Genome structure of *Brachionus asplanchnoidis*, a Eukaryote with intrapopulation variation in genome size

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

[1] Research Department for Limnology, University of Innsbruck, Mondsee, Austria

[2] Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA, USA

[3] Department of Organismal Biology, Uppsala University, Uppsala, Sweden

[4] Institute of Zoology, University of Cologne, Zülpicher Str. 47b, 50674 Cologne, Germany

[*] Corresponding author (claus-peter.stelzer@uibk.ac.at)

Abstract

Eukaryotic genomes vary greatly in size due to variation in the proportion of non-coding DNA, a pattern that emerges both in comparisons at a larger taxonomic scale and at the level of individuals within a species. The rotifer *Brachionus asplanchnoidis* represents one of the most extreme cases of intraspecific genome size variation among Eukaryotes, displaying almost 2-fold variation within a geographic population. 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. We identified large, contiguous copy number variable regions (CNVs), which exhibited significant coverage differences among individuals, and whose coverage overall scaled with genome size. CNVs were mainly composed of tandemly repeated satellite DNA, with only few interspersed genes or other sequences, and were characterized by an elevated GC-content. Judging from their distributions across contigs, some CNVs are fragments of accessory (B-)chromosomes while others resemble large insertions to normal chromosomes. 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. Our study provides unprecedented insights into genome size evolution at microevolutionary time scales and thus paves the way for studying genome size evolution in contemporary populations rather than inferring patterns and processes *a posteriori* from species comparisons.

42 # Introduction

43 The genomes of Eukaryotic organisms display remarkable diversity in size, overall spanning
44 approximately five orders of magnitude [1]. In addition, genome size may vary substantially among
45 closely related species [2, 3], within a species [e.g., 4, 5, 6], and sometimes even within a population
46 [7]. Most of the variation in genome size stems from differences in the proportion of various kinds of
47 non-coding DNA and/or transposable elements, which can reach excessive levels in species with giant
48 genomes [8, 9]. Studying genome size variation at the DNA-sequence level allows identification of
49 exactly those genomic elements that make up for the genome size difference, and it can suggest the
50 relative strength of mutation, selection, and drift - the underlying evolutionary forces ultimately
51 causing divergence in genome size.

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

66 Intraspecific genome size variation (IGV) has been described in several species of Eukaryotes [some
67 examples are summarized in 7, 14]. IGV may be associated with variation in the number of
68 chromosomes [e.g., B-chromosomes 14], but there are also examples where IGV is not reflected in
69 the karyotype [2, 6]. One of the most recent additions to IGV model organisms was the monogonont
70 rotifer *B. asplanchnoidis*, which displays a nearly 2-fold variation in genome size even among
71 individuals within a geographic population [7]. Monogonont rotifers are short-lived (1-2 weeks),
72 small aquatic metazoans, only a few hundred micrometers in size, common in fresh and brackish
73 water habitats throughout the world. They have a life-cycle involving cyclical parthenogenesis [15]
74 reproducing by ameiotic parthenogenesis for prolonged periods and inducing sexual reproduction
75 occasionally. A 'rotifer clone' consists of the asexual descendants of a single female that has hatched

76 from a single resting egg, which itself is the product of sexual reproduction. In many Monogonont
77 species, sex is triggered by crowding due to accumulation of mixis-inducing peptides released by the
78 animals [16]. In lab cultures, it is possible to suppress sexual reproduction by frequent dilution
79 intervals or large culture volumes, and to induce mixis in small culture volumes or through the use of
80 media drawn from dense cultures. Thus, it is possible to either deliberately cross two rotifer clones
81 sexually, or to keep them clonally for hundreds of generations.

82 In the present study, we focus on a population of *B. asplanchnoidis* from Obere Halbjochlacke (OHJ),
83 a shallow alkaline lake in Eastern Austria [7, 17]. Individuals of this population can be crossed with
84 each other - even if they substantially differ in genome size - and they will produce offspring with
85 intermediate genome sizes close to the parental mean. Genome size can be artificially selected up or
86 down with a heritability of 1 by breeding only individuals with large or small genome size. Genome
87 size variation in this system is mediated by relatively large genomic elements (several megabases in
88 size), which segregate independently from each other during meiosis. The smallest observed genome
89 size in *B. asplanchnoidis* was 404Mb (2C, nuclear DNA content). Individuals at or close to this basal
90 genome size are completely lacking independently segregating elements, while in larger individuals,
91 genome size scales with the amount of independently segregating elements [7].

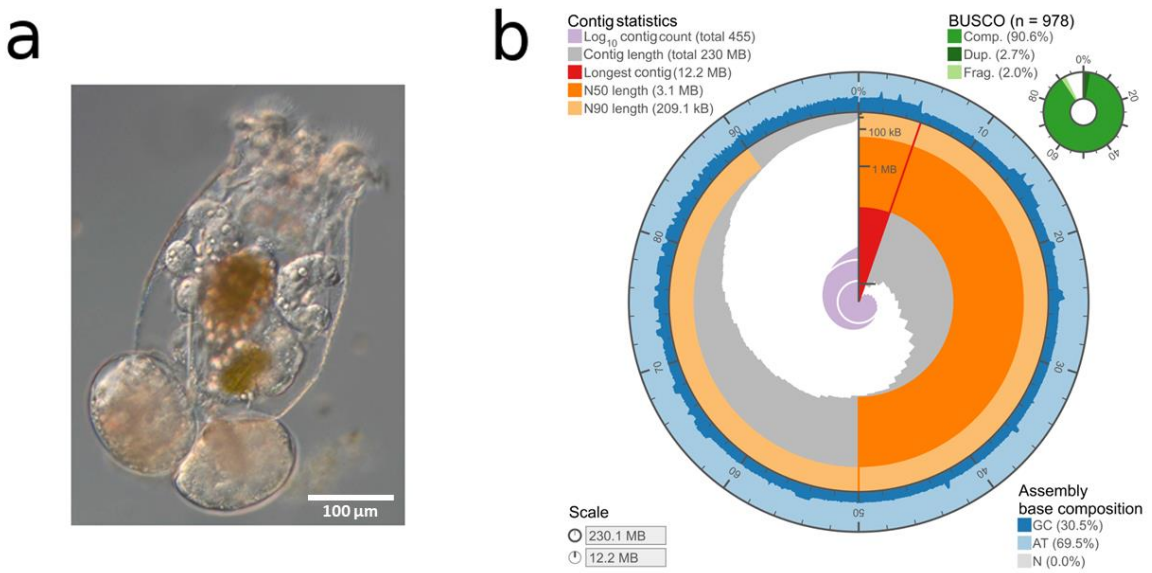
92 Here we used a whole-genome sequencing approach to identify the DNA sequence differences
93 responsible for intrapopulation genome size variation in this population of *B. asplanchnoidis*. Our
94 specific goal was to identify and characterize genomic regions that are present in one or multiple
95 copies in some individuals of the OHJ-population, but are missing in others. To this end, we
96 assembled a highly contiguous draft genome of a reference clone using long-read (PacBio)
97 technology and then mapped short-reads of 15 different clones with genome sizes from 404 to 644
98 Mbp to this assembly. To identify copy number variations (CNVs), we scanned for regions of
99 increased per-bp read coverage. To independently confirm CNVs, we used PCR to detect
100 presence/absence of selected CNVs across different clones of the OHJ-population, and droplet digital
101 PCR to determine the exact copy numbers of one specific locus. Finally, we annotated genes and
102 repetitive elements in the reference genome, and compared CNV regions to non-variable regions of
103 the genome.

104

105 # Results

106 **De novo assembly and annotation of the reference genome**

107 The rotifer clone (OHJ7i3n10) chosen for our reference genome derives from the natural isolate OHJ7
108 after three rounds of selfing (i.e., fertilizing sexual females by males of the same clone). As measured
109 by flow cytometry [7], OHJ7i3n10 has a 2C-genome size of 568 Mbp, and thus contains approximately
110 40% excess DNA, compared to the smallest genome size of the OHJ-population (~410 Mbp). The
111 total length of our reference assembly was 230.12 Mb, with 455 contigs and an N50 value of 3.065
112 Mb (**Fig. 1, Supplementary file 1**). Average GC content was 30.5%. However, the GC distribution was
113 not unimodal, but showed two major peaks at ~25% and ~35% GC, and a minor one at ~50% GC (**Fig.
114 S1**).



115

116 **Fig. 1 a** Photograph of *B. asplanchnoidis* female (with two attached asexual eggs). **b** Circular
117 visualization of the contiguity and completeness of the genome assembly. Full circle represents
118 the full assembly length of 230.1 Mb with the longest contig of 12.2 Mb in red and very few
119 contigs <100 kB. GC-content is varying along the assembly (blue).
120

121 Taxonomic partitioning of the polished genome assembly confirmed its purity. Most hits could
122 correctly be assigned to rotifers and remaining hits assigned to mollusks and arthropods can mostly
123 be explained by imbalanced availability of rotifer entries in the *nt* database (**Fig. S2, Supplementary
124 file 1**). We observed that 90.6 % of the metazoan BUSCO gene set collection was complete with low
125 levels of duplicated (2.7 %), fragmented (2.0 %), and missing (7.4 %) genes (**Tab. S2**). A visualization
126 of assembly contiguity and completeness was generated via assembly-stats [18] and is presented in
127 **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 (**Tab. 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 (Tab. 1, Tab. 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-fold for the different libraries, with the majority of libraries being above 20-fold (**Supplementary note, Fig. S3, Supplementary file 2**).

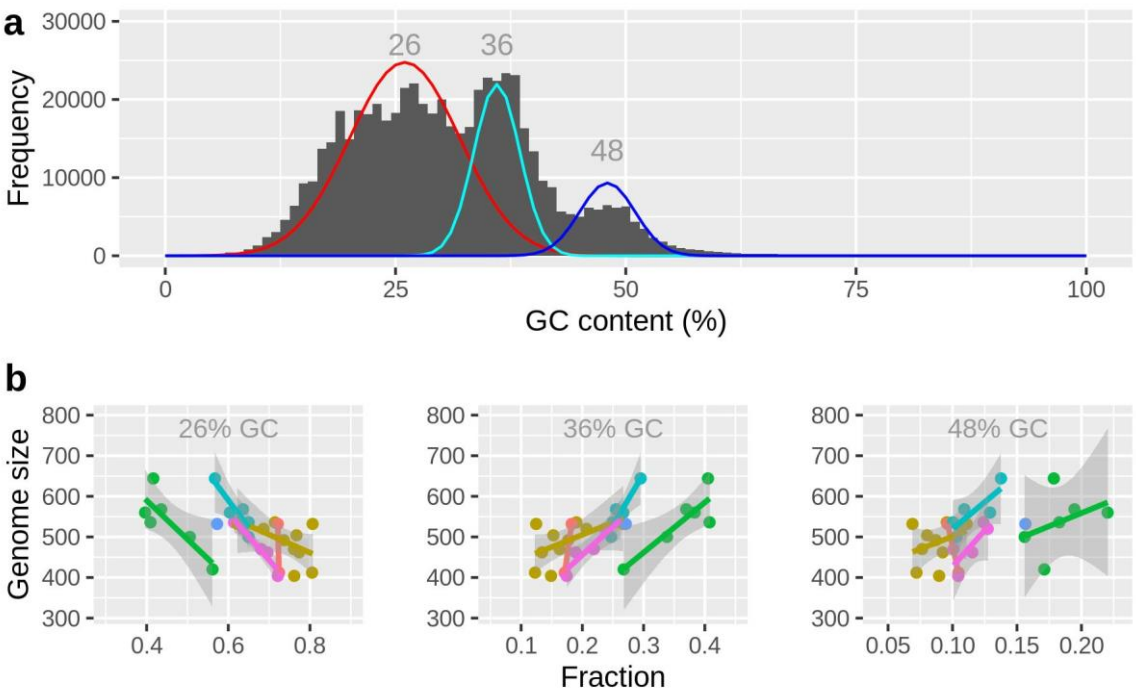
Tab. 1 Rotifer clones used in this study

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. 2019)
[2] same clone as reference genome
[3] hatched from an individual resting egg after three generations of selfing

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

155



156

157 **Fig. 2 Genome size variation is linked to sequences with elevated GC-content**. **a** GC-distribution of all short-
158 read libraries combined (29 libraries of 15 rotifer clones). The red, cyan, and blue lines designate a mixture
159 model fitted to these data, consisting of three normally distributed subpopulations (26 ±6, 36 ±2.5, and 48 ±3
160 %GC; means and sds). **b** Panels show the results of the same mixture model applied to each library individually,
161 thus estimating the proportion of the total reads per library in each GC-fraction. Genome size estimates are
162 based on flow cytometry and were taken from [7]. Colors in **b** correspond to the six library preps (A-F) listed in
163 Tab. S4: A=orange, B=gold, C=green, D=turquoise, E=blue, F=pink).

164

165 We also used two kmer-based tools, GenomeScope 2.0 [19] and findGSE [20], to obtain reference-
166 free estimates of genome size for each clone/library. Those estimates were generally lower than
167 their FCM based counterparts, approximately 0.8-fold in findGSE and 0.6-fold in GenomeScope

168 (**Supplementary file 3**). GenomeScope appeared to struggle at sequencing coverages below ca. 25-
169 fold, where it estimated extremely low genome sizes (compared to FCM) and unrealistically high
170 heterozygosities (6-10%). By contrast, findGSE performed constantly along the gradient of
171 sequencing coverages. Interestingly, there was a positive correlation between genome size (FCM
172 estimate) and the ratio of repeats, a fitted parameter of findGSE (**Fig. S4**; Spearman-rank correlation
173 test, ρ=0.609, *P*=0.0057).

174 Our assembly-based analyses rely on the alignment of cleaned reads of each of the 29 libraries to the
175 reference genome. Total alignment rates (TAR) of reads to the reference genome draft were
176 generally above 94%. TARs were highest in those rotifer clones that are most closely related to the
177 reference genome (**Fig. S5**). Concordant alignment rates, i.e., properly aligned reads with the correct
178 insert size, were >90% in all libraries (**Table S5**). However, only about 1/3 of the reads aligned
179 uniquely to one site in the genome. Only a small proportion of reads, usually well below 5%, showed
180 discordant alignment, either due to an incorrect insert size, or when only one of the mates aligning to
181 the reference genome. Discordant alignment rates were generally low (1.3-5.1%) and were not
182 correlated to genome size.

183 To identify CNV regions, we calculated the average per-base coverage for 5kbp and 50kbp windows.
184 To normalize coverage across libraries, we divided the per-base coverage of each window by 1/2 of
185 the (mean) exon coverage of the respective library. This yields a value of 2 for all diploid regions of
186 the genome, corresponding to two copies for those genomic regions (provided that our assembly is
187 unphased in these regions). This analysis of coverage variation revealed that large tracts in the
188 genome of *B. asplanchnoidis* display consistent patterns of coverage variation, which we quantified
189 as the standard deviation of coverage across libraries and clones (**Fig. 3**). For example, the first contig
190 (000F) exhibits large coverage variation with a standard deviation of 1.5-1.7 while the next three
191 contigs (001F to 003F) have much less coverage variation (< 0.5 s.d.) and a mean coverage value of
192 close to two. There are several other contigs showing consistently elevated coverage variation, like
193 004F, 031F, 032F,036F, 042F, and 044F. In addition, some of these variable contigs appear to be
194 more variable than others (e.g., 031F is more variable than 032F). These overall patterns were very
195 similar when a lower 5kbp- window resolution was applied (**Fig. S6**)

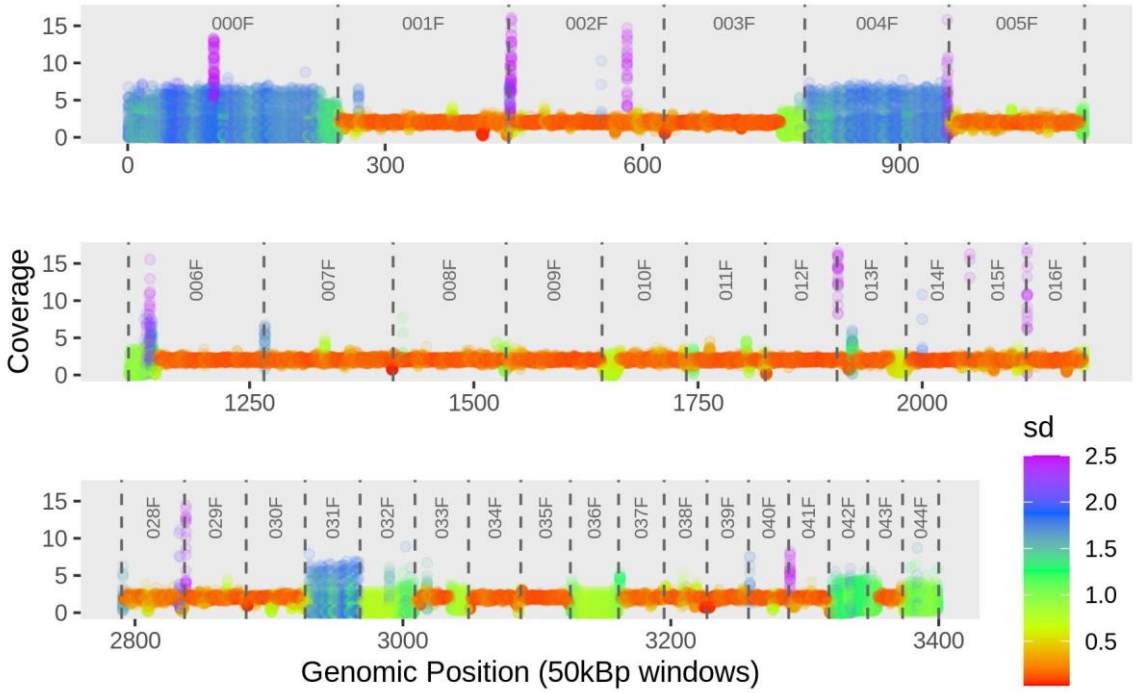


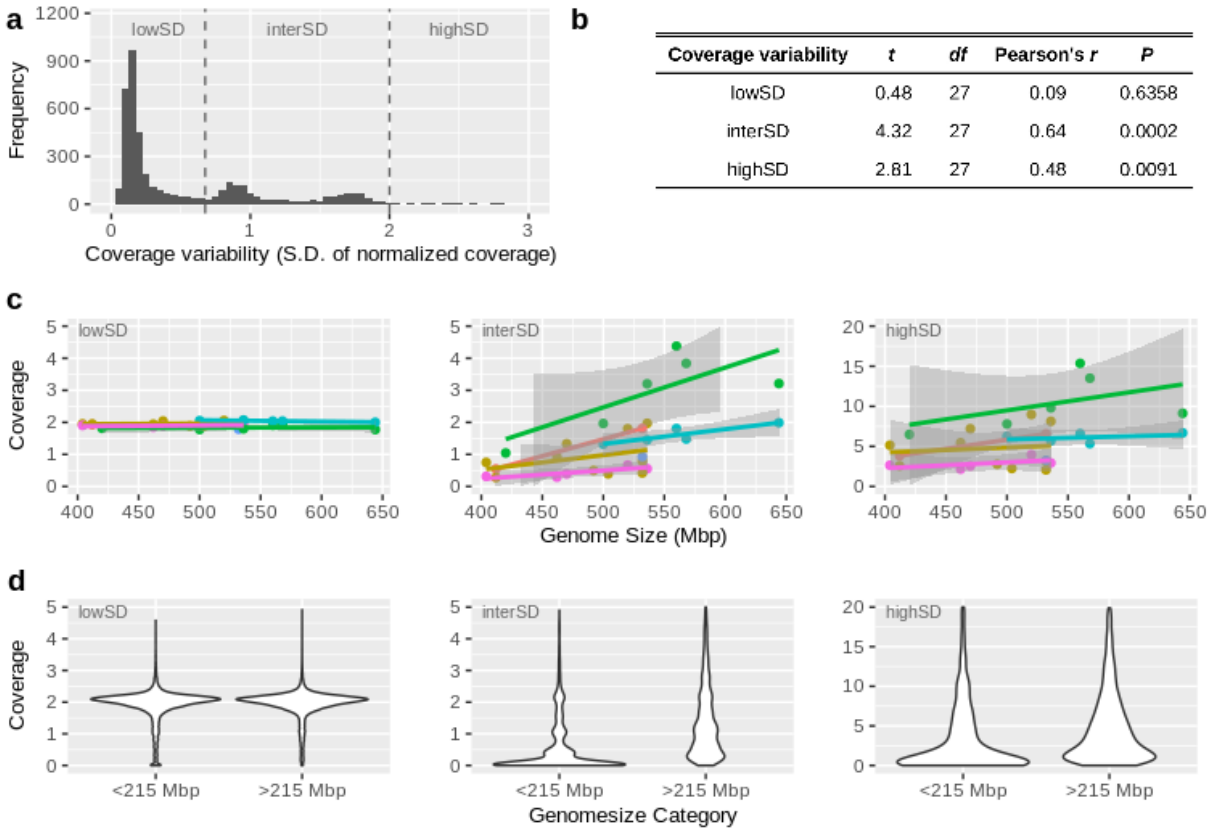
Fig. 3 Coverage variation along the *B. asplanchnoidis* genome. This graph shows 34 selected contigs spanning ~140 Mbp (i.e., about 60% of the assembly). Each circle represents a 50kbp window of one of the 29 sequencing libraries. Coverage (y-axis) was normalized by dividing the per-base coverage of 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 allow colored visualization.

