Comment

ILLUSTRATION BY DAVID PARKINS

A controlled trial for reproducibility

Marc P. Raphael, Paul E. Sheehan & Gary J. Vora

For three years, part of DARPA has funded two teams for each project: one for research and one for reproducibility. The investment is paying off.

In 2016, the US Defense Advanced Research Projects Agency (DARPA) told eight research groups that their proposals had made it through the review gauntlet and would soon get a few million dollars from its Biological Technologies Office (BTO). Along with congratulations, the teams received a reminder that their award came with an unusual requirement — an independent shadow team of scientists tasked with reproducing their results.

Thus began an intense, multi-year controlled trial in reproducibility. Each shadow team consists of three to five researchers, who visit the 'performer' team's laboratory and often host visits themselves. Between 3% and 8% of the programme's total funds go to this independent validation and verification (IV&V) work. But DARPA has the flexibility and resources for such herculean efforts to assess essential techniques. In one unusual instance, an IV&V laboratory needed a sophisticated US$200,000 microscopy and microfluidic set-up to make an accurate assessment.

These costs are high, but we think they are an essential investment to avoid wasting taxpayers' money and to advance fundamental research towards beneficial applications. Here, we outline what we've learnt from implementing this programme, and how it could be applied more broadly.

Engineering lessons

Engineers expect their work to be subject to an IV&V process, in which the organization conducting the research uses a separate set of engineers to test, for example, whether microprocessors or navigation software work as expected. NASA's IV&V facility was established more than 25 years ago and has around 300 employees testing code and satellite components.

In conventional electronics, IV&V relies on fundamental units such as transistors, diodes, capacitors and oscillators. The electronics industry takes great pains to promote the compatibility of these basic elements across platforms. For example, a given microprocessor developed by Intel can function in both

children's toys and advanced physics labs.

That compatibility does not apply to biological systems: proteins and cell lines are often not interchangeable, and even identical proteins can function differently in different environments. First-principle explanations of how they work often do not exist. This complicates efforts to apply IV&V approaches developed for electronics and software. What's more, increasingly sophisticated bioengineering tools are making cell-biology experiments more complicated, so thorough validation could take months or even years to complete. Few investigators have such resources.

Instead, the biological sciences have depended on other, less-reliable techniques for reproducibility. The most long-standing is the assumption that reproducibility studies will occur organically as different researchers work on related problems. In the past five years or so, funding agencies and journals have implemented more-stringent experimental-reporting and data-availability requirements for grant proposals and submitted manuscripts. A handful of initiatives have attempted to replicate select published studies. The peer-reviewed *Journal of Visualized Experiments* creates videos to disseminate details that are hard to convey in conventional methods sections.

Yet pitfalls persist. Scientists might waste resources trying to build on unproven techniques. And real discoveries can be labelled irreproducible because too few resources are available to conduct a validation. We were lucky enough to have the time, money and mandate to try something different.

Making it work

The synthetic-biology focus of DARPA's Biological Control programme is well suited to merging biological research with reproducibility studies. The programme aims to bring engineering principles of design and control to biology. By definition, this requires the adoption of best practices from the engineering community — such as IV&V — to improve the likelihood that technologies can advance.

Awardees were told from the outset that they would be paired with an IV&V team consisting of unbiased, third-party scientists hired by and accountable to DARPA. In this programme, we relied on US Department of Defense laboratories, with specific teams selected for their technical competence and ability to solve problems creatively. To get comfortable with the concept of IV&V, investigators needed reassurance that replicating teams would not steal ideas or derail publications. They also needed to get used to their results being challenged even before peer-review submission, and they needed reminders that cooperating with these teams was a programme requirement.

Results so far show a high degree of experimental reproducibility. The technologies investigated include using chemical triggers to control how cells migrate[1]; introducing synthetic circuits that control other cell functions[2]; intricate protein switches that can be programmed to respond to various cellular conditions[3]; and timed bacterial expression that works even in the variable environment of the mammalian gut[4]. In the future, we expect replication efforts will be reported as supplemental data submitted with manuscripts.

"Real discoveries can be labelled irreproducible because too few resources are available to conduct a validation."

Especially when claims border on the fantastical, it is helpful to show peer reviewers and editors that an independent party has confirmed the finding. So far, one publication co-authored by performer and IV&V teams has been accepted[5], and two more are nearing submission. Still, getting to this point was more difficult than we expected. It demanded intense coordination, communication and attention to detail.

