Eight Questions and Answers Defining Cloud Computing for Digital, Sequencing-Based Biological Research

Konstantinos Krampis*[1]agbiotec@gmail.com

20850, USA

Email:

*Corresponding author

Abstract

Background: Text for this section of the abstract. [**?**]

Results: Text for this section of the abstract . . .

Conclusions: Text for this section of the abstract . . .

Background

Next-Gen Sequencing, Cloud Computing, and The Digital Biology Research

Advances in high-throughput sequencing and the synthesis of genomes has defined ”The Digital Age of Biology” [**?**], where we have transitioned from the vision of Schrodinger that ”Life is code” [**?**] to applied processes that convert digital code into DNA that runs a living organism [**?**].

In recent years sequencing technologies continue to move in a direction where throughput per run is increasing while cost per basepair is decreasing (review in [1]). Several technologies available on the market today produce massive volumes of sequence data; for instance, one of the most widely used instruments in the field currently and for the past few years, Illumina’s GAIIx system can produce up to 95 Giga-base (Gb) of sequence per run [2] while the also broadly used SOLiD sequencer has yields of a similar range up to 90 Gb [3]. With the latest generation of instruments such as for example the HiSeq system, yield per run has reached 600 Gb [2], while the Pacific BioSciences sequencer yields 90 Gb in short amounts of time [**?**].

Recently, small-factor, benchtop sequencers became available all of which can be acquired at a fraction of the cost and be affordable for independent researchers running smaller laboratories. Examples in this category include GS Junior by 454, MiSeq by Illumina and Ion Proton by Life Technologies, providing sequencing capacity at 0.035Gb, 1Gb and 1.5Gb respectively for GS Junior, Ion Proton and MiSeq review in [**?**]) that is adequate for sequencing bacterial, small fungal or viral genomes. Taking this fact into account along with low cost per run (US $225 -$1100 depending on which benchtop sequencer is used and required throughput), sequencing has already started becoming a standard technique for basic biological research performed in smaller laboratories. Example applications of sequencing in biological research include Single Nucleotide Polymorphism (SNP) variation discovery , gene expression analysis (RNAseq) and DNA-protein interaction analysis (ChiPseq), (review in [4]).
The new generation of sequencing technologies is also being used in the area of metagenomics, for large-scale studies of uncultivated microbial communities. The J. Craig Venter Institute (JCVI) for example has been involved in several such metagenomic projects, including the Sorcerer II Global Ocean Sampling (GOS, [5]) expedition to study marine microbial diversity, and also the National Institutes of Health funded Human Microbiome Project to study human associated microbial communities [6].

While large datasets are generated sequencers, these instruments are typically bundled with only minimal computational and storage capacity for data capture during a sequencing run. For example, the un-assembled reads returned from a single lane of the Illumina GAIIx instrument after base calling are approximately 100 GigaByte (GB) in size. For small laboratories acquiring a sequencing instrument, the currently available online software tools are not and option for downstream sequence analysis, since they cannot provide the required compute capacity. The NCBI website for example [**?**], cannot accept input sequence data files of more than 0.5 GB size for BLAST sequence simijarity search. With this as an example, we see that scientific value cannot be obtained from investment in a sequencing instrument, unless it is accompanied by an almost equal or greater investment in informatics hardware infrastructure. Besides large capacity computing servers, also required are trained bioinformaticians competent to install, configure and use specific software to analyze the generated data, and store the data in appropriate formats for future use.

Due to the nature of next-generation sequencing data analysis, computationally intensive tasks of genome assembly or whole genome alignments that require extensive compute resources, alternate with tasks such as genome annotation that are less computationally demanding. This leads to sub-optimal utilization of computer hardware installed within bioinformatic data centers, while maintenance costs including electricity, cooling and informatics support personnel salaries stay at constant levels. For smaller academic institutions this can pose an impendiment for leveraging sequencing technology for research, as in addition to securing the funds for building a cluster with adequate capacity to handle large-scale genomic datasets, they also need to maintain a system that is not utilized at its full capacity most of the time. A second complication is related to the fact that databases with reference genomes [?] are constantly growing in size, and for most bioinformatic analysis they need to be downloaded to local storage systems, in order to be used as for example when conducting comparative genome annotations. As these databases grow larger the process becomes more time consuming, incurring in higher bandwidth and storage costs for replicating the data locally. Finally, building a data analysis infrastructure for next-generation sequencing also involves hiring trained bioinformatics engineers competent to install, configure and use specialized software tools and data analysis pipelines, which in the end can result to higher expenses than that of acquiring the computer hardware or maintaining the data center.

