References

1. Hug, L. A. *et al.* A new view of the tree of life. *Nature Microbiology* **1,** 16048 (2016).
2. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
3. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7,** e1002195 (2011).
4. Albertin, C. B. *et al.* The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature* **524,** 220–224 (2015).
5. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research* (2007).
6. Keeling, P. J. *et al.* The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol* **12,** e1001889 (2014).
7. Nekrutenko, A. & Taylor, J. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nat Rev Genet* **13,** 667–672 (2012).
8. O'Neil, S. T. & Emrich, S. J. Assessing De Novo transcriptome assembly metrics for consistency and utility. *BMC Genomics* **14,** 1 (2013).
9. Fisch, K. M. *et al.* Omics Pipe: a community-based framework for reproducible multi-omics data analysis. *Bioinformatics* **31,** 1724–1728 (2015).
10. Nevado, B. & Perez Enciso, M. Pipeliner: software to evaluate the performance of bioinformatics pipelines for next-generation resequencing. *Molecular Ecology Resources* **15,** 99–106 (2015).
11. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29,** 644–652 (2011).
12. Scott, C. dammit: an open and accessible de novo transcriptome annotater. *in prep* (2016).
13. Pierce, N. T., Hartwick, N., Burton, R. S. & Gaasterland, T. MakeMyTranscriptome: an Automated *De Novo* Transcriptome Analysis Framework. Available at: https://github.com/bluegenes/MakeMyTranscriptome. (Accessed: 1st November 2016)
14. MacManes, M. D. *Optimizing error correction of RNAseq reads*. 1–4 (2015). doi:10.1101/020123
15. MacManes, M. D. & Eisen, M. B. Improving transcriptome assembly through error correction of high-throughput sequence reads. *PeerJ* **1,** e113 (2013).
16. Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B. & Brom, T. H. A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. (2012).
17. Brown, C. T. Abundance Counting of sequences in graphs with graphalign. *ivory.idyll.org* (2015). Available at: http://ivory.idyll.org/blog/2015-wok-counting.html. (Accessed: 31st October 2016)
18. Wang, Q. *et al.* Xander: employing a novel method for efficient gene-targeted metagenomic assembly. *Microbiome* **3,** 32 (2015).
19. Holley, G. & Peterlongo, P. BlastGraph: intensive approximate pattern matching in string graphs and de-Bruijn graphs. (2012).
20. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Meth* **12,** 59–60 (2014).
21. Patro, R., Duggal, G. & Kingsford, C. *Accurate, fast, and model-aware transcript expression quantification with Salmon*. (2015). doi:10.1101/021592
22. Limasset, A., Cazaux, B., Rivals, E. & Peterlongo, P. Read mapping on de Bruijn graphs. *BMC Bioinformatics* **17,** 237 (2016).
23. Westbrook, A. *et al. PALADIN:Protein Alignment for Functional Profiling Whole Metagenome Shotgun Data*. *bioRxiv* 047712 (Cold Spring Harbor Labs Journals, 2016). doi:10.1101/047712
24. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
25. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Meth* **9,** 357–359 (2012).
26. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq

quantification. *Nature Biotechnology* **34,** 525–527 (2016).

27. Brown, C. T. Bashing on monstrous sequencing collections. *ivory.idyll.org* (2016). Available at: _http://ivory.idyll.org/blog/2016-mmetsp-a-first-look.html. (Accessed: 1st November 2016)
28. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *bioRxiv* 029827 (2016). doi:10.1101/029827
29. Solomon, B. & Kingsford, C. Fast search of thousands of short-read sequencing experiments. *Nature Biotechnology* **34,** 300–302 (2016).
30. Brown, C. T. Mixing Bloom Filters, Minhashes, and SBT. (2016). Available at: https://nbviewer.jupyter.org/github/luizirber/2016-sbt-minhash/blob/4a3a7fbfa55355e679183496d66d41fc26ae8c3f/SBT%20with%20MinHash%20leaves.ipynb. (Accessed: 31st October 2016)
31. Titus Brown, C. & Irber, L. sourmash: a library for MinHash sketching of DNA. *JOSS* **1,** (2016).
32. Brown, C. T. Applying MinHash to RNA-Seq Samples. *ivory.idyll.org* (2016). Available at: http://ivory.idyll.org/blog/2016-sourmash.html. (Accessed: 31st October 2016)
33. Hishiki, T., Kawamoto, S., Morishita, S. & Okubo, K. BodyMap: a human and mouse gene expression database. *Nucleic Acids Research* **28,** 136–138 (2000).
34. Chum, O., Philbin, J. & Zisserman, A. Near Duplicate Image Detection: min-Hash and tf-idf Weighting. *BMVC* (2008).
35. Shrivastava, A. Exact Weighted Minwise Hashing in Constant Time. (2016).
36. Hubbart, J., Link, T., Campbell, C. & Cobos, D. Evaluation of a low-cost temperature measurement system for environmental applications. *Hydrological Processes* **19,** 1517–1523 (2005).