TOOLBOX

A BIOINFORMATICS WORKSHOP IN A BOX

Bioconductor gathers genomics tools and data into a handy package that can run in the cloud. We took it for a test drive.

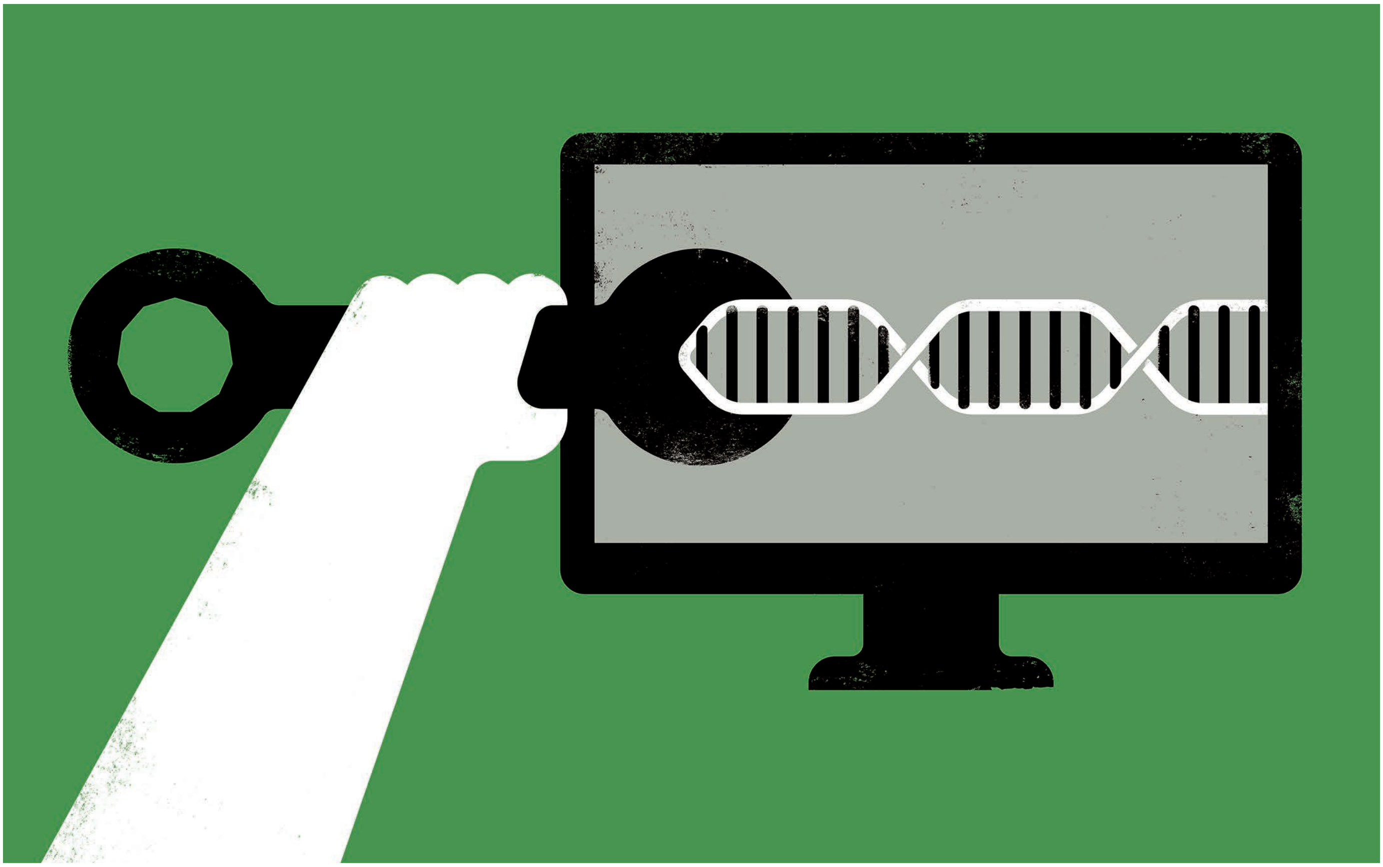

ILLUSTRATION BY THE PROJECT TWINS

BY W. WAYT GIBBS

Cloud-computing services offered by companies such as Amazon, Microsoft and Google have put high-performance computing in the hands of everyday researchers. And in the past decade, a project called Bioconductor has done something similar for the often complex field of bioinformatics.

