G3 Genes | Genomes | Genetics

GENOME REPORT

A High-Quality Genome Assembly from Short and Long Reads for the Non-biting Midge *Chironomus riparius* (Diptera)

Hanno Schmidt,*,† Sören Lukas Hellmann,‡ Ann-Marie Waldvogel,* Barbara Feldmeyer,* Thomas Hankeln,‡ and Markus Pfenninger*,§,1

*Molecular Ecology Group, Senckenberg Biodiversity and Climate Research Centre, Frankfurt am Main, Hessen, Germany, †Vector Genetics Laboratory, Department of Pathology, Microbiology and Immunology, School of Veterinary Medicine, University of California Davis, CA, ‡Molecular Genetics and Genome Analysis Group, Institute of Organismic and Molecular Evolutionary Biology, Johannes Gutenberg-University, Mainz, Germany, and §LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG), Frankfurt, Germany

ORCID IDs: 0000-0001-8915-891X (H.S.); 0000-0003-4958-1419 (S.L.H.); 0000-0003-2637-0766 (A.-M.W.); 0000-0002-0413-7245 (B.F.); 0000-0002-1547-7245 (M.P.)

ABSTRACT *Chironomus riparius* is of great importance as a study species in various fields like ecotoxicology, molecular genetics, developmental biology and ecology. However, only a fragmented draft genome exists to date, hindering the recent rush of population genomic studies in this species. Making use of 50 NGS datasets, we present a hybrid genome assembly from short and long sequence reads that make *C. riparius*' genome one of the most contiguous Dipteran genomes published, the first complete mitochondrial genome of the species, and the respective recombination rate among the first insect recombination rates at all. The genome assembly and associated resources will be highly valuable to the broad community working with dipterans in general and chironomids in particular. The estimated recombination rate will help evolutionary biologists gaining a better understanding of commonalities and differences of genomic patterns in insects.

KEYWORDS
hybrid genome assembly
Chironomus riparius
recombination rate

Non-biting midges (Chironomidae) are dipterans like the model organisms *Drosophila* fruit flies and *Anopheles* mosquitoes. The species *Chironomus riparius* (synonym *C. thummi* or *C. thummi thummi*) is particularly important in ecotoxicological (Williams *et al.* 1986; Lee *et al.* 2009), molecular genetic (Hankeln *et al.* 1997; Hankeln and Schmidt 1990), developmental (Klomp *et al.* 2015) and ecological (Armitage *et al.* 1995; Rosenberg 1992; Foucault *et al.* 2019) research. Recently, *C. riparius* has also emerged as a promising organism for transcriptomic (Schmidt *et al.* 2013; Nair *et al.* 2011; Marinković *et al.* 2012) and genomic studies (Oppold *et al.* 2017; Waldvogel *et al.* 2018). Although important population genomic parameters are already available for *C. riparius* (*e.g.*, the mutation rate μ; (Oppold and Pfenninger 2017)), analyses still rely on a fragmented Illumina-only genome assembly (Oppold *et al.* 2017). Here we present a high-quality hybrid genome assembly from short and long reads, along with an estimate for the species-specific recombination rate, the first complete mitochondrial genome for this species and a reference transcriptome based on several life stages. This is an important step forward to enable more complex genomic studies on *C. riparius* and hence understand variability in dipteran genome evolution patterns.

doi: https://doi.org/10.1534/g3.119.400710
Manuscript received January 11, 2020; accepted for publication February 7, 2020; published Early Online February 14, 2020.

Supplemental material available at figshare: https://doi.org/10.25387/g3.11575059.
[1]Corresponding author: Senckenberg Biodiversity and Climate Research Centre, Frankfurt am Main, Germany. E-mail: markus.pfenninger@senckenberg.de.

MATERIALS AND METHODS

Assembly strategy

The assembly of the Illumina-PacBio-hybrid genome consisted of five major steps: (1) De Bruijn graph assembly of the Illumina reads, (2) hybrid assembly of Illumina contigs and raw PacBio reads, (3) error correction of hybrid contigs by mapping of Illumina data, (4) scaffolding of the contigs using mate-pair reads, (5) closing the remaining gaps with corrected PacBio and Illumina paired end sequences.

Table 1 – *Characteristics of* C. riparius *genome assembly*. Shown are the improvements in quality by combining short and long reads in comparison to the previous Illumina-only assembly. Values are based on the nuclear genome only

	Illumina-only draft genome (Oppold *et al.* 2017)	Hybrid assembly (present study)	Degree of improvement
number of scaffolds	5,292	752	1/7
total scaffold length (bp)	180,652,019	178,167,951	equal
average scaffold length (bp)	34,136	236,926	x 7
longest sequence (bp)	2,056,324	2,626,431	+25%
N50	272,065	539,778	x 2
N content (%)	15.96	0.08	1/200
BUSCOs found (complete and fragmented)	92.8	93.7	+ 1%

Samples and PacBio sequencing

Long reads (Supplementary Table S1, dataset 01) were sequenced from 52 female imagines that originated from one egg clutch of a strict inbred line (described in (Oppold and Pfenninger 2017)) of the same *C. riparius* laboratory culture that has been used for previous draft genome sequencing (Oppold *et al.* 2017). DNA was isolated with the QIAGEN Gentra Puregene Tissue Kit according to manufacturer's instructions and sequenced on six SMRT Cells on a Pacific Biosciences RS II machine.

