Plectus sambesii assembly methods: Supplementary Note 1

We acquired ~250M 150bp paired-end genome sequencing read pairs with 550bp insert size through Edinburgh Genomics. Quality control of the genome sequencing data was performed using Fastqc v0.10.1 (Andrews, 2010), and reads were quality trimmed using skewer v0.2.2 (Jiang et al., 2014) with parameters –q 30 –l 51. K-mer plots were generated with kmc, which revealed extensive heterozygosity. A preliminary single-end assembly was generated with Velvet (Zerbino and Birney, 2008) to screen for contaminants but no significant hits were found using Blobtools (Kumar et al., 2013; Laetsch, 2016). The dataset was digitally normalised to 40x coverage and assembly was carried out with SPAdes v3.5.1 (Bankevich et al., 2012) with parameters --k 21,33,55,77,99 --cov-cutoff auto --careful. Error correction of the sequencing reads was done with BayesHammer that runs as part of the SPAdes pipeline. Examination of the resulting distribution of contig coverages revealed a bimodal distribution of coverage indicating again heterozygosity. Haplocontigs were collapsed and postprocessed with Redundans (Pryszcz and Gabaldón, 2016), that runs SSPACE3 (Boetzer et al., 2011) and SOAPdenovo Gapcloser internally. The identity percentage for the collapse of redundant contigs was chosen to be of 90%, ensuring the presence of 95.97% of CEGMA KOGs (Parra et al., 2007). Lower identity percentage values did not result in an appreciable reduction of the span suggesting that most heterozygous regions are collapsed at that threshold. Blobplot analysis after postprocessing revealed ~20Mb of bacterial contaminant sequence at 60% GC and lower coverage than the nematode contigs, which we therefore excluded from the final assembly.

To annotate genes, we first masked repeats in the assembly. We used RepeatModeler v1.0.8 (Smit and Hubley, 2008) to generate a species-specific repeat library that was concatenated with a nematode repeat library extracted from RepBase RepeatMasker edition update 27-01-2017. The combined library was used to mask the genome with RepeatMasker v4.0.7 (Smit et al., 2013). To annotate the genome, we acquired ~100M RNA-seq reads from rRNA-depleted paired-end RNA-seq libraries generated in the LMS Genomics Facility. Quality control of the RNA-seq reads was performed with Fastqc and Skewer with parameters –q 30 –l 50. Spliced alignment of the reads to the masked assembly was done with STAR v2.5.2 (Dobin et al., 2013) with default parameters except for --twopassMode Basic. RNA-seq alignments were used for automated HMM training and annotation using BRAKER v1.9 (Hoff et al., 2016) with parameters --softmasking 1 --UTR off --gff3, running internally GeneMark-ET v4.21 (Lomsadze et al., 2014) and augustus v2.3.2 (Stanke and Morgenstern, 2005).

References

Andrews, S. (2010). FastQC A Quality Control tool for High Throughput Sequence Data. Http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. a., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J. Comput. Biol. *19*, 455–477.

Boetzer, M., Henkel, C. V, Jansen, H.J., Butler, D., and Pirovano, W. (2011).

Scaffolding pre-assembled contigs using SSPACE. Bioinformatics *27*, 578–579.
Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.
Hoff, K.J., Lange, S., Lomsadze, A., Borodovsky, M., and Stanke, M. (2016). BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics *32*, 767–769.
Jiang, H., Lei, R., Ding, S.-W., and Zhu, S. (2014). Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics *15*, 182.
Kumar, S., Jones, M., Koutsovoulos, G., Clarke, M., and Blaxter, M. (2013). Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. Front. Genet. *4*, 237.
Laetsch, D. (2016). Blobtools. GitHub Repos.
Lomsadze, A., Burns, P.D., and Borodovsky, M. (2014). Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res. *42*, e119.
Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics *23*, 1061–1067.
Pryszcz, L.P., and Gabaldón, T. (2016). Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Res. *44*, e113–e113.
Smit, A., and Hubley, R. (2008). RepeatModeler Open-1.0.
Smit, A., Hubley, R., and Green, P. (2013). RepeatMasker Open-4.0.
Stanke, M., and Morgenstern, B. (2005). AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. *33*, 465–467.
Zerbino, D.R., and Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. *18*, 821–829.

Supplementary Note 2: Phylogeny References

Protozoa

Sierra, R., Matz, M. V., Aglyamova, G., Pillet, L., Decelle, J., Not, F., et al. (2013). Deep relationships of Rhizaria revealed by phylogenomics: A farewell to Haeckel's Radiolaria**,** *Molecular Phylogenetics and Evolution*, *67*(1), 53–59. http://doi.org/10.1016/j.ympev.2012.12.011

Cavalier-Smith, T., Fiore-Donno, A. M., Chao, E., Kudryavtsev, A., Berney, C., Snell, E. A., & Lewis, R. (2015). Multigene phylogeny resolves deep branching of the Amoebozoa. *Molecular Phylogenetics and Evolution*, *83*(C), 293–304. http://doi.org/10.1016/j.ympev.2014.08.011