Results and Discussion

Eight Questions and Answers to Define Cloud Computing Applications for Digital, Sequencing-Based Biological Research

What Role can Cloud Computing play in The Digital Biology Research Era?

As an alternative to building local informatics infrastructure at their laboratories, researchers can rent computational and storage capacity from cloud services such as Amazon EC2 [21]. This can potentially be a better economic model for smaller research laboratories, since the cost for hardware and data center maintenance cannot be justified for only a few sequencing experiments. The Amazon EC2 Cloud service for example, employs a charge model similar to traditional utilities such as electricity, and users renting servers are billed based on the amount of computational capacity consumed on an hourly basis [?]. This Cloud service consists of thousands of computer servers with petabytes of storage, leveraging economies of scale to achieve low operational costs that in turn offers as savings to users. Furthermore, the Amazon Cloud has data centers in US East and West regions, European Union and Asia [?], providing researchers worldwide with the ability to tap into a large pool of computational resources, outside of institutional, economic or geographic boundaries. Overall, renting computational capacity from Cloud services has the potential to eliminate many of the upfront capital expenses for building information technology infrastructures for next generation sequencing, and result in transformation of the analysis and data processing tasks into well defined operational costs. A large capacity server with 64GB memory and 8 processor (CPU) cores for example, that would suffice for most types of bioinformatic analysis costs $2 US to rent per hour on Amazon EC2 [?], and researchers worldwide can rent if required a number of servers for large-scale data analysis.

The current approach of data warehousing in public databases such as the Sequence Read Archive (SRA, [?]) provides little value to researchers that lack access to computational resources or informatics expertise. Leveraging the value of these or other genomic datasets deposited on public databases, requires multiple steps including downloading the data, provisioning high-performance computer servers and large-scale data storage, in addition to compiling and installing specialized bioinformatics software. For scientists that do not have access to computing infrastructures, alternative approaches are required in order to be able to analyze sequencing datasets generated locally at their laboratories, and to add scientific value by integrating with genomic data from other studies available on public databases. A project that has demonstrated distributing genomic datasets in combination with computational capacity, and essentially placing data where the compute cycles are, is the 1000 Human genomes project [?]. In this example, sequence read data have been deposited on Amazon Cloud storage that is directly accessible by rented servers on the same Cloud, allowing analysis using pre-installed bioinformatics software from the Cloud BioLinux project [19], [?], [?].

Computational analysis can be a the major bottleneck for smaller laboratories and academic institutes transitioning experimental techniques used for basic biology research to sequencing-based methods, including for example RNAseq for gene expression, ChipSeq for protein interactions, metagenomics for microbial population analysis. Furthermore, as sequencing capacities follow an uptrend that surpasses that of Moore's Law [?] while the cost per base pair follows an inverted trend, the genomics community can easily undertake projects at similar scales to that of the 1000 human genomes project. By placing such datasets on publicly accessible Cloud compute platforms, their value for other researchers in the community is immediatelly increased, reducing the number of required steps for scientific discovery to simply uploading their data on the Cloud storage and running analysis pipelines using rented servers with pre-configured software from the Cloud service.

Established research institutes that perform sequencing of new genomes on a regular basis, and consecutively have well-funded bioinformatic cores to process the sequence data, have the resource and expertise to deliver to the community beyond the production of finished genomes at each center and their upload to publicly-accessible repositories such as NCBI []. This can be achieved by capturing and distributing the bioinformatics expertise developed during each sequencing project, by making publicly available on the Cloud and will contain all software and customized data analysis pipelines developed for sequence analysis at each institute. Currently, many of these centers deposit developed software on open-source code repositories such as SourceForge [**?**] or GitHub [**?**]. One major bottleneck for community researchers leveraging these newly developed bioinformatics tools is the requirement to download and compile the code by first configuring all the dependencies including the type of operating system, code libraries and computer hardware used at the bioinformatics core where the software was originally developed. This creates an obstacle for laboratories lacking the required informatics expertise and computational infrastructure, that can be avoided by utilizing servers containing pre-configured bioinformatics tools. This approach in turn, can democratize access to computational resources requeired for high throughput sequencing data analysis and allow further adoption of sequencing technology for basic biology research. Specialized compute servers with bioinformatics tools and data pipelines available on the Cloud through the Amazon EC2 service for example, can provide publicly accessible, high performance data analysis platform for use by research groups acquiring sequencing capability, that can be rented on-demand at low cost, removing the need for implementing informatics infrastructure at each laboratory.