Genome assembly

Illumina data (Supplementary Table S1, datasets 02-06) was sequenced from approximately 50 larvae from a long-standing laboratory culture (Oppold *et al.* 2017). Quality processing of the reads was done using Trimmomatic v0.32 (Bolger *et al.* 2014) with default parameters and FastQC v0.11.3 (Andrews 2010). Additionally, we filtered out mitochondrial reads using BBDuk from the tool package BBMap v35.85 (Bushnell 2014) with k = 41 and hdist = 2.

We assembled the quality processed Illumina reads using the De Bruijn graph assembler Platanus v1.2.4 (Kajitani *et al.* 2014) with kmer-sizes between k = 32 and k = 84 and s = 6. The resulting contigs plus the raw PacBio reads were then used as input for the program DBG2OLC v1.0 (Ye *et al.* 2016) and assembled with recommended settings (k = 17, KmerCovTh = 2, MinOverlap = 20, AdaptiveTh = 0.002). The assembly was screened by BLASTN searches (Blast v2.3.0+ (Altschul *et al.* 1990)) for sequences originating from the common bacterial endosymbiont *Wolbachia* (Correa and Ballard 2016) and five contigs were removed, thereby getting rid of all *Wolbachia* contaminations. Since the raw PacBio reads were used for assembly to achieve highest contiguity of contig sequences, we subsequently used proovread v2.13.12 (Hackl *et al.* 2014) to correct the DBG2OLC contig sequences iteratively. In the first pass, we used the Platanus contig sequences described above, and in a second pass the additional Illumina reads (100x coverage; Supplementary Table S1, datasets 07-11). The Illumina data for error correction was sequenced from progeny of the same one egg clutch as the larvae for the PacBio sequences to allow for highest sequence conformity and thus correction confidence. Since hybrid assembly is a highly complex procedure and we did not want to miss any information, we screened Illumina-only and PacBio-only assemblies for additional sequence information lacking in the DBG2OLC contigs. The Platanus-derived contigs described above were compared to the DBG2OLC contigs by BLASTN searches with perc_identity = 80. The PacBio reads were assembled with Canu v1.0 (Berlin *et al.* 2015) using default settings and the output contigs (Supplementary Table S2) used for BLASTN searches as described for the Platanus contigs. All contigs from both approaches that did not match DBG2OLC contigs with at least 80% were then added to the DBG2OLC assembly. These sequences were then scaffolded using SSPACE v3.0 (Boetzer *et al.* 2011) with x = 0, n = 25 and mate-pair libraries with 3 and 5.5 kb insert size (S.D. 0.8). Scaffold gaps were addressed with an iterative gap closure process. First, we corrected PacBio raw reads with Illumina reads by proovread applying default settings, and then used them to close gaps applying PBJelly v15.2.20 (English *et al.* 2012) with default settings. Afterward, datasets 02-06 (Supplementary Table S1) were used as input to five iterative rounds of GapFiller v1.10 (Boetzer and Pirovano 2012) with default parameters and average insert size with insert size variation.

Estimation of the recombination rate

Recombination rate estimates (ρ) were derived from 20 field isolates from five European natural populations (Supplementary Table S1, datasets 31-50) by applying a *reversible jump Markov Chain Monte Carlo mechanism* (rjMCMC) implemented in the program LDhelmet v1.7 (Chan *et al.* 2012) individually for each scaffold. LDhelmet is a derivative of LDhat (Auton *et al.* 2012), especially modified to fit genomic characteristics that differ from hominids to *Drosophila* (for example higher SNP density). Since we anticipate similar patterns in *Chironomus*, we chose LDhelmet and mainly followed the parameter recommendations of the authors. The ultimate LDhelmet analysis with the *rjmcmc* command was run for each scaffold with a block penalty of 50.0 (as recommended; parameter of negligible influence on results (Smukowski Heil *et al.* 2015)) and a window size of 50 SNPs (as in the data preparation). We used a burn-in of 1,000,000 iterations and subsequently ran the Markov chain for 10,000,000 iterations (see Supplementary Methods S1 for details).

Using ancestral linkage disequilibrium-based methods for the estimation of recombination rates heavily profits from a genetic map provided at the stage of phasing the SNP data. Since there is no such a resource for *C. riparius*, a constant rate was used. Although this is the default of the phasing algorithm applied, it may introduce a bias into the estimation of ρ based on this data.

Genome annotation

The annotation of gene content along the genome was aided by construction of a reference transcriptome, which assembled from 19 cDNA sequence data sets (Supplementary Table S1, datasets 12-30). We used Illumina and 454 Roche sequence reads from embryos, larvae and adults (both sexes; treated and untreated; see Supplementary Table S1 and references therein for details) to reach a maximum of expressed genes in order to optimize gene annotation. First, data sets were pre-processed using fastqc (Andrews 2010) and BBDuk from the BBMap package v35.85 (Bushnell 2014). Thereby, sequence adapters were trimmed using k = 23, mink = 11, hdist = 1, tbo and tpe options. 3′ bases with phred quality below 20 were trimmed and reads with average phred quality below 20 discarded.

