microbial biotechnology

Microbial Biotechnology (2011) **4**(3), 308–313 doi:10.1111/j.1751-7915.2011.00261.x

Genomics update

Searching in microbial genomes for encoded small proteins

Jos Boekhorst,[1,2,3] Greer Wilson[4] and Roland J. Siezen[1,2,3]*

[1]*TI Food and Nutrition, 6700AN Wageningen, the Netherlands.*

[2]*NIZO food research, 6710BA Ede, the Netherlands.*

[3]*Netherlands Bioinformatics Centre and Centre for Molecular and Biomolecular Informatics, NCMLS, Radboud University Nijmegen Medical Centre, 6500HB Nijmegen, the Netherlands.*

[4]*Science Consultant, Bowlespark 30, 6701DS Wageningen, the Netherlands.*

As the price of genome sequencing has been rapidly decreasing and can be expected to keep on doing so in the next 10 years, the speed at which new microbial genome sequences become available will increase accordingly. In most genome projects, the first step after acquiring a genome sequence is predicting protein-encoding open reading frames (pORFs). Small proteins or peptides, loosely defined as less than 50 amino acids, encoded in microbial genomes have been largely underestimated. Recent focused functional genomics efforts have led to the identification of a number of new small proteins encoded in genomes of both Gram-negative and Gram-positive bacteria, and fungi (Kastenmayer *et al.*, 2006; Li *et al.*, 2008; Hemm *et al.*, 2010; Hobbs *et al.*, 2010; Bitton *et al.*, 2011). Increasing evidence demonstrates that small proteins participate in a wide array of cellular processes and exhibit great diversity in their mechanisms of action. A recent review (Hobbs *et al.*, 2011) highlights examples of small proteins that, in addition to the well-conserved small ribosomal proteins, participate in cell signalling or regulation, act as antibiotics and toxins/anti-toxins, alter membrane features, act as chaperones, stabilize protein complexes or serve as structural proteins (Table 1) (Fig. 1).

Failure to recognize a pORF encoding a small protein means that these important cell constituents will be missed. Here, we give a brief summary of which problems arise in searching for such encoded small proteins, and what we could do to improve the search process.

*For correspondence. E-mail r.siezen@cmbi.ru.nl; Tel. (+31) 2436 19559; Fax (+31) 2436 19395.

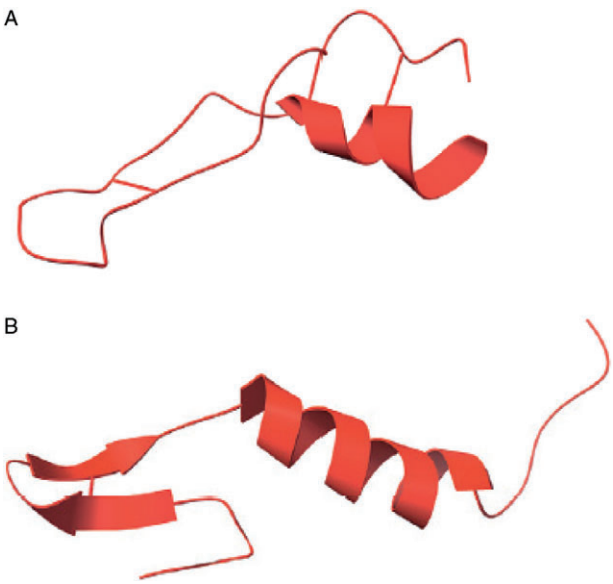

Fig. 1. Structure of small proteins. Small secreted proteins exhibit diversity in their three-dimensional structures and can contain unique intramolecular linkages or modified amino acids. For example, the mature form of (A) subtilosin (PDB: 1PXQ) is cyclized in a head-to-tail fashion (link omitted here for clarity) and contains three unique linkages between Cys sulfur atoms and α-carbons of Phe and Thr. In contrast, (B) the bacteriocin leucocin A (PDB: 1CW6) has a single disulfide bond. Reproduced from Hobbs and colleagues (2011), with permission from *Current Opinion in Microbiology*.

How are pORFs predicted?

Most commonly used genome annotation pipelines use Glimmer (Delcher *et al.*, 2007) or similar tools for the *ab-initio* prediction of pORFs. An overview of commonly used pipelines can be found in Siezen and van Hijum (2010). These tools use sequence characteristics like GC% and codon usage to differentiate between pORFs and non-coding DNA. Sometimes other sequence characteristics are included, e.g. recent versions of Glimmer can include the prediction of putative ribosome binding sites preceding pORFs. These *ab-initio* approaches have

Table 1. Types and functions of small proteins.

