BMC Biology

RESEARCH ARTICLE Open Access

Comparative analysis reveals within-population genome size variation in a rotifer is driven by large genomic elements with highly abundant satellite DNA repeat elements

C. P. Stelzer[1*], J. Blommaert[1,2], A. M. Waldvogel[3], M. Pichler[1], B. Hecox-Lea[4] and D. B. Mark Welch[4]

Abstract

Background: Eukaryotic genomes are known to display an enormous variation in size, but the evolutionary causes of this phenomenon are still poorly understood. To obtain mechanistic insights into such variation, previous studies have often employed comparative genomics approaches involving closely related species or geographically isolated populations within a species. Genome comparisons among individuals of the same population remained so far understudied—despite their great potential in providing a microevolutionary perspective to genome size evolution. The rotifer *Brachionus asplanchnoidis* represents one of the most extreme cases of within-population genome size variation among eukaryotes, displaying almost twofold variation within a geographic population.

Results: Here, we used a whole-genome sequencing approach to identify the underlying DNA sequence differences by assembling a high-quality reference genome draft for one individual of the population and aligning short reads of 15 individuals from the same geographic population including the reference individual. We identified several large, contiguous copy number variable regions (CNVs), up to megabases in size, which exhibited striking coverage differences among individuals, and whose coverage overall scaled with genome size. CNVs were of remarkably low complexity, being mainly composed of tandemly repeated satellite DNA with only a few interspersed genes or other sequences, and were characterized by a significantly elevated GC-content. CNV patterns in offspring of two parents with divergent genome size and CNV patterns in several individuals from an inbred line differing in genome size demonstrated inheritance and accumulation of CNVs across generations.

Conclusions: By identifying the exact genomic elements that cause within-population genome size variation, our study paves the way for studying genome size evolution in contemporary populations rather than inferring patterns and processes a posteriori from species comparisons.

Keywords: Genome size evolution, Genetic variation, Rotifer, C-value, Satellite DNA, B-chromosomes, Transposable elements, Comparative genomics

* Correspondence: claus-peter.stelzer@uibk.ac.at
[1]Research Department for Limnology, University of Innsbruck, Mondsee, Austria
Full list of author information is available at the end of the article

Background

The genomes of eukaryotic organisms display remarkable diversity in size, overall spanning approximately five orders of magnitude [1]. In addition, genome size may vary substantially among closely related species [2, 3], within a species (e.g., [4–6]), and sometimes even within a population [7]. Most of the variation in genome size stems from differences in the proportion of various kinds of non-coding DNA and/or transposable elements, which can reach excessive levels in species with giant genomes [8, 9]. Studying genome size variation at the DNA sequence level allows identification of exactly those genomic elements that make up for the genome size difference, and it can suggest the relative strength of mutation, selection, and drift—the underlying evolutionary forces ultimately causing divergence in genome size.

Much of our understanding of eukaryotic genome size variation comes from comparisons between closely related species. Recent studies suggested that the proliferation of repetitive elements (REs), in particular transposable elements, plays an important role in genome expansion, while their silencing or deletion has been implicated in the streamlining of genomes. In a few studies, it was possible to pinpoint individual REs as the driver of genome expansion [10, 11], whereas in other studies, differently sized genomes were found to differ in several classes of REs [12, 13]. In the latter case, it is difficult to decide whether multiple RE classes have expanded more or less simultaneously in evolutionary time, or whether the expansions of some REs have occurred *after* an initial genome size divergence driven by a single element. Without accurate dating of the expansions of individual elements, interspecific comparisons suffer from such a “blind spot” on the early stages of genome divergence. Ultimately, all genome size differences must have gone through a stage of intrapopulation variation followed by fixation, or loss, of these size variants. Thus, identifying genome size variants within populations and studying them on microevolutionary time scales may allow additional insights into the evolutionary dynamics of early genome divergences.

Intraspecific genome size variation (IGV) has been described in several species of eukaryotes (some examples are summarized in [7, 14]), although studies on the level of a single, natural population (i.e., unaltered by, e.g., inbreeding) are rare [15]. IGV may be associated with variation in the number of chromosomes (e.g., B-chromosomes [14]), but there are also examples where IGV is not reflected in the karyotype [2, 6]. One of the most recent additions to IGV model organisms was the monogonont rotifer *Brachionus asplanchnoidis*, which displays a nearly 2-fold variation in genome size even among individuals within a geographic population [7]. This level of variation is at the upper end of what has been found in other animals or plants (see Supplementary Table S4 in [7]). Monogonont rotifers are short-lived (1–2 weeks), small aquatic metazoans, only a few hundred micrometers in size, common in fresh and brackish water habitats throughout the world. They have a life cycle involving cyclical parthenogenesis [16] reproducing by ameiotic parthenogenesis for prolonged periods and inducing sexual reproduction occasionally. A “rotifer clone” consists of the asexual descendants of a single female that has hatched from a single resting egg, which itself is the product of sexual reproduction. In many Monogonont species, sex is triggered by crowding due to the accumulation of sex-inducing peptides released by the animals [17]. In lab cultures, it is possible to suppress sexual reproduction by frequent dilution intervals or large culture volumes or to induce sex in small culture volumes or through the use of media drawn from dense cultures. Thus, it is possible to either deliberately cross two rotifer clones sexually or to keep them clonally for hundreds of generations.

In the present study, we focus on a population of *B. asplanchnoidis* from Obere Halbjochlacke (OHJ), a shallow alkaline lake in Eastern Austria [7, 18]. Individuals of this population can be crossed with each other—even if they substantially differ in genome size—and they will produce offspring with intermediate genome sizes close to the parental mean. Genome size can be artificially selected up or down with a heritability of 1 by breeding only individuals with large or small genome sizes. Genome size variation in this system is mediated by relatively large genomic elements (several megabases in size), which segregate independently from each other during meiosis. The smallest observed genome size in *B. asplanchnoidis* was 404Mb (2C, nuclear DNA content). Individuals at or close to this basal genome size are completely lacking independently segregating elements, while in individuals with larger genome sizes, genome size scales with the amount of independently segregating elements [7].

Here, we used a whole-genome sequencing approach to identify the DNA sequence differences responsible for intrapopulation genome size variation in this population of *B. asplanchnoidis*. Our specific goal was to identify and characterize genomic regions that are present in one or multiple copies in some individuals of the OHJ population but are missing in others. To this end, we assembled a highly contiguous draft genome of a reference clone using long-read (PacBio) technology and then mapped short reads of 15 different clones with genome sizes from 404 to 644 Mbp to this assembly. To identify copy number variations (CNVs), we scanned for regions of increased per-bp read coverage. To independently confirm CNVs, we used PCR to detect the presence/absence of selected CNVs across different clones of the

OHJ population, and droplet digital PCR (ddPCR) to determine the exact copy numbers of one specific locus. Finally, we annotated genes and repetitive elements in the reference genome and compared CNV regions to non-variable regions of the genome.

Results

De novo assembly and annotation of the reference genome

The rotifer clone (OHJ7i3n10) chosen for our reference genome derives from the natural isolate OHJ7 after three rounds of selfing (i.e., fertilizing sexual females by males of the same clone). As measured by flow cytometry [7], OHJ7i3n10 has a 2C-genome size of 568 Mbp, and thus contains approximately 40% excess DNA, compared to the smallest genome size of the OHJ population (~410 Mbp). The total length of our reference assembly was 230.1 Mb, with 455 contigs and an N50 value of 3.065 Mb (Fig. 1). The average GC-content was 30.5%. However, the GC-distribution was not unimodal but showed two major peaks at ~25% and ~35% GC and a minor one at ~50% GC (Additional file 1: Figure S1).

Taxonomic partitioning of the polished genome assembly confirmed its purity. Most hits could correctly be assigned to rotifers and remaining hits assigned to mollusks and arthropods can mostly be explained by imbalanced availability of rotifer entries in the *nt* database (Additional file 1: Figure S2, Additional file 2). We observed that 90.6% of the metazoan BUSCO gene set collection was complete with low levels of duplicated (2.7%), fragmented (2.0%), and missing (7.4%) genes (Additional file 1: Table S2). A visualization of assembly contiguity and completeness was generated via assembly-stats [19] and is presented in Fig. 1. Protein-coding genes make up approx. 26% of the genome assembly length. In total, we annotated 16,667 genes with a median gene length of 1999 bp and approx. five exons per gene (Additional file 1: Table S3).

Comparison of short reads in 15 OHJ clones

To examine within-population genome size variation, we sequenced 29 short-read libraries from 15 different rotifer clones from the OHJ population (1–4 libraries per clone using different methods, described below). Nine clones were asexual descendants of individuals collected from the field, four (including the source of the reference genome) were each asexual descendants of the same three rounds of selfing of one of these clones, one was an asexual descendant from three rounds of selfing of a different clone, and one was an asexual descendant of a cross between clones derived from crossing two different selfed lineages from two natural isolates (Table 1, Table S4). Raw reads were passed through multiple preprocessing steps that included quality trimming, removal of PCR duplicates and mitochondrial DNA, and removal of contaminant DNA. Overall, preprocessing reduced the total sequence amount from 265.6 to 194.3 Gbp, resulting in per-base sequencing coverage of 9.4- to 79-

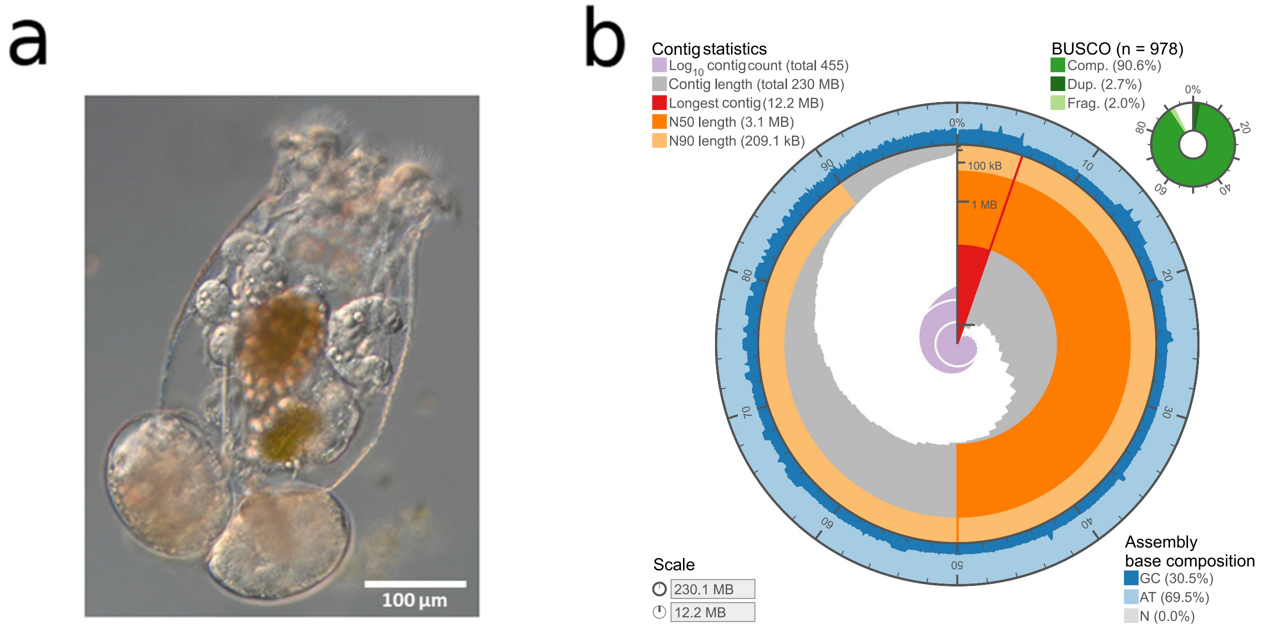


Fig. 1. a Photograph of *B. asplanchnoidis* female (with two attached asexual eggs). **b** Circular visualization of the contiguity and completeness of the genome assembly. The full circle represents the assembly length of 230.1 Mb with the longest contig of 12.2 Mb in red and very few contigs <100 kB. GC-content is varying across the assembly (blue)

