**Supplementary Material**

**“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, Markus Pfenninger

# Supplementary Methods S1: Estimation of the recombination rate

Short read sequence data from twenty unrelated female individuals (datasets 27-46 from Supplementary Table S1, published in (Waldvogel et al. 2018)) from five populations across Europe (MF = Rhône-Alpes, MG = Hessen, NMF = Lorraine, SI = Piemonte, SS = Andalucia) was used to generate a set of single nucleotide polymorphisms (SNP) for the analyses. Sampling locations and quality processing of the reads are described in (Oppold et al. 2017).

Reads were aligned to the genome described in this study using bwa mem algorithm from BWA v0.7.13-r1126 (Li and Durbin 2009; Li 2013) with standard parameters. PCR duplicate reads were marked with the MarkDuplicatesWithMateCigar algorithm from Picard tools v2.6.0 (<http://broadinstitute.github.io/picard/>). Afterwards, realignment around indels was performed with the IndelRealigner algorithm from the Genome Analysis Toolkit v3.6-0-g89b7209 (GATK; (McKenna et al. 2010)), followed by recalibration of bases. For the latter we applied the UnifiedGenotyper algorithm from GATK to identify an initial SNP set, searched for shared SNPs in all resulting VCF files with the isec algorithm from Bcftools v1.3.1 (Li et al. 2009) and finally recalibrated bases with the BaseRecalibrator algorithm from GATK.

The final SNPs were called on the recalibrated BAM files again using the *UnifiedGenotyper* algorithm from GATK with *stand\_call\_conf*=40 and *stand\_emit\_conf*=15. Phasing of haplotypes was performed scaffold-wise with SHAPEIT2 v2.r837 (Delaneau et al. 2013) using only diallelic SNPs. We converted the output from SHAPEIT2 to VCF files and extracted input files for LDhelmet with VCFtools (<http://vcftools.sourceforge.net>) and the *‑‑ldhelmet* option.

Recombination rates were estimated 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.

Using the information from the SNPS and POS files from the VCFtools’ output we generated FASTA files for every population and every scaffold individually, ending up with 3,220 input files (5 populations \* 644 scaffolds), containing 25,760 haplotype blocks (3,220 files \* 4 individuals \* 2 haplotypes), representing 99 % of the genome assembly. Data preparation for the main run in LDhelmet included creation of haplotype configuration files with the *find\_confs* command on a window size of 50 SNPs as recommended by the authors, as well as likelihood lookup tables and Padé coefficients. For these calculations we used population-specific mutation rates *theta* (θ = 4*Neμ*) that were individually derived from Pool-Seq data of the same five populations (Waldvogel et al. 2018) and calculated with PoPoolation v1.2.2 (Kofler et al. 2011). *Theta* values are: θ*MF*= 0.01236, θ*MG*= 0.01134, θ*NMF* = 0.01446, θ*SI* = 0.01029 and θ*SS* = 0.01021. Default values were applied to all other parameters.

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 (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 (in addition to the burn-in). Results were extracted from the binary output via the *post\_to\_text* command of LDhelmet.

To allow for a comparison of recombination rates between chromosomes, a number of scaffolds was correlated to their physical chromosomes by fluorescence-in-situ-hybridization (data not shown) applying protocols described previously (Hankeln and Schmidt 1987; Schmidt et al. 1988).

# Supplementary Methods S2: Genome annotation

The draft genome, the reference transcriptome described in the main text and the GFF file from a BUSCO run were uploaded to the Augustus v3.2.3 training (Stanke et al. 2008) at the University of Greifswald webserver (<http://bioinf.uni-greifswald.de/webaugustus/training/create>; accessed on 2017-02-24). This output then served as input for the first round of MAKER2. MAKER2 can work much more accurate when provided with genome-specific gene models at the beginning. Therefore we ran CEGMA v2.5 (Parra et al. 2007) on the draft genome and converted the output to a hidden Markov model using scripts from the SNAP gene finder v2006-07-28 (Korf 2004). Additionally, we created another hidden Markov model by running GeneMark v4.32 (Lomsadze et al. 2014) with min\_contig set to 20,000 on the draft genome.