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**. Interestingly, 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, in order to control for the effect of library

217 preparation, were calculated partial correlations between the variables 'mean coverage' and
218 'genome size' (**Fig. 4c**). Those correlations at 'intermediate' and 'high' variability were highly
219 significant. This result holds even if the libraries with elevated coverage and GC content at interSD
220 and highSD regions (green symbols in Fig. 2b and 4c) are excluded (**Fig. S7**)

221



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

222

223 **Fig. 4 Genomic regions of elevated coverage variability are responsible for genome size variation**. The data in
224 this figure is based on normalized coverage (calculated for 50kbp windows) of 29 short-read libraries from
225 15rotifer clones. **a:** Genomic regions differ in coverage variability (measured as standard deviation of coverage
226 among libraries) as there are regions of low, intermediate, and high variability. **b, c:** Mean coverage in regions
227 of elevated coverage variation (interSD, highSD) correlates positively with genome size (flow cytometry data on
228 genome size was taken from [7]. Dots represent the mean coverage per library. Colors indicate co-prepared
229 libraries (of which some were prepared from the same rotifer clone, but at different dates and with different
230 library preparation methods; see Table S4). Colors in **c** correspond to the six library preps (A-F) listed in Tab. S4:
231 A=orange, B=gold, C=green, D=turquoise, E=blue, F=pink). **d:** This panel shows the same data as in **c**, but rotifer
232 clones are categorized into small genome sizes (<215Mbp, which represent the basal genome size of *B.*
233 *asplanchnoidis* according to [7] vs. large genome sizes (>215Mbp, which contain additional genomic elements
234 that independently segregate during meiosis, according to [7]).

235

236 By merging adjacent 5kbp windows that show increased coverage variation and consistent coverage
237 pattern (significant correlation of coverages), we identified 509 CNV regions in the genome of *B.*
238 *asplanchnoidis* (**Supplementary file 4**). The total genome space classified hitherto as "copy number
239 variable" was 72.43Mbp (i.e., 31% of the genome assembly). Interestingly, large CNV regions
240 megabases in size made up a large fraction of this total, as can be seen by an "N50" of 0.455 Mbp for
241 the CNV-fraction of the genome. Two of the largest contigs, 000F and 004F, consisted almost entirely
242 of 3-4 large CNV regions, which were separated only by short "breakpoints" of lower coverage
243 variability (**Supplementary file 4**). Our 5kbp window-scanning algorithm also identified CNVs that
244 resided in contigs with otherwise "normal" coverage. In many cases, these CNV regions were near
245 the beginning or end of a contig (e.g., 003F, 006F, 010F). Interestingly, coverage values of clone IK1, a
246 cross between OHJ7i3n2 and OHJ22i3n14, were usually intermediate between those of the two
247 parental clones (Figs. S8, S11, S12). In addition, four clones that were derived by selfing from clone
248 OHJ7 displayed coverages that were largely consistent with their differences in genome size (Fig. S9).
249 For instance, the clone with the largest genome of the selfed line, OHJ7i3n5, had an additional
250 coverage peak at about 2.5 times the base coverage (set by IK1), which indicates that some CNV
251 regions have significantly higher coverage than any other clone of this selfed line.

252 To additionally classify CNV regions according to their length and contiguity, we considered contigs as
253 "B-contigs" if they contained a large fraction of CNV windows (in analogy to B-chromosomes). Setting
254 this threshold at 90% of contig length, 38 contigs are classified as "B-contigs", comprising 77% of all
255 CNV-windows, i.e., 55.8 Mbp of the assembly (**Fig. S10, S11, Table S6**). Thus, approximately three
256 quarters of the observed CNVs affect more or less an entire contig, while the remaining quarter of
257 CNV regions were found on contigs with otherwise low coverage variability.

258 To independently confirm CNVs, we chose four genomic loci for PCR-amplification, two in CNV
259 regions and two in non-CNV regions (**Table S7**). All four primer pairs yielded amplicons with the
260 correct size, with no signs of non-specific amplification (**Fig. 5a**). The two primer pairs targeted to
261 non-CNV regions (TA_001F and TA_003F) yielded an amplicon in all rotifer clones. In contrast, the
262 two primer pairs targeted to CNV loci (TA_000F and TA_032F) only amplified in some clones. In
263 particular, clones with the smallest genome sizes (OHJ82, OHJ22 and its descendent OHJ22i3n14)
264 apparently lack both CNV loci, and in others (OHJ 98, 104, 105) the TA_000F-locus was present, but
265 the TA_032F-locus was absent. Overall, these patterns were highly consistent with coverage of the
266 amplified regions in sequencing libraries (**Fig. 5c**). Copy numbers for TA_032, as estimated by ddPCR,
267 ranged from zero to six across the studied rotifer clones, including 3 copies in IK1, a cross between
268 OHJ7i3n2 (six copies) and OHJ22i3n14 (zero copies) (**Fig. 5b, Tab. S8**).

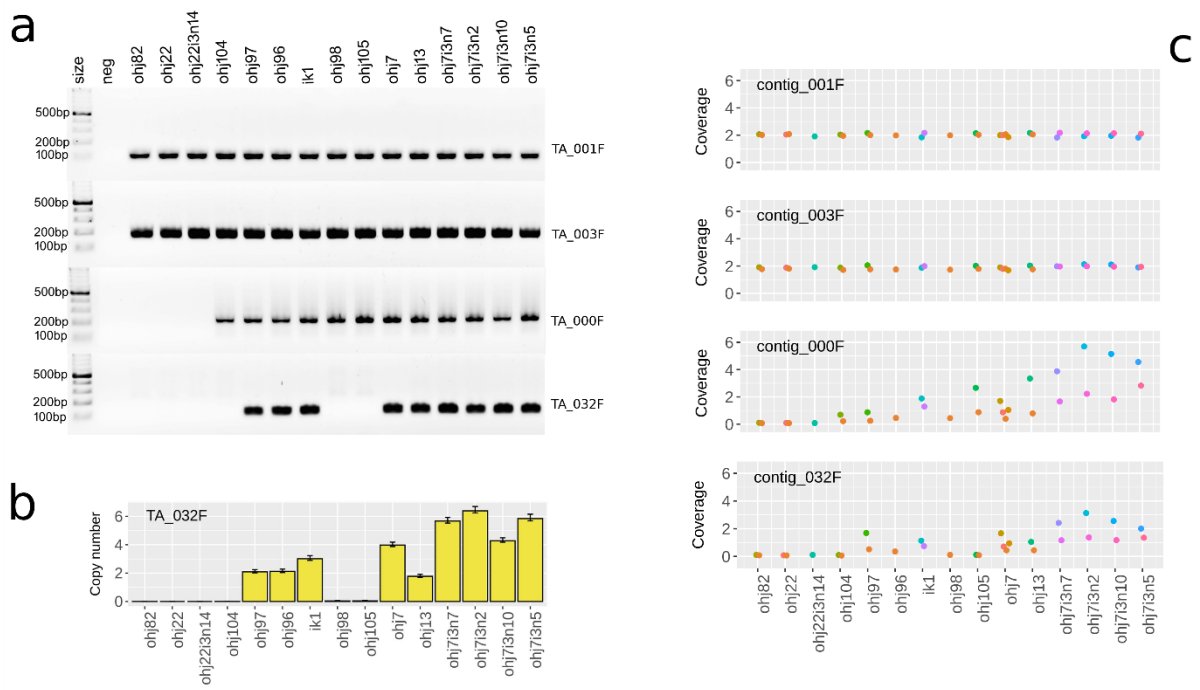


Fig. 5 PCR-based confirmation of CNV-loci. All 15 studied clones from the OHJ-population are displayed in order of ascending genome size (from left to right). **a** PCR-products of primers targeting the four candidate loci (non-CNV loci: TA_001F and TA_003F; CNV-loci: TA_000F and TA_032F). **b** Copy numbers of the TA_032F-locus as estimated by droplet digital PCR. 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.

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%) were 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 (**Supplementary file 6**). Overall, repetitive elements, and especially satDNAs were over-represented in CNV-regions (**Fig. 6c, Supplementary file 7**).

Of the satDNAs we identified in B. *asplanchnoidis*, three (rotiSat2, 8 and 9) are not present in any other sequenced *Brachionus* genome; three others (rotiSat1, 5, 10, 11) are shared with the *Brachionus plicatilis* ss. 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* ss. 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, Supplementary file 7**). Phylogenetic orthology inference based on 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 93063 genes (90.9% of all annotated proteins) to 17965 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.

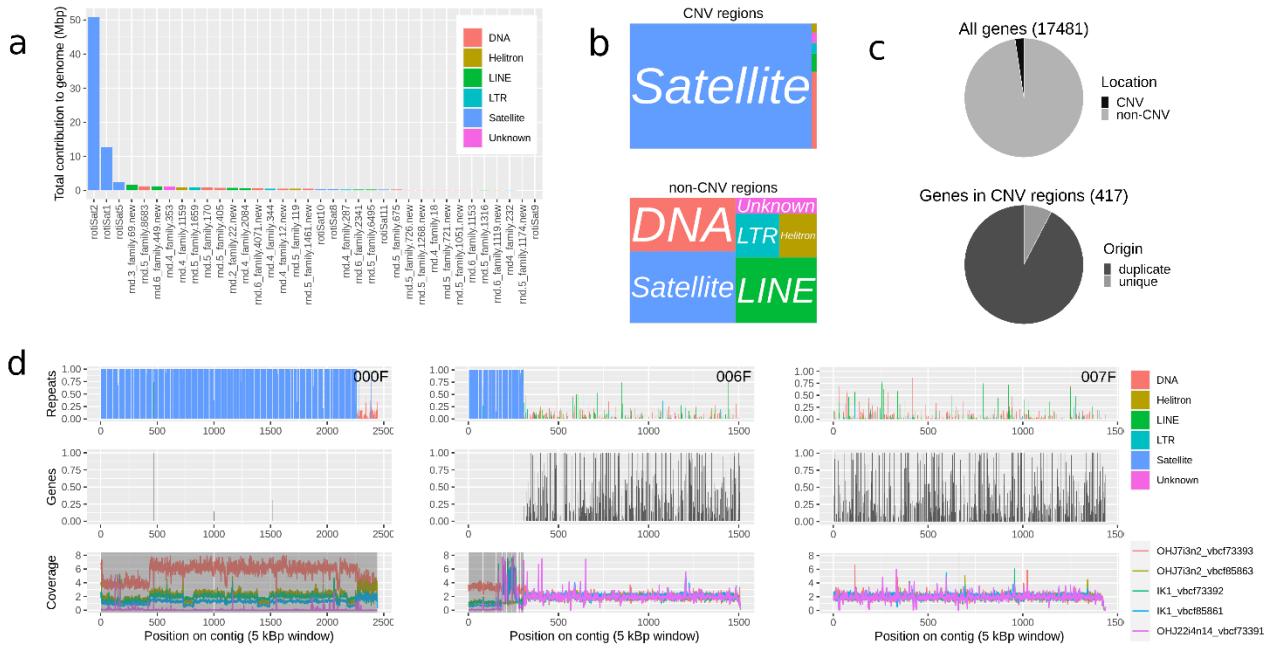


Fig. 6 Genome structure of *B. asplanchnoidis*. a The top-36 repeat elements ranked according to their contribution to the genome. **b** Differences in repeat element composition between CNV and non CNV regions. **c** Differences in the distribution of gene content between CNV and non CNV regions (top) 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

313 at one end (006F), and a non-CNV contig (007F). Top panels display the proportion of each 5kb window
314 occupied by repeats, intermediate panels display the proportion of each 5kb window occupied by annotated
315 genes, bottom panels display (exon-)normalized coverages of three rotifer clones (parents: OHJ7i3n2 and
316 OHJ22i3n14, crossed offspring: IK1). Shaded regions indicate CNVs.

317

318 Throughout this study we observed several conspicuous patterns related to GC-content. Regions of
319 elevated coverage variability (i.e., interSD and highSD regions), CNV regions, and B-contigs were
320 characterized by an elevated GC-content showing a main peak at ~37% GC and two additional peaks
321 at around 50% GC (Fig. S13, S14). By contrast, regions of low coverage variability had their main peak
322 at ~25% GC (Figure S14). Interestingly, these three peaks were also present in the GC-distributions of
323 unaligned sequencing reads from rotifer clones varying in genome size (Fig. 2b). The 37% GC peak,
324 which was the most prominent peak in genomic regions of elevated SD, in CNVs, and in B-contigs
325 (Fig. S14) could be attributed mainly to the satellites rotiSat1 and rotiSat2, while the higher peak at
326 ~55% GC could be attributed to rotiSat5 (Fig. S15-19). Overall, these three satellites are the most
327 abundant repeat elements in the *B. asplanchnoidis* genome, and their consensus sequences show
328 the same characteristically elevated GC contents compared to most other repeat elements
329 (Supplementary File 6). We also observed one minor but distinct peak at ~48% GC in highSD regions,
330 CNV regions, and B-contigs, which apparently consisted of sequences that were not classified as
331 repeats by repeatModeler2 (Figures S15-19).

332 ## Discussion

333 In this study, we provide a high-quality reference genome draft of the rotifer *Brachionus*
334 *asplanchnoidis* to shed light on one of the most extreme examples of intraspecific genome size
335 variation in the animal kingdom [7, 17]. Our genome assembly had a length of 230.1 Mbp. The 2C
336 DNA content of the same rotifer clone is 568 Mbp, according to flow-cytometry based estimates of
337 an earlier study [7]. This earlier study provided evidence that the genome of this particular clone
338 consists of a core haploid genome of 207Mbp and four copies of a segregating 34Mb element.
339 Assuming that our 230.1 Mbp assembly is completely unphased, our reference genome thus
340 amounts to 95% of the flow cytometry-based estimate of the haploid genome (241Mbp).

341 We aligned short-reads of 15 rotifer clones from the same geographic population, together
342 encompassing a 2C-genome size range of 404–644Mbp, to the reference genome. This analysis
343 uncovered multiple long tracts along the reference genome with increased coverage variation across
344 clones. Additionally, we found that the average coverage at CNV-regions strongly correlates with
345 genome size. CVN-regions also carried a distinct signature in terms of an increased GC-content (36%
346 and 48%, respectively, versus 26% for the rest of the genome), which was both apparent in the short-
347 read data and in the genome assembly. Strikingly, many CNVs had near zero coverage in three of the
348 studied clones (OHJ82, OHJ22, OHJ22i3n14), which was independently confirmed by our PCR-based
349 assays of two selected CNV regions. These results are highly consistent with previous evidence from
350 flow cytometry experiments, showing that these three clones are characterized by a 'basal' genome
351 size (i.e., are close do the smallest observed genome size in this species) and that they entirely lack
352 the independently segregating genomic elements (ISEs) that are present in many other members of
353 the OHJ population. Our results indicate the presence of multiple different ISEs in the OHJ
354 population. For example, both our PCR- and alignment data suggested that the two contigs 000F and
355 032F belong to two different ISEs, because only the former was detectable in the clones OHJ98,
356 OHJ104, and OHJ105 (**Fig. 5**). This observation is consistent with an earlier study, which suggested
357 such diversity based on the size of ISEs measured by flow-cytometry.

358 CNV regions account for megabase-long tracts in the genome of *B. asplanchnoidis*. Several contigs
359 displayed highly similar coverage patterns across almost their entire length, if one ignores the few
360 and very short breakpoints. Such contigs might be fragments of even larger elements, perhaps B-
361 chromosomes. Contig 032F might be a good candidate for this, ranging from zero to six copies across
362 the OHJ-population, as indicated by ddPCR. We also detected large CNVs, hundreds of kilobases in
363 length, that were located on contigs with otherwise normal (diploid) coverage and low coverage
364 variation (e.g. contig 006F in **Fig. 6d**). Such a genomic pattern is consistent with a stable diploid
365 chromosome that contains a large homozygous insertion in some clones, hemizygous insertions in

366 others, and a homozygous deletion in the remaining clones. Additionally, there might be length
367 polymorphisms in such genomic regions, which are dominated by tandemly repeated satellite DNA.
368 In the future, assembling multiple genomes from rotifer clones with different genome size, ideally
369 using long-read technologies in order to approach chromosome-level assemblies [21], might allow a
370 more precise delineation of individual CNVs into these two categories. Overall, our data is consistent
371 with a mixture of B-chromosomes and large insertions into normal chromosomes, and possibly a
372 dynamic exchange between both genomic fractions (since they are made up mostly by the same set
373 of tandemly repeated satellite DNA; see below).

374 There are a few technical caveats and limitations to be considered in our analysis. First, although we
375 found strong correlations between coverage variation at CNV regions and genome size variation, it
376 was not possible to quantitatively "predict" the genome size of individual clones based on coverage
377 along the reference assembly. Library preparation method seemed to introduce additional variation,
378 specifically at the CNV regions, that prevented us from determining exact copy numbers of these
379 genomic regions. This is a well-known limitation of short-read libraries from genomes that contain
380 large amounts of satellite DNA [22], in particular when combined with heterogeneous GC contents
381 [23]. We could alleviate this limitation and determine exact copy numbers by using digital droplet
382 PCR, finding that the targeted genomic region (contig 032F) is present in zero to six copies across the
383 OHJ-population. This approach is a promising strategy for future studies to accurately assess copy
384 number variation across many loci and in a large number of genomes from the same population. In
385 addition, the differences in copy numbers at CNV regions could be estimated by comparing clones
386 within a set of library preparations. For instance, if coverage at CNV regions is scaled by a reference
387 clone (like clone 'IK1' in Fig. S8 and S9), coverage at CNV regions of focal clones tends to fall into
388 discrete clusters, indicating n-fold differences in coverage relative to the reference. A second
389 limitation is that our reference genome might still not contain all the ISEs that contribute to genome
390 size variation in the natural OHJ population, simply because it represents just a sample from this
391 population. To obtain a more complete picture, more genomes of the OHJ-population would be
392 needed to be sequenced, preferably using long-read sequencing technologies that allow better
393 identification of structural variants [24]. Third, one might argue that our CNV-detection pipeline
394 could miss many of the smaller insertions and deletions, in particular those smaller than 5 kilobases.
395 However, it is quite unlikely that such small-scale structural variation has much influence on genome
396 size variation in *B. asplanchnoidis*, since the percentage of discordantly aligned reads was overall
397 rather low (1.3-5.1%) and it did not scale with genome size. With our reference genome being at the
398 upper end of the genome size distribution of the OHJ-population, we would expect to find a higher
399 percentage of discordant reads in smaller genomes (mainly due to deletions), which was not the
400 case.

