nature genetics

# Peer review in the cloud

**The migration of cancer genomics data to cloud computing is a great encouragement for data reuse and integration by bioinformaticians and other data symbionts. Because the cloud allows rapid, transparent and reproducible research on large data sets, we are keen to consider articles and analyses submitted to the journal that provide peer referee access to their constituent cloud projects.**



Recent funding initiatives to improve cancer diagnosis and treatment have been likened to a 'moonshot' (*Nat. Biotechnol.* **34**, 119, 2016). Although we do not think that the metaphor of a single engineering feat to achieve a defined goal is entirely appropriate to the aim of controlling cancers, the cloud computing infrastructure for the upcoming Genomic Data Commons (https://gdc.nci.nih.gov/index.html) and the three recently launched cancer cloud pilots (https://cbiit.nci.nih.gov/ncip/nci-cancer-genomics-cloud-pilots) is very much equivalent to building Mission Control to coordinate multifaceted and coherent programs. Not only does a cloud commons give broad access to petabyte data sets, which are beyond the capacity of many research institutes to even download, and offer huge reductions in the compute time and infrastructure costs of research projects, but it also reduces duplication of effort in reformatting, normalizing and relabeling data sets. Code, models and processed data can be readily found and built upon, and errors can be quickly corrected for constant improvement without the lags and version control problems of the previous hub-and-spoke model of central data holding and diverse data centers. The automatic accounting for every element, use and user within the cloud environment may be one of its most important aspects, as useful data and tools can be immediately seen and given resources and credit. Similarly, work that is unproductive or incompatible with advancing cancer research can be flagged and redirected.

Cloud infrastructure has been at the top of the wish list for cancer genomics researchers of the TCGA and ICGC consortia. Close behind are two standardization efforts enabled by the cloud that should help overcome the biggest problems currently holding back genomics research everywhere. Each data set in the cloud will have a manifest file detailing its standard data elements, generated upon upload to the resource. Metadata will thereby be established upon data upload rather than requiring exhortation to add them later. This incorporation of standards follows the advice of cancer researchers' experience with metadata protocols such as MIAME for transcriptome data sets (*Nature* **523**, 149–151, 2015). Although data access to cancer genome projects has been straightforward and has led to data reuse (page 224), a multitude of consents and access committees deters many would-be users. The best solution is a single, harmonized data access protocol connecting with consents for research reuse that applies to all human data in the genomic data cloud.

We too have our wish—to enable peer review in the cloud—as we see enormous potential for cloud commons research to improve the precision, transparency and reproducibility of research publications that provide periodic key results from and updated guides to the continuous knowledge production within the data commons. The publications also provide incentive and credit within the wider scientific community, above and beyond the reputation researchers can gain for coding and data deposition within their own commons. In the interest of refining the idea of a publishable unit and using expert review judiciously, some new peer refereeing conventions, tools and cloud pilots are therefore a priority.

Unlike supplementary data summaries and disparate data resources, the research cloud encourages explicit documentation of analysis and support for alternative decision paths taken as well as for version control of programs and data for multiple users. There are unique identifiers for every version of every element of every project, so a few well-placed 'peer here' URI links relating the cloud project to ongoing external publications will allow peer review of key constructions and decision points in every project within the cloud and from outside (for example, in preparation of manuscripts for which the cloud analysis project is only one component). By providing links to key junctures in projects in preparation where the criticism of peers is likely to be effective, cloud review has the potential to yield better papers. It will be most effective where referees can carry out spot checks for assumptions or readily display the robustness of results calculated on the fly with different input assumptions. Understandably, the three pilots (https://www.systemsbiology.org/research/cancer-genomics-cloud/, http://www.cancergenomicscloud.org/, https://software.broadinstitute.org/firecloud/) have initially been concerned with migrating data and troubleshooting the systems. Next, the systems must generate novel, reproducible and conceptually useful research results within the cloud platforms. Finally, the cloud must connect to laboratories in the outside world and to a much larger user group. We invite these groups to add links and to submit projects for peer review.

Having seen the full range of genomic data reanalysis projects over a number of years and having published some good ones, we remain enthusiastic that high-quality data give rise to several generations of reproducible research in the hands of the data producers as well as integrators and reanalysts, all of whom have their place in the research ecology. ■