Bench pressing with genomics benchmarkers

Some -omics tools can be more accurate, sensitive or efficient than others. Yet benchmarking is no tell-all.

Vivien Marx

Celebration is in order: a lab's new software tool is testing well and ready for wider distribution. Then, it fares poorly in a tool comparison by another lab. Later, in a larger competition, the tool ranks near the bottom. Gone is the lab's celebratory mood.

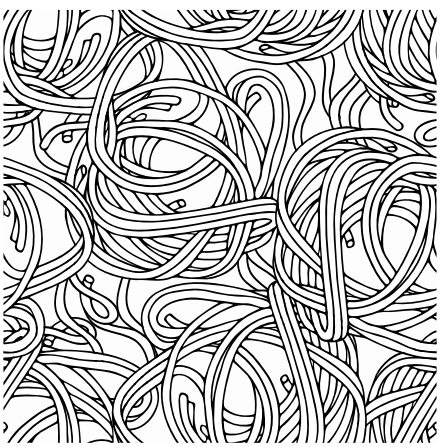
A tool might not rank as best overall, but if "it's really precise," that can be motivation to tweak the tool, says Serghei Mangul, a bioinformatician at the University of Southern California. Poorly chosen parameters can skew a comparison, says Kasper Lage, a computational biologist at Massachusetts General Hospital and the Broad Institute of MIT and Harvard. But even with well-established methods, optimized parameter settings can be hard to find. Some tools are too hard to set up. "You never get them to run," says Wolfgang Enard from Ludwig Maximilian University of Munich (LMU).

In a review[1], Mangul and colleagues note that 72% of benchmarking studies they surveyed exclude information about how computationally efficient a method is. "I was really surprised," says Mangul. In addition to knowing how precise or sensitive a tool is, it helps to know about its computational needs, such as how much memory is needed for it to work optimally. Often, bioinformatics methods are developed by postdoctoral fellows or students less familiar with best software engineering practices, says Lage. "So obviously there will be a certain 'spaghetti code' aspect to the pipelines," he says. Labs are on their own to untangle poorly organized and hard-to-debug code. Benchmarkers have some suggestions for users and developers, even those with pasta issues.

Benchmarking and access to benchmarking results for scientific tools should be easier, perhaps more qualitative and humanly informative, says Marc Salit, who is at SLAC National Accelerator Laboratory. "So long as benchmarking is difficult to access and esoteric, the methods will be limited in their impact," he says. Credit: TCmake_photo / iStock / Getty Images Plus

Gold, copper, truth

It's challenging to have a gold standard on hand for assessing what is 'good', says Justin Guinney, who directs the competition Dialogue on Reverse-Engineering Assessment and Methods (DREAM). He also heads computational biology at the non-profit Sage Bionetworks, on whose platform DREAM is run, and teaches at the University of Washington. Sometimes, says Gustavo Stolovitzky, the previous DREAM director and a researcher at IBM and Columbia University, the absence of a gold standard can be addressed by letting teams make predictions. Then the challenge organizers verify those on data that include additional measurements, as has happened in a challenge on RNA splice isoforms. The absence of gold standards leads organizers to think about "how to create good, even if imperfect, gold standards, which we sometime call 'copper' standards," says Stolovitzky.



Bioinformatics tools and pipelines can have a certain 'spaghetti code' aspect to them, says Kasper Lage, MGH/Broad Institute, since they are often not written by people who know software-engineering best practices. Credit: Zhemchuzhina / iStock / Getty Images Plus

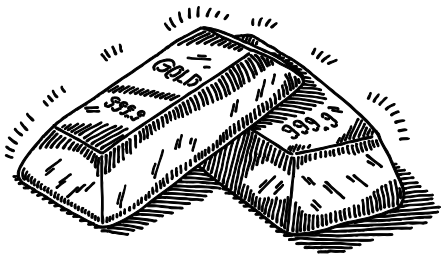
Ground-truth networks are "still a very thorny problem," says T. M. Murali, a bioinformatics researcher at Virginia Tech. As he and his group explored ways to infer gene regulatory networks from single-cell transcriptional data, they were dissatisfied and began benchmarking methods, which "mushroomed into the BEELINE[2] project," says Murali. It's a pipeline for labs to evaluate how accurate, robust and efficient algorithms are. The team's prep work included dockerizing all the algorithms to also enable uniform access in spite of the tools' language diversity: R, Matlab, Python, Julia and F#. Murali's group needed ground-truth networks to simulate datasets. Those ground truths had to be compiled: the team used fully synthetic networks curated from the literature, such as a comparison and pipeline of single-cell trajectory inference methods with datasets from researchers at Ghent University and elsewhere[3]; they also

