Open science is a research accelerator

Michael Woelfle, Piero Olliaro and Matthew H. Todd*

An open-source approach to the problem of producing an off-patent drug in enantiopure form serves as an example of how academic and industrial researchers can join forces to make new scientific discoveries that could have a huge impact on human health.

When we are faced with a challenging scientific problem we cannot solve, what do we do? Many of us would go to see our colleagues and ask for their advice. Our professional network is valuable. It is also limited. Perhaps there are people who are well-placed to help us, in another university or company, in a different country, but we unfortunately do not know them. Surely science would proceed faster if we could reach those people? Or, better, if they could find us? This Commentary describes a case study — a chemical project where open-source methodologies were employed to accelerate the process of discovery. The acceleration occurred because the project was open: relevant experts could identify themselves.

Open source has been responsible for many important software products used worldwide (including, for example, the Linux operating system, the Firefox web browser) and internet resources such as Wikipedia. The process of creating open-source products involves the iterative cycle of (1) a problem or need being identified, (2) a preliminary solution being posted to this problem, (3) an open appeal to the wider community being made, (4) inputs received from an unrestricted community and (5) the cycle beginning over again. Such a cycle can operate quickly because of the advent of online tools that strengthen the relevant networks.

In software development, traditional versus open-source methods of working are described by the analogy of the 'cathedral and the bazaar'[1]. Many academic and industrial groups operate along a cathedral model in that significant objects are built by a closed team of skilled artisans — the training of whom has consumed considerable resources. Cathedral projects operate in a hierarchical scheme — one person is in charge of a closed group. In a bazaar-type project, there is a low barrier to entry, and the operation is seemingly chaotic or self-organizing. Leadership is fluid, if it exists at all. The system is effective at what it does, yet requires little investment to start up and relies on the traffic of inherently interested strangers. We decided to apply this latter approach to a research problem — applying the principles of open-source software development to experimental science.

A scanning electron microscope image of a male and a female Schistosoma worm, the parasites that cause schistosomiasis in humans.

The drug praziquantel (PZQ) is used in the treatment of a serious parasitic infection, schistosomiasis (also known as bilharziasis) that affects the lives of hundreds of millions of people worldwide; the disease has been referred to as a 'silent pandemic'[2]. Praziquantel is highly effective, and is manufactured and distributed on a huge scale[3] — it is distributed for preventive chemotherapy through mass drug administration to school children or entire communities, for example by the Schistosomiasis Control Initiative[4]. As it is off-patent, this demand has driven down the price of the active pharmaceutical ingredient to approximately 10 US cents per gram and that of a 600 mg tablet to 8–14 US cents. The compound is made as a racemate, even though the inactive enantiomer has side effects and is responsible for a bitter taste[5]. A pill consisting of just the active enantiomer would not be bitter (hence more likely to be taken, especially by children), would be smaller (easier to ship and swallow) and generate fewer side effects. The World Health Organization, in its strategic plan for 2008–2013, listed the generation of PZQ as a single enantiomer as a priority[6]. How is it possible to produce only the active enantiomer while keeping the price very low?

This is a unique kind of problem. Racemates are always cheaper to make than enantiopure materials, unless the relevant drug is derivable from a natural source, which PZQ is not. This is a problem that both academia and industry are ill-equipped to solve. Academic research is not concerned with gradually reducing costs of anything, nor in incrementally improving a synthesis. Such aims are not generally suitable as the subject of a graduate thesis. Similarly, the pharmaceutical industry has little motive to assign research and development resources to a project that has a narrow profit margin.

How the project worked

In 2006 a website (on The Synaptic Leap forum) was started in which the problem of the production of PZQ as a single enantiomer was laid out[7]. There was some initial traffic, but there was little substantial community input. It is a fallacy that open-source products simply emerge — there are usually kernels of activity arising from funded work, to which the community then responds[8]. In mid-2008 the PZQ project was funded by a partnership between the World Health Organization and the Australian Government that enabled

