

Oatk: a de novo assembly tool for complex plant organelle genomes

**Chenxi Zhou^{1,2}, Max Brown^{2,3}, Mark Blaxter², Darwin Tree of Life Project Consortium²,
Shane A. McCarthy^{1,2}, and Richard Durbin^{1,2,*}**

¹Department of Genetics, University of Cambridge, Downing Street, Cambridge, CB2 3EH, UK

²Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

³Faculty of Science and Engineering, Anglia Ruskin University, East Road, Cambridge, CB1 1PT, UK

*Correspondence: rd109@cam.ac.uk

Supplementary Figures



Fig S1. continued on next page

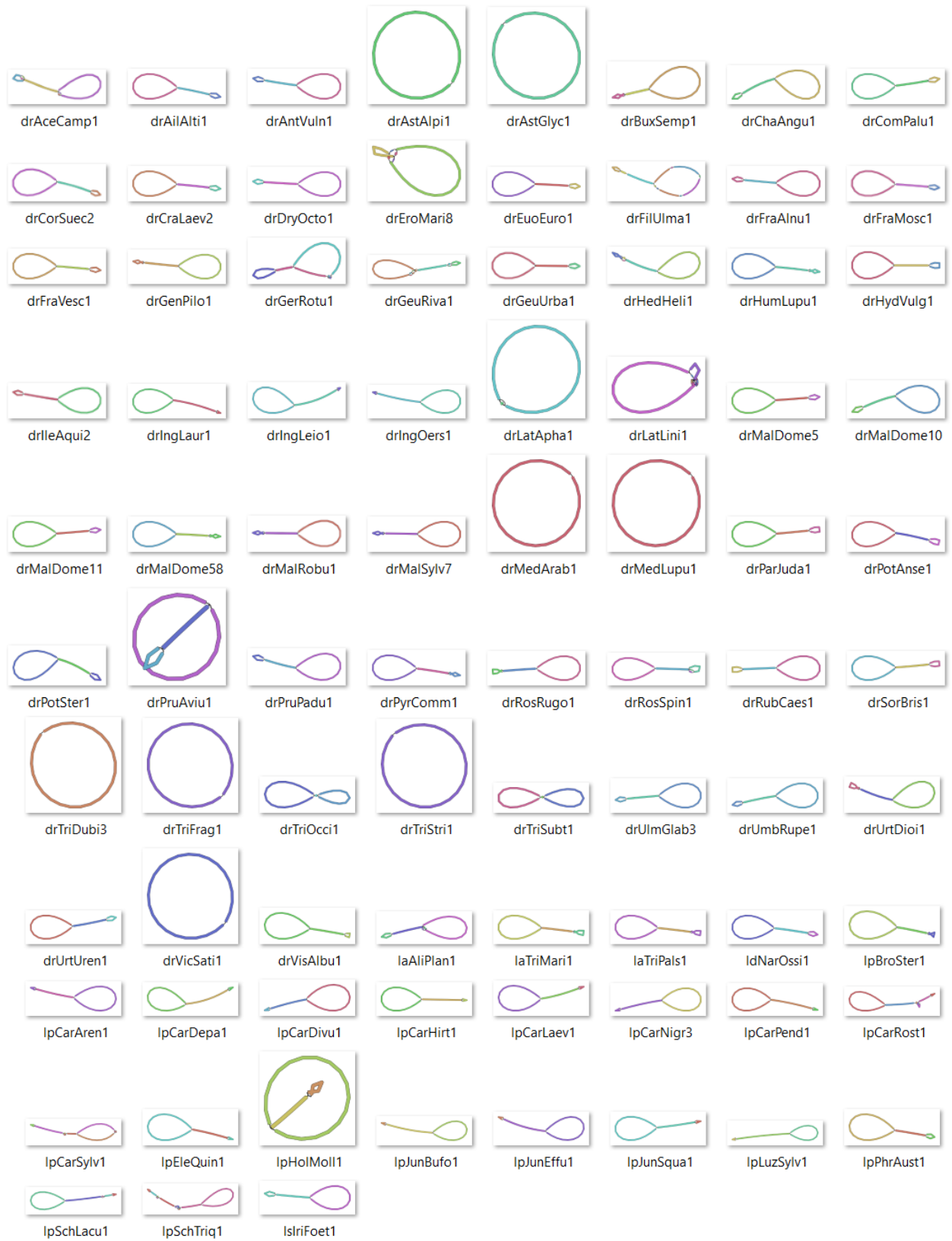


Fig S1. Assembly graphs for all plastomes. The graphs were produced using Bandage.