The first round of MAKER2 annotation was then run using MPICH2 v3.2 (https://www.mpich.org/) parallelization with the described transcriptome, SNAP, GeneMark and Augustus models, our custom repeat library plus the SwissProt database (as at 13.1.2016) as input. The MAKER2 pipeline was run with the programs Augustus v3.2.1, BLAST v2.2.28+ (Altschul et al. 1990), Repeatmasker v4.0.6 (Smit et al. 2013-2015), SNAP v2006-07-28 (Korf 2004), GeneMark v4.32 (Lomsadze et al. 2014) and Exonerate v2.2.0 (Slater and Birney 2005). We applied default parameters with only max\_dna\_len set to 500,000 to prevent loss of gene parts from genes with larger introns, min\_protein set to 10 to receive as much potential protein sequences as possible and fix\_nucleotides set as flag to allow for non-ACGT-characters in the genome file. The gff files of MAKER2’s output were merged to a single file using gff3\_merge from the MAKER2 distribution. Afterwards this gff file was converted to a hidden Markov model using SNAP scripts as described above for CEGMA with only the cegma2zff script being replaced by maker2zff. The information from the first MAKER2 run was also used for retraining the gene model in Augustus. The genome.ann file was first retransformed to gff and modified to match the gff format and then fed into the autoAug.pl script from Augustus v3.2.1 for retraining the EST-based gene models. The second round of the MAKER2 pipeline was then started with the same settings and input files as described above but with the updated Augustus and SNAP gene models and the parameters min\_protein set to 30 and alt\_splice on. Afterwards, a third round of the MAKER2 pipeline was run exactly as the second one, including another re-training of the Augustus gene model with autoAug.pl and again updating the SNAP gene model. From the resulting output of MAKER2’s third round we merged all gff files with gff3\_merge as described above and renamed the included gene tracks to ensure easier handling. To allow for assigning putative gene functions to the annotated gene tracks, we gathered all predicted proteins from the MAKER2 output folder, performed BLASTP searches against the SwissProt database (as at 13.1.2016) and then added the best BLAST hits to the accordant gene tracks.

To evaluate the reliability of the annotation produced by MAKER2, we used another annotation pipeline, BRAKER v2.1.0 (Hoff et al. 2015), to fully annotate *C. riparius’* genome sequence again. The completeness of the gene space was then checked by BUSCO searches on the coding genes’ protein sequences of both annotations. These were compared to a BUSCO search on the genome sequence itself. Most notably, the annotations of both pipelines contain almost the same number of genes that are complete and single copy (MAKER2 83.6%, BRAKER 83.9%; Supplementary Figure S3). One problem in MAKER2’s results was the formation of chimeric genes that occurred by concatenating neighboring genes (data not shown). On the other hand, BRAKER seemed to be quite aggressive in annotating, since it had fewer missing genes (MAKER2 9.2%, BRAKER 3.7%), but more duplicated ones (MAKER2 1.5%, BRAKER 9.1%), with the latter also being related to BRAKER’s potentially useful tendency to call several transcript variants of a gene. BRAKER’s aggressiveness is also demonstrated by the total number of genes found by the pipeline – with 18,690, the gene count is 40% higher than MAKER2’s result on *C. riparius’* genome, and also higher than in any other chironomid genome or the *Drosophila melanogaster* reference genome (Table 2). Overall, results between the two pipelines were similar enough to not question their reliability, but the differences present were too ambiguous to make a clear judgement.