401 CNV regions differed strongly from the remaining genome regarding repeat composition and gene
402 density. Strikingly, most CNV regions were composed of only three satellite repeat elements, some
403 unique to *B. asplanchnoidis* and others shared only with its closest congener, *B. plicatilis*, indicating
404 recent evolutionary origin. The two most prominent satellite repeats, rotiSat1 and rotiSat2, consisted
405 of 154 and 143bp monomers that were tandemly repeated for up to a few megabases. The low
406 sequence diversity of CNV regions nicely explains the characteristic trimodal GC-content signature
407 mentioned earlier, especially since the most abundant satellite elements display the same elevated
408 GC content (**Supplementary file 6**). Non-CNV parts of the genome contained a much higher diversity
409 of other repeat elements, which included DNA transposons, LINEs, LTR elements, and Helitrons. Gene
410 density was significantly lower in CNV regions compared to the rest of the genome, and genic regions
411 in CNVs were typically confined to short stretches scattered across the contig. Interestingly,
412 Orthofinder suggested that genes in CNV regions were three times more likely to derive from a
413 duplication event than genes found in other places of the genome. This indicates that duplications
414 had a role in the origin of these CNV-regions, which incidentally has some resemblance to the
415 proposed early evolution of B-chromosomes [25, 26].

416 In this study, we have identified, for the first time, genomic elements that can cause substantial
417 within-population genome size variation, ultimately allowing investigation of genome size evolution
418 at microevolutionary time scales. In the OHJ-population of *B. asplanchnoidis*, these genomic
419 elements consist of up to megabases-long arrays of satellite DNA, with only few interspersed genes
420 or other sequences. Though satellite arrays can form essential chromosome structures such as
421 centromeres and telomeres [27], the large intrapopulation variation of these elements in *B.*
422 *asplanchnoidis*, and their virtual absence in some individuals suggests that, overall, these DNA
423 additions do not provide an immediate fitness benefit to their carriers. Nevertheless, variable
424 amounts of 'bulk DNA' might influence the phenotype through subtler mechanisms independently of
425 the DNA information content, for example through potential causal relationships between genome
426 size and nucleus size, or cell size [28, 29]. Increased levels of structural variation in a genome, as
427 implied by our findings, may also constrain adaptive evolution or genome stability over
428 microevolutionary time scales. In this regard, the genome of *B. asplanchnoidis* should be a valuable
429 addition to existing models of genome evolution, enabling whole-genome analysis combined with
430 experimental evolution approaches [30, 31].

431

432 # Methods

433 **Origin of rotifer clones and DNA extraction**

434 Resting eggs of rotifers were collected in the field from Obere Halbjochlacke (OHJ), a small alkaline
435 playa lake in Eastern Austria (N 47°47'11", E 16°50'31"). Animals were kept in clonal cultures. A
436 rotifer clone consists of the asexual descendants of an individual female that hatched from a single
437 resting egg. Since resting eggs are produced sexually in monogonont rotifers, each clone has a unique
438 genotype. Our clones from the OHJ-population have been characterized previously with regard of
439 their genome size and other biological traits [7].

440 Rotifers were cultured in F/2 medium [32] at 16 ppt salinity and with *Tetraselmis suecica* algae as
441 food source (500–1000 cells μl^{-1}). Continuous illumination was provided with daylight LED lamps
442 (SunStrip, Econlux) at 30–40 µmol quanta m^{-2} s^{-1} for rotifers, and 200 µmol quanta m^{-2} s^{-1} for algae.
443 Stock cultures were kept either at 18 °C, re-inoculated once per week by transferring 20 asexual
444 females to 20ml fresh culture medium, or they were kept for long-term storage at 9°C, replacing
445 approximately 80% of the medium with fresh food suspension every 4 weeks.

446 To produce biomass for DNA extraction, rotifers were cultured at 23°C in aerated borosilicate glass
447 containers of variable size (250mL to 20L). Prior to DNA extraction, rotifers were starved overnight in
448 sterile-filtered F/2 medium, with 2-3 additional washes with sterile medium on the next day. In most
449 preparations we also added the antibiotics Streptomycin (Sigma-Aldrich: S6501) and Ampicillin
450 (Sigma-Aldrich: A9518) to the washing medium, both with an end concentration of 50mg/ml. DNA
451 was extracted using the Qiagen kits Dneasy (for short-read sequencing; from approximately 5000-
452 7000 rotifers) and GenomicTips 100 (for long read Pacbio sequencing; from >20000 rotifers), and
453 RNA was extracted from freshly prepared biomass with Rneasy.

454

455 **Sequencing of the reference clone**

456 We selected one rotifer clone (called: OHJ7i3n10) as DNA-donor for the reference genome. This
457 clone ultimately derives from an ancestor of the natural OHJ population. However, its immediate
458 ancestors were passed through three generations of selfing (i.e., mating one male and female of the
459 same clone). More details on the genealogy of this lineage and its biological characteristics can be
460 found in [7]. According to this study, OHJ7i3n10 has a 2C-genome size of 568 Mbp, and thus contains
461 approximately 40% excess genomic sequences, compared to the smallest genome size of the OHJ-
462 population (~410 Mbp).

463 Our *B. asplanchnoidis* reference genome is based on long-read sequencing technology (PacBio
464 SMRT® on the Sequel-platform). In total, we obtained 16.3 Gbp from two SMRTcells, which confers
465 to 57-fold coverage assuming a haploid genome size of 284 Mbp. Additionally, we obtained 35.5 Gbp
466 of Iso-Seq transcriptome data and 12 Gbp of short-read Illumina data of OHJ7i3n10. All sequencing
467 and library preps related to the reference genome were performed by the Next Generation
468 Sequencing Facility at Vienna BioCenter Core Facilities (VBCF), member of the Vienna BioCenter
469 (VBC), Austria.

470

471 **Sequencing of individuals of the OHJ-population**

472 To characterize genomic variation across the OHJ-population, we generated short-read sequencing
473 data (Illumina platforms HiSeq and NextSeq) from 15 different rotifer clones, both from the natural
474 OHJ population and from various clones of a selfed lineage (**Table S4**). Clones were selected such that
475 they covered the full range of genome size of the OHJ-population. Short-read libraries were
476 constructed either using the KAPA HyperPrep kit (Roche), or the NGS DNA Library Prep Kit
477 (Westburg). The KAPA library preparations were done at the Marine Biological Laboratory [for more
478 details, see12] while the Westburg preparations were done at VBCF. The two methods mainly differ
479 in the fragmentation method (ultrasonic fragmentation in KAPA *vs*. enzymatic fragmentation in
480 Westburg). Both methods are claimed to deliver sequence-independent fragmentation and to yield
481 consistent coverage across a wide range of GC-contents. While the peak fragment size was ~400bp in
482 both methods, we observed that the libraries constructed with the Westburg kit sometimes had a
483 pronounced right tail with some fragments up to ~2000bp. We accounted for these larger fragments
484 by adjusting the relevant parameters during short-read alignment (see below). In many of our clones,
485 we used both library construction methods, yielding in a total of 29 libraries (**Table S4**).

486

487 **Reference genome assembly and annotation**

488 Pacbio sequences of the reference genome were assembled using the HGAP4 pipeline [33] at VBCF,
489 and contamination was initially checked using CLARK [34] against all available bacterial genomes
490 from NCBI (**Supplementary file 1**). We polished this initial VBCF assembly with short-read Illumina
491 data of the identical *B. asplanchnoidis* clone using Pilon [35] in three rounds. To investigate the
492 assembly quality, we backmapped the Illumina data to the genome assembly using bwa mem [36]
493 and calculated summary statistics using QUAST [37] and Qualimap [38]. To provide an additional
494 check for contamination, we used Blobtools [39] based on the backmapping alignments and a blastn
495 search against the nt database (NCBI).To assess the assembly's completeness, we performed a

496 BUSCO v4.0.6 [40] using the metazoan gene set (n=978) in the *genome* mode applying the *–long*
497 option.

498 Genes were structurally annotated on the repeat masked genome assembly using the
499 MAKER2 annotation pipeline [v2.31.10, 41] using evidence from Pacbio Iso-Seq transcripts of the
500 same *B. asplanchnoidis* clone; protein homology evidence from the UniProt database (download
501 January 2020, The UniProt Consortium 2017) in combination with proteoms of different clones and
502 closely related *Brachionus* species (**Tab. S1**;, and *ab initio* gene predictions from SNAP [42],
503 GeneMark-ES [v4.48_3.60_lic, 43] and Augustus [v3.3.3, 44]. The SNAP model was initially trained on
504 the genome assembly with additional support of BUSCO complete hits (see above). GeneMark was
505 computed in the ES suite on the soft masked genome assembly. The ab initio training of the Augustus
506 model was computed on the genome assembly supported by the unassembled Iso-Seq transcripts. In
507 order to compensate for underrepresented rotifer protein representation in the UniProt database,
508 we combined this dataset with proteomes of another *B. asplanchnoidis* clone and four closely related
509 *Brachionus* species (**Tab. S1**). Completeness of these proteomes was checked with BUSCO in the
510 *protein* mode and missing orthologs were compared to determine if completeness could be
511 increased through the combination of proteomes. The combined proteome finally contained 97 % of
512 the BUSCO genes of the metazoan dataset.

513 MAKER was run over three rounds. For the first round of MAKER, we only used the Iso-Seq
514 data as EST evidence and the combined UniProt and proteome sequences as protein homology
515 evidence applying default parameters and est2genome=1, protein2genome=1 to infer gene
516 predictions directly from the transcripts and protein sequences. We used the gene prediction of
517 round 1 to retrain the Augustus and SNAP model, since the GeneMark model was exclusively
518 computed on the genome assembly and did not require retraining. Maker was run in the second
519 round using the retrained models and again est2genome=1, protein2genome=1 in order to increase
520 prediction fidelity. After a second round of retraining the Augustus and SNAP model, MAKER was run
521 through round 3 with switched off est2genome and protein2genome inference. After structural
522 annotation via MAKER, we functionally annotated the genes using functional classification of genes
523 from InterProScan [45] in combination with putative gene names derived from Swissprot [46].

524 To examine orthologous genes among rotifer species and identify gene duplication events,
525 we used Orthofinder [47]. This analysis was based on proteome information of the *Brachionus*
526 *plicatilis* species complex: *B. rotundiformis*, *B.* sp. 'Tiscar' and *B. plicatilis* [respectively "Italy2",
527 "TiscarSM28", "Tokyo1" in, 12], and *B. asplanchniodis* (annotation of this study). Proteome
528 information of *Adineta vaga* [48] was included as outgroup. We extracted the longest transcript
529 variant per gene to avoid duplicates in the input proteomes and followed the manual instructions of

530 Orthofinder. To analyze the genomic distribution of genes which were identified to derive from gene
531 duplication events, we extracted all genes of the *B. asplanchniodis* node (Orthofinder:
532 SpeciesTree_Gene_Duplications), removed duplicates to create a non-redundant list of genes.
533 Comparing this list of genes to genes that are located inside or outside of genomic regions with high
534 levels of CNVs, allowed the estimation of non-random gene distribution patterns via Monte-Carlo
535 permutation tests (1000 permutations).

536

537 **Annotation of repetitive elements**

538 The repeat library was produced using RepeatModeler2 [default settings, 49] and the polished
539 PacBio assembly. This produced 484 consensus sequences, 289 of which are "Unknown". These
540 consensus sequences were then used to mask the genome assembly using RepeatMasker with
541 default settings. From this, we identified top contributing repeat elements and used their consensus
542 sequences were used as blast queries against the genome assembly, and the top hits for each
543 consensus were used to produce alignments for manual curation [50] and classification [51]. Satellite
544 monomers were identified using Tandem Repeat Finder [52]. Consensus sequences of the satDNAs
545 are dimers of these identified monomers.

546 Contributions per repeat were estimated by summing up the total length covered by each
547 copy in the RepeatMasker output file. Top contributions were calculated over the whole assembly,
548 and over regions of the genome with distinct coverage variability (lowSD, interSD, and highSD in Fig.
549 4). For each region, the top 20 repeats were included, resulting in a total of 38 repeats to curate.
550 Curation was done by manually inspecting each alignment, identifying the ends of the aligning
551 regions, and producing a new consensus sequence. Additionally, classification was performed by
552 searching for TSDs, TIRs, LTRs, satellite structure, and conserved domains. RepBase searches were
553 rarely constructive since rotifer, and especially monogonont, TEs are not well-represented in any
554 databases. The final curated library of topREs contains 37 consensus sequences from 36 elements
555 (one sequence was removed because alignments were to only scattered AT-rich regions, one was an
556 rRNA gene, and one LTR sequence was split into the LTR and internal portions). Redundant
557 sequences between the topRE library and the RepeatModeler library were identified using
558 RepeatMasker. Sequences which were at least 95% identical and covered at least 80% of the
559 uncurated repeat were removed from the uncurated repeats prior to merging the libraries. Due to
560 the short length and difficulties in automatically creating consensus sequences for satDNA elements,
561 all uncurated elements that were identified as similar by RepeatMasker, regardless of similarity or
562 length of hit, were aligned to the curated consensus sequence and visually inspected for alignment.

563 Uncurated elements which aligned across most of the curated satDNA consensus sequences were
564 removed. The combined library (Supplementary file 8) of curated and uncurated elements contained
565 472 elements (262 unknowns), and was used to mask the *B. asplanchnoidis* genome and to
566 repeatmask related, high quality *Brachionus* genome assemblies to identify shared elements [53-55].
567 For each of the top contributing elements and each genomic region, an enrichment index was
568 calculated as the proportion of each repeat contribution in that region vs the whole genome (i.e.
569 repeat contribution in region/total contribution of repeat) divided by the proportion of the genome
570 represented by each region (i.e. region size/ assembly size). If this index was over 1, it meant that the
571 repeat in question was enriched in the region in question.

572

573 **Preprocessing of short-read data**

574 Trimming and adaptor removal was done using bbduk (v38.34, https://jgi.doe.gov/data-and-
575 tools/bbtools/) with the settings: k=23 ktrim=n mink=11 hdist=1 tpe tbo qtrim=rl trimq=20 maq=10
576 minlen=40. Trimmed reads were initially aligned to the reference genome using bowtie2 [56].
577 Preliminary tests with various parameter settings (local *vs*. end-to end alignment, different settings
578 for the sensitivity parameter, map as single reads *vs*. map as paired reads) did not indicate strong
579 differences. Thus, we used the default values for most parameters, except for fragment size
580 (parameter: X), which was set to the maximum observed fragment size of each library instead of the
581 default value of 500 bp (see **Tab. S4**).

582 We included three steps to remove contaminants, since DNA extracted from the whole
583 bodies of microscopic organisms might contain DNA from other organisms, such as bacteria. First, we
584 extracted all unmapped reads from an initial alignment to the reference genome and screened only
585 those for potential contaminant reads. Thus, we considered the mapped fraction as 'rotifer DNA',
586 since they mapped to the (contaminant-free) assembly. Second, the unmapped reads of all libraries
587 were combined and assembled using meta-genomic approaches. For this assembly, we used
588 metaVelvet [v1.2.02, 57] with a kmer-legth of 101bp, an insert size of 500 (+/- 200 standard
589 deviation), and a minimum reported contig length of 300bp. The resulting assembly was then
590 analyzed with metaQuast [58] using the option 'automatic pulling of reference sequences', which
591 restricts the search to bacteria and archaea. Subsequently, the complete genomes of putative
592 contaminants were downloaded, and the unmapped reads were mapped against those genomes. In
593 the third step, the remaining fraction (i.e., reads not mapping to the microbial metagenome) were
594 subjected to a kmer-based identification using the kraken2 pipeline [59], with the databases bacteria,
595 archaea, viral, UniVec-Core, and protozoa. We also performed checks on the false-discovery rate of

596 kraken2, by running the same pipeline on reads that initially did map to the reference genome.
597 Finally, all unmapped reads that could *not* be taxonomically assigned to contaminants, with either of
598 the two approaches above, were considered to be of 'rotifer origin' and were merged with the
599 mapped fraction. Those 'rotifer reads' were then further cleaned by removing mitochondrial DNA,
600 which was done by mapping them to the published mitochondrial genome of *B. plicatilis* [60], and by
601 removal of duplicates using FastUniq [61]. These 'final reads' were again mapped to the reference
602 genome, or they were analyzed using reference-free approaches that do not require alignment.

603

604 **Analysis of unaligned reads**

605 'Final reads' were subjected to kmer-based analyses using a kmer-size of 21bp with jellyfish [v2.1.4,
606 62]. Then GenomeScope 2.0 [19] and findGSE [20] were used to obtain kmer-based estimates of
607 coverage, heterozygosity and genome size. Per-base coverages C were computed from the kmer-
608 coverages

609 C_k with the formula:

610 $$C = \frac{C_K \cdot R}{(R - K + 1)}$$

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

615 To analyze GC-distributions among short-read libraries, the GC contents of individual reads
616 were extracted using the function fx2tab of SeqKit [63]. The resulting csv-files were further analyzed
617 with the R-package mixtools [64] using the function normalmixEM.

618

619 **Analysis of copy number variation**

620 Analysis of copy number variation was done separately for all 29 short-read libraries from 15
621 different clones of the rotifer *B. asplanchnoidis*. Average per-base depth-of-coverage (DOC) values
622 along 50-kbp and 5kbp windows, respectively, were extracted from the BAM alignment files ('final
623 reads' to reference genome) using the samtools function 'bedcov' [65]. In total, there are 4835
624 windows at 50kbp resolution, and 46255 windows at 5kbp resolution in the current genome
625 assembly. To allow comparisons among clones and libraries, DOC was normalized by dividing the per-

626 bp coverage of each window by 1/2 of the (mean) exon coverage of the respective library. In
627 unphased sections of the genome assembly, we expected DOC values of around 2, provided that
628 both alleles of a (diploid) genome map to the correct location. Coverage variation was quantified as
629 the standard deviation of DOC per window (50kbp or 5kbp) across all 29 libraries.

630 To identify individual CNVs and to locate possible breakpoints within contigs, we used a
631 custom-written R-algorithm involving the following criteria for merging adjacent windows (5kbp)
632 based on the similarity of coverage patterns. First, the coverage variation (measured as the standard
633 deviation of per-base normalized coverage across all libraries) had to be above a defined threshold
634 (i.e., 0.7, which was an *a-posteriori* determined threshold). Second, the read-depths of each library of
635 both windows had to be significantly correlated with each other at the $P<0.05$ level. This was done by
636 calculating the partial correlation coefficients. If both conditions applied, those two windows were
637 considered to belong to the same CNV. The very first and the last window of each contig, which often
638 showed deviant coverage patterns, were merged with their neighboring window. Third, adjacent
639 CNVs identified according to the above criteria were merged if they were separated by only one
640 window (CNV stopping breakpoint) AND if the coverages in the two windows surrounding the
641 breakpoint are significantly correlated with each other. Finally, we only considered CNVs with lengths
642 of at least three adjacent windows. Thus, we obtained a table of all CNVs along the genome of *B.*
643 *asplanchnoidis*, together with their size, and their location on individual contigs (i.e. in the middle of
644 the contig, at the edges, or spanning the entire contig). Data analysis related to coverage variation
645 and CNV detection was done using custom-written algorithms in the R environment [66] with the
646 base package (v3.6.3) and the add-on packages *stringr* [67] and *reshape2* [68]. For graphical
647 visualization, we used ggplot2 [69] and the add-on packages cowplot [70] and treemapify [71].

