Published online 12 May 2010 *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* ***W647–W651***
doi:10.1093/nar/gkq365

BAGEL2: mining for bacteriocins in genomic data

Anne de Jong[1], Auke J. van Heel[1], Jan Kok[1] and Oscar P. Kuipers[1,2,*]

[1]Department of Molecular Genetics, University of Groningen, Groningen Biomolecular Sciences and Biotechnology Institute, 9750 AA Haren and [2]Kluyver Center for Genomics of Industrial Fermentation, Delft/Groningen, The Netherlands

Received January 29, 2010; Revised April 9, 2010; Accepted April 24, 2010

ABSTRACT

Mining bacterial genomes for bacteriocins is a challenging task due to the substantial structure and sequence diversity, and generally small sizes, of these antimicrobial peptides. Major progress in the research of antimicrobial peptides and the ever-increasing quantities of genomic data, varying from (un)finished genomes to meta-genomic data, led us to develop the significantly improved genome mining software BAGEL2, as a follow-up of our previous BAGEL software. BAGEL2 identifies putative bacteriocins on the basis of conserved domains, physical properties and the presence of biosynthesis, transport and immunity genes in their genomic context. The software supports parameter-free, class-specific mining and has high-throughput capabilities. Besides building an expert validated bacteriocin database, we describe the development of novel Hidden Markov Models (HMMs) and the interpretation of combinations of HMMs via simple decision rules for prediction of bacteriocin (sub-)classes. Furthermore, the genetic context is automatically annotated based on (combinations of) PFAM domains and databases of known context genes. The scoring system was fine-tuned using expert knowledge on data derived from screening all bacterial genomes currently available at the NCBI. BAGEL2 is freely accessible at http://bagel2.molgenrug.nl.

INTRODUCTION

Bacteriocins are ribosomally synthesized antimicrobial peptides produced by bacteria. These compounds are of high interest to researchers in biotechnology and medicine, because of numerous (potential) applications based on their potent antimicrobial activity. Next to application as food preservatives (1), bacteriocins can be used for the development of novel antibiotics (2). Bacteriocin genes frequently have genes encoding proteins involved in bacteriocin regulation, self-immunity, transport and modification in their close genomic proximity. Discovery of new bacteriocins is shifting from classical screening for activity towards *in silico* analysis of genomic data. This is a challenging task due to the small size of bacteriocins and the fact that small ORFs are often omitted from annotations, especially when no homology to known proteins is found. Thorough bacteriocin prediction methods should therefore not only exploit homology to known bacteriocins and motifs therein, but they should also take into account the genomic context of the putative bacteriocin. To ensure that no putative bacteriocins are missed, an ORF calling procedure focusing on small ORFs should be employed. The only automated tool for discovery of bacteriocins until now is BAGEL (3), which has been used successfully in several studies (4–6). Here, we present the significantly improved BActeriocin GEnome mining tooL 2 (BAGEL2; Supplementary Table S1). Next to implementing current knowledge from the rapidly advancing molecular research in bacteriocins, the software now allows high-throughput screening to cope with the ever-increasing size of genomic data sets. This new feature of BAGEL2 enabled us to use data derived from a large input data set (>1000 genomes) to further optimize bacteriocin identification.

IMPLEMENTATION

Improvement of the software

Major improvements of BAGEL2 are the extended use of Hidden Markov Models (HMMs) and of the manually curated databases of known bacteriocins and context genes (encoding proteins for modification, immunity/transport and two component systems; 7–10). Furthermore, regular expressions (Supplementary Table S2) are used for improved recognition of bacteriocins. An advanced classification algorithm has been implemented that predicts sub-classes according to the

*To whom correspondence should be addressed. Tel: +31 50 3632093; Fax: +31(0)503632348; Email: o.p.kuipers@rug.nl

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

updated classification scheme (1) (Table 1). Due to many changes in BAGEL2 compared to the first version, BAGEL (3), among which the possibility for high-throughput screening, a new scoring algorithm has been implemented. A schematic overview of the program is presented in Figure 1. Finally, the web server now offers the possibility to BLAST peptides against the bacteriocin database and offers an improved ORF calling system.

