Stein *Genome Biology* 2010, **11**:207
http://genomebiology.com/2010/11/5/207

Genome Biology

REVIEW

The case for cloud computing in genome informatics

Lincoln D Stein*

Abstract

With DNA sequencing now getting cheaper more quickly than data storage or computation, the time may have come for genome informatics to migrate to the cloud.

The impending collapse of the genome informatics ecosystem

Since the 1980s, we have had the great fortune to work in a comfortable and effective ecosystem for the production and consumption of genomic information (Figure 1). Sequencing labs submit their data to big archival databases such as GenBank at the National Center for Biotechnology Information (NCBI) [1], the European Bioinformatics Institute EMBL database [2], DNA Data Bank of Japan (DDBJ) [3], the Short Read Archive (SRA) [4], the Gene Expression Omnibus (GEO) [5] and the microarray database ArrayExpress [6]. These databases maintain, organize and distribute the sequencing data. Most users access the information either through websites created by the archival databases, or through value-added integrators of genomic data, such as Ensembl [7], the University of California at Santa Cruz (UCSC) Genome Browser [8], Galaxy [9], or one of the many model organism databases [10-13]. Bioinformaticians and other power users download genomic data from these primary and secondary sources to their high performance clusters of computers ('compute clusters'), work with them and discard them when no longer needed (Figure 1).

The whole basis for this ecosystem is Moore's Law [14], a long-term trend first described in 1965 by Intel co-founder Gordon Moore. Moore's Law states that the number of transistors that can be placed on an integrated circuit board is increasing exponentially, with a doubling time of roughly 18 months. The trend has held up remarkably well for 35 years across multiple changes in semiconductor technology and manufacturing techniques. Similar laws for disk storage and network capacity have also been observed. Hard disk capacity doubles roughly annually (Kryder's Law [15]), and the cost of sending a bit of information over optical networks halves every 9 months (Butter's Law [16]).

Genome sequencing technology has also improved dramatically, and the number of bases that can be sequenced per unit cost has also been growing at an exponential rate. However, until just a few years ago, the doubling time for DNA sequencing was just a bit slower than the growth of compute and storage capacity. This

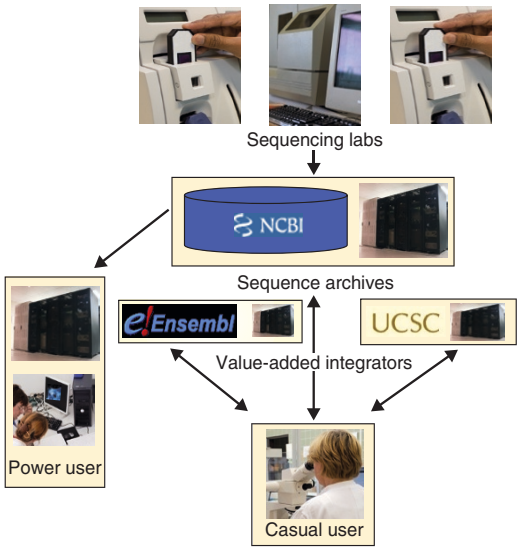

Figure 1. The old genome informatics ecosystem. Under the traditional flow of genome information, sequencing laboratories transmit raw and interpreted sequencing information across the internet to one of several sequencing archives. This information is accessed either directly by casual users or indirectly via a website run by one of the value-added genome integrators. Power users typically download large datasets from the archives onto their local compute clusters for computationally intensive number crunching. Under this model, the sequencing archives, value-added integrators and power users all maintain their own compute and storage clusters and keep local copies of the sequencing datasets.

