MAGpy: a reproducible pipeline for the downstream analysis of metagenome-assembled genomes (MAGs)

Robert Stewart[1], Marc Auffret[2], Tim Snelling[3], Rainer Roehe[2] and Mick Watson[1]

1. The Roslin Institute and R(D)SVS, University of Edinburgh, Easter Bush, EH25 9RG
2. Scotland's Rural College, Easter Bush, EH25 9RG
3. The Rowett Institute of Nutrition and Health, University of Aberdeen, King's College, Aberdeen, AB24 3FX

Abstract
Recent advances in bioinformatics have enabled the rapid assembly of genomes from metagenomes (MAGs), and there is a need for reproducible pipelines that can annotate and characterise thousands of genomes simultaneously. Here we present MAGpy, a Snakemake pipeline that takes FASTA input and compares MAGs to several public databases, checks quality, assigns a taxonomy and draws a phylogenetic tree.

Introduction
Discovering and studying microbes in the environment has been a goal of genomic technologies for many years [1,2], but advances in DNA sequencing [3–5] have enabled a revolution in metagenomics that has accelerated this area of research. Metagenomics refers to the whole-genome investigation of every organism within a particular environment, and is often used in microbiome studies to investigate changes in the taxonomic and functional profile of samples of interest. This method of simultaneously quantifying taxonomic and functional structure has been used in studies of age and geography in the human gut [6], release of carbon due to permafrost thawing [7], the environmental impact and feed efficiency of animal agriculture [8,9], environmental characterisation of Earth's oceans [10,11] and the extraction of industrially and commercially relevant enzymes from environmental samples [12].

Metagenomics also offers the ability to assemble near-complete and draft microbial genomes without the need for culture. Such "metagenomic binning" approaches involve the assembly of metagenomic sequence reads into contigs followed by clustering, or binning, of contigs into putative genomes, called metagenome-assembled genomes (MAGs) [13]. This was first demonstrated in 2004 when Tyson *et al* [14] assembled multiple near complete genomes from an environmental biofilm. More recently, this technique has been used to assemble complete and draft genomes from ruminants [15–18], oil reservoirs [19], aquifers [20] and the Tara oceans expedition [21].

A major challenge in the analysis of metagenome-assembled genomes is that researchers are often presented with hundreds or thousands of putative genomes, which need to be annotated, characterised, placed within a phylogenetic tree and assigned a putative function or role. This is further complicated by the fact that many of the putative genomes do not have close relatives with good quality reference genomes, making comparative genomics almost impossible.

Here we present MAGpy (pronounced 'magpie'), a reproducible pipeline for the characterisation of MAGs using open source and freely available bioinformatics software. MAGpy is implemented as a Snakemake [22] pipeline, enabling reproducible analyses, extensibility, HPC integration with high-performance-compute clusters (HPC) and restart capabilities. MAGpy annotates the genomes, predicts putative protein sequences, compares the MAGs to multiple genomic, proteomic and protein family databases, produces several reports and draws a taxonomic tree. We demonstrate the utility of MAGpy on a subset of 800 bacterial and archaeal MAGs recently published by Parks *et al* [23].

Methods and results
Snakemake is a workflow management system with a python-based workflow definition language, and works with any tool, including web services, that have well defined inputs and outputs. Workflows are defined within

Snakefiles, within which "rules" are defined that are essentially operations on data, converting inputs to outputs. That conversion can be carried out by a locally installed tool, or shell or Python code embedded within the Snakefile itself. The workflow and rule dependency is defined by setting the output of one tool as the input to another. Snakemake scales to 1000s of input datasets and complex workflows as it integrates with many HPC job management platforms, such as SGE or Slurm.

MAGpy makes use of Snakemake to define an analysis workflow for MAGs based on open source and freely available bioinformatics software. First, CheckM [24] is run, which uses a set of pre-computed core genes to assess the completeness and contamination of MAGs. CheckM also attempts to assign a taxonomic level to the MAGs, though in our experience this is often a conservative estimate. In tandem, MAGpy predicts the protein coding sequences of MAGs using Prodigal [25]. DIAMOND [26] blastp is used to compare the proteins to UniProt [27]. This has multiple purposes – the hits from UniProt provide a form of annotation of the putative proteins and may predict function; many MAGs may show little similarity to published genomes at the DNA level, but as proteins are more conserved, protein hits can help define the closest sequenced genome; and the length of the predicted protein and that protein's hits can be used to detect truncated genes and proteins in the MAG annotation. Reports of the DIAMOND results at the level of the MAG and for each contig within each MAG are produced. The proteins are also compared to protein families in Pfam [28] using PfamScan; and to draw a tree using PhyloPhlAn [29], which is subsequently visualised using GraPhlAn [30]. The MAG genome sequences are also compared to over 100,000 public genomes using MinHash signatures as implemented in Sourmash.