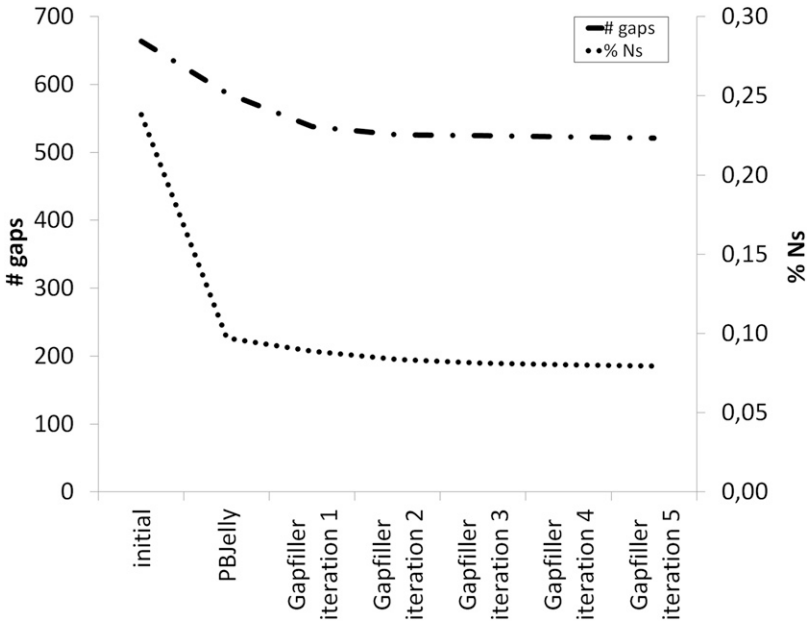

Figure 1 Effect of gap filling procedures. Gap filling with corrected PacBio reads (PBJelly) and Illumina paired end reads (Gapfiller). Shown is the decrease in number of gaps ("# gaps", dashed line) and fraction of undefined nucleotides ("% Ns", dotted line) in the scaffolds during the iterative gap filling process.

Assembly of the cleaned reads was then performed in two separate steps for Illumina and 454 data with Trinity v2.3.2 (Grabherr *et al.* 2011) using uneven k-mer sizes from 25 to 31. The best assembly was identified to be with k = 25 using assembly metrics like N50 and a search for core orthologous genes with BUSCO v1.2b (Simão *et al.* 2015) and used further on. The resulting assemblies for Illumina and 454 Roche data, respectively, were then merged and duplicate contigs removed using dedupe from the BBmap package with mid = 90. The resulting final transcriptome assembly was then used for gene annotation.

To ensure discovery of most repeat sequences in the draft genome, we extended the custom repeat library from (Oppold *et al.* 2017) with repeat sequences extracted manually from the draft genome presented in this study.

The whole annotation process was performed with the MAKER2 v2.31.8 (Cantarel *et al.* 2008; Holt and Yandell 2011) pipeline and affiliated programs. We used a three-round iterative process with the reference transcriptome and repeat library described above, three *C. riparius*-specific gene models and the SwissProt database as input (see Supplementary Methods S2 for details).

Assembly and annotation of the mitochondrial genome

The reconstruction of the mitochondrial genome sequence was performed using the program MITObim v1.8 (Hahn *et al.* 2013) on the large paired end dataset 03. MITObim applies a baiting and iterative mapping approach to short read data. Using a mitochondrial reference sequence (here an unpublished, partial Sanger sequence of *C. riparius*' mitochondrial genome), the program performs a mapping to gather all reads belonging to the mitochondrial genome and assembles them with MIRA v4.0.2 (Chevreux *et al.* 1999) in the first round. Then it uses the produced sequence to again fish for mitochondrial reads for further assembly. This is repeated until the number of mapped reads becomes stationary. MITObim was run four times with the reference

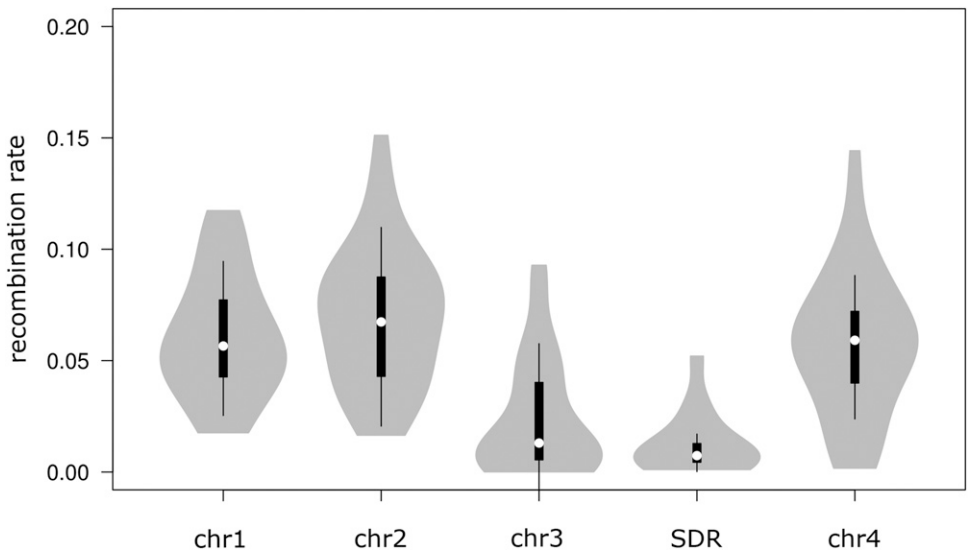

Figure 2 *C. riparius* chromosome-specific recombination rates. Recombination rates from all individuals across populations were pooled. Chromosome 3 is represented without the identified part of the sex determining region, which is displayed separately (SDR). White circles show medians, box limits indicate the 25th and 75th percentiles, whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, polygons represent density estimates of data and extend to extreme values. Kruskal-Wallis test (nonparametric for data without normal distribution) with Dunn's multiple comparison post-test (GraphPad Prism v5) revealed a significant difference ($P <$ 0.001) between chromosome 3 as well as the SDR relative to all other chromosomes. Chr = Chromosome.

