Sequencing solution: use volunteer annotators organized via Internet

Sir — When the *Pseudomonas aeruginosa* genome project began in 1997, one question facing us was how best to annotate gene descriptions and other information.We decided to take a community approach, in an attempt to improve the quality of annotation and to use the resources of all those researchers working with this versatile pathogenic bacterium.

Our results support aspects of 'open annotation' approaches for the human genome, as described in Correspondence[1] and News[2]. However, our experience has suggested that certain precautions must be taken.

For the community project, termed PseudoCAP[3], we recruited volunteers from the *Pseudomonas* research community, and later others, to submit annotations of genes or gene families with which they were familiar, through the direction of a single project moderator. Unlike the annotation jamboree for the *Drosophila* genome project[4], all communications with, and submissions by, the volunteer participants were made exclusively through the Internet[4].

The PseudoCAP annotations were overlaid on a genome viewer console developed by PathoGenesis, containing layers of other automatically generated analyses and literature reference information.

This resource, coupled with a critical, conservative annotation approach, was used to generate the final genome annotations, which were also classified according to whether they were based on (1) functional studies in *P. aeruginosa*; (2) high homology to functionally studied genes in other organisms; (3) low homology to functionally studied genes; or (4) homology to hypothetical genes (see the accompanying paper in this issue[5]).

We were pleasantly surprised at the enthusiasm for PseudoCAP — 61 participants made 1,741 submissions. Most of the later participants did not work on *Pseudomonas* but were researchers who wanted to examine genes of particular function. Judging from this response, an adequate number of annotators could probably be recruited for other community annotation projects.

Given the experimental nature of our approach, we allowed participants to submit whatever information they wished; as a result, variation in the quality of annotations led to numerous inconsistencies. Therefore, review of all annotations by a core group was essential. For the future, we recommend that community participants should be required to clearly define their annotation methods and criteria for using any particular functional description, and adopt a consistent, searchable format. Otherwise inconsistencies will not be easily detected, and useful information (for example, retrieval of all annotations based on a certain type of functional study) will not be readily available.

Final annotations for the genome project were based almost exclusively on functional studies of the gene in question, or on close homology of the encoded protein to functionally studied proteins. This method involved significant manual intervention, which could be automated if a sequence database based only on functional studies is created. SwissProt and the National Center for Biotechnology Information's RefSeq are beginning to develop in part along these lines.

In the meantime, we recommend that genome projects consider a community-aided annotation approach, coupled with critical, conservative annotation by a core group of project annotators. If such community involvement occurs through the Internet in a formal, well publicized setting, annotations can continue to be updated and corrected after a genome sequence is published.

Fiona S. L. Brinkman*, Robert E. W. Hancock*, C. Kendal Stover†

**Department of Microbiology and Immunology, University of British Columbia, Vancouver, British Columbia V6T 1Z3, Canada*

†PathoGenesis Corporation, 201 Elliott Ave West, Seattle, Washington 98119, USA

1. Hubbard, T. & Birney, E. *Nature* **403,** 825 (2000).
2. Butler, D. *Nature* **404,** 694 (2000).
3. http://www.cmdr.ubc.ca/bobh/PAAP.htm
4. Pennisi, E. *Science* **287,** 2182–2184 (2000).
5. Stover, C. K. *et al. Nature* **406,** 959–964 (2000).

Let's get the right man in the right job

Sir — As the late Peter Medawar[1] wrote "Scientists are entitled to be proud of their accomplishments, and what accomplishments can they call 'theirs' except the things they have done or thought of first?"

In this spirit I would like to correct the recent News and Views article[2] on the second Chapman Conference on Gaia. In an otherwise interesting article, Jim Gillon gets both my name and affiliation wrong, calling me D. Williamson of the University of Liverpool rather than D. Wilkinson of Liverpool John Moores University.

It may be of interest that some of my ideas on Gaia and evolution were published in *Oikos* last year[3]; since writing that paper I have become slightly less sceptical about the possibility of global regulation, for the reasons summarized in Gillon's article.

David M. Wilkinson

Biological and Earth Sciences, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, UK

1. Medawar, P. *Pluto's Republic* (Oxford Univ. Press, Oxford, 1982).
2. Gillon, J. *Nature* **406,** 685–686 (2000).
3. Wilkinson, D. M. *Oikos* **84,** 533–553 (1999).

Model already exists for fair use of gene data

Sir — I agree wholeheartedly with the analogies drawn between the advantages of open source and the genome projects by Russ, Aparicio and Carlton[1]. However, I would go further than Alberts and Klug[2,3] and suggest that it should never be possible to patent a gene. A gene is a pre-existing entity; it can be discovered, but not invented. Of course a drug invented to exploit a gene, or a method using a particular gene for therapy, is a different matter. These are clearly inventions.

I should also like to clear up the details of the GNU Public Licence. The GNU licence is not about software being free of charge. It is about freedom: allowing the user of the code to do with it what they will. The licence actually comes in two forms: the GPL and the LGPL. The GPL allows use of the software in commercial or freely distributable software. The restriction, however, is that the distributor of the software must make the source code available and must pass on the same rights of freedom; any code linked into an executable with GPL code must also fall under the GPL. Thus a product may be sold including GPL code, but the purchaser has the right to distribute the product without restriction, either for a fee or at no cost.

The LGPL, on the other hand, allows freedom of distribution (either commercially or at no cost), but allows executable programs to be created which do not require the LGPL code to be distributed and can be sold with the usual commercial restrictions. Thus the genome model better fits the LGPL. The original gene data should be completely free and shared throughout the community. Companies would then be free to exploit that information and to create patentable commercial products based on it. More than one company would be able to exploit the same gene, putting an end to the current 'land grab'[4].

Andrew C. R. Martin

School of Animal and Microbial Sciences, University of Reading, PO Box 228, Whiteknights, Reading RG6 6AJ, UK

1. Russ, A. P., Aparicio, S. A. J. R. & Carlton, M. B. L. *Nature* **404,** 809 (2000).
2. Alberts, B. & Klug, A. *Nature* **404,** 325 (2000).
3. Alberts, B. & Klug, A. *Nature* **404,** 542 (2000).
4. Gavaghan, H. *Nature* **404,** 684–686 (2000).