We applied MAGpy to 800 Bacterial and Archaeal MAGs from Parks *et al* [23]. The CheckM report (which uses Ete3 [31] to expand the taxonomic prediction) can be seen in Supplementary Table 1, the Sourmash report in Supplementary Table 2 and the Uniprot report in Supplementary Table 3. The PhyloPhlAn tree can be seen in Figure 1.

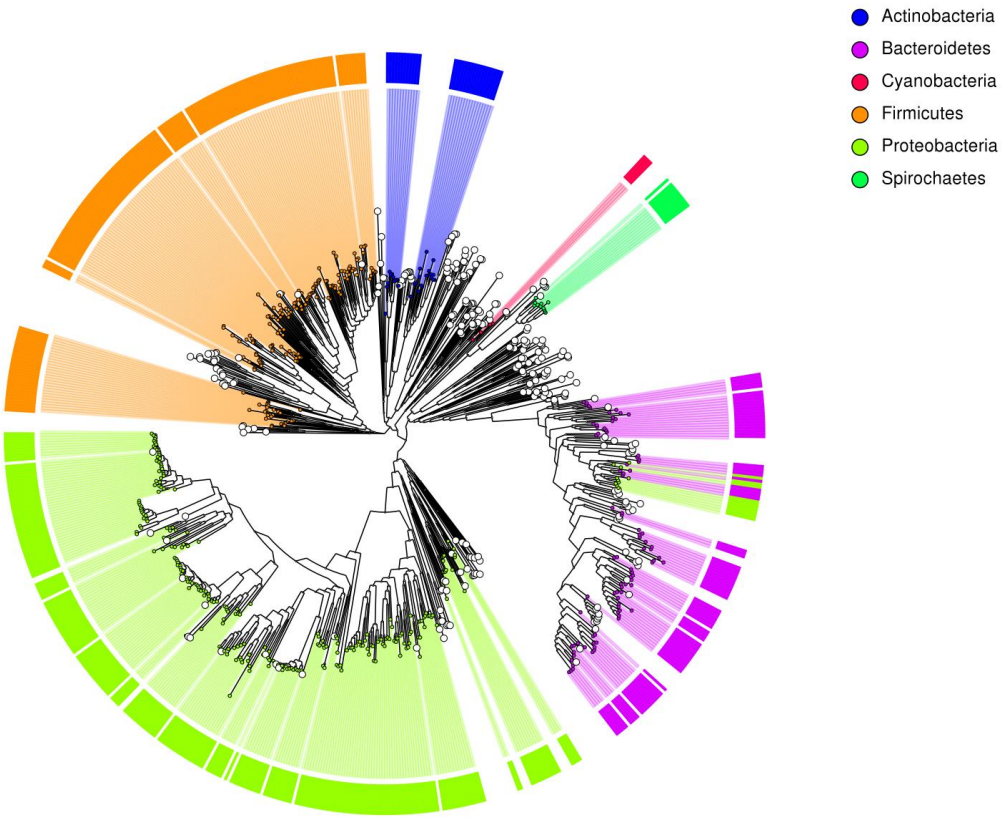


Figure 1. Phylogenetic tree of 800 MAGs created using PhyloPhlAn and produced by MAGpy.

Discussion and conclusion

Given the cost and difficulty of microbial culture, being able to assemble whole genomes from metagenomes without the need for culture is a hugely promising method which will enable researchers to populate the microbial tree of life rapidly and at relatively low cost. Short read assembly and binning are currently the most powerful method for creating MAGs; however, as long reads become more readily used for environmental metagenomes [32], this will enable the assembly of complex and difficult to assemble genomes directly from the environment, including those with large numbers of repeats and rearrangements [33,34]. Whichever sequencing technology is used, we predict that a huge amount of novel microbial genomes will need to be annotated and analyzed in a consistent and reproducible way, and publication of such pipelines is a key aspect of open science [35].

