Harnessing cloud computing with Galaxy Cloud

To the Editor:

Continuing evolution of DNA sequencing has transformed modern biology. Lower sequencing costs coupled with novel sequencing-based assays has led to rapid adoption of next-generation sequencing across diverse areas of life sciences research[1–4]. Sequencing has moved out of the genome centers into core facilities and individual laboratories where any investigator can access them for modest and progressively declining cost. Although easy to generate in tremendous quantities, sequence data are still difficult to manage and analyze. Sophisticated informatics techniques and supporting infrastructure are needed to make sense of even conceptually simple sequencing experiments, let alone the more complex analysis techniques being developed. The most pressing challenge facing the sequencing community today is providing the informatics infrastructure and accessible analysis methods needed to make it possible for all investigators to realize the power of high-throughput sequencing to advance their research.

A possible solution to this infrastructure challenge comes in the form of cloud computing, a model where computation and storage exist as virtual resources, accessed by means of the internet, which can be dynamically allocated and released as needed[5]. Where previously acquisition of large amounts of computing power required large initial and ongoing costs, the cloud model radically alters this by allowing computing resources and services to be acquired and paid for on demand. Importantly, cloud resources can provide storage and computation at far lower cost than dedicated resources for certain use cases. For several specific applications, effective use of cloud resources has already been demonstrated[6–8]. In general, however, cloud resources are not provided in a form that can be immediately used by a researcher without informatics expertise. Several commercial vendors provide cloud-based sequence analysis services through the web that hide all complexity of the underlying infrastructure. Yet these contain limited sets of analysis tools, and because they are proprietary solutions users must give up some control over their own data and risk vendor lock-in.[AU; reword or briefly explain "vendor lock-in".]

All 'battle-tested' next-generation sequencing analysis practices (e.g., analysis of human variation exemplified by the 1000 Genome Consortium publication[9] [AU: REF OK?]) are open source.

One popular open-source platform that has made substantial progress toward making complex analysis available to researchers is Galaxy[10,11]. Galaxy enables users to perform analysis using nothing more than a web browser. The environment automatically and transparently tracks every detail of the analysis, allows the construction of complex workflows and permits the results to be documented, shared and published with complete provenance, guaranteeing transparency and reproducibility. Importantly, Galaxy is an extensible platform; nearly any software tool can easily be integrated into Galaxy, and there is an active community of developers ensuring the latest tools are wrapped and made available through the Galaxy Tool Shed (http://usegalaxy.org/community). Galaxy is provided as a free public service with which thousands of users perform hundreds of thousands of analyses each

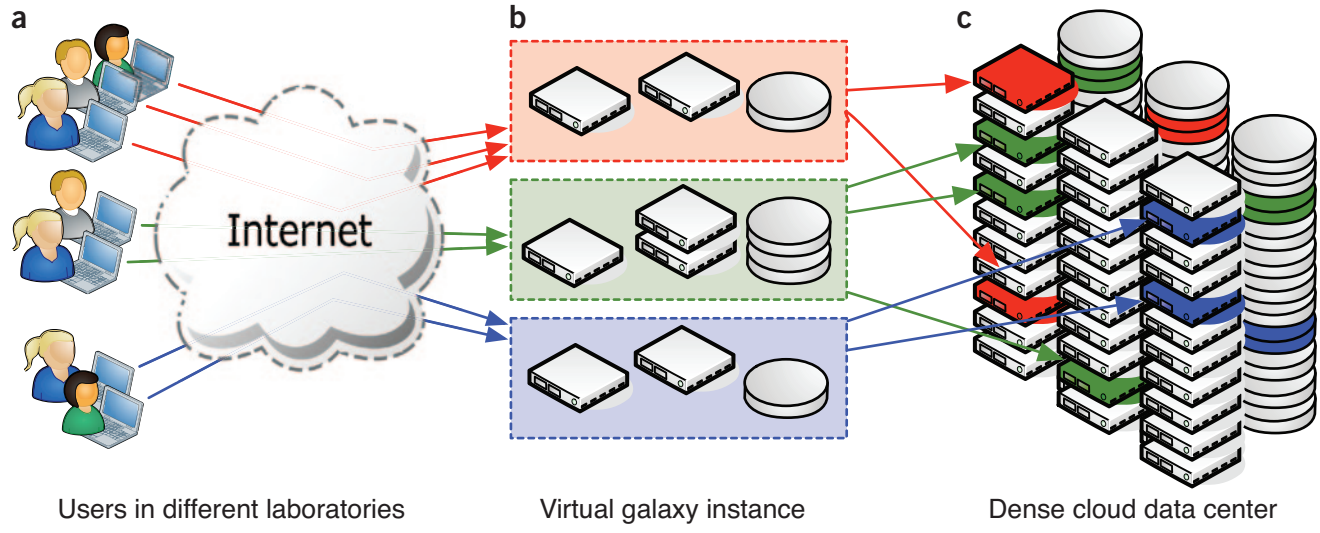

Figure 1 Overview of Galaxy instances running on cloud resources. (**a**) Users in different laboratories access a dedicated Galaxy instance over the internet with nothing more than a web browser. (**b**) These Galaxy instances appear to the users to be dedicated infrastructure with apparently infinite computing and storage resources. (**c**) In fact, they are virtual resources that Galaxy's autoscaling acquires and releases on demand in response to changing workloads.