Classification

One of the greatly improved features of BAGEL2 is the (sub-)classification of putative bacteriocins and, thus, the class-specific mining of genomes for bacteriocins. For BAGEL2, the classification scheme proposed by Cotter *et al*. (1) is implemented, with addition of the arguable Class III (for complete coverage), which contains relatively large antimicrobial proteins. Class I contains the lantibiotics; sub-classification of this class is according to Wiley *et al*. (11): IA, LanBC-modified; IB, LanM-modified; and IC, LanL-modified (12). For identification of sub-class IC through context screening an alignment of LanL was used to identify a motif for LanL encoding genes. Class II bacteriocins are sub-classified (A, B, C and D) according to Cotter *et al*. (1). Classification of bacteriocins in BAGEL2 is based on homology to known bacteriocins, presence of motifs, context genes and properties related to the specific class. BAGEL2 creates separate candidate lists for each class by assigning only class-relevant points. The sub-class is predicted by a simple interpretation of sub-class-specific features registered by BAGEL2.

Scoring algorithm

Eleven rules were implemented to add a specific weight factor (Table 2) to each putative bacteriocin: (i) based on a BLAST search against the bacteriocin database, high homology to a known bacteriocin Class I or Class II adds a weight factor, where a low homology adds half the weight factor; (ii) for bacteriocin of Class III, one threshold value is used; (iii) a match with the regular expression database; (iv) for each motif in the HMM database a weight factor is added on the basis of specificity and *P*-value cut-off in the HMM search (Supplementary Table S3); (v) presence of a processing site at the proper position in the bacteriocin; (vi) presence of cysteine in combination with serine/threonine; (vii) properties: iso-electric point, protein size and charged residues; (viii) presence of a biosynthesis or immunity gene in the genomic context on the basis of a BLAT hit in the database; (ix) presence of an ABC transporter, histidine kinase or a C39 type protease in the genomic context, using HMM for function detecting; (x) presence of genes for modification systems in the genomic context (see Supplementary Data); (xi) highly unlikely candidates (HUC) are masked using a BLAST search against the HUC database (only for Class III). These rules result in a score for each candidate bacteriocin. Candidates with $\geq$1800 points (threshold value) are considered 'putative bacteriocins', while candidates with a score below the threshold value but with a score of $\geq$1000 are considered 'interesting candidates'. The web server only displays the members of the groups of putative bacteriocins and interesting candidates.

Table 1. Classification scheme used by BAGEL2

Class I lanthionine	Class II non-lanthionine	Class III
A LanBC modified	A Pediocin-like	Large proteins
B LanM modified	B Two-component	
C LanL modified	C Cyclic peptides	
	D Miscellaneous	

Extended resources

New finished genomes are downloaded daily from NCBI. Unfinished genome sequences were retrieved from NCBI genome projects (ftp://ftp.ncbi.nih.gov/genbank/wgs/). To collect microbial genomes, only a list was generated on the NCBI website (www.ncbi.nlm.nih.gov/genomes/lproks.cgi) to determine which genomes should be downloaded. Subsequently, genomic data was downloaded selectively with an FTP client. A collection of plasmid sequences was downloaded from NCBI genomes by searching for 'replication type:plasmid' and limiting the results to bacterial data.

High-throughput screening

The BAGEL2 web server is designed for handling one data file per run. For high-throughput screening of multiple (unfinished) genomes, we use a local stand-alone version of BAGEL2. The output can be either tab-delimited text or html and is stored in a MySQL database. A typical search for bacteriocin genes in all genomes available at the NCBI takes 12–16 h on our server.

Validation of BAGEL2

The software was validated by close examination of the results obtained from screening all prokaryotic genomes available at NCBI: 1140 genomes and 738 plasmids. This resulted in 150 and 287 putative bacteriocins for Class I and Class II, respectively. Examining the original genomic annotation showed that 49% of the proteins identified by BAGEL2 were already known as (putative) bacteriocins and 48% were annotated as hypothetical proteins with unknown function. A closer inspection of some of these hypothetical proteins and their genomic contexts revealed that these are indeed putative bacteriocins, based on our expert knowledge. Also, BAGEL2 predicted the presence of a Class I bacteriocin in *Bacillus licheniformis* ATCC 14580, locus tag BL05375, which were recently proven to constitute the biologically active two component lantibiotic lichenicidin (13). Of the putative bacteriocins predicted by BAGEL2, 3% could be identified as false as they had an annotation that was clearly not bacteriocin-related. In the screen of 1140 genomes and 738 plasmids, 10 genes that were originally annotated as specifying (putative) bacteriocins were only annotated as interesting candidates by BAGEL2.

