Genomics in the clouds

Vivien Marx

Cloud computing can help busy genomics labs. But researchers will want to be cautious shoppers as they scan the skies for the cloud best suited to their needs.

When high-throughput genomics labs want to analyze and share data and software tools, cloud computing is an option. The cloud can be used transiently to add computing power to a project on variant calling, for example, or power an entire Web-based suite of genomics analysis tools for the longer term.

Some scientists might want to engineer software for the cloud themselves. Others may seek to efficiently use cloud-based software designed by others. For these groups and all those in between, researchers who have had their heads in the cloud for a while offer some advisable approaches and discuss habits to avoid.

What the cloud is and is not

Her group members are both enthusiastic and realistic about the cloud, says Carole Goble, a computer scientist at the University of Manchester. Goble has developed online collaborative platforms for the life sciences, such as the workflow collection myExperiment, which includes software pipelines that run on the cloud. The cloud "provides capacity to do all sorts of testing and benchmarking, but it requires expertise to use it," she says.

R. Whitrow

Carole Goble says she and her group are cloud enthusiasts and cloud realists.

Clouds are airy, fluffy collections of water droplets in the sky, but they are also informal names given to certain types of computing centers. Although loaded with heavy, hardwired hardware, the centers are not considered 'weighty' lab equipment because they are used remotely. Scientists need not maintain this hardware, pay the electric bill or answer the phone at 3 a.m. when something has crashed. Akin to renting a car, with cloud computing, researchers pay for only what they use. This car allows drivers to expand or contract the trunk and swap out motors of varying horsepower during the rental period.

Clouds are big business and are deployed around the world in the service of research institutions, government agencies and companies. In recent months, ten companies have landed cloud computing contracts with the US Department of the Interior, including IBM, AT&T, CGI, Lockheed Martin, Unisys and Verizon. Amazon Web Services (AWS) has a cloud designated solely for its work with the US government and recently garnered a cloud computing contract with the Central Intelligence Agency. Numerous small providers offer their services too, all of which leads to a sky of clouds with features and pricing options that are not easily compared. Prospective users can consider some basic recommendations about when to use the cloud and how to engineer for it (**Table 1**).

Some websites, such as CloudSleuth, compare cloud-based services. Even if such sites are not geared toward the life sciences, they can offer helpful clues. Cloud shoppers will want to consider factors such as monthly or annual fees, pricing per computing hour, cost of data transfer to and from the cloud, types of operating systems, number of processors and amounts of random-access memory (RAM), levels of support, ease of scale-up, amount of storage space and security levels.

For help deciding on the right cloud, researchers might poll their colleagues to see what they use. In genomics, the answer is often Amazon, says Anushka Brownley, a bioinformatician with the consulting company BioTeam. "It will take a while for other clouds to catch up."

iStockphoto

Cloud computing is not heavy genomics lab equipment.

As with many cloud providers, Amazon offers cloud computing resources with different types of configurations. To address the needs of life scientists, the company has added data sets on its cloud infrastructure, as have some other providers. This provision saves scientists from having to organize and pay for the transfer of common data such as reference genomes. To perform analysis, scientists transfer their own data and tools to the Amazon cloud and draw on these public data sets—such as the Ensembl annotated human genome; data from The US National Institute of Health's sequence database, GenBank; or data from the model organism Encyclopedia of DNA Elements (modENCODE).

But Amazon has limitations too, such as in the availability of large amounts of RAM. "For example, doing *de novo* assembly on large genomes requires a significant amount of RAM, which AWS currently does not offer," Brownley says.

When using the cloud, scientists pay for computing time and data transfer. Yet at the start of an analysis involving 50 whole genomes from 20 collaborators in seven countries and using five software tools, it

Table 1 | When to use the cloud

Application	Advantages of the cloud	Cautions
Data sharing	- Single copy of the data decreases management overhead - Broadly accessible	- Requires adequate network bandwidth to transfer data - Requires appropriate security protocols to manage data
Large-scale computing	- Not constrained by local infrastructure - Great for highly parallel jobs, such as jobs split by chromosome	- Requires clear understanding of cost - Cloud hardware is not suited for all types of analyses - May require reengineered code to enable parallelized computing
Data archive	- Cheap storage - Little IT overhead	- Requires smart backup policies - Costly to pull data
Bursting (for when local resources are saturated)	- Great for variable use - No need to buy hardware	- Tasks that run on local computing resources do not always run exactly the same in the cloud
Rapid testing	- Offers full control over the development environment - Allows quick launch or close of computing environments	- Easy to underestimate costs - Scientists must properly save what they need to keep

Source: A. Brownley, BioTeam

is hard to estimate how much computing power or data transfer will need to happen. Sticker shock can be the result both for scientists who use the cloud to build computing infrastructure and for those who use tools on the cloud, says Ravi Madduri, co-lead of the team behind the cloud-based genome analysis platform Globus Genomics at the Computation Institute (CI), which is run jointly by the University of Chicago and Argonne National Laboratory. For some scientists, awareness of these issues is enough to set them on the route to engineer their path to the cloud themselves.