648

649 **PCR-confirmation of CNV regions**

650 To independently confirm presence or absence of CNV regions in different rotifer clones, and
651 to estimate the copy numbers of these genomic regions, we used PCR-based methods. To identify
652 unique PCR-primer binding sites in in regions of high and of low coverage, we used Thermoalign [72]
653 searches in multiple regions of 5000 bp length, spread across the genome and on different contigs.
654 The exact search parameters for Thermoalign are given in **Supplementary file 5**. Candidate primers
655 were tested and optimized using PCR on template DNA from different OHJ clones, including the
656 reference clone OHJ7i3n10. PCR reactions consisted of 25µl HotStart Taq master mix (Qiagen), 0.1µM
657 Primer, 3mM $MgCl_2$ and 20 ng template DNA. PCR cycling conditions were: 95°C for 15 min, 30 cycles
658 of 94°C for 20 sec, 56°C for 20 sec, and 68°C for 10sec; followed by 68°C for 5 min and hold at 4°C.
659 Agarose gels were used to test for presence/absence of the associated loci across different members
660 of the OHJ population. In total we screened four loci located on different contigs of the reference
661 genome assembly. Two of them were located in copy number-invariable, diploid regions of the
662 genome (according to the coverage estimates from the short-read alignments) and two were located
663 in CNV-regions (**Table S7**). In addition to the qualitative PCR test, we used droplet-digital PCR (ddPCR)
664 to estimate the copy number of the CNV loci for each rotifer clone. For each 22µl reaction we used
665 EvaGreen Supermix (Biorad), 0.15µM Primer, 4 units EcoRI and template DNA equivalent of 1500
666 genome copies. PCR cycling conditions were: 95°C for 5 min, 40 cycles of 95°C for 30 sec, and 61°C
667 for 1 min with a ramp rate of 2°C/sec; 4°C for 5 min, followed by 90°C for 5 min and hold at 4°C. For
668 droplet generation and fluorescence readout, we used a QX200 Droplet Generator and Droplet
669 Reader (Biorad), respectively. Copy numbers (CN) of CNV loci were estimated for each rotifer clone
670 using the ratios of amplicon molecule concentrations, which were obtained with the QuantaSoft
671 Software (Biorad):

672 $$CN = \frac{T}{(R_1 + R_2)/2} N_R$$

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

678

Acknowlegements

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

689 # References

690 1. Elliott TA, Gregory TR. What's in a genome? The C-value enigma and the evolution of
691 eukaryotic genome content. Philosophical Transactions of the Royal Society B-Biological Sciences.
692 2015;370(1678). doi: 10.1098/rstb.2014.0331. PubMed PMID: WOS:000360552800011.

693 2. Jeffery NW, Hultgren K, Chak STC, Gregory R, Rubenstein DR. Patterns of genome size
694 variation in snapping shrimp. Genome. 2016;59(6):393-402. doi: 10.1139/gen-2015-0206.

695 3. Stelzer CP, Riss S, Stadler P. Genome size evolution at the speciation level: The cryptic species
696 complex *Brachionus plicatilis* (Rotifera). BMC evolutionary biology. 2011;11. PubMed PMID:
697 ISI:000290202100001.

698 4. Chia J-M, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, et al. Maize HapMap2
699 identifies extant variation from a genome in flux. Nature Genetics. 2012;44:803. doi:
700 10.1038/ng.2313. https://www.nature.com/articles/ng.2313#supplementary-information.

701 5. Ruiz-Ruano FJ, Ruiz-Estevez M, Rodriguez-Perez J, Lopez-Pino JL, Cabrero J, Camacho JPM.
702 DNA Amount of X and B Chromosomes in the Grasshoppers Eyprepocnemis plorans and Locusta
703 migratoria. Cytogenet Genome Res. 2011;134(2):120-6. doi: 10.1159/000324690. PubMed PMID:
704 WOS:000290305100007.

705 6. Šmarda P, Bureš P, Horová L, Rotreklová O. Intrapopulation Genome Size Dynamics in
706 Festuca pallens. Annals of Botany. 2008;102(4):599-607. doi: 10.1093/aob/mcn133.

707 7. Stelzer CP, Pichler M, Stadler P, Hatheuer A, Riss S. Within-Population Genome Size Variation
708 is Mediated by Multiple Genomic Elements That Segregate Independently during Meiosis. Genome
709 Biol Evol. 2019;11(12):3424-35. PubMed PMID: WOS:000514900600009.

710 8. Meyer A, Schloissnig S, Franchini P, Du K, Woltering J, Irisarri I, et al. Giant lungfish genome
711 elucidates the conquest of land by vertebrates. Nature. 2021;590(7845):284-9. doi: 10.1038/s41586-
712 021-03198-8.

713 9. Shah A, Hoffman JI, Schielzeth H. Comparative Analysis of Genomic Repeat Content in
714 Gomphocerine Grasshoppers Reveals Expansion of Satellite DNA and Helitrons in Species with
715 Unusually Large Genomes. Genome Biol Evol. 2020;12(7):1180-93. doi: 10.1093/gbe/evaa119.

716 10. Naville M, Henriet S, Warren I, Sumic S, Reeve M, Volff JN, et al. Massive Changes of Genome
717 Size Driven by Expansions of Non-autonomous Transposable Elements. Current Biology.
718 2019;29(7):1161-+. PubMed PMID: WOS:000462931400023.

719 11. Wong WY, Simakov O, Bridge DM, Cartwright P, Bellantuono AJ, Kuhn A, et al. Expansion of a
720 single transposable element family is associated with genome-size increase and radiation in the
721 genus Hydra. Proceedings of the National Academy of Sciences. 2019;116(46):22915-7.
722 doi: 10.1073/pnas.1910106116.

723 12. Blommaert J, Riss S, Hecox-Lea B, Mark Welch DB, Stelzer CP. Small, but surprisingly
724 repetitive genomes: Transposon expansion and not polyploidy has driven a doubling in genome size
725 in a metazoan species complex. BMC Genomics. 2019;20(466). doi: https://doi.org/10.1186/s12864-
726 019-5859-y.

727 13. McCann J, Macas J, Novák P, Stuessy TF, Villaseñor JL, Weiss-Schneeweiss H. Differential
728 Genome Size and Repetitive DNA Evolution in Diploid Species of Melampodium sect. Melampodium
729 (Asteraceae). Front Plant Sci. 2020;11:362-. doi: 10.3389/fpls.2020.00362. PubMed PMID: 32296454.

730 14. Smarda P, Bures P. Understanding intraspecific variation in genome size in plants. Preslia.
731 2010;82(1):41-61. PubMed PMID: ISI:000277220000004.

732 15. Nogrady T, Wallace RL, Snell TW. Rotifera: Biology, Ecology and Systematics. Dumont HJF,
733 editor. The Hague: SPB Academic Publishing; 1993. 137 p.

734 16. Gilbert JJ. Non-genetic polymorphisms in rotifers: environmental and endogenous controls,
735 development, and features for predictable or unpredictable environments. Biological Reviews.
736 2017;92(2):964-92. doi: https://doi.org/10.1111/brv.12264.

737 17. Riss S, Arthofer W, Steiner FM, Schlick-Steiner BC, Pichler M, Stadler P, et al. Do genome size
738 differences within Brachionus asplanchnoidis (Rotifera, Monogononta) cause reproductive barriers
739 among geographic populations? Hydrobiologia. 2017;796(1):59-75. PubMed PMID:
740 WOS:000403357600005.

741 18. Challis R. Assembly statistic visualization 2017. Available from:
742 https://github.com/rjchallis/assembly-stats.

743 19. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-
744 free profiling of polyploid genomes. Nature Communications. 2020;11(1):1432. doi: 10.1038/s41467-
745 020-14998-3.

746 20. Sun H, Ding J, Piednoël M, Schneeberger K. findGSE: estimating genome size variation within
747 human and Arabidopsis using k-mer frequencies. Bioinformatics. 2017;34(4):550-7. doi:
748 10.1093/bioinformatics/btx637.

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

752 22. Lower SS, McGurk MP, Clark AG, Barbash DA. Satellite DNA evolution: old ideas, new
753 approaches. Current Opinion in Genetics & Development. 2018;49:70-8. doi:
754 https://doi.org/10.1016/j.gde.2018.03.003.

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

757 24. De Coster W, De Rijk P, De Roeck A, De Pooter T, D'Hert S, Strazisar M, et al. Structural
758 variants identified by Oxford Nanopore PromethION sequencing of the human genome. Genome
759 research. 2019;29(7):1178-87. Epub 2019/06/11. doi: 10.1101/gr.244939.118. PubMed PMID:
760 31186302.

761 25. Ruiz-Ruano FJ, Navarro-Domínguez B, López-León MD, Cabrero J, Camacho JPM. Evolutionary
762 success of a parasitic B chromosome rests on gene content. bioRxiv. 2019:683417. doi:
763 10.1101/683417.

764 26. Ahmad SF, Martins C. The Modern View of B Chromosomes Under the Impact of High Scale
765 Omics Analyses. Cells. 2019;8(2):156. doi: 10.3390/cells8020156. PubMed PMID: 30781835.

766 27. Garrido-Ramos MA. Satellite DNA: An Evolving Topic. Genes (Basel). 2017;8(9):230. doi:
767 10.3390/genes8090230. PubMed PMID: 28926993.

768 28. Cavalier-Smith T. Economy, speed and size matter: Evolutionary forces driving nuclear
769 genome miniaturization and expansion. Annals of Botany. 2005;95(1):147-75. doi:
770 10.1093/aob/mci010. PubMed PMID: WOS:000226370900010.

771 29. Gregory TR. The bigger the C-value, the larger the cell: Genome size and red blood cell size in
772 vertebrates. Blood Cell Mol Dis. 2001;27(5):830-43. doi: DOI 10.1006/bcmd.2001.0457. PubMed
773 PMID: WOS:000172785800005.

30. Fussmann G. Rotifers: excellent subjects for the study of macro- and microevolutionary change. Hydrobiologia. 2011;662(1):11-8.

31. 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. doi: 10.1007/s10750-016-2782-y.

32. 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.

33. 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. Epub 2013/05/07. doi: 10.1038/nmeth.2474. PubMed PMID: 23644548.

34. 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. doi: 10.1186/s12864-015-1419-2.

35. 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. doi: 10.1371/journal.pone.0112963.

36. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60. Epub 2009/05/20. doi: 10.1093/bioinformatics/btp324. PubMed PMID: 19451168; PubMed Central PMCID: PMCPMC2705234.

37. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics. 2018;34(13):i142-i50. Epub 2018/06/29. doi: 10.1093/bioinformatics/bty266. PubMed PMID: 29949969; PubMed Central PMCID: PMCPMC6022658.

38. 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. Epub 2015/10/03. doi: 10.1093/bioinformatics/btv566. PubMed PMID: 26428292; PubMed Central PMCID: PMCPMC4708105.

39. Laetsch D, Blaxter M. BlobTools: Interrogation of genome assemblies [version 1; peer review: 2 approved with reservations]. F1000Research. 2017;6(1287). doi: 10.12688/f1000research.12232.1.

40. 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. doi: 10.1093/bioinformatics/btv351.

41. 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. Epub 2007/11/21. doi: 10.1101/gr.6743907. PubMed PMID: 18025269; PubMed Central PMCID: PMCPMC2134774.

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

43. 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. Epub 2020/05/23. doi: 10.1093/nargab/lqaa026. PubMed PMID: 32440658; PubMed Central PMCID: PMCPMC7222226.

817 44. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio
818 prediction of alternative transcripts. Nucleic Acids Res. 2006;34(suppl_2):W435-W9. doi:
819 10.1093/nar/gkl200.

820 45. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale
821 protein function classification. Bioinformatics. 2014;30(9):1236-40. Epub 2014/01/24. doi:
822 10.1093/bioinformatics/btu031. PubMed PMID: 24451626; PubMed Central PMCID:
823 PMCPMC3998142.

824 46. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, et al. UniProtKB/Swiss-
825 Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View.
826 Methods Mol Biol. 2016;1374:23-54. Epub 2015/11/01. doi: 10.1007/978-1-4939-3167-5_2. PubMed
827 PMID: 26519399.

828 47. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics.
829 Genome Biology. 2019;20(1):238. doi: 10.1186/s13059-019-1832-y.

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

833 49. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for
834 automated genomic discovery of transposable element families. Proceedings of the National
835 Academy of Sciences. 2020;117(17):9451-7. doi: 10.1073/pnas.1921046117.

836 50. Platt RN, II, Blanco-Berdugo L, Ray DA. Accurate Transposable Element Annotation Is Vital
837 When Analyzing New Genome Assemblies. Genome Biol Evol. 2016;8(2):403-10. doi:
838 10.1093/gbe/evw009.

839 51. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification
840 system for eukaryotic transposable elements. Nature Reviews Genetics. 2007;8(12):973-82. doi:
841 10.1038/nrg2165.

842 52. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res.
843 1999;27(2):573-80. doi: 10.1093/nar/27.2.573.

844 53. Park JC, Choi B-S, Kim M-S, Shi H, Zhou B, Park HG, et al. The genome of the marine rotifer
845 Brachionus koreanus sheds light on the antioxidative defense system in response to 2-ethyl-
846 phenanthrene and piperonyl butoxide. Aquatic Toxicology. 2020;221:105443. doi:
847 https://doi.org/10.1016/j.aquatox.2020.105443.

848 54. Han J, Park JC, Choi B-S, Kim M-S, Kim H-S, Hagiwara A, et al. The genome of the marine
849 monogonont rotifer Brachionus plicatilis: Genome-wide expression profiles of 28 cytochrome P450
850 genes in response to chlorpyrifos and 2-ethyl-phenanthrene. Aquatic Toxicology. 2019;214:105230.
851 doi: https://doi.org/10.1016/j.aquatox.2019.105230.

852 55. Kim H-S, Lee B-Y, Han J, Jeong C-B, Hwang D-S, Lee M-C, et al. The genome of the freshwater
853 monogonont rotifer Brachionus calyciflorus. Molecular Ecology Resources. 2018;18(3):646-55. doi:
854 10.1111/1755-0998.12768.

855 56. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods.
856 2012;9:357-9. doi: 10.1038/nmeth.1923. PubMed PMID: 22388286.

857 57. Namiki T, Hachiya T, Tanaka H, Sakakibara Y. MetaVelvet: an extension of Velvet assembler
858 to de novo metagenome assembly from short sequence reads. Nucleic Acids Res. 2012;40(20):e155.
859 Epub 2012/07/24. doi: 10.1093/nar/gks678. PubMed PMID: 22821567; PubMed Central PMCID:
860 PMCPMC3488206.

861 58. Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: evaluation of metagenome assemblies.
862 Bioinformatics. 2016;32(7):1088-90. Epub 2015/11/29. doi: 10.1093/bioinformatics/btv697. PubMed
863 PMID: 26614127.

864 59. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome
865 Biology. 2019;20(1):257. doi: 10.1186/s13059-019-1891-0.

866 60. Suga K, Mark Welch DB, Tanaka Y, Sakakura Y, Hagiwara A. Two Circular Chromosomes of
867 Unequal Copy Number Make Up the Mitochondrial Genome of the Rotifer Brachionus plicatilis.
868 Molecular Biology and Evolution. 2008;25(6):1129-37. doi: 10.1093/molbev/msn058.

869 61. Xu H, Luo X, Qian J, Pang X, Song J, Qian G, et al. FastUniq: A Fast De Novo Duplicates
870 Removal Tool for Paired Short Reads. PLOS ONE. 2012;7(12):e52249. doi:
871 10.1371/journal.pone.0052249.

872 62. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of
873 occurrences of k-mers. Bioinformatics. 2011;27(6):764-70. doi: 10.1093/bioinformatics/btr011.

874 63. Shen W, Le S, Li Y, Hu F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File
875 Manipulation. PLOS ONE. 2016;11(10):e0163962. doi: 10.1371/journal.pone.0163962.

876 64. Benaglia T, Chauveau D, Hunter DR, Young DS. mixtools: An R Package for Analyzing Mixture
877 Models. Journal of Statistical Software. 2009;32(6):29. Epub 2009-10-14. doi: 10.18637/jss.v032.i06.

878 65. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
879 Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9. doi:
880 10.1093/bioinformatics/btp352.

881 66. R Development Core Team. R: A Language and Environment for Statistical Computing. 2020.

882 67. Hadley W. stringr: Simple, Consistent Wrappers for Common String Operations 2019.
883 Available from: https://CRAN.R-project.org/package=stringr.

884 68. Hadley W. Reshaping Data with the reshape Package. Journal of Statistical Software.
885 2007;21(12):1-20.

886 69. Hadley W. ggplot2: Elegant Graphics for Data Analysis 2016. Available from:
887 https://ggplot2.tidyverse.org.

888 70. Wilke CO. cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2' 2020.
889 Available from: https://CRAN.R-project.org/package=cowplot.

890 71. Wilkins D. treemapify: Draw Treemaps in 'ggplot2' 2021. Available from: https://CRAN.R-
891 project.org/package=treemapify.

892 72. Francis F, Dumas MD, Wisser RJ. ThermoAlign: a genome-aware primer design tool for tiled
893 amplicon resequencing. Scientific Reports. 2017;7(1):44437. doi: 10.1038/srep44437.

894

895

896

897

898

Supplementary Information for:

Genome structure of *Brachionus asplanchnoidis*, a Eukaryote with intrapopulation variation in genome size

Stelzer, C.P., Blommaert J., Waldvogel A.M., Pichler M., Hecox-Lea, B. & Mark Welch D.B.

Correspondence to: claus-peter.stelzer@uibk.ac.at

This file includes:

Supplementary notes on short-read preprocessing

Figures S1 to S19

Tables S1 to S8

References for SI citations

Data to be submitted as electronic supplementary files:

Supplementary file 1 (VBCF report on genome assembly and contaminant filtering; pdf-file)
Supplementary file 2 (Summary of short-read preprocessing and fastqc-reports; xlsx-file)
Supplementary file 3 (Kmer-based analysis of cleaned Illumina reads, xlsx-file)
Supplementary file 4 (Ranges of all CNVs across the *B. asplanchnoidis* genome; csv-file)
Supplementary file 5 (Input parameters of the thermoalign pipeline; txt-file)
Supplementary file 6 (Detailed information on top-36 contributing repeat elements; xlsx-file)
Supplementary file 7 (Repeat profile, Gene density, and CNVs of the 50 largest contigs; html-file)
Supplementary file 8 (Combined library of curated and uncurated repeat elements; fasta-file)

Data to be submitted to public databases:

- Raw reads
- Assembly
- Gene annotation files
- Repeat annotation

Supplementary notes on short-read preprocessing

In total, we obtained 265.6 Gbp of Illumina short-read data from the 15 different *B. asplanchnoidis* clones. After all preprocessing steps had been completed, this amount was reduced to 194.3 Gbp. Total alignment rates to the reference genome improved from 90.9-98.4%, after quality-filtering only, to 94.1-98.4%, after contaminant removal (**Supplementary file 2**) Contamination rates were variable among the different libraries (**Fig. S1**). Contaminant DNA mostly derived from bacteria, of which *Pseudomonas toyotensis* was the only bacterial genome that could be identified with our metagenomic assembly-based approach. Among the libraries 3-30% of the unmapped reads could be assigned to other microbial contaminants. Contamination by the food algae *Tetraselmis* was low to absent depending on library (usually less than 1% of unmapped reads), and protozoan contamination was below the detection threshold in all libraries. After preprocessing, the genome coverages ranged between 9.4-79.0-fold, if the Gbp sequence of each library is expressed as multiples of the 230.1Mb *B. asplanchnoidis* reference genome, with the majority of libraries being above 20-fold (**Supplementary file 2**). Quality control using fastqc [1] indicated substantial improvements in most parameters in the course of the pre-processing pipeline, with the notable exception of GC-content (**Supplementary file 2**)

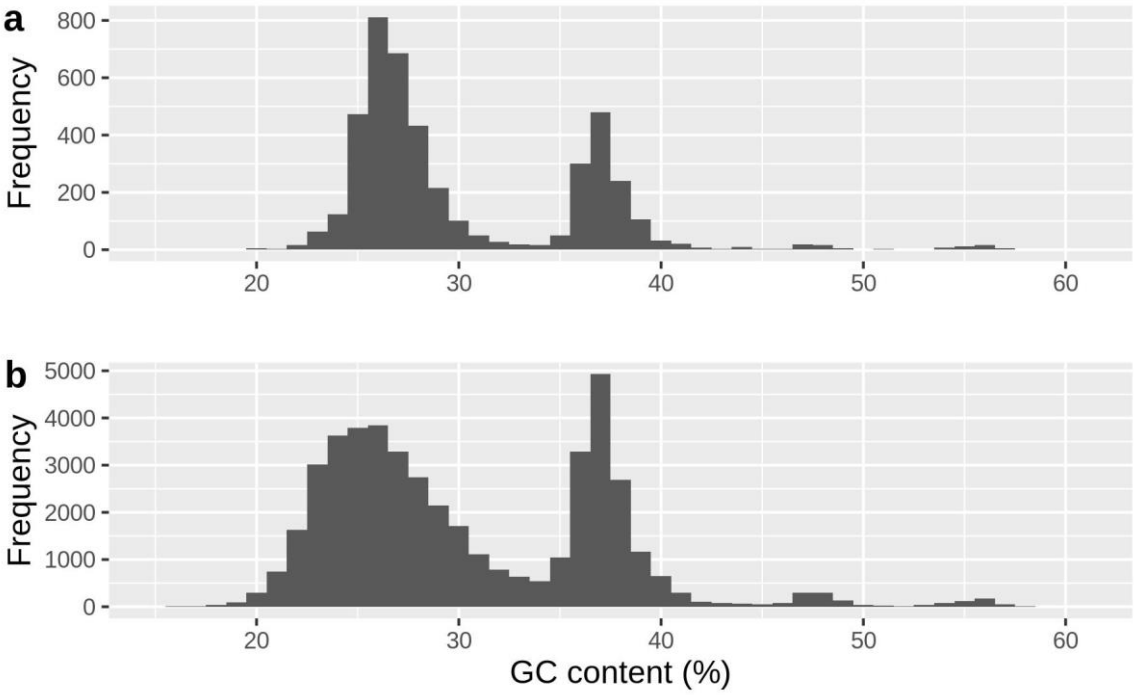


Fig. S1: GC-distribution in the reference assembly. **a** distribution based on 50kbp-windows, **b** distribution based on 5kbp-windows.