Types and functions	Length minimum[a]	References	Characteristics
Ribosomal proteins	37	Wilson and Nierhaus (2005)	High sequence similarity, easy to detect, but often missed
Bacteriocins	30	Bauer and Dicks (2005); de Jong *et al.* (2006)	Low sequence similarity; often modified post-translationally
Regulators	30[b]	Hemm *et al.* (2010)	
Signalling proteins	44[b]	Lopez *et al.* (2009)	Often modified post-translationally
Small membrane proteins	26[b]	Hemm *et al.* (2008); Prymula and Roterman (2009)	Hydrophobic
Chaperones and stress resistance proteins	20	Hemm *et al.* (2010)	

a. Number of amino acids (note that many families have members with a wide variety of lengths, the number given is the approximate lower boundary).
b. As reported by Hobbs and colleagues (2011).

difficulty accurately predicting small pORFs, as the lack of data makes it difficult to distinguish between signal and background noise. To prevent a large number of predicted false-positives (i.e. predicted pORFs that do not actually code for proteins), many pipelines include a minimal gene size threshold, typically picking a (quite arbitrary) size of around 150 bases, i.e. 50 amino acids. For genomes with a low GC content, this works relatively well, as non-coding DNA of these genomes contains a lot of stop codons. For genomes with a high GC content, the power of this approach is limited, as these genomes contain less stop codons, and genes are less obvious to find (Fig. 2) (Tech and Merkl, 2003).

The accuracy of pORF prediction can be increased by combining an *ab-initio* approach with similarity-based approaches. These approaches are based on the assumption that pORFs are under selective constraint relative to non-coding DNA: relatively high similarity of a putative pORF to an ORF in another species supports the hypothesis that the pORF encodes a protein. Some pipelines, e.g. RAST (Aziz *et al.*, 2008), utilize this principle by over-predicting pORFs, followed by a step in which small pORFs without significant similarity to ORFs from other species are deleted.

The *ab-initio* prediction of pORFs is sensitive to sequencing errors. Singe-nucleotide read errors can introduce in-frame stop codons or introduce frameshifts. Some pipelines include a step that detects such errors by analysing 5′ and 3′ prime ends of putative pORFs, attempting to generate a longer pORF by introducing a frameshift or removing a stop codon. Unfortunately, small ORFs (say < 300 bases) chopped in half by a frameshift or in-line stop codon resulting from a read error will result in two very small ORFs, which probably will be completely absent from the initial gene calls. These read errors will therefore not be picked up by these steps. A purely similarity-based approach, where either known protein-coding genes are compared against the genome using BLASTN, or amino acid sequence of known proteins are compared against the genome using TBLASTN, would detect these pORFs (and small pORFs in general), but unfortunately most pipelines do not include such a step. The detection of small pORFs based on sequence similarity is discussed in more detail in Poptsova and Gogarten (2010). When such an approach is taken, be careful not to propagate false-positive pORFs from other studies in your own genome!

Experimental support

A more elaborate approach uses whole-genome tiling arrays or RNA sequencing to confirm and refine pORF prediction. A recent study in *Candida albicans* identified as many as 2000 novel transcriptional segments, including both pORFs and non-coding RNAs (Sellam *et al.*, 2010). In these approaches, detection of messenger RNA

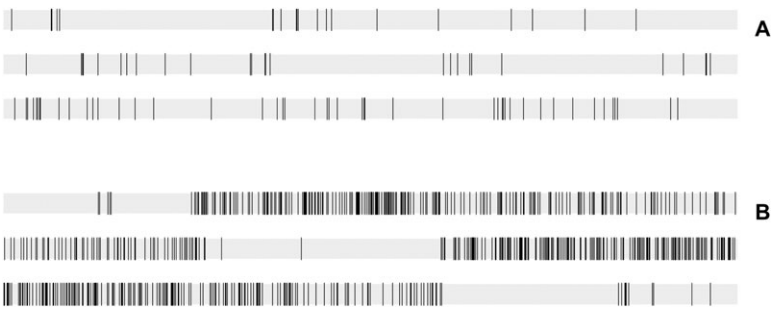

Fig. 2. The impact of GC percentage on ORF prediction. Horizontal grey bars are reading frames, vertical black lines are stop codons; (A) the reading frames in the 5′–3′ direction of the first 10 kb of the chromosome of *Streptomyces coelicolor* (high GC%), (B) the reading frames in the 5′–3′ direction of the first 10 kb of the chromosome of *Lactobacillus johnsonii* (low GC%). Figure generated using Artemis (Rutherford *et al.*, 2000).