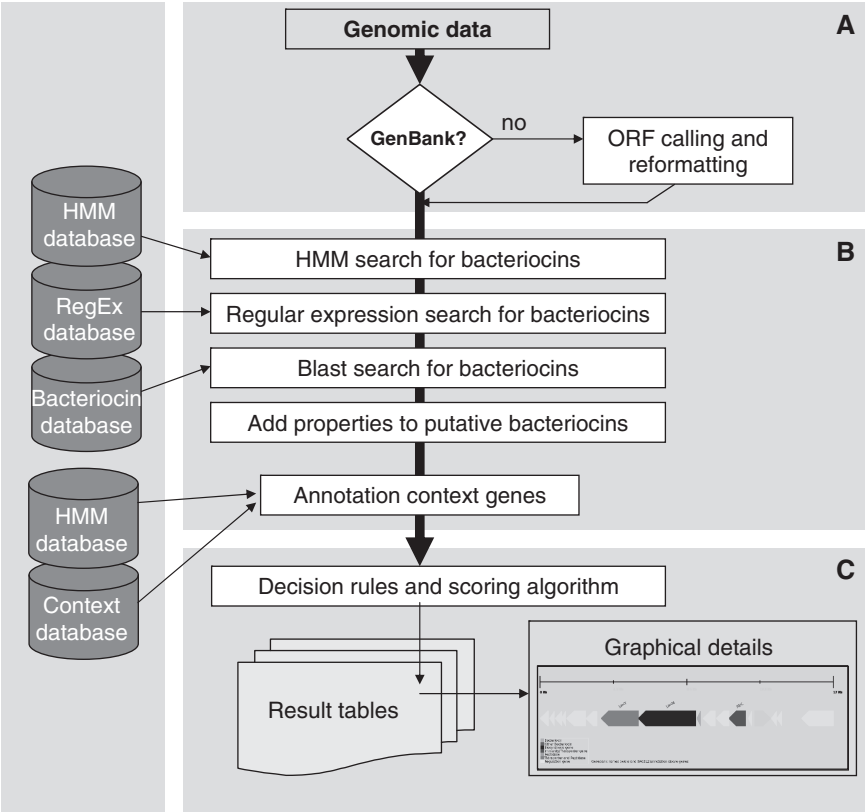

Figure 1. Process overview. (**A**) The input genome data can be a single or multi-entry GenBank file, in case of non-annotated data or, if re-annotation is desired, an ORF calling can be performed via the BAGEL2 web site. (**B**) Annotation of putative bacteriocins and their genomic context genes. (**C**) Calculation of score for each candidate and generation of detailed reports (including graphical representation).

Table 2. Weight factors used by the BAGEL2 scoring system

Property	Weight factor
Blast hit with bacteriocin Class I or II, stringent cut-off	10 000
Blast hit with bacteriocin Class I or II, non-stringent cut-off	5000
Blast hit with bacteriocin Class III	2000
[Cys]:[Thr,Ser] ratio 0.35: 0.80 and leader present	600
Bacteriocin regular expression match	500
Cysteine count between 2 and 8	400
PF05147 (1000 − 100 * distance)	200–900
PF03412 PF00005→LanT	300
PF04737 PF04738→LanB	300
PF00069 PF05147→LanL	300
Blast hit with context biosynthesis genes, Class I or II	200
Blast hit with context immunity genes, Class I or II	200
[Cys]:[Thr,Ser] ratio 0.25:0.55 and no leader	200
Presence of a leader processing site	100
HMM hit for context genes	100
PF00072 PF00486→Response	100
PF00512 PF02518→Sensor	100
Proper pI and charge	50
HMM hit for bacteriocin	see Supplementary Table S3A