*Correspondence: lincoln.stein@gmail.com
Ontario Institute for Cancer Research, Toronto, ON M5G 0A3, Canada

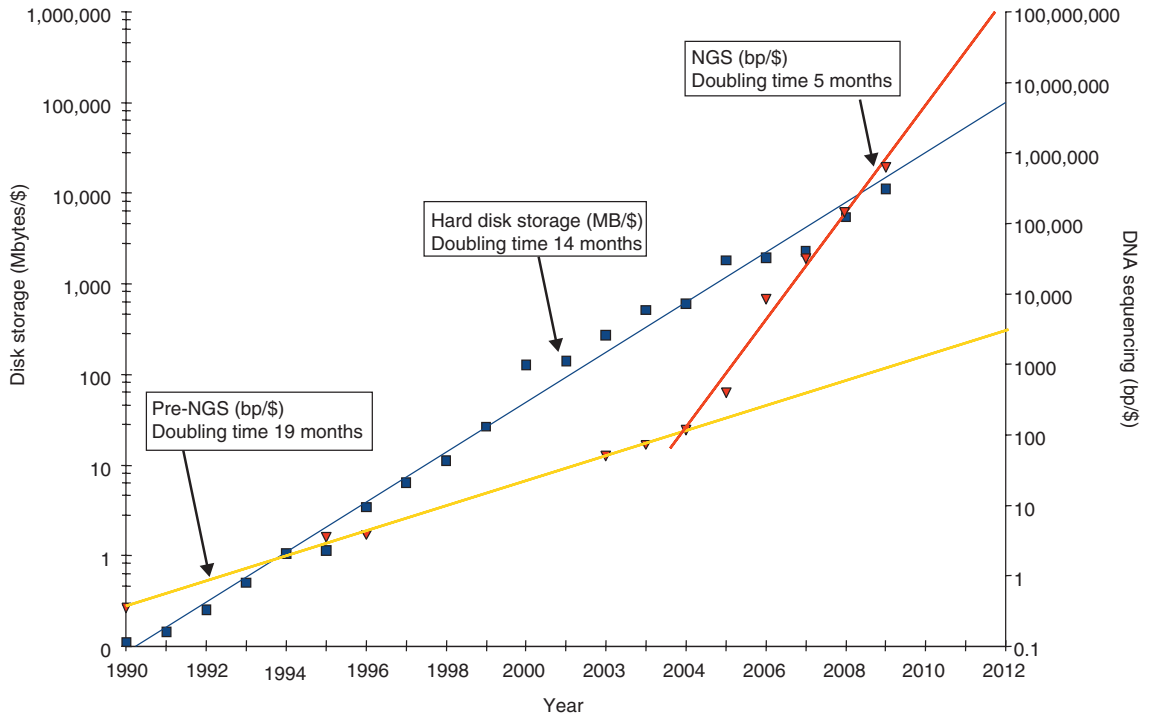

Figure 2. Historical trends in storage prices versus DNA sequencing costs. The blue squares describe the historic cost of disk prices in megabytes per US dollar. The long-term trend (blue line, which is a straight line here because the plot is logarithmic) shows exponential growth in storage per dollar with a doubling time of roughly 1.5 years. The cost of DNA sequencing, expressed in base pairs per dollar, is shown by the red triangles. It follows an exponential curve (yellow line) with a doubling time slightly slower than disk storage until 2004, when next generation sequencing (NGS) causes an inflection in the curve to a doubling time of less than 6 months (red line). These curves are not corrected for inflation or for the 'fully loaded' cost of sequencing and disk storage, which would include personnel costs, depreciation and overhead.

was great for the genome informatics ecosystem. The archival databases and the value-added genome distributors did not need to worry about running out of disk storage space because the long-term trends allowed them to upgrade their capacity faster than the world's sequencing labs could update theirs. Computational biologists did not worry about not having access to sufficiently powerful networks or compute clusters because they were always slightly ahead of the curve.

However, the advent of 'next generation' sequencing technologies in the mid-2000s changed these long-term trends and now threatens the conventional genome informatics ecosystem. To illustrate this, I recently plotted long-term trends in hard disk prices and DNA sequencing prices by using the Internet Archive's 'Wayback Machine' [17], which keeps archives of websites as they appeared in the past, to view vendors' catalogs, websites and press releases as they appeared over the past 20 years (Figure 2). Notice that this is a logarithmic plot, so exponential curves appear as straight lines. I made no attempt to factor in inflation or to calculate the cost of DNA sequencing with labor and overheads included, but the trends are clear. From 1990 to 2010, the cost of storing a byte of data has halved every 14 months, consistent with Kryder's Law. From 1990 to 2004, the cost of sequencing a base decreased more slowly than this, halving every 19 months - good news if you are running the bioinformatics core for a genome sequencing center.