Table 2 – *Comparative statistics of the nuclear genome's annotation.* Content of protein-coding genes in the genome of *C. riparius* compared to published genomes of other chironomids and *Drosophila melanogaster*

	gene count	average number of exons per gene	average exon length (bp)	protein coding part of the genome (%)
Chironomus riparius (this study)	13,449	5.1	378	14.6
Chironomus tentans (Kutsenko *et al.* 2014)	15,120	3.8	312	9
Polypedilum vanderplanki (Gusev *et al.* 2014)	17,137	4.3	324	20.2
Polypedilum nubifer (Gusev *et al.* 2014)	16,553	4.0	328	20.3
Belgica antarctica (Kim *et al.* 2017; Kelley *et al.* 2014)	11,005	5.0	321	19.6
Clunio marinus (Kaiser *et al.* 2016)	14,041	4.5	329	29.6
Parochlus steinenii (Kim *et al.* 2017)	13,468	6.2	215	13.0
Drosophila melanogaster (Adams *et al.* 2000)	13,907	5.5	538	18.3

sequences of *C. riparius* being modified in length to allow for different starting points of the procedure. All four output sequences were then aligned in MEGA v7.0.7 (Kumar *et al.* 2016) and manually integrated into a consensus sequence. This consensus sequence was annotated by MITOS WebServer (Bernt *et al.* 2013) using the genetic code *05 - invertebrate*. The whole sequence and all annotations were finally checked and, where necessary, corrected manually.

Data availability

Raw sequence reads are available through NCBI Sequence Read Archive, accession numbers are detailed in the Supplementary Table S1. Project number for the genome assembly is PRJEB27753 and for the mitochondrial genome assembly PRJEB27747. Supplemental material available at figshare: https://doi.org/10.25387/g3.11575059.

RESULTS AND DISCUSSION

Genome assembly

The 1,155,855 PacBio reads had an average length of 4,751 bp, and the longest read was 48,745 bp. The final hybrid assembly consisted of 752 scaffold sequences with a total length of 178,167,951 bp (Table 1). The total assembly length fits the published genome size of ~200 Mb estimated by flow cytometry (Schmidt-Ott *et al.* 2009), given that regions of low sequence complexity (*e.g.*, highly repetitive parts of centromeres and telomeres) are likely not to be resolved and thus missing. In light of the many tandem-repetitive element clusters interspersed in the genome of *C. riparius* (Schmidt 1984; Hankeln *et al.* 1994) it is therefore reasonable to assume that scaffold ends represent borders to internally repetitive heterochromatic regions in most cases. The N50 of 539,778 bp of the current genome draft is almost twice as high as for a previous version (Table 1). Gap filling drastically reduced the final unresolved base content (N's) of the assembly down to 0.08%, with the PacBio reads being especially helpful (Figure 1). On average 96.6% of the Illumina sequence reads could be mapped back to the draft genome (Supplementary Table S3), corroborating our assumption that only highly repetitive areas are underrepresented in the genome draft.

Recombination rate

Mean ρ values (always given per base pair within 50 kb windows) in *C. riparius* ranged between 0.04 and 0.07, thus lying within the range of those estimated for *Drosophila melanogaster* (0.01 to 0.11; (Chan *et al.* 2012)).

Recombination should be less frequent across sex-determining regions, because reciprocal exchange of chromosomal parts is only possible in the germ line of female individuals (Wright *et al.* 2016) (but see Rodrigues *et al.* 2018). *C. riparius* has a sex-determination system with heterogametic males, bearing a sex-determining region (SDR) on chromosome 3 that is being interpreted as an emerging sex chromosome (Kraemer and Schmidt 1993; Michailova *et al.* 2009). We identified the candidate SDR-containing scaffold by BLASTN searches with the sequence of the single copy gene *CpY* (NCBI accession number X82317.1) as query. Indeed, the SDR scaffold (# 549) showed a large region with lowered recombination rates around *CpY* (Supplementary Figure S1). When extracting the last 600,000 bp from this scaffold, we observed a mean ρ of 0.014 compared to overall estimates between 0.024 and 0.07 for the four chromosomes (Figure 2). This seems reasonable since recombination should be roughly halved in the SDR. Interestingly, the remaining parts of chromosome 3 without the identified part of the SDR also have relatively low recombination rates compared to all other chromosomes (Figure 2; Kruskal-Wallis test, $P < 0.001$), potentially due to further fragments of the SDR being present along this chromosome or impacts of the SDR on the genomic surroundings.

Annotation

13,449 protein-coding genes were annotated across the *C. riparius* nuclear genome (Table 2). There is a slight negative correlation between number of genes and number of exons per gene across chironomid genomes (Supplementary Figure S2). This may point toward multi-exonic genes being split up into several genes in draft genomes

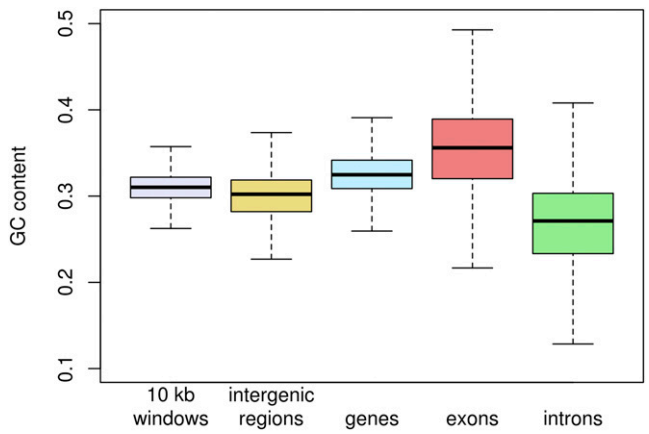

Figure 3 GC content for genomic features. Different genomic features revealed differences in GC content. GC content in exons resided above genome average, while the opposite was found for introns. 10 kb windows were generated without regard to their content. Box limits indicate the 25th and 75th percentiles, whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles.

