Supplemental Methods for:

Causes and consequences of a complex recombinational landscape in the ant *Cardiocondyla obscurior*

Mohammed Errbii[1], Jürgen Gadau[1], Kerstin Becker[2], Lukas Schrader[1*], Jan Oettler[3*]

[1]Institute for Evolution and Biodiversity, University of Münster, 48149, Münster, Germany

[2]Cologne Center for Genomics (CCG), Medical Faculty, University of Cologne, 50931 Cologne, Germany

[3]Lehrstuhl für Zoologie/Evolutionsbiologie, University Regensburg, 93053 Regensburg, Germany

*Correspondence: Lukas Schrader & Jan Oettler

Email: lukas.schrader@uni-muenster.de and joettler@gmail.com

Gene and (tandem) repeat annotation

For gene annotation, liftoff (v.1.6.3) (Shumate and Salzberg 2021) was used to transfer the Cobs2.1 annotation (Errbii et al. 2021) to Cobs3.1 with default parameters and the *-polish* option. For repeat annotation, we used RepeatMasker (v.4.0.7) (http://www.repeatmasker.org) and a previous TE library (Errbii et al. 2021). TE islands were defined as previously described (Errbii et al. 2021) by calculating the TE content in 250 kb nonoverlapping windows and grouping consecutive windows with more than 50% TE content. The boundaries of the TE islands were then manually curated and refined.

The Tandem Repeats Finder tool (v.4.09.1) (Benson 1999) was used with recommended parameters (2 5 7 80 10 50 2000) to identify tandem repeats (TRs) in the genome. The annotation was used with RepeatMasker (options: -s -norna -gff -u -pa 20 -a -nolow) to mask Cobs3.1 and TR content was calculated as the proportion of masked bases per 250 kb nonoverlapping windows. Additionally, TRs from five ant species (Melters et al. 2013; Huang et al. 2016) were used with RepeatMasker to explore their enrichment in the genome of *C. obscurior*.

Calculation of absolute divergence

To explore genetic divergence between lineages of *C. obscurior*, the average number of nucleotide substitutions between DNA sequences (d_{XY}) was estimated. Absolute divergence was also calculated between grandparental haplotypes (gp-d_{XY}).

Published paired-end short reads data from single workers of the Old World (Taiwan and Leiden) and New World (Itabuna and Una) lineages of *C. obscurior* (Errbii et al. 2021) were filtered and mapped to Cobs3.1 as described above. Note that reads from both lineages can be aligned to the reference genome with high confidence (Supplemental Fig. S27). SNP calling and initial hard filtering were performed using the HaplotypeCaller and VariantFiltration tools within GATK (v.4.1.2.0), with all positions including invariant sites retained (set1). For the grandparents (set2), we combined the high-quality markers produced above

with invariant sites obtained using parental sequencing. The two sets were then merged to generate a single VCF file with variant and invariant sites. Using VCFtools, we retained positions that were genotyped in >80% of the samples, had a maximum of two alleles, had a genotype quality of at least 20 and read coverage of at least 5. This resulted in a total of 154,202,917 successfully genotyped positions. Absolute divergence was then calculated in 250 kb windows using scripts available at https://github.com/simonhmartin/genomics_general.

Structural variant identification

To identify SVs, we adopted an approach by Mérot et al. (Mérot et al. 2023) using published paired-end short read data from 16 workers from five populations (Errbii et al. 2021). Reads were filtered and mapped to Cobs3.1, and read duplicates were removed using Picard's MarkDuplicates. SVs were identified using three different tools: Manta (v.1.6.0) (Chen et al. 2016), smoove (v.0.2.8) (https://github.com/brentp/smoove) which is based on LUMPY (Layer et al. 2014), and DELLY (v.0.8.1) (Rausch et al. 2012), all run with default parameters.

We excluded SVs that were shorter than 50 bp or longer than 100 kb, and those overlapping a gap of more than 10 Ns. After excluding all SVs flagged as "LowQual", 1,130, 3,060, and 9,343 SVs were identified by Manta, smoove, and DELLY, respectively. We then used Jasmine (Kirsche et al. 2023) to merge the three sets (options: *--min_dist=50 --mutual_distance --ignore_strand*). Using BCFtools, we retained SVs that were identified by at least two different tools, and with VCFtools we excluded SVs with a MAF <0.05 or genotyped in less than 80% of the individuals. This resulted in a total of 1,410 SVs, including 660 deletions, 174 duplications, and 576 inversions. RepeatMasker and the TE library (Errbii et al. 2021) were used to annotate the SVs and remove those with more than 80% repeat content. SVs were then assigned the recombination rate of the corresponding 250 kb window. When a SV matched more than one window

(two at most), the mean recombination rate of those windows was assigned. A total of 771 SVs were retained, of which 481 SVs were located outside TE islands.

