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

Konstantinos Krampis*[1], Granger Sutton[1], Vivek Sarangi

[1] Informatics Department, J. Craig Venter Institute, 9704 Medical Center Dr., Rockville, MD 20850, USA

Email: Konstantinos Krampis*- agbiotec@gmail.com;

*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-Generation Sequencing, Computing, and Digital Biological Research

Advances in recent years in the areas of high-throughput sequencing and synthetic genomics has defined "The Digital Age of Biology" [1], where Schrodinger's vision of "Life is code" [2] is now implemented in technologies that can be used to convert digital code into DNA that runs a living organism [3].

The latest generation of sequencing technologies is also being used in the area of metagenomics, where large-scale studies of uncultivated microbial communities are performed. The J. Craig Venter Institute (JCVI) for example has been involved in several such metagenomic projects, including the Sorcerer II Global Ocean Sampling (GOS, [4]) expedition to study marine microbial diversity, and also the National Institutes of Health funded Human Microbiome Project to study human associated microbial

communities [5].

Sequencing technologies continue to move in a direction where throughput per run is increasing while cost per basepair is decreasing (review in [6]). For instance, one of the most widely used instruments in the field currently, Illumina's GAIIx system can produce up to 95 Giga-base (Gb) of sequence per run [7] while the ABI SOLiD sequencer has yields of a similar range [8]. With the latest generation of instruments such as for example the HiSeq system, yield has reached 600 Gb [7], while the Pacific BioSciences sequencer can produce 90 Gb in short amounts of time [9].

Small-factor, benchtop sequencers are also available which can be acquired at a fraction of the cost, making them affordable for independent researchers running small laboratories. Examples in this category include the 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 [10]). That level of throughput is adequate for sequencing bacterial, small fungal or viral genomes and along with the low cost per run (US $225 -$1100), sequencing has started to become a standard technique even in small laboratories. Example applications of sequencing for basic biological research include Single Nucleotide Polymorphism (SNP) variation discovery , gene expression analysis (RNAseq) and DNA-protein interaction analysis (ChiPseq), (review in [11]).

While sequencers generate datasets of significant size, they are typically bundled with only minimal computational and storage capacity for data capture during a run of the instrument. For example, the un-assembled reads returned from a single lane, single run of the Illumina GAIIx instrument after base calling are approximately 100 GigaByte (GB) in size. Therefore, for laboratories acquiring a sequencer, scientific value cannot be obtained from this investment, unless it is accompanied by an almost equal or greater expense for informatics hardware infrastructure. In addition, significant software engineering bioinformatic data analysis expertise is required [12], in order to go from sequence data to valuable information such as assembled and annotated genomes. This means that besides large capacity computing servers, trained personnel competent to install, configure and use specific software to analyze and store the generated data is required. Furthermore, public databases and software tools currently available online are not an option for downstream sequence data analysis, since BLAST [13] available from NCBI for

example [14], cannot accept input sequence files of more than 0.5 GB size for sequence similarity search.

An additional conundrum specific to bioinformatic analysis of sequencing data, is that computationally intensive tasks that require extensive compute resources such as genome assembly or whole genome alignments for example, are followed by genome annotation that is much less computationally demanding. This leads to sub-optimal utilization of computer hardware installed within data centers, while maintenance costs including electricity, cooling and salaries for informatics support personnel are at constant or even increasing levels. For smaller academic institutions this can pose a significant impediment in leveraging sequencing technology for research, as in addition to securing the funds for purchasing computing hardware with adequate capacity to handle large-scale genomic datasets, they also need to maintain informatics systems that are not utilized at its full capacity most of the time. A second complication is related to the fact that databases with reference genomes [15] are constantly growing in size, and for most bioinformatic analysis tasks they need to be downloaded to local storage systems when performing 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 implement specialized software tools and data analysis pipelines, which can incur higher costs than that of acquiring the computer hardware or maintaining a data center.

Results and Discussion

What Role can Cloud Computing play for Digital Biological Research ?

Computational data analysis can become a major bottleneck for smaller laboratories transitioning their experimental techniques to sequencing-based methods. Furthermore, sequencing instrument capacities follow an uptrend that surpasses that of Moore's Law [16] while the cost per base pair follows an inverted trend, resulting in the genomics community to constantly take on sequencing projects of increasing scale and scope. Currently, most federally-funded projects conclude with upload of the sequences or annotated genomes to NCBI [17] databases such as the Sequence Read Archive (SRA, [18]). This provides little value to researchers that do not have access to computational resources or informatics expertise, since multiple steps including data download, provision of high-performance computer servers and large-scale data storage, in addition to compiling and installing specialized bioinformatics software are required in order to

utilize these datasets.