Availability

All code and documentation is available at https://github.com/WatsonLab/MAGpy

Acknowledgements and funding

The Roslin Institute forms part of the Royal (Dick) School of Veterinary Studies, University of Edinburgh. This project was supported by the Biotechnology and Biological Sciences Research Council (BBSRC; BB/N016742/1, BB/N01720X/1), including institute strategic programme and national capability awards to The Roslin Institute (BBSRC: BB/P013759/1, BB/P013732/1, BB/J004235/1, BB/J004243/1); and by the Scottish Government (RESAS) as part of the 2016-2021 commission.

References

1. Watson M, Dukes J, Abu-Median A-B, King DP, Britton P. DetectiV: visualization, normalization and significance testing for pathogen-detection microarray data. Genome Biol. [Internet]. 2007;8:R190. Available from: http://dx.doi.org/10.1186/gb-2007-8-9-r190

2. Brodie EL, Desantis TZ, Joyner DC, Baek SM, Larsen JT, Andersen GL, et al. Application of a high-density oligonucleotide microarray approach to study bacterial population dynamics during uranium reduction and reoxidation. Appl. Environ. Microbiol. [Internet]. 2006;72:6288–98. Available from: http://dx.doi.org/10.1128/AEM.00246-06

3. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. [Internet]. 2016;17:333–51. Available from: http://dx.doi.org/10.1038/nrg.2016.49

4. Loman NJ, Watson M. Successful test launch for nanopore sequencing. Nat. Methods [Internet]. 2015;12:303–4. Available from: http://dx.doi.org/10.1038/nmeth.3327

5. Watson M. Illuminating the future of DNA sequencing. Genome Biol. [Internet]. 2014;15:108. Available from: http://dx.doi.org/10.1186/gb4165

6. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut microbiome viewed across age and geography. Nature [Internet]. 2012;486:222–7. Available from: http://dx.doi.org/10.1038/nature11053

7. Mackelprang R, Waldrop MP, DeAngelis KM, David MM, Chavarria KL, Blazewicz SJ, et al. Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. Nature [Internet]. 2011;480:368–71. Available from: http://dx.doi.org/10.1038/nature10576

8. Wallace RJ, Rooke JA, McKain N, Duthie C-A, Hyslop JJ, Ross DW, et al. The rumen microbial metagenome associated with high methane production in cattle. BMC Genomics [Internet]. 2015;16:839. Available from: http://dx.doi.org/10.1186/s12864-015-2032-0

9. Roehe R, Dewhurst RJ, Duthie C-A, Rooke JA, McKain N, Ross DW, et al. Bovine Host Genetic Variation Influences Rumen Microbial Methane Production with Best Selection Criterion for Low Methane Emitting and Efficiently Feed Converting Hosts Based on Metagenomic Gene Abundance. PLoS Genet. [Internet]. Public Library of Science; 2016 [cited 2017 Nov 25];12:e1005846. Available from: http://journals.plos.org/plosgenetics/article/file?id=10.1371/journal.pgen.1005846&type=printable

10. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, et al. Environmental genome shotgun sequencing of the Sargasso Sea. Science [Internet]. 2004;304:66–74. Available from: http://dx.doi.org/10.1126/science.1093857

11. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Ocean plankton. Structure and function of the global ocean microbiome. Science [Internet]. 2015;348:1261359. Available from: http://dx.doi.org/10.1126/science.1261359

12. Roumpeka DD, Wallace RJ, Escalettes F, Fotheringham I, Watson M. A Review of Bioinformatics Tools for Bio-Prospecting from Metagenomic Sequence Data. Front. Genet. [Internet]. 2017;8:23. Available from: http://dx.doi.org/10.3389/fgene.2017.00023

13. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat. Biotechnol. [Internet]. 2017;35:725–31. Available from: http://dx.doi.org/10.1038/nbt.3893

14. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature [Internet]. 2004;428:37–43. Available from: http://dx.doi.org/10.1038/nature02340

15. Stewart R, Auffret M, Warr A, Snelling T, Dewhurst R, Walker A, et al. Assembly of hundreds of microbial genomes from the cow rumen reveals novel microbial species encoding enzymes with roles in carbohydrate metabolism [Internet]. bioRxiv. 2017 [cited 2017 Nov 25]. p. 162578. Available from: https://www.biorxiv.org/content/early/2017/07/13/162578 (DOI: 10.1101/162578) http://dx.doi.org/10.1101/162578