Table 1 Rotifer clones used in this study

Clone	Genome size[1] (Mbp)	Origin	No. of libraries
OHJ82	404	Natural clone	2
OHJ22	412	Natural clone	2
OHJ104	462	Natural clone	2
OHJ97	470	Natural clone	2
OHJ96	492	Natural clone	1
OHJ98	504	Natural clone	1
OHJ105	520	Natural clone	2
OHJ7	532	Natural clone	4
OHJ13	536	Natural clone	2
OHJ22i3n14	420	Clone derived from selfing[3] OHJ22	1
OHJ7i3n7	536	Clone derived from selfing[3] OHJ7	2
OHJ7i3n2	560	Clone derived from selfing[3] OHJ7	2
OHJ7i3n10[2]	568	Clone derived from selfing[3] OHJ7	2
OHJ7i3n5	644	Clone derived from selfing[3] OHJ7	2
IK1	500	Clone from cross of OHJ7i3n2 and OHJ22i3n14	2

[1] 2C-genome size estimated by flow cytometry (Stelzer et al. [7])
[2] Same clone as the reference genome
[3] Hatched from an individual resting egg after three generations of selfing

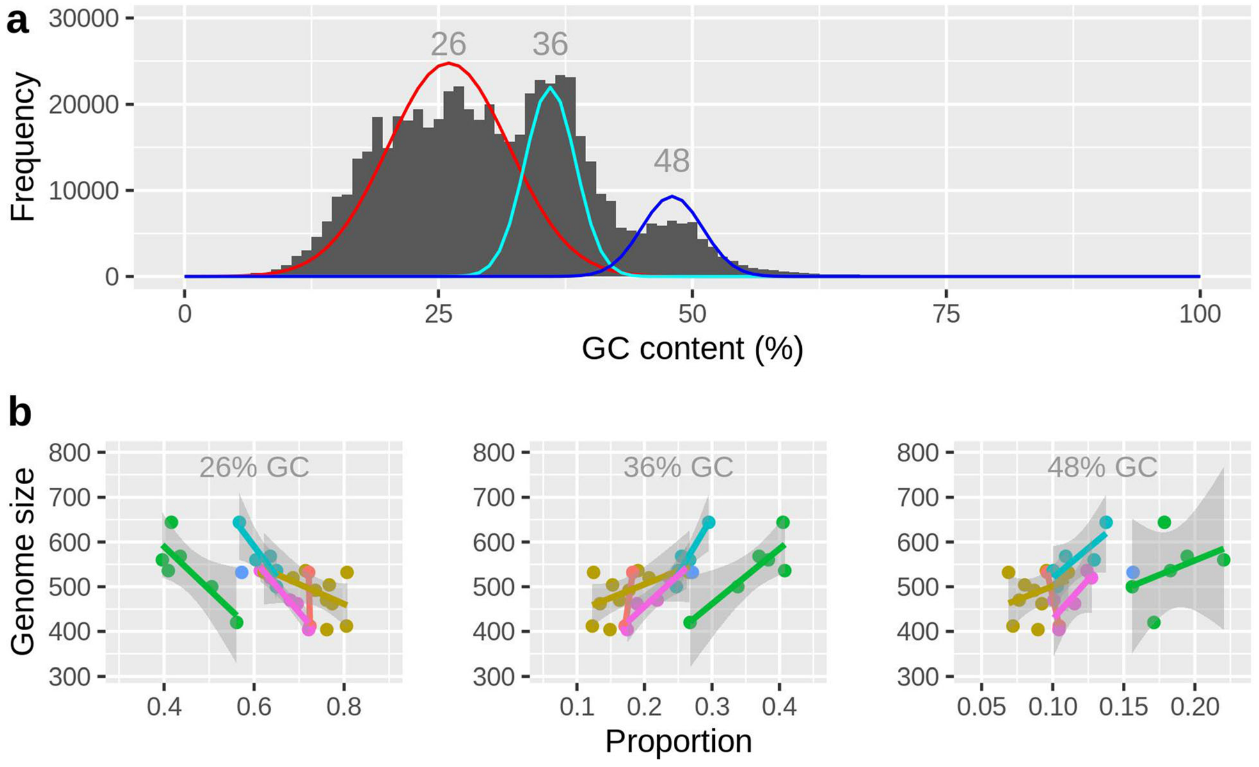


Fig. 2. Genome size variation is linked to sequences with elevated GC-content. **a** GC-distribution of all short-read libraries combined (29 libraries of 15 rotifer clones). The red, cyan, and blue lines designate a mixture model fitted to these data, consisting of three normally distributed subpopulations (26 ± 6, 36 ± 2.5, and 48 ± 3 %GC; means and SDs). **b** Panels show the results of the same mixture model applied to each library individually, thus estimating the proportion of the total reads per library in each GC-fraction. 2C-genome size estimates are based on flow cytometry and were taken from [7]. Colors in **b** correspond to the six library preps (A–F) listed in Table S4: A = orange, B = gold, C = green, D = turquoise, E = blue, and F = pink

fold for the different libraries, with the majority of libraries being above 20-fold (Additional file 1: Supplementary results, Figure S3, Additional file 3).

Pooling short reads from all libraries revealed the same three-peak pattern of GC-content apparent in the reference assembly, with GC maxima at 26%, 36%, and 48% (Fig. 2a). Since these three GC peaks are indicative of three discrete fractions among the genomic reads, we applied a mixture model to the short-read data, which allowed estimating the relative proportion of each fraction per library. Overall, there was substantial variation in the relative proportions of the three fractions, both among rotifer clones and between libraries of the same clone. Overall, the 26% GC-fraction was negatively correlated with genome size based on flow cytometry (FCM), and the 36% and 48% GC-fractions were positively correlated (Fig. 2b, Additional file 1: Table S5).

We also used two kmer-based tools, GenomeScope 2.0 [20] and findGSE [21], to obtain reference-free estimates of genome size for each clone/library. Those estimates were generally lower than their FCM-based counterparts, approximately 0.8-fold in findGSE and 0.6-fold in GenomeScope (Additional file 4). GenomeScope appeared to underperform at sequencing coverages below ca. 25-fold, where it estimated extremely low genome sizes (compared to FCM) and unrealistically high heterozygosities (6–10%). By contrast, findGSE performed consistently along the gradient of sequencing coverages. Overall, there was a positive correlation between genome size (FCM estimate) and the ratio of repeats, a fitted parameter of findGSE (Additional file 1: Figure S4; Pearson's r_{27} = 0.53, p = 0.004).

Our assembly-based analyses rely on the alignment of cleaned reads of each of the 29 libraries to the reference genome. Total alignment rates (TARs) of reads to the reference genome draft were generally above 94%. TARs were highest in those rotifer clones that are most closely related to the reference genome (Additional file 1: Figure S5). Concordant alignment rates, i.e., properly aligned reads with the correct insert size, were > 90% in all libraries (Additional file 1: Table S6). However, only about 1/3 of the reads aligned uniquely to one site in the genome. Only a small proportion of reads, usually well below 5%, showed discordant alignment, either due to incorrect insert size, or when only one of the mates aligning to the reference genome. Discordant alignment rates were generally low (1.3–5.1%) and were not correlated to genome size (r_{27} = −0.124, p = 0.5).

To identify CNV regions, we calculated the average per-base coverage for 5-kbp and 50-kbp windows. To normalize coverage across libraries, we divided the per-base coverage of each window by 1/2 of the (mean) exon coverage of the respective library. This yields a value of 2 for all diploid regions of the genome, corresponding to two copies for those genomic regions (provided that our assembly is unphased in these regions). This analysis of coverage variation revealed that large tracts in the genome of *B. asplanchnoidis* display consistent patterns of coverage variation, which we quantified as the standard deviation of coverage across libraries and clones (Fig. 3). For example, the first contig (000F) exhibits large coverage variation with a standard deviation of 1.5–1.7 while the next three contigs (001F to 003F) have much less coverage variation (< 0.5 s.d.) and a mean coverage value of close to 2. There are several other contigs showing consistently elevated coverage variation, like 004F, 031F, 032F,036F, 042F, and 044F. In addition, some of these variable contigs appear to be more variable than others (e.g., 031F is more variable than 032F). These overall patterns were very similar when a lower 5-kbp window resolution was applied (Additional file 1: Figure S6).

Combining the values of coverage variation of all windows (n = 4380 for 50kbp, n = 45800 for 5kbp) reveals a multimodal distribution with a prominent peak located at low coverage variations of ~0.15 (lowSD in Fig. 4a). This peak corresponds to the genomic sections that are colored in orange/red with a mean coverage of 2 in Fig. 3. At intermediate coverage variations (interSD; 0.7 < s.d. < 2.0), there appear to be at least two peaks, which correspond to the green and blue regions in Fig. 3, respectively. There are also a few windows showing high coverage variation (highSD; s.d. > 2.0), which form the right tail in Fig. 4a.

To test for an effect of these coverage variations on genome size, we calculated the mean coverage for each clone/library for these three categories of coverage variation (lowSD, interSD, highSD) and calculated their correlations with genome size. Notably, there were substantial differences in coverage at the interSD and high SD regions among the different libraries, even in some that had been prepared from the same rotifer clone (Fig. 4c). Thus, to control for the effect of library preparation, we calculated partial correlations between the variables "mean coverage" and "genome size" (Fig. 4c). Those correlations at "intermediate" and "high" variability were highly significant. This result holds even if the libraries with elevated coverage and GC-content at interSD and highSD regions (green symbols in Figs. 2b and 4c) are excluded (Additional file 1: Figure S7).