Successfully combining reproducibility studies with fundamental research required a level of coordination between laboratories and with the programme manager (P.E.S.) that none of us had experienced before. The manager worked with each project team to determine which of their many results merited validation on the basis of the desired impact and application. We wanted to know that the engineered organism — yeast, bacteria, slime moulds, mammalian cells or something else — could be modified reliably and that these modifications performed as expected, as well as what environmental conditions were essential for that performance.

A typical academic lab trying to reproduce another lab's results would probably limit itself to a month or so and perhaps three or four permutations before giving up. Our effort needed capable research groups that could dedicate much more time (in one case, 20 months) and that could flexibly follow evolving research.

Ultimately, the technologies that DARPA is developing should end up being applied by many people for a broad range of uses. So in addition to assessing whether the technologies worked, IV&V teams had to assess robustness. For instance, we needed to know what fraction of cells would incorporate new genetic material, especially when multiple genes and control elements were involved. We tested whether cells would still work in the same way if frozen and thawed months later, and whether they would retain their functionality after being grown continuously. One IV&V team checked whether migration in a genetically modified cell line was faster than in its precursor, and fabricated guidance chips to determine what surfaces best directed cell migration.

Achieving verification means communicating effectively. Performer teams, particularly those with several principal investigators, had to designate someone to facilitate teleconferences and site visits. Both teams present jointly to the programme manager at least twice a year.

A key component of the IV&V teams' effort has been to spend a day or more working with the performer teams in their laboratories. Often, members of a performer laboratory travel to the IV&V laboratory as well. These interactions lead to a better grasp of methodology than reading a paper, frequently revealing person-to-person differences that can affect results. This is especially true when the IV&V investigator does not regularly work with the same cell type as the performer team, and thus approaches experiments in a similar way to other researchers who are building on a newly reported technique.

Real-time collaboration minimizes or avoids logistical roadblocks that are known to prevent basic research validation (for example, when the original samples cannot be located, or the postdoctoral researcher with the necessary expertise is no longer with the laboratory). Still, our IV&V efforts have been derailed for weeks at a time for trivial reasons (see 'Hard lessons'), such as a typo that meant an ingredient in cell media was off by an order of magnitude. We lost more than a year after discovering that commonly used biochemicals that were thought to be interchangeable are not. A five-laboratory consortium testing how cultured cells responded to cancer drugs reported similar experiences, with minor differences causing major effects[6].

Now, our IV&V efforts begin by cataloguing all chemicals, media and cell types, their suppliers and, for animal-derived extracts, lot numbers. Instruments are calibrated

HARD LESSONS

Recommendation	What to do	Our experience
Document reagents	Include the vendor, product number and lot number for all reagents.	We lost weeks of work and performed useless experiments when we assumed that identically named reagents (for example, polyethylene glycol or fetal bovine serum) from different vendors could be used interchangeably.
See it live	Watch an experiment carried out by another team. In our hands, washing cells too vigorously or using the wrong-size pipette tip changed results unpredictably.	Site visits are mandatory because witnessing experiments in action reveals valuable information, such as how to trap *Hydra* without harming them, or how to tilt a cell plate. The benefits of site visits in terms of achieving reproducibility are worth the cost of plane tickets and lodging.
State a range	Rather than a single number, state a range of acceptable conditions for temperature, convection and other control standards.	Knowing whether 21 °C means 20.5–21.5 °C or 20–22 °C can tell you whether cells will thrive or wither, and whether you'll need to buy an incubator to make an experiment work.
Test, then ship	Immediately before shipping cells or a genetic construct for testing, check them or it.	Incorrect, outdated or otherwise diminished products were sent to the IV&V team for verification many times.
Double check	If a standard protocol does not work, the performer and independent valuation and verification (IV&V) teams should work together on a step-by-step review.	A typo in one protocol cost us four weeks of failed experiments, and in general, vague descriptions of formulation protocols (for example, for expressing genes and making proteins without cells) caused months of delay and cost thousands of dollars in wasted reagents.
Pick a person	Each performer team should designate one person to keep communication open, accurate and timely.	The projects that lacked a dedicated and stable point of contact were the same ones that took the longest to reproduce. That is not coincidence.
Keep *in silico* analysis up to date	Data-analysis pipelines are replete with configuration decisions, assumptions, dependencies and contingencies that move quickly beyond documentation, making troubleshooting incredibly difficult.	Teams had to visit each others' labs more than once to understand and fully implement computational-analysis pipelines for large microscopy data sets.

