

Genome Analysis

ExplorePipolin: reconstruction and annotation of piPolB-encoding bacterial mobile elements from draft genomes

L. Chuprikova^{1,†}, V. Mateo-Cáceres¹, M. de Toro² and M. Redrejo-Rodríguez^{1,*}

¹Department of Biochemistry, School of Medicine, Universidad Autónoma de Madrid and Instituto de Investigaciones Biomédicas ‘Alberto Sols’ (UAM-CSIC), Madrid, Spain and ²Plataforma de Genómica y Bioinformática, CIBIR (Centro de Investigación Biomédica de La Rioja), Logroño, La Rioja 26006, Spain

*To whom correspondence should be addressed.

[†]Present address: Division of Virus-associated Carcinogenesis, German Cancer Research Center (DKFZ), Heidelberg, Germany

Associate Editor: Cecilia Arighi

Received on June 20, 2022; revised on July 18, 2022; editorial decision on July 23, 2022; accepted on August 6, 2022

Abstract

Motivation: Detailed and accurate analysis of mobile genetic elements (MGEs) in bacteria is essential to deal with the current threat of multiresistant microbes. The overwhelming use of draft, contig-based genomes hinder the delineation of the genetic structure of these plastic and variable genomic stretches, as in the case of pipolins, a superfamily of MGEs that spans diverse integrative and plasmidic elements, characterized by the presence of a primer-independent DNA polymerase.

Results: ExplorePipolin is a Python-based pipeline that screens for the presence of the element and performs its reconstruction and annotation. The pipeline can be used on virtually any genome from diverse organisms and of diverse quality, obtaining the highest-scored possible structure and reconstructed out of different contigs if necessary. Then, predicted pipolin boundaries and pipolin encoded genes are subsequently annotated using a custom database, returning the standard file formats suitable for comparative genomics of this mobile element.

Availability and implementation: All code is available and can be accessed here: github.com/pipolinlab/ExplorePipolin.

Contact: modesto.redrejo@uam.es

Supplementary information: [Supplementary data](#) are available at *Bioinformatics Advances* online.

1. Introduction

Infectious diseases caused by multidrug-resistant pathogens pose a major threat to global health, food security and development today, a situation aggravated by the insufficient production of new drugs (Murray *et al.*, 2022; Zhu *et al.*, 2021). Antimicrobial resistance (AMR) and virulence factors are largely encoded by itinerant nucleic acid fragments known as mobile genetic elements (MGEs) that act as a platform for their transference and spreading (Partridge *et al.*, 2018; von Wintersdorff *et al.*, 2016). Thus, MGEs constitute an essential part of the bacterial genome and can contribute to their evolution, fitness and pathogenicity. However, despite recent efforts, many MGEs are overlooked or misannotated in population genomics analyses, and they are generally underestimated (Hua *et al.*, 2021; Oliveira Alvarenga *et al.*, 2018; Ross *et al.*, 2021).

The recent accessibility of high-throughput sequencing methods as part of surveillance programs of bacterial pathogens allows the genomic and metagenomic monitoring of the expansion of bacterial strain-specific markers, including virulence and AMR genes.

However, these fast-evolving methods also generated a large amount of data that must be thoroughly processed and analyzed (Mitchell and Simmer, 2019). This is particularly problematic regarding the study of dynamics and plasticity of MGEs, as they can range in size from very simple and small elements, such as insertion elements (IS), coding for only the transposase necessary for their relocation, to large prophages, transposons and plasmids, which can be tens or hundreds of kilobase pairs in length and also can interact among themselves, as has been recently highlighted by the widespread presence of defense mechanisms (Benler *et al.*, 2021; Durrant *et al.*, 2020; Partridge *et al.*, 2018). Further, MGEs prediction and analysis is hindered by their great modularity and rapid evolution through gene acquisition and gene loss. Many pipelines designed for the analysis of MGEs are specialized and rely on the identification of hallmark genes, like plasmid replication proteins (Carattoli and Hasman, 2020), relaxases (Alvarado *et al.*, 2012) or specific transposase or recombinases for integrative elements (Cury *et al.*, 2016, 2017; Moura *et al.*, 2009; Ross *et al.*, 2021; Siguier *et al.*, 2012). Some works have focused on the use of high-quality reference

