The difficulty of a fair comparison

Comparing methods in a fair and informative manner is often not straightforward. Benchmark data sets, thoughtfully applied metrics and clear reporting can help.

Contributors to *Nature Methods* will be familiar with the request to put their new or improved method or tool in the context of existing ones. How does its performance compare? Is it faster, brighter, more accurate, more specific, easier to use or implement? Does it make compromises on some aspects of performance so that it may soar with others? It is rare, after all, for a method to access an entirely new biological space such that these comparisons cannot be made.

Yet a fair comparison of methods performance is no easy matter. Even if researchers undertake a direct, side-by-side comparison, they may often be testing their newly developed approach against other methods in which they are not expert or for which documentation is insufficient. Furthermore, many scientists understandably do not wish to cast the work of others in a negative light by showing deficiencies in existing methods.

Even comparisons of relatively simple tools such as fluorescent proteins or affinity binders, which can in principle be based on defined, measurable properties, are easily confounded. Which conditions are used to measure photostability will affect the reported photobleaching rate of a fluorescent protein, for instance. Or, a binder with high affinity for a linear epitope presented *in vitro* may not recognize a folded protein or may have poor penetration into tissue; comparisons of binding affinity may thus not be informative about performance in specific applications.

And most methods are far more complex than these examples. Performance typically depends on several other experimental or computational steps in the overall approach, for instance, the quality of the library used for a profiling method, the sample preparation for structure determination, or differences in performance between older and newer instruments. It is not trivial to identify the contribution of a particular methodological development in comparisons where more than one component of the approach has changed.

And yet, it is undoubtedly important that the performance of a new method be vetted in light of what is already possible and that guidance be provided to potential users. Though we are very conscious of the difficulty of good methodological comparisons, and temper our expectations accordingly, some comparison to current methods—whether directly or via the literature—is unavoidable. In addition, authors should report clearly and in detail what was actually compared: describe modifications to the original protocol for experimental comparisons, or report which version of previous software was used for computational ones.

Studies expressly designed for methods comparison are a more rigorous solution to the problem. These are often the result of community-wide competitions (*Nat. Methods* **11**, 695, 2014), with the compared methods implemented by expert users on a common data set. Other studies test intra- and interlaboratory variability by treating the same sample with variations on the same overall methodology—stem cell culture methods, for example, or analysis of protein complexes with affinity purification–mass spectrometry. We continue to encourage systematic comparisons of methods and tools and to consider them for publication in the Analysis format (*Nat. Methods* **9**, 111, 2012).

Though systematic methods evaluation remains the exception rather than the rule, community-approved benchmarks can be tremendously useful even outside the confines of such a study. For proper methods comparison, typically needed are both reference data sets to which any relevant method can be applied and quantitative measures of performance.

Such benchmarks have been developed in some fields. High-throughput methods to identify protein-protein interactions, for example, are now widely assessed using literature-curated reference data sets together with statistical performance measures. For genomic analysis tools of various types, simulated data with known properties constitute useful benchmarks, though these cannot entirely replace performance tests on experimental data. As for metrics to quantify a method's performance, these can take the form of common statistical measures such as the false discovery rate or metrics designed specifically for a particular application. For example, in stem cell culture and differentiation—an area in which methods are bedeviled by variability—gene expression–based scores to define pluripotency or differentiation help to quantitatively assess a method's performance *vis-à-vis* other available options.

What constitutes a good benchmark will obviously vary depending on the methods being compared and the biological applications in question. The definition of such benchmarks should be an ongoing discussion in every field. Though the ultimate test of a method's power is how it performs 'in the wild', standards that any scientist can use to test a (new or existing) method will promote reproducibility and help researchers decide which methods are worth an investment of their money, energy and time.