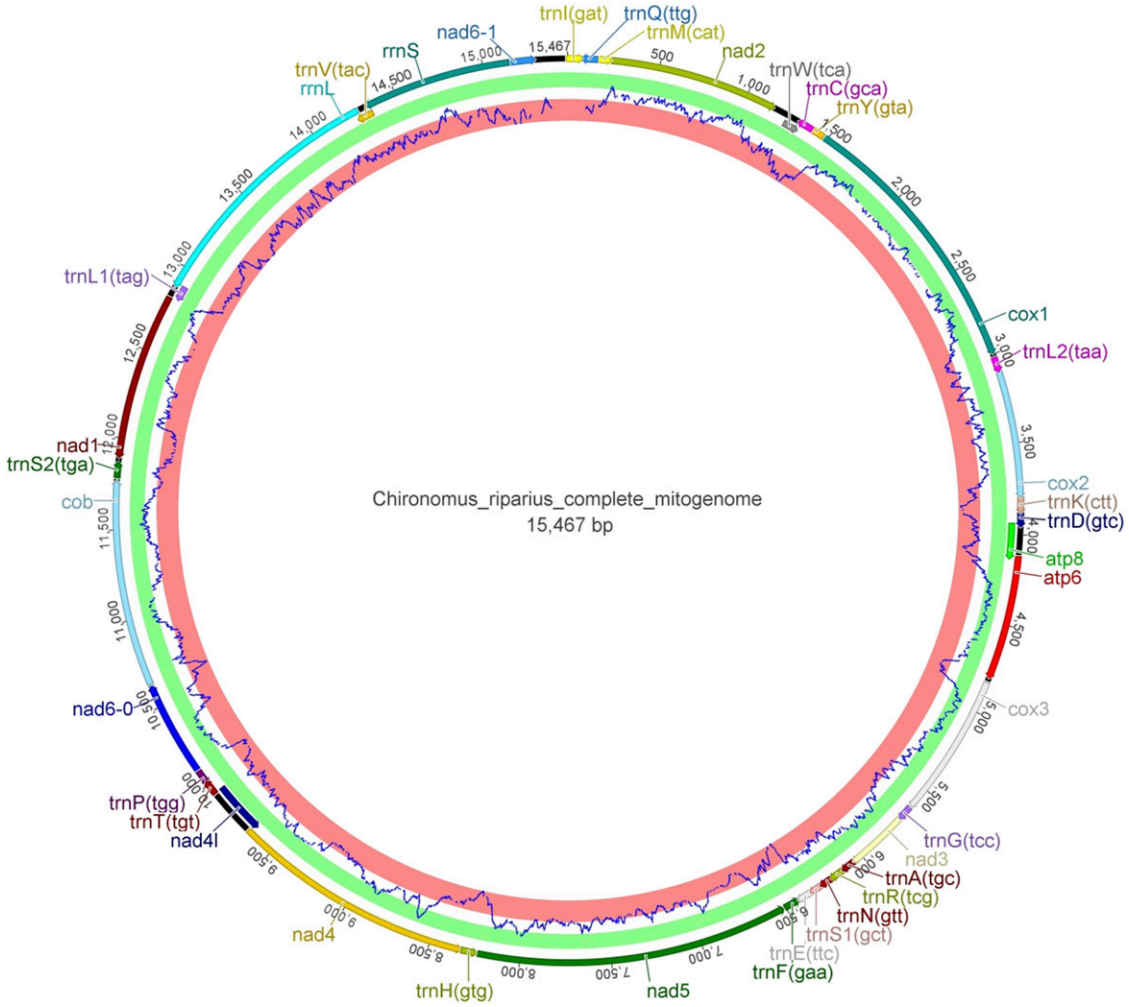

Figure 4 Mitochondrial genome of *C. riparius*. The circular genome consists of 15,467 bp. Prediction of protein-coding sequences by the EMBOSS tool tcode (blue graph at the inner edge of the genome; green ring = coding, red ring = non-coding) mainly is consistent with the annotation from MITOS.

with large numbers of annotated genes. The relatively high number of exons per gene (5.1) in a relatively low number of genes in the *C. riparius* genome annotation compared to other chironomid genomes therefore suggest a high quality due to the assembly's contiguity. Applying the algorithm BUSCO v1.2b (Simão *et al.* 2015), we found orthologous sequences to 93.7% of arthropod core genes (Table 1). Therefore, we can assume the genome to be almost complete also in terms of gene space.