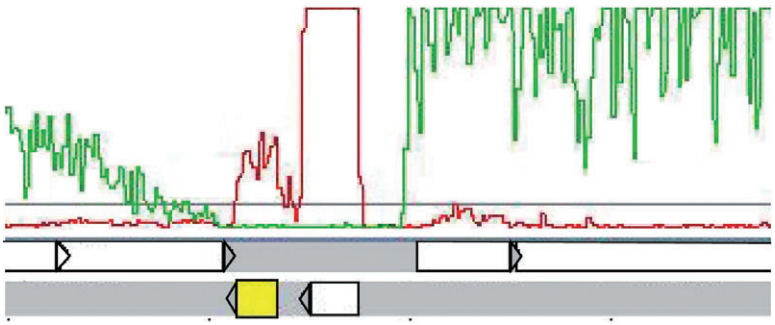

Fig. 3. Detection of a small pORF in *Lactobacillus plantarum* WCFS1 using tiling array data. The green line represents expression in the + direction, the red line expression in the – direction (T. Todt, M. Wels, R.S. Bongers, R.J. Siezen, S.A.F.T. van Hijum and M. Kleerebezem, unpubl. data). The white boxes are putative pORFs present in the GenBank annotation of *L. plantarum*, the yellow box indicates a new putative small pORF not present in the GenBank annotation, but supported by the tiling array data.

is taken as evidence that the ORF is transcribed (Fig. 3). As this requires relatively elaborate experiments, they are not routinely part of pORF prediction.

New tools and methods

In addition to the methods mentioned above, new tools have been generated for targeted detection of small pORFs. Some tools are targeted at a specific group of proteins, for example Bagel, a bioinformatics tool for mining bacterial genomes for bacteriocins (de Jong *et al.*, 2006). Other tools aim at small pORFs in general, by combining *ab-initio* methods as described earlier with estimates of selection pressure, e.g. sORF Finder (Hanada *et al.*, 2010). In addition to these bioinformatics approaches, there are also experimental techniques particularly suited or specifically designed for the identification and functional characterization of small pORFs. Many small proteins can be inserted into membranes with relative ease (Kuhn *et al.*, 2010), and protein characteristics like hydrophobicity can be used to infer protein function (Prymula and Roterman, 2009; Prymula *et al.*, 2010). Techniques are emerging that allow us to differentiate between non-functional ORFs and (conditionally) essential or beneficial genes (Dinger *et al.*, 2008). These methods involve the generation of large sets of gene inactivation mutants, followed by essays measuring growth characteristics (Bijlsma *et al.*, 2007; Hobbs *et al.*, 2010) (Fig. 4).

Unfortunately, these new methods are not routinely applied in microbial genome annotation. If the functions most often encoded by small pORFs are of particular interest to you, you should consider including one or more of these dedicated analysis methods in you annotation pipeline.

Why should we care about small protein-encoding ORFs?

Small pORFs are often the first to be removed in genome (re)-annotation, even though closer inspection reveals that many of them have a high coding potential (Li *et al.*, 2008), and studies targeted specifically at identifying small pORFs often identify significant numbers of novel pORFs. A study in *Schizosaccharomyces pombe* identified 39 likely functional proteins (Bitton *et al.*, 2011), and a