month. However, this free public resource cannot meet increasing demand without implementing limits on data transfer and computer usage, resulting in delays that users may find unacceptable. Fortunately, the Galaxy platform is easily deployed on local resources, and many groups working with large-scale sequence data now run their own Galaxy instances. **[AU: Please explain briefly or use synonym for 'instance' for the general reader who is not a computer scientist]** However, this still requires local computing resources **[AU: Although 'compute resources' can be found on the internet, 'computing resources' is also used (see, for instance, NCSA http://www.ncsa.illinois.edu/AboutUs/) and is preferred usage for our nonspecialized audience]**and informatics knowledge.

To bring the virtually unlimited resources of cloud computing into the hands of biomedical researchers, we have developed Galaxy Cloud. It allows anyone to run a private Galaxy installation on the Cloud exactly replicating functionality of the main site (http://usegalaxy.org/) but without the need to share computing resources with other users (**Fig. 1**). With Galaxy Cloud, unlike software service solutions, the user can customize their deployment as well as retain complete control over their instance and associated data; the analysis can also be moved to other cloud providers or local resources, avoiding concerns about vendor lock-in.

Currently, we provide a public Galaxy Cloud deployment on the popular Amazon Web Services (AWS) cloud; however, it is compatible with Eucalyptus and other clouds. If starting for the first time, the instance is configured by the user (e.g., by specifying the amount of initial storage allocated; exact step-by-step instructions are provided at http://usegalaxy.org/cloud). Once configured, users can then access their Galaxy, which will function exactly like the Galaxy public site. Every analysis tool that is available through the public Galaxy instance is installed and available for immediate use, as well as all the necessary supporting data (e.g., genome sequences, alignments or indexes). In addition, several tools that are too computationally intensive to provide on the public Galaxy are also included. This ready-to-use environment is combined with the ability to allocate practically unlimited computing power on demand thanks to use of cloud computing. When the user has finished analysis and the instance is no longer needed, all computing resources can be released, whereas the user's data and instance state are preserved to be used later.

Galaxy Cloud's deployment is achieved by coupling the Galaxy framework to CloudMan[12], which automates management of the underlying infrastructure cloud resources (**Supplementary Notes** and **Supplementary Fig. 1**). CloudMan handles all aspects of infrastructure management, including resource acquisition, configuration and data persistence, necessary to support the Galaxy application. In the above scenario, CloudMan has allocated dedicated storage for the user's own data, initialized the Galaxy database, as well as composed additional data volumes containing the tools and secondary data they require. As with any instance of Galaxy, additional tools and data can easily be added by the owner of the instance and shared with others.

We consider, as a case study into the use of Galaxy Cloud, the problem of identifying heteroplasmic sites—variation among the multiple copies of the mitochondrial genome (mtDNA) within a cell or individual. Mutations in mtDNA have been implicated in hundreds of diseases and, in many cases, the disease-causing variants can be heteroplasmic, with manifestation dependent on the relative proportion of variants[13,14]. Furthermore, this task emphasizes many of the key motivations for Galaxy Cloud. First, it involves the use of clinical samples, which often involve strict privacy concerns and should not be analyzed on a public site, but can be analyzed on a secure public or private cloud resource. Second, it is both a data-intensive problem and one with computing needs that vary over the course of the analysis. Third, it requires different methodology than the related problem of single nucleotide polymorphism calling in diploid genomes, showing the power of Galaxy's workflow system to construct solutions to nonstandard tasks. Fourth, there is currently no commonly accepted approach, which has led to questions about the validity of published heteroplasmic sites, emphasizing the need for a system that makes analysis completely transparent and reproducible.

Using mtDNA sequence data from nine individuals across three families[15], we developed Galaxy workflows to perform the identification of heteroplasmic sites. These workflows map the sequencing reads, separate them by strand, transform data sets from read-centric to genome-centric form and perform several filtering and thresholding steps before merging the branches and generating a list of sites that contain allelic variants above a certain frequency supported by high-quality reads on both strands. Running the workflows identified four heteroplasmic sites in two of the three examined families.

This analysis was computed entirely using Galaxy Cloud on AWS, and can be replicated exactly by importing the data sets and workflow available at http://s3.amazonaws.com/heteroplasmy/index.html. For a complete description and explanation of the acquired data as well as how to use, import and modify workflows used for the described heteroplasmy study see the Galaxy Page[10] at http://usegalaxy.org/heteroplasmy/.

To perform the analysis, we uploaded 45 GB of sequence data sets to S3, which took 9 h at an average transfer rate of 1.5 MB/s and cost $5. During the execution of the analysis workflow, the cluster size was managed by CloudMan's autoscaling feature and the cluster size varied between 1 and 16 nodes. It took ~6 h and cost $20 to complete the workflow. With autoscaling disabled, for fixed cluster sizes of 5 and 20 nodes, the run time was 9 h at a cost of $20 and 6 h at a cost of $50, respectively. By adapting the computing resource as the workflows demands change, autoscaling is able to provide both the shortest total run time and lowest cost. Once the workflow is executed, the obtained results can be further analyzed directly on the cloud, downloaded or left on the cloud for future reference. Overall, a complete analysis using a computing cluster and a variety of open-source, next-generation sequencing tools was performed within 15 h for a cost of $25 using nothing but a web browser.