genomes but at the cost of diversity loss (Jiang et al., 2019). Similarly, a few recent methods for annotation of prophages are able to identify insertion boundaries on complete genomes (Arndt et al., 2019; Guo et al., 2021). However, the great majority of currently available genomes are based in short sequencing reads technologies, resulting in draft genomes, made up of tens or hundreds of heterogeneous contigs. This makes accurate analysis of the genetic structure and dynamics of integrated elements virtually impossible (Arredondo-Alonso et al., 2017). The usage of raw reads to scaffold integrative elements has been successfully applied to reconstruct some elements (Durrant et al., 2020), though it entails large computational resources, hindering its application for massive screenings. More recently, the conserved sequences of the transposase binding sites and their unique architecture are shown to carry a signal that is sufficient to delineate the 5' and 3' boundaries of Tn7-like elements (Benler et al., 2021).

Pipolins are a novel group of MGEs found in diverse major bacterial phyla but also mitochondria (Redrejo-Rodríguez et al., 2017). The only gene shared by all pipolins encodes for a subgroup of DNA polymerases from family B, named piPolB (for *primer-independent PolB*). This relates pipolins to other elements commonly referred to as ‘self-synthesizing’ by the presence of a PolB, the eukaryotic polintons and casposons, detected in archaea and some bacteria (Kapitonov and Jurka, 2006; Krupovic and Koonin, 2016; Krupovic et al., 2014). However, the biological role of the piPolB is still unclear; although a putative function in host genome damage tolerance/repair has been shown (Redrejo-Rodríguez et al., 2017).

Pipolins in *Escherichia coli* are diverse and plastic but, besides the piPolB gene, they share conserved att-like direct repeats (hereafter *atts*), overlapping with a tRNA gene. We previously took advantage of that common structural features to delineate their structure out of contig-scale *E. coli* draft genomes using a preliminary customized Python pipeline (Chuprikova, 2020). This allowed us to confirm the common presence of tyrosine recombinases in *E. coli* pipolins that further supported an att-mediated excision/integration mechanism, which still needs to be confirmed *in vivo*. This highlighted the genetic diversity of pipolins, sometimes increased by the presence of additional overlapping transposases and insertion sequences (Flament-Simon et al., 2020). However, beyond *E. coli*, the att sequences are not conserved and indeed, pipolins have been identified as circular plasmids without direct repeats in several fungi mitochondria as well as in a number of bacteria, spanning very diverse species, like *Enterobacter hormaechei*, *Staphylococcus epidermidis* or *Lactobacillus fermentum*. Accordingly, plasmidic pipolins harbor relaxases and resolvases and they often have lost the integrase gene (Redrejo-Rodríguez et al., 2017).

In this work, we now present ExplorePipolin, a full-fledged Python-based pipeline, that allows robust reconstruction and homogeneous characterization of pipolins from diverse bacterial genomes. The pipeline analyzes each contig for the presence of a pipolin marker and proposes the highest-scored possible structure, reconstructed out of different contigs if necessary. Then, the element is delimited by predicted *atts* and pipolin encoded genes are subsequently annotated using a custom database. The workflow is publicly available and can be installed from its repository on the GitHub or via Conda. It aims to be a documented, composable and reusable application (Brack et al., 2022), easily translatable to other elements, such as integrative and conjugative elements (ICEs) or casposons, among others.

2. Methods

2.1 Pipeline outline

As pipolins are defined by the presence of a piPolB gene (Redrejo-Rodríguez et al., 2017), ExplorePipolin pipeline workflow (Fig. 1) must start by searching the (multi)fasta query sequence for a piPolB coding sequence. To achieve high sensitivity and accuracy in this limiting step, we updated the previous dataset of diverse piPolBs (Redrejo-Rodríguez et al., 2017) using Jackhmmer searches on several databases (Potter et al., 2018) and generated an HMM profile spanning all possible diversity. This profile was then used to scan for