Are Cloud-Based Bioinformatics Software Suites Available on the Cloud ?

A number of systems for deploying bioinformatics tools through community accessible web portals have been developed during the past decade, including the Biology Workbench [7], PISE [8], wEMBOSS [9], Mobyle [10] BioManager [11] and BioExtract [12]. Some of those systems are not actively developed anymore; most require significant software development effort and system customization in order to deploy the tools through the web interface; with the exception of Mobyle, none of these portals provides users with an intuitive option to create and edit complex data analysis workflows; users have the option to download the source code for each system, but need to provision the hardware and software engineering expertise to install the portal on a compute server; management and sharing of datasets among users is difficult, while available storage space is limited by the capacity of the server where the portal is installed; finally, these portals do not leverage the Cloud's scalability but rather run on computational hardware with fixed capacity, that poses a limitation on the dataset size that can be processed.

On the other hand, web portals that can analyze large-scale bioinformatic datasets, have been developed by well-funded institutions including IMG/M [13], CAMERA [14], EBI [15] and MG-RAST [16]. While these portals are backed by considerable compute resources and data storage, due to their centralized approach they cannot possibly support the increasing numbers of laboratories that purchase benchtop sequencers and produce genomic datasets. Furthermore, researchers are required to go through an application process in order to get access for uploading their data and have allocated computational capacity for data analysis within their account. Another major drawback is that the software for most of these sites is not open-source, while researchers often have to perform multiple submitions of their datasets to all the different portals, since each offers a different sequence data analysis pipeline.

An alternative to these centralized services are Cloud-based, scalable data analysis portals such as Galaxy [17], CloVR [18], Cloud BioLinux [19] and BioKepler [20] that are open-source and accessible to any laboratory or research group through the Amazon EC2 computer Cloud [21]. The Galaxy bioinformatics workbench, includes a range of tools, from scripts that extract entries from sequence files, to software for processing next-generation sequence data. Galaxy is a self-contained platform including a web portal software stack and a set of bioinformatics tools with graphical user interfaces, in addition to an intuitive, drag and drop canvas for composing complex data analysis workflows. Galaxy was designed as a framework for easy deployment through a web portal of command-line only software that lacks a user interface, and for that purpose provides a standardized method to deploy through the web portal command-line only bioinformatics tools with only minimal expertise for editing simple configuration files describing the interface [?]. Furthermore, it allows to leverage computational capacity beyond a single compute server through the Galaxy-Cloudman [22] framework that provides compute clusters for parallel data processing on Cloud services such as Amazon EC2.

Another community-centered, public access offering for computing on the Cloud is through our own Cloud Biolinux [19], [?], that provides on-demand bioinformatics computing and a set of pre-configured sequence analysis tools within a high-performance Virtual Machine (VM) server that runs on Cloud computing services such as Amazon EC2. The project is targeted to researchers that without access to large-scale informatics infrastructures for sequencing data analysis, but can rent instead computational capacity from Cloud services. Users can access the tools by starting the Cloud BioLinux VM through the Amazon cloud console web page [?], and easily perform large-scale data analysis as we have demonstrated with the 1000 Human genomes data [?], [?], [?]. An intuitive [?] our VM with the pre-installed tools on the Amazon EC2

Cloud, they will only need to follow four simple steps through their web browser: visit the Amazon Cloud website and create a new account, start the VM execution wizard through the Cloud's control console [58], choose computational capacity for the VM (memory, processor, cores, storage capacity), and specify username and password credentials for accessing the running VM. Each running VM receives a unique web address, and by using their web browser to access the address, a researcher can login to the portal interface with the assembly tools. These four steps are described in detail in our Cloud BioLinux publication and the project's documentation [59].