System requirement and the web interface

BAGEL2 runs on a Linux platform (SentOS; http://www.centos.org/) with Apache web server (version2.2), MySQL server (version 5.1), PHP 5.0 (http://www.php.net/), Perl 5.10 (http://www.perl.org/) and BioPerl 1.6.0 (http://www.bioperl.org/). Furthermore, the following softwares were used: BLAST 2.2.9 (14); BLAT (15); HMMsearch (HMMER3; http://hmmer.janelia.org/); Glimmer v2.13 (16); RBSfinder (http://www.tigr.org/software/genefinding.shtml); and Prodigal v1.20 (http://compbio.ornl.gov/prodigal/; publication in progress). We implemented the recently released HMMER3, which uses a different but much faster algorithm than HMMER2, mainly to keep the search time within 1 min. The web interface consists of two separate sections: (i) one for uploading or selecting of an already available GenBank file. This section also displays the status of the run; and (ii) a result section with a condensed table of top hits found by BAGEL2 and links to detailed reports of each individual putative bacteriocin and interesting candidates, including a graphical presentation of the gene cluster.

Availability

The BAGEL2 web server is freely accessible at http://bagel2.molgenrug.nl. Next to bacteriocin mining, the

website offers ORF prediction tools and a BLAST server that can be used to BLAST search against the bacteriocin database. High-throughput data analysis can be performed upon request.

RESULTS AND DISCUSSION

Prodigal ORF prediction improves bacteriocin discovery

Bacterial genomes are mostly annotated using Glimmer as ORF calling algorithm. Recently, Prodigal ORF prediction also became available at NCBI. For ORF prediction in prokaryotic genomic data, we compared the performance of Glimmer and Prodigal on four very well-annotated Gram positive bacterial genomes, i.e. *B. subtilis* 168, *Lactobacillus plantarum* WCFS1, *Streptococcus pneumoniae* TIGR4 and *Enterococcus faecalis* V583. Glimmer predicted more smaller ORFs (<360 bp) than Prodigal, but the Prodigal re-annotation of these genomes contained all the bacteriocins annotated in the original data whereas the Glimmer predictions lacked two. Moreover, re-annotation of several genomes using Prodigal instead of Glimmer resulted in more novel ORFs encoding putative bacteriocins (data not shown).

Identification and benchmarking of novel bacteriocin motifs

For classification purposes, we identified motifs in several sub-classes/groups of lantibiotics, which also proved to be useful in the identification of new potential lantibiotics. The motifs were obtained using the amino acid sequences of known lantibiotics (7) as input for MEME (17). Subsequently, the obtained motifs were checked by MAST (18) against the UniProtKB database and against a database of proteins created for validation purposes, which contained random *Lactococcus lactis* MG1363 genes and all known bacteriocins. The added value of these motifs is demonstrated by the identification of relevant gene clusters (Figure 2; Table 3; Supplementary Table S4). From 10 novel motifs, which were found to be discriminating for bacteriocins and useful for classification, an HMM was built (Supplementary Table S3A). Additionally, one motif based on the lacticin 481 group described in literature was implemented (19). Furthermore, regular expressions were deduced from these new motifs (Supplementary Table S2B).

Analysis of example genomes

The genomes of *S. pneumoniae* ATCC 700669, *B. clausii* KSM-K16 (re-annotated with Prodigal) and *Corynebacterium matruchotii* ATCC 33806 were analysed using BAGEL2, resulting in three lists with putative bacteriocins, one for each class. Here, we only discuss the Class I output (Table 4). From the six annotated genes identified by BAGEL2 as encoding Class I bacteriocins, only one gene was previously annotated as 'putative lantibiotic', the rest were annotated as 'hypothetical protein'. Manual examination of the gene clusters suggests that the latter five might indeed be lantibiotics. An additional candidate gene, *orf3711* of *B. clausii* KSM-K16, could only be identified by BAGEL2 after re-annotation of its genome with Prodigal. Furthermore, BAGEL2 was able to predict a sub-class for all putative bacteriocins obtained in this screening. Within the group of significant putative bacteriocins of Class I no false positives were observed. In this example of three genomes, ~80% of the putative bacteriocins of Class I identified by BAGEL2 are not annotated as such with current methods.