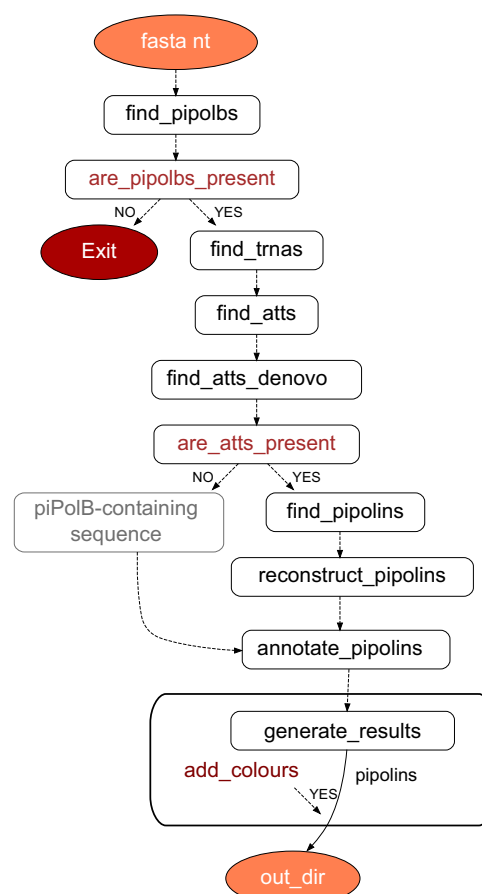
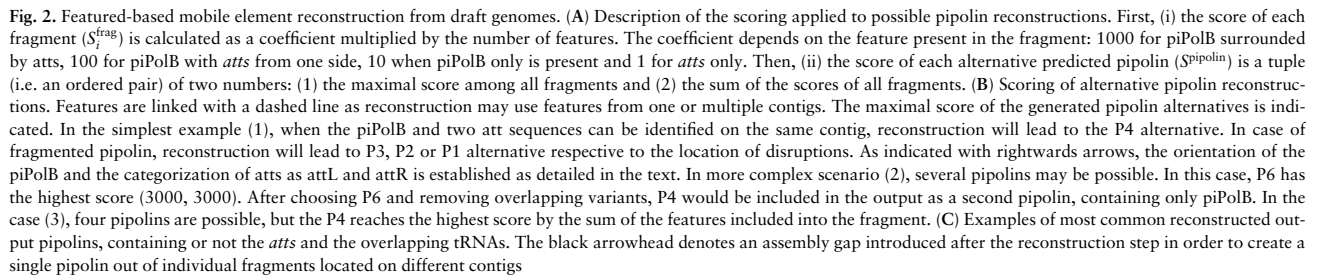


Fig. 1. Workflow of ExplorePipolin. Main nodes and tasks are indicated. Dataflow throughout the pipeline is managed with Prefect (<https://www.prefect.io/>)

piPolBs in the translated Fasta nucleotide sequences using as cutoff an E-value of 10^{-50} , selected to detect all known piPolBs and reduce the false positives hits from sequences that belong to pPolBs and other related B-family DNA polymerase groups (Kazlauskas et al., 2020).

When a piPolB-containing fragment is identified, DNA sequences are subsequently scanned for the pipolin boundaries. First, as pipolins often integrate into tRNA genes, we search for nearby tRNA or tmRNA gene, using the Aragorn tool (Laslett and Canback, 2004) and that will be as a verification step for the identification of the att-like direct repeats. In *E. coli*, pipolin *atts* are highly conserved among diverse strains (Flament-Simon et al., 2020), which allow their quick identification with BLAST using the sequence of the attR from *E. coli* strain 3-373-03_S1_C2 pipolin (Redrejo-Rodríguez et al., 2017; Richter et al., 2018). Alternatively, when known att sequences are not found, we search for ‘*de novo*’ direct repeats that might participate in the pipolin mobilization. Briefly, we first look for short direct repeats around a piPolB gene using BLAST (85% identity with word size 6, plus strand only to exclude inverted repeats). This strategy is also implemented in other tools, as Phaster (Arndt et al., 2019). All the predicted repeats are annotated and situations with an unexpected number or repeats (i.e. one or more than two) are handled during the element reconstruction (see below). To facilitate the comparison among diverse pipolin genomic structures, we established that whenever an att overlapped with a tRNA, that will be the attR and the other repetition will be denoted as attL. Importantly, unsuccessful repeats identification will not lead to the pipeline failure. Notwithstanding technical issues, as mentioned above, pipolins have been also identified as circular plasmids. Thus, in the absence of known or predicted repetitions, the boundaries of the pipolin are arbitrarily settled in a maximum of 30 kb at each side of the piPolB gene (Fig. 2C, bottom pipolin), which may be modified by the user (option `–max-inflate`),



approach, ExplorePipolin could be used for a wide diversity of pipolin structures, from the already studied ones in *E.coli* (Flament-Simon *et al.*, 2020) to a broad range present in other bacterial genomes (see below).