However, from 2005 the slope of the DNA sequencing curve increases abruptly. This corresponds to the advent of the 454 Sequencer [18], quickly followed by the Solexa/Illumina [19] and ABI SOLiD [20] technologies. Since then, the cost of sequencing a base has been dropping by half every 5 months. The cost of genome sequencing is now decreasing several times faster than the cost of storage, promising that at some time in the not too distant future it will cost less to sequence a base of DNA than to store it on a hard disk. Of course there is no guarantee that this accelerated trend will continue indefinitely, but recent and announced offerings from Illumina [21], Pacific Biosystems [22], Helicos [23] and Ion Torrent [24], among others, promise to continue the trend until the middle of the decade.

This change in the long-term trend overthrows the assumptions that support the current ecosystem. The various members of the genome informatics ecosystem are now facing a potential tsunami of genome data that will swamp our storage systems and crush our compute clusters. Just consider this one statistic: the first big genome project based on next generation sequencing technologies, the 1000 Genomes Project [25], which is cataloguing human genetic variation, deposited twice as much raw sequencing data into GenBank's SRA division during the project's first 6 months of operation as had been deposited into all of GenBank for the entire 30 years preceding (Paul Flicek, personal communication). But the 1000 Genomes Project is just the first ripple of the tsunami. Projects like ENCODE [26] and modENCODE [27], which use next generation sequencing for high-resolution mapping of epigenetic marks, chromatin-binding proteins and other functional elements, are currently generating raw sequence at tremendous rates. Cancer genome projects such as The Cancer Genome Atlas [28] and the International Cancer Genome Sequencing Consortium [29] are an order of magnitude larger than the 1000 Genomes Project, and the various Human Microbiome Projects [30,31] are potentially even larger still.

Run for the hills?

First, we must face up to reality. The ability of laboratories around the world to produce sequence faster and more cheaply than information technology groups can upgrade their storage systems is a fundamental challenge that admits no easy solution. At some future point it will become simply unfeasible to store all raw sequencing reads in a central archive or even in local storage. Genome biologists will have to start acting like the high energy physicists, who filter the huge datasets coming out of their collectors for a tiny number of informative events and then discard the rest.

Even though raw read sets may not be preserved in their entirety, it will remain imperative for the assembled genomes of animals, plants and ecological communities to be maintained in publicly accessible form. But these are also rapidly growing in size and complexity because of the drop in sequencing costs and the growth of derivative technologies such as chromatin immunoprecipitation with sequencing (ChIP-seq [32]), DNA methylation sequencing [33] and chromatin interaction mapping [34]. These large datasets pose significant challenges for both the primary and secondary genome sequence repositories who must maintain the data, as well as the 'power users' who are accustomed to downloading the data to local computers for analysis.

Reconsider the traditional genome informatics ecosystem of Figure 1. It is inefficient and wasteful in several ways. For the value-added genome integrators to do their magic with the data, they must download it from the archival databases across the internet and store copies in their local storage systems. The power users must do the same thing: either downloading the data directly from the archive, or downloading it from one of the integrators. This entails moving the same datasets across the network repeatedly and mirroring them in multiple local storage systems. When datasets are updated, each of the mirrors must detect that fact and refresh their copies. As datasets get larger, this process of mirroring and refreshing becomes increasingly cumbersome, error prone and expensive.

A less obvious inefficiency comes from the need of the archives, integrators and power users to maintain local compute clusters to meet their analysis needs. NCBI, UCSC and the other genome data providers maintain large server farms that process genome data and serve it out via the web. The load on the server farm fluctuates hourly, daily and seasonally. At any time, a good portion of their clusters is sitting idle, waiting in reserve for periods of peak activity when a big new genome dataset comes in, or a major scientific meeting is getting close. However, even though much of the cluster is idle, it still consumes electricity and requires the care of a systems administration staff.

Bioinformaticians and other computational biologists face similar problems. They can choose between building a cluster that is adequate to meet their everyday needs, or build one with the capacity to handle peak usage. In the former case, the researcher risks being unable to run an unusually involved analysis in reasonable running time and possibly being scooped by a competitor. In the latter case, they waste money purchasing and maintaining a system that they are not using to capacity much of the time.