Fig S2. continued on next page

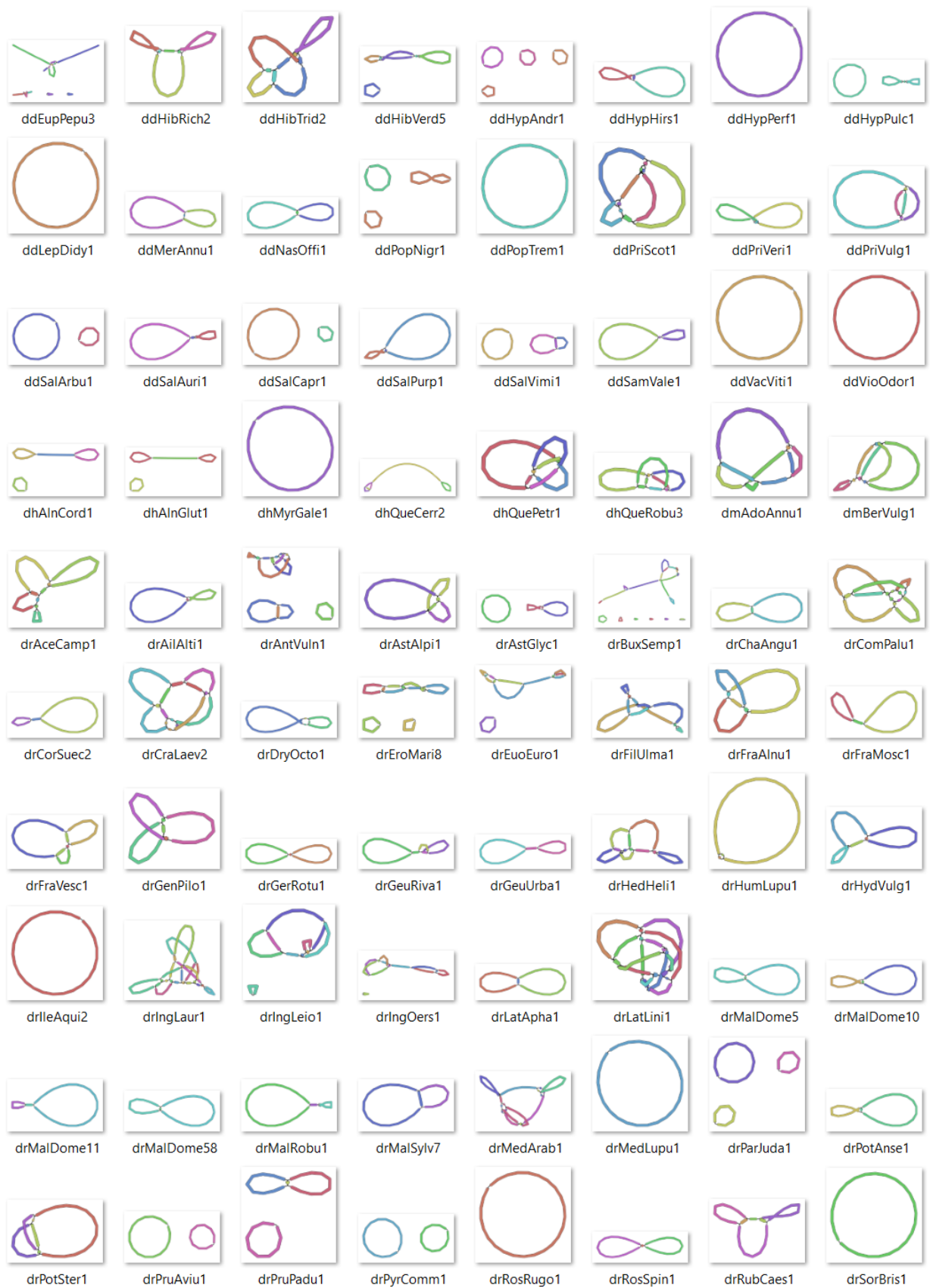


Fig S2. continued on next page

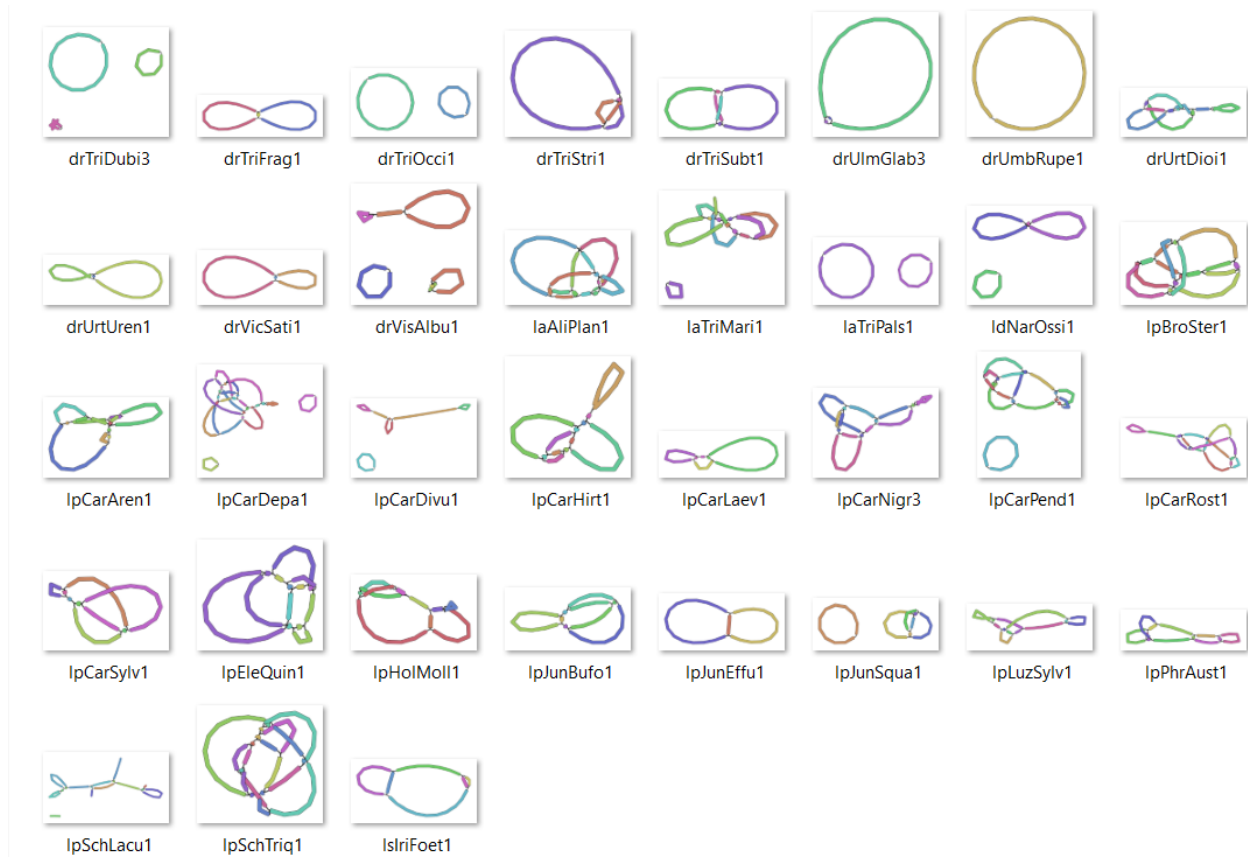
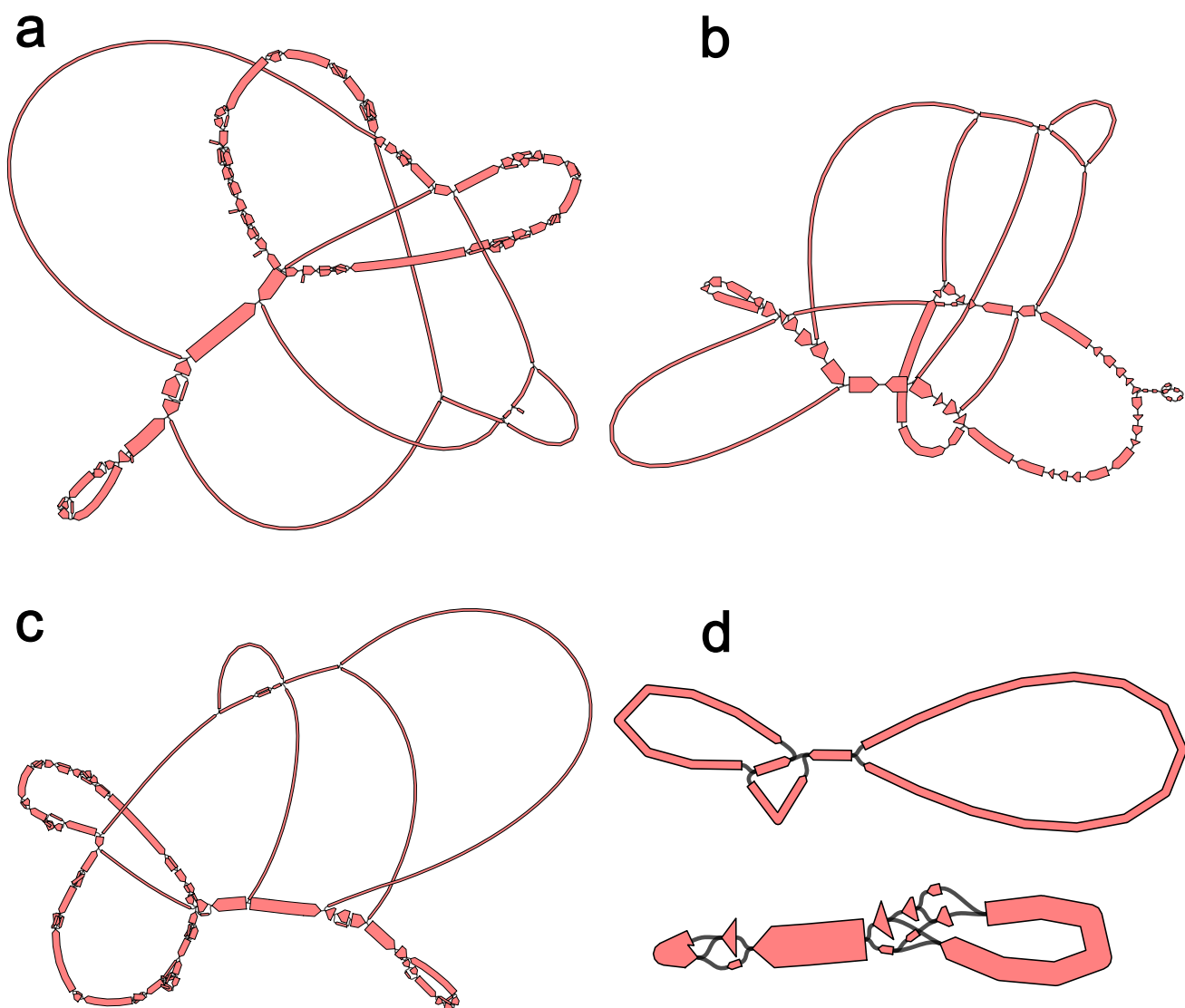