By merging adjacent 5-kbp windows that show increased coverage variation and consistent coverage pattern (significant correlation of coverages), we identified 509 CNV regions in the genome of *B. asplanchnoidis* (Additional file 5). The total genome space classified hitherto as "copy number variable" was 72.43Mbp (i.e., 31% of the genome assembly). Most CNV regions tended to be rather long, as can be seen by an "N50" of 0.455 Mbp for the CNV fraction of the genome. Two of

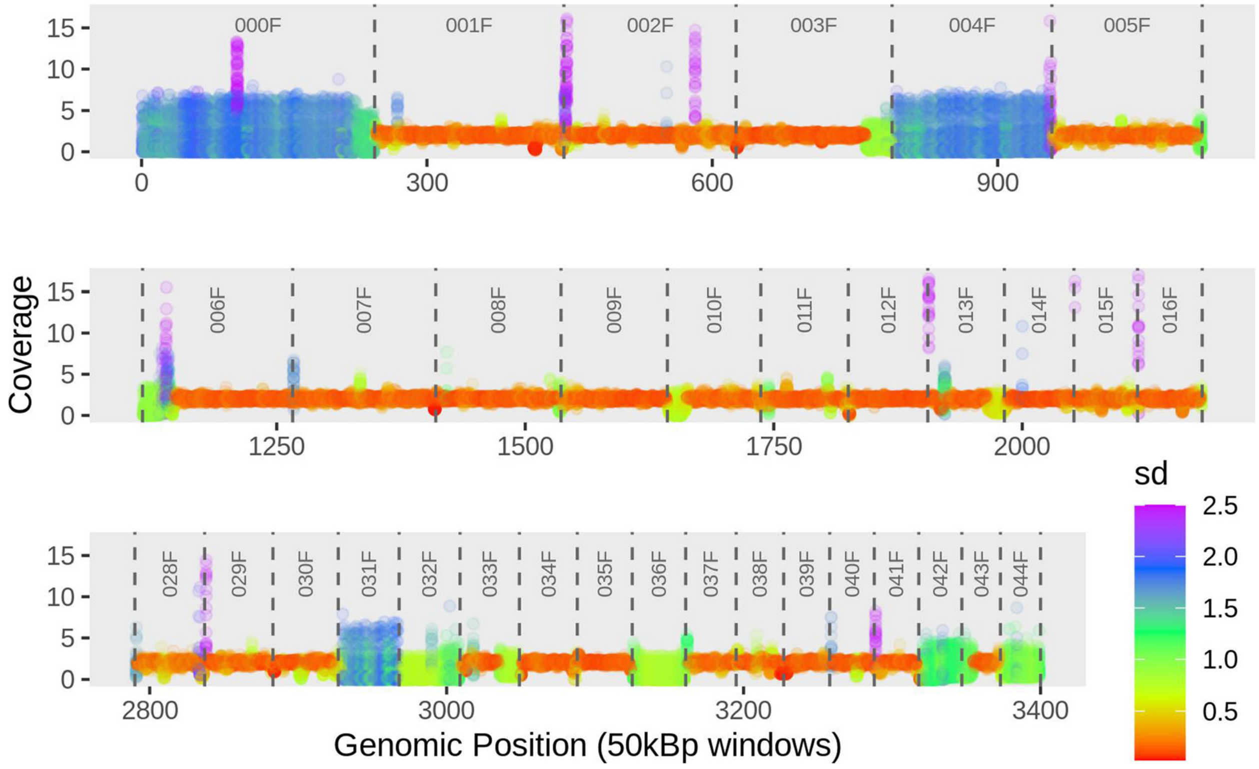
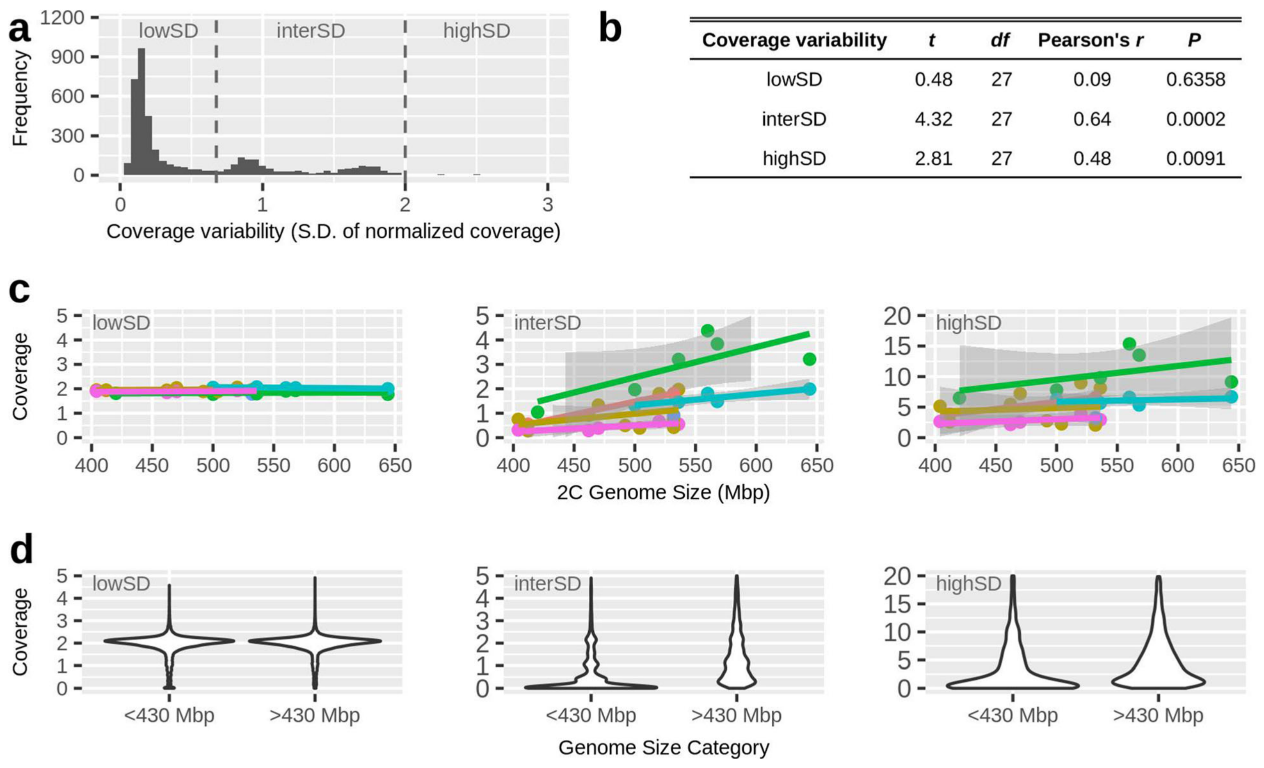


Fig. 3. Coverage variation along the *B. asplanchnoidis* genome. This graph shows 34 selected contigs spanning ~140 Mbp (i.e., about 60% of the assembly). Each circle represents a 50-kbp window of one of the 29 sequencing libraries. Coverage (*y*-axis) was normalized by dividing the per-base coverage of each window by 1/2 of the (mean) exon coverage of the respective library. Contig borders are indicated by vertical dashed lines, and contig IDs are listed on the top. Contigs are ordered by size rather than reflecting biological contiguity. Variation across clones/libraries is indicated by color (based on standard deviations). Standard deviations larger than 2.5 (about 4% of the data) were capped to a value of 2.5, to limit outliers affecting the color scale

the largest contigs, 000F and 004F, consisted almost entirely of 3–4 large CNV regions, which were separated only by short “breakpoints” of lower coverage variability (Additional file 5). Our 5-kbp window-scanning algorithm also identified CNVs that resided in contigs with otherwise “normal” coverage. In many cases, these CNV regions were near the beginning or end of a contig (e.g., 003F, 006F, 010F). Coverage values of clone IK1, a cross between OHJ7i3n2 and OHJ22i3n14, were usually intermediate between those of the two parental clones (Figures S8, S11, S12). In addition, four clones that were derived by selfing from clone OHJ7 displayed coverages that were largely consistent with their differences in genome size (Figure S9). For instance, the clone with the largest genome of the selfed line, OHJ7i3n5, had an additional coverage peak at about 2.5 times the base coverage (set by IK1), which indicates that some CNV regions have significantly higher coverage than any other clone of this selfed line.

To additionally classify CNV regions according to their length and contiguity, we considered contigs as “B-contigs” if they contained a large fraction of CNV windows (in analogy to B-chromosomes). Setting this threshold at 90% of contig length, 38 contigs are classified as “B-contigs,” comprising 77% of all CNV windows, i.e., 55.8 Mbp of the assembly (Additional file 1: Figures S10, S11, Table S7). Thus, approximately three-quarters of the observed CNVs affect more or less an entire contig, while the remaining quarter of CNV regions were found on contigs with otherwise low coverage variability.

To independently confirm CNVs, we chose four genomic loci for PCR amplification, two in CNV regions and two in non-CNV regions (Additional file 1: Tables S8, S9). All four primer pairs yielded amplicons with the correct size, with no signs of non-specific amplification (Fig. 5a). The two primer pairs targeted to non-CNV regions (TA_001F and TA_003F) yielded an amplicon in all rotifer clones. In contrast, the two primer pairs targeted to CNV loci (TA_000F and TA_032F) amplified only in some clones. In particular, clones with the smallest genome sizes (OHJ82, OHJ22, and its descendent OHJ22i3n14) seem to lack both CNV loci, and in others



Coverage variability	*t*	*df*	Pearson's *r*	*P*
lowSD	0.48	27	0.09	0.6358
interSD	4.32	27	0.64	0.0002
highSD	2.81	27	0.48	0.0091

Fig. 4. Genomic regions of elevated coverage variability are responsible for genome size variation. The data in this figure are based on normalized coverage (calculated for 50-kbp windows) of 29 short-read libraries from 15 rotifer clones. **a** Genomic regions differ in coverage variability (measured as the standard deviation of coverage among libraries) as there are regions of low, intermediate, and high variability. **b**, **c** Mean coverage in regions of elevated coverage variation (interSD, highSD) correlates positively with genome size (flow cytometry data on genome size was taken from [7]). Dots in **c** represent the mean coverage per library. Colors indicate co-prepared libraries (of which some were prepared from the same rotifer clone, but at different dates and with different library preparation methods; see Table S4). Colors in **c** correspond to the six library preps (A–F) listed in Table S4: A = orange, B = gold, C = green, D = turquoise, E = blue, and F = pink. **d** This panel shows the same data as in **c**, but rotifer clones are categorized into small 2C genome sizes (<430Mbp, which represent the basal genome size of *B. asplanchnoidis* according to [7]) vs. large genome sizes (>430Mbp, which contain additional genomic elements that independently segregate during meiosis, according to [7])

(OHJ 98, 104, 105) the TA_000F-locus was present, but the TA_032F-locus was absent. Overall, these patterns were highly consistent with coverage of the amplified regions in sequencing libraries (Fig. 5c). Copy numbers for TA_032, as estimated by ddPCR, ranged from zero to six across the studied rotifer clones, including 3 copies in IK1, a cross between OHJ7i3n2 (six copies) and OHJ22i3n14 (zero copies) (Fig. 5b, Table S8).

