur and flow cytometry. J Mol Biol Res. 3(1):47–54. doi:10.5539/jmbr.v3n1p47.
Johnston JS, Zapalac ME, Hjelmen CE. 2020. Flying high—muscle-specific underreplication in *Drosophila*. Genes (Basel). 11(3):246. doi:10.3390/genes11030246.
Kalendar R, Tanskanen J, Immonen S, Nevo E, Schulman AH. 2000. Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proc Natl Acad Sci U S A. 97(12):6603–6607. doi:10.1073/pnas.110587497.
Kassambara A. 2018. ggpubr:'ggplot2' based publication ready plots. R package version 2. [accessed 2024 Feb 27]. https://CRAN.R-project.org/package=ggpubr.
Kidwell MG. 2002. Transposable elements and the evolution of genome size in eukaryotes. Genetica. 115(1):49–63. doi:10.1023/A:1016072014259.
Kim BY, Gellert HR, Church SH, Suvorov A, Anderson SS, Barmina O, Beskid SG, Comeault AA, Crown KN, Diamond SE, *et al.* 2023. Single-fly assemblies fill major phylogenomic gaps across the *Drosophilidae* Tree of Life. bioRxiv 2023.10.02.560517. doi:10.1101/2023.10.02.560517, preprint: not peer reviewed.
Kim BY, Wang JR, Miller DE, Barmina O, Delaney E, Thompson A, Comeault AA, Peede D, D'Agostino ERR, Pelaez J, *et al.* 2021. Highly contiguous assemblies of 101 drosophilid genomes. eLife. 10:e66405. doi:10.7554/eLife.66405.
Knight CA, Molinari NA, Petrov DA. 2005. The large genome constraint hypothesis: evolution, ecology and phenotype. Ann Bot. 95(1):177–190. doi:10.1093/aob/mci011.
Kong W, Wang Y, Zhang S, Yu J, Zhang X. 2023. Recent advances in assembly of complex plant genomes. Genomics Proteomics Bioinformatics. 21(3):427–439. doi:10.1016/j.gpb.2023.04.004.
Kress WJ, Soltis DE, Kersey PJ, Wegrzyn JL, Leebens-Mack JH, Gostel MR, Liu X, Soltis PS. 2022. Green plant genomes: what we know in an era of rapidly expanding opportunities. Proc Natl Acad Sci U S A. 119(4):e2115640118. doi:10.1073/pnas.2115640118.
Kronenberg ZN, Rhie A, Koren S, Concepcion GT, Peluso P, Munson KM, Porubsky D, Kuhn K, Mueller KA, Low WY, *et al.* 2021. Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. Nat Commun. 12(1):1935. doi:10.1038/s41467-020-20536-y.
Kullman B, Tamm H, Kullman K. 2005. Fungal genome size database. [accessed 2024 Feb 27]. http://www.zbi.ee/fungal-genomesize/.
Lang DT, Team C. 2012. XML: tools for parsing and generating XML within R and S-Plus. R package version 3.99-0.16. [accessed 2024 Feb 27]. https://CRAN.R-project.org/package=XML.
Leinaas HP, Jalal M, Gabrielsen TM, Hessen DO. 2016. Inter- and intraspecific variation in body- and genome size in calanoid copepods from temperate and arctic waters. Ecol Evol. 6(16):5585–5595. doi:10.1002/ece3.2302.
Leitch IJ, Chase MW, Bennett MD. 1998. Phylogenetic analysis of DNA C-values provides evidence for a small ancestral genome size in flowering plants. Ann Bot. 82:85–94. doi:10.1006/anbo.1998.0783.
Lewin HA, Richards S, Lieberman Aiden E, Allende ML, Archibald JM, Bálint M, Barker KB, Baumgartner B, Belov K, Bertorelle G, *et al.* 2022. The Earth BioGenome Project 2020: starting the clock. Proc Natl Acad Sci U S A. 119(4):e2115635118. doi:10.1073/pnas.2115635118.
Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, Durbin R, Edwards SV, Forest F, Gilbert MTP, *et al.* 2018. Earth BioGenome Project: sequencing life for the future of life. Proc Natl Acad Sci U S A. 115(17):4325–4333. doi:10.1073/pnas.1720115115.
Liao Y, Zhang X, Chakraborty M, Emerson JJ. 2021. Topologically associating domains and their role in the evolution of genome structure and function in *Drosophila*. Genome Res. 31(3):397–410. doi:10.1101/gr.266130.120.
Liu G, Chang Z, Chen L, He J, Dong Z, Yang J, Lu S, Zhao R, Wan W, Ma G, *et al.* 2020. Genome size variation in butterflies (Insecta, Lepidotera, Papilionoidea): a thorough phylogenetic comparison. Syst Entomol. 45(3):571–582. doi:10.1111/syen.12417.
Liu B, Shi Y, Yuan J, Hu X, Zhang H, Li N, Li Z, Chen Y, Mu D, Fan W. 2013. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. arXiv:1308.2012. doi:10.48550/arXiv.1308.2012, preprint: not peer reviewed.
Lower SS, Spencer Johnston J, Stanger-Hall K, Hjelmen CE, Hanrahan SJ, Korunes K, Hall D. 2017. Genome size in North American fireflies: substantial variation likely driven by neutral processes. Genome Biol Evol. 9(6):1499–1512. doi:10.1093/gbe/evx097.