# Supplementary Tables

Supplementary Table S1 – Sequence data used in the study

| **dataset****ID** | **used in part** | **sequencing technology** | **read length (bp)** | **number of raw reads** | **accession number** | **publication** |
| --- | --- | --- | --- | --- | --- | --- |
| 01 | genome assembly | PacBio RS II | up to 48,745 | 1,155,855 | ERR2696325 | **this study** |
| 02 | genome assembly | Illumina paired end | 100 | 29,136,088 | SAMEA4560833 | (Oppold et al. 2017) |
| 03 | genome assembly | Illumina paired end | 100 | 167,264,372 | SAMEA4560834 | (Oppold et al. 2017) |
| 04 | genome assembly | Illumina paired end (MiSeq) | 300 | 58,447,092 | SAMEA4560835 | (Oppold et al. 2017) |
| 05 | genome assembly + scaffolding | Illumina mate-pair | 100(3 kb insert size) | 48,853,530 | SAMEA4560836 | (Oppold et al. 2017) |
| 06 | genome assembly + scaffolding | Illumina mate-pair | 100(6 kb insert size) | 47,978,148 | SAMEA4560837 | (Oppold et al. 2017) |
| 07 | error correction | Illumina paired end | 150 | 29,694,824 | PRJEB18039 | (Oppold and Pfenninger 2017) |
| 08 | error correction | Illumina paired end | 150 | 34,118,424 | PRJEB18039 | (Oppold and Pfenninger 2017) |
| 09 | error correction | Illumina paired end | 150 | 35,803,108 | PRJEB18039 | (Oppold and Pfenninger 2017) |
| 10 | error correction | Illumina paired end | 150 | 38,534,360 | PRJEB18039 | (Oppold and Pfenninger 2017) |
| 11 | error correction | Illumina paired end | 150 | 39,512,980 | PRJEB18039 | (Oppold and Pfenninger 2017) |
| 12 | annotation (transcriptome) | 454 Roche | up to 679 | 430,916 | SRR834592 | (Schmidt et al. 2013) |
| 13 | annotation (transcriptome) | 454 Roche | up to 914 | 456,584 | SRR834593 | (Schmidt et al. 2013) |
| 14 | annotation (transcriptome) | 454 Roche | up to 690  | 1,549,146 | SRR496839 | (Marinković et al. 2012) |
| 15 | annotation (transcriptome) | 454 Roche | up to 606 | 4,041 | SRX022389 | (Nair et al. 2011) |
| 16 | annotation (transcriptome) | 454 Roche | up to 679 | 138,103 | SRR1049908 |  |
| 17 | annotation (transcriptome) | 454 Roche | up to 802 | 266,129 | SRR1049909 |  |
| 18 | annotation (transcriptome) | 454 Roche | up to 600 | 189,271 | SRR1049910 |  |
| 19 | annotation (transcriptome) | 454 Roche | up to 662 | 235,202 | SRR1049911 |  |
| 20 | annotation (transcriptome) | Illumina paired end | 100 | 6,863,480 | SRR1028867 | (Klomp et al. 2015) |
| 21 | annotation (transcriptome) | Illumina paired end | 100 | 10,222,374 | SRR1032319 | (Klomp et al. 2015) |
| 22 | annotation (transcriptome) | Illumina paired end | 100 | 9,684,342 | SRR1032320 | (Klomp et al. 2015) |
| 23 | annotation (transcriptome) | Illumina paired end | 100 | 9,305,342 | SRR1032321 | (Klomp et al. 2015) |
| 24 | annotation (transcriptome) | Illumina paired end | 100 | 10,861,756 | SRR1032322 | (Klomp et al. 2015) |
| 25 | annotation (transcriptome) | Illumina paired end | 100 | 10,783,776 | SRR1032323 | (Klomp et al. 2015) |
| 26 | annotation (transcriptome) | Illumina paired end |  |  | unpublished | 1KITE project |
| 27 | annotation (transcriptome) | Illumina paired end | 100 | 105,429,480 | ERS1472439 | (Oppold et al. 2017) |
| 28 | annotation (transcriptome) | Illumina paired end | 100 | 197,899,590 | ERS1472440 | (Oppold et al. 2017) |
| 29 | annotation (transcriptome) | Illumina paired end | 100 | 34,987,537 | ERS1472441 | (Oppold et al. 2017) |
| 30 | annotation (transcriptome) | Illumina paired end | 100 | 46,557,991 | ERS1472442 | (Oppold et al. 2017) |
| 31MF1 | recombination rate  | Illumina paired end | 150 | 36,869,618 | ERR2528543 | (Waldvogel et al. 2018) |
| 32MF2 | recombination rate  | Illumina paired end | 150 | 30,387,616 | ERR2528544 | (Waldvogel et al. 2018) |
| 33MF3 | recombination rate  | Illumina paired end | 150 | 32,170,704 | ERR2528545 | (Waldvogel et al. 2018) |
| 34MF4 | recombination rate  | Illumina paired end | 150 | 32,312,656 | ERR2528546 | (Waldvogel et al. 2018) |
| 35MG2 | recombination rate  | Illumina paired end | 150 | 29,057,938 | ERR2528547 | (Waldvogel et al. 2018) |
| 36MG3 | recombination rate  | Illumina paired end | 150 | 28,788,518 | ERR2528548 | (Waldvogel et al. 2018) |
| 37MG4 | recombination rate  | Illumina paired end | 150 | 27,094,554 | ERR2528549 | (Waldvogel et al. 2018) |
| 38MG5 | recombination rate  | Illumina paired end | 150 | 34,100,258 | ERR2528550 | (Waldvogel et al. 2018) |
| 39NMF1 | recombination rate  | Illumina paired end | 150 | 38,469,060 | ERR2528551 | (Waldvogel et al. 2018) |
| 40NMF2 | recombination rate  | Illumina paired end | 150 | 32,269,454 | ERR2528552 | (Waldvogel et al. 2018) |
| 41NMF3 | recombination rate  | Illumina paired end | 150 | 29,389,106 | ERR2528553 | (Waldvogel et al. 2018) |
| 42NMF4 | recombination rate  | Illumina paired end | 150 | 26,446,596 | ERR2528554 | (Waldvogel et al. 2018) |
| 43SI1 | recombination rate  | Illumina paired end | 150 | 37,946,848 | ERR2528555 | (Waldvogel et al. 2018) |
| 44SI2 | recombination rate  | Illumina paired end | 150 | 42,327,576 | ERR2528556 | (Waldvogel et al. 2018) |
| 45SI3 | recombination rate  | Illumina paired end | 150 | 32,922,356 | ERR2528557 | (Waldvogel et al. 2018) |
| 46SI4 | recombination rate  | Illumina paired end | 150 | 30,872,238 | ERR2528558 | (Waldvogel et al. 2018) |
| 47SS1 | recombination rate  | Illumina paired end | 150 | 31,078,600 | ERR2528559 | (Waldvogel et al. 2018) |
| 48SS2 | recombination rate  | Illumina paired end | 150 | 39,785,504 | ERR2528560 | (Waldvogel et al. 2018) |
| 49SS3 | recombination rate  | Illumina paired end | 150 | 36,018,158 | ERR2528561 | (Waldvogel et al. 2018) |
| 50SS4 | recombination rate  | Illumina paired end | 150 | 28,776,202 | ERR2528562 | (Waldvogel et al. 2018) |