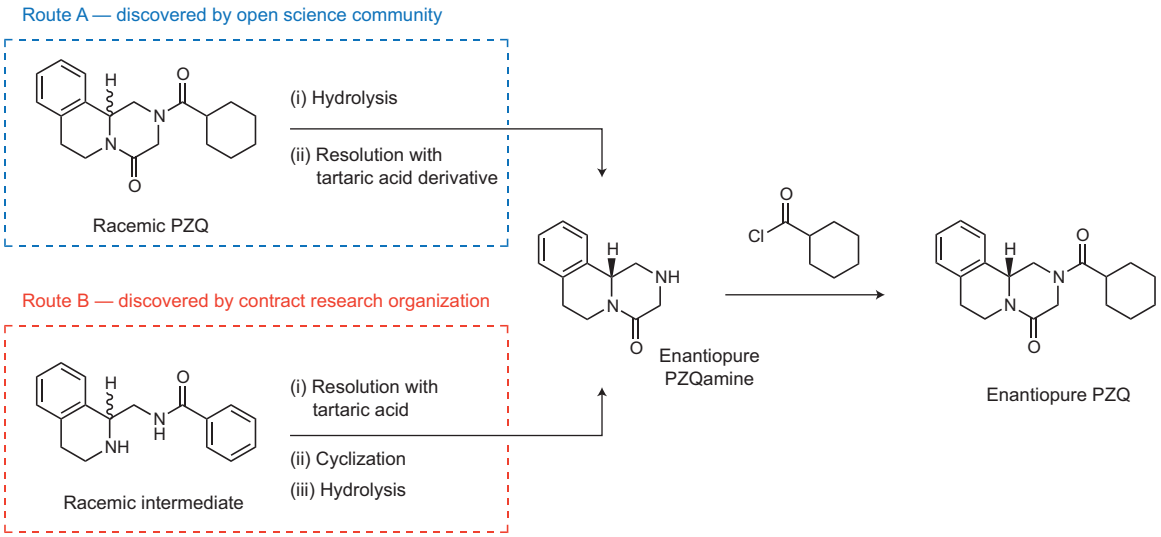


Figure 1 | Routes to enantiopure PZQ discovered by an open-science community and a contract research organization.

preliminary experiments to be performed and all data deposited in an open-source online electronic lab notebook (ELN) which could be properly curated[9]. Our ELN was based on an open-source platform, Labtrove, developed[10] by a team at the University of Southampton in the UK.

Experimental work began in earnest in January 2010. Our early inroads into the problem were only partially successful, but it was probably this incompleteness that stimulated what was a much greater input to the project from people unknown to us at the start — in some respects an open lab notebook is the scientific equivalent of the software development mantra 'release early, release often'[11]. The result of these inputs was eventually a change in direction of the project, away from a catalytic, asymmetric synthesis of PZQ *de novo*, towards an approach based on resolution that was less academically interesting but more likely to succeed.

It became clear that PZQ could be efficiently hydrolysed to give an amine (PZQamine, see Route A, Fig. 1) that might be resolvable. Two problems arose: (1) with the standard chiral stationary phases available to us we were unable to effect a baseline separation of PZQamine's enantiomers; (2) we had no experience in resolutions and did not have an intuitive feeling about a good place to start in the landscape of relevant variables — chiral acids, solvents, concentrations, temperatures and times. This was exactly the kind of specific problem we felt could be solved through an open approach, because this was a highly technical issue where we did not yet know the relevant experts, but required their input.

In April 2010 a request was posted to a (closed, 2,500-member) process chemistry networking forum[12] on LinkedIn for suggestions, but also for people who might be willing to contribute more materially. This stimulated 20 comments (from 11 different people) and four private e-mails (via the website). None of these contributors were previously known to us. From the advice and offers, we chose to send one gram of racemic PZQamine to a Dutch contract research organization, Syncom, which arrived in mid-May. On 25 May, the company posted the identification of several chiral columns and conditions that enabled the baseline separation of the PZQamine enantiomers, permitting an assay for the effectiveness of any resolution attempts. On 25 August the company posted a lead chiral acid that had been identified (actually two months earlier) that effected the resolution of PZQamine. The company was not paid for this work.

The lead chiral acid that was identified (dianisoyl tartaric acid) was fairly expensive to buy, and its purification when synthesized was challenging. In addition, the desired enantiomer of PZQamine was present in the mother liquor of the resolution, rather than the solid. Nevertheless, this was a valuable lead. Optimization was performed in Sydney. All results were posted openly, resulting in the identification of dibenzoyl tartaric acid as a superior resolving agent on 8 November. Not only was this resolving agent easier to make, but gave the desired enantiomer of PZQamine in the solid. The overall process[13] now delivers PZQ with an enantiomeric excess of 97% in 27% overall yield for the three-step process of hydrolysis, resolution and re-attachment of the cyclohexanoyl group. The resolving agent can be recycled in 87% yield. The project is currently seeking ways to racemize the unwanted enantiomer of PZQamine to regenerate racemic PZQ that can be re-entered into the process[14].