3. Results

The pipolin reconstructed structure(s) are subsequently annotated using a custom Prokka-based pipeline that includes previously known pipolin-related genes. The output includes files in standard formats GBK and GFF generated for wide compatibility for downstream analysis of pipolins genetic structure. An additional file,

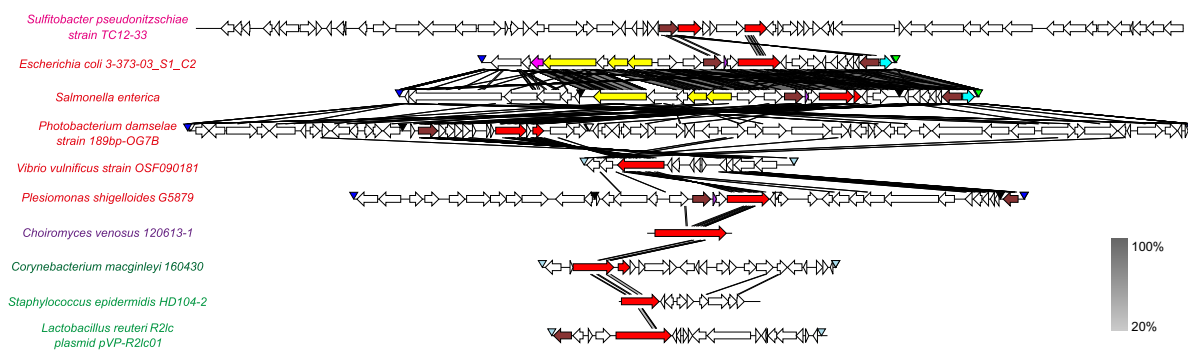


Fig. 3. Genetic structure of diverse pipolins from genomes from a wide range of bacteria reconstructed by ExplorePipolin. Predicted protein-coding genes are represented by arrows, indicating the direction of transcription and more common genes are colored following Prokka annotation (piPolB in red, tyrosine recombinase in brown, UDG in cyan, excisionase in purple and metallohydrolase in magenta). When detected, other features are indicated as colored arrowheads: sequence gaps (black), *E. coli* related-atts (navy blue), *de novo* detected direct repeats (steel) and tRNAs (green). The grayscale on the right reflects the percent of amino acid identity between pairs of sequences. The image was generated by EasyFig software using tBlastX for elements comparison. Selected genome assemblies were downloaded from GenBank database (IDs GCA_020905835.1, GCA_000700265.1, GCA_917083495.1, GCA_013377875.1, GCA_015790535.1, GCA_009183495.1, GCA_003788595.2, GCA_016628745.1, GCA_016859455.1 and GCA_003703875.1). Names of the analyzed genomes are indicated on the left and colored by taxonomy: magenta, Alphaproteobacteria; red, Gammaproteobacteria; purple, fungi, forest green, Actinobacteria; and green, Firmicutes

appended with the suffix ‘single_record’, contains a final reconstructed pipolin comprising individual pipolin fragments from different contigs joined by an assembly gap feature type. The annotated *att* direct repeats are also included in the output files and labeled as ‘pipolin conserved’ or ‘*de novo*’, according to their relationship to known *att*-like sequences of *E. coli* pipolins or identified as a novel direct repeat fragment.

Moreover, we also include color information of common pipolin features (i.e. piPolB, *att*, tyrosine recombinases, restriction-modification systems, uracil DNA glycosylase...) for straightforward representation with EasyFig (Sullivan et al., 2011). Default colors from the previously used scheme can be modified providing a TSV file (option `-colors`) or removed (`-skip-colors`). We found EasyFig a simple customizable visualization alternative, suitable for comparison of genetic structure and diversity of pipolins and convenient for non-bioinformaticians. We have also tested other recent related applications such as the web-based cliniker (Gilchrist and Chooi, 2021) or the R package gggenomes (Hackl and Ankenbrand, 2022). The latter seemed enough versatile for pipolins representation (Supplementary Fig. S1). We have included in the GitHub repository scripts for plotting ExplorePipolin output in R using the gggenomes. In short, a custom Python script adapts the GBK files with some required modifications and calculates pipolin synteny and GC content according to the gggenomes recipe. Then, an R script uses these data to plot the sequences and the desired features. It is only necessary the final pipolin GBK files (‘single_record’), a file specifying the order of the pipolins to compare in the representation, and the software requirements specified in the README file.