curated various models of cellular processes from the scientific literature; and they built BoolODE, a computational environment in which a Boolean function describes every gene in the network. When they simulated single-cell transcriptional data from these networks and tested the algorithms, some were more accurate than others. The ranking of algorithms according to how well they did on simulated datasets "sort of flipped when we ranked them on the Boolean networks," he says. The ranking of the methods on Boolean models was similar to rankings achieved with experimental single-cell RNA-seq datasets, which suggests the Boolean models are similar to "the real ground-truth networks," he says. There's a need for methods in this space, even though "the ground truth is still very hard to define."

Real data, simulated data

"Simulated data is imperfect," says Guinney, but it's been essential in many of the DREAM Challenges, especially for genomics tool evaluations. With simulated data, one has to carefully state all the assumptions that went into data generation so the tool developers and users can understand the limitations of what is known and what can be evaluated. The data used for benchmarking should represent the problems faced by people interested in the solutions, says the Broad Institute's Juan Caicedo, who has run a competition in micrograph image analysis. "Simulated data is not good for benchmarking, but sometimes there is no other option because the data is expensive to create or cannot be made public," he says. Data availability is definitely a limitation with comprehensive benchmarks. There are roles for real data and for simulated data, "but the limitations of both should be understood," says Justin Zook, who co-leads the public–private consortium Genome in a Bottle (GIAB), hosted by the National Institute of Standards and Technology (NIST) and focused on genomics reference materials and data. Purely simulated data usually cannot represent all the biases and errors that occur in real data, he says, though

Whether or not to use synthetic data has been present since the first DREAM Challenges discussions, says Gustavo Stolovitzky. Credit: IBM



For assessing what is good, one does not always have a gold standard on hand, says Justin Guinney, who directs the DREAM Challenges. Credit: F. Ramspott / DigitalVision Vectors / Getty

there are approaches that modify real data. That includes introducing variants into reads or mixing real data from two samples together. Those do a better job of representing the biases and errors that occur in real data. The disadvantages of real data are that the truth may not be known, says Zook, such as in the 'truth' about the most difficult regions of the genome. Scientists might not be able to get real data from samples that have the variants of interest to a lab, such as "rare, clinically interesting variants." Not only should performance from real and simulated data be interpreted with care, a method's true performance will likely be worse than the performance benchmarked with real or simulated data. The discussion of whether or not to use synthetic data has been present since the first DREAM Challenges discussions, says Stolovitzky, "and given the complexities of biology, it will continue to play a role in the foreseeable future." Before launching the first DREAM competition he and his colleagues thought about consensus, and in a paper[4] on the subject they point out that in silico networks offer an "ideal model." "However, even the most biologically inspired among synthetic models is far removed from an actual biological counterpart," they noted. But given the role those networks play in the assessment of reverse-engineering methods, they decided to include them in the DREAM Challenges.

Mangul thinks labs should try to use both simulated and experimental data, if available. With simulated data, labs can explore many parameters, such as different levels of genomic coverage or parameters related to the sequencer instrument itself. If experimental data are not available, one can try to use the properties of the real data to give the simulated data some of those properties. One can insert mutations into simulated reads and, with high-coverage whole-genome data, one can use the many reads to call single-nucleotide polymorphisms at 1× coverage. "That can potentially be a gold standard," he says. Then one can computationally 'subsample': reduce the coverage to 1× so labs can benchmark tools on how well they call SNPs at 1× coverage. "You need both," says LMU researcher Ines Hellmann of simulated and real data. "We tried to make simulations as realistic as possible," she says. The team used powsimR to estimate, simulate and evaluate single-cell RNA-seq experiments. A lab will want to check whether the algorithm can do what the developers expect it to do and assess simulations to check which real situations are captured. One needs a ground truth — a true-positive rate and a false-negative rate. But, she says, in single-cell techniques it's difficult to get replicates for a single cell, so there's no way to get around simulations.

