Published in final edited form as:
Nat Rev Genet. 2018 April ; 19(4): 208–219. doi:10.1038/nrg.2017.113.

Cloud computing as a platform for genomic data analysis and collaboration

Ben Langmead[1,2] and **Abhinav Nellore**[3,4,5]

[1]Department of Computer Science, Johns Hopkins University

[2]Center for Computational Biology, Johns Hopkins University

[3]Department of Biomedical Engineering, Oregon Health & Science University

[4]Department of Surgery, Oregon Health & Science University

[5]Computational Biology Program, Oregon Health & Science University

Abstract

DNA sequencing made huge strides in the last decade. Studies based on large sequencing datasets appear frequently, and public archives for raw sequencing data have been doubling in size every 18 months. Meanwhile, commercial and academic cloud computing have matured, leading to more providers, greater total capacity, and a larger variety of services. Here we describe how cloud computing is used for large-scale genomics collaborations and research and argue how cloud computing will likely be a basic underpinning for future large-scale genomics collaborations and for efforts to re-analyze archived data, including privacy-protected data.

Introduction

Commercial cloud computing and DNA sequencing have been twin technology success stories of the past decade. Commercial cloud services have matured rapidly, creating new data centers, lowering prices, adding services, and generating notable profits (https://www.nytimes.com/2017/04/27/technology/quarterly-earnings-cloud-computing-amazon-microsoft-alphabet.html). DNA sequencers improved dramatically, and large genomics collaborations like The Cancer Genome Atlas (TCGA), The Genotype-Tissue Expression (GTEx) Project[1], and TopMed (https://www.nhlbiwgs.org) generated ever-larger datasets. Public archives for sequencing data such as the Sequence Read Archive (SRA)[2] grew rapidly and now have a doubling time of 10–18 months.

How are these stories related? While cloud computing was not invented with science and genomics in mind --- its major users are technology companies and other businesses --- there is a growing list of cases where cloud computing helped to achieve important scientific goals [TODO: cite]. Here we review cloud computing's role in recent genomics research, focusing on two areas where the cloud is making major contributions. The first involves the vast datasets now available in sequencing data archives like the SRA. The SRA currently

Correspondence to: langmea@cs.jhu.edu, anellore@gmail.com.

contains over 12 petabases of raw data, more than double its size 10 months ago (https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=announcement) (Figure 1). These data are available to download, with the caveat that most of the human data, and a majority of the nucleotides in the SRA overall, are protected by Database of Genotype and Phenotype (dbGaP) measures for maintaining privacy of donors. In short, the SRA gives researchers ready access to valuable data, some of it quite unique – e.g. from individuals with rare phenotypes or from hard-to-obtain tissues -- and at a scale no one laboratory or institution could recreate.

But the archived data is not easy for typical researchers to use. To get new results from archived datasets, a researcher must secure sufficient storage space, perform large and time-consuming downloads, then perform a compute-intensive re-analysis of the data from scratch. Many labs aren't equipped for this, so valuable data go unused[3]. But the cloud's elasticity and the increasing availability of cloud-enabled software tools are easing the way. We will discuss how the cloud has been applied in this area and why its elasticity, reproducibility and privacy features make it an apt fit for large-scale reanalysis of public data.

The second area where cloud computing has made inroads is in enabling collaborations on large amounts of shared data. The distributed nature of the cloud has made it a natural venue for driving the collaborative computing effort, and for otherwise facilitating the collaboration. This has been happening most visibly in projects like the International Cancer Genome Consortium (ICGC)[4,5], the Cancer Genomics Cloud (CGC) Pilot[6], and the Encyclopedia of DNA Elements (ENCODE and modENCODE)[7].

The cloud model

In the cloud-computing model, computational resources like processors and hard disks are thought of as utilities to be rented from a service provider like Amazon Web Services, Microsoft Azure or Google Cloud Platform. The providers control vast pools of computers and storage, organized into data centers scattered across the world. Users request resources, use them, then release them back into the pool when the work is complete. Fees are incurred according to usage. Storage incurs a per-gigabyte-per-month fee and computers incur a per-computer-per-hour fee. Users are billed monthly, just as for a home utility.

The cloud's hallmarks are elasticity and convenience. Elasticity refers to the ability to rent and pay for the exact resources needed. The user is not compelled to downscale the task to fit the confines a local cluster ("under provisioning"), nor must the user incur the cost of "over provisioning": purchasing an amount of computing to match the largest possible future need.

Because cloud datacenters are vast, computational requests large and small can be fulfilled quickly, sometimes immediately. A user requesting a 1,000-computer cluster for 1 hour is about as likely to succeed as a user requesting a single computer for 1,000 hours. This is not as true for smaller institutional clusters. The ability to recruit vast resources is an extraordinary advantage; instead of waiting for the trickle of computer-hours available on a busy institutional cluster, the user can rent a cluster the size of your entire institutional cluster for a day. Work completes in a fraction of the time and you pay only for what you use (Figure 2).

The cloud also frees the user from maintaining computer hardware. Cloud providers maintain data centers in a way that achieves economies of scale. Users need not be concerned with outages, software patches, service contracts, or damaged parts. That said, the task of recruiting and maintaining cloud resources appropriate for one's needs is itself a complex administrative task that can require a professional administrator or a dedicated effort to learn.

For scientific users, the cloud has two other major advantages: reproducibility and global access. Cloud resources are rented in virtualized slices called *instances*. Providers advertise a stable menu of instance types, each with defined capabilities: a certain processor speed, amount of disk space, amount of memory, etc. This predictability extends to the software running on the instances; the user decides exactly which software catalog should be pre-installed, including the operating system and software. This makes it easy for two users, perhaps on opposite ends of the globe, to create near-identical hardware and software setups. Similar reproducibility advantages are possible using virtual machines[8,9] and Docker containers[10,11], but the cloud makes it particularly easy to do this for entire clusters of computers, plus the software used to tie them together. Frameworks like StarCluster (http://star.mit.edu), Elastic MapReduce (https://aws.amazon.com/emr/) make this easier, and genomics frameworks like Galaxy[12,13] and Omics Pipe[14] extend these advantages to typical genomics workflows.