Cloud computing resources may not be as cost effective for all usage scenarios. The heteroplasmy workflow was already developed, which made it straightforward to execute in its entirety. The interactive analysis and trial and error involved in building and refining the workflows is less cost effective, though autoscaling helps avoid excessive waste. This particular workflow has steps that could be executed in parallel, which allowed it to take advantage of cloud elasticity. Cloud instances of Galaxy will be limited by the resources available from a given cloud provider. For example, the largest memory instances currently provided by AWS are not sufficient to run certain *de novo* assemblers. However, these are limitations of the provider used, not the cloud model. An advantage of the virtualization-based cloud model is the ability to move to a different cloud provider or to local resources. Cloud computing offers a new avenue for accessing computational infrastructure and Galaxy Cloud helps harness the potential in a very general way,

but may not be appropriate or cost effective for some workloads.

As next-generation sequencing becomes an indispensible tool for biomedical research, it is crucial to provide analysis solutions that are usable and cost effective for biomedical researchers. Galaxy Cloud addresses this by combining the accessible Galaxy interface with automated management of cloud computing resources. Unlike purpose-built solutions, Galaxy allows users either to use existing tested best practices in the form of workflows or to construct their own analyses for novel tasks. Galaxy Cloud instances are owned and controlled entirely by the user who created them and can be used effectively in secure private clouds. Thus, Galaxy Cloud provides a solution that retains user control and privacy, makes complex analysis accessible and enables the use of practically limitless on-demand computing resources.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

The authors are grateful to J. Beiler for coordinating sample collection, to clinical nurses from Penn State College of Medicine's Pediatric Clinical Research Office for collecting the samples and to volunteers for donating the samples. Efforts of the Galaxy Team (E.A., D.B., D. Blankenberg, N.C., J. Goecks, G. Von Kuster, R. Lazarus, K. Li & K. Vincent) were instrumental for making this work happen. This work was funded by US National Institutes of Health (NIH) grants HG005133 and HG005542 to J.T. and A.N., US National Science Foundation grant DBI 0543285 and NIH grant HG004909 to A.N. and J.T., and NIH grant GM07226405S2 to K.D.M. Additional funding is provided, in part, under a grant with the Pennsylvania Department of Health using Tobacco Settlement Funds. The Department specifically disclaims responsibility for any analyses, interpretations or conclusions.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Enis Afgan[1], Dannon Baker[1], Nate Coraor[2], Hiroki Goto[2], Ian M Paul[3], Kateryna D Makova[2], Anton Nekrutenko[2] & James Taylor[1]

[1]Departments of Biology and Mathematics & Computer Science, Emory University, Atlanta, Georgia, USA. [2]Center for Comparative Genomics and Bioinformatics, Penn State University, University Park, Pennsylvania, USA. [3]Department of Pediatrics, Penn State College of Medicine, Hershey, Pennsylvania, USA.

e-mail: james.taylor@emory.edu or anton@bx.psu.edu or kmakova@bx.psu.edu

1. Schatz, M.C., Langmead, B. & Salzberg, S.L. *Nat. Biotechnol.* **28**, 691–693 (2010).
2. Lieberman-Aiden, E. *et al. Science* **326**, 289–293 (2009).
3. Pepke, S., Wold, B. & Mortazavi, A. *Nat. Methods* **6**, S22–S32 (2009).
4. Wang, Z., Gerstein, M. & Snyder, M. *Nat. Rev. Genet.* **10**, 57–63 (2009).
5. Stein, L. *Genome Biol.* **11**, 207 (2010).
6. Langmead, B., Hansen, K.D. & Leek, J.T. *Genome Biol.* **11**, R83 (2010).
7. Langmead, B., Schatz, M.C., Lin, J., Pop, M. & Salzberg, S.L. *Genome Biol.* **10**, R134 (2009).
8. Schatz, M.C. *Bioinformatics* **25**, 1363–1369 (2009).
9. The 1000 Genomes Project Consortium. *Nature* **467**, 1061–1073 (2010).
10. Goecks, J., Nekrutenko, A. & Taylor, J. *Genome Biol.* **11**, R86 (2010).
11. Afgan, E. *et al.* in *Guide to e-Science: Next Generation Scientific Research and Discovery* (ed. Yang, K.) 35 (Springer, New York, in the press).
12. Afgan, E. *et al. BMC Bioinformatics* **11** Suppl 12, S4 (2010).
13. DiMauro, S. *Mitochondrion* **4**, 799–807 (2004).
14. Taylor, R.W. & Turnbull, D.M. *Nat. Rev. Genet.* **6**, 389–402 (2005).
15. Goto, H. *et al. Genome Biol.* (in the press). **[AU: Is there a volume, page number or Advanced Online Publication doi?]**