The Cloud BioLinux VM is open-source, can be downloaded and modified, while advanced users can install and run it on a private instance of the Eucalyptus [?] or Openstack [?] Cloud platforms. A diverse community of researchers from both the US (Massachusetts General Hospital, Harvard School of Public Health, Emory University) and Europe (National Environmental Research Center, King's College London, Denmark Technical University, Netherlands Wageninen University) has been already established around the project [?]. Finally, we have recently expanded Cloud BioLinux by adding support for advanced users through a developer's framework for building and distributing customized bioinformatics VMs, providing a toolkit for development of Cloud-based bioinformatics data analysis solutions. The framework includes a software management system that automates building a VM with a set of bioinformatics tools specified by the user and seamlessly deploys it across different Cloud platforms. The framework is freely available from the GitHub code repository [?]. The overall goal is to offer a platform for maintaining a range of specialized VM setups for serving different computing needs within the bioinformatics community, and allow researchers to focus on the next challenges of providing data, documentation, and the development of scalable analysis pipelines.

The solutions presented above provide public access to scalable sequence analysis for the genomic community, and users can get access to pre-configured software and on-demand computing platforms using Cloud infrastructures. While this is a great solution for smaller laboratories that lack informatics resources for sequencing data analysis, these cloud solutions simply offer within VM servers bioinformatics applications with monolithic designs that process data serially, and are not designed to leverage specific characteristics of Cloud computing platforms such as highly parallelism and distributed computing. On the other hand, specialized, high-performance bioinformatics applications and data pipelines that have been implemented by bioinformatics core teams at large institutions, are usually coupled with specific cluster computing hardware and data storage infrastructure at each institution, requiring extended effort to refactor the code and run data analysis pipelines on the Cloud [23].

Unique characteristics of Cloud Computing versus traditional Bioinformatics Infrastructures

The cloud offers two great advantages:

CHEAP, NO GEO BOUNDARIES DATA STORAGE:

Data storage using a Cloud service has the advantage that large-scale sequencing datasets can be easily exchanged among collaborators worldwide. Inherent in the design of Amazon S3 (http://aws.amazon.com/s3) service is replication of data across several physical storage locations for disaster prevention, available currently on US East and West regions, European Union (Ireland) and Asia Pacific (Singapore). For the data upload a researcher can choose the closest region for minimizing data transfer latency over the internet. Following that, the Cloud service automatically replicates the data to different locations as part of the disaster prevention policy, allowing collaborating researchers to download the data from their closest region.

Currently Amazon S3 has offers a community program (http://aws.amazon.com/datasets) to host a variety of widely used public datasets at no charge for researchers. Bioinformatics-related datasets hosted for free come from the 1000 human genomes, the complete Genbank, Ensembl and Unigene databases, in addition to the Ensembl human genome annotation data. public data Researchers can then access, copy, modify and perform computation on these data volumes directly using pre-configured VMs on Amazon EC2 instances such as JCVI's Cloud Biolinux, and just pay for the compute and additional storage resources they use. We are in negotiations with Amazon to get support for hosting the data from this project as part of this program.

Stein LD. (2010) The case for Cloud computing in genome informatics. Genome biology, 2010, 11, 207+

Amazon EC2 Cloud pricing : http://aws.amazon.com/ec2/pricing

Amazon EC2 Cloud console : https://console.aws.amazon.com

Cloud BioLinux project documentation : http://tinyurl.com/Cloud-docu

AND SECOND, DATA HAVE NO VALUE UNLESS THEY RESIDE WHERE THE COMPUTE IS AND THE CLOUD OFFERS VM SERVERS.