After having identified the CNV regions that contribute to intrapopulation genome size variation in the OHJ population, we annotated repetitive elements of these regions and compared them to the rest of the genome. A custom repeat library was created using RepeatModeler2, and the top-contributing TEs were curated. In total, 123 Mbp of the assembly (53.6%) was masked by this library. The highest contributing element (rotiSat2) accounts for just over 50 Mbp of this (Fig. 6a). The 36 most abundant repeats represent 67% of masked repeats and 82.6 Mbp of the assembly. Of the 36 highly contributing repeats, 7 were enriched in the interSD region, 12 in the highSD region, 5 in allCNVs, 6 in the B90 region, and 5 in the B95 region (Additional file 6). Overall, repetitive elements, and especially satDNAs, were over-represented in CNV regions (Fig. 6c, Additional file 7).

Of the satDNAs we identified in B. *asplanchnoidis*, three (rotiSat2, 8, 9) are not present in any other sequenced *Brachionus* genome; three others (rotiSat1, 5, 10, 11) are shared with the *Brachionus plicatilis* genome. All but two other repeats in the topRE library were found in at least two other *Brachionus* genomes in varying levels. In addition, we identified a DNA/MITE element and an uncharacterized element in the *Brachionus plicatilis* genome that are not found in other *Brachionus* genomes.

CNV regions differed strongly from non-CNV regions by having a much lower gene density (Fig. 6c, Additional file 7). Phylogenetic orthology inference based on

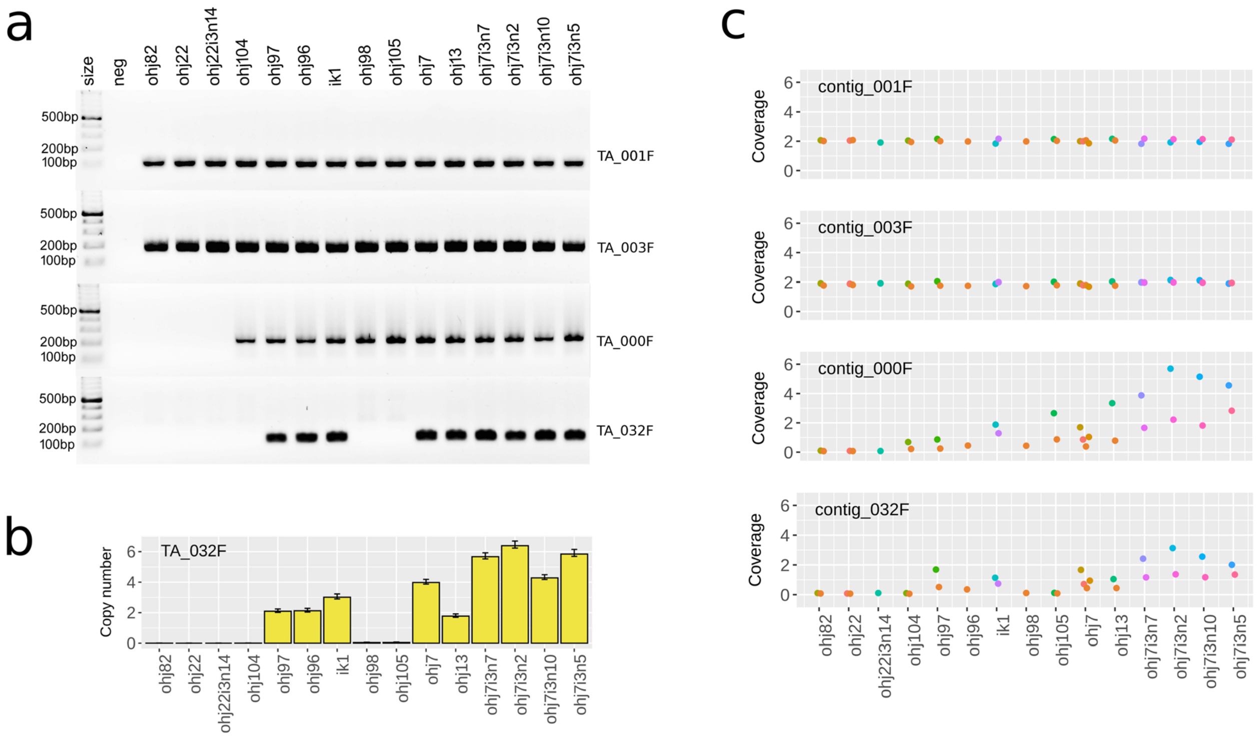


Fig. 5. PCR-based confirmation of CNV loci. All 15 studied clones from the OHJ population are displayed in order of ascending genome size (from left to right). **a** PCR products of primers targeting the four candidate loci (non-CNV loci: TA_001F and TA_003F; CNV loci: TA_000F and TA_032F). **b** Copy numbers of the TA_032F-locus as estimated by ddPCR. Error bars are 95% confidence intervals. **c** Coverage of the amplified region in sequencing libraries of the same 15 rotifer clones (for comparison). Dots are the average coverages across an entire contig. Colors identify co-prepared sequencing libraries

proteomes (OrthoFinder analysis) of four species in the *B. plicatilis* species complex, with the bdelloid rotifer *Adineta vaga* as an outgroup, resulted in the assignment of 93,063 genes (90.9% of all annotated proteins) to 17,965 orthogroups. Fifty percent of genes were in orthogroups with 6 or more genes and were contained in the largest 5228 orthogroups. There were 3953 orthogroups represented in all species and 299 of these consisted entirely of single-copy genes. Many duplication events appear to be species-specific, with an especially high number of genes (5634, or 32% of all protein-coding genes) derived from gene duplication in *B. asplanchnoidis*. While gene density is significantly reduced within CNV regions (417 genes located within CNV regions = 2.39% of all annotated protein-coding genes, p < 0.001, Fig. 6c), a significant number of these genes derive from gene duplication events (385 genes = 92.33% of all genes located within CNVs, p < 0.001, Fig. 6c). The overall pattern of gene distribution thus shows that CNV regions almost exclusively contain gene copies.

GO enrichment analysis of genes throughout the *B. asplanchnoidis* genome that derived from duplication events identified 29 significantly enriched GO terms (Additional file 1: Table S10). When restricting the gene set to only those genes derived from a duplication event that were found within CNV regions, we identified eleven significantly enriched GO terms (Additional file 1: Table S11).

Throughout this study, we observed several conspicuous patterns related to GC-content. Regions of elevated coverage variability (i.e., interSD and highSD regions), CNV regions, and B-contigs were characterized by an elevated GC-content showing the main peak at ~37% GC and two additional peaks at around 50% GC (Additional file 1: Figures S13, S14). By contrast, regions of low coverage variability had their main peak at ~25% GC (Additional file 1: Figure S14). Those three peaks were also present in the GC-distributions of unaligned sequencing reads from rotifer clones varying in genome size (Fig. 2b). The 37% GC peak, which was the most prominent peak in genomic regions of elevated SD, CNVs, and B-contigs (Additional file 1: Figure S14), could be attributed mainly to the satellites rotiSat1 and rotiSat2, while the higher peak at ~55% GC could be attributed to rotiSat5 (Additional file 1: Figure S15-S19).

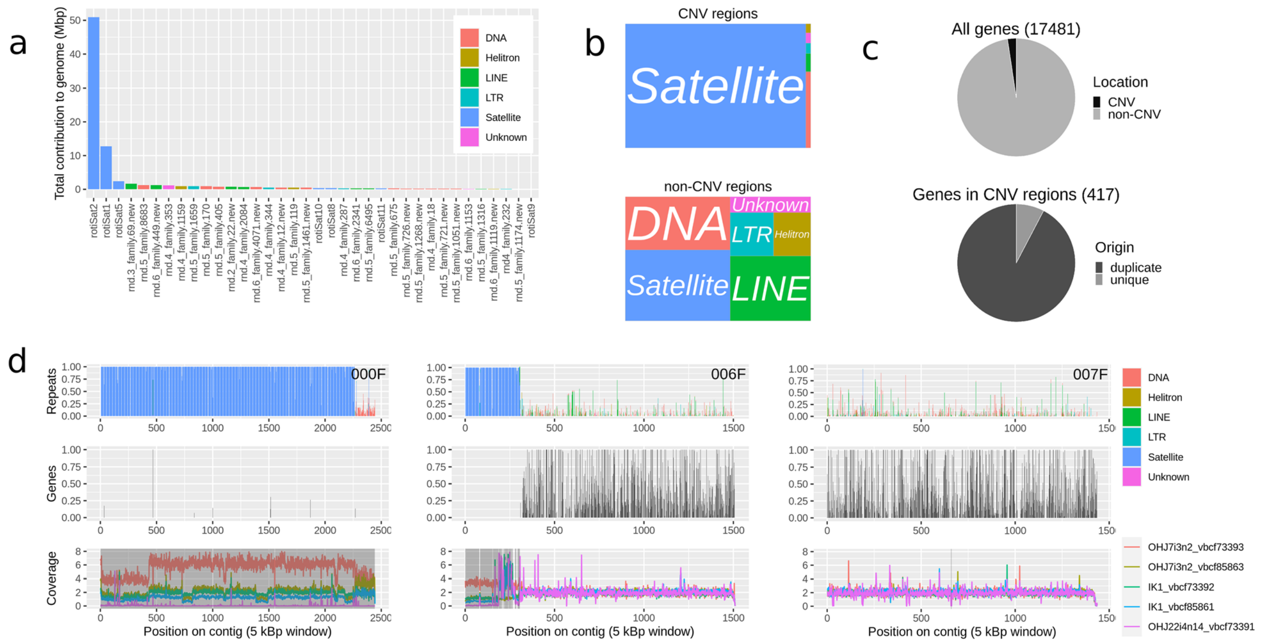


Fig. 6 Genome structure of *B. asplanchnoidis*. **a** The top 36 repeat elements ranked according to their contribution to the genome. **b** Differences in repeat element composition between CNV and non-CNV regions. **c** Differences in the distribution of gene content between CNV and non-CNV regions (top) and the respective proportion of genes derived from gene duplication events (bottom). **d** Three representative examples of contigs: a "B-contig" consisting almost exclusively of CNV regions (000F), a contig containing a large CNV region at one end (006F), and a non-CNV contig (007F). Top panels display the proportion of each 5-kb window occupied by repeats, intermediate panels display the proportion of each 5-kb window occupied by annotated genes, bottom panels display (exon-)normalized coverages of three rotifer clones (parents: OHJ7i3n2 and OHJ22i3n14, crossed offspring: IK1). Shaded regions indicate CNVs

Overall, these three satellites are the most abundant repeat elements in the *B. asplanchnoidis* genome, and their consensus sequences show the same characteristically elevated GC-contents compared to most other repeat elements (Additional file 6). We also observed one minor but distinct peak at ~48% GC in highSD regions, CNV regions, and B-contigs, which consisted of sequences that were not classified as repeats by repeatModeler2 (Additional file 1: Figure S15-S19).