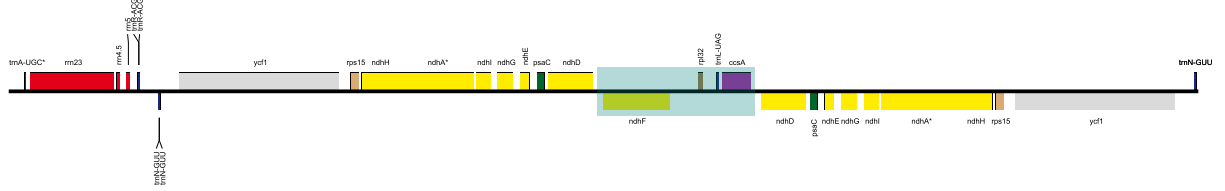


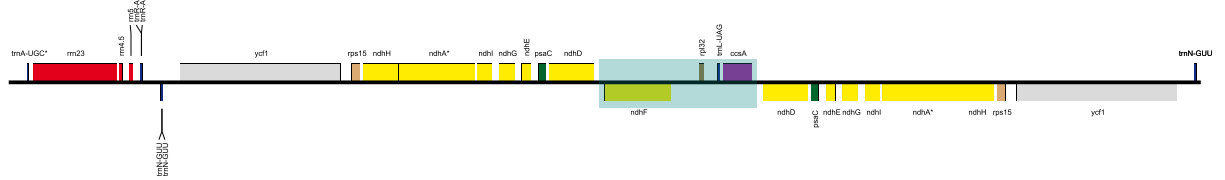
Fig S2. Assembly graphs for all mitogenomes. The graphs were produced using Bandage.