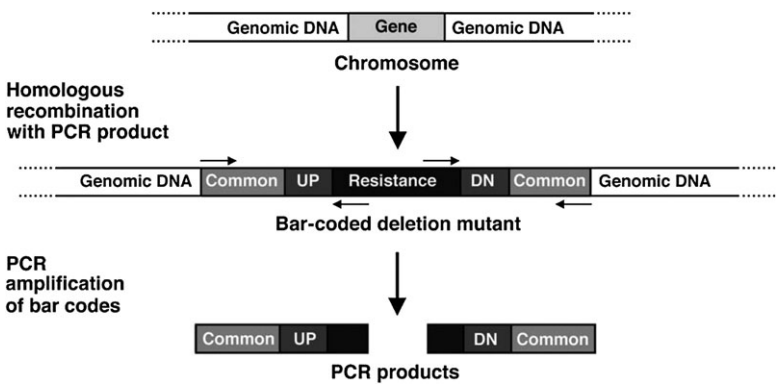

Fig. 4. Generation of bar-coded deletion mutants. Kanamycin-resistance cassettes flanked by two unique 20-mer DNA bar code sequences (UP and DN) were generated by a two-step PCR process for each deleted gene. For the analysis of the large-scale competition experiments, bar codes upstream and downstream of every kanamycin-resistance cassette are amplified by means of common primer sequences (indicated by small black arrows) encoded within the regions bordering the UP and DN bar codes. The amplified bar codes were then hybridized to a DNA microarray to score each bar-coded deletion mutant within the population. Reproduced from Hobbs and colleagues (2010) with permission from the American Society of Microbiology.

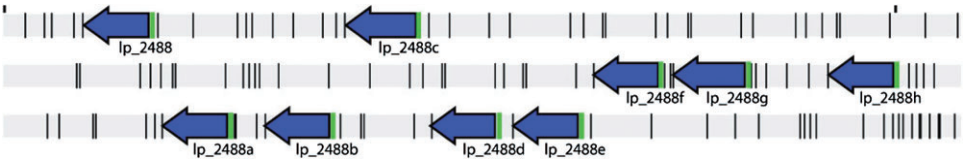

Fig. 5. Conserved small putative pORFs in *Lactobacillus plantarum* WCFS1. The nine ORFs (blue arrows) are located in different reading frames (horizontal grey bars). Although the ORFs are very small (approximately 40 AA), all nine ORFs are preceded by a good ribosome binding site (vertical green lines), supporting the hypothesis that these ORFs encode proteins. The vertical black lines are stop codons. Only the leftmost ORF is annotated in the *L. plantarum* WCFS1 genome currently in GenBank (locus tag lp_2488) (Kleerebezem *et al*., 2003). An update release will contain all nine ORFs (R.J. Siezen, B. Renckens, C. Francke, J. Boekhorst, M. Wels, M. Kleerebezem and S.A.F.T. van Hijum, unpubl. data). Figure generated using Artemis (Rutherford *et al*., 2000).

study in *Bacillus subtilis* revealed 11 transcriptional units linked to sporulation, many containing functional pORFs (Schmalisch *et al*., 2010). Still, if most small pORFs had rather boring and uninteresting functions, missing most of them in genome annotation would not be too much of a problem. However, the function of many, if not most, small ORFs remain uncertain. Systematic studies into small pORFs reveal novel gene families with no similarity to known proteins, providing a pool of genes that could be responsible for as yet unexplained regulatory or phenotypic complexity (Warren *et al*., 2010). An intriguing example of small pORFs with unknown function can be found in *Lactobacillus plantarum* WCFS1 (Fig. 5). Nine consecutive putative pORFs are highly similar to each other, and have excellent ribosome binding sites, yet lack any significant similarity to genes with known or predicted functions.

Setting thresholds

If we cannot get it exactly right, should we aim for over-prediction or under-prediction?

As in many bioinformatics analysis, pORF prediction involves setting a lot of thresholds: what is the minimal length of a pORF? How much is codon usage allowed to deviate from the norm? Choosing liberal thresholds will result in over-prediction, while being strict will mean you are likely to miss real pORFs. Which of these two evils to choose from depends on you research question. If you are designing custom microarray slides to measure gene expression, liberal thresholds are probably the way to go (assuming you can squeeze in the additional probes on the slide). Over-predicted pORFs will simply result in no signal for these 'ORFs', while not recognizing real pORFs means you will not detect (changes in) expression for these pORFs, in turn potentially meaning you might not be able to answer your research question.

Wherever you place your thresholds, it is crucial to accurately describe the procedure followed. Although standardization initiatives like the Standards In Genome Sciences (http://sigen.org/index.php/sigen) are gaining ground, it can be non-trivial to figure out how exactly a specific study or annotation pipeline predicts pORFs. This can be especially problematic in comparative genomics studies. Statements like '30% of the pORFs in genome A do not have a homologue in any other species, while for species B this is only 5%' become quite meaningless if they reflect differences in pORF calling rather than biological differences. It has been argued that a common standard for pORF prediction would greatly benefit comparative analysis (Nielsen and Krogh, 2005).