Launched in 2001 by a group of bioinformaticians led by Robert Gentleman, then at the Harvard T. H. Chan School of Public Health in Boston, Massachusetts, the Bioconductor project offers a collection of software that makes it easy for researchers and engineers to analyse, visualize and share genomic data. The project has assembled thousands of tools for computational molecular biology, all of which work together in the R statistical-programming language in conjunction with the RStudio programming environment.

Like R and RStudio, Bioconductor is open source; all three can be downloaded and installed for free. But genomics data sets can be large and require more processing power, memory or disk space than researchers have at hand. Fortunately, Bioconductor also comes in the form of ready-to-use configurations that can be run using the near-limitless resources available on the Amazon cloud or similar services, at prices starting below US$0.20 an hour.

In effect, the cloud has made computing hardware cheap, and Bioconductor has made genomics software free. Good documentation, tutorials and courses make these technologies accessible to informatics experts and beginners alike. The promise of Bioconductor running in the cloud is to cut the cost of computational biology while flattening the learning curve for molecular biologists and boosting productivity.

TEST DRIVE

But is it really as straightforward as it seems? I talked to Bioconductor experts and novices — and then took this genomics workshop-in-a-box for a spin myself.

The first step was to choose a task and some experimental data. There's no shortage of options. The latest update (3.6), released at ▶

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

BEGINNER'S GUIDE

Dive into Bioconductor

Whether you prefer formal training or going it alone, you'll find many helpful resources for climbing the Bioconductor learning curve.

Bioconductor is described thoroughly in a 2015 Perspective in *Nature Methods*[4]. But for beginners, geneticist Michael Love of the University of North Carolina at Chapel Hill recommends attending a Bioconductor meeting or workshop. The event schedule, as well as slides and code from previous meetings, is available online.

A companion video to this article (see go.nature.com/2antzqz) covers the basic steps for setting up a Bioconductor server on the Amazon cloud and using it to do RNA transcriptomic and differential-expression analyses. Many other tutorials are available on the Bioconductor YouTube channel.

Johns Hopkins University in Baltimore, Maryland, offers an introductory course on the online classroom system Coursera taught by bioinformatician Kasper Hansen, one of the most active and longest-serving contributors to Bioconductor.

A free, self-paced option is available from the online-learning site edX and Harvard University in Cambridge, Massachusetts, as part of a series of courses that also cover the mathematics and data-science concepts that underlie computational genomics.

Dozens of basic and advanced task-oriented workflows are on the Bioconductor website. Many workflows have been peer-reviewed and published on the Bioconductor gateway at F1000Research.com. The Bioconductor site also includes complete documentation for all packages and data sets, as well as user-reviewed tutorials. W.W.G.

▶ the end of October, includes nearly 1,500 software packages, 326 experimental data sets and 911 annotation tools. The PubMed literature database indexes roughly 1,000 articles that mention using Bioconductor in diverse ways. One 2017 study[1] deployed the software to study gene-expression patterns in individuals with a potentially fatal parasitic disease before and after treatment. Another analysed metabolite profiles from drought-stricken vineyards[2].

According to Michael Love, a geneticist at the University of North Carolina at Chapel Hill, Bioconductor has become "fairly standard for various tasks in high-throughput genomic data, such as gene-expression analysis, and for harmonizing genomic annotations from various sources". The packages are also increasingly used for epigenetics and metagenomics, image processing and proteomics.

Love now requires students in his introductory graduate-level course in computational biology to use Bioconductor. So, I sought the advice of Anqi Zhu, a doctoral student of Love's who started using Bioconductor about a year ago to perform differential-expression analysis of transcriptomic data. Zhu recommends working through the many tutorials and manuals, as well as how-to demos called vignettes, at Bioconductor.org (see 'Dive into Bioconductor').

Also available are peer-reviewed tutorials, known as workflows, which are updated as the platform evolves. One, co-authored[3] by Love, walks readers through a differential-expression analysis of RNA-sequencing data. I used his workflow to guide my exploration.