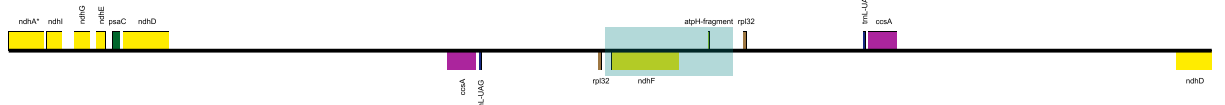
drIngOers1 [175176, 4951]



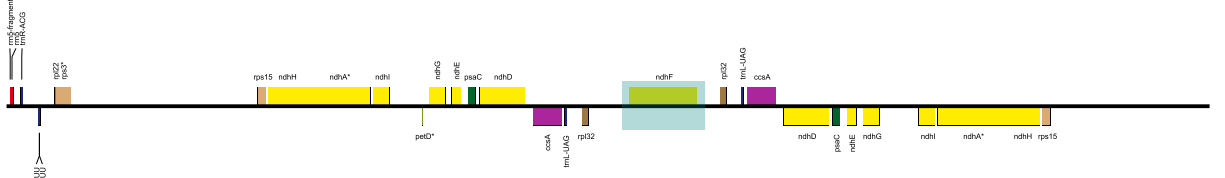
drIngLeio1 [175513, 4948]



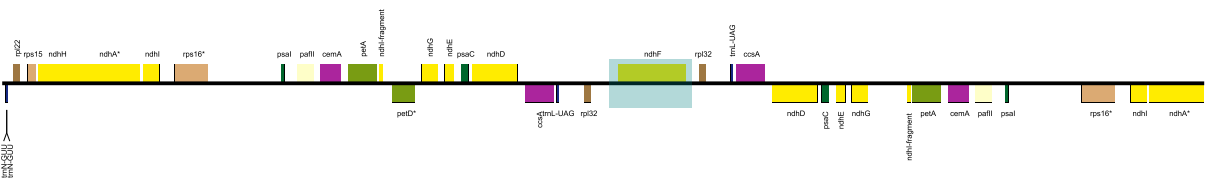
daJasMont1 [200376, 4129]



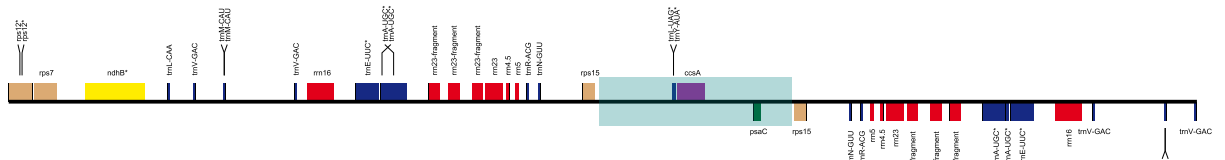
ddVacViti1 [180516, 3058]



ddCalVulg4 [208465, 2769]



lpLuzSylv1 [201321, 6186]



drVisAlbu1 [128925, 8632]

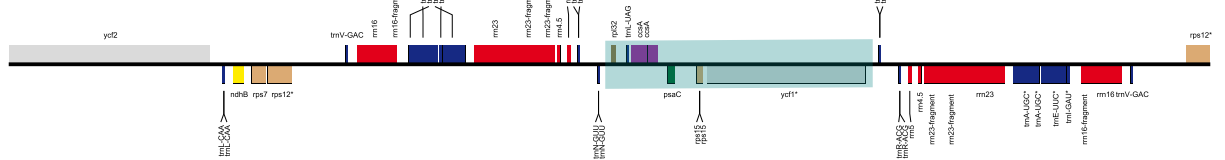


Fig S4. Gene annotation results in the SSC region for species with SSCs smaller than 10 kb. The numbers in square brackets next to the species ID denote the genome size and the SSC size. The light blue box represents the approximate location of the SSC. The annotation was performed using GeSeq and visualised with OGDRAW.