These inefficiencies have been tolerable in a world in which most genome-scale datasets have fit on a DVD (uncompressed, the human genome is about 3 gigabytes). When datasets are measured in terabytes these inefficiencies add up.

Cloud computing to the rescue

Which brings us, at last, to 'cloud computing.' This is a general term for computation-as-a-service. There are various different types of cloud computing, but the one that is closest to the way that computational biologists currently work depends on the concept of a 'virtual machine.' In the traditional economic model of computation, customers purchase server, storage and networking hardware, configure it the way they need, and run software on it. In computation-as-a-service, customers essentially rent the hardware and storage for as long or as short a time as they need to achieve their

goals. Customers pay only for the time the rented systems are running and only for the storage they actually use.

This model would be lunatic if the rented machines were physical ones. However, in cloud computing, the rentals are virtual: without ever touching a power cable, customers can power up a fully functional 10-computer server farm with a terabyte of shared storage, upgrade the cluster in minutes to 100 servers when needed for some heavy duty calculations, and then return to the baseline 10-server system when the extra virtual machines are no longer needed.

The way it works is that a service provider puts up the capital expenditure of creating an extremely large compute and storage farm (tens of thousands of nodes and petabytes of storage) with all the frills needed to maintain an operation of this size, including a dedicated system administration staff, storage redundancy, data centers distributed to strategically placed parts of the world, and broadband network connectivity. The service provider then implements the infrastructure to give users the ability to create, upload and launch virtual machines on this compute farm. Because of economies of scale, the service provider can obtain highly discounted rates on hardware, electricity and network connectivity, and can pass these savings on to the end users to make virtual machine rental economically competitive with purchasing the real thing.

A virtual machine is a piece of software running on the host computer (the real hardware) that emulates the properties of a computer: the emulator provides a virtual central processing unit (CPU), network card, hard disk, keyboard and so forth. You can run the operating system of your choice on the virtual machine, log into it remotely via the internet, configure it to run web servers, databases, load management software, parallel computation libraries, and any other software you favor. You may be familiar with virtual machines from working with consumer products such as VMware [35] or open source projects such as KVM [36]. A single physical machine can host multiple virtual machines, and software running on the physical server farm can distribute requests for new virtual machines across the server farm in a way that intelligently distributes load.

The experience of working with virtual machines is relatively painless. Choose the physical aspects of the virtual machine you wish to make, including CPU type, memory size and hard disk capacity, specify the operating system you wish to run, and power up one or more machines. Within a couple of minutes, your virtual machines are up and running. Log into them over the network and get to work. When a virtual machine is not running, you can store an image of its bootable hard disk. You can then use this image as a template on which to start up multiple virtual machines, which is how you can launch a virtual compute cluster in a matter of minutes.

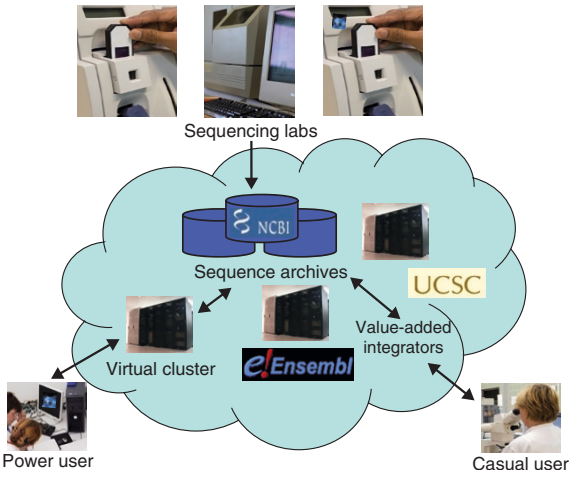

Figure 3. The 'new' genome informatics ecosystem based on cloud computing. In this model, the community's storage and compute resources are co-located in a 'cloud' maintained by a large service provider. The sequence archives and value-added integrators maintain servers and storage systems within the cloud, and use more or less capacity as needed for daily and seasonal fluctuations in usage. Casual users continue to access the data via the websites of the archives and integrators, but power users now have the option of creating virtual on-demand compute clusters within the cloud, which have direct access to the sequencing datasets.

