Essential information for synthetic DNA sequences

To the Editor:

Following a discussion by the workgroup for Data Standards in Synthetic Biology, which met in June 2010 during the Second Workshop on Biodesign Automation in Anaheim, California, we wish to highlight a problem relating to the reproducibility of the synthetic biology literature. In particular, we have noted the very small number of articles reporting synthetic gene networks that disclose the complete sequence of all the constructs they describe.

To our knowledge, there are only a few examples where full sequences have been released. In 2005, a patent application[1] disclosed the sequences of the toggle switches published four years earlier in a paper by Gardner *et al.*[2]. The same year, Basu *et al.*[3] deposited their construct sequences for programmed pattern formation into GenBank[3]. Examples of synthetic DNA sequences derived from standardized parts that have been made available in GenBank include the refactored genome of the bacteriophage phage T7 (ref. 4) and a BioBrick-based plasmid[5]. More recently, the full genome sequence of synthetic *Mycoplasma mycoides* JCVI-syn1.0 clone sMmYCp235-1 also has been made available in GenBank (accession no. CP002027)[6].

In contrast, most publications provide a variety of methods, information and/or partial sequences to explain the constructs used in a paper; for the research community, piecing together the full sequences of constructs is thus laborious, error-prone and sometimes impossible. A paper from your journal provides a recent example; although Kemmer *et al.*[7] provided admirably detailed Supplementary Information on the construction methods for their plasmids, they failed to provide access to the final sequences. Indeed, the gaps between key components are almost never reported, presumably because they are not considered crucial to the report. Yet, synthetic biology relies on the premise that synthetic DNA can be engineered with base-level precision.

Missing sequence information in papers hurts reproducibility, limits reuse of past work and incorrectly assumes that we know fully which sequence segments are important. For example, many synthetic biologists are currently realizing that translation initiation rates are dependent on more than the Shine-Dalgarno sequence[8]. Sequences upstream of the start codon are crucial for translation rates, yet are underreported. Similarly, it has been demonstrated that intron length can affect the dynamics of genetic oscillators[9]. Many more such examples are likely to emerge.

Because full sequence disclosure is critical, we wonder why the common requirement by many journals to provide GenBank entries for genomes and natural sequences has not been enforced for synthetic DNA and engineered genetic constructs. In an environment where word count is a constant battle, replacing plasmid construction method sections with references to annotated GenBank entries would be a welcome change. We therefore feel that including a completely annotated sequence of the construct would greatly contribute to the development of our discipline. We hope that in the future you will encourage the authors you publish to submit this information to GenBank or other appropriate databases. In the long term, we hope to establish a minimal information guideline around the Minimal Information about a Biomedical or Biological Investigation (MIBBI; http://mibbi.org/index.php/Main_Page) project and welcome contributions from the greater community.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Jean Peccoud[1], J Christopher Anderson[2], Deepak Chandran[3], Douglas Densmore[4], Michal Galdzicki[5], Matthew W Lux[1], Cesar A Rodriguez[6], Guy-Bart Stan[7] & Herbert M Sauro[3]

[1]Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, Virginia, USA. [2]Department of Bioengineering, QB3: California Institute for Quantitative Biological Research, University of California, Berkeley, California, USA. [3]Department of Bioengineering, University of Washington, Seattle, Washington, USA. [4]Department of Electrical and Computer Engineering, Boston University, Boston, Massachusetts, USA.[5]Biomedical and Health Informatics, University of Washington, Seattle, Washington, USA. [6]BIOFAB, Emeryville, California, USA. [7]Department of Bioengineering and Centre for Synthetic Biology and Innovation, Imperial College London, London, UK.
e-mail: peccoud@vt.edu

1. Gardner, T.S. & Collins, J.J. US patent 6,841,376 (2005).
2. Gardner, T.S., Cantor, C.R. & Collins, J.J. *Nature* **403**, 339–342 (2000).
3. Basu, S., Gerchman, Y., Collins, C.H., Arnold, F.H. & Weiss, R. *Nature* **434**, 1130–1134 (2005).
4. Chan, L.Y., Kosuri, S. & Endy, D. *Mol. Syst. Biol.* **1**, 2005.0018 (2005).
5. Shetty, R.P., Endy, D. & Knight, T.F. Jr. *J. Biol. Eng.* **2**, 5 (2008).
6. Gibson, D.G. *et al. Science* **329**, 52–56 (2010).
7. Kemmer, C. *et al. Nat. Biotechnol.* **28**, 355–360 (2010).
8. Salis, H.M., Mirsky, E.A. & Voigt, C.A. *Nat. Biotechnol.* **27**, 946–950 (2009).
9. Swinburne, I.A., Miguez, D.G., Landgraf, D. & Silver, P.A. *Genes Dev.* **22**, 2342–2346 (2008).

***Nature Biotechnology* replies:**

Kemmer *et al.*[1] have now lodged the sequences of the constructs used in their paper with GenBank HQ644133, HQ644134, HQ644135, HQ644136 and HQ644137. *Nature Biotechnology* and other Nature research journals currently require disclosure only of the sequences of genomes, deep sequencing and short-read data, short stretches of novel