CONCLUSIONS

Using BAGEL2, we showed that the annotation systems currently used by NCBI, TIGR and JGI fail to discover all putative bacteriocin genes in bacterial genomes. We demonstrated that the BAGEL2 web server is able to identify (small) bacteriocin-coding ORFs quickly and precisely using several crucial features in addition to homology. The incorporation of increasingly more bacteriocin knowledge into the software enables it to accurately predict the sub-classes to which putative bacteriocins belong. The development of the automated bacteriocin prediction tool BAGEL2 significantly improves the prediction of putative bacteriocins and should be of added value to prokaryotic genome annotation pipelines. Rapid advances in sequencing methods and computational meta-genomics (20), improving contig lengths, will soon

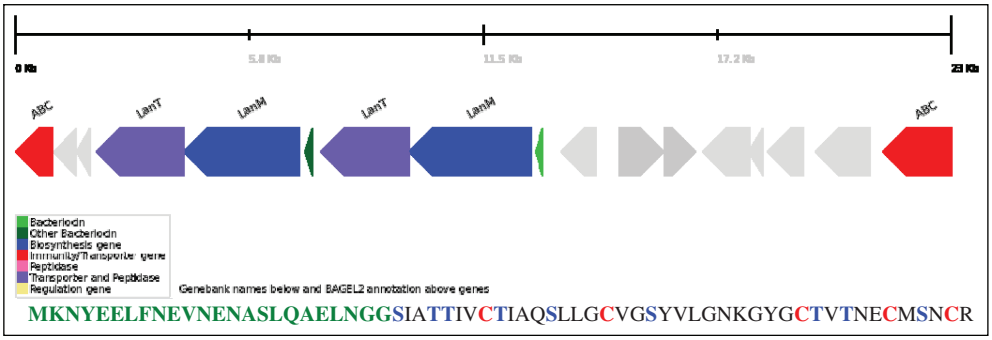

Figure 2. BAGEL2 graphical output for putative bacteriocin (light green). ClocelDRAFT_0418 from *Clostridium cellulovorans* 743B, which was identified through the new MA-2PEPA motif. Amino acids in the leader sequence of the putative bacteriocin are indicated in green. Amino acids potentially involved in lanthionine ring formation are marked in red (cysteine) and blue (serine and threonine).

Table 3. Putative bacteriocins identified by BAGEL2 using new HMMs

Motif	Putative bacteriocin identified by new motif	Organism
LE-MER1	bpmyx0001_45460	*B. pseudomycoides* DSM 12442
MA-2PEPA	ClocelDRAFT_0418	*C. cellulovorans* 743B
LE-LAC481	G11MC16DRAFT_3402	*Geobacillus* sp. G11MC16
LE-LanBC	SnasDRAFT_14510	*Stackebrandtia nassauensis* DSM 44728
MA-2PEPb	CORMATOL_02550	*C. matruchotii* ATCC 3380
MA-DUF	bcere0025_31210	*B. cereus* F65185

Table 4. Putative Class I bacteriocins identified with BAGEL2

Gene	Product	Protein_ID	Class	Score	Organism
SPN23F_12701	Putative lantibiotic precursor	YP_002511205.1	IB	8025	*S. pneumoniae*
SPN23F_19710	Hypothetical protein	YP_002511831.1	IB	2950	*S. pneumoniae*
SPN23F_19700	Hypothetical protein	YP_002511830.1	IB	1950	*S. pneumoniae*
orf3711	Not annotated	orf3711	IA	10 400	*B. clausii* KSM
CORMATOL_2550	Hypothetical protein	EEG25958.1	IB	4775	*C. matruchotii*
CORMATOL_2549	Hypothetical protein	EEG25957.1	IB	3575	*C. matruchotii*
CORMATOL_2551	Hypothetical protein	EEG25959.1	IB	3375	*C. matruchotii*

Displayed are the combined data from mining the genomes of *S. pneumoniae* 23F, *B. clausii KSM-K16* (re-annotated using Prodigal) and *C. matruchotii* ATCC 33806.

allow bacteriocin mining in meta-genomic data with BAGEL2.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors are grateful to H.C. Pietersma for his valuable contribution on Perl scripting.