At the same time as this process was being discovered with an open approach, another contract research organization was funded in parallel during 2010 specifically to devise a solution to the same problem. A consideration of available routes, as well as the commercial availability of intermediates on process scale, led to an alternative resolution not using PZQ itself as the starting point, but an intermediate available on a large scale (Fig. 1, route B). On completion of the project the results of this work were posted[15] (also on The Synaptic Leap) along with those generated through the open project. It is interesting to note how similar the eventual solutions that arose are.

It is difficult to evaluate accurately the resources that went into the open versus contract approaches. Contracted research here was used to complement open contributions with the view that eventually all results must converge in an open-source system. Open science can compete with traditional science, yet there may be other projects, in which the relevant research is intellectual-property sensitive, where contract synthesis may be the preferred mechanism. In the present case, the question of which of the two routes identified may eventually be taken on to scale-up depends to some extent on which synthetic route is used to generate PZQ worldwide. Perhaps surprisingly, this information is not clear — even to those purchasing the active pharmaceutical ingredient — owing to the relevant industrial processes involving separate companies making specific intermediates in series, as well as a degree of corporate secrecy.

Currently both resolution processes are being examined on a kilogram scale for economic viability. The open-source approach is now the basis of an open educational project[16] in which students from around the world are free to collaborate in further optimization by posting their data to an online ELN and, eventually, publishing their work.

Publicity and how to get people involved

Open projects rely on traffic, and to generate traffic the participants must engage in raising awareness of the project. In advance of academic papers, this means creating publicity. The traffic at our websites notably increased when the project was featured in news articles[17], popular blogs[18] and online videos[19,20]. For people to actually take part, we found two things to be crucial. The first is that there must be a kernel of data or activity with which people can become involved. Without a starting point, people have little to go on and no incentive to contribute. Second, the barrier to entry must be low. Thus it is essential that project summaries are up to date, and that what is required from the community is clear. It is also important that the technology and software people use to contribute is simple.

Although our use of open-source blogging and ELN platforms is a good start (because anyone can contribute without having to purchase software), a great deal of work is needed to develop a powerful, intuitive front end to an open-standard ELN. Online ELNs as a repository for primary research data should also be complemented by coordination sites, and posts in diverse other websites, to alert interested parties. Reliance on a single site ('build it and they will come') is probably unwise and ineffective. The employers of participants may need to consider how best to track and archive data generated and contributed by their staff to open projects. As we move towards an age where science is increasingly recorded in digital form, and inter-organization collaboration is more common, this is a fundamental issue.

A strategy that was considered to increase community involvement in the research was to offer a financial reward; this is a model that is currently being discussed[21] and used[22] elsewhere. In our project we wanted to operate explicitly with no reward other than peer recognition for having solved a problem and contributed to something philanthropically valuable. With typical financial reward models, the research is still conducted by isolated laboratories competing with each other and it is only the incentive, rather than the process of research, that is different. In open source, all data are shared, and there are no 'teams' as such that are aiming for a prize. The well-known analogy in software development and social policy circles is the so-called gift relationship, referring to a study showing that blood donated was of a higher quality than blood solicited[23]. This does not exclude the future possibility of combining open networks of participants with financial incentives for milestones.