16. Svartström O, Alneberg J, Terrapon N, Lombard V, de Bruijn I, Malmsten J, et al. Ninety-nine de novo assembled genomes from the moose (Alces alces) rumen microbiome provide new insights into microbial plant biomass degradation. ISME J. [Internet]. 2017;11:2538–51. Available from: http://dx.doi.org/10.1038/ismej.2017.108

17. Watson M. Hi-C genomes from "Assembly of hundreds of microbial genomes from the cow rumen reveals novel microbial species encoding enzymes with roles in carbohydrate metabolism" [Internet]. The Roslin Institute; 2017 [cited 2017 Nov 25]. Available from: http://hdl.handle.net/10283/2911 (DOI: 10.7488/ds/2222) http://dx.doi.org/10.7488/ds/2222 https://datashare.is.ed.ac.uk/handle/10283/2911

18. Watson M. Assembly of hundreds of microbial genomes from the cow rumen reveals novel microbial species encoding enzymes with roles in carbohydrate metabolism [Internet]. The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh; 2017 [cited 2017 Nov 25]. Available from: http://hdl.handle.net/10283/2772 (DOI: 10.7488/ds/2099) http://dx.doi.org/10.7488/ds/2099 https://datashare.is.ed.ac.uk/handle/10283/2772

19. Hu P, Tom L, Singh A, Thomas BC, Baker BJ, Piceno YM, et al. Genome-Resolved Metagenomic Analysis Reveals Roles for Candidate Phyla and Other Microbial Community Members in Biogeochemical Transformations in Oil Reservoirs. MBio [Internet]. 2016;7. Available from: http://mbio.asm.org/content/7/1/e01669-15.abstract

20. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, et al. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat. Commun. [Internet]. 2016;7:13219. Available from: http://dx.doi.org/10.1038/ncomms13219

21. Tully BJ, Sachdeva R, Graham ED, Heidelberg JF. 290 Metagenome-assembled Genomes from the Mediterranean Sea: Ongoing Effort to Generate Genomes from the Tara Oceans Dataset [Internet]. bioRxiv. 2016 [cited 2017 Nov 25]. p. 069484. Available from: https://www.biorxiv.org/content/early/2016/08/16/069484

22. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics [Internet]. Oxford University Press; 2012 [cited 2017 Nov 25];28:2520–2. Available from: https://academic.oup.com/bioinformatics/article-abstract/28/19/2520/290322

23. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nature Microbiology [Internet]. 2017;2:1533–42. Available from: https://doi.org/10.1038/s41564-017-0012-7

24. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. [Internet]. 2015;25:1043–55. Available from: http://dx.doi.org/10.1101/gr.186072.114

25. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics [Internet]. 2010;11:119. Available from: http://dx.doi.org/10.1186/1471-2105-11-119

26. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat. Methods [Internet]. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2014;12:59. Available from: http://dx.doi.org/10.1038/nmeth.3176

27. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. [Internet]. 2004;32:D115–9. Available from: http://dx.doi.org/10.1093/nar/gkh131

28. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, et al. The Pfam protein families database. Nucleic Acids Res. [Internet]. 2004;32:D138–41. Available from: http://dx.doi.org/10.1093/nar/gkh121

29. Segata N, Börnigen D, Morgan XC, Huttenhower C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat. Commun. [Internet]. 2013;4:2304. Available from: http://dx.doi.org/10.1038/ncomms3304

30. Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ [Internet]. 2015;3:e1029. Available from: http://dx.doi.org/10.7717/peerj.1029