As an alternative to investing in informatics infrastructure, researchers can rent computational and storage capacity from Cloud services such as Amazon EC2 [19]. This can potentially be a better economic model for smaller research laboratories, as the cost for hardware and data center maintenance cannot be justified for only a few sequencing experiments. The Amazon EC2 Cloud, employs a charge model similar to traditional utilities such as electricity and users are billed based on the amount of computational capacity consumed on an hourly basis [20]. This particular 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, Asia and Australia [21], 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 the Cloud 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.

The 1000 Human genomes project [22] has demonstrated a new approach for distributing genomic datasets in combination with computational capacity, by depositing sequence reads and reference genome mapping data on the Amazon Cloud storage. These datasets can be directly accessed by renting servers loaded with pre-installed bioinformatics software on the Cloud, and such examples have been provided by the Cloud BioLinux project [23], [24], [25]. Overall, by placing data from public projects on Cloud compute platforms, value for the community is immediately increased; no infrastructure other than a computer with internet access is required for a researcher to upload data generated at his or her laboratory , rent computational capacity and perform comparative analysis with data available on the Cloud storage.

On the other hand, large research institutes with in-house bioinformatic core facilities that complete sequencing of new genomes on a regular basis, also have the resources and expertise to deliver more to the community than just sequence reads and annotated genomes uploaded to publicly-accessible repositories [15]. Currently, while many of the centers deposit software on open-source code repositories such as SourceForge [26] or GitHub [27], an impediment for researchers trying to utilize the released software is the requirement to download, compile and configure all the dependencies including the type of operating system, software libraries, and computer hardware. This creates a bottleneck for laboratories

lacking the required informatics expertise or computational infrastructure, that could be alleviated if bioinformatic core facilities capture and distribute the bioinformatics expertise developed during each sequencing project on Cloud servers containing pre-configured bioinformatics tools, made publicly available along with the data on the Cloud. This approach, can democratize access to computational resources required 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, could provide publicly accessible, high performance data analysis platforms for use by research groups acquiring sequencing capability. These platforms could 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 that allow making bioinformatics tools accessible online through web portals have been developed during the past years, including the Biology Workbench [28], PISE [29], wEMBOSS [30], Mobyle [31] BioManager [32] and BioExtract [33]. Some of those systems are not actively developed anymore; most require significant software development effort and back-end system customization in order to deploy the tools through a web interface; with the exception of Mobyle, none of these portals provides users with a straightforward 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 set up their own instance of the portal; large-scale data management or collaborative data sharing and exchange among users is cumbersome, and available storage space is fized; finally, these portals do not leverage the Cloud's scalability but are rather limited by the capacity of the server where the portal is installed, posing a limitation on the dataset size that can be processed.

On the other hand, centralized web portals for large-scale data analysis of bioinformatic datasets have been developed by well-funded institutions, including IMG/M [34], CAMERA [35], EBI [36] and MG-RAST [37]. While these portals are backed by considerable compute resources and data storage, their centralized nature is what eventually becomes a bottleneck. First, for those of the centralized portals that provide access to their resources, researchers are required to apply for an account, in order to receive allocation of computational capacity for performing data analysis. Second, given the constantly increasing scale of genomic datasets, despite their capacity these portal cannot possibly support all small laboratories that

purchase low-cost, benchtop sequencers and generate sequence data. Finally, a major drawback is that the software for most of these sites is not open-source, while researchers often have to perform multiple submissions of their datasets, since each site offers a different sequence data analysis pipeline.

An alternative to these centralized services are Cloud-based, scalable bioinformatics data analysis systems such as Galaxy [38], CloVR [39], Cloud BioLinux [25] and BioKepler [40]. These systems are open-source and accessible to any laboratory or research group through Amazon EC2, but also are available for download an execution on private compute Clouds [19]. As an example, the Galaxy bioinformatics workbench includes a range of tools, from simple scripts that extract entries from sequence files, to complex algorithms for processing next-generation sequence data. Furthermore, Galaxy is a complete platform including a web portal software stack that provides the user interface for executing the bioinformatics tools, in addition to an intuitive, drag and drop canvas for composing workflows and data analysis pipelines with the available tools. Finally, it provides a standardized method for easy deployment on the portal of command-line only bioinformatics software, by editing simple configuration files to specify the interface design [41]. Through the Galaxy-Cloudman [42] framework, compute clusters for parallel data processing on Cloud services such as Amazon EC2 and private cloud can be instantiated.