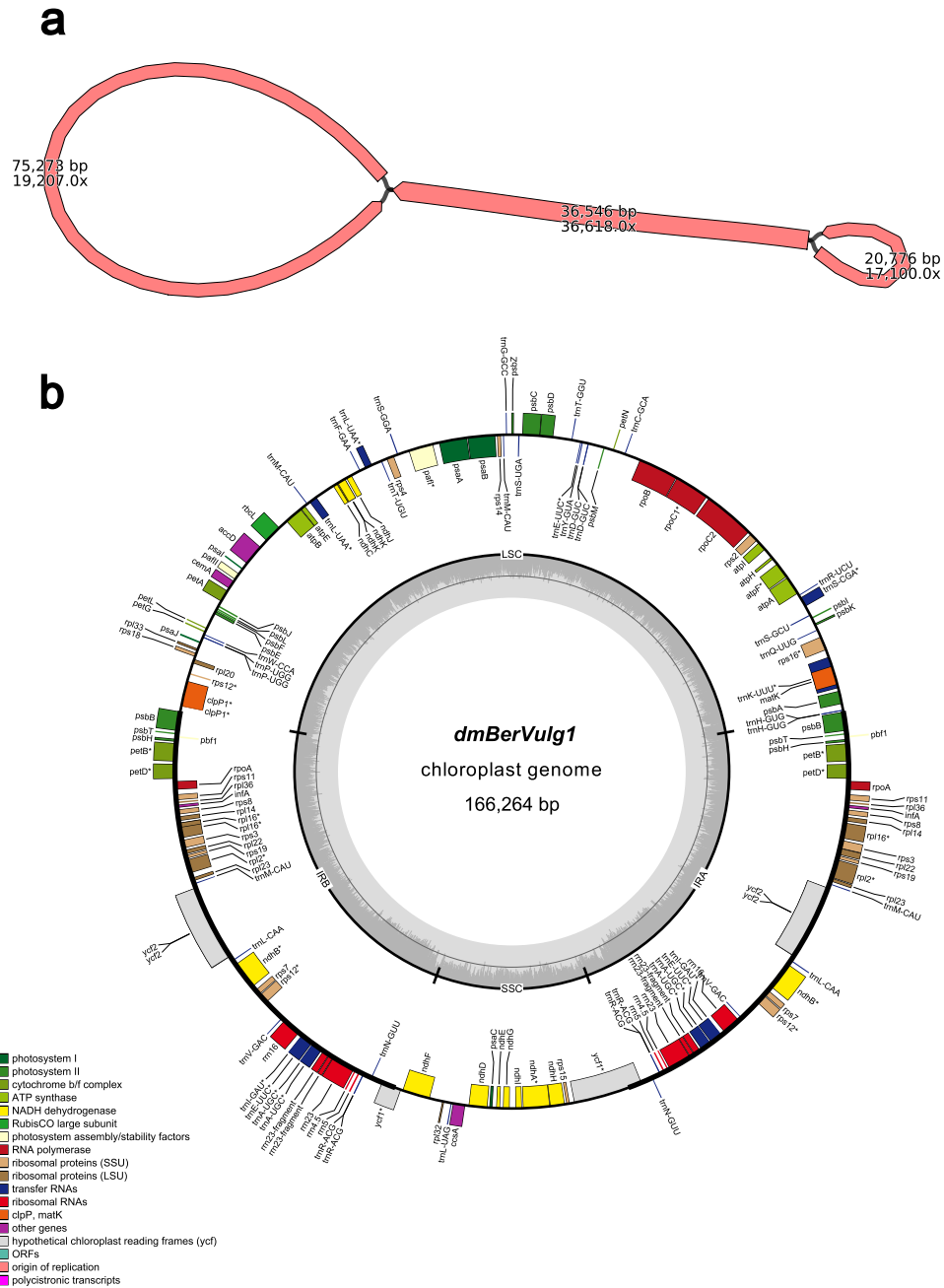


Fig S5. Plastome structure of *Berberis vulgaris*. **a**, the genome assembly graph. The numbers on the bar indicate the sequence length and coverage. **b**, the genome annotation. The assembly graph was produced using Bandage with additional manual adjustments. The annotation was performed using GeSeq and visualised with OGDRAW.

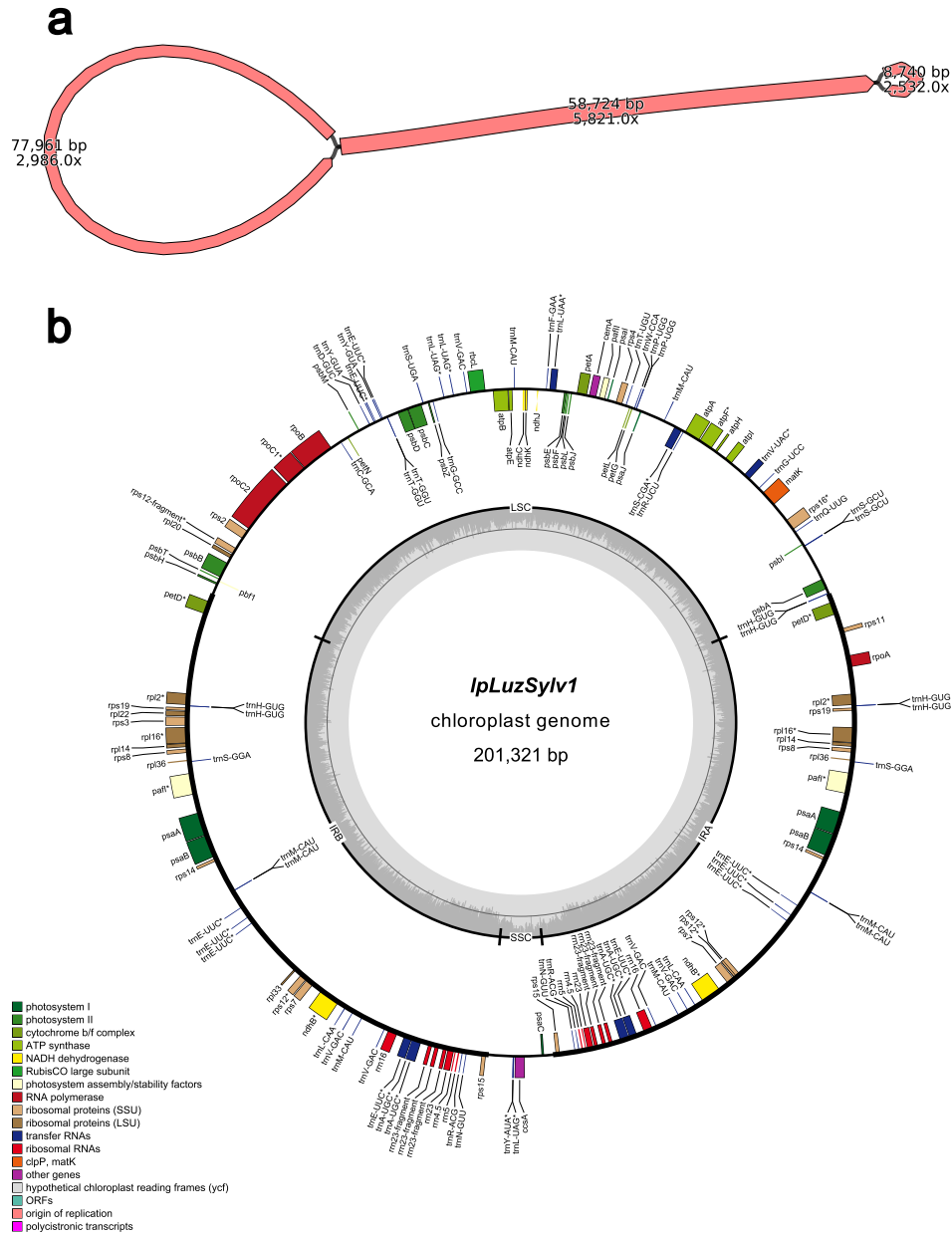


Fig S6. Plastome structure of *Luzula sylvatica*. **a**, the genome assembly graph. The numbers on the bar indicate the sequence length and coverage. **b**, the genome annotation. The assembly graph was produced using Bandage with additional manual adjustments. The annotation was performed using GeSeq and visualised with OGDRAW.