The cloud is also globally accessible. A user anywhere in the world can rent resources from a provider, regardless of whether the user is near a data center. Data can be secured and controlled by the collaborators, without having to navigate several institutions' firewalls. Team members can use the same commands to run the same analysis on the same (virtualized) hardware and software. This makes the cloud an attractive venue for large genomics collaborations, and an important tool in the effort to break down data silos and promote robust sharing of genomics data[15]. The cloud is also the substrate for the NIH Data Commons Pilot, an effort to increase availability and utility of data and software from NIH-funded efforts[16,17].

The cloud has disadvantages as well. Depending on the user's financial incentives, the cloud might be more expensive than a local cluster[18]. While building and maintaining a local cluster is work-intensive, initial costs are amortized over the cluster's lifetime. The marginal cost of doing one more computation is low; once hard disks have been purchased and set up, storing one more gigabyte of data is practically free. In the cloud, on the other hand, each experiment incurs a usage-based cost. That said, cloud providers do allow a degree of cost amortization through sustained use discounts, through use of partially prepaid "dedicated" instances, or through resellers or enterprise agreements. While funding agencies are increasingly willing to support these costs, as evidenced by the Commons Credit Pilot (https://www.commons-credit-portal.org) and the NSF Jetstream science cloud[19], this represents only a fraction of the computational cost of science.

Another key issue is data privacy. Later we discuss ways in which the cloud can make it easier to adhere to privacy standards like dbGaP. But the overall issue of privacy is not so simple. For one thing, stronger privacy standards like HIPAA add substantial complication and expense on top of what's needed for dbGaP compliance. And while it may be possible to adhere to the relevant standards, the user must often invest substantial time and effort to convince other bodies, such as Internal Review Boards or Information Technology administrators, that the standards are indeed being followed[20].

Revitalizing archived data

The Sequence Read Archive (SRA)[2], the central archive in the US for published sequencing data hosted at the National Center for Biotechnology Information (NCBI), now contains over 12 petabases of data from over 100,000 distinct studies. These data are heterogeneous, spanning many species, individuals, tissues, sequencing instruments, and assays. Vast archives like this allow researchers to reproduce past studies or to borrow data from those studies to address new questions. This is an avenue more research groups are taking. For example, one recent study gathered human RNA sequencing samples from various projects, including TCGA and ENCODE[21], to create a catalog of long non-coding RNAs (lncRNAs)[22]. Another reanalyzed RNA sequencing data from modENCODE[23,24], which studied transcription in *Drosophila melanogaster* over developmental time. The reanalysis focused on gene expression in an endosymbiont, *Wolbachia pipientis*[22,25], yielding new insights into gene expression patterns related to symbiosis.

In short, public archives are comprehensive enough to allow researchers to ask and answer a broad range of sophisticated questions without generating new data. But there are major obstacles to using archived data. One is computational. Downloading, storing and analyzing the data is resource-intensive, subject to bottlenecks like Internet uplink speeds, and beyond the skill set of many researchers.

A second major obstacle is data quality, both of raw data and metadata. Not all datasets in the SRA are of high quality, or annotated with informative (or even correct) metadata. In the absence of reliable and automatic methods for dealing with poor quality data and metadata --- a problem that is receiving more attention[26,27] --- researchers must approach public data cautiously. We return to this issue in the following section.

Cloud computing addresses many problems posed by data archives. The cloud's elasticity allows users to scale computing resources in proportion to the amount of data being analyzed, sidestepping constraints imposed by local clusters. Input data can be downloaded directly to the cloud computers that will process it, without first traversing a particular investigator's cluster. If data are protected, e.g. by dbGaP, existing protocols make it possible to craft a compliant cloud-based computational setup[28]. Commands used to rent the cluster and run the software can be published or shared so that collaborators can do the same, sidestepping inter-cluster compatibility issues.

Why has cloud computing not been more popular? We and others have advocated for cloud computing in genomics since the early days of Amazon Web Services[29–31]. Archives like the

SRA were far smaller at the time, so public data was not the main concern. Instead the prevailing argument was that sequencing throughput was growing much faster than Moore's law, forcing labs to grow their computational resources quickly with no upper bound in sight[30,31]. Computational studies[32–35] showed cloud computing could be used to analyze data fresh from the sequencer at reasonable speed and cost. But this assumed data could be transferred to the cloud conveniently. Researchers found this cloud-transfer step to be a challenge, worrying because of the specter of having the data stolen, or simply too slow depending on Internet uplink speed. Though there are now commercial services that similarly advocate cloud-based analysis of freshly-sequenced samples (https://www.dnanexus.com, https://basespace.illumina.com, https://www.sevenbridges.com/platform/, http://globusgenomics.org), we expect reanalysis of public data to be a more durable motivation for use of cloud computing in genomics research.

Recent years have seen a series of studies applying cloud computing can be used to study large collections of public archived data. The ReCount effort[36] used Myrna[34] and computing resources rented from the AWS cloud to re-analyze 475 RNA-seq samples from 18 different experiments which; this is a small study by modern standards but constituted a large fraction of publicly available RNA-seq data in 2011. The study summarized the raw data into tables giving expression levels summarized genomewide at the level of genes and exons. The Intropolis[37] and recount2[38] efforts used AWS and a custom cloud-enabled RNA-seq aligner[39] to re-analyze over 70,000 human samples spanning TCGA, GTEx, and the SRA, releasing expression-level summaries at the level of splice junctions, exons, genes and individual genomic bases. The Toil effort[40] used Spark-based computational pipeline and AWS to analyze nearly 20,000 samples spanning 4 major studies, including TCGA and GTEx. After a total of about 1.3M core hours of work, the estimated cost was about $1.30 per sample, but a much lower per-sample cost of $0.19 was achieved using an alternate pipeline based on the Kallisto pseudo-alignment tool[41]. Another effort[42] used Google Cloud Platform to quantify transcript expression levels for over 12,000 RNA-seq samples from large cancer projects. One the pipelines proposed in this work, again based on Kallisto, was shown to achieve a cost of just $0.09 per sample, the lowest of any effort to date and a small fraction of the cost of sequencing[43,44].