Another community-centered, public access offering for computing on the Cloud is through our own work on Cloud Biolinux [25], [**?**]. This offering 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 a host of Cloud and virtualization platforms. The project is targeted to researchers that do not have access to large-scale informatics infrastructures for sequencing data analysis, but can instead rent on demand computational capacity from the Cloud. Users can access the tools by starting the Cloud BioLinux VM through the Amazon console web page [43], and easily perform large-scale data analysis as we have demonstrated for examplewith the 1000 Human genomes [23], [24], [**?**]. Furthermore, 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 [44] or Openstack [45] 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 [46]. Finally, we have recently expanded Cloud BioLinux by adding support for software developers

through a framework for building and distributing bioinformatics VMs, that essentially provides a toolkit for implementation of customized, 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, and is freely available from the GitHub code repository [47]. 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.

Some more tool suites on the Cloud that have become available recently, including the non-profit/ open-source GenoSpace by Broad Institute [48] and SeqWare [49], in addition to commercial offerings such as Illumina's BaseSpace [50], DNAnexus [51] and Nimbus Informatics [**?**]. GenomeSpace essentially integrates a set of tools and databases developed at the Broad Institute through a unified graphical interface for end-users, in addition to offering Application Programming Interfaces (API) for programmatic access by developers, backed by Cloud VMs and storage. Users can access the public GenomeSpace instance or create their local instance by retrieving the source code, in addition to accessing Virtual Machines (VM) with the complete system pre-installed and ready to execute with minimal configuration on the Amazon EC2 cloud [42].

SequeWare.... Similarly to GenomeSpace SeqWare provides source code and an Amazon Cloud VM with everything pre-installed and ready to execute with only minimal configuration. DNANexus 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.

The solutions presented above provide public access to scalable sequence data analysis resources for the genomic community, through which users can get access to pre-configured software and on-demand computing using Cloud infrastructures. Nonetheless, specialized, high-performance bioinformatics applications and data pipelines implemented by bioinformatics core teams at large institutions, are usually coupled with specific hardware and the informatics infrastructure at each institution. As a result, significant effort might be required to refactor data analysis pipelines to run at a different site from where they were originally developed or port them on the cloud [52]. Finally, while the current Cloud-based

solutions are great for smaller laboratories that lack informatics resources and in addition the VM servers provide enhanced portability across sites, they are simply a sophisticated container for bioinformatics software that in most cases has a monolithic design and does not leverage the distributed computing characteristics of the Cloud.

Unique characteristics of Cloud Computing versus traditional Bioinformatics Infrastructures

One of the building blocks of cloud computing technology is virtualization [53], that allows entire compute servers including the operating system and all the necessary software packages for data analysis to be encapsulated 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 executes independently of the underlying hardware architecture, on both Cloud and desktop computers. Cloud services such as Amazon EC2 [19] provide high-performance computer hardware with a virtualization layer, on top of which users run VM servers. Since all software components and dependencies are encapsulated within the VM, it is possible to distribute data analysis pipelines, databases, website portals, and all their required code libraries and configuration files in a ready to execute, compact and easy to download format. This approach can remove many of the technical roadblocks encountered when performing complex installations of open-source bioinformatics software, and consecutively make bioinformatics tools more accessible to the research community.

In our experience with development of bioinformatics projects, it is difficult to provide long-term software support or maintain web portals that provide online access to data analysis tools and databases, especially for projects funded by government grants that have an expiration date. Alternatively, by using Cloud VM servers to build and maintain a bioinformatics system and subsequently create Whole System Snapshots (WSSE, [25, 54]) of the VM servers, bioinformatics web portals and online databases that are build on the Cloud can be preserved in their precise state when the snapshot was created. A snapshot essentially is a compressed, exact replica of a VM server, capturing all of the software configuration, bioinformatics pipelines, input data and sequence assemblies, genome annotations and all other sequence data analysis results. A snapshot is an executable binary file as the original VM, and by using it as a template the virtualization layer of a Cloud platform can instantiate multiple replicas of the original VM server [55]. Finally, a researcher can set her snapshots to be publicly accessible or share them only with specific users within the same Cloud, therefore providing access for collaborators to both data and software in a ready to

execute and compact format.

Regarding costs, the Amazon EC2 Cloud [19] charges for VM snapshots $0.01 US per GigaByte (GB) of storage used per month. Such low costs can allow researchers to allocate a relatively small amount compared to their overall informatics budgets for sequencing data analysis, and maintain a VM server snapshot for a number of years past the end of a funding cycle. For projects involving data release online though a web portal that is usually decommissioned shortly after funding for a project ends, using VM snapshots for archiving the portal on a Cloud platform, enables other researchers to lease compute time on the Cloud, and create fully-functional instances of the original VM server from the snapshots. Therefore, using VM technology and the Cloud for building bioinformatics systems and then creating compressed snapshots to reduce costs for long-term storage, offers an economical and flexible solution throughout and past the life cycle of a research project.