Sometimes the choice of experimental techniques and design can circumvent ORF calling all together. In high-throughput mass-spectrometry-based proteomics, the database against which peptides are searched (Perkins *et al*., 1999) can be filled in such a say that it includes virtually all potential protein-coding ORFs, as in the database tsORFdb for theoretical small ORFs (Heo *et al*., 2010). In gene expression analysis, the use of tiling arrays, on which every nucleotide of a genome is represented in at least one probe (Mockler *et al*., 2005) as well as RNA sequencing (Wang *et al*., 2009) circumvent ORF calling altogether, and the data produced in these types of experiments can in fact be used to identify pORFs (Fig. 3). Both proteomics and RNA sequencing are rapidly advancing techniques, potentially making the impact of ORF calling issues less of a problem in studies where these experimental techniques could be applied. In contrast, new developments in methods for function elucidation heavily rely on ORF predictions, making the issue far from obsolete.

Future

The cost of sequencing bacterial genomes will continue its journey downward, resulting in an ever-increasing speed at which new sequences become available. This will in turn increase the power of comparative methods for the identification of small pORFs. Wet-lab studies, both high-throughput and low-throughput, will provide experimental confirmation of putative pORFs, allowing the creation, training and validation of more accurate bioinformatics tools for the prediction of pORFs. The downside of

the dramatic reduction in the cost of sequencing a microbial genome is that the speed at which new genome sequences become available keeps increasing, while there is no corresponding increase in man-hours for manual curation. Pioneering genome projects had many man-years available for painstaking checking and correction of automated pORF predictions, while recent genomes are generally annotated completely automatically. In principle, this lack of curation could be offset by the increase in quality of the automated methods (or at least in part), but this requires that scientists pay attention to the use of tools and template genomes and are aware of the pitfalls.

What do these small proteins or large peptides do? Where are they located? Do they reside inside the cell, in the membrane, on the cell surface or are they secreted? How do they get to where they should be? Are short hydrophobic proteins inserted directly into the membrane after ribosomal synthesis (Kuhn *et al*., 2010)? How are their structures stabilized? Which are subject to post-translational modification and where? Clearly, experimentalists still have lots of high-throughput analyses to complete, and bioinformaticians will need to continuously fine-tune their search algorithms. Exciting times and more still to come.

Acknowledgements

We thank Michiel Wels and Tilman Todt for use of unpublished tiling array data of *L. plantarum*. This project was carried out within the research programmes of the Netherlands Bioinformatics Centre, which is part of the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research.

References

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., *et al*. (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9:** 75.

Bauer, R., and Dicks, L.M. (2005) Mode of action of lipid II-targeting lantibiotics. *Int J Food Microbiol* **101:** 201–216.

Bijlsma, J.J., Burghout, P., Kloosterman, T.G., Bootsma, H.J., de Jong, A., Hermans, P.W., and Kuipers, O.P. (2007) Development of genomic array footprinting for identification of conditionally essential genes in *Streptococcus pneumoniae*. *Appl Environ Microbiol* **73:** 1514–1524.

Bitton, D.A., Wood, V., Scutt, P.J., Grallert, A., Yates, T., Smith, D.L., *et al*. (2011) Augmented annotation of the *Schizosaccharomyces pombe* genome reveals additional genes required for growth and viability. *Genetics* (Epub ahead of print).

Delcher, A.L., Bratke, K.A., Powers, E.C., and Salzberg, S.L. (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23:** 673–679.

Dinger, M.E., Pang, K.C., Mercer, T.R., and Mattick, J.S. (2008) Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput Biol* **4:** e1000176.

Hanada, K., Akiyama, K., Sakurai, T., Toyoda, T., Shinozaki, K., and Shiu, S.H. (2010) sORF finder: a program package to identify small open reading frames with high coding potential. *Bioinformatics* **26:** 399–400.

Hemm, M.R., Paul, B.J., Schneider, T.D., Storz, G., and Rudd, K.E. (2008) Small membrane proteins found by comparative genomics and ribosome binding site models. *Mol Microbiol* **70:** 1487–1501.

Hemm, M.R., Paul, B.J., Miranda-Rios, J., Zhang, A., Soltanzad, N., and Storz, G. (2010) Small stress response proteins in *Escherichia coli*: proteins missed by classical proteomic studies. *J Bacteriol* **192:** 46–58.

Heo, H.S., Lee, S., Kim, J.M., Choi, Y.J., Chung, H.Y., and Oh, S.J. (2010) tsORFdb: theoretical small open reading frames (ORFs) database and massProphet: peptide mass fingerprinting (PMF) tool for unknown small functional ORFs. *Biochem Biophys Res Commun* **397:** 120–126.