For the field of genome informatics, a key feature of cloud computing is the ability of service providers and their customers to store large datasets in the cloud. These datasets typically take the form of virtual disk images that can be attached to virtual machines as local hard disks and/or shared as networked volumes. For example, the entire GenBank archive could be (and in fact is, see below) stored in the cloud as a disk image that can be loaded and unloaded as needed.

Figure 3 shows what the genome informatics ecosystem might look like in a cloud computing environment. Here, instead of there being separate copies of genome datasets stored at diverse locations and groups copying the data to their local machines in order to work with them, most datasets are stored in the cloud as virtual disks and databases. Web services that run on top of these datasets, including both the primary archives and the value-added integrators, run as virtual machines within the cloud. Casual users, who are accustomed to accessing the data via the web pages at NCBI, DDBJ, Ensembl or UCSC, continue to work with the data in their accustomed way; the fact that these servers are now located inside the cloud is invisible to them.

Power users can continue to download the data, but they now have an attractive alternative. Instead of moving the data to the compute cluster, they move the compute cluster to the data. Using the facilities provided by the

service provider, they configure a virtual machine image that contains the software they wish to run, launch as many copies as they need, mount the disks and databases containing the public datasets they need, and do the analysis. When the job is complete, their virtual cluster sends them the results and then vanishes until it is needed again.

Cloud computing also creates a new niche in the ecosystem for genome software developers to package their work in the form of virtual machines. For example, many genome annotation groups have developed pipelines for identifying and classifying genes and other functional elements. Although many of these pipelines are open source, packaging and distributing them for use by other groups has been challenging given their many software dependencies and site-specific configuration options. In a cloud computing environment these pipelines can be packaged into virtual machine images and stored in a way that lets anyone copy them, run them and customize them for their own needs, thus avoiding the software installation and configuration complexities.

But will it work?

Cloud computing is real. The earliest service provider to realize a practical cloud computing environment was Amazon, with its Elastic Cloud Computing (EC2) service [37] introduced in 2005. It supports a variety of Linux and Windows virtual machines, a virtual storage system, and mechanisms for managing internet protocol (IP) addresses. Amazon also provides a virtual private network service that allows organizations with their own compute resources to extend their local area network into Amazon's cloud to create what is sometimes called a 'hybrid' cloud. Other service providers, notably Rackspace Cloud [38] and Flexiant [39], offer cloud services with similar overall functionality but many distinguishing differences of detail.

As of today, you can establish an account with Amazon Web Services or one of the other commercial vendors, launch a virtual machine instance from a wide variety of generic and bioinformatics-oriented images and attach any one of several large public genome-oriented datasets. For virtual machine images, you can choose images prepopulated with Galaxy [40], a powerful web-based system for performing many common genome analysis tasks, Bioconductor [41], a programming environment that is integrated with the R statistics package [42], GBrowse [43], a genome browser, BioPerl [44], a comprehensive set of bioinformatics modules written in the Perl programming language, JCVI Cloud BioLinux [45], a collection of bioinformatics tools including the Celera Assembler, and a variety of others. Several images that run specialized instances of the UCSC Genome Browser are under development [46].

In addition to these useful images, Amazon provides several large genomic datasets in its cloud. These include a complete copy of GenBank (200 gigabytes), the 30X coverage sequencing reads of a trio of individuals from the 1000 Genomes Project (700 gigabytes) and the genome databases from Ensembl, which includes the annotated genomes of human and 50 other species (150 gigabytes of annotations plus 100 gigabytes of sequence). These datasets were contributed to Amazon's repository of public datasets by a variety of institutions and can be attached to virtual machine images for a nominal fee.