3.2 Examples of use

We have successfully applied the pipeline to diverse genomic sequences. Complete analysis of a medium size bacterial draft assemblies (5–6 MB) takes approximately 30–60 s per genome using multiple threads for annotation (option `-cpus`). Further optimization may be achieved using additional parallelization strategies, like GNU Parallel (Tange, 2018). Moreover, most time-consuming step is the annotation of the reconstructed elements, and it can be passed up (option `-skip-annotation`) for saving time, for example, when large screening of thousands of genomes is done.

Notwithstanding a throughout modification of the original pipeline, ExplorePipolin analyzed the genomes of our laboratory *E. coli*-harboring pipolins collection obtaining nearly identical pipolin structures than in the original analysis (Flament-Simon et al., 2020) (Supplementary Fig. S1).

Furthermore, the new reconstruction strategy allows successful reconstruction and analysis of pipolins from a wide range of distant organisms, spanning different bacteria phyla and Fungi genomes (Fig. 3). The usage of HMM profile searches gives rise to a highly sensitive and specific identification of piPolB coding sequences that

are subsequently delineated and annotated, regardless of the presence of the known *att* sequences or any direct repeat at all.

4. Conclusion

ExplorePipolin is an example of a comprehensive pipeline that not only looks for MGE presence by marker screening but also reconstructs the whole pipolin element structure and attempt to provide precise boundaries, making further analysis more accurate and straightforward. Importantly, it can be widely used to screen for pipolins on virtually any genome of a wide range of organisms and diverse assembly qualities and it returns putative reconstructed pipolins with homogenous annotation suitable for comparative genomics of this mobile element.

Furthermore, the modular structure of the pipeline will facilitate its future update for bulk analysis of other elements with one or few universal hallmark features, particularly those related with pipolins, such as casposons or polintons.

Acknowledgements

We thank Mario R. Mestre for the generation of the piPolBs HMM profile. We are also grateful to members of the MR-R lab for discussions and suggestions.

Funding

This work has been supported by Grant PGC2018-093723-A-I00 MCIN/AEI/10.13039/501100011033/ and FEDER ‘A way to make Europe’.

Conflict of Interest: none declared.

References

- Alvarado, A. et al. (2012) A degenerate primer MOB typing (DPMT) method to classify Gamma-Proteobacterial plasmids in clinical and environmental settings. *PLoS One*, 7, e40438.
- Arndt, D. et al. (2019) PHAST, PHASTER and PHASTEST: tools for finding prophage in bacterial genomes. *Brief. Bioinform.*, 20, 1560–1567.
- Arredondo-Alonso, S. et al. (2017) On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microb. Genom.*, 3, e000128.
- Benler, S. et al. (2021) Cargo genes of Tn7-like transposons comprise an enormous diversity of defense systems, mobile genetic elements, and antibiotic resistance genes. *mBio*, 12, e0293821.
- Brack, P. et al. (2022) Ten simple rules for making a software tool workflow-ready. *PLoS Comput. Biol.*, 18, e1009823.