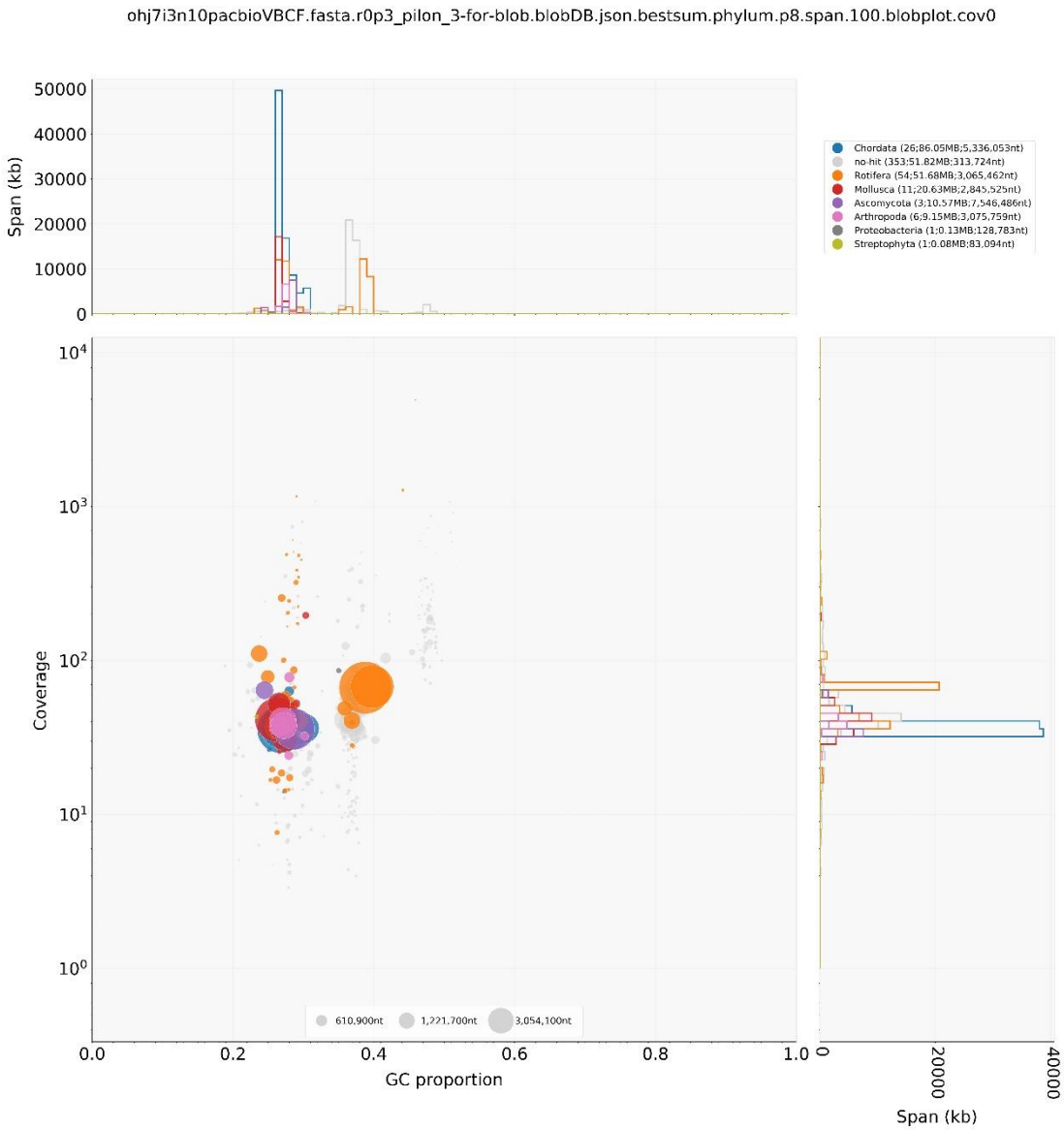


Fig. S2: Blobplot of cleaned OHJ7i3n10 reads against reference genome

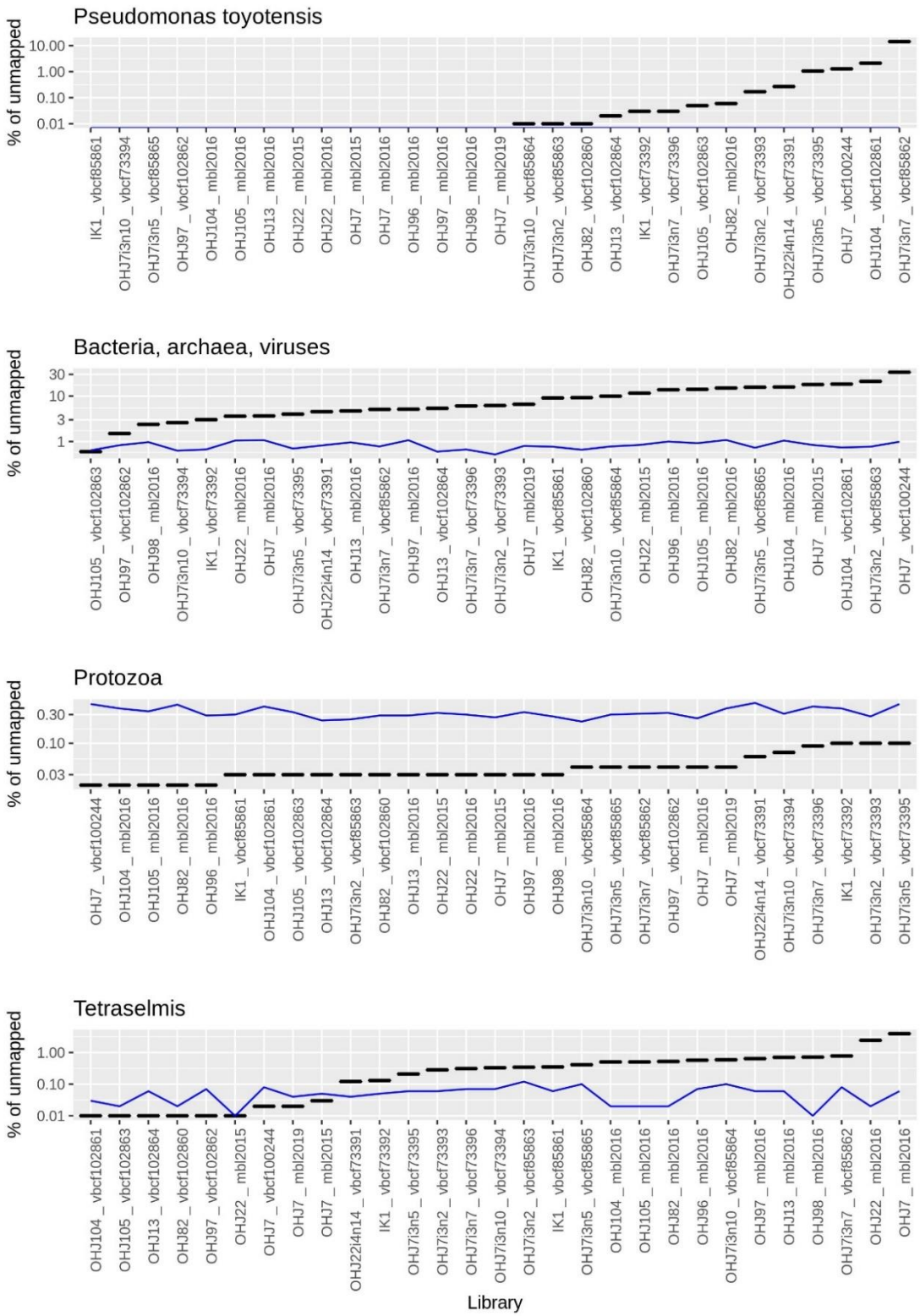


Fig. S3: Contamination rates of short-read libraries (before contaminant removal). Horizontal dashes indicate the percentage of reads in the unmapped-reads fraction (i.e., quality-trimmed reads that did *not* align to the reference genome in the first alignment) that could be assigned to four classes of contaminants: *Pseudomonas toyotensis*, (b) other bacteria, archaea, viruses, (c) protozoan contaminants, and (d) *Tetraselmis* (food algae). The blue line indicates the "false-discovery rate", which was obtained by applying the same pipeline to the mapped-reads fraction of each library.

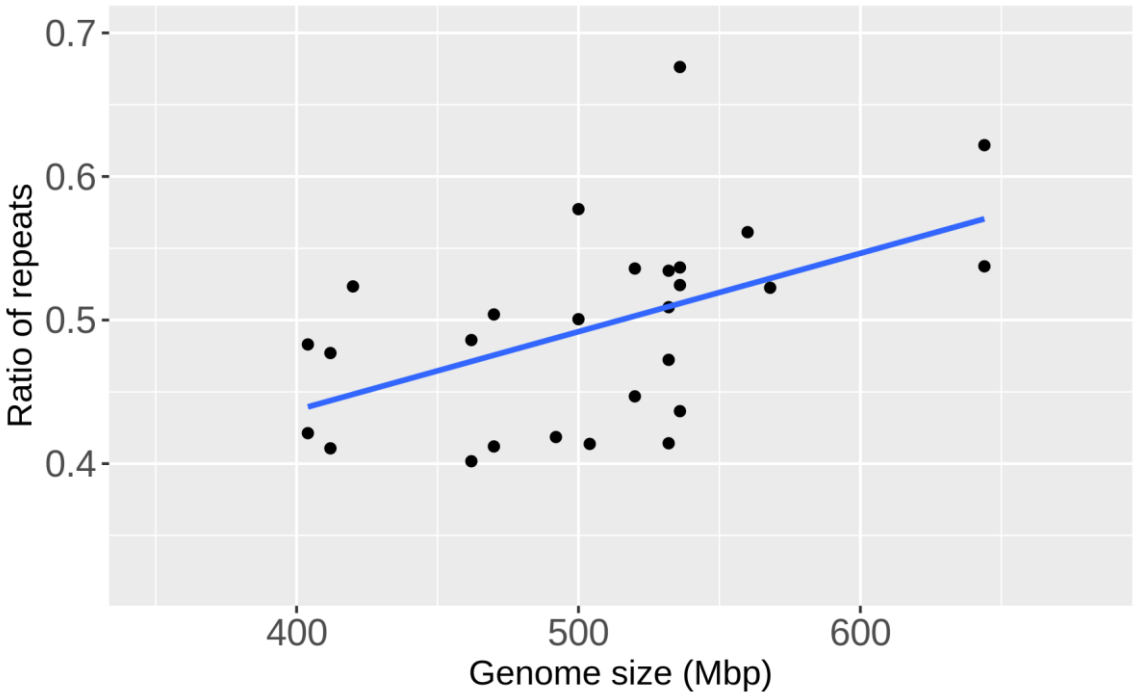


Fig. S4: Genome size (2C, as measured by flow cytometry) versus Ratio of repeats, a fitted parameter of findGSE [2]. The two variables are significantly correlated with each other (Spearman-rank correlation test, ρ=0.609, *P*=0.0057)

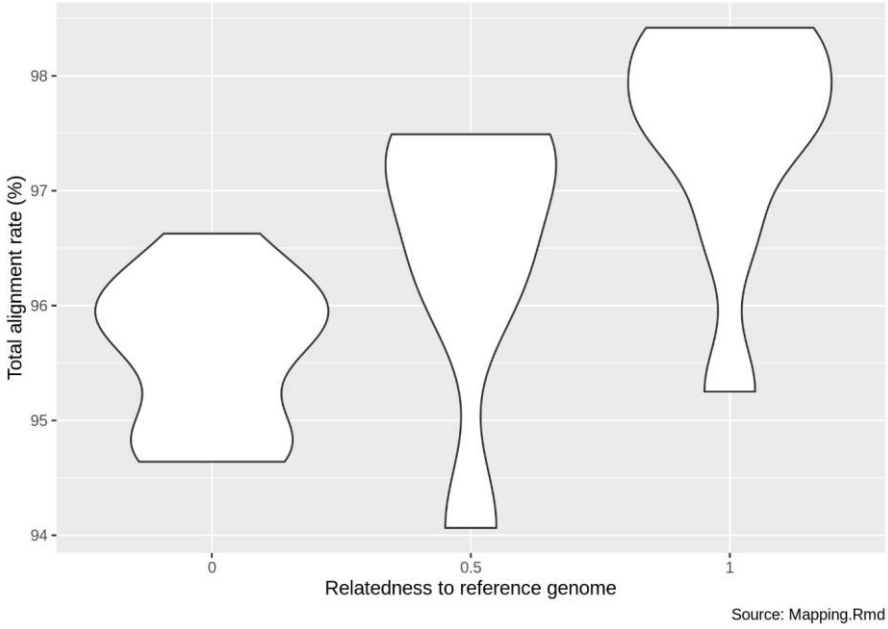


Fig. S5: Total alignment rate *vs*. relatedness of DNA source to the reference genome. Relatedness was assumed to be 1 for all clones within the OHJ7i3-selfed line, 0.5 for the inbred line cross (IK1) and the natural ancestor of the inbred line (OHJ7), and 0 for all other clones which hatched from the natural population.

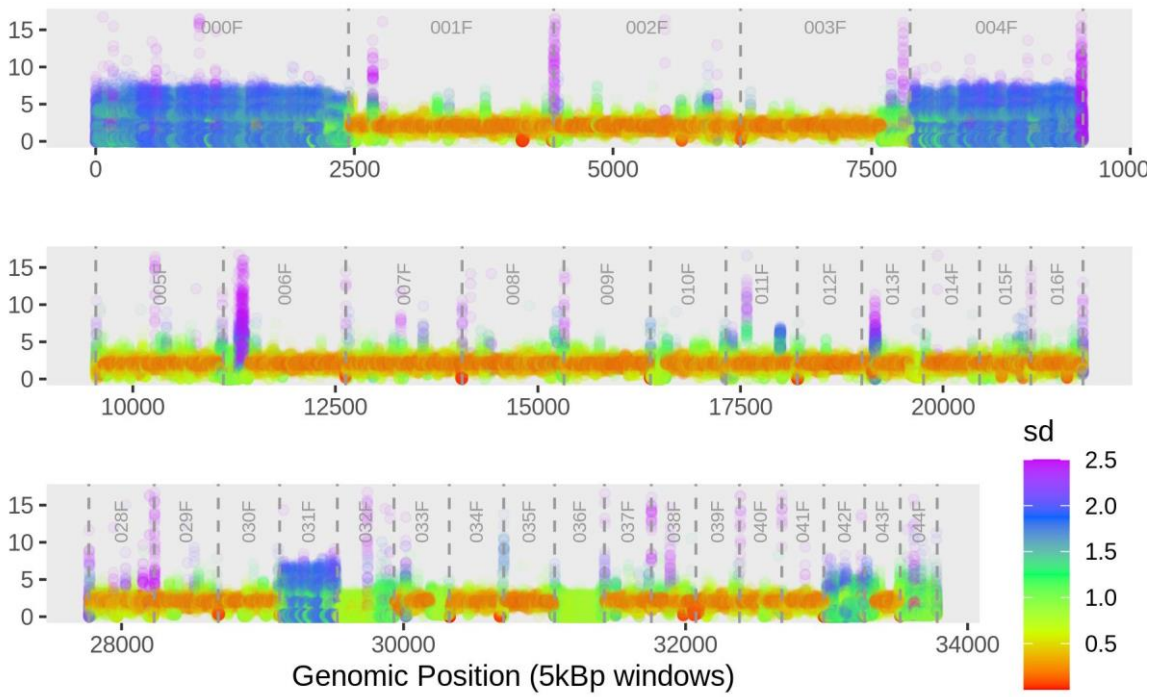
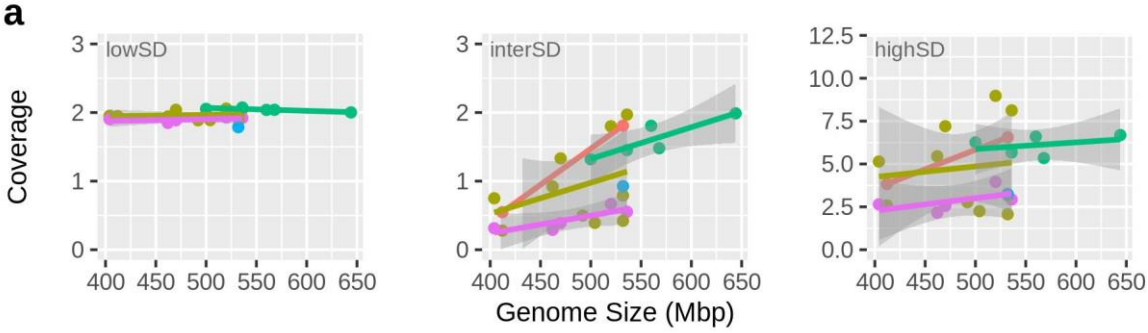


Fig. S6: Coverage and coverage variation along representative contigs of the *B. asplanchnoidis* genome (5kBp resolution). Each circle represents a 5kBp window of one of the 31 sequencing libraries. Coverage (y-axis) was normalized by dividing the per-bp coverage of each window by 1/2 of the (mean) exon coverage of the respective library. Contig borders are indicated by vertical dashed lines, and contig IDs are listed on the top. All contigs displayed here span ~140 Mbp in total, i.e. about 60% of the assembly. Variation across clones/libraries is indicated by color (based on standard deviations). Standard deviations larger than 2.5 (about 4% of the data) had to be capped to a value 2.5, to allow colored visualization.

a



b

Coverage variability	*t*	*df*	Pearson's *r*	*P*
lowSD	1.75	21	0.36	0.0939
interSD	4.38	21	0.69	0.0003
highSd	2.12	21	0.42	0.0459

Fig. S7: Genomic regions of elevated coverage variability are responsible for genome size variation. This figure shows the same analysis as in Fig. 4 but it excludes the library preps "C" (c.f., Tab. S4), which showed considerably higher coverage and higher GC content in interSD and highSD regions than the other libraries. **a** Mean coverage versus genome size [flow cytometry data on genome size was taken from 3]. Dots represent the mean coverage per library. Colors indicate co-prepared libraries (as in **Tab. S4**): Orange = A, Pink = B, Green = D, Blue= E, Gold = F. **b** partial correlations between the variables 'mean coverage' and 'genome size'.