Supplementary Table S2 – Statistics for the PacBio-only assembly with Canu

|  |  |
| --- | --- |
| number of sequences | 8,488 |
| total sequence length (bp) | 229,089,447 |
| average sequence length (bp) | 26,990 |
| longest sequence (bp) | 1,085,725 |
| N50 | 56,198 |

Supplementary Table S3 – Back-mapping of reads used to obtain the assembly onto the draft genome

Only intact pairs after all quality processing steps were mapped.

|  |  |  |
| --- | --- | --- |
| dataset ID | mapped read pairs | % mapped read pairs |
| 02 | 6,363,845 | 96.9 |
| 03 | 42,770,555 | 98.1 |
| 04 | 25,635,946 | 99.5 |
| 05 | 13,462,849 | 93.3 |
| 06 | 18,452,418 | 92.2 |

Supplementary Table S4 – Annotation of the mitochondrial genome

The annotation was generated by MITOS and manually curated afterwards.

|  |  |  |  |
| --- | --- | --- | --- |
| **Gene name** | **Start** | **Stop** | **Strand** |
| trnI(gat) | 1 | 68 | + |
| trnQ(ttg) | 82 | 150 | - |
| trnM(cat) | 159 | 227 | + |
| nad2 | 249 | 1154 | + |
| trnW(tca) | 1256 | 1323 | + |
| trnC(gca) | 1323 | 1391 | - |
| trnY(gta) | 1398 | 1463 | - |
| cox1 | 1474 | 2988 | + |
| trnL2(taa) | 3028 | 3093 | + |
| cox2 | 3105 | 3785 | + |
| trnK(ctt) | 3802 | 3873 | + |
| trnD(gtc) | 3880 | 3948 | + |
| atp8 | 3949 | 4125 | + |
| atp6 | 4125 | 4790 | + |
| cox3 | 4837 | 5616 | + |
| trnG(tcc) | 5639 | 5704 | + |
| nad3 | 5705 | 6049 | + |
| trnA(tgc) | 6060 | 6125 | + |
| trnR(tcg) | 6138 | 6204 | + |
| trnN(gtt) | 6205 | 6272 | + |
| trnS1(gct) | 6273 | 6339 | + |
| trnE(ttc) | 6347 | 6415 | + |
| trnF(gaa) | 6432 | 6498 | - |
| nad5 | 6522 | 8222 | - |
| trnH(gtg) | 8247 | 8313 | - |
| nad4 | 8335 | 9657 | - |
| nad4l | 9654 | 9935 | - |
| trnT(tgt) | 9949 | 10014 | + |
| trnP(tgg) | 10015 | 10081 | - |
| nad6-0 | 10104 | 10613 | + |
| cob | 10642 | 11754 | + |
| trnS2(tga) | 11793 | 11860 | + |
| nad1 | 11902 | 12801 | - |
| trnL1(tag) | 12826 | 12892 | - |
| rrnL | 12888 | 14267 | - |
| trnV(tac) | 14266 | 14337 | - |
| rrnS | 14337 | 15138 | - |
| nad6-1 | 15140 | 15280 | + |