Combined with similar efforts that used large local clusters rather than cloud computing[45,46], the field has produced an array of uniformly processed and summarized datasets, summarized in Table 1. While these developments are recent, we expect resources like these to become popular starting points for studies that derive new conclusions from archived data. Microarrays are an instructive precedent: once appropriate resources and methods were available, it became common to combine and reanalyze large collections of microarray data, e.g. for platform comparisons[47], improved methods[48,49], meta-analyses[50,51] and clinical predictors[52,53]. Cloud computing enables this for sequencing data as well.

A shared computational laboratory

Because of the complexity of genomics studies and the need to enroll patients in geographically dispersed labs, collaboration on large-scale genomics sequencing projects across multiple sites is quite common. Before a computational analysis begins, all relevant data is gathered at whichever site has the requisite computing capacity and expertise. In practice, it is common for more than one site to analyze the full dataset, necessitating copies and transfers. The larger and more decentralized the project, the more copies must be made, yielding an adverse multiplier effect[31] (Figure 2a). The cloud combats this by providing a single common venue for data and computation (Figure 2b). Collaborators at the various sites can use computers located near the data.

This approach is exemplified by the NCI Cancer Genomics Cloud Pilots, initiated in 2015 and managed separately by the Broad Institute, the Institute for Systems Biology, and Seven Bridges Genomics (https://cbiit.nci.nih.gov/ncip/nci-cancer-genomics-cloud-pilots/nci-cloud-initiative). The goal was to facilitate querying and reanalysis of large cancer datasets like TCGA (https://medium.com/@theNCI/advancing-a-national-cancer-knowledge-system-a1c016046fbe). Two of the three efforts, the Broad's FireCloud and Seven Bridges' Cancer Genomics Cloud (CGC), additionally allow users to upload their own data and perform analyses in workspaces that can be shared privately with collaborators or publicly with the broader community (https://software.broadinstitute.org/firecloud/documentation/, http://docs.cancergenomicscloud.org/docs). Common analysis workflows are provided, but users can develop and share new workflows using the Common Workflow Language (CWL)[54]. One recently published workflow uses this platform to detect patient-specific tumor neoantigens from sequencing data[55].

FireCloud and CGC rely on Amazon Web Services and the Google Cloud Platform for compute and data storage (https://software.broadinstitute.org/firecloud/documentation/, http://docs.cancergenomicscloud.org/docs). In addition to charges for these commercial services, users pay convenience surcharges. By contrast, since 2007, the Galaxy Project[56] has allowed executing sharable analysis workflows for free through its main public server, which uses computing hardware at the Texas Advanced Computing Center (TACC), part of the NSF-supported Extreme Science and Engineering Discovery Environment (XSEDE)[57]. According to the project website (https://galaxyproject.org/main/), hardware allocated exclusively for Galaxy users spans the Rodeo cluster, with 256 processing cores and 2 TB of memory, and Corral, with 20 PB of disk space. Hardware shared with non-Galaxy users includes the Stampede cluster, with over 400,000 processing cores and 205 TB of memory. However, a registered user has a 250-GB disk space quota for their own data and is limited to running 6 concurrent jobs, each using at most 16 processing cores. Galaxy can also be used with Jetstream[57], a large-scale cloud computing resource that is also part of XSEDE and broadly serves as an alternative to commercial cloud computing services. Investigators can request XSEDE allocations to use Jetstream (https://www.xsede.org/allocations), and analyses can be run there through the Galaxy main server.

Collaboration using Galaxy is less direct than on FireCloud and CGC. Multiple collaborators do not manage an analysis in a shared workspace; rather, a single user completes all or part of an analysis, and Galaxy records its history, or the sequence of intermediate and final outputs (https://galaxyproject.org/tutorials/histories/). The history is then shared with another Galaxy user, and they import it, essentially creating a new branch of the history. The history can also be published for public consumption.

The Galaxy interface can run atop different computing infrastructures, and the community has set up over 80 public servers besides the main server (https://galaxyproject.org/public-galaxy-servers/). Users can also install Galaxy on cloud clusters using CloudMan[12], which not only supports Amazon Web Services for pay-as-you-go storage and compute but also private and public clouds that leverage OpenStack[58] or OpenNebula[59]. Galaxy servers may also be launched on Jetstream (https://galaxyproject.org/cloud/jetstream/). Globus Genomics is an alternative way to use Galaxy on Amazon Web Services. An initiative by the Computation Institute at the University of Chicago, Globus Genomics not only combines an enhanced Galaxy instance for workflow management with Amazon Web Services for powering computational analyses[60], but also uses the GridFTP-backed Globus Online[61] to speed data transfer among endpoints. These endpoints include Amazon's cloud storage service S3 and major sequencing centers[62].

Another way to combat the multiplier effect would be to distribute different parts of a dataset across multiple sites while leveraging the computational resources at those sites (Figure 2c). A layer of software coordinates computational activity across sites, turning them into a federated cloud that can run a common analysis workflow while enforcing co-location of data and the compute infrastructure used to analyze them. This is the approach of projects such as Beacon from the Global Alliance for Genomics and Health (GA4GH)[63]. It is also used by the Collaborative Cancer Cloud (CCC), a partnership among Intel, Oregon Health and Science University, the Dana-Farber Cancer Institute, and the Ontario Institute for Cancer Research (http://www.dana-farber.org/Newsroom/News-Releases/dana-farber-and-ontario-institute-for-cancer-research-join-collaborative-cancer-cloud.aspx). The CCC will be a platform that allows cancer researchers to search for patient omics data across multiple sites and perform analyses while keeping identifying information about patients secure (http://siliconangle.com/blog/2016/12/16/collaborative-cancer-cloud-intel-ohsu-team-cancer-research-thecube/).

Many collaborations already use the cloud to consolidate project data. ENCODE uses DNANexus for cloud-based analysis and data sharing, and DNANexus in turn uses Amazon Web Services (https://aws.amazon.com/solutions/case-studies/dnanexus/). modENCODE and ICGC host their datasets in the cloud through Amazon Web Services (http://data.modencode.org/modencode-cloud.html, http://docs.icgc.org/cloud/about/).