Derelle, R., L pez-Garc a, P. N., Timpano, H. L. N., & Moreira, D. (2016). A Phylogenomic Framework to Study the Diversity and Evolution of Stramenopiles (=Heterokonts). *Molecular Biology and Evolution*, *33*(11), 2890–2898. http://doi.org/10.1093/molbev/msw168

Jinkerson, R. E., Radakovits, R., & Posewitz, M. C. (2014). Genomic insights from the oleaginous model alga Nannochloropsis gaditana. *Bioengineered*, *4*(1), 37–43. http://doi.org/10.4155/bfs.11.7

Fritz-Laylin, L. K., Prochnik, S. E., Ginger, M. L., Dacks, J. B., Carpenter, M. L., Field, M. C., et al. (2010). The Genome of Naegleria gruberi Illuminates Early Eukaryotic Versatility. *Cell*, *140*(5), 631–642. http://doi.org/10.1016/j.cell.2010.01.032

Read, B. A., Kegel, J., Klute, M. J., Kuo, A., Lefebvre, S. C., Maumus, F., et al. (2014). Pan genome of the phytoplankton Emiliania underpins its global distribution. *Nature*, *499*(7457), 209–213. http://doi.org/10.1038/nature12221

Microsporidia

Vossbrinck, C.R., Debrunner-Vossbrinck, B.A., &Weiss, L.M., 2014 Phylogeny of the Microsporidia (2014). In *Microsporidia, Pathogens of opportunity*, Edited Weiss, L and Becnel, J, Wiley.

Fungi

Sugiyama ,I. Hosaka, K., Suh S.O (2006). Early diverging Ascomycota: phylogenetic divergence and related evolutionary enigmas. *Mycologia*, 98(6), 996-1005.

Hibbert D.S., (2006). A phylogenetic overview of the Agaricomycotina, *Mycologia*. 98(6):917-25.

James, T. Y., Kauff, F., Schoch, C. L., Matheny, P. B., Hofstetter, V. R., Cox, C. J., et al. (2006). Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature*, *443*(7113), 818–822. http://doi.org/10.1038/nature05110

Masneuf-Pomarede, I., Bely, M., Marullo, P., & Albertin, W. (2016). The Genetics of Non-conventional Wine Yeasts: Current Knowledge and Future Challenges. *Frontiers in Microbiology*, *6*(1569), 702. http://doi.org/10.1016/j.ijfoodmicro.2011.08.026

Supplementary Note 3: sample numbers for genome-wide comparisons

CGs covered by >10 reads used for significance testing in Figure 1c
Plectus sambesii
gene 1230239
rRNA 1193
LTR 7069
DNA 1599
unannotated 727548

Romanomermis culicivorax
gene 656710
rRNA 561
LTR 4897
DNA 1830
unannotated 124399

Trichuris muris
gene 2115909
rRNA 3183
LTR 79067
DNA 51875
unannotated 1247125

Trichinella spiralis
gene 738591
rRNA 291
LTR 21480
DNA 3540
unannotated 193453

Caenorhabditis briggsae
gene 82526
rRNA 708
LTR 590
DNA 3725
unannotated 59570

Nippostrongylus brasiliensis
Gene 20765
rRNA 926
DNA 1199
LTR 2879
unannotated 12940

Analysed genomes: number of genome-wide Cs and CGs

P. sambesii
total_C 29615596
total_CG 7728644
total_CGA 2355653
total_CGC 2169535
total_CGG 1402387
total_CGT 1793453

R. culicivorax
total_C 48473942
total_CG 7829396
total_CGA 2785794
total_CGC 1457635
total_CGG 1529381
total_CGT 2050824

T. spiralis
total_C 9943967
total_CG 1710706
total_CGA 539942
total_CGC 348898
total_CGG 309852
total_CGT 511933

T. muris
total_C 18839135
total_CG 4505970
total_CGA 1184822
total_CGC 1125042
total_CGG 921658
total_CGT 1274273

N. brasiliensis
total_C 58220992
total_CG 12740122
total_CGA 4131434
total_CGC 2803568
total_CGG 2582339
total_CGT 3221088

C. briggsae
total_C 19684666
total_CG 3455622
total_CGA 1266130
total_CGC 588973
total_CGG 765446
total_CGT 834762

B. mori
total_C 79039767
total_CG 16946963
total_CGA 4852595
total_CGC 3693029
total_CGG 3551168
total_CGT 4848034

M. musculus
total_C 38179275
total_CG 1428189
total_CGA 354635
total_CGC 299758
total_CGG 374931
total_CGT 398865

Analysed genomes: CGs within +-1000bp of TSS with sufficient coverage to estimate methylation (see Supplemental Figure 3)
P. sambesii 2002411
R. culicivorax 6231622
T. spiralis 1055855
T. muris 1632567

CGs in 1st exon vs introns analysed (Supplemental Figure 3)
P. sambesii 480833e,4352323i
R. culicivorax 2506380e,5629056i
T. spiralis 502608e,1992647i
T. muris 287818e,2699672i

total number of genes analysed with DNA methylation coverage for repeat homology (Supplemental Figure 2)
P. sambesii 38467
R. culicivorax 43622
T. spiralis 12913
T. muris 10932