Finally, use of virtualization and VM technology can provide two additional advantages: first, by depositing data and pre-configured software on a publicly accessible, Cloud-based VM, allows for reproducibility, provenance and openness of the bioinformatics research. For example, following publication of assembly and annotation results from a genome sequencing project, researchers in the community might require to re-run the assembly with additional data generated at their own laboratory or to change algorithmic parameters and fine tune outputs such as gene predictions. Furthermore, lowering the barrier to access high- performance informatics infrastructures required for working with next-generation sequencing datasets, is key for allowing researchers in the community to extract value from data released from publicly funded projects, while also to add value as similar studies take place. Second, by using the Cloud researchers have the capability to scale computational resources on-demand according to the amount of data generated from a sequencing project, through provision of the appropriate number of VM servers. With this approach usage of resources can be adjusted accordingly during the different analysis phases: while initially extensive computing resources will be necessary to perform assembly and annotation of the sequence data, computational resources (defined by the number of running VM servers), can be scaled down for less computationally demanding tasks such as visualization and browsing of the sequence annotations. After funding has ended, a lab can further lower the usage of the Cloud's computational resources and cut its informatics costs by archiving the VMs using snapshots [54].

What the Public, Private, Open-Source or Commercial Cloud Solutions Available to Biologists Today ?
Accessing Computational Cycles on The Cloud

Amazon Web Services (AWS, [56]) is one of the better established Cloud computing vendors, running on a similar infrastructure with the one that powers Amazon.com's e-commerce web portals. Cloud computing services offered by this vendor that are most applicable to bioinformatics, include the Amazon Elastic Compute Cloud (EC2, [19]), Elastic Block Store (EBS, [55]) and Simple Storage Service (S3, [57]), while a complete list of the available services can be found at [56]. These option provide respectively compute cycles through Virtual Machine (VM) servers, multiple virtual hard drives up to 1TeraByte (TB) that can attached to a running VM, and web-accessible data storage. None of these options is tied to any specific operating system or programming model and each comes with affordable pricing, as for example a large capacity VM server with 64GB memory and 8 processor (CPU) cores that would suffice for many different types of bioinformatic analysis costs, $2 US to rent per hour (for a complete price list see [20]). A large software developer community with a lot of expertise has formed during the past few years around the Amazon Web Services discussion forums [58], and based on our experience questions regarding the different services or requests for technical advice, are always answered in a day or less.

Using the Cloud for bioinformatic data analysis is not limited only to the Amazon EC2 platform, since researchers with access to a local computing cluster at their home institution have the option to run VM servers (and VM snapshots downloaded from Amazon) on a private Cloud, such as Eucalyptus [44] or OpenStack [45]. While OpenStack is the official Cloud of the Ubuntu Linux operating system [59] and is included by default on a compute cluster that runs this particular Linux flavor, it can also be installed on clusters running other Linux versions [60], and similarly for Eucalyptus [61]. These Cloud platforms are essentially open-source replicas of Amazon EC2 and offer identical Application Programming Interfaces (API), meaning that applications developed on Eucalyptus or OpenStack will work seamlessly on the Amazon Cloud and vice-versa. This allows to seamlessly transfer VM server snapshots such as for example the Cloud BioLinux VM [25] across installations of these Cloud platforms, providing researchers with ready-to-execute bioinformatics tools and data analysis pipelines pre-configured and installed on the VM.

While Amazon Web Services [56] was the first vendor to offer public Cloud computing access at a large scale, many alternative platforms became available during recent years including the Google App Engine [62], Microsoft Azure [63], GoGrid [64], FlexiScale [65] and the IBM SmartCloud [66].Multiple technical differences exist between these offerings, but they can be categorized into two broad groups: the

first includes Cloud platforms that provide users with access to Virtual Machines (VM) that are no different than standard Unix servers, available by Amazon, GoGrid, FlexiScale and IBM. On the other hand, Clouds that users access as abstracted computational resources that run software, without need for logging into Unix filesystems or provisioning VMs, are found on the Google and Microsoft offerings. This approach might sound as a better solution for smaller research groups with limited software engineering and informatics expertise, but is not without drawbacks. Specifically, the seamless execution of code and automatic scalability stems from the fact that the Google and Microsoft Clouds require software developers to implement their applications using programming frameworks based on Python [67] and .NET [68] respectively. While this might be an option for the development of new software or web-based applications that are most suitable for these frameworks, the majority of existing bioinformatics software is designed to run on standard Unix servers and filesystems. In this case, Clouds that provide direct access to VMs are a better choice, while additionally a framework-specific Cloud implementation despite its advantages could eventually lead to vendor lock-in.