Engineering for the cloud

Scientists can use the cloud to broaden the accessibility of a freshly minted software tool. Researchers can also leverage the cloud to help them scale up an analysis: they can throw a large amount of computing power at a problem or allow many tool users at once to crank through their data.

Researchers might explore several cloud options before settling on one. Rob Knight of the University of Colorado says he and his team set up metagenomic data analysis tools on Google's App Engine and Microsoft's Azure before settling on Amazon. "Azure and Google App Engine were not useful for the workflows we needed to do the last time we benchmarked, which admittedly was a while ago," he says.

Getting a genomics analysis tool to run on the cloud takes software engineering skills. Knight and his group have those skills, which he says his students pick up quickly, although he admits it takes some effort. "To go from 'It's good enough for me and my research' to 'It's good enough in general' is very time consuming and a big step change," Goble says.

As her lab members Alan Williams and Robert Haines explain, there are many different ways of mounting tools on the cloud. For example, a toolmaker might want to make visible the full application programming interface of a given software tool. That approach gives other Web-based software developers an opportunity to interact with the software.

And then, Goble says, researchers will need to reproduce a tool's environment on the cloud. Without the additional libraries,

Whenever deploying software, developers need to remember documentation.

code, services and data sets, the software can crash when run. Another facet to consider is how users will access the software: will there be a user interface, and how will the access architecture be prepared? It could be by representational state transfer, which is a way to present software through servers and clients that can be queried.

A useful and usable tool needs documentation, tests, examples, a release process and mechanisms for community interaction, says Reece Hart, a computational biologist at genetic testing company InVitae. The company uses secure cloud computing for its research tools and clinical variant interpretation pipeline. A resource also needs support service, backups, monitoring, upgrades and ways to prevent or handle abuse, he says.

Although software code written for the cloud in the life sciences is often identical to the version written for laptops, desktops and traditional compute clusters, an 'instance' remains "a façade or emulation of a computer," he says. What is unique about developing for the cloud is that performance of an application can vary dramatically across environments that have different processing, memory and storage speed profiles.

I/O hog

Much genomics analysis software is input-output (I/O) intensive, which means the tools involve the input, processing and output of large amounts of data. Bottlenecks are common. "In our case, most of the I/O is attributed to filtering and annotating variants," says Hart. Although these bottlenecks can be mitigated by plenty of approaches, they become more apparent in cloud computing than on owned hardware. Computing and storage options vary widely among cloud providers and even within the offerings of one cloud provider. Therefore, software that reads and writes lots of data to local storage on owned hardware may perform very differently on a cloud-based instance with cloud-based storage, says Hart. Also, on a given cloud, any number of the virtual machines might share the same physical hardware. Performance dips on the cloud can stem from wrestling for resources such as network bandwidth.

Virtual machines make a software tool portable and help with distribution. Users might still wonder how to use the tool on their data. And if source code is not exposed, users might wonder how it accomplishes its tasks. According to a tweet by @ianholmes from the amusing but sobering Twitter exchange #overlyhonestmethods: "You can download our code from the URL supplied. Good luck downloading the only postdoc who can get it to run."

Computation Institute

Maintaining an IT analysis environment can be a four-person job, says Paul Davé.

The cloud can improve the accessibility of a tool for other researchers. But to fully understand, extend or reimplement the software requires knowing the actual code, says Goble. Developers should fight the temptation to host a black box behind a Web interface: they should make all algorithms and source code available, she says. They can do so through repositories such as GitHub or make the source code visible on the cloud.

Scale is another aspect to keep in mind. If many people want to use the tool, or if some want to use it with large data sets, the developer will need to engineer the software to expand across many virtual machines as needed. A number of websites help scientists explore and expand their computational skills: for example, Software Carpentry.

And then there is the wallet to remember. A lab needs the resources to upload, maintain, update and optimize software to run on a cloud on many samples or for many users at once. "That starts to feel like a four- or five-person team just to maintain your analysis environment," says Paul Davé, a computer scientist at the CI who co-leads the Globus Genomics team. "Who will pay for hosting it?" Goble asks. "If you want to make the tool or service available for direct use, then it has to be running all the time. Who pays?"