It's still early days for the single-cell field, says Guinney. At some point the field may coalesce in a way that mirrors events with microarrays over a decade ago. The MicroArray Quality Control project addressed issues related to reliability and reproducibility. "Right now, it feels like there are too many degrees of freedom to robustly benchmark single-cell pipelines, but hopefully these will reduce in time to make the comparison problem tractable," he says. He and his colleagues are running single-cell challenges today, but these focus on a downstream question using single-cell datasets derived from a common platform.

Making benchmarks

Especially in genomics, with its wide tool-range, users save time when a benchmark includes metrics such as accuracy, running speed, ease of use or deployment, reproducibility and stability, says Caicedo. Benchmarks need rigor, and if whoever ran the benchmark "forgot to incorporate realistic assumptions in the experiments, then the benchmark may not be completely useful," he says. For an individual lab's focused needs, flexible benchmarks might work well. All benchmarks are good, he says, but results need to be interpreted within a benchmark's scope. That avoids faulty generalizations about a software tool. There are software benchmarks, benchmark samples and combinations thereof. When using benchmark samples, labs can use their analysis method of choice and then test software tools to see how they perform against the benchmark set, says Zook. For variant calling, labs can use NIST GIAB reference materials, analyze with their preferred software tools, then compare using a benchmarking framework for 'calling' genomic variants developed by the Global Alliance for Genomics and Health (GA4GH) and others including GIAB[5,6]. Illumina offers a 'truth set' called Platinum Genomes, a

whole genome sequenced at 50× depth. There are 'small variant truth sets' from GIAB and Platinum Genomes, such as variant calls from a pedigree of 17 people from 3 generations in the so-called CEPH/UTAH Pedigree 1463. Academics, government agencies and companies, too, can join in on benchmarking. As Zook explains, both for GIAB and for the GA4GH Benchmarking Team, companies have contributed valuable data analyses and expertise. "They also help us minimize biases in our benchmarks against particular technologies," he says. When GIAB evaluates draft benchmarks, the team asks community members to compare results obtained with their method to the benchmark and manually curate differences. "This helps ensure the benchmark accurately identifies errors in results from a variety of methods," he says.

Backstage peek

As Stolovitzky explains, a member of his group took part in a DREAM Challenge only after leaving the group. He was confident his tool would rank well, but its placement was mediocre. "He told me that the experience made him appreciate the pitfalls of overfitting," says Stolovitzky, who invented DREAM. Stolovitzky hopes that all participants, independent of ranking, can enjoy that the competition's collective experience usually leads to a high-profile paper.

Lage says that a benchmarking competition does not always capture the diversity of a tool's value, which is comparable to asking how good a baseball player is by testing how quickly he or she hits or runs under very controlled circumstances. One needs realistic expectations about benchmarking and data challenges, says

A benchmarking competition does not always capture the diversity of a tool's value. It's like asking how good a baseball player is by testing how quickly he or she hits or runs in very controlled circumstances, says Kasper Lage, MGH/Broad Institute. Credit: E. Dewalt / Springer Nature

Ines Hellmann, a computer scientist, and Wolfgang Enard, a biologist, enjoy working together. He feels like the luckier one. Hellmann could work with any experimental biologist, but he would have a harder time finding the right computational colleague. Credit: C. Bleese

Guinney. Both are good when posing specific questions with a set of associated metrics. Whether the question is a good one or metrics are correct or weighted appropriately "will always be open to criticism," he says. Around 70% of DREAM Challenge prep, he says, is spent on metrics, because the organizers know these will heavily determine how people approach the question's solution and the expected learning from the challenge. "Many participants are very good at 'gaming' the metrics," says Guinney. They optimize tools to perform well in a challenge, but "it doesn't always reflect 'real life,' as there are often many other factors to consider that cannot be captured in one — or even several — metrics." Designing a challenge, he says, "gets people thinking deeply about what 'good' performance means and how it can be evaluated in a specific domain." As an unbiased assessment, it convenes a community at a specific time and place to define and to try to exceed the state of the art.

Benchmarking competitions, challenges, and benchmark dataset and tool development complement one another, says Zook. The work in GIAB and GA4GH is about building ways for the community to assess and optimize the performance of sequencing and analysis tools, be it for an individual lab or for a benchmarking competition. "At their best, benchmarking competitions energize the community to represent the state of the art at a point in time, but it is challenging to keep up to date with continually evolving sequencing technologies and analysis methods," he says. A limitation of both approaches — benchmarking competitions and challenges and benchmark set and tool development — is that "they tend to ignore the most challenging regions of the genome, where no benchmarks yet exist," he says, "so results should be interpreted critically." There is a need for ongoing benchmark set and benchmark tool development as new technologies and analysis methods enable characterization of increasingly challenging variants and genomic regions.