Discussion

In this study, we provide a high-quality reference genome draft of the rotifer *Brachionus asplanchnoidis* with a total length of 230.1 Mbp. The 2C DNA content of the same rotifer clone is 568 Mbp, according to flow-cytometry-based estimates of an earlier study [7]. This earlier study provided evidence that the genome of this particular clone consists of a core haploid genome of 207Mbp and four copies of a segregating 34-Mb element. Assuming that our 230.1-Mbp assembly is completely unphased, our reference genome thus amounts to 95% of the flow-cytometry-based estimate of the haploid genome (241Mbp).

Aligning short reads of 15 rotifer clones from the same geographic population to the reference genome revealed multiple long tracts along the reference genome with increased coverage variation across clones. Additionally, we found that the average coverage at CNV regions strongly correlates with genome size. CVN regions also carried a distinct signature in terms of an increased GC-content (36% and 48%, respectively, versus 26% for the rest of the genome), which was both apparent in the short-read data and the genome assembly. Strikingly, many CNVs had near-zero coverage in three of the studied clones (OHJ82, OHJ22, OHJ22i3n14), which was independently confirmed by our PCR-based assays of two selected CNV regions. These results are highly consistent with previous evidence from flow cytometry experiments, showing that these three clones are characterized by a "basal" genome size (i.e., they are close to the smallest observed genome size in this species) and that they entirely lack the independently segregating genomic elements (ISEs) that are present in many other members of the OHJ population. Our results indicate the presence of multiple different ISEs in the OHJ population. For example, both our PCR and alignment data suggested that the two contigs 000F and 032F belong to two different ISEs because only the former was detectable in the clones OHJ98, OHJ104, and OHJ105 (Fig. 5). This observation is consistent with an earlier study, which suggested such diversity based on the size of ISEs measured by flow cytometry.

CNV regions account for megabase-long tracts in the genome of *B. asplanchnoidis*. Several contigs displayed highly similar coverage patterns across almost their entire length if one ignores the few and very short breakpoints. Such contigs might be fragments of even larger elements, perhaps B-chromosomes. Contig 032F might be a good candidate for this, ranging from zero to six copies across the OHJ population, as indicated by ddPCR. We also detected large CNVs, hundreds of kilobases in length, that were located on contigs with otherwise normal (diploid) coverage and low coverage variation (e.g., contig 006F in Fig. 6d). Such a genomic pattern is consistent with a stable diploid chromosome that contains a large homozygous insertion in some clones, hemizygous insertions in others, and a homozygous deletion in the remaining clones. Additionally, there might be length polymorphisms in such genomic regions, which are dominated by tandemly repeated satellite DNA. In the future, assembling multiple genomes from rotifer clones with different genome sizes, ideally using long-read technologies to approach chromosome-level assemblies [22], might allow a more precise delineation of individual CNVs into these two categories. Overall, our data is consistent with a mixture of B-chromosomes and large insertions into normal chromosomes, and possibly a dynamic exchange between both genomic fractions (since they are made up mostly by the same set of tandemly repeated satellite DNA; see below). Interestingly, GO enrichment analysis of genes that derived from a duplication event identified one term, GO: 0015074 (DNA integration), as the most significant term (p < 1e−30), which might indicate elevated transposon activity during the early evolution of the *B. asplanchnoidis* genome.

There are a few technical caveats and limitations to be considered in our analysis. First, although we found strong correlations between coverage variation at CNV regions and genome size variation, it was not possible to quantitatively “predict” the genome size of individual clones based on coverage along the reference assembly. The library preparation method seemed to introduce additional variation, specifically at the CNV regions, that prevented us from determining exact copy numbers of these genomic regions. This is a well-known limitation of short-read libraries from genomes that contain large amounts of satellite DNA [23], in particular when combined with heterogeneous GC-contents [24]. We could alleviate this limitation and determine exact copy numbers by using ddPCR, finding that the targeted genomic region (contig 032F) is present in zero to six copies across the OHJ population. This approach is a promising strategy for future studies to accurately assess copy number variation across many loci and in a large number of genomes from the same population. In addition, the differences in copy numbers at CNV regions could be estimated by comparing clones within a set of library preparations. For instance, if coverage at CNV regions is scaled by a reference clone (like clone “IK1” in Figures S8, S9), coverage at CNV regions of focal clones tends to fall into discrete clusters, indicating n-fold differences in coverage relative to the reference. A second limitation is that our reference genome might still not contain all the ISEs that contribute to genome size variation in the natural OHJ population, simply because it represents just a sample from that population. To obtain a more complete picture, more genomes of the OHJ population would be needed to be sequenced, preferably using long-read sequencing technologies that allow better identification of structural variants [25]. Third, one might argue that our CNV detection pipeline could miss many of the smaller insertions and deletions, in particular those smaller than 5 kb. However, it is quite unlikely that such small-scale structural variation has much influence on genome size variation in *B. asplanchnoidis*, since the percentage of discordantly aligned reads was overall rather low (1.3–5.1%) and it did not scale with genome size. With our reference genome being at the upper end of the genome size distribution of the OHJ population, we would expect to find a higher percentage of discordant reads in smaller genomes (mainly due to deletions), which was not the case.

CNV regions differed strongly from the remaining genome regarding repeat composition and gene density. Strikingly, most CNV regions were composed of only three satellite repeat elements, some unique to *B. asplanchnoidis* and others shared only with its closest congener, *B. plicatilis*, indicating recent evolutionary origin. The two most prominent satellite repeats, rotiSat1 and rotiSat2, consisted of 154- and 143-bp monomers that were arranged as tandem repeats of a length of up to a few megabases. The low sequence diversity of CNV regions explains the characteristic trimodal GC-content signature mentioned earlier, especially since the most abundant satellite elements display the same elevated GC-content (Additional file 6). Non-CNV parts of the genome contained a much higher diversity of other repeat elements, which included DNA transposons, LINEs, LTR elements, and Helitrons. Gene density was significantly lower in CNV regions compared to the rest of the genome, and genic regions in CNVs were typically confined to short stretches scattered across the contig. Interestingly, Orthofinder suggested that genes in CNV regions were three times more likely to derive from a duplication event than genes found in other places of the genome. This indicates that duplications had a role in the origin of these CNV regions, which incidentally has some resemblance to the proposed early evolution of B-chromosomes [26, 27].

Conclusions

In this study, we have identified, for the first time, genomic elements that can cause substantial within-population genome size variation, ultimately allowing investigation of genome size evolution at microevolutionary time scales. In the OHJ population of *B. asplanchnoidis*, these genomic elements consist of up to megabases-long arrays of satellite DNA, with only a few interspersed genes or other sequences. Though satellite arrays can form essential chromosome structures such as centromeres and telomeres [28], the large intrapopulation variation of these elements in *B. asplanchnoidis*, and their virtual absence in some individuals, suggests that, overall, these DNA additions do not provide an immediate fitness benefit to their carriers. Nevertheless, variable amounts of "bulk DNA" might influence the phenotype through subtler mechanisms independently of the DNA information content, for example through potential causal relationships between genome size and nucleus size, or cell size [15, 29, 30]. Increased levels of structural variation in a genome, as implied by our findings, may also constrain adaptive evolution or genome stability over microevolutionary time scales. In this regard, the genome of *B. asplanchnoidis* should be a valuable addition to existing models of genome evolution, enabling whole-genome analysis combined with experimental evolution approaches [31, 32] to disentangle the contributions of selection and drift to phenotypic change (and potentially concomitant changes in the genome size distribution) of evolving populations.

Methods

Origin of rotifer clones and DNA extraction

Resting eggs of rotifers were collected in the field from Obere Halbjochlacke (OHJ), a small alkaline playa lake in Eastern Austria (N 47° 47′ 11″, E 16° 50′ 31″). The hatchlings of these resting eggs were used to found clonal lines. Since resting eggs are produced sexually in monogonont rotifers, each clone has a unique genotype. Our clones from the OHJ population have been characterized previously concerning their genome size and other biological traits [7].

Rotifers were cultured in F/2 medium [33] at 16 ppt salinity and with *Tetraselmis suecica* algae as the food source (500–1000 cells μl^{-1}). Continuous illumination was provided with daylight LED lamps (SunStrip, Econlux) at 30–40 μmol quanta m^{-2} s^{-1} for rotifers, and 200 μmol quanta m^{-2} s^{-1} for algae. Stock cultures were kept either at 18 °C, re-inoculated once per week by transferring 20 asexual females to 20-mL fresh culture medium, or they were kept for long-term storage at 9°C, replacing approximately 80% of the medium with fresh food suspension every 4 weeks.

To produce biomass for DNA extraction, rotifers were cultured at 23°C in aerated borosilicate glass containers of variable size (250mL to 20L). Before DNA extraction, rotifers were starved overnight in sterile-filtered F/2 medium, with 2–3 additional washes of the sterile medium on the next day. In most preparations, we also added the antibiotics streptomycin (Sigma-Aldrich: S6501) and ampicillin (Sigma-Aldrich: A9518) to the washing medium, both with an end concentration of 50mg/mL. DNA was extracted using the Qiagen kits Dneasy (for short-read sequencing; from approximately 5000–7000 rotifers) and GenomicTips 100 (for long-read Pacbio sequencing; from >20,000 rotifers), and RNA was extracted from freshly prepared biomass with Rneasy.