There are also a growing number of academic compute cloud projects based on open source cloud management software, such as Eucalyptus [47]. One such project is the Open Cloud Consortium [48], with participants from a group of American universities and industrial partners; another is the Cloud Computing University Initiative, an effort initiated by IBM and Google in partnership with a series of academic institutions [49], and supplemented by grants from the US National Science Foundation [50], for use by themselves and the community. Academic clouds may in fact be a better long-term solution for genome informatics than using a commercial system, because genome computing has requirements for high data read and write speeds that are quite different from typical business applications. Academic clouds will likely be able to tune their performance characteristics to the needs of scientific computing.

The economics of cloud computing

Is this change in the ecosystem really going to happen? There are some significant downsides to moving genomics into the cloud. An important one is the cost of migrating existing systems into an environment that is unlike what exists today. Both the genome databases and the value-added integrators will need to make significant changes in their standard operating procedures and their funding models as capital expenditures are shifted into recurrent costs; genomics power users will also need to adjust to the new paradigm.

Another issue that needs to be dealt with is how to handle potentially identifiable genetic data, such as that produced by whole genome association studies or disease sequencing projects. These data are currently stored in restricted-access databases. In order to move such datasets into a public cloud operated by Amazon or another service provider, they will have to be encrypted before entering the cloud and a layer of software developed that allows authorized users access to them. Such a system would be covered by a variety of privacy regulations and would take time to get right at both the technological and the legal level.

Then there is the money question. Does cloud computing make economic sense for genomics? It is difficult to

make blanket conclusions about the relative costs of renting versus buying computational services, but a good discussion of the issues can be found in a technical report on Cloud Computing published about a year ago by the UC Berkeley Reliable Adaptive Distributed Systems Laboratory [51]. The conclusion of this report is that when all the costs of running a data center are factored in, including hardware depreciation, electricity, cooling, network connectivity, service contracts and administrator salaries, the cost of renting a data center from Amazon is marginally more expensive than buying one. However, when the flexibility of the cloud to support a virtual data center that shrinks and grows as needed is factored in, the economics start to look downright good.

For genomics, the biggest obstacle to moving to the cloud may well be network bandwidth. A typical research institution will have network bandwidth of about a gigabit/second (roughly 125 megabytes/second). On a good day this will support sustained transfer rates of 5 to 10 megabytes/second across the internet. Transferring a 100 gigabyte next-generation sequencing data file across such a link will take about a week in the best case. A 10 gigabit/second connection (1.25 gigabytes/second), which is typical for major universities and some of the larger research institutions, reduces the transfer time to under a day, but only at the cost of hogging much of the institution's bandwidth. Clearly cloud services will not be used for production sequencing any time soon. If cloud computing is to work for genomics, the service providers will have to offer some flexibility in how large datasets get into the system. For instance, they could accept external disks shipped by mail the way that the Protein Database [52] once accepted atomic structure submissions on tape and floppy disk. In fact, a now-defunct Google initiative called Google Research Datasets once planned to collect large scientific datasets by shipping around 3-terabyte disk arrays [53].

The reversal of the advantage that Moore's Law has had over sequencing costs will have long-term consequences for the field of genome informatics. In my opinion the most likely outcome is to turn the current genome analysis paradigm on its head and force the software to come to the data rather than the other way around. Cloud computing is an attractive technology at this critical juncture.

Acknowledgements

I thank Mark Gerstein, Dan Stanzione, Robert Grossman, John McPherson, Kamran Shazand and David Sutton for helpful discussions during the research and preparation of this article.