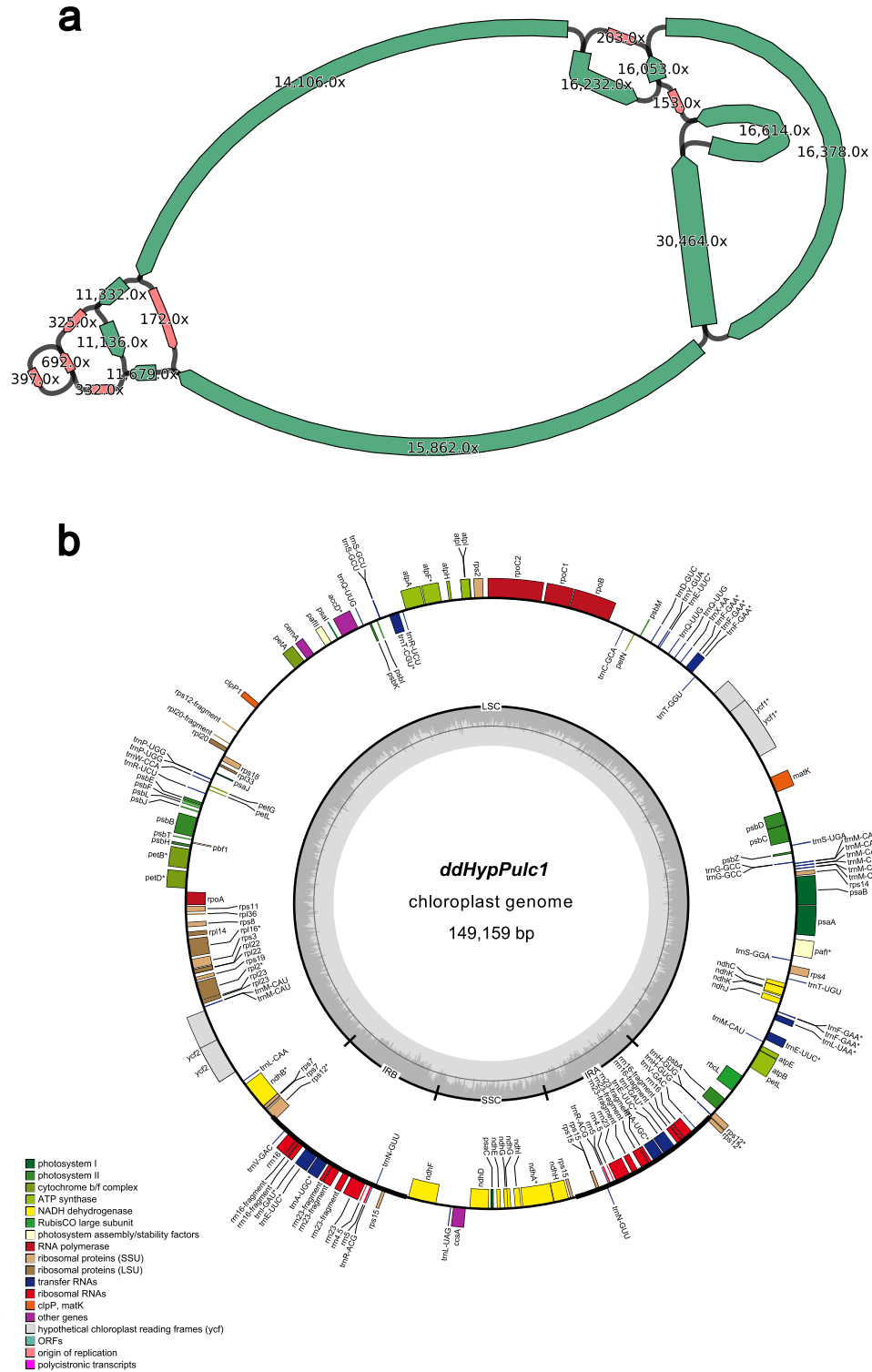


Fig S7. Plastome structure of *Hypericum pulchrum*. **a**, the genome assembly graph. The numbers on the bar indicate the sequence length and coverage. Sequences in green represent a path in the graph for the most abundant haplotype, while sequences in red represent heteroplasmy. **b**, the genome annotation. The assembly graph was produced using Bandage with additional manual adjustments. The annotation was performed using GeSeq and visualised with OGDRAW.