Conclusions

Cloud computing is a popular and enduring paradigm that has changed how large and small companies manage computational resources. Increasingly, it is also changing how scientists in genomics and in other fields collaborate and deal with vast archive datasets. As the cloud makes inroads into genomics, it is important for researchers to understand it and the new modes of analysis and collaboration it enables. For readers looking for further information about cloud computing and its uses in genomes, we recommend the "Galaxy on the Cloud" tutorial (https://galaxyproject.org/cloud/), the "Informatics for RNA-seq" site (https://github.com/griffithlab/rnaseq_tutorial/wiki).

While the cloud can be viewed as an alternate way of running existing genomics software, it has also spurred software deeper thinking about how to design software for large datasets. A typical approach is to take software originally designed to analyze a single sample on a single computer, then attempt to scale it by launching many simultaneous copies, each analyzing a distinct sample. But the rise of cloud computing has driven a rise in certain programming frameworks, MapReduce in particular[64], that make it easier to scale software to the kinds of large clusters available from cloud providers. In return, the programmer must adhere to certain permitted programming patterns. MapReduce is discussed in other reviews[30,65,66] but a prime advantage is that it allows programs to scalably aggregate and sort data in between computational steps[64]. For genomics this makes it particularly easy to analyze many samples at once, naturally allowing "strength borrowing" across replicates[39]. MapReduce and similar frameworks have been used in many types of sequencing data analysis, including variant calling[33], RNA-seq[34,39], and ChIP-seq analysis[67]. Notably, the popular GATK variant caller[68] uses MapReduce to parallelize population-scale variant calling.

By making it easier to leverage public data, cloud computing encourages another dimension of "strength borrowing." That is, researchers can use public data to boost the power available to analyze a locally-generated dataset, a paradigm that already prevails in microarray data analysis[49]. We expect this trend to continue and evolve, even to the point new sequencing data analyses are performed in the cloud and with the benefit of being able to "see across" many past studies with important variables in common. The RNASeq-er API[46] and Xena (http://xena.ucsc.edu) resources show how this is already possible for RNA-seq.

As the cloud gains users, funding agencies must learn about and facilitate this new kind of computing. Funders have traditionally viewed computing as a one-time expense yielding benefits that can be spread over future projects. The cloud model is radically different: all costs are ongoing, and little or nothing is subsidized or amortized. Funders do seem to be making adjustments for cloud computing, though. The NIH is funding cloud computing activities through its Commons Credit Pilot (https://www.commons-credit-portal.org) as is the National Science Foundation through its BIGDATA program (https://www.nsf.gov/news/news_summ.jsp?cntn_id=190830&WT.mc_ev=click). Both programs allow investigators to apply for cloud credits redeemable with major cloud providers. The NSF-funded Jetstream resource[19] is a cloud geared toward sciences such as genomics that have not traditionally made heavy use of supercomputers. While smaller than the commercial clouds, Jetstream can handle cloud workloads without charge, via grants through the XSEDE program[57].

Funders can further control costs for grantees by working to keep key genomic datasets in cloud storage, as suggested previously[5]. NIH and others are beginning to do this, as evidenced by the NIMH Data Archive (https://ndar.nih.gov/index.html), and Cancer Genomics Cloud (CGC) Pilot[6]. Staging the raw data in the cloud at the outset greatly reduces the effort and cost of moving the data to the cloud computers doing the work. Cloud providers are aware of this advantage, with AWS and Google advertising availability of data from several large projects, including the 1000 Genomes Project[69] on their cloud storage services (https://aws.amazon.com/public-datasets/, https://cloud.google.com/genomics/v1/public-data).

Acknowledgements

We thank James Taylor, Enis Afgan, Mike Schatz, Jeremy Goecks, and Adam Margolin for reading through a draft of this work and providing helpful comments.

Funding

BL was supported by NIH/NIGMS grant 1R01GM118568 to BL.

Bibliography

1. Melé M et al. Human genomics. The human transcriptome across tissues and individuals. Science 348, 660–665, doi:10.1126/science.aaa0355 (2015). [PubMed: 25954002]
2. Leinonen R, Sugawara H, Shumway M & on behalf of the International Nucleotide Sequence Database, C. The Sequence Read Archive. Nucleic Acids Res 39, D19–D21, doi:10.1093/nar/gkq1019 (2010). [PubMed: 21062823]
3. Denk F Don't let useful data go to waste. Nature 543, 7, doi:10.1038/543007a (2017). [PubMed: 28252084]
4. Yung CK et al. Abstract 3605: ICGC in the cloud. Cancer Res 76, 3605–3605, doi:10.1158/1538-7445.am2016-3605 (2016).
5. Stein LD, Knoppers BM, Campbell P, Getz G & Korbel JO Data analysis: Create a cloud commons. Nature 523, 149–151, doi:10.1038/523149a (2015). [PubMed: 26156357]
6. Davis-Dusenbery BN Petabyte-Scale Cancer Genomics in the Cloud. Cancer Genet 208, 360, doi:10.1016/j.cancergen.2015.05.012 (2015).
7. Trinh QM et al. Cloud-based uniform ChIP-Seq processing tools for modENCODE and ENCODE. BMC Genomics 14, 494, doi:10.1186/1471-2164-14-494 (2013). [PubMed: 23875683]
8. Angiuoli SV et al. CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. BMC Bioinformatics 12, 356, doi:10.1186/1471-2105-12-356 (2011). [PubMed: 21878105]
9. Krampis K et al. Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community. BMC Bioinformatics 13, 42, doi:10.1186/1471-2105-13-42 (2012). [PubMed: 22429538]
10. Beaulieu-Jones BK & Greene CS Reproducibility of computational workflows is automated using continuous analysis. Nat. Biotechnol 35, 342–346, doi:10.1038/nbt.3780 (2017). [PubMed: 28288103]
11. Boettiger C An introduction to Docker for reproducible research. Oper. Syst. Rev 49, 71–79, doi:10.1145/2723872.2723882 (2015).
12. Afgan E et al. Galaxy CloudMan: delivering cloud compute clusters. BMC Bioinformatics 11 Suppl 12, S4, doi:10.1186/1471-2105-11-S12-S4 (2010).
13. Sloggett C, Goonasekera N & Afgan E BioBlend: automating pipeline analyses within Galaxy and CloudMan. Bioinformatics 29, 1685–1686, doi:10.1093/bioinformatics/btt199 (2013). [PubMed: 23630176]
14. Fisch KM et al. Omics Pipe: a community-based framework for reproducible multi-omics data analysis. Bioinformatics 31, 1724–1728, doi:10.1093/bioinformatics/btv061 (2015). [PubMed: 25637560]
15. Clinical Cancer Genome Task Team of the Global Alliance for, G. et al. Sharing Clinical and Genomic Data on Cancer - The Need for Global Solutions. N. Engl. J. Med 376, 2006–2009, doi:10.1056/NEJMp1612254 (2017). [PubMed: 28538124]
16. Bonazzi VR & Bourne PE Should biomedical research be like Airbnb? PLoS Biol 15, e2001818, doi:10.1371/journal.pbio.2001818 (2017). [PubMed: 28388615]
17. Bourne PE, Lorsch JR & Green ED Perspective: Sustaining the big-data ecosystem. Nature 527, S16–17, doi:10.1038/527S16a (2015). [PubMed: 26536219]
18. Marx V Genomics in the clouds. Nat. Methods 10, 941–945, doi:10.1038/nmeth.2654 (2013). [PubMed: 24076987]