- Carattoli, A. and Hasman, H. (2020) PlasmidFinder and in silico pMLST: identification and typing of plasmid replicons in whole-genome sequencing (WGS). *Methods Mol. Biol.*, **2075**, 285–294.
- Chuprikova, L. (2020) ExplorePipolin: a pipeline for identification and exploration of pipolins, novel mobile genetic elements widespread among bacteria. Universidad Autónoma de Madrid. Master's Thesis. URI: hdl.handle.net/10486/692516.
- Cury, J. *et al.* (2016) Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res.*, **44**, 4539–4550.
- Cury, J. *et al.* (2017) Integrative and conjugative elements and their hosts: composition, distribution and organization. *Nucleic Acids Res.*, **45**, 8943–8956.
- Durrant, M.G. *et al.* (2020) A bioinformatic analysis of integrative mobile genetic elements highlights their role in bacterial adaptation. *Cell Host Microbe*, **27**, 140–153.e9.
- Flament-Simon, S.-C. *et al.* (2020) High diversity and variability of pipolins among a wide range of pathogenic *Escherichia coli* strains. *Sci. Rep.*, **10**, 12452.
- Gilchrist, C.L.M. and Chooi, Y.-H. (2021) Clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics*, **37**, 2473–2475.
- Guo, J. *et al.* (2021) VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome*, **9**, 37.
- Hackl, T. and Ankenbrand, M. (2022) gggenomes: a grammar of graphics for comparative genomics. <https://github.com/thackl/gggenomes> (17 July 2022, date last accessed).
- Hua, X. *et al.* (2021) BacAnt: a combination annotation server for bacterial DNA sequences to identify antibiotic resistance genes, integrons, and transposable elements. *Front. Microbiol.*, **12**, 649969.
- Jiang, X. *et al.* (2019) Comprehensive analysis of chromosomal mobile genetic elements in the gut microbiome reveals phylum-level niche-adaptive gene pools. *PLoS One*, **14**, e0223680.
- Kapitonov, V.V. and Jurka, J. (2006) Self-synthesizing DNA transposons in eukaryotes. *Proc. Natl. Acad. Sci. USA*, **103**, 4540–4545.
- Kazlauskas, D. *et al.* (2020) Diversity and evolution of B-family DNA polymerases. *Nucleic Acids Res.*, **48**, 10142–10156.
- Krupovic, M. *et al.* (2014) Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. *BMC Biol.*, **12**, 36.
- Krupovic, M. and Koonin, E.V. (2016) Self-synthesizing transposons: unexpected key players in the evolution of viruses and defense systems. *Curr. Opin. Microbiol.*, **31**, 25–33.
- Laslett, D. and Canback, B. (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.*, **32**, 11–16.
- Mitchell, S.L. and Simner, P.J. (2019) Next-generation sequencing in clinical microbiology: are we there yet? *Clin. Lab. Med.*, **39**, 405–418.
- Moura, A. *et al.* (2009) INTEGRALL: a database and search engine for integrons, integrases and gene cassettes. *Bioinformatics*, **25**, 1096–1098.
- Murray, C.J. *et al.* (2022) Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet*, **399**, 629–655.
- Oliveira Alvarenga, D. *et al.* (2018) A practical guide for comparative genomics of mobile genetic elements in prokaryotic genomes. *Methods Mol. Biol.*, **1704**, 213–242.
- Partridge, S.R. *et al.* (2018) Mobile genetic elements associated with antimicrobial resistance. *Clin. Microbiol. Rev.*, **31**, e00088–17.
- Potter, S.C. *et al.* (2018) HMMER web server: 2018 update. *Nucleic Acids Res.*, **46**, W200–W204.
- Redrejo-Rodríguez, M. *et al.* (2017) Primer-independent DNA synthesis by a family B DNA polymerase from self-replicating mobile genetic elements. *Cell Rep.*, **21**, 1574–1587.
- Richter, T.K.S. *et al.* (2018) Temporal variability of *Escherichia coli* diversity in the gastrointestinal tracts of Tanzanian children with and without exposure to antibiotics. *mSphere*, **3**, e00558–18.
- Ross, K. *et al.* (2021) TnCentral: a prokaryotic transposable element database and web portal for transposon analysis. *mBio*, **12**, e0206021.
- Siguier, P. *et al.* (2012) Exploring bacterial insertion sequences with ISfinder: objectives, uses, and future developments. *Methods Mol. Biol.*, **859**, 91–103.
- Sullivan, M.J. *et al.* (2011) Easyfig: a genome comparison visualizer. *Bioinformatics*, **27**, 1009–1010.
- Tange, O. (2018) GNU Parallel. doi: [10.5281/zenodo.1146014](https://doi.org/10.5281/zenodo.1146014).
- von Wintersdorff, C.J. *et al.* (2016) Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Front. Microbiol.*, **7**, 173.
- Zhu, M. *et al.* (2021) The future of antibiotics begins with discovering new combinations. *Ann. N. Y. Acad. Sci.* **1496**, 82–96.