Small-lab perspective

Benchmarking is not what Enard, Hellmann and then-PhD student Beate Vieth and colleagues set out to do. The young technique of single-cell RNA sequencing intrigued them, says Enard, as a way of exploring differentially expressed genes relevant to brain development and characterizing induced pluripotent stem cells. Their assessment of tools[7] converted them, a small team without a server farm, to benchmarkers for a little over two years. "We just kind of did it somehow," says Hellmann. The LMU team took both computational and wet-lab aspects into account as they evaluated a total of around 3,000 potential analysis pipelines used to analyze scRNA-seq data, including mapping, imputation, normalization and differential expression approaches. The choice of normalization method turned out to have the largest effect on performance, the magnitude of which was unexpected, says Hellmann. "Using a bad normalization method is similar to having to sequence four times more cells," says Enard.

Benchmarking can yield results that go against the intuition of a wet-lab biologist like himself, says Enard. A lab might pick the most sensitive method. But if the sensitive one is twice as sensitive yet ten times pricier, the same sum of money gives the lab five times more cells. When researchers look at genes expressed at low levels, one method might detect the genes in ten percent of the cells and the other method in five percent of the cells. That's when it's better to make measurements in more cells, he says. Wet-lab intuition might not lead to the best decisions — for example, about sample size or choice of method. The process gave them a deep understanding of tools for processing scRNA-seq data, says Enard,

The most expensive method might not always be the best one in all situations, and wet-lab intuitions don't always work, says Wolfgang Enard, LMU Munich. Credit: F. Ramspott / DigitalVision Vectors / Getty

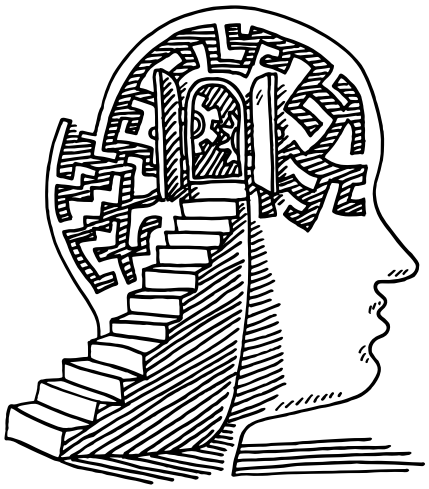
which is relevant for individual labs. A user wants to know what question a tool was developed for and know about assumptions built into the tool, says Hellmann. "That's why I got into benchmarking, because I don't like to use things I don't understand," she says. Most tools will have their benefits and applications, she says. "There might be a niche for most of them." When a technology begins to emerge, benchmarking is crucial for individual labs, says Enard. Over time, usage converges on standards. Then big consortia enter the picture, as has happened with RNA-sequencing, and they can do large-scale comparisons that small labs cannot.

Hellmann points out that many in the computational space like methods that run faster than others. But some labs have precious samples, which is when it makes little difference if a method takes a few extra hours, she says. Enard, a biologist, and Hellmann, a computer scientist, say they enjoy working together. In benchmarking both skill sets are needed, especially as tools get more complex, he says. He feels like the luckier one. Hellmann could choose to work with any experimental biologist, but he would have a harder time finding the right computational colleague.

Continuous benchmarking

Marc Salit, a colleague of Zook's, says that benchmarking is hard because "we have no 'perfect' samples or datasets." Salit used to be at NIST and now directs the Joint Initiative for Metrology in Biology at SLAC National Accelerator Laboratory and teaches in Stanford University's departments of bioengineering and pathology. Synthetic data may be redundant but are also not perfect. Many aspects of performance need to be measured and benchmarked in genomics, and sometimes those aspects are mutually exclusive. There is plenty that makes benchmarking imperfect: that lack of perfect samples or data, he says, and the fact that "the metrics aren't very predictive of performance on an arbitrary sample being queried for an arbitrary question." Another factor is "a constantly shifting landscape," he says. Some benchmarks reported at a single point in time become obsolete quickly. This has led to the concept of ongoing evaluation.