To identify genomic inversions potentially affecting recombination in the F1 queen, we followed the same approach described above. Using the three tools, we identified inversions in the two grandparents and the two F1 parents, resulting in 425 and 118 candidates detected by DELLY and smoove, respectively. Manta did not detect any inversions in this data set. We then filtered these candidate inversions to only retain those that were heterozygous in the F1 queen. Furthermore, we required these inversions to be homozygous for different alleles in the grandparents or heterozygous in the grandmother (i.e., presenting no Mendelian errors).

For indels, we used the raw variant calls generated by GATK's HaplotypeCaller for the 16 workers. Indels were extracted and subjected to GATK's recommended hard filtering criteria (QD < 2.0; QUAL < 30.0; FS > 200 and ReadPosRankSum < -20.0). We further refined the dataset using VCFtools, retaining only biallelic indels with a MAF >0.05 and genotyped across >80% of the samples. The resulting set of 95,841 indel variants was then used to calculate diversity in nonoverlapping 250 kb windows with VCFtools. We also separated intergenic indels from those affecting genes using BEDTools and calculated diversity for both sets.

Estimating substitution rates

To explore the correlation between recombination and the efficiency of selection, we calculated omega (ω) for single-copy ortholog genes between *C. obscurior* and three ant species (*Monomorium pharaonis* (GCA013373865v2), *Solenopsis invicta* (UNIL_Sinv_3.0) and *Ooceraea biroi* (Obir_v5.4)) and the parasitoid wasp *Nasonia vitripennis* (Nvit_psr_1.1). Genomic and protein sequences, and gene annotations, were downloaded from the Ensembl database (release-54).

We used SeqKit (v.2.3.0) (Shen et al. 2016) to extract the longest protein isoform and corresponding CDS for each gene and retrieved all single-copy orthologs using OrthoFinder (v.2.5.4) (Emms and Kelly 2019). Using ClustalO (v.1.2.4) (Sievers et al. 2011), amino acid alignments for each of the 6,562 single-copy ortholog groups were generated and converted into nucleic acid alignments using Pal2Nal (v.14) (140 genes failed) (Suyama et al. 2006). After removing poorly aligned positions from each of the alignments using Gblocks (Talavera and Castresana 2007), we generated a phylogenetic tree for the 6,421 remaining genes (longer than 100 bases) using RAxML (v.8.2.12) (Stamatakis 2014).

To estimate the branch-wide rate ω and rates of synonymous (d_N) and synonymous mutations (d_S), we applied HyPhy's aBSREL (v.2.5.48) (adaptive Branch-Site Random Effects Likelihood) (Smith et al. 2015) to the 6,421 orthologous coding sequences. This model tests whether a proportion of sites have evolved under positive selection for each branch in the phylogeny. We also applied the FitMG94 model, implemented in the HyPhy package, to fit the Muse-Gaut model (Muse and Gaut 1994) of DNA sequence evolution, which estimates the number of synonymous and nonsynonymous substitutions per nucleotide site for each branch on the tree. To ensure the quality of subsequent analyses, we excluded genes with high ω values exceeding 2, resulting in 5,630 genes, using rates for the branch leading to *C. obscurior*. Using BEDTools' *map* function, genes were then assigned to the recombination rate of the 250 kb window where they were located. We excluded TE islands as well as introgressed regions from the correlation analysis due to their distinct evolutionary dynamics compared to the recipient populations used for linkage mapping, retaining a final set of 3,916 genes. Since the two models produced similar rates (ρ = 0.96, $p < 2.2\times 10^{-16}$), we restricted the correlation analyses to rates produced by the FitMG94 model.