Overall, Cloud platforms that provide access to VM servers can be a better choice for deploying existing software, in order to make it accessible to groups that do not have local informatics infrastructure but could instead rent compute time from the Cloud. In addition, for groups that have a longer-term vision for building Cloud-based bioinformatics infrastructures outside of traditional in-house data centers, portability across Cloud platforms should be a top criterion. While standalone software can runs on VMs that can be easily converted and ported across public, private Clouds that provide a virtualization layer [25], Cloud informatics infrastructures are essentially encoded in scripts that setup VM-based compute clusters and storage by issuing directives to the Cloud's API ([42], [69]). Therefore, for achieving portability of the complete infrastructure across different Cloud platforms with minimal software re-engineering effort, compatible and inter-operable APIs across the platforms is a requirement. Currently, this is only fulfilled amongst the open-source Eucalyptus / OpenStack Clouds, and the commercial Amazon Cloud. This fact should be carefully considered before choosing a Cloud provider for longer-term, especially if instances of the infrastructure is might be required to exist on both a private and a publicly-accessible, commercial Cloud.

For researchers who do not have access to a compute cluster, neither have the available funds to lease computing time from Amazon, the government-funded Magellan Cloud [70] provides an OpenStack cluster

where researchers can apply for a user account. In addition, a number of academic computing centers in both the US and other countries have similar clusters with OpenStack installed [71], where scientistics could access a Cloud platform. Finally, an option for users is to to run the VM servers on a desktop computer, using virtualization software such as VirtualBox [72], that is also open-source and can be installed in a single step on Windows, Mac or Linux computers. The Cloud BioLinux project for example, provides VM server snapshots that run on both private clouds and VirtualBox [73].

Finally, professors that teach courses on Cloud computing, can apply for the Amazon Web Services educational grants [74], that provide free computing and data storage resources for educational use. Furthermore, the Amazon Cloud has established a program [75] that hosts free of charge a variety of large-scale public datasets that have significant value for the scientific community. For example, genomic datasets available through this program include the 1000 human genomes data ([**?**], [76]) the NCBI flu genomes [77] and Ensembl human genome annotation database [78]. Users of the Amazon Cloud can access, copy, and perform computation on the data using VM servers, and just pay for the compute and storage resources they lease.

Accessing Storage Capacity on The Cloud

Cloud data storage services provide the advantage of data centers distributed across the globe, and therefore can make available connection endpoints in various geographic regions that reduce network transfer latency for data exchange among remotely located research sites. The Amazon S3 storage for example [57] has data centers located on the US East and West coast, European Union (Ireland), South America (Brazil), and Asia-Pacific (Japan, Singapore and Australia). A researcher working with large-scale sequencing datasets can choose to upload data to their closest data center, then initiate replication across the different regions through the S3 CloudFront service [79], and subsequently allow collaborators worldwide to retrieve the data from the nearest location. Furthermore, data stored using this service are protected from physical disasters, since they replicated by default across three different regions. Note that triple-replication is not synonymous with backup that prevents accidental deletion by the user, but instead refers to protection from permanent loss of a single data center. The S3 storage model deviates from that of POSIX-compliant hard drives [**?**], but instead uses data objects that have unique Uniform Resource Location (URL) identifiers across the Amazon Cloud and are also web browser-accessible. In detail, S3 is organized using a two-level namespace with top-level folders called "buckets", and while each Amazon

account may have up to 100 buckets, a bucket can store an unlimited number of data objects. A file on S3 with open access permissions can be accessed by simply pasting its URL on a web browser, but for upload, modification, bulk operations or for files that require authentication, one of the two Application Programming Interfaces (API) must be used to access the service: the first is based on the Representational State Transfer protocol (REST, [80]) and the second on the Simple Object Access Protocol (SOAP, [81]), both with a rich set of programmatic access code libraries available for developers [82]. Alternatively, users can interact with the service through various desktop client applications that have graphical front-ends, available for all operating systems ([**?**], [**?**], [83]).

A second storage option on the Amazon Cloud is Elastic Block Store (EBS, [55]), that provides POSIX-compliant [**?**] hard drives up to 1 TeraByte (TB), multiples of which can be attached to a running VM. While these data volumes persist after a Virtual Machine (VM) shutdown and can be re-attached to a new VM booted at a later time, data stored on EBS are not triple-replicated and are stored only on a single Amazon data center. This makes them prone to loss during physical disasters or hardware failures. Another option for data storage is the transient virtual hard-drive available as the file system of a running VM, but given that a VM would be used by a cost-conscious researcher only during execution of software (unless it is a 24/7 uptime web server), its storage should only be used as a temporary holding for the software's data outputs.

Overall, storage costs can be reduced by maintaining on high-availability, low-latency storage such as Amazon S3 only non-processed data for which value will be generated following compute. Alternatively, for data that have been already processed and is loss is not critical an option is the S3 Reduced Redundancy Storage [84], that allows users to cut down on the costs by storing non-critical, reproducible data at lower levels of redundancy and disaster-protection than Amazon S3s standard storage. Finally, if data preservation is important but the data are rarely accessed, the Amazon Glacier service [85] provides an archival storage option at one-tenth of the cost.