Ongoing benchmarking is tough but something to strive for, says Murali. He is setting up a continuous integration framework such that BEELINE can integrate a new dataset or algorithm and generate results. Some methods involve many parameter searches. Add in the number of datasets and algorithms and "it's a pretty massive computation," he says. "I hope by the end of this semester we will have something up and ready." Given that new



When developers of a tool test it, they might test it in an unsystematic way, which leads to the 'self-assessment trap', says Serghei Mangul, USC. Credit: F. Ramspott / DigitalVision Vectors / Getty

technologies and analysis methods enable characterization of increasingly challenging variants and genomic regions, ongoing benchmark set and benchmark tool development is needed, says Zook.

At DREAM, Guinney says he and his colleagues are developing ways to set up continuous benchmarking. For example, they ask participants to submit their algorithm as a re-runnable Docker container. "This allows us to evaluate old algorithms over time as new data become available," he says. "Similarly, we can evaluate new algorithms on old data." They call this approach "model-to-data" (M2D). Cloud-ready software packages can help to avoid issues such as the diversity of software architectures or file formats, which can make it hard to reproduce and reuse methods in a given competition. With M2D data, the underlying dataset is not visible to users and the computing runs in environments that, for example, heed data privacy rules so they can still get an assessment of an algorithm's performance while maintaining data privacy. DREAM has run several M2D challenges and encountered a bundle of logistical and technical issues, such as estimating computational resources the methods need. Continuous benchmarking avoids frozen-in-time benchmarked results. The Kaggle imaging competition that Caicedo and colleagues run is "never ending," says Caicedo. As labs develop methods, people can try new things and the leaderboard updates continuously. "You wake up one morning and your method is not in the top of the list anymore," he says. The field of computer vision has a few live benchmarks, "but I'm not aware of such benchmarks in biology." Perhaps this is because biologists do not like to reuse data to optimize methods, and there is a danger of overfitting algorithms to data. Fresh data might present difficulties for methods, he says. That means one would need to keep adding new test cases to the challenge to keep it alive and prevent saturation.

Die-hard habits

Sometimes habits die hard. TopHat was long a highly used tool, a spliced read mapper for RNA-seq reads. But it was highly used even after the developers released a successor. The developers clearly advise to use TopHat's successor, HISAT. They note on their web page that TopHat is in a "low maintenance, low support stage as it is now superseded by HISAT2," which serves the same purpose more efficiently. "I just had a paper to review," says Hellmann. "They use TopHat." It's not the fault of the TopHat developers, says Enard. It's that other scientists are using it. Perhaps it's on a Galaxy server and someone neglected to check for the release of a newer tool or a different version of the tool, says Hellmann. "At some point they probably installed TopHat when it was still the tool of choice."

Bioinformatics does not talk much about tool retirement, says Mangul, and labs use TopHat, which may or may not be the best tool. At one time it, or any tool, was perhaps the best or only available one and people liked the results. He believes that bioinformatics can shape science in a methodological way, especially now that so many public datasets are available. These data can be used to make the kinds of discoveries possible in a wet lab. "In that sense, benchmarking is crucial because we want to use the best tools." Many aspects can get in the way, such as the "self-assessment trap," which is when the developers of a tool test it themselves in an unsystematic way. "Best," he says, can be defined in many different ways in many matrices, but there will be a "most suitable" tool for a given task. ❐

Vivien Marx
Technology editor for Nature Methods.
e-mail: v.marx@us.nature.com

Published online: 20 February 2020
https://doi.org/10.1038/s41592-020-0768-1

References
1. Mangul, S. et al. *Nat. Commun.* **10**, 1393 (2019).
2. Pratapa, A. et al. *Nat. Methods* https://doi.org/10.1038/s41592-019-0690-6 (2020).
3. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. *Nat. Biotechnol.* **37**, 547–554 (2019).
4. Stolovitzky, G., Monroe, D. & Califano, A. *Ann. NY Acad. Sci.* **1115**, 1–22 (2007).
5. Krusche, P. et al. *Nat. Biotechnol.* **37**, 555–560 (2019).
6. Zook, J. M. et al. *Nat. Biotechnol.* **37**, 561–566 (2019).
7. Vieth, B. et al. *Nat. Commun.* **10**, 4667 (2019).