Industrial versus academic input

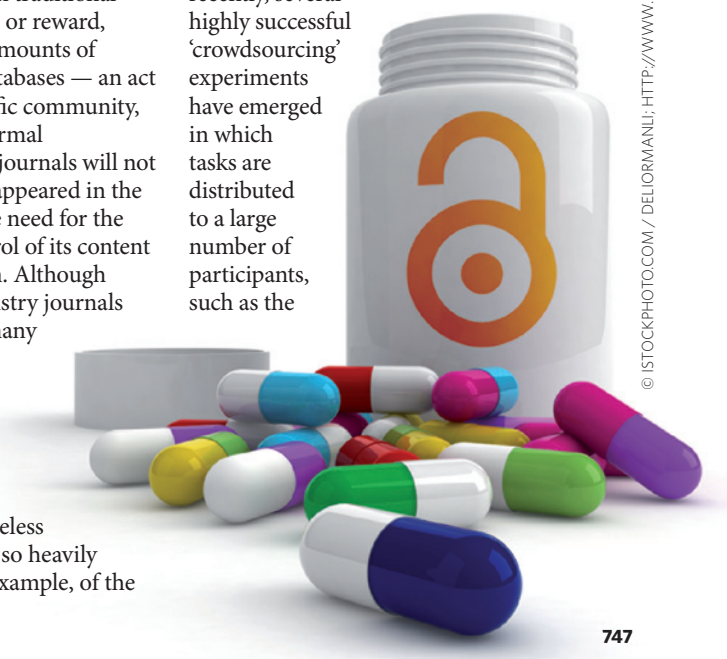
Academia is associated with the free transmission of data and resources, but in many ways this is no longer how it operates. The scientific community generally works towards common goals by competition between closed groups of scientists and communicates research results through publications relying on pre-publication peer-review. Papers frequently omit some experimental information, or ignore negative results. The delays involved in publication of papers, or reviewing of grants are significant. Many of us still publish papers in journals where comments on papers are not permitted, meaning that technical errors can remain uncorrected because rebuttals are usually required to be substantial works in their own right. Improvements to the existing state of the art are made through subsequent, substantial and stand-alone articles where there can be significant delays arising from the peer-review process of both the papers and the grant proposals required to fund the work. There have recently been isolated examples of post-publication peer-review using social networking tools[24,25], implying that post-publication peer review is gaining in popularity and acceptance.

There is perhaps also a problem with the recent rise of metrics to assess academic performance. If we assess impact based on a product of [number of papers] × [impact factor of journals] there is little room for academic activity beyond such traditional outputs. How are we to judge, or reward, someone who donates large amounts of experimental data to open databases — an act of immense use to the scientific community, yet an act that results in no formal publications[26]? Indeed, many journals will not accept work that has already appeared in the public domain, because of the need for the journal to have absolute control of its content to guarantee a revenue stream. Although many of the traditional chemistry journals follow this model, there are many others that do accept public-domain work, and where the peer-reviewed paper can act as an important summary of a project.

Industry suffers less from such metrics, but it is nevertheless surprising that industry were so heavily involved in this project. For example, of the roughly 100 comments since January 2010 on The Synaptic Leap website, around 60 came from readers not involved in the kernel project at Sydney, and of those approximately 42 came from industry, 16 from academia. Besides the input described above to the resolution experiments, a different company contributed samples of PZQ enantiomers isolated by chromatography for analytical purposes, and another company is currently determining the phase diagram of PZQamine. Why would companies choose to be involved, particularly in a project in neglected tropical diseases where there is little profit margin and no new intellectual property available? One can appeal to human nature — we see a problem we can help solve, and we find it impossible to resist stepping in, partly to showcase our abilities to our peers, and partly, in this case, because of the philanthropic nature of the project. These motivations also work on a corporate, rather than a personal level. Participation in open projects allows companies to demonstrate a commitment to worthy causes for public-relations reasons. More pragmatically, however, open projects enable companies to showcase their core competencies in real time, without the burden of client confidentiality. Companies can show the world that they can solve real problems, and quickly.

Data and citizen science

Many initiatives advocating 'open data' have emerged in which large amounts of data are deposited to assist groups of researchers[27–35]. These immensely important ventures still employ the internet as an information resource, rather than as a means for active collaboration, and groups using the data do not have to work together[36]. More recently, several highly successful 'crowdsourcing' experiments have emerged in which tasks are distributed to a large number of participants, such as the



Foldit[37] and Galaxy Zoo[38] projects. What is notable about such cases is the speed with which the science progresses through the harnessing of what has been termed the 'cognitive surplus'[39].

The active engagement of scientists in the design and implementation of open projects is rarer, but has been shown to give rise to a similar acceleration in the production of results. Examples include the Polymath project[40] in mathematics as well as the open generation of cheminformatics tools by the Blue Obelisk group[41]. In such cases the number of participants is smaller than in crowdsourcing projects, but that is because more is being asked of them. Similarly with our project, accelerating the research did not take thousands of participants, merely a small number of experienced, naturally motivated people. Nevertheless, many open-science projects so far have involved text- or code-based interchanges between scientists, and as such are easily achievable online.