The enabling technology towards is direction is virtualization, which allows data analysis pipelines, databases, website portals and all software dependencies including entire operating systems, code libraries and configuration files to be encapsulated in Virtual Machines (VMs). A Virtual Machine is essentially an emulation of a compute server, albeit a full-featured one, with virtual processors, memory and storage capacity, in the form of a single binary file. Cloud services such as Amazon web-services (http://aws.amazon.com) offer high-performance computer hardware with a virtualization layer, upon which a user executes VMs. Since the VMs are full-featured compute servers in the form of a single binary file, researchers with access to local computing clusters have also the option to download and run the VM servers and therefore instantiate a local version of the pipelines on a private Cloud. For example, a researcher can install the Eucalyptus or OpenStack Cloud (http://open.eucalyptus.com, http://www.openstack.org) that are an open-source replicas of the Amazon Cloud on their clusters, while also foregoing the need to perform any software installation since the VMs contain pre-installed all required dependencies for running the pipelines.

Meanwhile advances in cyber-infrastructure and information technology are changing the landscape of biological computing. Virtualization technologies allow entire compute servers including the operating system replete with all the necessary software packages for data analysis, to be archived within a Virtual Machine (VM). A VM is an emulation of a compute server, with virtual processors, memory and storage capacity, in the form of a single binary file that can be executed independently of the hardware architecture available [3]. Since all software components and dependencies are encapsulated within the VM, it is possible to distribute pre-installed, ready-to-execute bioinformatics tools and data analysis pipelines, in the format of a single binary, down-loadable VM file. This addresses one of the main hurdles to make open-source software with complex dependencies and installation procedures widely accessible to the research community. Virtualization technologies led to the development of Cloud computing services, where remote computer server farms can be rented on an hourly basis by researchers and used for scalable, on-demand computation. Cloud services offer high-performance computer hardware with virtualization, upon which a user executes VM servers [3].

In our experience, there is no ideal solution for being able to provide long-term support and maintaining hardware servers that provide public access to bioinformatics software, especially once the funding cycle of a project runs out. By using VMs however, we can preserve an exact replica of the AIP portal and databases in its precise state at the time when the VM was generated. This allows us to archive the system, and users have the ability to re-instantiate AIP to a fully functional state at any time in the future by installing a local Cloud or leasing computing time on the Amazon Cloud. Our approach offers a way to keep the system readily accessible with minimal cost (only hosting the VMs on an FTP site, which can be long-term provided by JCVIs IT for free) past the funding cycle, while researchers who receive funding for their own projects can either allocate on their budget compute costs on the Amazon Cloud or budget in their grant hardware for a local Cloud. Following the concept of Whole System Snapshot Exhange (WSSE, Dudley and Butte 2010) we will create Virtual Machines (VMs) that contain replicas of the bioinformatics systems used for data analysis in this study. The VMs will be made publicly available on the Amazon EC2 Cloud computing platform, for the purpose of fulfilling two goals: first, to allow reproducibility of the bioinformatic analysis performed in this project through placing the data along with pre-configured software, on a Cloud compute platform accessible by the worldwide genomics community outside of institutional, economic or national boundaries. Given the complexity of the current project, following publication of our results we expect for example researchers to require re-running part of the bioinformatic analysis for adjusting the algorithmic parameters. Second, as bioinformatics computing is a key aspect for researchers in the community to extract value from the data released from this project, while also build additional value on top as similar studies are performed or by integrating with results from existing studies, placing our data on a publicly accessible compute platform will facilitate this process. Besides the current study, we believe that this approach will demonstrate a new approach for sharing research results by distributing the data together with the computational capacity, which will provide an option for researchers from smaller institutions without access to extensive infrastructure for their data analysis needs.

For researchers that would like to leverage the advantages of a VM with the pre-installed assembly portal for working with the completed assemblies but consider the public Cloud as not secure option, we will offer the alternative of returning to them by mail an external hard drive with a VM containing with the portal and assembly data. Users will then be able to load and execute the VM on a local computer cluster with a Eucalyptus/OpenStack Cloud or on a PC using Virtualbox. We are currently offering a similar solution with Cloud BioLinux [3], where the project's VM is available for download and execution on a local Cloud or a PC from our website [4]. Upon local execution of a VM users will simply need to point their browser to the portal's Internet or local IP address [55] assigned automatically by either the Cloud or Virtualbox (Fig.1B). The IP address is available through each Cloud platform's or the Virtualbox software administrative interface, and we will provide extensive documentation (see subsequent paragraph) on uploading, running and accessing a local VM on the different platforms by extending the available Cloud BioLinux project documentation.

Available Options to Buying Software as A Service on the Cloud ?