The proportionate genome-wide GC content in *C. riparius* is 0.311, which is close to the average of 0.332 across chironomid genomes, including *C. tentans* (0.312), *Polypedilum vanderplanki* (0.28), *P. nubifer* (0.39), *Belgica antarctica* (0.39), *Clunio marinus* (0.317) and *Parochlus steinenii* (0.322). Average GC content in chironomids is thus at the lower end compared to other insect genomes (Samanta 2007). Across a broad phylogenetic range including plants, invertebrates and vertebrates, GC content has been shown to be higher in exons than introns and might have evolved as a determinant of exon selection (Amit *et al.* 2012). To assess this for the *C. riparius* genome, we inferred GC content for all non-overlapping 10 kb windows throughout the genome assembly (N = 17,566), all coherent regions without genes (N = 24,771), all protein-coding genes (N = 13,449), all exons (N = 68,943) and all introns (N = 54,860). GC content across the random 10 kb windows was on average 0.310 +/− 0.026 (mean +/− s.d.), perfectly mirroring the GC content of 0.311 for the whole genome assembly. Windows containing genes had a slightly higher GC content than genome average (0.327 +/− 0.032) and windows without genes had a slightly lower GC content than genome average (0.299 +/− 0.044). This difference was much more pronounced between exons (0.355 +/− 0.061) and introns (0.269 +/− 0.056), with exons being the feature with the highest and introns with the lowest GC content (Figure 3). The differences between categories were highly significant ($P < 0.0001$) for all pairwise comparisons applying Mann-Whitney tests with Bonferroni correction.

101,693 regions of up to 30 kb in length were annotated as repetitive sequences (9.14%). Compared to the Illumina-only genome draft (Oppold *et al.* 2017), the inclusion of long sequence reads has significantly increased detection of repeats by 41%. Given the heavy load of repetitive sequences in the *C. riparius* genome (Schaefer and Schmidt 1981), however, this value most likely still underestimates the true repeat content due to unresolved large heterochromatic regions.

The mitochondrial genome's length is 15,467 bp, which is in line with other dipteran values. All 37 genes of the mitochondrial genome could be annotated (Figure 4, Supplementary Table S4). Gene order follows the one conserved across Diptera (with the exception of

Culicidae having the *trnA* and *trnR* genes switched (Behura *et al.* 2011; Hao *et al.* 2017)), sharing complete synteny even with drosophilids (Montooth *et al.* 2009) from which they split an estimated 250-300 Myr ago (Cranston *et al.* 2012; Bolshakov *et al.* 2002).

ACKNOWLEDGMENTS

We gratefully acknowledge support in LDhelmet usage by Paul Jenkins and Mathias Weber and assistance in genome annotation by Florian Dolze. Parts of this research were conducted using the supercomputer Mogon and/or advisory services offered by Johannes Gutenberg University Mainz (hpc.uni-mainz.de), which is a member of the AHRP and the Gauss Alliance e.V.

LITERATURE CITED

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne *et al.*, 2000 The genome sequence of *Drosophila melanogaster*. Science 287: 2185–2195. https://doi.org/10.1126/science.287.5461.2185

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990 Basic local alignment search tool. J. Mol. Biol. 215: 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Amit, M., M. Donyo, D. Hollander, A. Goren, E. Kim *et al.*, 2012 Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. Cell Reports 1: 543–556. https://doi.org/10.1016/j.celrep.2012.03.013

Andrews, S., 2010 FastQC: a quality control tool for high throughput sequence data; https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Armitage, P. D., P. S. Cranston, and L. C. V. Pinder, 1995 *The Chironomidae: Biology and Ecology of Non-Biting Midges*, Springer-Verlag GmbH. Chapman and Hall, London. https://doi.org/10.1007/978-94-011-0715-0

Auton, A., A. Fledel-Alon, S. Pfeifer, O. Venn, L. Ségurel *et al.*, 2012 A fine-scale chimpanzee genetic map from population sequencing. Science 336: 193–198. https://doi.org/10.1126/science.1216872

Behura, S. K., N. F. Lobo, B. Haas, D. D. Lovin, M. F. Shumway *et al.*, 2011 Complete sequences of mitochondria genomes of *Aedes aegypti* and *Culex quinquefasciatus* and comparative analysis of mitochondrial DNA fragments inserted in the nuclear genomes. Insect Biochem. Mol. Biol. 41: 770–777. https://doi.org/10.1016/j.ibmb.2011.05.006

Berlin, K., S. Koren, C.-S. Chin, J. P. Drake, J. M. Landolin *et al.*, 2015 Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat. Biotechnol. 33: 623–630. Erratum: 1109. https://doi.org/10.1038/nbt.3238

Bernt, M., A. Donath, F. Jühling, F. Externbrink, C. Florentz *et al.*, 2013 MITOS: Improved de novo metazoan mitochondrial genome annotation. Mol. Phylogenet. Evol. 69: 313–319. https://doi.org/10.1016/j.ympev.2012.08.023

Boetzer, M., C. V. Henkel, H. J. Jansen, D. Butler, and W. Pirovano, 2011 Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27: 578–579. https://doi.org/10.1093/bioinformatics/btq683

Boetzer, M., and W. Pirovano, 2012 Toward almost closed genomes with GapFiller. Genome Biol. 13: R56. https://doi.org/10.1186/gb-2012-13-6-r56

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

Bolshakov, V. N., P. Topalis, C. Blass, E. Kokoza, A. della Torre *et al.*, 2002 A comparative genomic analysis of two distant diptera, the fruit fly, *Drosophila melanogaster*, and the malaria mosquito, *Anopheles gambiae*. Genome Res. 12: 57–66. https://doi.org/10.1101/gr.196101

Bushnell, B., 2014 BBMap: A Fast, Accurate, Splice-Aware Aligner. https://sourceforge.net/projects/bbmap/.