19. Stewart CA et al. in Proceedings of the 2015 XSEDE Conference on Scientific Advancements Enabled by Enhanced Cyberinfrastructure - XSEDE ‘15 (2015).

20. Datta S, Bettinger K & Snyder M Secure cloud computing for genomic data. Nat. Biotechnol 34, 588–591, doi:10.1038/nbt.3496 (2016).

21. Consortium EP An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74, doi:10.1038/nature11247 (2012). [PubMed: 22955616]

22. Iyer MK et al. The landscape of long noncoding RNAs in the human transcriptome. Nat. Genet 47, 199–208, doi:10.1038/ng.3192 (2015). [PubMed: 25599403]

23. Brown JB et al. Diversity and dynamics of the Drosophila transcriptome. Nature 512, 393–399, doi:10.1038/nature12962 (2014). [PubMed: 24670639]

24. Graveley B The developmental transcriptome of Drosophila melanogaster. Genome Biol 11, I11, doi:10.1186/gb-2010-11-s1-i11 (2010).

25. Gutzwiller F et al. Dynamics of Wolbachia pipientis Gene Expression Across the Drosophila melanogaster Life Cycle. G3 5, 2843–2856, doi:10.1534/g3.115.021931 (2015). [PubMed: 26497146]

26. Bernstein MN, Doan A & Dewey CN MetaSRA: normalized sample-specific metadata for the Sequence Read Archive. doi:10.1101/090506 (2016).

27. Galeota E & Pelizzola M Ontology-based annotations and semantic relations in large-scale (epi)genomics data. Brief. Bioinform 18, 403–412, doi:10.1093/bib/bbw036 (2017). [PubMed: 27142216]

28. Nellore A, Wilks C, Hansen KD, Leek JT & Langmead B Rail-dbGaP: analyzing dbGaP-protected data in the cloud with Amazon Elastic MapReduce. Bioinformatics 32, 2551–2553, doi:10.1093/bioinformatics/btw177 (2016). [PubMed: 27153614]

29. Bateman A & Wood M Cloud computing. Bioinformatics 25, 1475, doi:10.1093/bioinformatics/btp274 (2009). [PubMed: 19435745]

30. Schatz MC, Langmead B & Salzberg SL Cloud computing and the DNA data race. Nat. Biotechnol 28, 691–693, doi:10.1038/nbt0710-691 (2010). [PubMed: 20622843]

31. Stein LD The case for cloud computing in genome informatics. Genome Biol 11, 207, doi:10.1186/gb-2010-11-5-207 (2010). [PubMed: 20441614]

32. Schatz MC CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics 25, 1363–1369, doi:10.1093/bioinformatics/btp236 (2009). [PubMed: 19357099]

33. Langmead B, Schatz MC, Lin J, Pop M & Salzberg SL Searching for SNPs with cloud computing. Genome Biol 10, R134, doi:10.1186/gb-2009-10-11-r134 (2009). [PubMed: 19930550]

34. Langmead B, Hansen KD & Leek JT Cloud-scale RNA-sequencing differential expression analysis with Myrna. Genome Biol 11, R83, doi:10.1186/gb-2010-11-8-r83 (2010). [PubMed: 20701754]

35. Kelly BJ et al. Churchill: an ultra-fast, deterministic, highly scalable and balanced parallelization strategy for the discovery of human genetic variation in clinical and population-scale genomics. Genome Biol. 16, 6, doi:10.1186/s13059-014-0577-x (2015). [PubMed: 25600152]

36. Frazee AC, Langmead B & Leek JT ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. BMC Bioinformatics 12, 449, doi:10.1186/1471-2105-12-449 (2011). [PubMed: 22087737]

37. Nellore A et al. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. Genome Biol 17, 266, doi:10.1186/s13059-016-1118-6 (2016). [PubMed: 28038678]

38. Collado-Torres L et al. Reproducible RNA-seq analysis using recount2. Nat. Biotechnol 35, 319–321, doi:10.1038/nbt.3838 (2017). [PubMed: 28398307]

39. Nellore A et al. Rail-RNA: scalable analysis of RNA-seq splicing and coverage. Bioinformatics, doi:10.1093/bioinformatics/btw575 (2016).

40. Vivian J et al. Rapid and efficient analysis of 20,000 RNA-seq samples with Toil. doi:10.1101/062497 (2016).

41. Schaeffer L, Pimentel H, Bray N, Melsted P & Pachter L Pseudoalignment for metagenomic read assignment. arXiv [q-bio.QM] (2015).