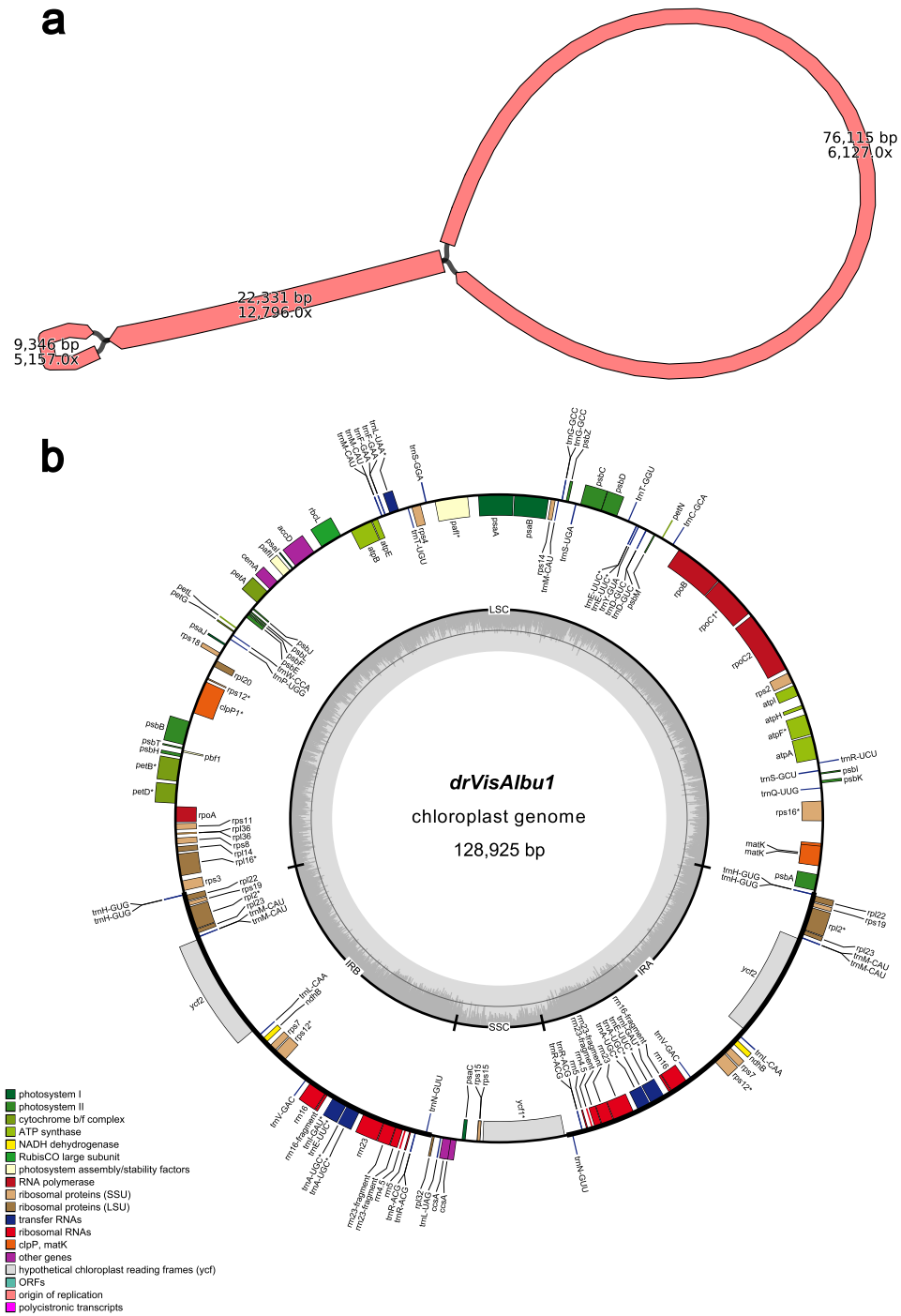


Fig S9. Plastome structure of *Viscum album*. **a**, the genome assembly graph. The numbers on the bar indicate the sequence length and coverage. **b**, the genome annotation. The assembly graph was produced using Bandage with additional manual adjustments. The annotation was performed using GeSeq and visualised with OGDRAW.

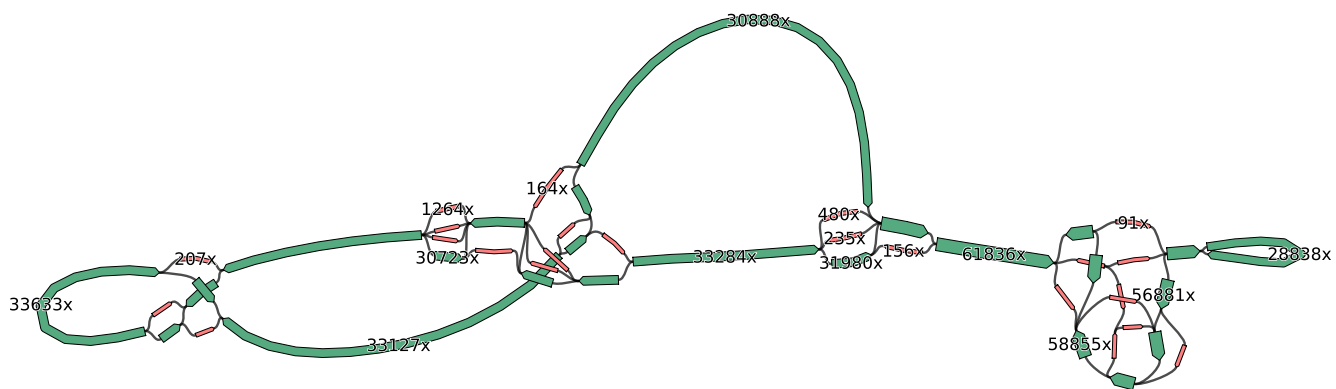


Fig S10. Plastome structure of *Hypericum perforatum*. In the graph, the width of each bar is proportional to the sequence coverage. The number on the bar indicates the actual coverage of the sequence, with some omitted for clarity. Sequences in green represent a path in the graph for the most abundant haplotype, while sequences in red represent heteroplasmy. The assembly graph was produced using Bandage with additional manual adjustments.