Lucek K, Gompert Z, Nosil P. 2019. The role of structural genomic variants in population differentiation and ecotype formation in *Timema cristinae* walking sticks. Mol Ecol. 28(6):1224–1237. doi:10.1111/mec.15016.

Lynch M, Conery JS. 2003. The origins of genome complexity. Science. 302(5649):1401–1404. doi:10.1126/science.1089370.

Makalowski W. 2000. Genomic scrap yard: how genomes utilize all that junk. Gene. 259(1-2):61–67. doi:10.1016/S0378-1119(00)00436-4.

Makalowski W. 2003. Not junk after all. Science. 300(5623): 1246–1247. doi:10.1126/science.1085690.

Makova KD, Pickett BD, Harris RS, Hartley GA, Cechova M, Pal K, Nurk S, Yoo DA, Li Q, Hebbar P, *et al.* 2024. The complete sequence and comparative analysis of ape sex chromosomes. Nature. 630: 401–411. doi:10.1038/s41586-024-07473-2.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021b. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 38(10): 4647–4654. doi:10.1093/molbev/msab199.

Manni M, Berkeley MR, Seppey M, Zdobnov EM. 2021a. BUSCO: assessing genomic data quality and beyond. Curr Protoc. 1(12):e323. doi:10.1002/cpz1.323.

Mérot C, Oomen RA, Tigano A, Wellenreuther M. 2020. A roadmap for understanding the evolutionary significance of structural genomic variation. Trends Ecol Evol. 35(7):561–572. doi:10.1016/j.tree.2020.03.002.

Millard SP. 2013. EnvStats: an R package for environmental statistics. New York (NY): Springer Science & Business Media.

Morelli MW, Blackmon H, Hjelmen CE. 2022. Diptera and *Drosophila* karyotype databases: a useful dataset to guide evolutionary and genomic studies. Front Ecol Evol. 10:832378. doi:10.3389/fevo.2022.832378.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, *et al.* 2022. The complete sequence of a human genome. Science. 376(6588): 44–53. doi:10.1126/science.abj6987.

Ohno S. 1972. So much "junk" DNA in our genome. Brookhaven Symp Biol. 23:366–370.

Palazzo AF, Gregory TR. 2014. The case for junk DNA. PLoS Genet. 10(5):e1004351. doi:10.1371/journal.pgen.1004351.