# Supplementary Figures



Supplementary Figure S1 – Recombination rates in the scaffold containing the sex-determining region

Scaffold 549 of the draft genome contains a part of the putative sex-determining region of *C. riparius*. The grey box to the right marks the position of gene CpY, a gene associated with sex-determination in chironomids. The SDR-specific calculations were done on the last 600,000 bp of the scaffold. Population codes are MF = Rhône-Alpes, MG = Hessen, NMF = Lorraine, SI = Piemonte, SS = Andalucia.



Supplementary Figure S2 – Correlation between number of genes and exons



Supplementary Figure S3 – Comparison of the annotations with MAKER and BRAKER

“Genome” refers to a BUSCO search on the actual genome sequence, whereas the two other runs were performed on the protein-coding genes’ sequences output by the two annotation pipelines. C = complete, S = single-copy, D = duplicated, F = fragmented, M = missing.

**References**

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, 1990 Basic local alignment search tool. *Journal of Molecular Biology* 215 (3):403-410.

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 (6078):193-198.

Chan, A.H., P.A. Jenkins, and Y.S. Song, 2012 Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. *PLoS Genetics* 8 (12):e1003090.

Delaneau, O., J.-F. Zagury, and J. Marchini, 2013 Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods* 10 (1):5-6.

Hankeln, T., and E. Schmidt, 1987 Cotransposition of a highly repetitive DNA element with flanking sequences in the genome of the midge *Chironomus thummi*. *Journal of Molecular Evolution* 26 (4):311-319.

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 Biology and Evolution* 7 (10):2829-2842.

Hoff, K.J., S. Lange, A. Lomsadze, M. Borodovsky, and M. Stanke, 2015 BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*:btv661.

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 (6238):1040-1042.

Kofler, R., P. Orozco-terWengel, N. De Maio, R.V. Pandey, V. Nolte *et al.*, 2011 PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS One* 6 (1):e15925.

Korf, I., 2004 Gene finding in novel genomes. *BMC Bioinformatics* 5 (1):1.

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

Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25 (14):1754-1760.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The sequence alignment/map format and SAMtools. *Bioinformatics* 25 (16):2078-2079.

Lomsadze, A., P.D. Burns, and M. Borodovsky, 2014 Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research* 42 (15):e119-e119.

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 (10):e48096.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20:1297-1303.

Nair, P.M.G., S.Y. Park, and J. Choi, 2011 Analyses of Expressed Sequence Tags from *Chironomus riparius* Using Pyrosequencing : Molecular Ecotoxicology Perspective. *Environmental Health and Toxicology* 26.

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 (2):86-92.

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. *Molecular Ecology* 26 (12):3256-3275.

Parra, G., K. Bradnam, and I. Korf, 2007 CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23 (9):1061-1067.

Schmidt, E.R., H.G. Keyl, and T. Hankeln, 1988 In situ localization of two haemoglobin gene clusters in the chromosomes of 13 species of *Chironomus*. *Chromosoma* 96 (5):353-359.

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 (1):384.

Slater, G.S.C., and E. Birney, 2005 Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6 (1):31.

Smit, A., R. Hubley, and P. Green, 2013-2015 RepeatMasker Open-4.0.

Stanke, M., M. Diekhans, R. Baertsch, and D. Haussler, 2008 Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* 24 (5):637-644.

Waldvogel, A.-M., A. Wieser, T. Schell, S. Patel, H. Schmidt *et al.*, 2018 The genomic footprint of climate adaptation in *Chironomus riparius*. *Molecular Ecology* 27:1439-1456.