Which Factors Challenge Adoption of Cloud-Based Solutions for Bioinformatics

In the case of Cloud computing and similarly to a large-scale institutional server clusters, besides availability of compute resources that in both of these cases should not be a concern, three major facts affect users planning to access the computational system for data analysis: first, flexibility of moving data inputs and outputs to and from the system; second, available interfaces for the user interacting with the

system, and third; since both the Cloud and institutional clusters are multi-tenant systems, what are the mechanisms safeguarding eacy user's data integrity, privacy and isolation. While many studies are available in the literature that review available large-scale institutional clusters based on these criteria (ref,ref,ref), here we attempt to provide an overview of the Cloud in these respect, using as basis the Amazon EC2 platform.

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).

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.

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].

In the Cloud BioLinux work we combined the convenience of SaaS for end-users with the power of cloud computing, in order to bring pre-installed specialized bioinformatics application which need large computational capacity such as those for genome assembly while simplifying the the way users can get onto the cloud. A user can start and access the OSMF Frame VM instance in three simple steps by using the Amazon EC2 cloud console graphical user interface that is accessible via a web browser: first the user signs up for an Amazon EC2 account and after she obtains the credentials logins to the cloud console (http://aws.amazon.com/console); within the Amazon console the users clicks the Launch Instance Wizard button and specifies the OSMF Frame VM volume identifier (our project website will provide the VM identifier for the most recent update, but the latest VM will be also identifiable by the meta-data added to the volume); following the steps of the wizard within the web browser the users selects computational capacity and storage for the OSMF VM, and specifies a username and password for the OSMF WebInterface login (Fig.1, additional users can be created after the initial login); finally, once the wizard steps are complete and the VM status shows running, the user copies the assigned URL address of the VM from the Amazon cloud console in a new web browser window in order to access the OSMF interface. Through the URL users can get access to CloudMan and Galaxy (ref Enis and Brad)The process of starting a VM on the cloud and connecting to it has been documented for the JCVI Cloud BioLinux VM

instances (REF), but nonetheless more detailed documentation, video tutorials and user support will be available from the proposed project's website and discussion forum (see Education Outreach section).

Bioinformatics Computing and Science as A Service on the Cloud ?

Science as a Service (ScaaS) for bioinformatics, can be defined along similar lines with the Software as a Service computing model (SaaS, [86]), where software is running on remote datacenters. In this model, users access the software through a web browser or a desktop client application and there is no requirement other than a desktop computer with Internet connection, since the SaaS service provisions and manages the computing infrastructure, in addition to setting up the software and all its dependencies.

A Science as a Service (ScaaS) model could be of benefit to the bioinformatics community, where currently a unified approach is not available for researchers to access software or datasets, most of which have been generated as a result of federal grants awarded to individual investigators [87]. The different approaches range from websites created by small laboratories and which provide online access to specialized bioinformatic tools, to centralized web portals developed by large institutions such as NCBI [17] where web-based versions of mainstream applications including BLAST [**?**] are available. Nonetheless, in most cases datasets and software as source code are simply made available for download from FTP sites. In every case there are significant drawbacks including restrictions on the size of datasets that can be processed on the websites provided by small laboratories, as they are backed by compute servers with limited computational capacity; on the other hand, web portals set up by large institutions are also restricted on the computational resources that they can offer to the public (for example, users cannot upload multiple Gbp of sequence to the NCBI-BLAST website), while also NCBI cannot possibly provide online access to all available bioinformatic tools; finally, in the case of software available only as source code, provisioning informatics infrastructure and the technical expertise for performing specialized installation procedures can be a burden for non computationally-savvy investigators, with a prime example being genome assembly software.

Edit section on additional ScaaS (tools suites on Cloud), that has been moved on the tool suites on Cloud Section.

For users of the Cloud who require more control and additional flexibility to customize the computational infrastructure were their software is running, an alternative computing model is Infrastructure as a Service (IaaS, [88]). The Amazon Elastic Compute Cloud (EC2) is one of the most popular providers with this

model, and essentially has become a standard for IaaS service providers followed on the open-source replicas OpenStack and Eucalyptus ([45], [44]). The Amazon Cloud uses Virtual Machine (VM) servers as the basic unit for computational resource allocation, that are available in different capacities [89] allowing users the option to lease a portions of the underlying physical compute server capacity according to their budget and data processing needs. The specific term for the Amazon VMs is Amazons Machine Images (AMIs), and while these run on top of a virtualization layer user interaction is no different than accessing a physical server with a full operating system, processors and memory depending on the VM capacity selected. In a few recent studies VM performance characteristics such as read-write speed of the virtual hard drives, processors speeds and network latency for inter-communication of nodes of cluster instantiated using EC2 VMs, was found to be lagging that of in-house built compute clusters and specialized network interconnects such as Myrinet or Infiniband ([90], [91], [92], [93]). Nonetheless, users can rent at higher cost specialized VMs [89] that are connected within the Cloud through high-speed network or connect to virtual hard drives that are physically backed by Solid State Disks (SSD), and which have shown promising results [90]. Overall, other than specialized scientific applications that require specialized networking and configuration within a cluster, the on-demand availability of the Amazon EC2 Cloud service can provide a viable alternative to dedicated clusters, as it is similar to commodity hardware clusters often built in smaller labs without the expense or labor the build process requires.