Sequencing of the reference clone

We selected one rotifer clone (called: OHJ7i3n10) as DNA donor for the reference genome This clone ultimately derives from the natural population since it is a descendant of clone OHJ7, which was hatched from sediments of Obere Halbjochlacke. However, its immediate ancestors were passed through three generations of selfing (i.e., mating one male and female of the same clone). More details on the genealogy of this lineage and its biological characteristics can be found in [7]. Using long-read sequencing technology (PacBio SMRT® on the Sequel-platform), we obtained 16.3 Gbp from two SMRTcells, which corresponds to a 57-fold coverage assuming a haploid genome size of 284 Mbp. Additionally, we obtained 35.5 Gbp of Iso-Seq transcriptome data and 12 Gbp of short-read Illumina data of OHJ7i3n10. All sequencing and library preps related to the reference genome were performed by the Next Generation Sequencing Facility at Vienna BioCenter Core Facilities (VBCF), a member of the Vienna BioCenter (VBC), Austria.

Sequencing of individuals of the OHJ population

To characterize genomic variation across the OHJ population, we generated short-read sequencing data (Illumina platforms HiSeq and NextSeq) from 15 different rotifer clones, both from the natural OHJ population and from various clones of a selfed lineage (Additional file 1: Table S4). Clones were selected such that they covered the full range of genome size of the OHJ population. Short-read libraries were constructed using either the KAPA HyperPrep kit (Roche) or the NGS DNA Library Prep Kit (Westburg). The KAPA library preparations were done at the Marine Biological Laboratory [for more details, see [12]] while the Westburg preparations were done at VBCF. The two methods mainly differ in the fragmentation method (ultrasonic fragmentation in KAPA vs. enzymatic fragmentation in Westburg). Both methods are claimed to deliver sequence-independent fragmentation and to yield consistent coverage across a wide range of GC-contents. While the peak fragment size was ~400bp in both methods, we observed that the libraries constructed with the Westburg kit sometimes had a pronounced right tail with some fragments up to ~2000bp. We accounted for these larger fragments by adjusting the relevant parameters during short-read alignment (see below). In many of our clones, we used both library construction methods, yielding a total of 29 libraries (Additional file 1: Table S4).

Reference genome assembly and annotation

Pacbio sequences of the reference genome were assembled using the HGAP4 pipeline [34] at VBCF, and contamination was initially checked using CLARK [35] against all available bacterial genomes from NCBI (Additional file 2). We polished this initial VBCF assembly with short-read Illumina data of the identical *B. asplanchnoidis* clone using Pilon [36] in three rounds. To investigate the assembly quality, we backmapped the Illumina data to the genome assembly using bwa mem [37] and calculated summary statistics using QUAST [38] and Qualimap [39]. To provide an additional check for contamination, we used Blobtools [40] based on the backmapping alignments and a blastn search against the nt database (NCBI).To assess the assembly's completeness, we performed a BUSCO v4.0.6 [41] using the metazoan gene set (n = 978) in the *genome* mode applying the *–long* option.

Genes were structurally annotated on the repeat masked genome assembly using the MAKER2 annotation pipeline (v2.31.10, [42]) using evidence from Pacbio Iso-Seq transcripts of the same *B. asplanchnoidis* clone, protein homology evidence from the UniProt database (download January 2020, The UniProt Consortium 2017) in combination with proteoms of different clones and closely related *Brachionus* species (Additional file 1: Table S1), and ab initio gene predictions from SNAP [43], GeneMark-ES (v4.48_3.60_lic, [44]), and Augustus (v3.3.3, [45]). The SNAP model was initially trained on the genome assembly with the additional support of BUSCO complete hits (see above). GeneMark was computed in the ES suite on the soft masked genome assembly. The ab initio training of the Augustus model was computed on the genome assembly supported by the unassembled Iso-Seq transcripts. To compensate for underrepresented rotifer protein representation in the UniProt database, we combined this dataset with proteomes of another *B. asplanchnoidis* clone and four closely related *Brachionus* species (Table S1). The completeness of these proteomes was checked with BUSCO in the *protein* mode and missing orthologs were compared to determine if completeness could be increased through the combination of proteomes. The combined proteome finally contained 97% of the BUSCO genes of the metazoan dataset.

MAKER was run over three rounds. For the first round of MAKER, we only used the Iso-Seq data as EST evidence and the combined UniProt and proteome sequences as protein homology evidence applying default parameters and est2genome=1, protein2genome=1 to infer gene predictions directly from the transcripts and protein sequences. We used the gene prediction of round 1 to retrain the Augustus and SNAP model since the GeneMark model was exclusively computed on the genome assembly and did not require retraining. Maker was run in the second round using the retrained models and again est2genome=1, protein2genome=1 to increase prediction fidelity. After a second round of retraining the Augustus and SNAP model, MAKER was run through round 3 with switched off est2genome and protein2genome inference. After structural annotation via MAKER, we functionally annotated the genes using functional classification of genes from InterProScan [46] in combination with putative gene names derived from Swissprot [47].

To examine orthologous genes among rotifer species and identify gene duplication events, we used Orthofinder [48]. This analysis was based on proteome information of the *Brachionus plicatilis* species complex: *B. rotundiformis*, *B.* sp. "Tiscar," and *B. plicatilis* [respectively "Italy2," "TiscarSM28," and "Tokyo1" in [12]], and *B. asplanchniodis* (annotation of this study). Proteome information of *Adineta vaga* [49] was included as an outgroup. We extracted the longest transcript variant per gene to avoid duplicates in the input proteomes and followed the manual instructions of Orthofinder. To analyze the genomic distribution of genes that were identified to derive from gene duplication events, we extracted all genes of the *B. asplanchniodis* node (Orthofinder: SpeciesTree_Gene_Duplications) and removed duplicates to create a non-redundant list of genes. Comparing this list of genes to genes that are located inside or outside of genomic regions with high levels of CNVs allowed the estimation of non-random gene distribution patterns via Monte Carlo permutation tests (1000 permutations).

To characterize the putative functional properties of genes derived from a duplication event, we performed gene ontology (GO) enrichment analyses on the set of all multiple copy genes (n = 5634) and the more exclusive set of duplicated genes found within CNV regions (n = 385). The reference list consisted of 10,083 protein-coding genes (60% of all annotated genes) with GO annotation via INTERPROSCAN (see above). Enrichment analysis was done with the topGO R package (v2.24.0, [50]) in the category "biological process" using the weight01 algorithm and Fisher statistics (significance level $p < 0.05$).

Annotation of repetitive elements

The repeat library was produced using RepeatModeler2 (default settings, [51]) and the polished PacBio assembly. This produced 484 consensus sequences, 289 of which are "Unknown." These consensus sequences were then used to mask the genome assembly using RepeatMasker with default settings. From this, we identified the top contributing repeat elements and used their consensus sequences to blast queries against the genome assembly, and the top hits for each consensus were used to produce alignments for manual curation [52] and classification [53]. Satellite monomers were identified using Tandem Repeat Finder [54]. Consensus sequences of the satDNAs are dimers of these identified monomers.

Contributions per repeat were estimated by summing up the total length covered by each copy in the RepeatMasker output file. Top contributions were calculated over the whole assembly, and over regions of the genome with distinct coverage variability (lowSD, interSD, and highSD in Fig. 4). For each region, the top 20 repeats were included, resulting in a total of 38 repeats to curate. Curation was done by manually inspecting each alignment, identifying the ends of the aligning regions, and producing a new consensus sequence. Additionally, classification was performed by searching for TSDs, TIRs, LTRs, satellite structure, and conserved domains. RepBase searches were rarely constructive since rotifer, and especially monogonont, TEs are not well-represented in any databases. The final curated library of topREs contains 37 consensus sequences from 36 elements (one sequence was removed because alignments were to only scattered AT-rich regions, one was an rRNA gene, and one LTR sequence was split into the LTR and internal portions). Redundant sequences between the topRE library and the RepeatModeler library were identified using RepeatMasker. Sequences that were at least 95% identical and covered at least 80% of the uncurated repeat were removed from the uncurated repeats before merging the libraries. Due to the short length and difficulties in automatically creating consensus sequences for satDNA elements, all uncurated elements that were identified as similar by RepeatMasker, regardless of similarity or length of hit, were aligned to the curated consensus sequence and visually inspected for alignment. Uncurated elements that aligned across most of the curated satDNA consensus sequences were removed. The combined library (Additional file 8) of curated and uncurated elements contained 472 elements (262 unknowns) and was used to mask the *B. asplanchnoidis* genome and to repeat-mask related, high-quality *Brachionus* genome assemblies to identify shared elements [55–57]. For each of the top contributing elements and each genomic region, an enrichment index was calculated as the proportion of each repeat contribution in that region vs. the whole genome (i.e., repeat contribution in region/total contribution of repeat) divided by the proportion of the genome represented by each region (i.e., region size/assembly size). If this index was over 1, it meant that the repeat in question was enriched in the region in question.

Preprocessing of short-read data

Trimming and adaptor removal was done using bbduk (v38.34, https://jgi.doe.gov/data-and-tools/bbtools/) with the settings: k = 23 ktrim = n mink = 11 hdist = 1 tpe tbo qtrim = rl trimq = 20 maq = 10 minlen = 40. Trimmed reads were initially aligned to the reference genome using bowtie2 [58]. Preliminary tests with various parameter settings (local vs. end-to-end alignment, different settings for the sensitivity parameter, map as single reads vs. map as paired reads) did not indicate strong differences. Thus, we used the default values for most parameters, except for fragment size (parameter: X), which was set to the maximum observed fragment size of each library instead of the default value of 500 bp (see Additional file 1: Table S4).

We included three steps to remove contaminants, since DNA extracted from the whole bodies of microscopic organisms might contain DNA from other organisms, such as bacteria. First, we extracted all unmapped reads from an initial alignment to the reference genome and screened only those for potential contaminant reads. Thus, we considered the mapped fraction as "rotifer DNA," since they mapped to the (contaminant-free) assembly. Second, the unmapped reads of all libraries were combined and assembled using metagenomic approaches. For this assembly, we used metaVelvet (v1.2.02, [59]) with a kmer length of 101bp, an insert size of 500 (± 200 standard deviation), and a minimum reported contig length of 300bp. The resulting assembly was then analyzed with metaQuast [60] using the option "automatic pulling of reference sequences," which restricts the search to bacteria and archaea. Subsequently, the complete genomes of putative contaminants were downloaded, and the unmapped reads were mapped against those genomes. In the third step, the remaining fraction (i.e., reads not mapping to the microbial metagenome) were subjected to a kmer-based identification using the kraken2 pipeline [61], with the databases bacteria, archaea, viral, UniVec-Core, and protozoa. We also performed checks on the false-discovery rate of kraken2, by running the same pipeline on reads that initially did map to the reference genome. Finally, all unmapped reads that could *not* be taxonomically assigned to contaminants, with either of the two approaches above, were considered to be of "rotifer origin" and were merged with the mapped fraction. Those "rotifer reads" were then further cleaned by removing mitochondrial DNA, which was done by mapping them to the published mitochondrial genome of *B. plicatilis* [62], and by removal of duplicates using FastUniq [63]. These "final reads" were again mapped to the reference genome, or they were analyzed using reference-free approaches that do not require alignment.

