EDITORIAL

If you build it, they might come

Roy and Laura Welch examine why researchers seem reluctant to be more directly involved in the annotation of microbial genomes.

To annotate an organism's genome, biological information about the organism must be matched to the genes and genetic elements in the sequenced genome. The process is iterative and open-ended: new information is constantly incorporated into the annotation. It can also be recursive: analysis of the annotation may provide insight about the organism that in turn leads to changes to the annotation. Unfortunately, the generation of new information and annotation of the genome are at present completely separate processes. Often new information does not become incorporated into the annotation in a timely manner, a costly loss for those who rely on it to advance their research.

The community of expert researchers who study an organism produce most of the information that becomes part of the annotation and are also the primary group of end-users. It is therefore curious that the annotation process is circuitous and inefficient: researchers communicate new information not as direct updates to the annotation, but as research papers that must later be interpreted and incorporated into the annotation separately — most often by a third party! Indeed, some information never finds its way into the annotation. It would be far more efficient for the research community to contribute directly to genome annotation. Yet the life science community as a whole remains stuck in the old, inefficient paradigm.

Technology is not the impediment. The Internet is now well equipped to enable a collaborative information repository (CIR). In fact, a successful example already exists: Wikipedia. The Wikipedia online encyclopaedia is written by volunteer contributors, who have created more than 10 million articles. It has approximately 75,000 editors, and it has attracted more than 50 million unique visitors each month throughout 2008. Its impact is enormous, as it is substantially replacing edited proprietary encyclopaedias and its collaborative principles are profound in their simplicity. Anyone with Internet access can edit Wikipedia content, and yet, by some metrics, it is as accurate as the Encyclopaedia Britannica.

Wikipedia can be reduced to just three fundamental editing principles: all content is available for editing by any member of the community; all edits are saved in perpetuity; and any member can easily undo all changes simply by reverting to a previous state. Superficially, these principles seem to invite chaos, but over time, Wikipedia has demonstrated that they act as a stabilizing force. It is enticing to imagine the positive impact that this kind of activity could have on the completeness and efficiency of the annotation of a genome through the participation of researchers.

The failure of direct, collaborative genome annotation has not been caused by lack of funding. Both the National Institutes of Health and the National Science Foundation have supported projects to create databases using Wikipedia principles. These databases include EcoliWiki, GONUTS (gene ontology normal usage tracking system), xanthusBase and WikiPathways.

Strangely, the impediment to the creation of a successful CIR seems to be sociological. It is clear that the success of community genome annotation depends on certain prerequisites. First, a group of participants must operate within a set of shared behavioural principles. Second, novices must be given opportunities to become members of the group, and thereby learn its behavioural principles by participating first in simple activities and then, as they gain experience, more complex ones. Third, there must be a means of rewarding participants who contribute to enhance their credibility within the group. This last point is particularly crucial, as it provides direct incentives for participation.

Academic researchers already operate within this paradigm. We are a group that shares a rigid set of behavioural principles regarding research, publishing and education. Graduate and postgraduate training provides the opportunity for new members to join. The rewards are publication, tenure and funding; credibility is enhanced by anything that appears on a curriculum vitae. It seems inevitable that annotation of CIRs will suffer from a lack of participation because CIRs are not part of the system of rewards and credibility. Contributing time and effort to a CIR does not currently enhance academic credibility, or move a researcher closer to a doctoral degree, tenure or successful funding applications. In fact, in a competitive scientific environment, there are substantial incentives not to share information until doing so will result in a publication.

Until contributions to a genome-annotation CIR can be credited by inclusion in a PhD thesis, curriculum vitae, tenure application or grant proposal, direct collaborative annotations are unlikely to fulfil their promise and potential to accelerate scientific achievement.

Roy Welch is at the Department of Biology, Syracuse University, 130 College Place, BRL Room 702A, Syracuse, New York, New York 13244, USA. Laura Welch is Deputy Director at the Center for Advanced Systems and Engineering, 2-212 CST, Syracuse University, Syracuse, New York, New York 13244, USA. Correspondence to R.W. e-mail: rowelch@gmail.com