During the past year, both public and commercial offerings pre-configured sequence analysis applications on the Cloud have become available. On the commercial side, DNAnexus (online ref. 3) currently includes tools for ChiPseq, RNAseq, 3'-end sequencing for expression quantification (3SEQ) and enzyme restriction analysis. DNAnexus runs on the Amazon Elastic Compute Cloud (EC2, online ref. 4), which provides on-demand virtual servers with various compute capacities. Another offering is iNquiry, which is a port to the Cloud of the bioinformatics software suite that used to be bundled together with the computer clusters built by the BioTeam (online ref. 5), but is no longer maintained. This platform is essentially a web-server for pre-installed open-source tools such as EMBOSS, HMMER, BLAST and the R statistical package, also on the EC2 Cloud platform. And many consulting firms such as cycle computing and BioTeam provide custom solutions at high cost.

Can Non-Computationally Savvy Researchers Easily Access the Cloud?

Text for this sub-section. More results ...

VM servers with the pre-configured pipelines and data will be publicly available for download. Researchers will have the option to execute them on a desktop computer, using virtualization software such as VirtualBox (http://www.virtualbox.org). VirtualBox is free and can be installed with a single step on Windows, Mac or Linux desktop computers. Alternatively, research teams with informatics expertise and access to a local cluster could choose to download our VM servers and perform large-scale data analysis by running them on a private Cloud installation, after installing Eucalyptus or OpenStack and converting part of the cluster to a local Cloud.

What the Public, Private, Open-Source or Commercial Clouds Available to Biologists Today ?

ur system will not be limited only to the Amazon EC2 Cloud platform, since researchers that have access to a local cluster at their home institution will have the option to download the VM and run, without being required to perform any software installation. The only dependency will be a virtualization layer that can run the VM on the cluster such as for example the OpenStack open-source Cloud (http://www.openstack.org). OpenStack is available as part of widely used Ubuntu Linux (http://www.ubuntu.com/Cloud) and included by default on a compute cluster set-up to run this operating system, while it can be easily installed as a package on clusters running other Linux versions. Alternatively, for researchers that do not have access to local compute Clouds, OpenStack installations can be accessed through the government-funded Argonne National Lab Magellan Cloud http://www.alcf.anl.gov/magellan) that provides compute allocations to researchers, in addition to a number of academic computing centers in both the US and abroad (http://openstack.org/user-stories/), or commercial Cloud providers such as RackSpace.

Which Factors Challenge Adoption of Cloud-Based Solutions for Bioinformatics

Another important concern for Cloud-based bioinformatic tools is related to the data transfer bottleneck from the local sequencing machines to the Cloud servers. According to the data published for the Amazon Cloud platform (online ref.11), 600GB of data would require approximately one week to upload on to the remote Cloud servers, when using an average broadband connection of 10Mbps. With a faster T3 connection which is usually easily obtainable even at small research institutions, within one week 2TB of data can be uploaded or approximately 600GB in 2 days. Solutions addressing this issue are available both as software that maximizes data transfer over the network compared to traditional File Transfer Protocol (FTP), or physical disk drive import/export services offered by the Cloud provider to its customers. Aspera's server (online ref. 12) has been recently integrated to NCBI's infrastructure, and researchers can download a free client that allows increased upload speeds to the Short Read Archive (online ref. 13). Through the Aspera software, transfer bandwidth between NCBI and the European Bioinformatics Institute for data sharing in the 1000 Genomes Project, has been increased from 20Mbps to 1000Mbps (see online ref. 14).

Finally, the Amazon offers the option for its users to physically ship disk drives to the company's offices and have the data copied to their servers (online ref. 11). With only 80 import cost for disk drives up to 4TB of data (4000GB), this is the most efficient method if we take into account the charge by Amazon for 0.10 per GB of bandwith consumed, which would add up to 60 for a 600GB data upload. In addition to that cost, the expense for obtaining a high-bandwidth internet connection for the data upload should be taken into account. We expect the Microsoft Azure Cloud platform to offer a similar service in the near future, given the requests on the Azure developer forums and the immediate consideration of the matter by Microsoft (online ref. 15).

Conclusions

Text for this section . . .

Methods

Methods sub-heading for this section

Text for this sub-section . . .