In this Commentary we have described an example of a project involving experimental science being conducted in the open. The other notable and pioneering example of this approach in organic chemistry is the Usefulchem project[42], and in biotechnology research the CAMBIA BiOS initiative has pioneered the use of licensing to protect the usage of shared, experimental tools for the acceleration of innovation[43]. Inputs consisting of text-based advice were still important in the PZQ project because there was a funded kernel of activity taking place in the lab, and all data were being shared. However, what we also showed was that having effective means of sharing research data in full stimulated a distribution of the real, experimental lab work. With advances in technology, it will only become easier to collaborate in this way.

The advantages of openness

The crucial message of the open project is this: the research was accelerated by being open. Experts identified themselves, and spontaneously contributed based on what was being posted online. The research therefore inevitably proceeded faster than if we had attempted to contact people in our limited professional circle individually, in series. Perhaps this is not surprising, but if it is the case that 'none of us is as smart as all of us' and if we wish to reach scientific goals quickly, why is so much science not practised this way?

Besides speed, there are several other advantages of conducting science in the open. The process is transparent, meaning the public can be assured that funding for science, arising from their taxes, is being used responsibly and there is no suggestion of political interference in the scientific process[44]. Secondly, in open projects everything is available on the web; the project need not cease with the graduation of students, the termination of a grant or the demise of a principle investigator. Funding for the kernel effort of such a project, crucial in generating activity to which others may respond, can leverage extra input that is unfunded, and this should be attractive for funding agencies keen to maximize the impact of the relevant science. Open science is subject to the most rigorous peer review because the review process never ends, essentially because there will always be a commenting function on results, and a mechanism for the community to police those comments. The results of open science, freely available on the web, can still be published in pre-publication peer-reviewed journals that accept work that has previously been made public, because this serves as an important mechanism to summarize the research for future participants, and to reward those who have contributed with authorship along a traditional model.

Open-source drug discovery?

Although this project essentially involved open sourcing process chemistry, one cannot help but ask the question: what about open-source drug discovery? The potential impact of an open approach on the pharmaceutical industry should not be underestimated. Although there is interest in 'open innovation' in this industry (because of its current crisis regarding weak pipelines of new drugs and falling revenues) it is not clear that the science will be conducted open to the outside world[45]. There has been discussion of open-source drug discovery[46–52], but no coordinated efforts at compound discovery. Whether completely open-science efforts can provide a complementary — yet disruptive — alternative to the traditional process of drug discovery is the next interesting question. That the answer is unclear makes it worth trying. ❐

Michael Woelfle and Matthew H. Todd are in the School of Chemistry, The University of Sydney, NSW 2006, Australia; Piero Olliaro is in the UNICEF/UNPD/World Bank/WHO Special Programme for Research and Training in Tropical Diseases (TDR), World Health Organization, Avenue Appia 20, 1211 Geneva 27, Switzerland. The authors alone are responsible for the views expressed in this publication and they do not necessarily represent the decisions, policy or views of the World Health Organization. *e-mail: matthew.todd@sydney.edu.au*

