LENA BJÖRK BLIXT/LUND UNIV.

No more hidden solutions in bioinformatics

Precision medicine cannot advance without full disclosure of how commercial genome sequencing and interpretation software works, says **Mauno Vihinen**.

Last month, California became the latest in a series of places to launch a 'precision medicine' project, aiming to develop diagnostic tools and treatments based on individual genomic data. Advances in sequencing technology have already made the US$1,000 genome a reality.

Producing genomic data is now relatively easy, but analysing these data is not. For precision medicine to fulfil its potential we need to identify genetic variation between individuals, and then work out which variants have a role in disease.

Human genomes are very similar, but the 0.1% difference between them still leaves millions of variations between individuals. Most such variations have little or no effect. Working out whether a particular deviation from the reference genome is important, and how, is complex and time consuming. It has become the crucial bottleneck in the precision-medicine process.

Drawing the connections between genetic variants and disease is largely the work of bioinformatics. Conventionally, the computer software used has been written and shared by academics. But as the production of genomic data has exploded, commercial firms have increasingly started to offer their own software. This growing market is evident to anyone who attends major genetics conferences. Three or four years ago, just a handful of these companies exhibited; now there are dozens.

The appeal of commercial bioinformatics packages is obvious. They are relatively simple to use, with well-designed interfaces that allow even non-experts to process complex genomic sequence information. Some commercial software streamlines the whole process, from sequencing to analysis and interpretation. The companies guarantee technical support, which is not always available with open-source software.

However, there is a major problem. Companies are generally unwilling to reveal how their software works: they do not wish to disclose the methods and data used to construct the algorithms or the details of performance. It is impossible to check the programs' quality and to compare them. Companies are selling a pig in a poke.

In such tools, the details really matter. Short segments of sequence (reads) must be joined to build complete genomes. This is difficult, and fast sequencing methods have quite high error rates, which must be taken into account when software flags possible variants. The more overlapping reads that a sequencing project includes, the better the results. Typically the coverage is in the tens, but in really deep sequencing it can be in the thousands.

Once possible variants are identified, a different set of techniques is used to filter and sort them, and then to annotate them to suggest possible clinical relevance. There is no single correct way to do this, and various academic groups have produced distinct tools that all perform these tasks slightly differently. That is why the output alone — the variants and their link to disease — is not sufficient to judge their clinical relevance. We must know how the result was obtained and how the raw data were processed.

Academics are up-front about this, and are happy to show their working. This allows comparison, and a number of studies have been published in which the performance of several methods has been checked against independent benchmark data sets. These studies allow end-users to select the most suitable tool and get an idea of how reliable it is. This information should be included when data are published, especially if it has a direct clinical relevance. The journal *Human Mutation* demands it for studies that use and develop these tools.

RESEARCH MUST BE BASED ON OPENNESS AND FULL ACCOUNTS OF THE TOOLS USED.

At present, it is impossible to check the performance of commercial software in this way. I have asked companies to give me the relevant details, but they have refused. They all say that their method is the best, but offer no way for customers to verify that. As the market grows for these commercial packages (many of which, ironically, are based on open-source academic programs), so will the scale of the problem.

The way to sort this out is to test each of the different commercial programs with established benchmarks — data sets with known variant outcomes. But even if I were to buy a licence to use each of them (and these are not cheap), I would still be unable to do the comparison. The algorithms that drive such software are often developed using the same data sets. To make the tests fair, we need to know how the algorithm was trained, so as to avoid using the same variants for both training and testing. This is something that the companies are unwilling to reveal.

Companies expect users to accept their 'black-box' solutions without knowing anything about the algorithm, training details, data sets used, method performance and use of benchmark data. This is not acceptable. Research must be based on openness and full accounts of the tools used.

Precision medicine must be evidence-based medicine. And evidence-based medicine is exactly what the name says. I understand that companies need to keep some trade secrets, but disclosing the information I discuss here will not jeopardize their competitive edge. These are details that we as the community have to demand if companies want to sell their products and services to us. ■

➲ NATURE.COM
Discuss this article online at:
go.nature.com/elxesv

Mauno Vihinen *is professor of medical structural biology at Lund University, Sweden.*
e-mail: mauno.vihinen@med.lu.se