42. Tatlow PJ & Piccolo SR A cloud-based workflow to quantify transcript-expression levels in public cancer compendia. Sci. Rep 6, 39259, doi:10.1038/srep39259 (2016). [PubMed: 27982081]

43. Lohman BK, Weber JN & Bolnick DI Evaluation of TagSeq, a reliable low-cost alternative for RNAseq. Mol. Ecol. Resour 16, 1315–1321, doi:10.1111/1755-0998.12529 (2016). [PubMed: 27037501]

44. Combs PA & Eisen MB Low-cost, low-input RNA-seq protocols perform nearly as well as high-input protocols. PeerJ 3, e869, doi:10.7717/peerj.869 (2015). [PubMed: 25834775]

45. Rahman M et al. Alternative preprocessing of RNA-Sequencing data in The Cancer Genome Atlas leads to improved analysis results. Bioinformatics 31, 3666–3672, doi:10.1093/bioinformatics/btv377 (2015). [PubMed: 26209429]

46. Petryszak R et al. The RNASeq-er API—a gateway to systematically updated analysis of public RNA-seq data. Bioinformatics, doi:10.1093/bioinformatics/btx143 (2017).

47. Kuo WP, Jenssen T-K, Butte AJ, Ohno-Machado L & Kohane IS Analysis of matched mRNA measurements from two different microarray technologies. Bioinformatics 18, 405–412 (2002). [PubMed: 11934739]

48. Leek JT et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat. Rev. Genet 11, 733–739, doi:10.1038/nrg2825 (2010). [PubMed: 20838408]

49. McCall MN, Bolstad BM & Irizarry RA Frozen robust multiarray analysis (fRMA). Biostatistics 11, 242–253, doi:10.1093/biostatistics/kxp059 (2010). [PubMed: 20097884]

50. Franke A et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn’s disease susceptibility loci. Nat. Genet 42, 1118–1125, doi:10.1038/ng.717 (2010). [PubMed: 21102463]

51. Zeggini E et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat. Genet 40, 638–645, doi:10.1038/ng.120 (2008). [PubMed: 18372903]

52. Rhodes DR et al. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc. Natl. Acad. Sci. U. S. A 101, 9309–9314, doi:10.1073/pnas.0401994101 (2004). [PubMed: 15184677]

53. Marchionni L, Afsari B, Geman D & Leek JT A simple and reproducible breast cancer prognostic test. BMC Genomics 14, 336, doi:10.1186/1471-2164-14-336 (2013). [PubMed: 23682826]

54. Amstutz P et al. Common Workflow Language, v1.0. (2016).

55. Bais P, Namburi S, Gatti DM, Zhang X & Chuang JH CloudNeo: A cloud pipeline for identifying patient-specific tumor neoantigens. Bioinformatics, doi:10.1093/bioinformatics/btx375 (2017).

56. Afgan E et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res 44, W3–W10, doi:10.1093/nar/gkw343 (2016). [PubMed: 27137889]

57. Towns J et al. XSEDE: Accelerating Scientific Discovery. Comput. Sci. Eng 16, 62–74, doi:10.1109/mcse.2014.80 (2014).

58. Sefraoui O, Aissaoui M & Eleuldj M OpenStack: Toward an Open-source Solution for Cloud Computing. Int. J. Comput. Appl. Technol 55, 38–42, doi:10.5120/8738-2991 (2012).

59. Molina D et al. in Open Source Cloud Computing Systems 19–43.

60. Liu B et al. Cloud-based bioinformatics workflow platform for large-scale next-generation sequencing analyses. J. Biomed. Inform 49, 119–133, doi:10.1016/j.jbi.2014.01.005 (2014). [PubMed: 24462600]

61. Foster I Globus Online: Accelerating and Democratizing Science through Cloud-Based Services. IEEE Internet Comput 15, 70–73, doi:10.1109/mic.2011.64 (2011).

62. Madduri RK et al. Experiences Building Globus Genomics: A Next-Generation Sequencing Analysis Service using Galaxy, Globus, and Amazon Web Services. Concurr. Comput 26, 2266–2279, doi:10.1002/cpe.3274 (2014). [PubMed: 25342933]

63. Global Alliance for, G. & Health. GENOMICS. A federated ecosystem for sharing genomic, clinical data. Science 352, 1278–1280, doi:10.1126/science.aaf6162 (2016). [PubMed: 27284183]

64. Dean J & Ghemawat S MapReduce. Commun. ACM 51, 107, doi:10.1145/1327452.1327492 (2008).

65. Taylor RC An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. BMC Bioinformatics 11 Suppl 12, S1, doi:10.1186/1471-2105-11-S12-S1 (2010).

66. O'Driscoll A, Daugelaite J & Sleator RD 'Big data', Hadoop and cloud computing in genomics. J. Biomed. Inform 46, 774–781, doi:10.1016/j.jbi.2013.07.001 (2013). [PubMed: 23872175]

67. Feng X, Grossman R & Stein L PeakRanger: a cloud-enabled peak caller for ChIP-seq data. BMC Bioinformatics 12, 139, doi:10.1186/1471-2105-12-139 (2011). [PubMed: 21554709]

68. McKenna A et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20, 1297–1303, doi:10.1101/gr.107524.110 (2010). [PubMed: 20644199]

69. Genomes Project, C. et al. A global reference for human genetic variation. Nature 526, 68–74, doi:10.1038/nature15393 (2015). [PubMed: 26432245]

70. Cancer Genome Atlas Research, N. et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet 45, 1113–1120, doi:10.1038/ng.2764 (2013). [PubMed: 24071849]

71. Vivian J et al. Toil enables reproducible, open source, big biomedical data analyses. Nat. Biotechnol 35, 314–316, doi:10.1038/nbt.3772 (2017). [PubMed: 28398314]

72. Barretina J et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483, 603–607, doi:10.1038/nature11003 (2012). [PubMed: 22460905]