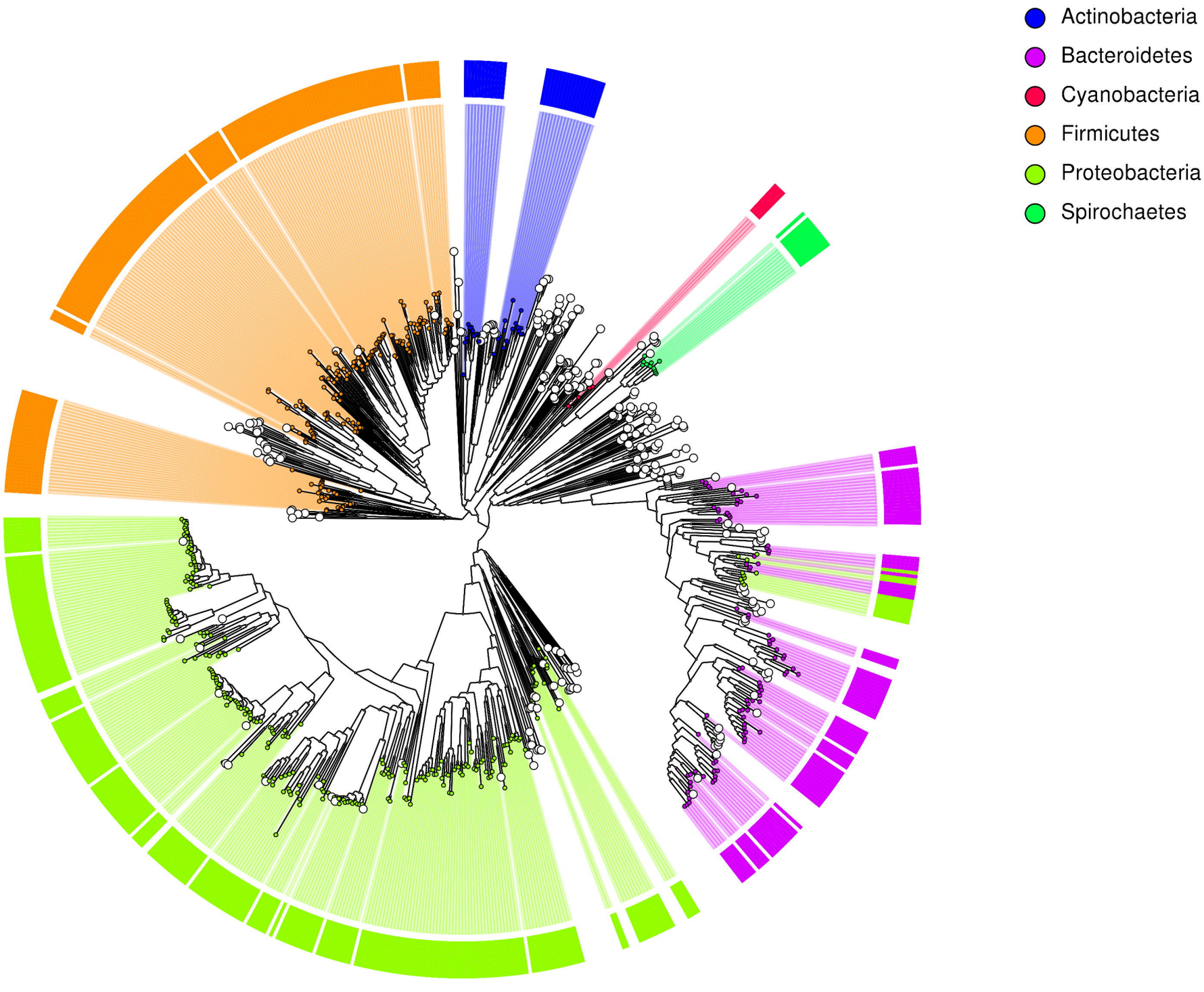
31. Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol. Biol. Evol. [Internet]. 2016;33:1635–8. Available from: http://dx.doi.org/10.1093/molbev/msw046

32. Brown BL, Watson M, Minot SS, Rivera MC, Franklin RB. MinION^{TM} nanopore sequencing of environmental metagenomes: a synthetic approach. Gigascience [Internet]. 2017;6:1–10. Available from: http://dx.doi.org/10.1093/gigascience/gix007

33. Gallagher LA, Coughlan S, Black NS, Lalor P, Waters EM, Wee B, et al. Tandem Amplification of the Staphylococcal Cassette Chromosome mec Element Can Drive High-Level Methicillin Resistance in Methicillin-Resistant Staphylococcus aureus. Antimicrob. Agents Chemother. [Internet]. 2017;61. Available from: http://dx.doi.org/10.1128/AAC.00869-17

34. Risse J, Thomson M, Patrick S, Blakely G, Koutsovoulos G, Blaxter M, et al. A single chromosome assembly of Bacteroides fragilis strain BE1 from Illumina and MinION nanopore sequencing data. Gigascience [Internet]. 2015;4:60. Available from: https://doi.org/10.1186/s13742-015-0101-6

35. Watson M. When will “open science” become simply “science”? Genome Biol. [Internet]. 2015;16:101. Available from: http://dx.doi.org/10.1186/s13059-015-0669-2


Actinobacteria
Bacteroidetes
Cyanobacteria
Firmicutes
Proteobacteria
Spirochaetes