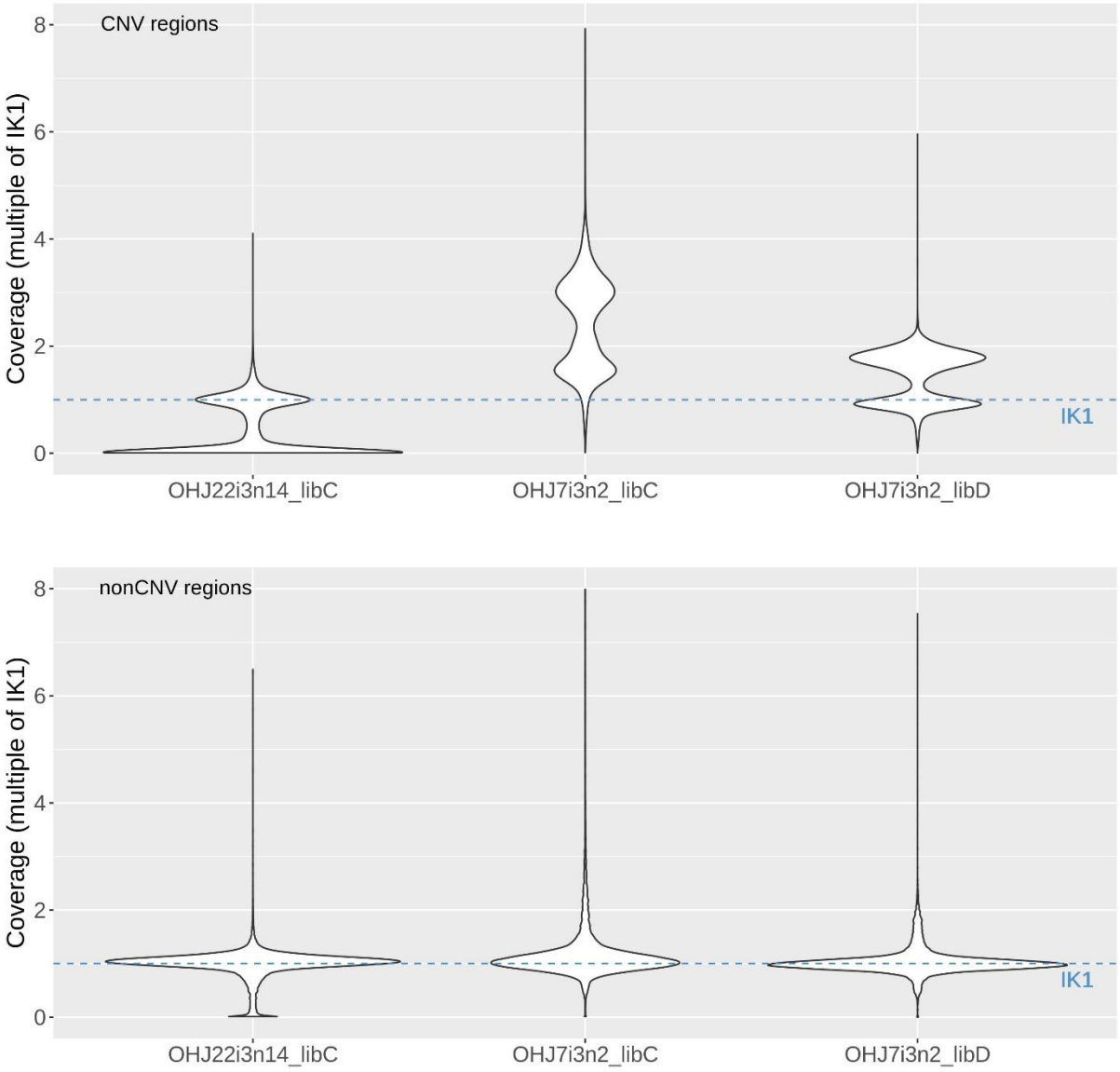


Fig. S8: Genome-wide coverage patterns of two parental clones and their cross. This figure summarizes per-base coverage estimates of all 5000bp windows along the genome (subdivided in CNV regions and non-CNV regions, respectively). Coverages of the two parental clones (OHJ22i3n14, OHJ7i3n2) are expressed as multiple of the coverage of the crossed offspring clone (IK1). The dashed line at 1 refers to the crossed offspring IK1. The extensions “libC” and “libD” designate co-prepared libraries as in **Tab S4**.

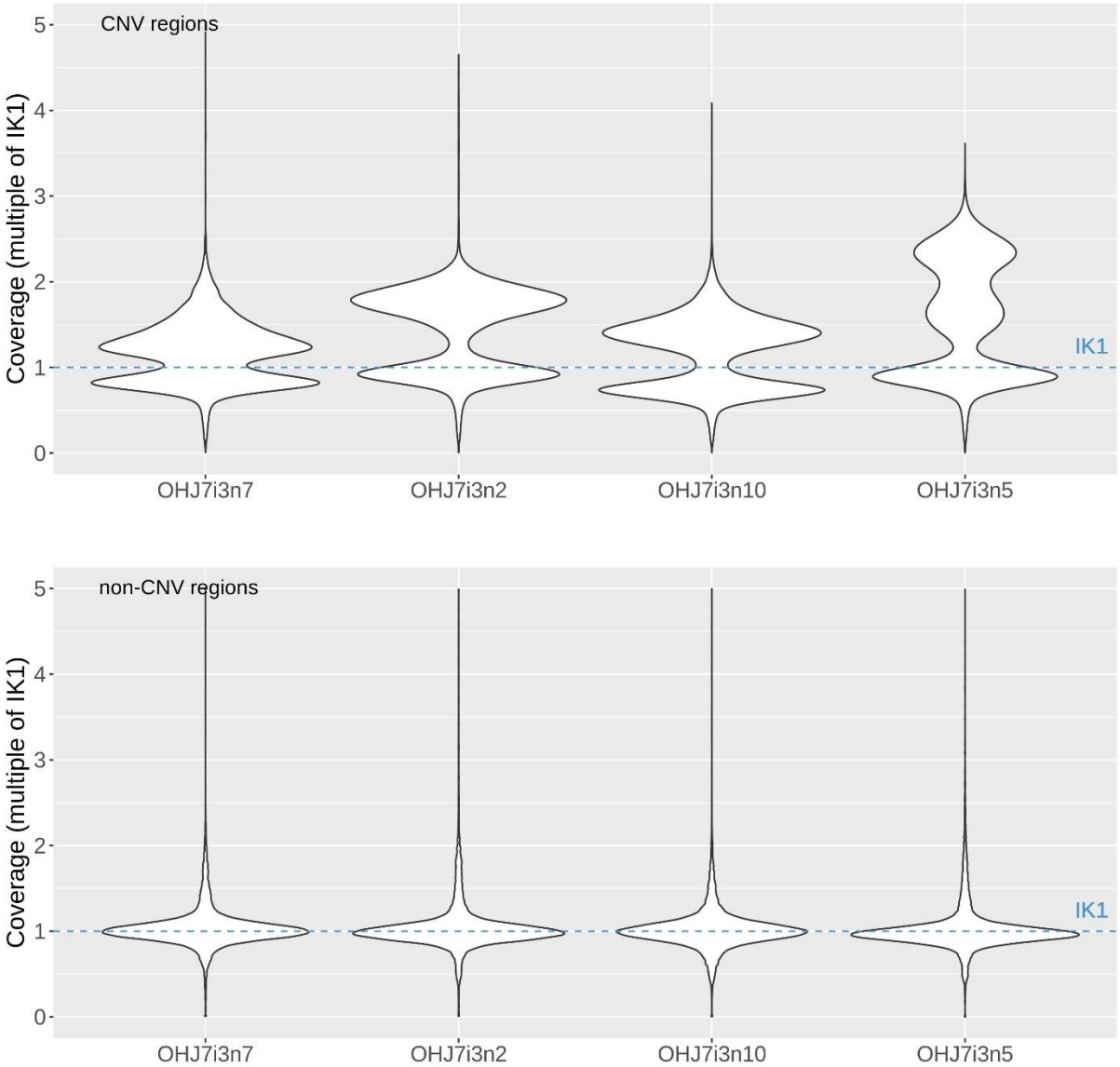


Fig. S9: Genome-wide coverage patterns of multiple clones of a selfed line. This figure summarizes per-base coverage estimates of all 5000bp windows along the genome (subdivided in CNV regions and nonCNV regions, respectively). Coverages of the four clones are expressed as multiple of the coverage of the crossed offspring clone (IK1). The dashed line at 1 refers to the crossed offspring IK1. Clones are ordered in ascending genome size: OHJ7i3n7 (536Mbp), OHJ7i3n2 (560MBp), OHJ7i3n10 (568Mbp), OHJ7i3n5 (644Mbp). Only the library preparations "D" Tab. S4 are displayed here, but library preparations "C" exhibited a very similar pattern (not shown).

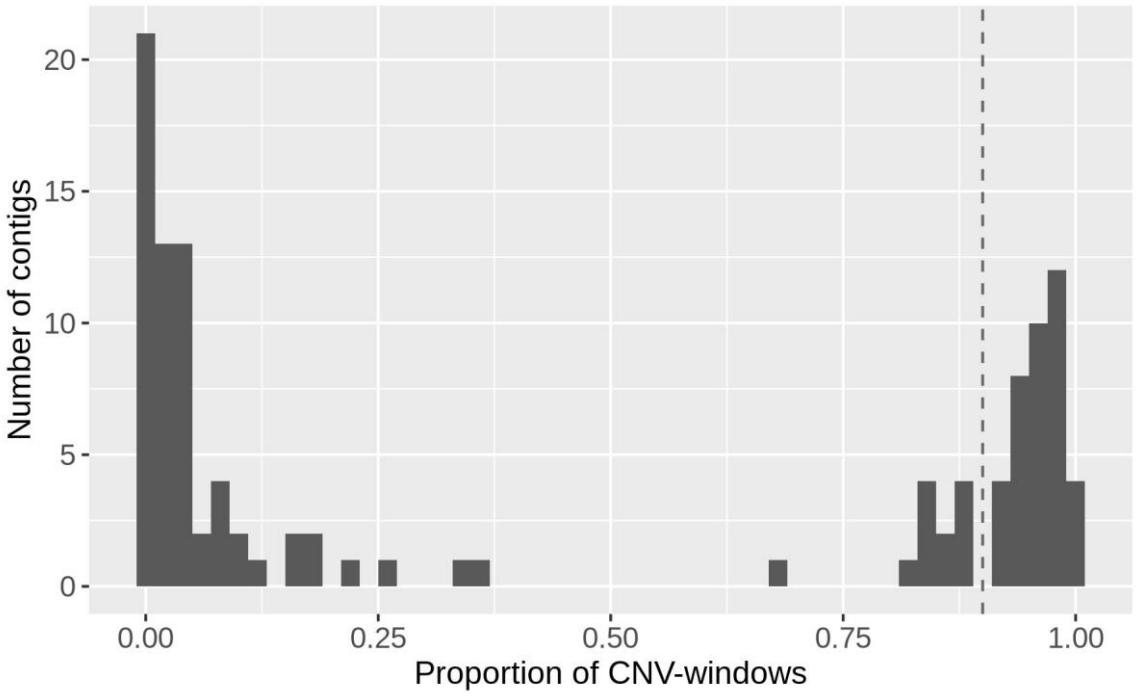


Fig. S10: Contig space that is occupied by CNVs. The X-axis displays the proportion of all 5kbp windows in a contig that have been classified as part of a CNV. For this graph, we only used contigs with at least 50 windows in length (i.e., contig length > 250kbp; n=114). The dashed line is our threshold at 0.9 for defining 'B-contigs'.

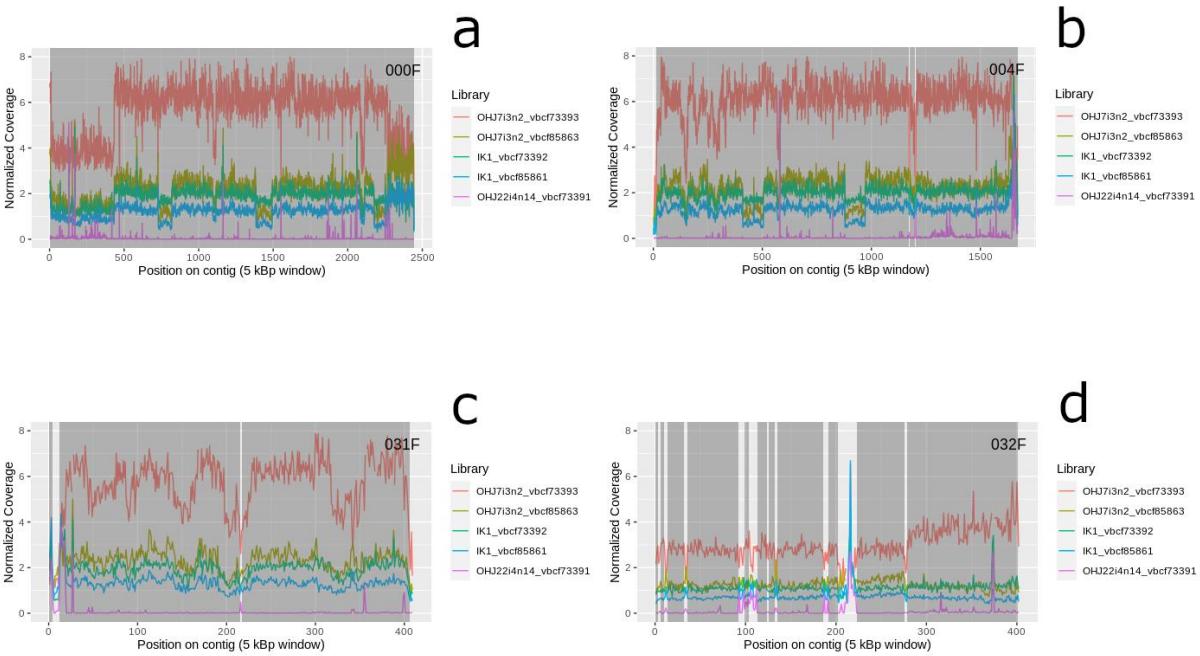


Fig. S11: Examples of "B-contigs". B-contigs consist almost exclusively of coverage-variable regions, with a consistent coverage pattern**.** Shaded regions highlight calls of individual CNVs, while open regions are coverage breakpoints, either due to low coverage variation and/or shifts in the coverage patterns among clones. Colored lines represent normalized coverage values of different rotifer clones/libraries (with "2" equaling the coverage at exonic regions). The displayed clones are a "trio" consisting of the parents (OHJ7i3n2, 2C genome size: 560 Mbp; OHJ22i4n14, 420 Mbp) and their sexual offspring (IK1, 500 Mbp).

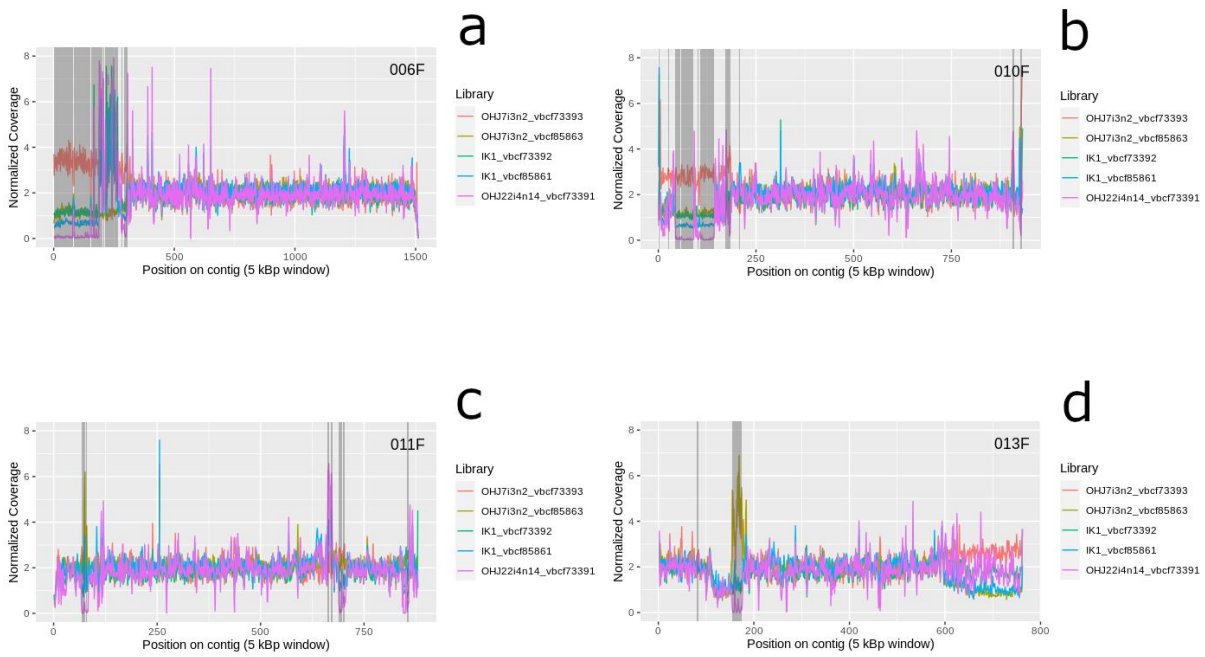


Fig. S12: Example of small-scale coverage variation. The contigs displayed here contain shorter sections that have been identified as CNVs. Shaded regions highlight calls of individual CNVs, while open regions are coverage breakpoints, either due to low coverage variation and/or shifts in the coverage patterns among clones. Colored lines represent normalized coverage values of different rotifer clones/libraries (with "2" equaling the coverage at exonic regions). The displayed clones are a "trio" consisting of the parents (OHJ7i3n2, 2C genome size: 560 Mbp; OHJ22i4n14, 420 Mbp) and their sexual offspring (IK1, 500 Mbp).