I had previously opened an account on Amazon's Elastic Compute Cloud (EC2) and had configured cloud servers before. But for those who are new to cloud computing, Bioconductor provides step-by-step instructions. Installing Bioconductor on an EC2 server with 4 processing cores and 16 gigabytes of memory required simply entering the code for my desired Bioconductor configuration, selecting a few options and clicking 'launch'. In less than an hour, I had booted up the server, logged into the RStudio software running on that server and begun my tour of Love's workflow. It mines RNA-sequencing data collected from human airway cells to identify genes that are differentially expressed when the cells are treated with a corticosteroid.

LEARNING CURVE

Most work in Bioconductor is done by typing R code directly into the RStudio command prompt, rather than by pointing-and-clicking with a mouse. The RStudio development environment executes the R commands and spits out the results. It also offers interactive explanations of both R and Bioconductor functions, and it can display the values of variables and data structures — useful features for debugging code.

Many of these features smoothed my path through the workflow (see go.nature.com/2antzqz). If you type a few characters of a command or variable name, a menu of possible matches will pop up. Other pop-up tips reveal a command's syntax and function. Graphic plots appear alongside the code window and can be easily shared online. Thanks to the workflow guidance and these features, it took just two hours — and less than the cost of a latte — to convert raw RNA-seq data into a series of summary tables and publication-quality figures, including gene-expression heat maps and principal-component analyses.

Using graphics only where necessary, such as for charting and visualization packages, keeps the barrier to entry low, so that researchers can more easily contribute their own Bioconductor tools, which are called 'packages'. Zhu grew comfortable enough with the platform in just a few months to develop her own package for quantifying changes in gene expression. The command-line focus also helps the core Bioconductor development team enforce good software-development practices that ensure all of the components work as documented and play nicely together.

The design choice can limit user-friendliness, however. Zhu says she found it challenging to understand how Bioconductor bundles together disparate kinds of data — variables for patient gender and age, biopsy location, gene position and expression levels, for instance — into 'objects' that can be analysed or manipulated as units. "For beginners who are used to the basic data types in R (matrix, vector, et cetera)," Love says, "there is a little bit of [a] learning curve to remember all the ways of manipulating or accessing these rich data objects."

I experienced that confusion myself with Love's workflow. Nearly all of the sample code worked as expected, although a few commands did throw errors noting that a necessary package wasn't installed — easily solved by clicking on the 'install package' button in RStudio and typing in its name. But to depart from the workflow script and attempt a new kind of analysis on my own would take considerably more practice.

Zhu says she read and reread vignettes and user guides until she understood each function's behaviour and memorized default settings and data types. "Once an analyst reaches the stage that they know what the next steps are without having to look them up in the documentation," Love adds, "she can perform very complex manipulations of data, and generate more or less publication-ready results and figures very quickly using Bioconductor."

That simplicity is what draws many to the project. Love, who wrote the DESeq2 package underlying the workflow I used, says he was hooked right away. "I was fascinated that you could discover sub-phenotypes of disease just by looking at the measurements from the microarrays," he recalls. "I ended up contributing my first package to Bioconductor in 2011, and since then, getting more deeply involved in contributing packages and workflows, going to the meetings, involvement on the support site and so on."

So, give it a try. With Bioconductor in the cloud, it's never been easier to dive into bioinformatics. ■

W. Wayt Gibbs *is a science writer based in Seattle, Washington.*

1. Salih, M. A. M. *et al. Parasite Immunol.* http://dx.doi.org/10.1111/pim.12431 (2017).
2. Gago, J. *et al. Plant Methods* **13,** 90 (2017).
3. Love, M. I., Anders, S., Kim, V. & Huber, W. *F1000Research* **4,** 1070 (2016).
4. Huber, W. *et al. Nature Methods* **12,** 115–121 (2015).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

CORRECTION

The Toolbox ‘A bioinformatics workshop in a box’ (*Nature* **552,** 137–138; 2017) erroneously affiliated Robert Gentleman with the Cambridge campus of Harvard University. He was, in fact, at the Harvard T. H. Chan School of Public Health in Boston, Massachusetts.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.