Another methods sub-heading for this section

Text for this sub-section . . .

Yet another sub-heading for this section

Text for this sub-section . . .

Authors contributions

Text for this section . . .

Acknowledgements

Text for this section . . .

References

1. Mason C, Elemento O: **Faster sequencers, larger datasets, new challenges**. *Genome Biology* 2012.
2. Illumina I: **Technical Specifications of Illumina Sequencers**.
3. Inc AB: **The SOLiD 5500 sequencing system.**
4. Mardis E: **Next-generation DNA sequencing methods**. *Annu. Rev. Genomics Hum. Genet.* 2008, **9**:387–402.
5. Rusch D, Halpern A, Sutton G, Heidelberg K, Williamson S, Yooseph S, Wu D, Eisen J, Hoffman J, Remington K: **The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific**. *PLoS biology* 2007, **5**(3):e77.
6. Nelson K, Weinstock G, Highlander S, Worley K, Creasy H, Wortman J, Rusch D, Mitreva M, Sodergren E, Chinwalla A: **A catalog of reference genomes from the human microbiome**. *Science (New York, NY)* 2010, **328**(5981):994.
7. Subramaniam S: **The biology workbenchA seamless database and analysis environment for the biologist**. *Proteins* 1998, **32**:1–2.
8. Letondal C: **A Web interface generator for molecular biology programs in Unix**. *Bioinformatics* 2001, **17**:73–82.
9. Sarachu M, Colet M: **wEMBOSS: a web interface for EMBOSS**. *Bioinformatics* 2005, **21**(4):540–541.
10. Neron B, and H Menager, Maufrais C, Joly N, Maupetit J, Letort S, Carrere S, Tuffery P, Letondal C: **Mobyle: a new full web bioinformatics framework Bioinformatics**. *Bioinformatics* 2009, **25**(22):3005–3011.
11. Cattley S, Arthur JW: **BioManager: the use of a bioinformatics web application as a teaching tool in undergraduate bioinformatics training**. *Briefings in Bioinformatics* 2007, **8**(6):457–465.
12. Lushbough CM, Bergman MK, Lawrence C, Jennewein D, Volker B: **Implementing bioinformatic workflows within the BioExtract Server**. *International Journal of Computational Biology and Drug Design* 2008, **1**(3):302–312.
13. Grigoriev I, Nordberg H, Shabalov I, Aerts A, Cantor M, Goodstein D, Kuo A, Minovitsky S, Nikitin R, Ohm R: **The Genome Portal of the Department of Energy Joint Genome Institute**. *Nucleic Acids Research* 2012, **40**(D1):D26–D32.
14. *Camera 2.0: A data-centric metagenomics community infrastructure driven by scientific workflows*, IEEE 2010.
15. Hunter C, Cochrane G, Apweiler R, Hunter S: **The EBI Metagenomics Archive, Integration and Analysis Resource**. *Handbook of Molecular Microbial Ecology I* 2011, :333–340.

16. Aziz R: **Subsystems-based servers for rapid annotation of genomes and metagenomes**. *BMC Bioinformatics* 2010, **11**(Suppl 4):O2.

17. Goecks J, Nekrutenko A, Taylor J: **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences**. *Genome Biol* 2010, **11**(8):R86.

18. Angiuoli S, Matalka M, Gussman A, Galens K, Vangala M, Riley D, Arze C, White J, White O, Fricke W: **CloVR: A virtual machine for automated and portable sequence analysis from the desktop using cloud computing**. *BMC Bioinformatics* 2011, **12**:356.

19. Krampis K, Booth T, Chapman B, Tiwari B, Bicak M, Field D, Nelson KE: **Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community**. *BMC Bioinformatics* 2012, **13**:42.

20. *Distributed workflow-driven analysis of large-scale biological data using biokepler*, ACM 2011.

21. Services AW: **Elastic Compute Cloud (EC2)**.

22. Afgan E, Baker D, Coraor N, Chapman B, Nekrutenko A, Taylor J: **Galaxy CloudMan: delivering cloud compute clusters**. *BMC Bioinformatics* 2010, **11**(Suppl 12):S4.

23. *Using clouds for metagenomics: A case study*, IEEE 2009.