Cantarel, B. L., I. Korf, S. M. C. Robb, G. Parra, E. Ross *et al.*, 2008 MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18: 188–196. https://doi.org/10.1101/gr.6743907

Chan, A. H., P. A. Jenkins, and Y. S. Song, 2012 Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. PLoS Genet. 8: e1003090. https://doi.org/10.1371/journal.pgen.1003090

Chevreux, B., T. Wetter, and S. Suhai, 1999 Genome sequence assembly using trace signals and additional sequence information. *German conference on bioinformatics* 99:45–56.

Correa, C. C., and J. Ballard, 2016 *Wolbachia* associations with insects: winning or losing against a master manipulator. Front. Ecol. Evol. 3: 153. https://doi.org/10.3389/fevo.2015.00153

Cranston, P. S., N. B. Hardy, and G. E. Morse, 2012 A dated molecular phylogeny for the Chironomidae (Diptera). Syst. Entomol. 37: 172–188. https://doi.org/10.1111/j.1365-3113.2011.00603.x

English, A. C., S. Richards, Y. Han, M. Wang, V. Vee *et al.*, 2012 Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One 7: e47768. https://doi.org/10.1371/journal.pone.0047768

Foucault, Q., A. Wieser, A. M. Waldvogel, and M. Pfenninger, 2019 Establishing laboratory cultures and performing ecological and evolutionary experiments with the emerging model species *Chironomus riparius*. J. Appl. Entomol. 143: 584–592. https://doi.org/10.1111/jen.12606

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson *et al.*, 2011 Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29: 644–652. https://doi.org/10.1038/nbt.1883

Gusev, O., Y. Suetsugu, R. Cornette, T. Kawashima, M. D. Logacheva *et al.*, 2014 Comparative genome sequencing reveals genomic signature of extreme desiccation tolerance in the anhydrobiotic midge. Nat. Commun. 5: 4784. https://doi.org/10.1038/ncomms5784

Hackl, T., R. Hedrich, J. Schultz, and F. Förster, 2014 proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. Bioinformatics 30: 3004–3011. https://doi.org/10.1093/bioinformatics/btu392

Hahn, C., L. Bachmann, and B. Chevreux, 2013 Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads - a baiting and iterative mapping approach. Nucleic Acids Res. 41: e129. https://doi.org/10.1093/nar/gkt371

Hankeln, T., H. Friedl, I. Ebersberger, J. Martin, and E. R. Schmidt, 1997 A variable intron distribution in globin genes of *Chironomus*: evidence for recent intron gain. Gene 205: 151–160. https://doi.org/10.1016/S0378-1119(97)00518-0

Hankeln, T., A. Rohwedder, B. Weich, and E. R. Schmidt, 1994 Transposition of minisatellite-like DNA in *Chironomus* midges. Genome 37: 542–549. https://doi.org/10.1139/g94-077

Hankeln, T., and E. Schmidt, 1990 New foldback transposable element TFB1 found in histone genes of the midge *Chironomus thummi*. J. Mol. Biol. 215: 477–482. https://doi.org/10.1016/S0022-2836(05)80159-7

Hao, Y.-J., Y.-L. Zou, Y.-R. Ding, W.-Y. Xu, Z.-T. Yan *et al.*, 2017 Complete mitochondrial genomes of *Anopheles stephensi* and *An. dirus* and comparative evolutionary mitochondriomics of 50 mosquitoes. Sci. Rep. 7: 7666. https://doi.org/10.1038/s41598-017-07977-0

Smukowski Heil, C. S. S., C. Ellison, M. Dubin, and M. A. Noor, 2015 Recombining without hotspots: A comprehensive evolutionary portrait of recombination in two closely related species of *Drosophila*. Genome Biol. Evol. 7: 2829–2842. https://doi.org/10.1093/gbe/evv182

Holt, C., and M. Yandell, 2011 MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12: 491. https://doi.org/10.1186/1471-2105-12-491

Kaiser, T. S., B. Poehn, D. Szkiba, M. Preussner, F. J. Sedlazeck *et al.*, 2016 The genomic basis of circadian and circalunar timing adaptations in a midge. Nature 540: 69–73. https://doi.org/10.1038/nature20151

Kajitani, R., K. Toshimoto, H. Noguchi, A. Toyoda, Y. Ogura *et al.*, 2014 Efficient de novo assembly of highly heterozygous genomes from

whole-genome shotgun short reads. Genome Res. 24: 1384–1395. https://doi.org/10.1101/gr.170720.113

Kelley, J. L., J. T. Peyton, A.-S. Fiston-Lavier, N. M. Teets, M.-C. Yee *et al.*, 2014 Compact genome of the Antarctic midge is likely an adaptation to an extreme environment. Nat. Commun. 5: 4611. https://doi.org/10.1038/ncomms5611

Kim, S., M. Oh, W. Jung, J. Park, H.-G. Choi *et al.*, 2017 Genome sequencing of the winged midge, *Parochlus steinenii*, from the Antarctic Peninsula. Gigascience 6: giw009. https://doi.org/10.1093/gigascience/giw009

Klomp, J., D. Athy, C. W. Kwan, N. I. Bloch, T. Sandmann *et al.*, 2015 A cysteine-clamp gene drives embryo polarity in the midge *Chironomus*. Science 348: 1040–1042. https://doi.org/10.1126/science.aaa7105

Kraemer, C., and E. R. Schmidt, 1993 The sex determining region of *Chironomus thummi* is associated with highly repetitive DNA and transposable elements. Chromosoma 102: 553–562. https://doi.org/10.1007/BF00368348

Kumar, S., G. Stecher, and K. Tamura, 2016 MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33: 1870–1874. https://doi.org/10.1093/molbev/msw054