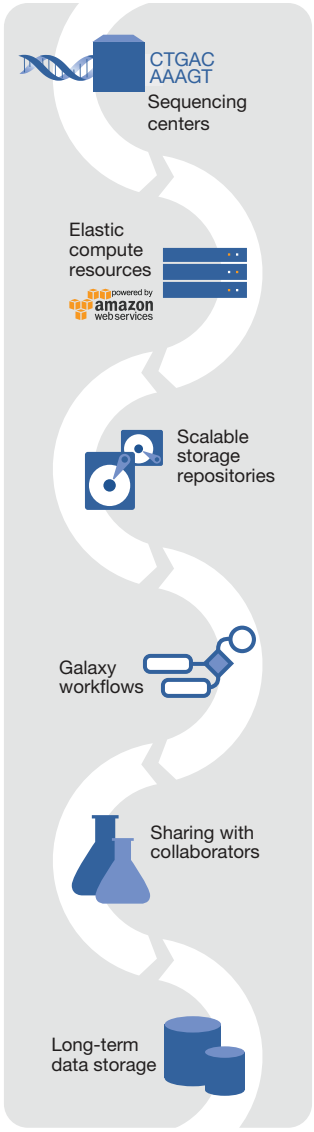
Private clouds and public clouds

To some scientists, a cloud's security and privacy is the aspect that matters most: a need to which a number of services cater with different kinds of private or secured clouds. For example, BT Cloud Compute, a large provider, has cloud offerings geared to the regulated environments of the biotech and pharma industries. DNAnexus, too, which offers cloud-based genomics analysis, has developed its cloud in line with the regulations governing these environments.

InVitae's Hart says that his company heeds the regulations protecting patient data as it runs genome analysis on deidentified sequence data to interpret genomic variants. With patient consent, the company plans to

deposit clinically observed variants into public databases such as ClinVar. The company's security includes isolation of computers with patient data, encryption, secure transmission, extensive monitoring and logging, and encrypted backups to secure locations, he says.

Private clouds exist in the nonprofit world, too. This spring, the University of Chicago launched the Bionimbus Protected Data Cloud, where authorized researchers can work with data from The Cancer Genome Atlas. The European Bioinformatics Institute (EBI) is piloting Embassy Cloud, a private cloud-computing environment in which scientists can analyze their data and compare them with EBI data sets. The data center is physically located at CSC–IT Center for Science in Finland, a government-run high-performance computing center.



V. Cuplinkas, Computation Institute/ Univ. Chicago

The cloud-based genome analysis platform Globus Genomics wants to save researchers time.

Engineered for users

The data wealth in the life sciences and an openness to cloud computing draws the interest of companies and organizations outside of the life sciences. These groups build cloud-based platforms and engineer software to help users who would otherwise have to do the engineering themselves. A data analysis platform from the company Appistry helps FedEx route packages. Appistry expanded into genomics data analysis for university labs with a view to the future. Solutions that work for a few genomes do not scale when working on scores or hundreds of genomes, says Trevor Heritage, who directs the company's business development and product strategy.

In partnership with the Broad Institute, his team has begun distributing the Genome Analysis Toolkit (GATK) for commercial applications. These tools are used by pharmaceutical and biotech companies as well as academic labs, says Heritage. The toolbox can currently be downloaded from the Broad Institute website for installation on local hardware or on a cloud.

If a lab wants to scale up how it uses GATK, researchers can try Appistry's proprietary computing environment, Ayrris. Customers access the platform on a subscription or on a fee-per-sample basis. Whenever the Broad issues updated GATK versions, Appistry packages them. The company has readied genome analysis tools for high-throughput use, says Heritage. Beyond GATK, the platform accommodates around 70 open-source and commercial software tools and lets scientists use an open-source pipeline builder from the collaborative platform Galaxy.

Researchers can launch the company's software environment on a local compute cluster, on a cloud, on Appistry's private cloud or on a combination of these.

"They use their internal cluster up to the point that they're comfortable with, and thereafter any of the other processing basically—I think the phrase is—'cloudbursts' out to us." To burst out, scientists do not need to shift software parameters. The data are transferred securely, and computing scales according to the data volume.

Although Appistry's developers considered using the Amazon cloud environment, they decided the company could more easily handle security at its own data center. The Appistry private cloud is compliant with the regulations concerning handling of patient data.

Many cloud providers charge for cloud-to-cloud data transfer if data move from one geographic region to another, a common occurrence in data sharing. As part of the subscription fee, the company lets researchers define the geography of their collaborative effort. The territory can span several continents.

Heritage sees customers struggle to put all of the "pieces of the genomics puzzle together," such as the tools, compute architectures and data storage. "People have made a choice of their compute architecture, and then they find out, 'Now I can't run this specific tool because it's not supported on that compute architecture,'" he says. "In almost every organization people have hit a stumbling block of that type."

Appistry allows scientists to upload data directly from the sequencer into the cloud environment instead of accumulating them and transferring them later. "Let's just upload it as the data comes off, in meaningful chunks," Heritage says.