Pellicer J, Hidalgo O, Dodsworth S, Leitch IJ. 2018. Genome size diversity and its impact on the evolution of land plants. Genes (Basel). 9(2):88. doi:10.3390/genes9020088.

Pellicer J, Leitch IJ. 2014. The application of flow cytometry for estimating genome size and ploidy level in plants. In: Besse P, editors. Molecular Plant Taxonomy: Methods and Protocols. Totowa (NJ): Humana Press. p. 279–307.

Pellicer J, Leitch IJ. 2020. The plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. New Phytol. 226(2):301–305. doi:10.1111/nph.16261.

Perkins RD, Gamboa JR, Jonika MM, Lo J, Shum A, Adams RH, Blackmon H. 2019. A database of amphibian karyotypes. Chromosome Res. 27(4):313–319. doi:10.1007/s10577-019-09613-1.

Pflug JM, Holmes VR, Burrus C, Johnston JS, Maddison DR. 2020. Measuring genome sizes using read-depth, k-mers, and flow cytometry: methodological comparisons in beetles (Coleoptera). G3 (Bethesda). 10(9):3047–3060. doi:10.1534/g3.120.401028.

Prunier J, Giguère I, Ryan N, Guy R, Soolanayakanahally R, Isabel N, MacKay J, Porth I. 2019. Gene copy number variations involved in balsam poplar (*Populus balsamifera* L.) adaptive variations. Mol Ecol. 28(6):1476–1490. doi:10.1111/mec.14836.

R Core Team. 2016 R: A Language and environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, *et al.* 2021. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 592(7856):737–746. doi:10.1038/s41586-021-03451-0.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 74(12): 5463–5467. doi:10.1073/pnas.74.12.5463.

Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S, *et al.* 2022. Database resources of the national center for biotechnology information. Nucleic Acids Res. 50(D1):D20–D26. doi:10.1093/nar/gkab1112.

Schielzeth H, Streitner C, Lampe U, Franzke A, Reinhold K. 2014. Genome size variation affects song attractiveness in grasshoppers: evidence for sexual selection against large genomes. Evolution. 68(12):3629–3635. doi:10.1111/evo.12522.

Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, Murphy TD, Pruitt KD, Thibaud-Nissen F, Albracht D, *et al.* 2017. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 27(5):849–864. doi:10.1101/gr.213611.116.

Sedlazeck FJ, Lee H, Darby CA, Schatz MC. 2018. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat Rev Genet. 19(6):329–346. doi:10.1038/s41576-018-0003-4.

Sessegolo C, Burlet N, Haudry A. 2016. Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies. Biol Lett. 12(8):20160407. doi:10.1098/rsbl.2016.0407.

Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, Waterston RH. 2017. DNA sequencing at 40: past, present and future. Nature. 550(7676):345–353. doi:10.1038/nature24286.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31(19): 3210–3212. doi:10.1093/bioinformatics/btv351.

Song J-M, Xie W-Z, Wang S, Guo Y-X, Koo D-H, Kudrna D, Gong C, Huang Y, Feng J-W, Zhang W, *et al.* 2021. Two gap-free reference genomes and a global view of the centromere architecture in rice. Mol Plant. 14(10):1757–1767. doi:10.1016/j.molp.2021.06.018.

Sylvester T, Hjelmen CE, Hanrahan SJ, Lenhart PA, Johnston JS, Blackmon H. 2020. Lineage-specific patterns of chromosome evolution are the rule not the exception in Polyneoptera insects. Proc Biol Sci. 287(1935):20201388. doi:10.1098/rspb.2020.1388.

The Darwin Tree of Life Project Consortium. 2022. Sequence locally, think globally: the Darwin Tree of Life Project. Proc Natl Acad Sci U S A. 119(4):e2115642118. doi:10.1073/pnas.2115642118.