Kutsenko, A., T. Svensson, B. Nystedt, J. Lundeberg, P. Björk *et al.*, 2014 The *Chironomus tentans* genome sequence and the organization of the Balbiani ring genes. BMC Genomics 15: 819. https://doi.org/10.1186/1471-2164-15-819

Lee, S.-W., S.-M. Kim, and J. Choi, 2009 Genotoxicity and ecotoxicity assays using the freshwater crustacean *Daphnia magna* and the larva of the aquatic midge *Chironomus riparius* to screen the ecological risks of nanoparticle exposure. Environ. Toxicol. Pharmacol. 28: 86–91. https://doi.org/10.1016/j.etap.2009.03.001

Marinković, M., W. C. de Leeuw, M. de Jong, M. H. S. Kraak, W. Admiraal *et al.*, 2012 Combining Next-Generation Sequencing and Microarray Technology into a Transcriptomics Approach for the Non-Model Organism *Chironomus riparius*. PLoS One 7: e48096. https://doi.org/10.1371/journal.pone.0048096

Michailova, P., B. Krastanov, T. Hankeln, E. Schmidt, and C. Kraemer, 2009 In situ localization of the evolutionary conserved Cpy/Cty gene in the subfamily Chironominae (Chironomidae, Diptera): establishment of chromosomal homologies. J. Zoological Syst. Evol. Res. 47: 298–301. https://doi.org/10.1111/j.1439-0469.2008.00494.x

Montooth, K. L., D. N. Abt, J. W. Hofmann, and D. M. Rand, 2009 Comparative genomics of *Drosophila* mtDNA: novel features of conservation and change across functional domains and lineages. J. Mol. Evol. 69: 94–114. https://doi.org/10.1007/s00239-009-9255-0

Nair, P. M. G., S. Y. Park, and J. Choi, 2011 Analyses of Expressed Sequence Tags from *Chironomus riparius* Using Pyrosequencing: Molecular Ecotoxicology Perspective. Environ. Health Toxicol. 26: e2011010. https://doi.org/10.5620/eht.2011.26.e2011010

Oppold, A.-M., and M. Pfenninger, 2017 Direct estimation of the spontaneous mutation rate by short-term mutation accumulation lines in *Chironomus riparius*. Evolution Letters 1: 86–92. https://doi.org/10.1002/evl3.8

Oppold, A.-M., H. Schmidt, M. Rose, S. L. Hellmann, F. Dolze *et al.*, 2017 *Chironomus riparius* (Diptera) genome sequencing reveals the impact of minisatellite transposable elements on population divergence. Mol. Ecol. 26: 3256–3275. https://doi.org/10.1111/mec.14111

Rodrigues, N., T. Studer, C. Dufresnes, and N. Perrin, 2018 Sex-Chromosome Recombination in Common Frogs Brings Water to the Fountain-of-Youth. Mol. Biol. Evol. 35: 942–948. https://doi.org/10.1093/molbev/msy008

Rosenberg, D. M., 1992 Freshwater biomonitoring and Chironomidae. Neth. J. Aquat. Ecol. 26: 101–122. https://doi.org/10.1007/BF02255231

Samanta, M. P., 2007 Nucleotide distribution patterns in insect genomes. *Systemix Reports* 1:arXiv:q-bio/0702036.

Schaefer, J., and E. R. Schmidt, 1981 Different repetition frequencies of a 120 base-pair DNA-element and its arrangement in *Chironomus thummi thummi* and *Chironomus thummi piger*. Chromosoma 84: 61–66. https://doi.org/10.1007/BF00293363

Schmidt-Ott, U., A. M. Rafiqi, K. Sander, and J. S. Johnston, 2009 Extremely small genomes in two unrelated dipteran insects with shared early developmental traits. Dev. Genes Evol. 219: 207–210. https://doi.org/10.1007/s00427-009-0281-0

Schmidt, E. R., 1984 Clustered and interspersed repetitive DNA sequence family of *Chironomus*: The nucleotide sequence of the Cla-elements and of various flanking sequences. J. Mol. Biol. 178: 1–15. https://doi.org/10.1016/0022-2836(84)90227-4

Schmidt, H., B. Greshake, B. Feldmeyer, T. Hankeln, and M. Pfenninger, 2013 Genomic basis of ecological niche divergence among cryptic sister species of non-biting midges. BMC Genomics 14: 384. https://doi.org/10.1186/1471-2164-14-384

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, 2015 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Waldvogel, A.-M., A. Wieser, T. Schell, S. Patel, H. Schmidt *et al.*, 2018 The genomic footprint of climate adaptation in *Chironomus riparius*. Mol. Ecol. 27: 1439–1456. https://doi.org/10.1111/mec.14543

Williams, K. A., D. W. J. Green, D. Pascoe, and D. E. Gower, 1986 The acute toxicity of cadmium to different larval stages of *Chironomus riparius* (Diptera: Chironomidae) and its ecological significance for pollution regulation. Oecologia 70: 362–366. https://doi.org/10.1007/BF00379498

Wright, A. E., R. Dean, F. Zimmer, and J. E. Mank, 2016 How to make a sex chromosome. Nat. Commun. 7: 12087. https://doi.org/10.1038/ncomms12087

Ye, C., C. M. Hill, S. Wu, J. Ruan, and Z. S. Ma, 2016 DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. Sci. Rep. 6: 31900. https://doi.org/10.1038/srep31900

Communicating editor: S. Celniker