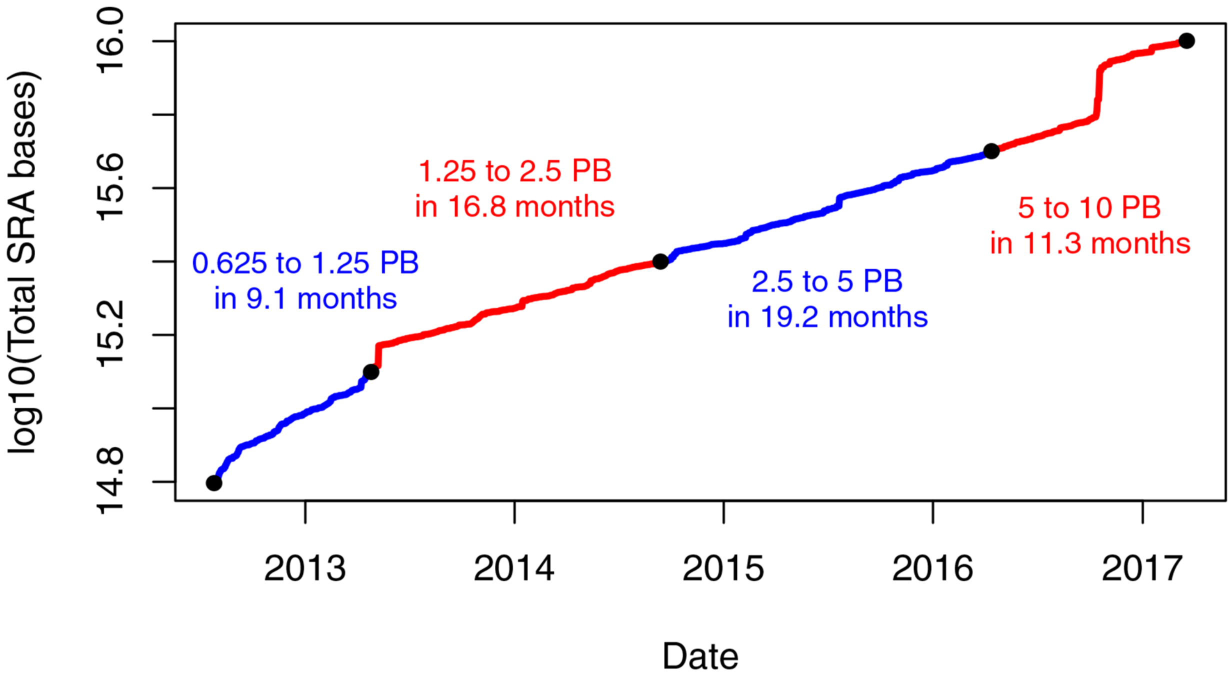

Figure 1:

Four doublings of the Sequence Read Archive from July 2012 to March 2017. The large jump in October 2016 is chiefly due to the TopMed project. As of June 2017, the SRA contains over 12 petabases (millions of billions of bases) of data.

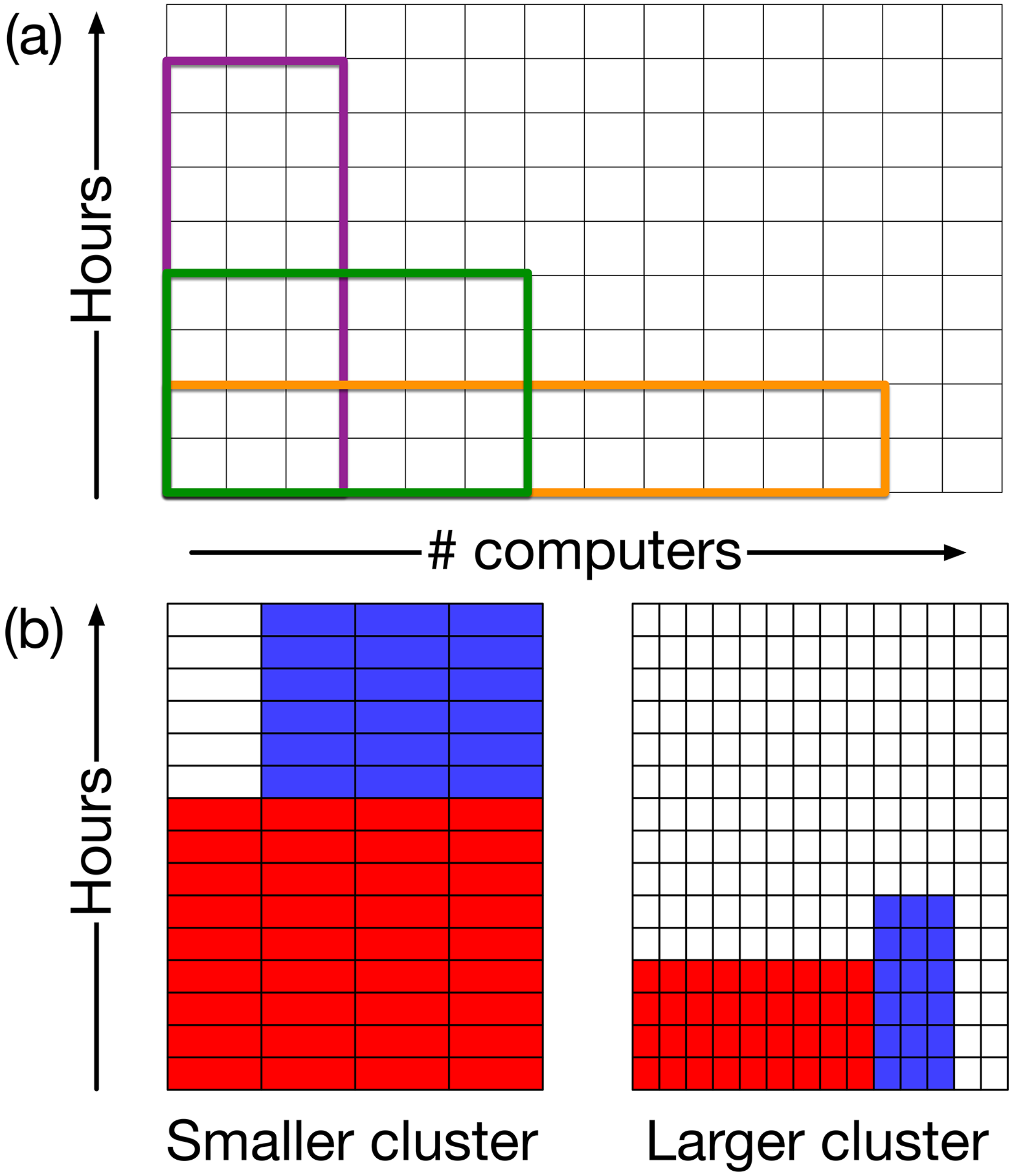

Figure 2:

Elasticity allows the user to rent resources while paying only for what gets used. Panel (a) illustrates a scenario with two computational tasks to perform, colored red and blue. The red task requires 36 computer-hours and runs on up to 8 computers simultaneously. The blue task requires 18 computer-hours and runs on 3 computers simultaneously. On a smaller cluster (left) both the tasks run sequentially and require 15 hours to complete. On a larger cluster (right), representing a cloud cluster, the tasks can run simultaneously and the red task can use its full complement of 8 computers. As a result, both complete within 6 hours. This ignores the fact that many more users are contending for cloud clusters than are contending for an institutional cluster. The greater number of users is mitigated by the fact that needs and timing vary from user to user. Cloud providers also provide incentives, such as spot pricing, to encourage renting at less busy times.