Analysis of unaligned reads

"Final reads" were subjected to kmer-based analyses using a kmer size of 21bp with jellyfish (v2.1.4, [64]). Then, GenomeScope 2.0 [20] and findGSE [21] were used to obtain kmer-based estimates of coverage, heterozygosity, and genome size. Per-base coverages C were computed from the kmer-coverages

C_k with the formula:

$$C = \frac{C_K \cdot R}{(R - K + 1)}$$

where R is the average read size, obtained from dividing the total number of base pairs in each library by the total number of reads. The coverage estimates from these two programs were contrasted with the naïve coverage estimate, based on sequencing effort (total number of bp in a library) and 1C genome size estimated from flow cytometry, assuming a diploid genome.

To analyze GC-distributions among short-read libraries, the GC-contents of individual reads were extracted using the function fx2tab of SeqKit [65]. The resulting csv files were further analyzed with the R-package mixtools [66] using the function normalmixEM.

Analysis of copy number variation

Analysis of copy number variation was done separately for all 29 short-read libraries from 15 different clones of the rotifer *B. asplanchnoidis*. Average per-base depth-of-coverage (DOC) values along 50-kbp and 5-kbp windows, respectively, were extracted from the BAM alignment files ("final reads" to reference genome) using the samtools function "bedcov" [67]. In total, there are 4835 windows at 50-kbp resolution and 46,255 windows at 5-kbp resolution in the current genome assembly. To allow comparisons among clones and libraries, DOC was normalized by dividing the per-bp coverage of each window by 1/2 of the (mean) exon coverage of the respective library. In unphased sections of the genome assembly, we expected DOC values of around 2, provided that both alleles of a (diploid) genome map to the correct location. Coverage variation was quantified as the standard deviation of DOC per window (50kbp or 5kbp) across all 29 libraries.

To identify individual CNVs and to locate possible breakpoints within contigs, we used a custom-written R-algorithm involving the following criteria for merging adjacent windows (5kbp) based on the similarity of coverage patterns. First, the coverage variation (measured as the standard deviation of per-base normalized coverage across all libraries) had to be above a defined threshold (i.e., 0.7, which was an a posteriori determined threshold). Second, the read depths of each library of both windows had to be significantly correlated with each other at the $p < 0.05$ level. This was done by calculating the partial correlation coefficients. If both conditions applied, those two windows were considered to belong to the same CNV. The very first and the last window of each contig, which often showed deviant coverage patterns, were merged with their neighboring window. Third, adjacent CNVs identified according to the above criteria were merged if they were separated by only one window (CNV stopping breakpoint) AND if the coverages in the two windows surrounding the breakpoint were significantly correlated with each other. Finally, we only considered CNVs with lengths of at least three adjacent windows. Thus, we obtained a table of all CNVs along the genome of *B. asplanchnoidis*, together with their size, and their location on individual contigs (i.e., in the middle of the contig, at the edges, or spanning the entire contig). Data analysis related to coverage variation and CNV detection was done using custom-written algorithms in the R environment [68] with the base package (v3.6.3) and the add-on packages stringr v1.4.0 [69] and reshape2 v1.4.4 [70]. For graphical visualization, we used ggplot2 v3.3.0 [71] and the add-on packages cowplot v1.0.0 [72] and treemapify v2.5.5 [73].

PCR confirmation of CNV regions

To independently confirm the presence or absence of CNV regions in different rotifer clones, and to estimate the copy numbers of these genomic regions, we used PCR-based methods. To identify unique PCR-primer binding sites in regions of high or low coverage, we used Thermoalign [74] searches in multiple regions of 5000-bp length, spread across the genome and on different contigs. The exact search parameters for Thermoalign are given in Additional file 9. Candidate primers were tested and optimized using PCR on template DNA from different OHJ clones, including the reference clone OHJ7i3n10. PCR reactions consisted of 25μl HotStart Taq master mix (Qiagen), 0.1μM Primer, 3mM $MgCl_2$, and 20 ng template DNA. PCR cycling conditions were 95°C for 15 min, 30 cycles of 94°C for 20 s, 56°C for 20 s, and 68°C for 10s, followed by 68°C for 5 min and hold at 4°C. Agarose gels were used to test for the presence or absence of the associated loci across different members of the OHJ population. In total, we screened four loci located on different contigs of the reference genome assembly. Two of them were located in copy number-invariable, diploid regions of the genome (according to the coverage estimates from the short-read alignments), and two were located in CNV regions (Additional file 1: Table S7). In addition to the qualitative PCR test, we used ddPCR to estimate the copy number of the CNV loci for each rotifer clone. For each 22-μl reaction, we used EvaGreen Supermix (Biorad), 0.15μM Primer, 4 units EcoRI, and template DNA equivalent of 1500 genome copies. PCR cycling conditions were 95°C for 5 min, 40 cycles of 95°C for 30 s, and 61°C for 1 min with a ramp rate of 2°C/s; 4°C for 5 min, followed by 90°C for 5 min and hold at 4°C. For droplet generation and fluorescence readout, we used a QX200 Droplet Generator and Droplet Reader (Biorad), respectively. Copy numbers (*CN*) of CNV loci were estimated for each rotifer clone using the ratios of amplicon molecule concentrations, which were obtained with the QuantaSoft Software (Biorad):

$$CN = \frac{T}{(R_1 + R_2)/2} N_R$$

where T is the amplicon concentration of the target locus (the one showing copy number variation across the OHJ population), R_1 and R_2 are the amplicon concentrations of the two reference loci, and N_R is the number of copies of the reference loci in the genome (in this case, N_R=2, since the reference loci were both diploid). The 95% confidence intervals obtained from QuantaSoft were used as an indication of the measurement error.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-021-01134-w.

Additional file 1. Supplementary figures and tables.

Additional file 2. Report on genome assembly and contaminant filtering provided by VBCF (Vienna Biocenter Core Facility).

Additional file 3. A summary of short-read preprocessing and fastqc-reports.

Additional file 4. Kmer-based analysis of cleaned Illumina reads.

Additional file 5 Ranges of all CNVs across the *B. asplanchnoidis* genome.

Additional file 6. Detailed information on top-36 contributing repeat elements.

Additional file 7. Repeat profile, Gene density, and CNVs of the 50 largest contigs.

Additional file 8. Combined library of curated and uncurated repeat elements.

Additional file 9. Input parameters of the thermoalign pipeline.

Acknowledgements

We thank Hermann Schwärzler for his bioinformatic support in establishing the Thermoalign pipeline. Pia Scheu and Martina Nindl provided valuable advice for the optimization of ddPCR conditions. The participants of the "Uppsala TE Jamboree" provided support in classifying repetitive elements.

Authors' contributions

CPS, JB, and DMW conceived the study. BHL contributed to library preparations, sequencing, and the preliminary quality control of some of the samples. AMW contributed to gene annotation, orthofinder, and GO enrichment analysis. JB contributed the annotation of repetitive elements. MP contributed ddPCR analysis. CPS performed the preprocessing of short-read data, read-based analyses, alignments, and bioinformatic analyses related to coverage variation and wrote the draft. CPS, JB, AMW, and DMW revised the manuscript. All authors read and approved the final manuscript.

Funding

Funding for this project was provided by the Austrian Science Fund, grant number P26256 (PI: CPS), and by a scholarship for the promotion of young researchers by the University of Innsbruck to JB. JB is currently supported by funding from SciLifeLab, Uppsala.

Availability of data and materials

Raw sequencing reads, genome assembly, and the annotation files were deposited at the NCBI under the BioProject PRJNA755169 [75].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Research Department for Limnology, University of Innsbruck, Mondsee, Austria. [2]Department of Organismal Biology, Uppsala University, Uppsala, Sweden. [3]Institute of Zoology, University of Cologne, Cologne, Germany. [4]Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA, USA.

Received: 7 June 2021 Accepted: 27 August 2021
Published online: 16 September 2021

References

1. Elliott TA, Gregory TR. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. Philos Trans R Soc B Biol Sci. 2015; 370(1678). https://doi.org/10.1098/rstb.2014.0331.
2. Jeffery NW, Hultgren K, Chak STC, Gregory R, Rubenstein DR. Patterns of genome size variation in snapping shrimp. Genome. 2016;59(6):393–402. https://doi.org/10.1139/gen-2015-0206.
3. Stelzer CP, Riss S, Stadler P. Genome size evolution at the speciation level: the cryptic species complex *Brachionus plicatilis* (Rotifera). BMC Evol Biol. 2011;11(1). https://doi.org/10.1186/1471-2148-11-90.
4. Chia J-M, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, et al. Maize HapMap2 identifies extant variation from a genome in flux. Nat Genet. 2012; 44(7):803–7. https://doi.org/10.1038/ng.2313.
5. Ruiz-Ruano FJ, Ruiz-Estevez M, Rodriguez-Perez J, Lopez-Pino JL, Cabrero J, Camacho JPM. DNA amount of X and B chromosomes in the grasshoppers *Eyprepocnemis plorans* and *Locusta migratoria*. Cytogenet Genome Res. 2011;134(2):120–6. https://doi.org/10.1159/000324690.
6. Šmarda P, Bureš P, Horová L, Rotreklová O. Intrapopulation genome size dynamics in *Festuca pallens*. Ann Bot. 2008;102(4):599–607. https://doi.org/10.1093/aob/mcn133.
7. Stelzer CP, Pichler M, Stadler P, Hatheuer A, Riss S. Within-population genome size variation is mediated by multiple genomic elements that segregate independently during meiosis. Genome Biol Evol. 2019;11(12): 3424–35. https://doi.org/10.1093/gbe/evz253.
8. Meyer A, Schloissnig S, Franchini P, Du K, Woltering J, Irisarri I, et al. Giant lungfish genome elucidates the conquest of land by vertebrates. Nature. 2021;590(7845):284–9. https://doi.org/10.1038/s41586-021-03198-8.
9. Shah A, Hoffman JI, Schielzeth H. Comparative analysis of genomic repeat content in gomphocerine grasshoppers reveals expansion of satellite DNA

and Helitrons in species with unusually large genomes. Genome Biol Evol. 2020;12(7):1180–93. https://doi.org/10.1093/gbe/evaa119.

10. Naville M, Henriet S, Warren I, Sumic S, Reeve M, Volff JN, et al. Massive changes of genome size driven by expansions of non-autonomous transposable elements. Curr Biol. 2019;29(7):1161.