Conclusions

Text for this section ...

Authors contributions

Text for this section ...

Acknowledgements

Text for this section ...

References

1. *Syntetic Biology, Euroscience Open Forum 2012.*
2. *What is life?: With mind and matter and autobiographical sketches.* Cambridge University Press 1992.
3. Gibson D, Glass J, Lartigue C, Noskov V, Chuang R, Algire M, Benders G, Montague M, Ma L, Moodie M: **Creation of a bacterial cell controlled by a chemically synthesized genome**. *science* 2010, **329**(5987):52–56.

4. 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.
5. 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.
6. Mason C, Elemento O: **Faster sequencers, larger datasets, new challenges**. *Genome Biology* 2012.
7. Illumina I: **Technical Specifications of Illumina Sequencers**.
8. Inc AB: **The SOLiD 5500 sequencing system.**
9. Inc PB: **PacBio RS sequencing technology**.
10. Loman N, Misra R, Dallman T: **Performance comparison of benchtop high-throughput sequencing platforms** 2012.
11. Mardis E: **Next-generation DNA sequencing methods**. *Annu. Rev. Genomics Hum. Genet.* 2008, **9**:387–402.
12. Gogol-Döring A, Chen W, et al.: **An overview of the analysis of next generation sequencing data**. *Methods in Molecular Biology* 2012, **802**:249–257.
13. Altschul S, Gish W, Miller W, Myers E, Lipman D, et al.: **Basic local alignment search tool**. *Journal of molecular biology* 1990, **215**(3):403–410.
14. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden T: **NCBI BLAST: a better web interface**. *Nucleic acids research* 2008, **36**(suppl 2):W5–W9.
15. Pruitt K, Tatusova T, Klimke W, Maglott D: **NCBI Reference Sequences: current status, policy and new initiatives**. *Nucleic acids research* 2009, **37**(suppl 1):D32–D36.
16. Schaller R: **Moore's law: past, present and future**. *Spectrum, IEEE* 1997, **34**(6):52–59.
17. **National Center for Biotechnology Information:**[[http://www.ncbi.nlm.nih.gov]].
18. Information NCFB: **Sequence Read Archive**.
19. Services AW: **Elastic Compute Cloud (EC2)**.
20. Services AW: **EC2 Pricing**.
21. **Amazon EC2 Cloud global regions**.
22. Clarke L, Zheng-Bradley X, Smith R, Kulesha E, Xiao C, Toneva I, Vaughan B, Preuss D, Leinonen R, Shumway M: **The 1000 Genomes Project: data management and community access**. *nature methods* 2012, **9**(5):459–462.
23. Krampis K: **Accessing 1000 Human Genomes Data with Cloud BioLinux**.
24. Krampis K: **Accessing 1000 Human Genomes Data with Cloud BioLinux**.
25. 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.
26. **SourceForge open-source code repository**.
27. Inc G: **Open-Source code repository**.
28. Subramaniam S: **The biology workbench—A seamless database and analysis environment for the biologist**. *Proteins* 1998, **32**:1–2.
29. Letondal C: **A Web interface generator for molecular biology programs in Unix**. *Bioinformatics* 2001, **17**:73–82.
30. Sarachu M, Colet M: **wEMBOSS: a web interface for EMBOSS**. *Bioinformatics* 2005, **21**(4):540–541.
31. 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.

