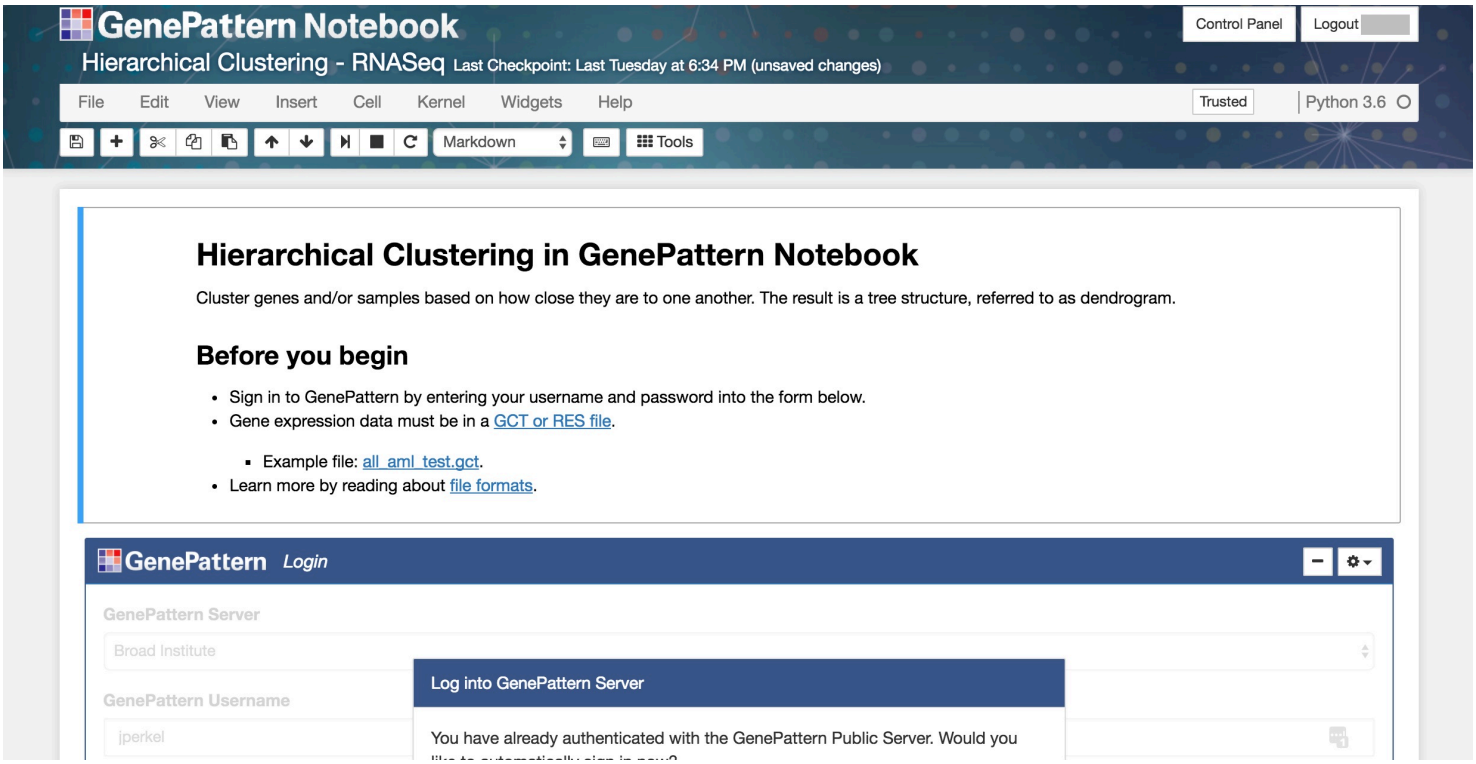
NATUREJOBS | NATUREJOBS BLOG

TechBlog: Jupyter powers bioinformatics, again

15 Sep 2017 | 12:00 BST | Posted by Jeffrey Perkel | Category: Blog, Technology



Bioinformatics isn't easy for newbies. It's typically done on the Linux command line, where users direct the computer using text-based instructions rather than clicking a mouse.