and settings (for example, microscope light source power) coordinated to avoid unintended side effects such as phototoxicity. Routine measurements in cell culture, such as pH, osmolarity and testing for *Mycoplasma*, which often fall by the wayside, are prioritized. Each project creates a customized checklist depending on its cell lines, equipment and experiments. Without this essential level of research hygiene, troubleshooting efforts would become an uninformative time sink.

We have learnt to note the flow rates used when washing cells from culture dishes, to optimize salt concentration in each batch of medium and to describe temperature and other conditions with a range rather than a single number. This last practice came about after we realized that diminished slime-mould viability in our Washington DC facility was due to lab temperatures that could fluctuate by 2 °C on warm summer days, versus the more tightly controlled temperature of the performer lab in Baltimore 63 kilometres away. Such observations can be written up in a protocol paper.

Sometimes, validation requires new equipment. For the slime moulds, independent validation meant buying an incubator that could keep cells stably at 21.5 °C, slightly below the IV&V laboratory's ambient temperature. In another case, the performer team had to help install customized microfluidic and optical equipment at the IV&V lab because the standard microscopes and analysis software used for live-cell imaging were not up to the task.

All this makes for a considerably more variable IV&V programme than is found in microelectronics. But without these efforts, some promising technologies could have been abandoned prematurely as seeming dead ends.

Big dividends

We think that the IV&V programme brings benefits beyond reproducing any individual project. Now, there is a process to make investigations of disparate results more transparent. Performing reproducibility studies invariably forces scientists to think more deeply about their own experimental protocols and techniques. As one of our scientists said, "IV&V forces performers to think more critically about what qualifies as a successful system, and facilitates candid discussion about system performance and limitations." Trainees told us that they have gained skill in analysing data, providing constructive criticism and designing and documenting their own research so that it can be reproduced.

IV&V teams gained further advantages. For example, because service laboratories become well-versed in the mindset and protocols for new technologies even before publications appear, they are well-poised to integrate them into their offerings, predict future directions for the field and move research more quickly to applications. The IV&V programme also expands networking opportunities between DARPA scientists and the top-quality labs DARPA funds, including the potential to recruit postdocs and graduate students across laboratories. Not surprisingly, many DARPA BTO programmes in recent years have incorporated some form of IV&V to help validate programme results.

As we continue the Biological Control IV&V programme, we expect to find more ways to improve it, to better quantify its benefits and to codify best practices, such as incorporating automation and robotics where possible and keeping an open line of communication between performer groups and IV&V teams. Although some of the lessons learnt from the first stages might seem obvious and trite, that also reinforces their necessity.

We think that a dedicated shift towards the IV&V model by more research institutions and funding agencies will bring more reliable and cost-effective science. Programme officers at other granting agencies should consider allocating a portion of their funding stream to independent reproducibility efforts. This will both reduce the number of papers that cannot be replicated and improve the quality of work that funding agencies support. Metrics will need to be established to quantify the cost savings of applying this model to synthetic biology and bioengineering, but given its successful integration throughout more conventional engineering disciplines, we are optimistic that the returns will be worth it.

The authors

Marc P. Raphael is a biophysicist at the Naval Research Laboratory in Washington DC, USA. **Paul E. Sheehan** is a programme manager in DARPA's Biological Technologies Office in Arlington, Virginia, USA. **Gary J. Vora** is a biologist at the Naval Research Laboratory in Washington DC, USA.
e-mail: marc.raphael@nrl.navy.mil

1. Miao, Y. C. *et al. Nature Cell Biol.* **19**, 329–340 (2017).
2. Ng, A. H. *et al. Nature* **572**, 265–269 (2019).
3. Langan, R. A. *et al. Nature* **572**, 205–210 (2019).
4. Riglar, D. T. *et al. Nature Commun.* **10**, 4665 (2019).
5. Ames, C. L. *et al. Commun. Biol.* **3**, 67 (2020).
6. Niepel, M. *et al. Cell Sys.* **9**, 35–48 (2019).