FUNDING

Funding for open access charge: Department of Molecular Genetics, University of Groningen, The Netherlands.

Conflict of interest statement. None declared.

REFERENCES

1. Cotter,P.D., Hill,C. and Ross,R.P. (2005) Bacteriocins: Developing innate immunity for food. *Nat. Rev. Microbiol.*, **3**, 777–788.
2. Gillor,O., Nigro,L.M. and Riley,M.A. (2005) Genetically engineered bacteriocins and their potential as the next generation of antimicrobials. *Curr. Pharm. Des.*, **11**, 1067–1075.
3. de Jong,A., van Hijum,S.A.F.T., Bijlsma,J.J.E., Kok,J. and Kuipers,O.P. (2006) BAGEL: A web-based bacteriocin genome mining tool. *Nucleic Acids Res.*, **34**, W273.
4. Knoll,C., Divol,B. and du Toit,M. (2008) Genetic screening of lactic acid bacteria of oenological origin for bacteriocin-encoding genes. *Food Microbiol.*, **25**, 983–991.
5. Navarro,L., Rojo-Bezares,B., Sáenz,Y., Díez,L., Zarazaga,M., Ruiz-Larrea,F. and Torres,C. (2008) Comparative study of the pln locus of the quorum-sensing regulated bacteriocin-producing *L. plantarum J51 strain*. *Int. J. Food Microbiol.*, **128**, 390–394.
6. Holtsmark,I., Eijsink,V.H. and Brurberg,M. (2008) Bacteriocins from plant pathogenic bacteria. *FEMS Microbiol. Lett.*, **280**, 1–7.
7. Bierbaum,G. and Sahl,H.G. (2009) Lantibiotics: mode of action, biosynthesis and bioengineering. *Curr. Pharm. Biotechnol.*, **10**, 2–18.
8. Nissen-Meyer,J., Rogne,P., Oppegard,C., Haugen,H.S. and Kristiansen,P.E. (2009) Structure-function relationships of the non-lanthionine-containing peptide (class II) bacteriocins produced by Gram-positive bacteria. *Curr. Pharm. Biotechnol.*, **10**, 19–37.
9. Maqueda,M., Sanchez-Hidalgo,M., Fernandez,M., Montalban-Lopez,M., Valdivia,E. and Martinez-Bueno,M. (2008) Genetic features of circular bacteriocins produced by gram-positive bacteria. *FEMS Microbiol. Rev.*, **32**, 2–22.
10. Drider,D., Fimland,G., Hechard,Y., McMullen,L.M. and Prevost,H. (2006) The continuing story of class IIa bacteriocins. *Microbiol. Mol. Biol. Rev.*, **70**, 564.
11. Willey,J.M. and Van der Donk,W.A. (2007) Lantibiotics: peptides of diverse structure and function. *Annu. Rev. Microbiol.*, **61**, 477–501.
12. Goto,Y., Li,B., Claesen,J., Shi,Y., Bibb,M.J. and van der Donk,W.A. (2010) Discovery of unique lanthionine synthetases reveals new mechanistic and evolutionary insights. *PLoS Biology*, **8**, e1000339.
13. Begley,M., Cotter,P.D., Hill,C. and Ross,R.P. (2009) Identification of a novel two-peptide lantibiotic, lichenicidin, following rational genome mining for LanM proteins. *Appl. Environ. Microbiol.*, **75**, 5451.
14. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
15. Kent,W.J. (2002) BLAT-the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
16. Delcher,A.L., Bratke,K.A., Powers,E.C. and Salzberg,S.L. (2007) Identifying bacterial genes and endosymbiont DNA with glimmer. *Bioinformatics*, **23**, 673.
17. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
18. Bailey,T.L. and Gribskov,M. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, **14**, 48–54.
19. Dufour,A., Hindre,T., Haras,D. and Le Pennec,J.P. (2007) The biology of lantibiotics from the lacticin 481 group is coming of age. *FEMS Microbiol. Rev.*, **31**, 134–167.
20. Wooley,J.C., Godzik,A. and Friedberg,I. (2010) A primer on metagenomics. *PLoS Comput. Biol.*, **6**, e1000667.