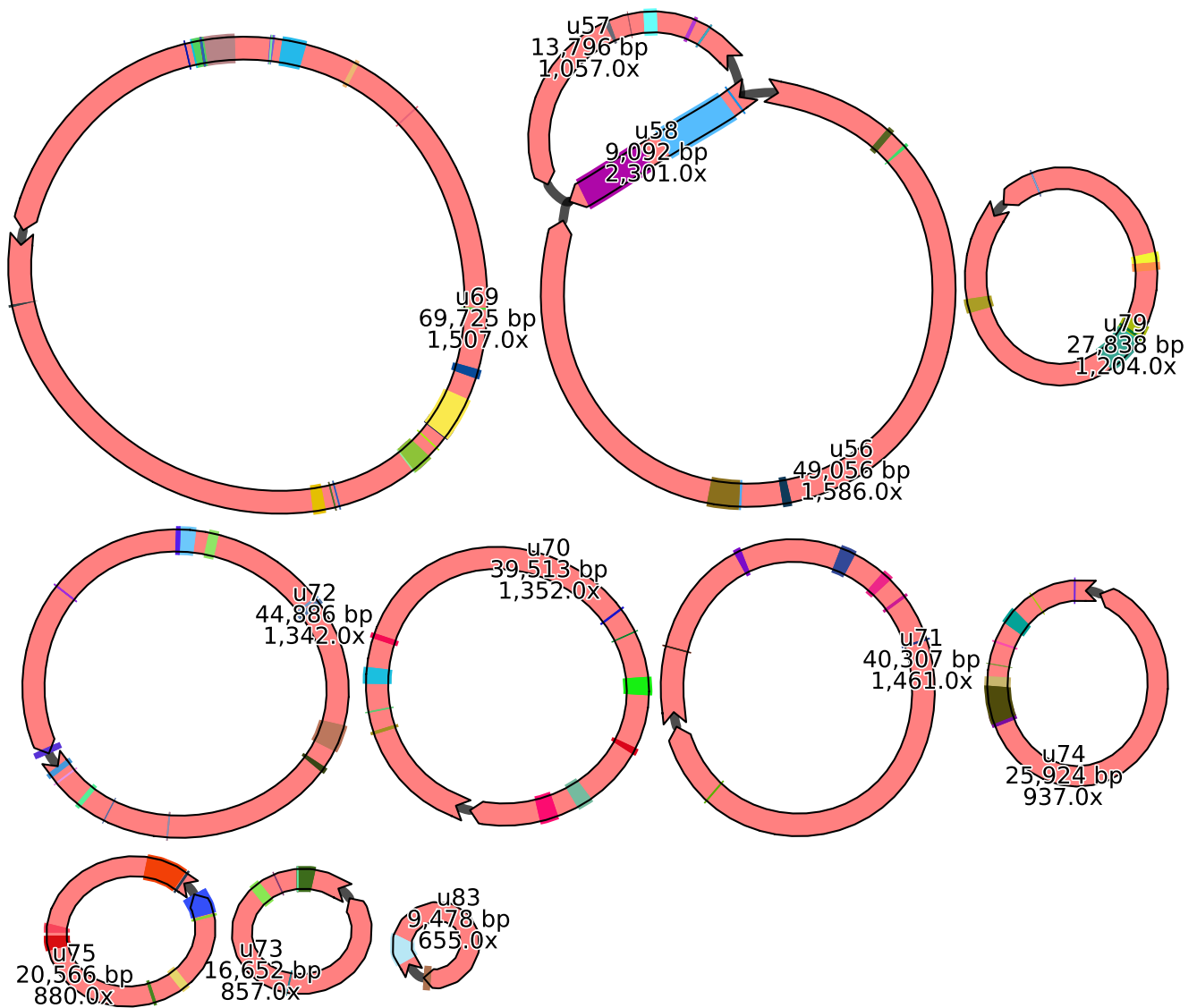


Fig S11. Mitogenome assembly of *Galeopsis tetrahit*. The genome assembly consists of ten components, with coloured blocks over the bars representing annotated genes. The numbers on the bar indicate the sequence name, length and coverage. The assembly graph was produced using Bandage with additional manual adjustments.

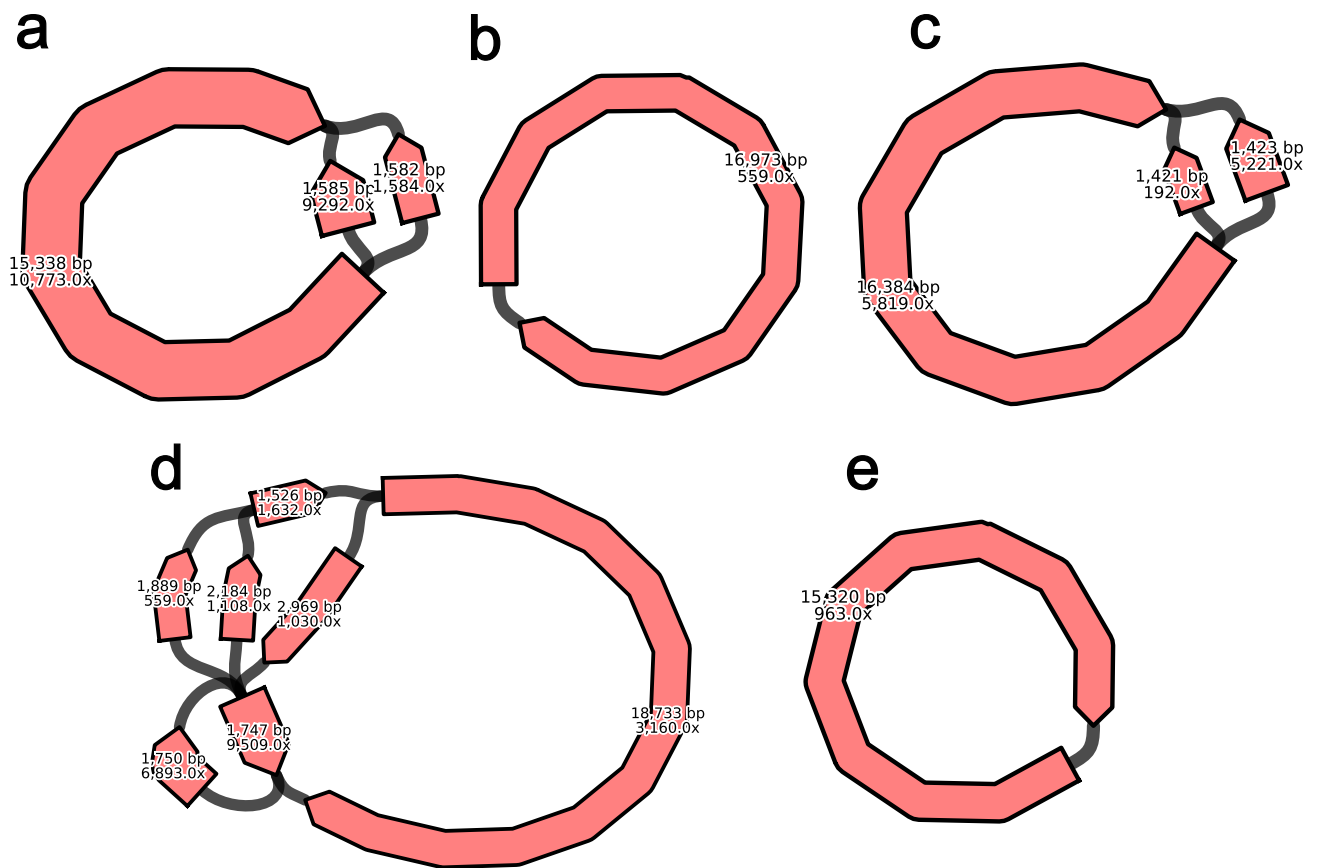


Fig S12. Mitogenome assembly of five animal species. **a**, *Micromys minutus* (TOLID: mMicMin1; ACCESSION: PRJEB72093). **b**, *Abramis brama* (TOLID: fAbrBra2; ACCESSION: PRJEB73975). **c**, *Anas platyrhynchos* (TOLID: bAnaPla2; ACCESSION: PRJEB76742). **d**, *Podarcis gaigeae* (TOLID: rPodGai1; ACCESSION: PRJEB76108). **e**, *Erebia medusa* (TOLID: ilEreMedu1; ACCESSION: PRJEB76717). The numbers on the bar indicate the sequence length and coverage. The assembly graph was produced using Bandage with additional manual adjustments.