32. 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.
33. 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.
34. 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.
35. *Camera 2.0: A data-centric metagenomics community infrastructure driven by scientific workflows*, IEEE 2010.
36. 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.
37. Aziz R: **Subsystems-based servers for rapid annotation of genomes and metagenomes**. *BMC Bioinformatics* 2010, **11**(Suppl 4):O2.
38. 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.
39. 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.
40. *Distributed workflow-driven analysis of large-scale biological data using biokepler*, ACM 2011.
41. team Gd: **Galaxy Bioinformatics Wiki**.
42. 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.
43. Services AW: **Amazon Cloud Console**.
44. **Eucalyptus Open Source Cloud Platform:**[[http://open.eucalyptus.com]].
45. **OpenStack Open Source Cloud Platform:**[[http://www.openstack.org]].
46. Groups G: .
47. framework CBD: .
48. **Broad Institute Genomespace portal:**[[http://www.genomespace.org]].
49. D OConnor B, Merriman B, Nelson S: **SeqWare Query Engine: storing and searching sequence data in the cloud**. *BMC bioinformatics* 2010, **11**(Suppl 12):S2.
50. **Illumina BaseSpace:**[[http://basespace.illumina.com]].
51. **DNANexus Inc.:**[[http://www.dnanexus.com]].
52. *Using clouds for metagenomics: A case study*, IEEE 2009.
53. Uhlig R, Neiger G, Rodgers D, Santoni A, Martins F, Anderson A, Bennett S, Kagi A, Leung F, Smith L: **Intel virtualization technology**. *Computer* 2005, **38**(5):48–56.
54. Dudley JT, Butte AJ: **In silico research in the era of cloud computing**. *Nature biotechnology* 2010, **28**(11):1181–1185.
55. **Elastic Block Store:**[[http://aws.amazon.com/ebs]].
56. **Amazon Web Services:**[[http://aws.amazon.com]].
57. **Simple Storage Service:**[[http://aws.amazon.com/s3]].
58. **Amazon Web Services Forums:**[[https://forums.aws.amazon.com]].
59. **Ubuntu Linux Operating System:**[[http://www.ubuntu.com/Cloud]].
60. **OpenStack Installation Documentation:**[[http://docs.openstack.org/essex/openstack-compute/starter/content/CentOS-de1592.html]].
61. **Eucalyptus Cloud Installation Packages:**[[http://www.eucalyptus.com/download/eucalyptus]].

62. **Google Apple Engine**.
63. **Microsoft Windows Azure**.
64. **Go Grid Cloud**.
65. **Flexiscale Cloud**.
66. **IBM SmartCloud**.
67. **Python Programming Language**.
68. **Microsoft .NET Framework**.
69. **BioTeam: Scriptable Cloud Infrastructures**.
70. **OpenStack at Magellan Cloud:**[[http://www.alcf.anl.gov/magellan]].
71. **OpenStack Cloud Installations:**[[http://openstack.org/user-stories/]].
72. **VirtualBox desktop virtualization software:**[[http://www.virtualbox.org]].
73. **Cloud BioLinux community site:**[[http://www.cloudbiolinux.org]].
74. **Educational Grants:**[[http://aws.amazon.com/grants/]].
75. **Community datasets hosting:**[[http://aws.amazon.com/datasets]].
76. **1000 Human Genomes dataset on the cloud:**[[http://aws.amazon.com/1000genomes/]].
77. **NCBI Flue sequence datasets:**[[http://aws.amazon.com/datasets/2419]].
78. **Ensembl Human Genome Annotation:**[[http://aws.amazon.com/datasets/3841]].
79. **Amazon CloudFront:**[[http://aws.amazon.com/cloudfront/]].
80. Fielding R: **Representational state transfer (REST)**. *Architectural Styles and the Design of Network-based Software Architectures. University of California, Irvine* 2000, :120.
81. **Simple Object Access Protocol (SOAP):**.
82. Libraries SD: .
83. **Amazon Cloud Storage S3 Add-On**.
84. **S3 Reduced Redudancy Storage**.
85. **Amazon Glacier storage service**.
86. *Service-oriented computing: Concepts, characteristics and directions Web Information Systems Engineering, 2003. WISE 2003. Proceedings of the Fourth International Conference on*, IEEE 2003.
87. Stein LD: **The case for cloud computing in genome informatics**. *Genome Biology* 2010, **11**(5):207.
88. Bhardwaj S, Jain L, Jain S: **Cloud computing: A study of infrastructure as a service (IAAS)**. *International Journal of engineering and information Technology* 2010, **2**:60–63.
89. **Amazon EC2 Cloud instance types:**.
90. Jackson K, Ramakrishnan L, Muriki K, Canon S, Cholia S, Shalf J, Wasserman H, Wright N: **Performance analysis of high performance computing applications on the amazon web services cloud**. In *Cloud Computing Technology and Science (CloudCom), 2010 IEEE Second International Conference on*, IEEE 2010:159–168.
91. Hill Z, Humphrey M: **A quantitative analysis of high performance computing with Amazon's EC2 infrastructure: The death of the local cluster?** In *Grid Computing, 2009 10th IEEE/ACM International Conference on*, IEEE 2009:26–33.
92. Boden N, Cohen D, Felderman R, Kulawik A, Seitz C, Seizovic J, Su W: **Myrinet: A gigabit-per-second local area network**. *Micro, IEEE* 1995, **15**:29–36.
93. Association IT: *InfiniBand Architecture Specification: Release 1.0*. InfiniBand Trade Association 2000.