Hobbs, E.C., Astarita, J.L., and Storz, G. (2010) Small RNAs and small proteins involved in resistance to cell envelope stress and acid shock in *Escherichia coli*: analysis of a bar-coded mutant collection. *J Bacteriol* **192:** 59–67.

Hobbs, E.C., Fontaine, F., Yin, X., and Storz, G. (2011) An expanding universe of small proteins. *Curr Opin Microbiol* **14:** 1–7.

de Jong, A., van Hijum, S.A., Bijlsma, J.J., Kok, J., and Kuipers, O.P. (2006) BAGEL: a web-based bacteriocin genome mining tool. *Nucleic Acids Res* **34:** W273–W279.

Kastenmayer, J.P., Ni, L., Chu, A., Kitchen, L.E., Au, W.C., Yang, H., *et al*. (2006) Functional genomics of genes with small open reading frames (sORFs) in *S. cerevisiae*. *Genome Res* **16:** 365–373.

Kleerebezem, M., Boekhorst, J., van Kranenburg, R., Molenaar, D., Kuipers, O.P., Leer, R., *et al*. (2003) Complete genome sequence of *Lactobacillus plantarum* WCFS1. *Proc Natl Acad Sci USA* **100:** 1990–1995.

Kuhn, A., Stiegler, N., and Schubert, A.K. (2010) Membrane insertion of small proteins. *Methods Mol Biol* **619:** 39–62.

Li, Q.R., Carvunis, A.R., Yu, H., Han, J.D., Zhong, Q., Simonis, N., *et al*. (2008) Revisiting the *Saccharomyces cerevisiae* predicted ORFeome. *Genome Res* **18:** 1294–1303.

Lopez, D., Vlamakis, H., Losick, R., and Kolter, R. (2009) Paracrine signaling in a bacterium. *Genes Dev* **23:** 1631–1638.

Mockler, T.C., Chan, S., Sundaresan, A., Chen, H., Jacobsen, S.E., and Ecker, J.R. (2005) Applications of DNA tiling arrays for whole-genome analysis. *Genomics* **85:** 1–15.

Nielsen, P., and Krogh, A. (2005) Large-scale prokaryotic gene prediction and comparison to genome annotation. *Bioinformatics* **21:** 4322–4329.

Perkins, D.N., Pappin, D.J., Creasy, D.M., and Cottrell, J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20:** 3551–3567.

Poptsova, M.S., and Gogarten, J.P. (2010) Using comparative genome analysis to identify problems in annotated microbial genomes. *Microbiology* **156:** 1909–1917.

Prymula, K., and Roterman, I. (2009) Functional characteristics of small proteins (70 amino acid residues) forming

protein–nucleic acid complexes. *J Biomol Struct Dyn* **26:** 663–677.

Prymula, K., Salapa, K., and Roterman, I. (2010) 'Fuzzy oil drop' model applied to individual small proteins built of 70 amino acids. *J Mol Model* **16:** 1269–1282.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., and Barrell, B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics* **16:** 944–945.

Schmalisch, M., Maiques, E., Nikolov, L., Camp, A.H., Chevreux, B., Muffler, A., *et al*. (2010) Small genes under sporulation control in the *Bacillus subtilis* genome. *J Bacteriol* **192:** 5402–5412.

Sellam, A., Hogues, H., Askew, C., Tebbji, F., Hoog, M., Lavoie, H., *et al*. (2010) Experimental annotation of the human pathogen *Candida albicans* coding and noncoding transcribed regions using high-resolution tiling arrays. *Genome Biol* **11:** R71.

Siezen, R.J., and van Hijum, S.A. (2010) Genome (re-) annotation and open-source annotation pipelines. *Microb Biotechnol* **3:** 362–369.

Tech, M., and Merkl, R. (2003) YACOP: enhanced gene prediction obtained by a combination of existing methods. *In Silico Biol* **3:** 441–451.

Wang, Z., Gerstein, M., and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10:** 57–63.

Warren, A.S., Archuleta, J., Feng, W.C., and Setubal, J.C. (2010) Missing genes in the annotation of prokaryotic genomes. *BMC Bioinformatics* **11:** 131.

Wilson, D.N., and Nierhaus, K.H. (2005) Ribosomal proteins in the spotlight. *Crit Rev Biochem Mol Biol* **40:** 243–267.