Gene expression bias

To investigate the relationship between local recombination rate and developmental gene expression bias, we calculated the plasticity index (Schrader et al. 2017), capturing gene expression bias across

multiple caste/morph contrasts. High values indicate caste-specific gene expression, while low values suggest uniform expression across castes. RNA-seq data of worker, queen, wingless male, and winged male third instar larvae (n = 7 individuals each) were retrieved from NCBI (BioProject: PRJNA237579) (Schrader et al. 2015, 2017). Reads were quality-filtered for adapter contamination using Trimmomatic (options: ILLUMINACLIP:TruSeq3-SE.fa:2:30:10 SLIDINGWINDOW:4:20 MINLEN:30), and ribosomal RNA sequences were removed with SortMeRNA (v.4.3.4) (Kopylova et al. 2012) and the Silva databases (Quast et al. 2013). The filtered single-end reads were mapped to Cobs3.1 using STAR (v.2.6.0c) (Dobin et al. 2013) with default parameters, and mapping quality was evaluated with Qualimap. We used featureCounts (v.2.0.3) (Liao et al. 2014) to count reads per annotated gene, and differential expression analysis was performed with the R package limma (Ritchie et al. 2015) following Smyth et al. (2018). Read counts were converted to $\log_2$ counts per million and genes with fewer than 20 reads were removed. We applied the method of trimmed mean of M-values (Robinson and Oshlack 2010) from edgeR (Robinson et al. 2010) for normalization.

Plasticity indices were computed following Schrader et al. (2017) for genes to assess overall gene expression plasticity. Genes were then assigned to the recombination rate of the corresponding 250 kb window using BEDTools' *map* function. After excluding genes located in TE islands and introgressed regions, we retained 6,807 genes for the correlation analysis.

Supplemental References

Benson G. 1999. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573–580.

Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT. 2016. Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**: 1220–1222.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

Emms DM, Kelly S. 2019. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**: 238.

Errbii M, Keilwagen J, Hoff KJ, Steffen R, Altmüller J, Oettler J, Schrader L. 2021. Transposable elements and introgression introduce genetic variation in the invasive ant Cardiocondyla obscurior. In *Molecular Ecology*, Vol. 30 of, pp. 6211–6228, John Wiley & Sons, Ltd.

Huang YC, Lee CC, Kao CY, Chang NC, Lin CC, Shoemaker D, Wang J. 2016. Evolution of long centromeres in fire ants. *BMC Evol Biol* **16**.

Kirsche M, Prabhu G, Sherman R, Ni B, Battle A, Aganezov S, Schatz MC. 2023. Jasmine and Iris: population-scale structural variant comparison and analysis. *Nat Methods* **20**: 408–417.

Kopylova E, Noé L, Touzet H. 2012. SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**: 3211–3217.

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: A probabilistic framework for structural variant discovery. *Genome Biol* **15**: 1–19.

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**: 923–930.

Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, Sebra R, Peluso P, Eid J, Rank D, et al. 2013. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol* **14**: R10.

Mérot C, Stenløkk KSR, Venney C, Laporte M, Moser M, Normandeau E, Árnyasi M, Kent M, Rougeux C, Flynn JM, et al. 2023. Genome assembly, structural variants, and genetic differentiation between lake whitefish young species pairs (Coregonus sp.) with long and short reads. *Mol Ecol* **32**: 1458–1477.

Muse S V., Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol* **11**: 715–724.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**: D590–D596.

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012. DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**: i333–i339.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**: e47–e47.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139–140.

Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**: 1–9.

Schrader L, Helantera H, Oettler J. 2017. Accelerated evolution of developmentally biased genes in the tetraphenic ant cardiocondyla obscurior. *Mol Biol Evol* **34**: 535–544.

Schrader L, Simola DF, Heinze J, Oettler J. 2015. Sphingolipids, transcription factors, and conserved toolkit genes: Developmental plasticity in the ant cardiocondyla obscurior. *Mol Biol Evol* **32**: 1474–1486.

Shen W, Le S, Li Y, Hu F. 2016. SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* **11**: 1–10.

Shumate A, Salzberg SL. 2021. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**: 1639–1643.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**: 539.

Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL. 2015. Less is more: An adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol Biol Evol* **32**:

1342–1353.
Smyth GK, Ritchie ME, Law CW, Alhamdoosh M, Su S, Dong X, Tian L. 2018. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. *F1000Research* **5**: 1408.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.
Suyama M, Torrents D, Bork P. 2006. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**.
Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56**: 564–577.