But there are alternatives. One popular choice is Galaxy; another is GenePattern. Both allow researchers to execute complex bioinformatics tools via open-source, point-and-click, web-based interfaces, freeing them from the burdens of the command line, programming, and software installation. As such, they make bioinformatics workflows relatively user-friendly. And that trend is continuing.

Several weeks ago, I wrote about an extension to the Galaxy environment that allows researchers to launch an in-environment programming notebook called Jupyter, access and document their data using Python or R, and push the

new results back into Galaxy, all without ever leaving the Galaxy environment.

Now GenePattern users have similar functionality. In the 23 August issue of *Cell Systems*, Michael Reich of the University of California, San Diego, and colleagues describe the GenePattern Notebook, with which researchers can build, execute, and annotate GenePattern workflows using the Jupyter notebook in place of the standard GenePattern interface.

James Taylor, a computational biologist at Johns Hopkins University in Baltimore, Maryland, whose lab co-leads Galaxy development, says this strategy represents the "opposite approach" from the one his team took. Whereas GenePattern Notebook incorporates GenePattern tools inside Jupyter, Taylor's team opted to integrate Jupyter within Galaxy.

"So, two different approaches, both have good features, which one someone wants to use is probably dependent on their usage style and the tools that are available in each environment [*sic*]," he writes in an email.

According to Reich, who led development of both GenePattern and GenePattern Notebook, notebook systems such as Jupyter are increasingly popular in genomics research thanks to their ability to effectively document, reproduce, and share computational workflows. "It is an extremely effective medium for communicating science, because it allows you to interleave text, graphics, multimedia, with the actual code that's being executed." But, Jupyter is designed primarily for programmers — users need to know how to code.

Similarly, GenePattern hides complex tools behind a simple interface, but offers users limited opportunity to document what they're doing. That makes it difficult for researchers to return to a workflow months or years later and understand what they did. Ditto for shared notebooks: researchers may have a hard time following their colleagues' work.

"Each of those two different types of software, the notebook metaphor and the bioinformatics tool aggregation portal, is very effective at supporting reproducible science in its own way," Reich says. "What we wanted to do was to combine those two, to derive the benefits from both of the them."

To use the GenePattern Notebook, users create a free account at genepattern-notebook.org. That provides access to any public notebook — the web site includes a repository of 13 notebooks at present — as well as the public GenePattern server at the Broad Institute, on which the jobs run. Alternatively, researchers can install a local copy of the notebook software on their own hardware using Docker.

The system is implemented as an extension of Jupyter itself. In standard Jupyter, tasks are performed by keying code blocks into interactive cells. In GenePattern Notebook, cells can contain code or informatics tools; in the latter case, the cells are point-and-click dialog boxes in which users configure and execute the specified software. As with standard Jupyter, users can document what they're doing to create a detailed record of their steps, and even access and manipulate the data outputs within the document using Python.

The result, Reich explains, is "a complete research narrative, where essentially the entire contents of a paper plus the lab analyses are all available within one document. And that makes it very easy for people to understand the science that was done, as well as to adapt it for their own use." Those narratives can be stored and shared, allowing users to replicate or adapt other researchers' work.

GenePattern Notebook users can access any of the several hundred bioinformatics tools already available for GenePattern itself; new tools can be incorporated via a simple web form, Reich says (though only in local installations of the notebook system).

The GenePattern system has some 50,000 registered users, and hundreds have already registered to use the notebook

software as well, Reich says. Future versions will endow those users with tools for collaborative notebook development, hierarchical notebook structures supporting the concept of “chapters”, as well as public commentary — a feature that supports “a more public form of peer review,” he says.

A video tutorial on the GenePattern Notebook is available here.

Jeffrey Perkel *is Technology Editor,* Nature

Image: screenshot/Jeffrey Perkel

Suggested posts

Mike Goodstadt: A circuitous route to bioinformatics

Jupyter joins the Galaxy

Smartphone science, no programming required

Previous post
Naturejobs podcast: How to start and run a lab

Next post
Do it for science – not for tenure

Comments

There are currently no comments.

You need to log in or register to comment.