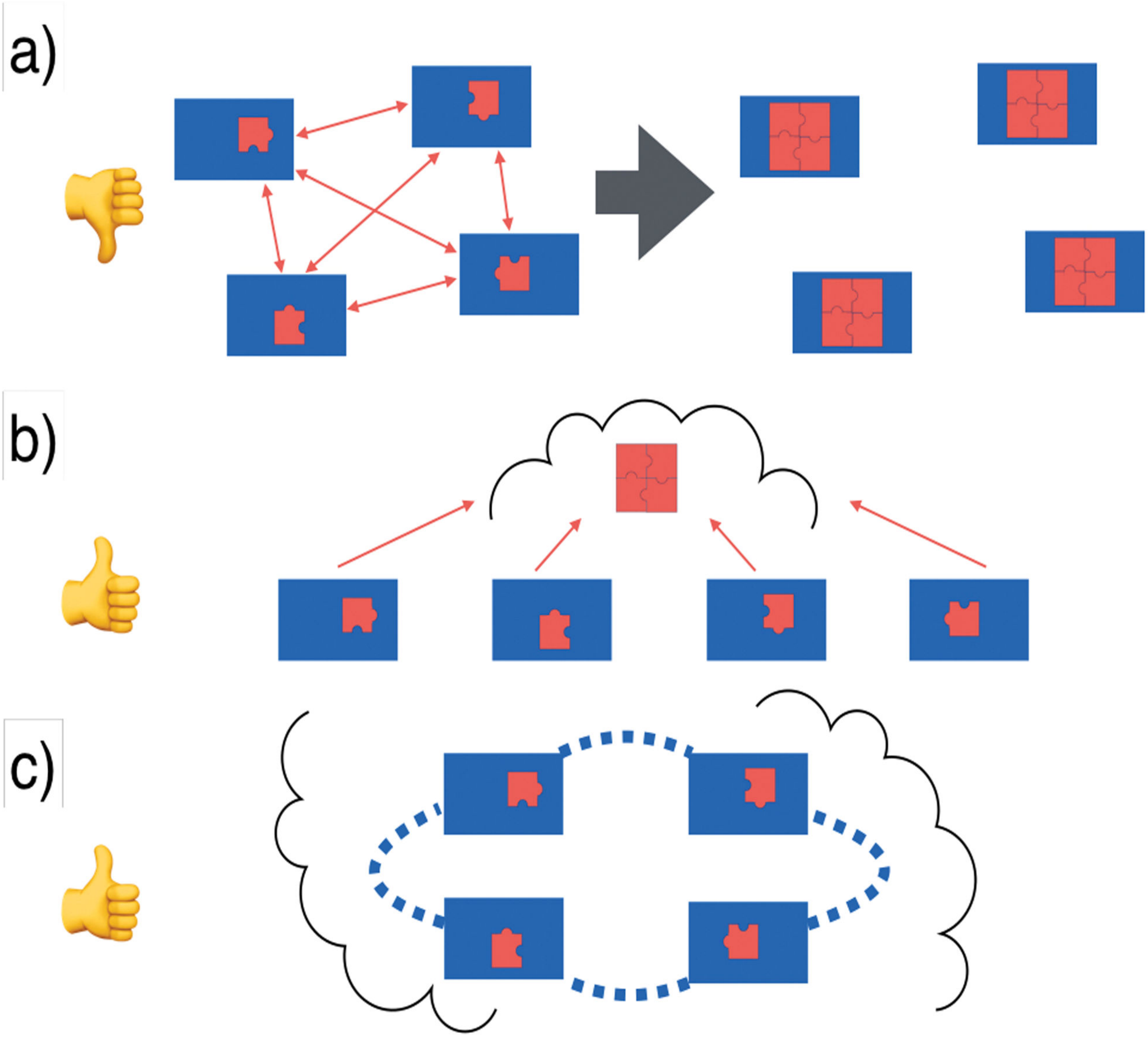

Figure 3:

Each site (blue rectangle) has some computational resources and also generates a portion of the data (red puzzle pieces). In (a) analysis that require the full datasets are to be performed at multiple sites, requiring each of these sites to gather all portions of the data. As more sites join the analysis, more copies must be made. (b) and (c) are alternate solutions. In (b) sites consolidate their data in a cloud-based data center, where all analyses are performed. In (c), multiple sites organize themselves into a federated cloud, where each analysis of the full dataset is automatically coordinated to minimize data transfer. Where possible, the computers located where data are generated are used to analyze that subset.

Table 1:

Cloud-based efforts to re-analyze RNA sequencing samples from large consortium projects and public archives.

Effort	Data summarized	Summaries	URL	Resources
recount2 [38]	• 9,662 human RNA-seq samples in the V6 release of GTEx [1]. Various tissues, many tissues per individual. • 11,350 RNA-seq cases from TCGA [70] • 50,186 human RNA-seq runs from the Sequence read archive	Junction-, exon-, gene- level expression measurements. Genomewide coverage.	https://jhubiostatistics.shinyapps.io/recount/	Amazon Web Services
Toil [71]	• 19,952 human mRNA-seq samples from 4 large studies: TCGA, TARGET (https://ocg.cancer.gov/programs/target), PNOC (http://www.pnoc.us/) and GTEx	Gene- and isoform-level expression measurements. Genomewide coverage for GTEx.	http://xena.ucsc.edu https://genome.ucsc.edu/	Amazon Web Services
Tatlow and Piccolo [42]	• 12,307 RNA-Sequencing samples from the Cancer Cell Line Encyclopedia [72] and TCGA	Isoform-level expression measurements	https://osf.io/gqrz9/	Google Cloud Platform