References

1. Raymond, E. S. *The Cathedral and the Bazaar* (O'Reilly, 1999); available via http://go.nature.com/EpPqah
2. Hotez, P. J., Engels, D., Fenwick, A. & Savioli, L. *Lancet* **376,** 496–498 (2010).
3. Meyer, T. *et al. Plos Neglect. Trop. D.* **3,** e357 (2009).
4. http://www3.imperial.ac.uk/schisto
5. *Drug Development and Evaluation for Helminths and Other Neglected Tropical Diseases: Business Plan 2008–2013* (WHO/TDR, May 2007); available via http://go.nature.com/HXNQH6
6. King, C. H., Dickman, K. & Tisch, D. J. *Lancet* **365,** 1561–1569 (2005).
7. http://www.thesynapticleap.org/schisto/community
8. Kittur, A., Chi, E. H., Pendleton, B. A., Suh, B. & Mytkowicz, T. in *25th Annual ACM Conf. Human Factors in Computing Systems* (CHI 2007); available via http://go.nature.com/cgdyYu
9. http://www.ourexperiment.org/
10. http://www.labtrove.org/
11. http://en.wikipedia.org/wiki/Release_early,_release_often
12. http://www.linkedin.com/groups?mostPopular=&gid=1061737
13. Woelfle, M. *et al. Plos Neglect. Trop. D.* http://dx.doi.org/10.1371/journal.pntd.0001260 (in the press).
14. http://www.ourexperiment.org/rac_pza
15. http://www.thesynapticleap.org/node/333
16. http://pzq.ourexperiment.org/.
17. http://www.nature.com/news/2010/100204/full/news.2010.50.html
18. http://pipeline.corante.com/archives/2010/08/03/know_how_to_make_praziquantel_tell_the_world.php
19. http://www.youtube.com/watch?v=yWnTJw042OM
20. http://igniteshow.com/videos/open-science-we-can-all-help
21. Love J. & Hubbard, T. *Chicago-Kent Law Rev.* **82,** 3 (2007); available at http://accessvector.org/oldkei/content/view/38/
22. http://www2.innocentive.com/
23. Titmuss, R. M. *The Gift Relationship: From Human Blood to Social Policy* (Oakley, A. & Ashton, J. Eds.) (The New Press, 1997).
24. Mandavilli, A. *Nature* **469,** 286–287 (2011).
25. http://www.nature.com/news/2011/110809/full/news.2011.469.html
26. *Nature Med.* **17,** 137 (2011).
27. http://www.ncbi.nlm.nih.gov/genbank/
28. http://pubchem.ncbi.nlm.nih.gov/
29. http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml
30. http://snp.cshl.org/
31. http://www.sgc.utoronto.ca/sgc-webpages/sgc-toronto.php
32. http://sagebase.org/
33. http://www.chemspider.com/
34. http://www.nature.com/news/2010/100120/full/news.2010.20.html
35. Murray-Rust, P. *Nature* **451,** 648–651 (2008).
36. http://en.wikipedia.org/wiki/Open_science_data
37. Cooper, S. *et al. Nature* **466,** 756–760 (2010).
38. http://www.galaxyzoo.org/
39. Shirky, C. *Cognitive Surplus: Creativity and Generosity in a Connected Age* (Penguin, 2010).
40. Gowers, T. & Nielsen, M. *Nature* **461,** 879–881 (2009).
41. Guha, R. *J. Chem. Inf. Model.* **46,** 991–998 (2006).
42. http://usefulchem.wikispaces.com/
43. Jefferson, R. *Innovations: Technol. Govern. Global.* **1,** 13–44 (2006).
44. *Nature* **468,** 345 (2010).
45. Hunter, J. & Stephens, S. *Nature Rev. Drug Discov.* **9,** 87–88 (2010).
46. http://www.osdd.net/
47. Masum, H. & Harris, R. *Open Source for Neglected Diseases: Magic Bullet or Mirage?* (Results for Development Institute, Washington DC, 2011).
48. Munos, B. *Clin. Pharmacol. Ther.* **87,** 534–536 (2010).
49. Orti, L. *et al. Nature Biotechnol.* **27,** 320–321 (2009).
50. Munos, B. *Nature Rev. Drug Discov.* **5,** 723–729 (2006).
51. Maurer, S. M., Rai, A. & Sali, A. *Plos Med.* **1,** e56 (2004).
52. *An Open-Source Shot in the Arm?* (Economist, 10 June 2004); available at http://www.economist.com/node/2724420

Acknowledgements

We thank all the (sometimes anonymous) participants at The Synaptic Leap website and at other places on the internet for their support of this project, and in particular thanks go to Jean-Paul Seerden (Syncom), Harald Sekljic (Intervet Innovation) and Nick Tyrrell (Almac Sciences). A full list is given in the supporting information of ref. 13. We thank Michael Nielsen and Jonathan Eisen (University of California, Davis, USA) for helpful comments on the manuscript. Particular thanks go to Ginger Taylor and Marc Marti-Renom (Bioinformatics Department, CIPF, Valencia, Spain), co-founders of The Synaptic Leap and Tropical Disease Initiative. This research was funded by the Australian Research Council (Linkage Grant LP0883419) and UNICEF/UNPD/World Bank/WHO Special Programme for Research and Training (TDR) grants A70050 and A90461.