Published: 5 May 2010

References

1. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DJ: **GenBank.** *Nucleic Acids Res* 2005, **33:**D34-D38.
2. Brooksbank C, Cameron G, Thornton J: **The European Bioinformatics Institute's data resources.** *Nucleic Acids Res* 2010, **38:**D17-D25.
3. Sugawara H, Ogasawara O, Okubo K, Gojobori T, Tateno Y: **DDBJ with new system and face.** *Nucleic Acids Res* 2008, **36:**D22-24.
4. Shumway M, Cochrane G, Sugawara H: **Archiving next generation sequencing data.** *Nucleic Acids Res* 2010, **38:**D870-D871.
5. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Edgar R: **NCBI GEO: archive for high-throughput functional genomic data.** *Nucleic Acids Res* 2009, **37:**D885-D890.
6. Kapushesky M, Emam I, Holloway E, Kurnosov P, Zorin A, Malone J, Rustici G, Williams E, Parkinson H, Brazma A: **Gene expression atlas at the European bioinformatics institute.** *Nucleic Acids Res* 2010, **38:**D690-D698.
7. Flicek P, Aken BL, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Gräf S, Haider S, Hammond M, Howe K, Jenkinson A, Johnson N, Kähäri A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Koscielny G, Kulesha E, Lawson D, Longden I, Massingham T, McLaren W, *et al.*: **Ensembl's 10th year.** *Nucleic Acids Res* 2010, **38:**D557-D662.
8. Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, Pohl A, Pheasant M, Meyer LR, Learned K, Hsu F, Hillman-Jackson J, Harte RA, Giardine B, Dreszer TR, Clawson H, Barber GP, Haussler D, Kent WJ: **The UCSC Genome Browser database: update 2010.** *Nucleic Acids Res* 2010, **38:**D613-D619.
9. Taylor J, Schenck I, Blankenberg D, Nekrutenko A: **Using Galaxy to perform large-scale interactive data analyses.** *Curr Protoc Bioinformatics* 2007, **10:**10.5.
10. Engel SR, Balakrishnan R, Binkley G, Christie KR, Costanzo MC, Dwight SS, Fisk DG, Hirschman JE, Hitz BC, Hong EL, Krieger CJ, Livstone MS, Miyasato SR, Nash R, Oughtred R, Park J, Skrzypek MS, Weng S, Wong ED, Dolinski K, Botstein D, Cherry JM: ***Saccharomyces*** **Genome Database provides mutant phenotype data.** *Nucleic Acids Res* 2010, **38:**D433-D436.
11. Harris TW, Antoshechkin I, Bieri T, Blasiar D, Chan J, Chen WJ, De La Cruz N, Davis P, Duesbury M, Fang R, Fernandes J, Han M, Kishore R, Lee R, Müller HM, Nakamura C, Ozersky P, Petcherski A, Rangarajan A, Rogers A, Schindelman G, Schwarz EM, Tuli MA, Van Auken K, Wang D, Wang X, Williams G, Yook K, Durbin R, Stein LD, Spieth J, Sternberg PW: **WormBase: a comprehensive resource for nematode research.** *Nucleic Acids Res* 2010, **38:**D463-D467.
12. Fey P, Gaudet P, Curk T, Zupan B, Just EM, Basu S, Merchant SN, Bushmanova YA, Shaulsky G, Kibbe WA, Chisholm RL: **dictyBase - a** ***Dictyostelium*** **bioinformatics resource update.** *Nucleic Acids Res* 2009, **37:**D515-D519.
13. Liang C, Jaiswal P, Hebbard C, Avraham S, Buckler ES, Casstevens T, Hurwitz B, McCouch S, Ni J, Pujar A, Ravenscroft D, Ren L, Spooner W, Tecle I, Thomason J, Tung CW, Wei X, Yap I, Youens-Clark K, Ware D, Stein L: **Gramene: a growing plant comparative genomics resource.** *Nucleic Acids Res* 2008, **36:**D947-D953.
14. Moore GE: **Cramming more components onto integrated circuits.** *Electronics* 1965, **38:**4-7.
15. Walter C: **Kryder's Law.** *Sci Am* August 2005, **293:**32-33.
16. Tehrani R: **As we may communicate.** TMCNet, 2000. [http://www.tmcnet.com/articles/comsol/0100/0100pubout.htm]
17. **Internet Archive** [http://www.archive.org/]
18. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, *et al.*: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437:**376-380.
19. Bennett S: **Solexa Ltd.** *Pharmacogenomics* 2004, **5:**433-438.
20. McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC, Zhang Z, Ranade SS, Dimalanta ET, Hyland FC, Sokolsky TD, Zhang L, Sheridan A, Fu H, Hendrickson CL, Li B, Kotler L, Stuart JR, Malek JA, Manning JM, Antipova AA, Perez DS, Moore MP, Hayashibara KC, Lyons MR, Beaudoin RE, *et al.*: **Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding.** *Genome Res* 2009, **19:**1527-1541.
21. **Illumina** [http://www.illumina.com/]
22. **Pacific Biosciences** [http://www.pacificbiosciences.com/]
23. **Helicos Biosciences Corporation** [http://www.helicosbio.com/]
24. **Ion Torrent** [http://www.iontorrent.com/]
25. **The 1000 Genomes Project** [http://www.1000genomes.org/]
26. ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A,

Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, *et al.*: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447:**799-816.