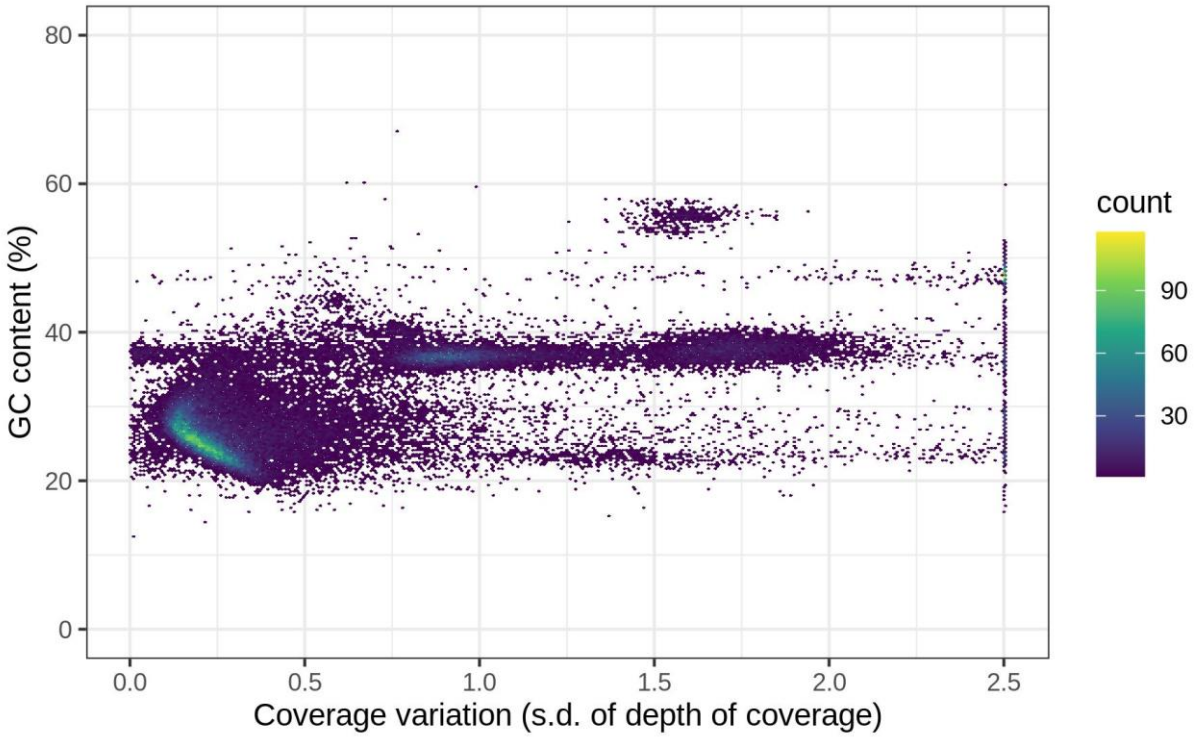


Fig. S13: GC-content vs. coverage variability in different regions of the genome of *B. asplanchnoidis*. Hexbin chart is based on all 5kbp windows of the assembly (bins=250, in ggplot2).

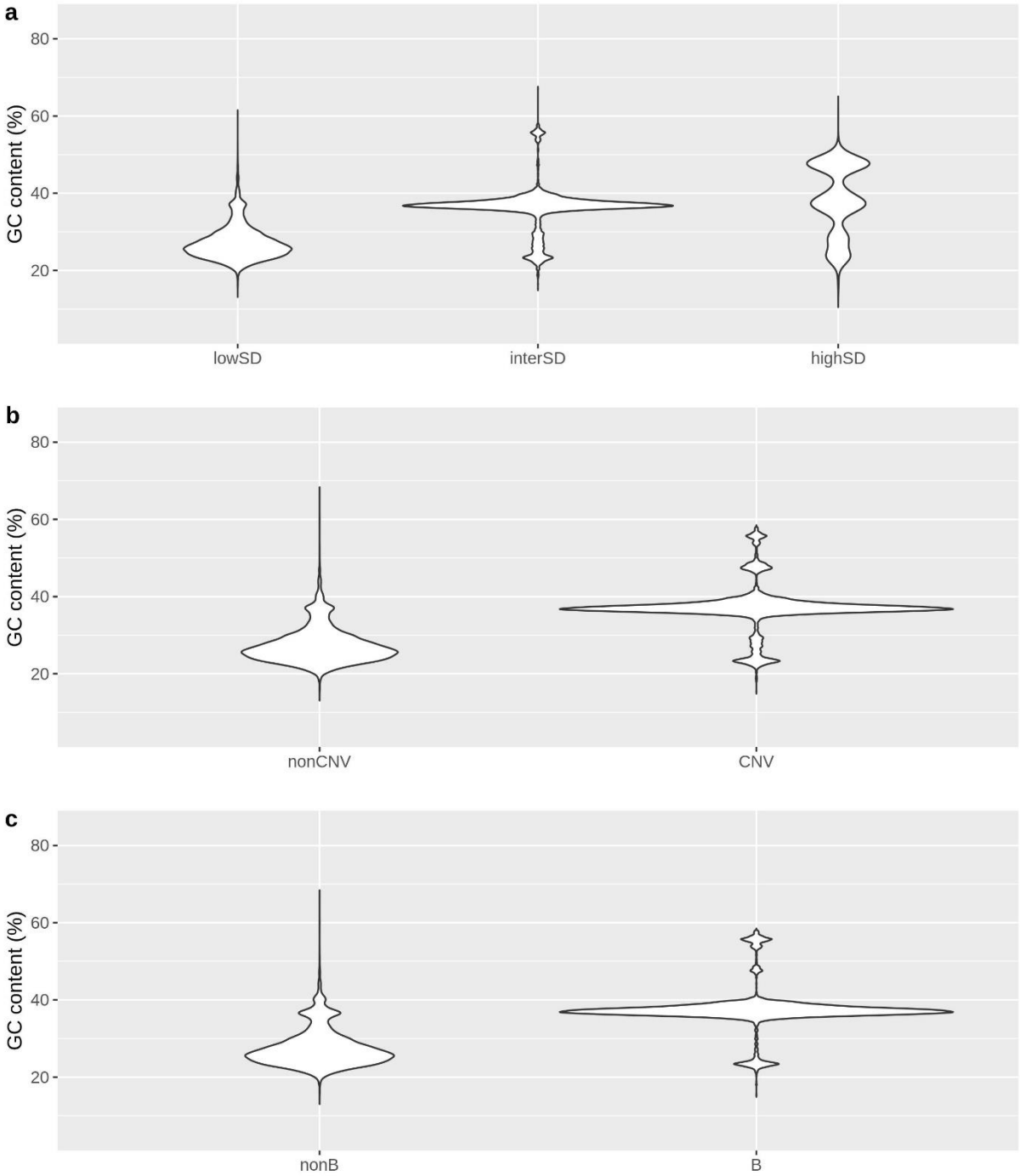


Fig. S14: GC-content of genomic regions for in the *B. asplanchnoidis* genome. **a** Regions defined by coverage variability, **b** CNV regions (i.e., elevated coverage variability & locally consistent coverage patterns), **c** B-contigs (contigs consisting of >90% CNVs). For **c**, only contigs with at least 50 windows were considered. The distributions shown here are based on GC-contents calculated from 5000-bp windows. To prevent losses in precision, the last window of a contig was excluded (since it was always shorter than 5000 bp).

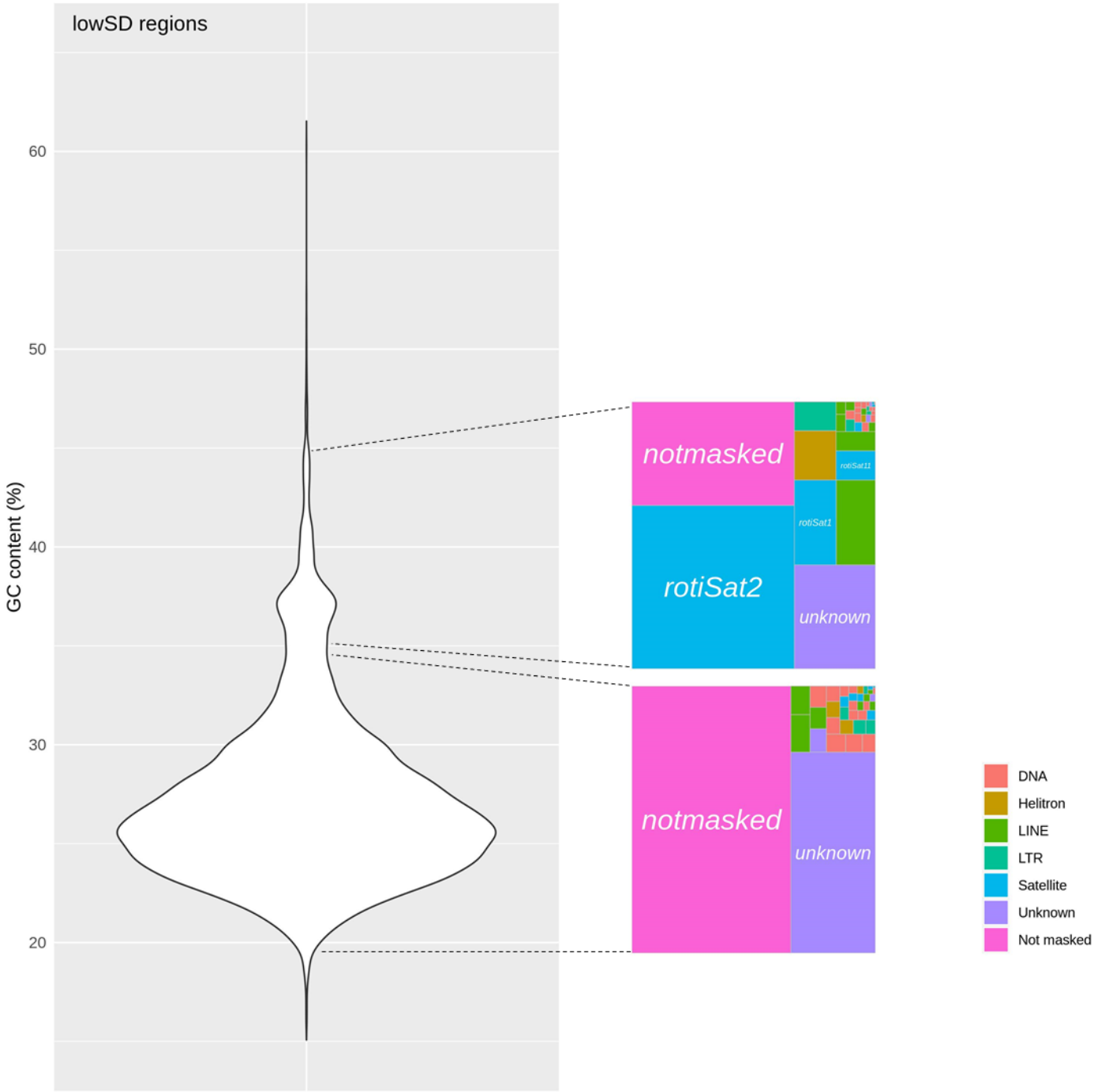


Fig. S15: Repeat composition of various GC fractions of lowSD regions. 'Unknown' refers to regions that were masked as repeats but could not be ascribed to any of the above categories. 'Not masked' refers to regions that were not identified as repeats by repeatModeler2.

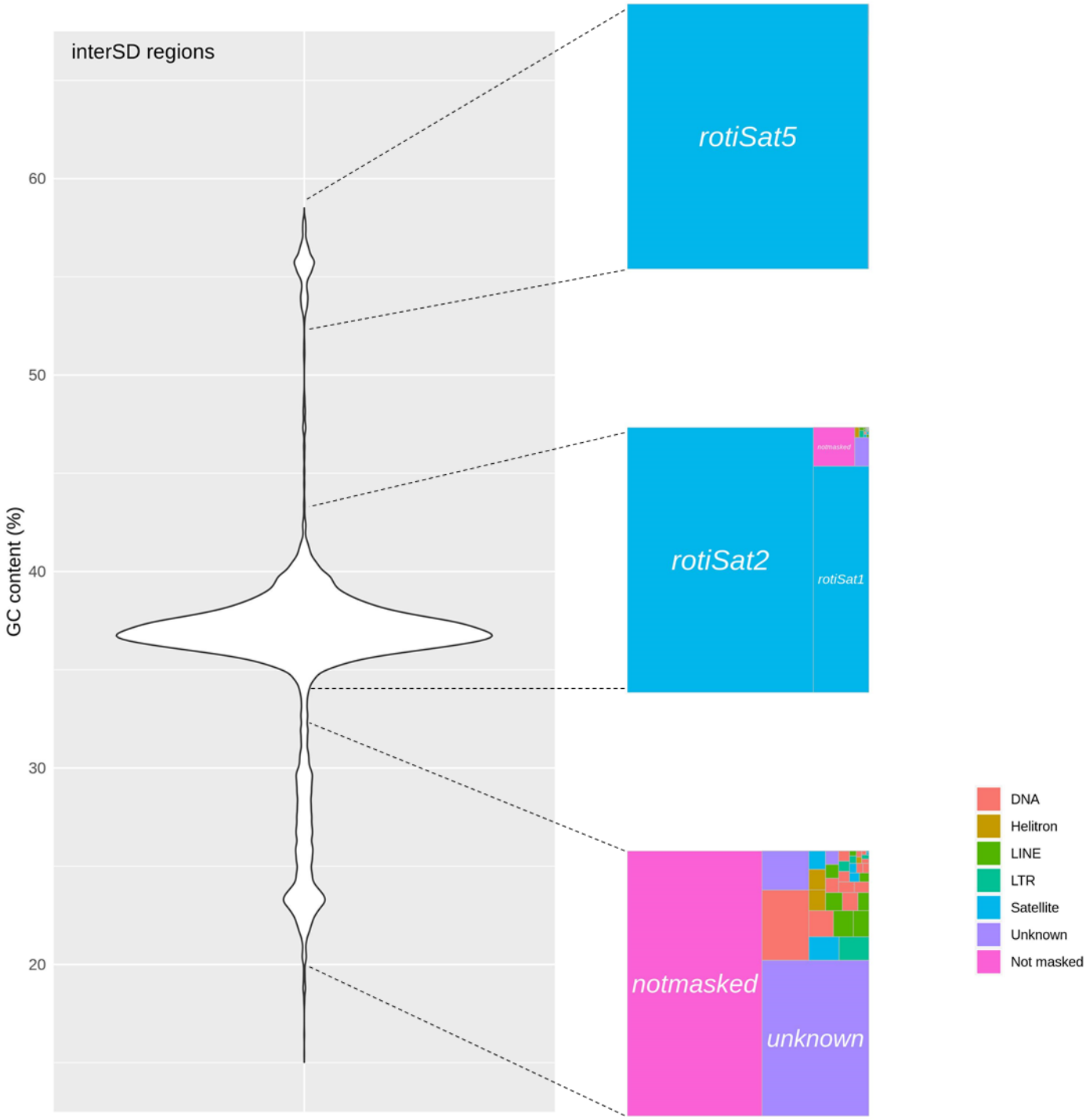


Fig. S16: Repeat composition of various GC fractions of interSD regions.

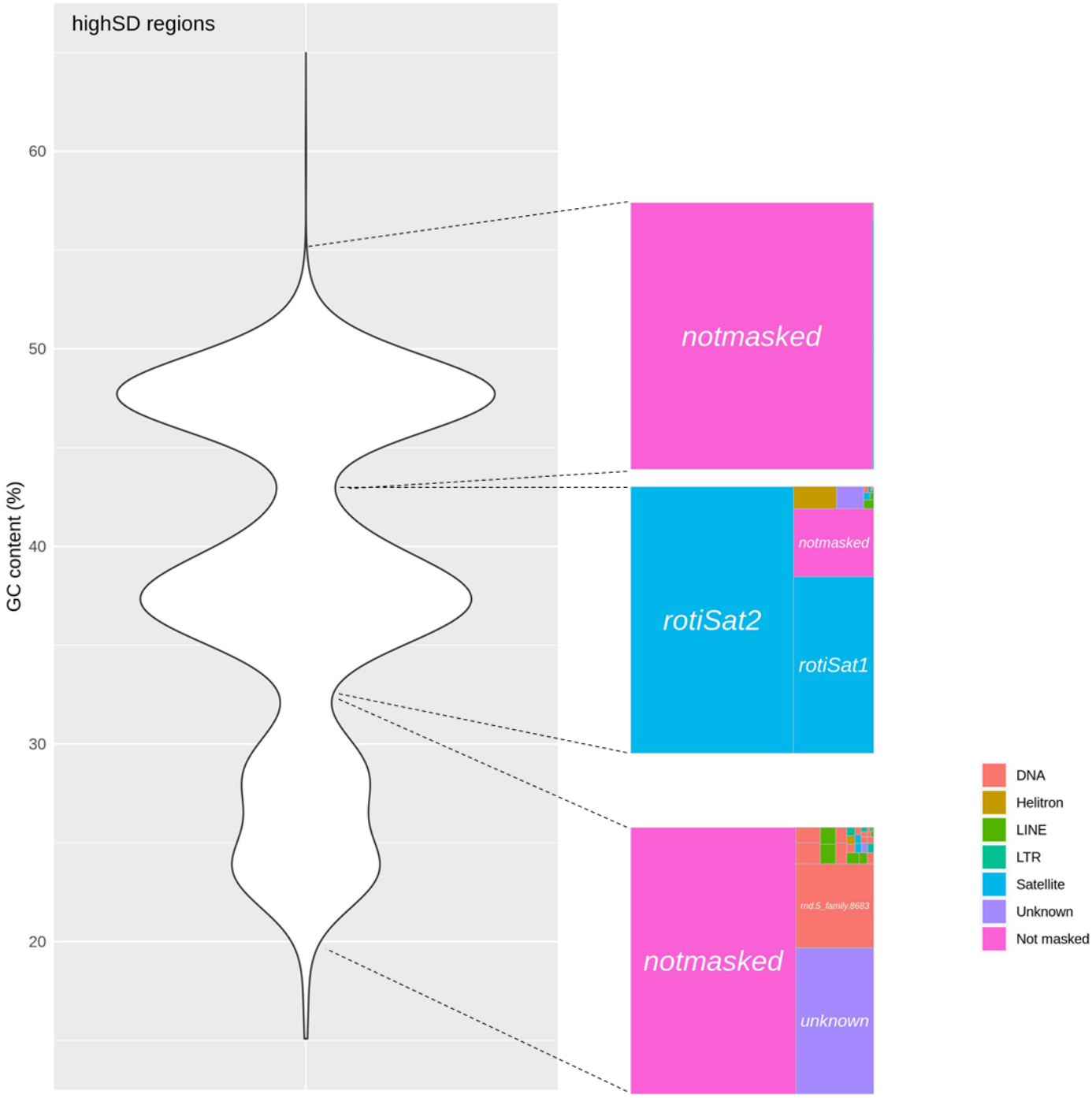


Fig. S17: Repeat composition of various GC fractions of highSD regions.

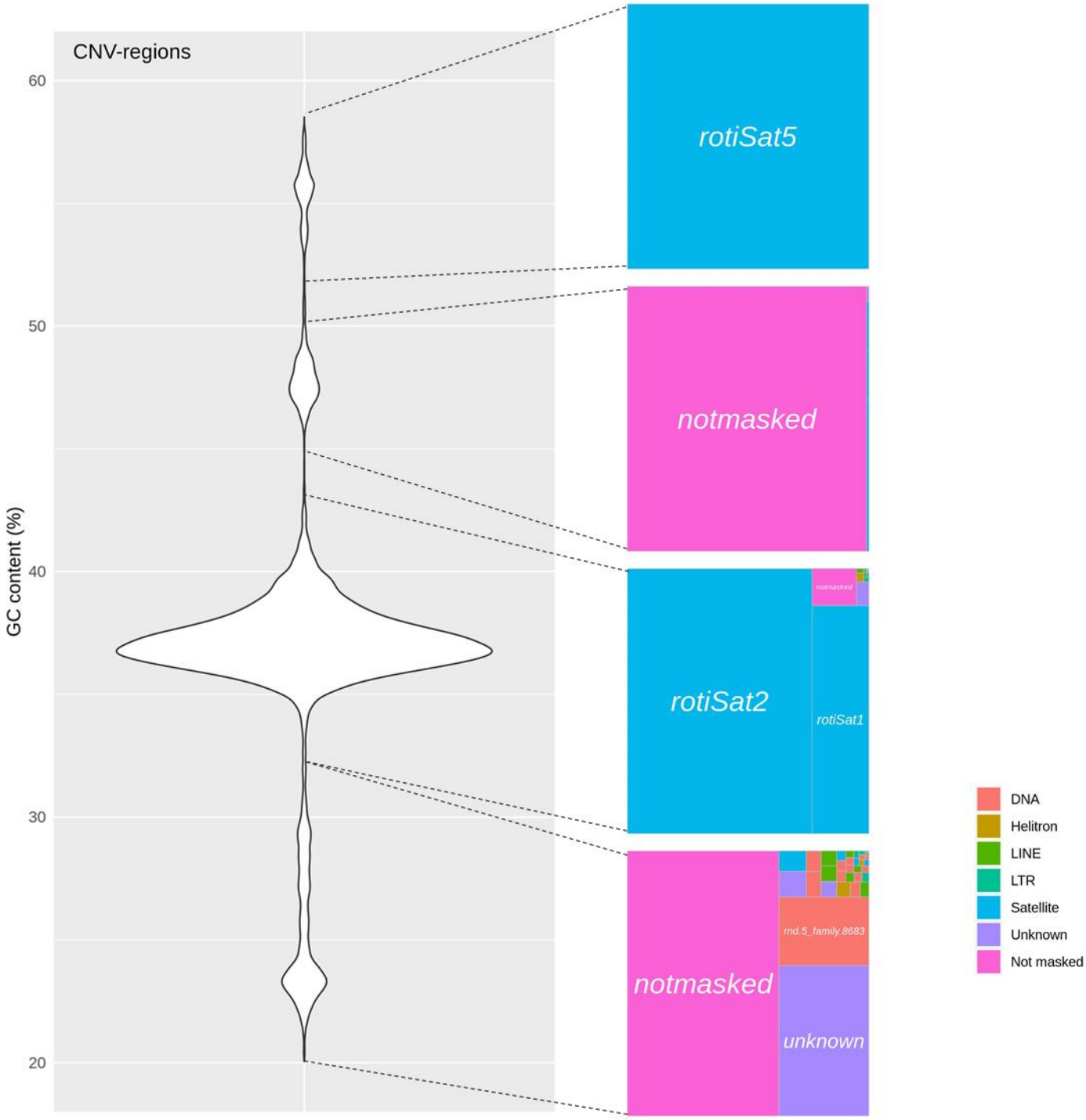


Fig. S18: Repeat composition of various GC fractions of CNV regions.