A Galaxy of clouds

High-throughput genomics labs with data streaming off the sequencer can also use other types of clouds. These clouds may involve, for example, academic initiatives that have engineered a commercial cloud—notably, the Amazon cloud—into their platforms.

The "Galaxy ecosystem is growing," says James Taylor, a computational biologist at Emory University. He is a co-developer of Galaxy, the open-source collaborative genomics analysis platform developed at Emory and Penn State University. Galaxy software can be downloaded to use locally and also can be used on many types of clouds.

CloudMan and CloudLaunch are software applications that let users create their own Galaxy virtual machines on the Amazon cloud without an intermediary[1]. "There are no additional costs beyond what they pay directly to the cloud provider," Taylor says.

Lincoln Stein and his group at the Ontario Institute for Cancer Research have set up virtual computing environments with Galaxy and all modENCODE data on both the Amazon cloud and Bionimbus[2].

Galaxy has also been built into the cloud-based Globus Genomics platform, which is intended to help big-data scientists in all disciplines. The biggest data crunchers are traditionally cosmologists, says CI computer scientist Madduri, but "biology has recently become big-data science."


Emory Univ.

The "Galaxy ecosystem is growing," says James Taylor.

Ian Foster, the CI's director, initiated grid computing in the late 1990s to distribute computing across many computing centers. Now cloud computing expands the grid. CI scientists can expand their efforts from writing scheduling algorithms for a set grid of computing resources to working in a world in which computing capacity can be scaled as needed, Madduri says.

The team launched the cloud-based Globus Genomics platform for data analysis in April on the Amazon cloud. Users can choose from a workbench of Galaxy tools, add their own and run an analysis pipeline. The team has also added reference genomes for users.

Academic groups have begun using the platform. "It's for fee but without a profit motive," says Davé. "It's for sustaining the effort and the development we have put in."

The team tested the platform with the lab of William Dobyns, at Seattle Children's Hospital, who studies the genetics of brain malformation. His patient samples were sequenced at the Broad Institute and by a service offered through PerkinElmer. Previously, the lab received raw data on hard drives and then struggled with storage and analysis. "To run one of their exome analysis pipelines might take them 20 or 30 hours," Davé says. Running the analyses one after the next could mean many months of computational time.

In their test, the Globus Genomics team first streamlined data movement so that the data ran securely from sequencing facilities to the analysis pipeline on the cloud as they were generated. This way, says Davé, "they wouldn't queue up raw data from 20 patients and then put it all on one hard disk and send it out."

Data transfer is handled by a protocol called GridFTP-based Globus Online, an open-source standard for high performance data transfer that is also used by research teams working on Large Hadron Collider, says Madduri. Getting data to and from the cloud means pushing for speed without killing the machine or losing data in transit, he says. "We do check-sums on the files, and we do end-to-end encryption that is based on public-key cryptography to make sure that the data is encrypted and nobody can snoop the data," Madduri says.

The Globus Genomics team members have engineered analysis to scale to the needs of a given pipeline and sample size. And they have optimized genomics analysis tools for their cloud, too. "We take a tool and see: how many processors is it using?" says Madduri. They also look at how much memory tools require and how to match the software settings to those of Amazon's machines. This profiling helps with analyzing multiple data sets in parallel on multiple cloud-based processors, which saves researchers time. "It used to take them 20 hours to get through one exome; they can run 20 exomes in the same amount of time," says Davé, referring to the test with the Dobyns lab.

Davé says that Amazon updates its hardware and uses cost-effective, high-performance machines "because that's their business to do so," which is to researchers' advantage. In recent months, the team has been testing Amazon's new, fast compute instance.

For now, Globus Genomics is not set up for the analysis of personally identifiable data. But the CI group is currently building a secure environment on Amazon. Sharing data with other collaborators has also been engineered. "It's a trivial drag and drop," says Davé. The idea behind Globus Genomics is to reduce the amount of time researchers spend on building infrastructure, which frees them to spend more time on developing algorithms and modes of analysis, the team says.

Cloud computing is increasingly making inroads in genomics, says BioTeam's Brownley. For users of cloud-based software, analyses can happen even in labs without large computing resources. Bypassing the need to set up a physical infrastructure saves time for both tool developers and users. But qualified technical staff are still needed to support any development activities on the cloud. "Cloud adoption will become more prevalent as better solutions are available on it," she says.

1. Afgan, E. *et al. BMC Bioinformatics* **11** (suppl. 12), S4 (2010).
2. Trinh, Q.M. *et al. BMC Genomics* **14**, 494 (2013).

Vivien Marx is technology editor for *Nature* and *Nature Methods* (v.marx@us.nature.com).