27. Celniker SE, Dillon LA, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, Micklem G, Piano F, Snyder M, Stein L, White KP, Waterston RH; modENCODE Consortium: **Unlocking the secrets of the genome.** *Nature* 2009, **459:**927-930.
28. Cancer Genome Atlas Research Network: **Comprehensive genomic characterization defines human glioblastoma genes and core pathways.** *Nature* 2008, **455:**1061-1068.
29. International Cancer Genome Consortium: **International network of cancer genome projects.** *Nature* 2010, **464:**993-998.
30. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI: **The human microbiome project.** *Nature* 2007, **449:**804-810.
31. **Human Microbiome Project** [http://nihroadmap.nih.gov/hmp/]
32. Johnson DS, Mortazavi A, Myers RM, Wold B: **Genome-wide mapping of in vivo protein-DNA interactions.** *Science* 2007, **316:**1497-1502.
33. El-Maarri O: **Methods: DNA methylation.** *Adv Exp Med Biol* 2003, **544:**197-204.
34. Li G, Fullwood MJ, Xu H, Mulawadi FH, Velkov S, Vega V, Ariyaratne PN, Mohamed YB, Ooi HS, Tennakoon C, Wei CL, Ruan Y, Sung WK: **ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing.** *Genome Biol* 2010, **11:**R22.
35. **VMware** [http://www.vmware.com/]
36. **KVM** [http://www.linux-kvm.org/page/Main_Page]
37. **Amazon Elastic Compute Cloud** [http://aws.amazon.com/ec2]
38. **The Rackspace Cloud** [http://www.rackspacecloud.com/]
39. **Flexiant** [http://www.flexiant.com/]
40. **Galaxy** [http://main.g2.bx.psu.edu/]
41. **Bioconductor** [http://www.bioconductor.org/]
42. **The R Project for Statistical Computing** [http://www.r-project.org/]
43. **GBrowse** [http://gmod.org/wiki/Gbrowse]
44. **Bioperl** [http://www.bioperl.org/wiki/Main_Page]
45. **JCVI Cloud BioLinux** [http://www.jcvi.org/cms/research/projects/jcvi-cloud-biolinux/overview/]
46. **Amazon Cloud Instance** [http://genomewiki.ucsc.edu/index.php/Amazon_Cloud_Instance]
47. **Eucalyptus** [http://open.eucalyptus.com/]
48. **Open Cloud Consortium** [http://opencloudconsortium.org/]
49. **Google and IBM Announce University Initiative to Address Internet-Scale Computing Challenges**. Press release 2007. [http://www.google.com/intl/en/press/pressrel/20071008_ibm_univ.html]
50. **National Science Foundation Awards Millions to Fourteen Universities for Cloud Computing Research** [http://www.nsf.gov/news/news_summ.jsp?cntn_id=114686]
51. Armbrust M, Fox A, Griffith R, Joseph AD, Katz RH, Konwinski A, Lee G, Patterson DA, Rabkin A, Stoica I, Zaharia M: **Above the clouds: a Berkeley view of cloud computing.** Technical Report No. UCB/EECS-2009-28**.** Electrical Engineering and Computer Sciences University of California at Berkeley [http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.pdf]
52. Kouranov A, Xie L, de la Cruz J, Chen L, Westbrook J, Bourne PE, Berman HM: **The RCSB PDB information portal for structural genomics.** *Nucleic Acids Res* 2006, **34:**D302-D305.
53. Madrigal A: **Google to host terabytes of open-source science data.** *Wired Science* 2008. [http://www.wired.com/wiredscience/2008/01/google-to-provi/]

doi:10.1186/gb-2010-11-5-207
Cite this article as: Stein LD: **The case for cloud computing in genome informatics.** *Genome Biology* 2010, **11:**207.