11. Wong WY, Simakov O, Bridge DM, Cartwright P, Bellantuono AJ, Kuhn A, et al. Expansion of a single transposable element family is associated with genome-size increase and radiation in the genus *Hydra*. Proc Natl Acad Sci. 2019;116(46):22915–7. https://doi.org/10.1073/pnas.1910106116.

12. Blommaert J, Riss S, Hecox-Lea B, Mark Welch DB, Stelzer CP. Small, but surprisingly repetitive genomes: transposon expansion and not polyploidy has driven a doubling in genome size in a metazoan species complex. BMC Genomics. 2019;20(466).

13. McCann J, Macas J, Novák P, Stuessy TF, Villaseñor JL, Weiss-Schneeweiss H. Differential genome size and repetitive DNA evolution in diploid species of *Melampodium* sect. *Melampodium* (Asteraceae). Front Plant Sci. 2020;11:362. https://doi.org/10.3389/fpls.2020.00362.

14. Smarda P, Bures P. Understanding intraspecific variation in genome size in plants. Preslia. 2010;82(1):41–61.

15. Stelzer C-P, Pichler M, Hatheuer A. Linking genome size variation to population phenotypic variation within the rotifer *Brachionus asplanchnoidis*. Commun Biol. 2021;4(1):596. https://doi.org/10.1038/s42003-021-02131-z.

16. Nogrady T, Wallace RL, Snell TW. Rotifera: biology, ecology and systematics, vol. 1. The Hague: SPB Academic Publishing; 1993.

17. Gilbert JJ. Non-genetic polymorphisms in rotifers: environmental and endogenous controls, development, and features for predictable or unpredictable environments. Biol Rev. 2017;92(2):964–92. https://doi.org/10.1111/brv.12264.

18. Riss S, Arthofer W, Steiner FM, Schlick-Steiner BC, Pichler M, Stadler P, et al. Do genome size differences within *Brachionus asplanchnoidis* (Rotifera, Monogononta) cause reproductive barriers among geographic populations? *Hydrobiologia*. 2017;796(1):59–75. https://doi.org/10.1007/s10750-016-2872-x.

19. Assembly statistic visualization [https://github.com/rjchallis/assembly-stats].

20. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. 2020;11(1):1432.

21. Sun H, Ding J, Piednoël M. Schneeberger K: findGSE: estimating genome size variation within human and *Arabidopsis* using k-mer frequencies. Bioinformatics. 2017;34(4):550–7.

22. Simion P, Narayan J, Houtain A, Derzelle A, Baudry L, Nicolas E, et al. Homologous chromosomes in asexual rotifer *Adineta vaga* suggest automixis. bioRxiv. 2020; https://doi.org/10.1101/2020.06.16.155473.

23. Lower SS, McGurk MP, Clark AG, Barbash DA. Satellite DNA evolution: old ideas, new approaches. Curr Opin Genet Dev. 2018;49:70–8. https://doi.org/10.1016/j.gde.2018.03.003.

24. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012;40(10):e72. https://doi.org/10.1093/nar/gks001.

25. De Coster W, De Rijk P, De Roeck A, De Pooter T, D'Hert S, Strazisar M, et al. Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. Genome Res. 2019;29(7):1178–87. https://doi.org/10.1101/gr.244939.118.

26. Ruiz-Ruano FJ, Navarro-Domínguez B, López-León MD, Cabrero J, Camacho JPM. Evolutionary success of a parasitic B chromosome rests on gene content. bioRxiv. https://doi.org/10.1101/683417.

27. Ahmad SF, Martins C. The modern view of B chromosomes under the impact of high scale omics analyses. Cells. 2019;8(2):156. https://doi.org/10.3390/cells8020156.

28. Garrido-Ramos MA. Satellite DNA: an evolving topic. Genes (Basel). 2017;8(9):230. https://doi.org/10.3390/genes8090230.

29. Cavalier-Smith T. Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. Ann Bot. 2005;95(1):147–75. https://doi.org/10.1093/aob/mci010.

30. Gregory TR. The bigger the C-value, the larger the cell: genome size and red blood cell size in vertebrates. Blood Cell Mol Dis. 2001;27(5):830–43. https://doi.org/10.1006/bcmd.2001.0457.

31. Fussmann G. Rotifers: excellent subjects for the study of macro- and microevolutionary change. Hydrobiologia. 2011;662(1):11–8. https://doi.org/10.1007/s10750-010-0515-1.

32. Declerck SAJ, Papakostas S. Monogonont rotifers as model systems for the study of micro-evolutionary adaptation and its eco-evolutionary implications. Hydrobiologia. 2017;796(1):131–44. https://doi.org/10.1007/s10750-016-2782-y.

33. Guillard RRL. Culture of phytoplankton for feeding marine invertebrates. In: Smith WL, Chanley MH, editors. Culture of marine invertebrate animals. New York: Pleum Pub. Co.; 1975. p. 29–60.

34. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013;10(6):563–9. https://doi.org/10.1038/nmeth.2474.

35. Ounit R, Wanamaker S, Close TJ, Lonardi S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics. 2015;16(1):236. https://doi.org/10.1186/s12864-015-1419-2.

36. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9(11):e112963. https://doi.org/10.1371/journal.pone.0112963.

37. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60. https://doi.org/10.1093/bioinformatics/btp324.

38. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics. 2018;34(13):i142–50. https://doi.org/10.1093/bioinformatics/bty266.

39. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 2016;32(2):292–4. https://doi.org/10.1093/bioinformatics/btv566.

40. Laetsch D, Blaxter M. BlobTools: jnterrogation of genome assemblies [version 1; peer review: 2 approved with reservations]. F1000Research. 2017;6(1287).

41. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2. https://doi.org/10.1093/bioinformatics/btv351.

42. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008;18(1):188–96. https://doi.org/10.1101/gr.6743907.

43. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004;5(1):59. https://doi.org/10.1186/1471-2105-5-59.

44. Brůna T, Lomsadze A, Borodovsky M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. NAR Genom Bioinform. 2020;2(2):lqaa026.

45. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006;34(suppl_2):W435–9.

46. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9):1236–40. https://doi.org/10.1093/bioinformatics/btu031.

47. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, et al. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. Methods Mol Biol. 2016;1374:23–54. https://doi.org/10.1007/978-1-4939-3167-5_2.

48. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20(1):238. https://doi.org/10.1186/s13059-019-1832-y.

49. Flot J-F, Hespeels B, Li X, Noel B, Arkhipova I, Danchin EGJ, et al. Evidence for the absence of meiosis from the genome of the bdelloid rotifer *Adineta vaga*. Nature. 2013;500(7463):453–7. https://doi.org/10.1038/nature12326.

50. topGO: Enrichment analysis for gene ontology [http://bioconductor.org/packages/release/bioc/html/topGO.html].

51. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci. 2020;117(17):9451–7. https://doi.org/10.1073/pnas.1921046117.

52. Platt RN II, Blanco-Berdugo L, Ray DA. Accurate transposable element annotation is vital when analyzing new genome assemblies. Genome Biol Evol. 2016;8(2):403–10. https://doi.org/10.1093/gbe/evw009.

53. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8(12):973–82. https://doi.org/10.1038/nrg2165.

54. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27(2):573–80. https://doi.org/10.1093/nar/27.2.573.
55. Park JC, Choi B-S, Kim M-S, Shi H, Zhou B, Park HG, et al. The genome of the marine rotifer *Brachionus koreanus* sheds light on the antioxidative defense system in response to 2-ethyl-phenanthrene and piperonyl butoxide. Aquat Toxicol. 2020;221:105443. https://doi.org/10.1016/j.aquatox.2020.105443.
56. Han J, Park JC, Choi B-S, Kim M-S, Kim H-S, Hagiwara A, et al. The genome of the marine monogonont rotifer *Brachionus plicatilis*: genome-wide expression profiles of 28 cytochrome P450 genes in response to chlorpyrifos and 2-ethyl-phenanthrene. Aquat Toxicol. 2019;214:105230. https://doi.org/10.1016/j.aquatox.2019.105230.
57. Kim H-S, Lee B-Y, Han J, Jeong C-B, Hwang D-S, Lee M-C, et al. The genome of the freshwater monogonont rotifer *Brachionus calyciflorus*. Mol Ecol Resour. 2018;18(3):646–55. https://doi.org/10.1111/1755-0998.12768.
58. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9. https://doi.org/10.1038/nmeth.1923.
59. Namiki T, Hachiya T, Tanaka H, Sakakibara Y. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res. 2012;40(20):e155. https://doi.org/10.1093/nar/gks678.
60. Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: evaluation of metagenome assemblies. Bioinformatics. 2016;32(7):1088–90. https://doi.org/10.1093/bioinformatics/btv697.
61. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20(1):257. https://doi.org/10.1186/s13059-019-1891-0.
62. Suga K, Mark Welch DB, Tanaka Y, Sakakura Y, Hagiwara A. Two circular chromosomes of unequal copy number make up the mitochondrial genome of the rotifer *Brachionus plicatilis*. Mol Biol Evol. 2008;25(6):1129–37. https://doi.org/10.1093/molbev/msn058.
63. Xu H, Luo X, Qian J, Pang X, Song J, Qian G, et al. FastUniq: a fast *de novo* duplicates removal tool for paired short reads. PLoS One. 2012;7(12):e52249. https://doi.org/10.1371/journal.pone.0052249.
64. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27(6):764–70. https://doi.org/10.1093/bioinformatics/btr011.
65. Shen W, Le S, Li Y, Hu F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One. 2016;11(10):e0163962. https://doi.org/10.1371/journal.pone.0163962.
66. Benaglia T, Chauveau D, Hunter DR. Young DS: mixtools: an R package for analyzing mixture models. J Stat Softw. 2009;32(6):29.
67. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352.
68. R Development Core Team: R: a language and environment for statistical computing. 2020.
69. stringr: simple, consistent wrappers for common string operations [https://CRAN.R-project.org/package=stringr].
70. Hadley W. Reshaping data with the reshape package. J Stat Soft. 2007;21(12):1–20.
71. ggplot2: elegant graphics for data analysis [https://ggplot2.tidyverse.org].
72. cowplot: streamlined plot theme and plot annotations for 'ggplot2' [https://CRAN.R-project.org/package=cowplot].
73. treemapify: draw treemaps in 'ggplot2' [https://CRAN.R-project.org/package=treemapify].
74. Francis F, Dumas MD, Wisser RJ. ThermoAlign: a genome-aware primer design tool for tiled amplicon resequencing. Sci Rep. 2017;7(1):44437. https://doi.org/10.1038/srep44437.
75. Brachionus asplanchnoidis genome sequencing. NCBI BioProject accession: PRJNA755169. 2021. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA755169.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.