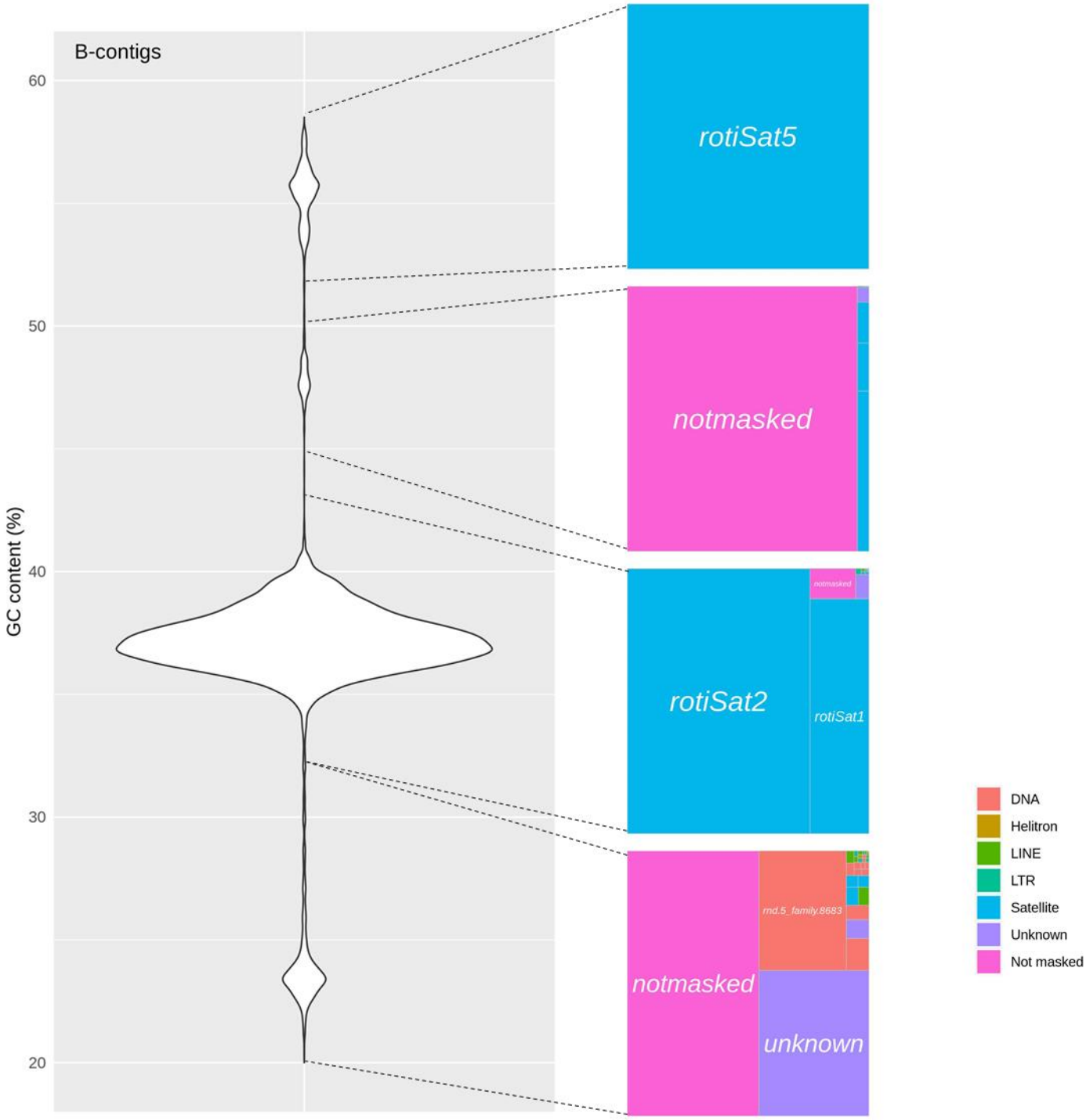


Fig. S19: Repeat composition of various GC fractions of B-contigs.

Tab. S1: Outline of transcriptome and proteome libraries used for gene annotation.

Species | clone	N transcripts (T) | protein seqs (P)	Reference
***B. asplanchnoidis* | ohj7i3n10**	T: 21157	
***B. asplanchnoidis* | ohj22**	P: 12547	Blommaert et al. 2019
***B. plicatilis* | Tokyo1**	P: 12484	Blommaert et al. 2019
***B. rotundiformis* | Italy2**	P: 11050	Blommaert et al. 2019
B. spec | TiscarSM28	P: 12085	Blommaert et al. 2019

Table S2: BUSCO completeness statistics of *B. asplanchniodis* genome in comparison to two previously published genomes of related *Brachionus* species.

BUSCO (Metazoa n=978)	*B. asplanchnoidis*	*B. plicatilis*	*B. calyciflorus*
Complete %	90.6	95.0	90.7
Single complete %	87.9	91.7	88.0
Duplicated complete %	2.7	3.3	2.7
Fragmented %	2.0	2.0	2.0
Missing %	7.4	3.0	7.3

Data sources: Published genome of *B. plicatilis* by [4] and of *B. calyciflorus* by [5]

Table S3: Summary statistics of B. *asplanchnoidis* annotation in comparison to annotations of previously published genomes from related *Brachionus* species [6]. For reasons of data availability, this comparison does not include more species of the *B. plicatilis* species complex.

	B. asplanchnoidis	*B. asplanchnoidis* ohj22	*B. rotundiformis* Italy2	*B.* spec TiscarSM28	*B. plicatilis* Tokyo1
Number					
Gene	16667	12547	11050	12085	12484
mRNA	17419	12547	11050	12085	12484
Exon	92799	87600	91192	96546	93603
CDS	96847	87188	90882	96132	93249
Mean					
mRNAs/gene	1.05	1	1	1	1
Exons/mRNA	5.33	6.98	8.25	7.99	7.5
CDSs/mRNA	5.56	6.95	8.22	7.95	7.47
Median length					
Gene	1999	1839	1874	1912	1926.5
mRNA	2060	1839	1874	1912	1926.5
Exon	153	62	70	61	60
Intron	59	66	58	63	67
CDS	141	62	70	62	60
Total space					
Gene	59373627	29922141	25565912	29432444	31741372
mRNA	59373549	29922141	25565912	29432444	31741372
Exon	28793789	15401203	16387408	16563187	16266105
CDS	27027022	15373829	16368700	16537800	16239684
Single					
Exon mRNA	2881	514	249	315	444
CDS mRNA	2907	528	257	323	460

Table S4: Information on short-read sequencing libraries

Clone	Origin	GS[1] (Mbp)	Method	Prep.[2]	max. FS[3] (Bp)	avg. FS[3] (Bp)
OHJ82	natural clone	404	westburg	F	2000	503
			kapa	B	600	450
OHJ22	natural clone	412	kapa	A	600	450
			kapa	B	600	450
OHJ104	natural clone	462	westburg	F	2000	499
			kapa	B	600	450
OHJ97	natural clone	470	westburg	F	2000	499
			kapa	B	600	450
OHJ96	natural clone	492	kapa	B	600	450
OHJ98	natural clone	504	kapa	B	600	450
OHJ105	natural clone	520	westburg	F	2000	493
			kapa	B	600	450
OHJ7	natural clone	532	westburg	E	1000	386
			kapa	A	600	450
			kapa	B[4]	600	450
			kapa	B[4]	600	450
OHJ13	natural clone	536	westburg	F	2000	516
			kapa	B	600	450
OHJ22i3n14	selfed line	420	westburg	C	1000	545
OHJ7i3n7	selfed line	536	westburg	C	1000	495
			westburg	D	3000	1149
OHJ7i3n2	selfed line	560	westburg	C	1000	632
			westburg	D	3000	1228
OHJ7i3n10	selfed line	568	westburg	C	1000	589
			westburg	D	3000	1093
OHJ7i3n5	selfed line	644	westburg	C	1000	496
			westburg	D	3000	1129
IK1	selfed line cross	500	westburg	C	1000	514
			westburg	D	2000	1001

[1] genome size (2C) estimated by flow cytometry in [3]

[2] Indicates co-prepared libraries

[3] Fragment size (FS) of the library

[4] Both derive from the same original preparation, but one was re-sequenced and had an intermediate PCR-step

Table S5: Alignment statistics of short reads to reference genome.

Rotifer clone	Library ID	TAR (%)	CCA (%)	CCA1 (%)	DAS (%)
IK1	vbcf73392	96.5	91.4	30.9	5.1
IK1	vbcf85861	96.1	94.0	36.5	2.1
OHJ104	vbcf102861	94.7	91.9	42.1	2.8
OHJ105	vbcf102863	95.9	93.0	34.6	2.9
OHJ13	vbcf102864	96.1	93.6	34.2	2.5
OHJ22i4n14	vbcf73391	94.9	90.8	36.7	4.1
OHJ7	vbcf100244	94.1	92.8	34.1	1.3
OHJ7i3n10	vbcf73394	98.4	95.8	25.4	2.6
OHJ7i3n10	vbcf85864	97.7	96.3	37.8	1.4
OHJ7i3n2	vbcf73393	98.4	96.4	22.8	2.0
OHJ7i3n2	vbcf85863	95.3	93.5	33.5	1.8
OHJ7i3n5	vbcf73395	97.6	93.1	26.5	4.5
OHJ7i3n5	vbcf85865	96.7	95.1	32.2	1.5
OHJ7i3n7	vbcf73396	97.5	93.6	25.9	3.9
OHJ7i3n7	vbcf85862	98.2	96.9	38.1	1.3
OHJ82	vbcf102860	94.7	91.8	43.0	3.0
OHJ97	vbcf102862	95.8	92.9	38.2	2.9
OHJ104	mbl2016	94.6	92.9	52.0	1.8
OHJ105	mbl2016	95.7	94.0	42.0	1.6
OHJ13	mbl2016	96.6	95.0	45.5	1.7
OHJ22	mbl2015	94.8	91.6	48.4	3.3
OHJ22	mbl2016	95.6	93.3	51.4	2.3
OHJ7	mbl2015	97.2	95.1	49.6	2.1
OHJ7	mbl2016	97.5	96.2	52.9	1.3
OHJ82	mbl2016	95.3	93.5	49.4	1.8
OHJ96	mbl2016	95.9	94.3	47.4	1.6
OHJ97	mbl2016	96.4	94.8	49.6	1.6
OHJ98	mbl2016	96.1	94.3	50.3	1.8
OHJ7	mbl2019	97.4	96.0	42.9	1.4

TAR = Total alignment rate; CCA = Concordantly aligned reads; CCA1 = Concordantly aligned reads (exactly once); DAS = Discordantly aligned reads and singletons

Table S6: Classification of 'B-contigs' in terms of the proportion of 5kbp windows that are CNV

Contig	Total 5kbp windows	CNV-windows	Proportion
000F	2444	2436	0.997
057F	166	165	0.994
058F	163	162	0.994
071F	109	108	0.991
004F	1671	1651	0.988
083F	79	78	0.987
084F	78	77	0.987
086F	72	71	0.986
048F	210	207	0.986
091F	66	65	0.985
099F	59	58	0.983
059F	155	152	0.981
109F	50	49	0.980
061F	146	143	0.979
055F	175	171	0.977
022F	564	550	0.975
031F	409	396	0.968
094F	62	60	0.968
097F	61	59	0.967
095F	60	58	0.967
062F	146	141	0.966
053F	175	169	0.966
047F	232	224	0.966
110F	52	50	0.962
036F	353	337	0.955
044F	262	250	0.954
064F	135	128	0.948
070F	114	108	0.947
068F	108	102	0.944
054F	176	166	0.943
067F	122	115	0.943
106F	52	49	0.942
107F	52	49	0.942
080F	83	78	0.940
088F	69	64	0.928
074F	107	99	0.925
056F	173	160	0.925
045F	252	231	0.917

Table S7: PCR-primers used for verification of selected CNV-loci

Locus	Position in genome			PCR primers					
	Contig	Start	End	Orientation	Sequence (5'->3')	Length (mer)	GC (%)	Tm (°C)	Amplicon (bp)
TA_000F	000F	471	681	forward	TCCAAAGGACGTTGATTAACCTG	23	43	58.9	210
				reverse	GGTTGATCAAGCTCCTTAGGAC	22	50	60.3	
TA_003F	003F	3642291	3642472	forward	TTTCGGGTAAGTCGACTTATTCC	23	43	58.9	181
				reverse	ATCCGGCATGTAGTATGAAACTG	23	43	58.9	
TA_001F	001F	841341	841438	forward	GCGGTCCAAATTCCCAAATTC	21	47	57.9	97
				reverse	GCGTCGCAATGTTGGAAAGC	20	55	59.4	
TA_032F	032F	62	163	forward	GCTCATCTCTGATGGTATTTGTCC	24	45	61.0	101
				reverse	CAAGGCAAAGTCATCAGTGAAGTC	24	45	61.0	

Table S8: Data from ddPCR-measurements.

		Estimated copies/nl			Droplet counts		
Clone	**Primer**	**Estimate**	**ConfMax**	**ConfMin**	**Positives**	**Negatives**	**Accepted**
ik1	TA_001F	40.1	43.6	38.3	497	14340	14837
ik1	TA_003F	44.1	47.8	42.1	523	13708	14231
ik1	TA_032F	63.4	67.9	61	751	13574	14325
ik1	TA_000F	0.58	1.12	0.38	7	14308	14315
neg	TA_001F	0	0.22	0	0	15814	15814
neg	TA_003F	0	0.24	0	0	14881	14881
neg	TA_032F	0.07	0.35	0.02	1	15833	15834
neg	TA_000F	1.2	2	1	16	15155	15171
ohj104	TA_001F	80.1	85	77.7	1072	15206	16278
ohj104	TA_003F	58.2	62.6	56	700	13791	14491
ohj104	TA_032F	0.15	0.48	0.06	2	15859	15861
ohj104	TA_000F	0.53	1.04	0.35	7	15476	15483
ohj105	TA_001F	59.7	64	57.4	715	13746	14461
ohj105	TA_003F	45.5	49.8	43.4	444	11256	11700
ohj105	TA_032F	1.3	2	1	16	14976	14992
ohj105	TA_000F	1.3	2	1	16	14668	14684
ohj13	TA_001F	53.2	57.3	51.1	658	14229	14887
ohj13	TA_003F	48.2	52.1	46.3	602	14382	14984
ohj13	TA_032F	44.9	48.6	43	552	14195	14747
ohj13	TA_000F	0.76	1.34	0.54	10	15477	15487
ohj22	TA_001F	89.6	94.8	87	1160	14650	15810
ohj22	TA_003F	72	76.8	69.6	865	13705	14570
ohj22	TA_032F	0	0.23	0	0	15068	15068
ohj22	TA_000F	0.57	1.11	0.38	7	14428	14435
ohj7	TA_001F	61	65.4	58.7	726	13645	14371
ohj7	TA_003F	50	54	48	608	13992	14600
ohj7	TA_032F	109.9	115.5	107	1463	14942	16405
ohj7	TA_000F	1.27	1.97	0.98	17	15772	15789
ohj7i3n10	TA_001F	68.5	73.2	66.1	815	13589	14404
ohj7i3n10	TA_003F	56.7	61.1	54.4	636	12888	13524
ohj7i3n10	TA_032F	133.4	139.9	130.1	1629	13570	15199
ohj7i3n10	TA_000F	0.59	1.1	0.4	8	15982	15990
ohj7i3n2	TA_001F	41	44.5	39.2	539	15200	15739
ohj7i3n2	TA_003F	38.9	42.4	37.2	502	14917	15419
ohj7i3n2	TA_032F	127.6	133.7	124.5	1680	14664	16344
ohj7i3n2	TA_000F	0.76	1.33	0.54	10	15523	15533
ohj7i3n5	TA_001F	46.2	50.1	44.2	529	13218	13747
ohj7i3n5	TA_003F	40.7	44.5	38.7	430	12222	12652
ohj7i3n5	TA_032F	126.7	133.3	123.4	1429	12565	13994
ohj7i3n5	TA_000F	0.43	0.94	0.26	5	13668	13673
ohj7i3n7	TA_001F	65.1	70.9	62.1	481	8458	8939
ohj7i3n7	TA_003F	46.5	50.3	44.6	600	14874	15474
ohj7i3n7	TA_032F	156	163	152	1919	13578	15497
ohj7i3n7	TA_000F	0.51	0.99	0.34	7	16221	16228
ohj82	TA_001F	74.5	79.2	72.1	949	14519	15468
ohj82	TA_003F	65.1	69.8	62.7	727	12783	13510
ohj82	TA_032F	0.07	0.35	0.02	1	16175	16176
ohj82	TA_000F	1.1	1.79	0.83	14	14950	14964
ohj96	TA_001F	57	61.3	54.8	688	13859	14547
ohj96	TA_003F	45.7	49.4	43.8	590	14901	15491
ohj96	TA_032F	54.2	58.3	52.1	653	13858	14511

ohj96	TA_000F	0.33	0.79	0.19	4	14106	14110
ohj97	TA_001F	62	66.1	59.9	857	15846	16703
ohj97	TA_003F	44.5	48	42.7	620	16079	16699
ohj97	TA_032F	55	58.9	53	749	15652	16401
ohj97	TA_000F	0.3	0.7	0.17	4	15757	15761
ohj98	TA_001F	57.4	61.7	55.2	689	13772	14461
ohj98	TA_003F	51.1	55.2	49.1	604	13594	14198
ohj98	TA_032F	1.16	1.86	0.89	15	15193	15208
ohj98	TA_000F	1.3	2	1	16	15009	15025

References for SI-citations

1. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010.
2. 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. doi: 10.1093/bioinformatics/btx637.
3. 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. PubMed PMID: WOS:000514900600009.
4. 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. Aquatic Toxicology. 2019;214:105230. doi: https://doi.org/10.1016/j.aquatox.2019.105230.
5. 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. Molecular Ecology Resources. 2018;18(3):646-55. doi: 10.1111/1755-0998.12768.
6. 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). doi: https://doi.org/10.1186/s12864-019-5859-y.