Thrash A, Hoffmann F, Perkins A. 2020. Toward a more holistic method of genome assembly assessment. BMC Bioinformatics. 21(Suppl 4):249. doi:10.1186/s12859-020-3382-4.

Verlinden H, Sterck L, Li J, Li Z, Yssel A, Gansemans Y, Verdonck R, Holtof M, Song H, Behmer ST, *et al.* 2021. First draft genome assembly of the desert locust, *Schistocerca gregaria*. F1000Res. 9: 775. doi:10.12688/f1000research.25148.2.

Vieira C, Aubry P, Lepetit D, Biémont C. 1998. A temperature cline in copy number for 412 but not roo/B104 retrotransposons in populations of *Drosophila simulans*. Proc Biol Sci. 265(1402):1161–1165. doi:10.1098/rspb.1998.0413.

Wang X, Fang X, Yang P, Jiang X, Jiang F, Zhao D, Li B, Cui F, Wei J, Ma C, *et al.* 2014. The locust genome provides insight into swarm formation and long-distance flight. Nat Commun. 5(1):2957. doi:10.1038/ncomms3957.

Wang B, Hou M, Shi J, Ku L, Song W, Li C, Ning Q, Li X, Li C, Zhao B, *et al.* 2023. De novo genome assembly and analyses of 12 founder

inbred lines provide insights into maize heterosis. Nat Genet. 55(2):312–323. doi:10.1038/s41588-022-01283-w.
Wang P, Wang F. 2023. A proposed metric set for evaluation of genome assembly quality. Trends Genet. 39(3):175–186. doi:10.1016/j.tig.2022.10.005.
Wang Y, Yu J, Jiang M, Lei W, Zhang X, Tang H. 2023. Sequencing and assembly of polyploid genomes. In: Van de Peer Y, editors. Polyploidy: Methods and Protocols. New York (NY): Springer. p. 429–458.
Ward BL, Heed WB. 1970. Chromosome phylogeny of *Drosophila pachea* and related species. J Hered. 61(6):248–258. doi:10.1093/oxfordjournals.jhered.a108095.
Whitney KD, Garland T Jr. 2010. Did genetic drift drive increases in genome complexity? PLoS Genet. 6(8):e1001080. doi:10.1371/journal.pgen.1001080.
Wickham H. 2007. Reshaping data with the reshape package. J Stat Softw. 21(12):1–20. doi:10.18637/jss.v021.i12.
Wickham H. 2016. ggplot2: elegant graphics for data analysis. New York (NY): Springer
Wickham H. 2023a. Forcats: tools for working with categorical variables (factors). R package version. 05:1. [accessed 2024 Feb 27]. https://CRAN.R-project.org/package=forcatsForcats: tools for working with categorical variables (factors).
Wickham H. 2023b. Httr: tools for working with URLs and HTTP (R package version 1.4. 2) [Computer software]. [accessed 2024 Feb 27]. https://CRAN.R-project.org/package=httr.
Wickham H, François R, Henry L, Müller K, Vaughan D. 2023. dplyr: a grammar of data manipulation. R package version. 1.1. 2. [accessed 2024 Feb 27]. https://CRAN.R-project.org/package=dplyr.
Winter DJ. 2017. rentrez: an R package for the NCBI eUtils API. The R Journal. 9(2):520–526. doi:10.32614/RJ-2017-058.
Yuan H, Huang Y, Mao Y, Zhang N, Nie Y, Zhang X, Zhou Y, Mao S. 2021. The evolutionary patterns of genome size in Ensifera (Insecta: Orthoptera). Front Genet. 12:693541. doi:10.3389/fgene.2021.693541.
Zhang G. 2015. Bird sequencing project takes off. Nature. 522(7554):34. doi:10.1038/522034d.
Zhang X, Zhang S, Zhao Q, Ming R, Tang H. 2019. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nat Plants. 5(8):833–845. doi:10.1038/s41477-019-0487-8.

Editor: J. Stajich