Six 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" [1], where we have transitioned from the vision of Schrodinger that "Life is code" [2] to applied processes that convert digital code into DNA that runs a living organism [3].

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 [4]). 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 [5] while the also broadly used SOLiD sequencer has yields of a similar range up

to 90 Gb [6]. With the latest generation of instruments such as for example the HiSeq system, yield per run has reached 600 Gb [5], while the Pacific BioSciences sequencer yields 90 Gb in short amounts of time [7].

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 [8]).
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, [9]) expedition to study marine microbial diversity, and also the National Institutes of Health funded Human Microbiome Project to study human associated microbial communities [10].

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. Furthermore, the different types of analysis and software engineering technical expertise required [11], to convert raw sequence data to valuable information such as genome assembly can be daunting. Furthermore real scientific value is found through comparative studies and annotation of assembled genomes, that translates to more requirements for technical expertise and computing hardware. Currently available publicly databases and software tools accessible online such as the NCBI website for example [**?**], are not an option for downstream sequence analysis, since it cannot accept input sequence data files of more than 0.5 GB size for BLAST [12] sequence similarity search. For small laboratories acquiring a sequencing instrument, scientific value cannot be obtained from this investment, 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 [13] 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

Six 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?

Computational analysis can be a 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 and metagenomics for microbial population analysis. Furthermore, as sequencing capacities follow an uptrend that surpasses that of Moore's Law [**?**] and the cost per base pair follows an inverted trend, the genomics community can undertake sequencing projects of increasing scale and scope. Currently, publicly-funded projects in most cases conclude with upload of the sequences or annotated genomes in NCBI [14] databases such and the

Sequence Read Archive (SRA, [15]). Nonetheless, this provides little value to other 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 leverage these datasets. As an alternative to investing in informatics infrastructure researchers can rent computational and storage capacity from Cloud services such as Amazon EC2 [16]. 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 are billed based on the amount of computational capacity consumed on an hourly basis [17]. 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 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.

The 1000 Human genomes project [18] has demonstrated distribution of genomic datasets in combination with computational capacity, by depositing sequence reads and reference genome mapping data on the Amazon Cloud storage. There datasets are directly accessible by rented servers loaded with pre-installed bioinformatics software on the same Cloud, and this has been demonstrated with through the Cloud BioLinux project [19], [20], [21]. By placing data from public projects on Cloud compute platforms, value for other researchers in the community is immediately increased, since no infrastructure is required other than a computer with internet access to upload local data on the Cloud, rent computational capacity and perform for example comparative analysis with the 1000 Human genome data.

Established research institutes that perform sequencing of new genomes on a regular basis and have well-funded bioinformatic core facilities, also have the resources and expertise to deliver more to the community than just sequence reads and annotated genomes uploaded to publicly-accessible repositories such as NCBI [13]. Currently, while many of these centers deposit software on open-source code repositories such as SourceForge [**?**] or GitHub [22], a major bottleneck for researchers trying to utilize these tools 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, which can be overcome 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. In addition, 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, 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 [23], PISE [24], wEMBOSS [25], Mobyle [26] BioManager [27] and BioExtract [28]. 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 [29], CAMERA [30], EBI [31] and MG-RAST [32]. 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 [33], CloVR [34], Cloud BioLinux [21] and BioKepler [35] that are open-source and accessible to any laboratory or research group through the Amazon EC2 computer Cloud [16]. 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 [36]. Furthermore, it allows to leverage computational capacity beyond a single compute server through the Galaxy-Cloudman [37] 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 [21], [**?**], 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 [38], and easily perform large-scale data analysis as we have demonstrated with the 1000 Human genomes data [19], [20], [**?**].

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 [39] or Openstack [40] 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 [41]. 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 [42]. 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 [43].

Unique characteristics of Cloud Computing versus traditional Bioinformatics Infrastructures

One major component of cloud computing technology is virtualization [**?**] that allows entire compute servers including the operating system and 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 executes independently of the underlying hardware architecture on both Cloud and desktop computers. Cloud services such as Amazon EC2 [16] provide high- performance computer hardware with a virtualization layer, where users can 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, in addition to all required code libraries and configuration files in a ready to execute, compact and easy to download format. This approach can remove many of the roadblocks researchers encounter when trying to use open-source bioinformatics software that requires complex installation procedures, and consecutively make bioinformatics tools widely accessible to the research community.

In our experience with development of bioinformatics projects, it is difficult to provide long-term software

support or maintain for example web portals with data analysis tools and online 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, [21, 44]) 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 configuration such as installed software and bioinformatic pipelines, uploaded data and results generated by running the pipelines. A snapshot is an executable binary file as the original VM, and by using a snapshot as template the virtualization layer of a Cloud platform can instantiate multiple replicas of the original VM server [45]. Finally, a researcher can set her snapshots to be publicly accessible or share for example with specific users within the same Cloud see for example [45], therefore providing access for collaborators to both data and software in a ready to execute and compact format.

Regarding cost, the Amazon EC2 Cloud [16] for example charges VM snapshots at $0.01 US per GigaByte(GB) of storage used per month. Therefore, such low pricing can allow researchers to allocate a relatively small amount compared to that required on their informatics budgets for maintaining VM server snapshots for a number of years past the end of the funding cycle. For projects involving data release online though a web portal that is usually decommissioned shortly after the funding for a project ends due to the maintenance costs, using VM snapshots for archiving the portal on a Cloud platform allows other researchers in the community to lease computing time on the Cloud, and create fully-functional instances of the original VM server based on its snapshots. Overall, renting a VM from Cloud a providers for data analysis 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, leveraging virtualization and VM technology can provide two additional advantages: first, by placing data and pre-configured software on a publicly accessible, Cloud-based VM, allows for reproducibility, provenance and openness of the performed bioinformatic data analysis. For example, following publication of assembly and annotation results from a genome sequencing project, researchers in the community might need to re-run part of the bioinformatic analysis with additional data generated at their laboratories or to change algorithmic parameters to fine tune outputs such as gene predictions. Lowering the barrier of technical expertise and access to high-performance informatics infrastructure required for bioinformatics computing with next-generation sequencing datasets, is key for allowing researchers in the community to extract value from data released from publicly funded projects, and also

build additional value on top of existing data as similar studies take place. Second, with the Cloud researchers now have the capability to scale computational resources on-demand and based on the amount of data generated within a sequencing project, through provision of an appropriate number of VM servers. With this approach usage of resources can be scaled 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 (number of VM servers) can be reduced for the less computationally demanding tasks such as visualization and browsing of the sequence annotations. After funding has ended, a lab can further scale down the computational resources and cut its informatics costs by archiving the VMs using Whole System Snapshots [44], resulting in significant cost savings at all levels.

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

Amazon Web Services (AWS, [46]) is one of the largest cloud computing vendors, providing a Cloud backed by the same infrastructure that powers Amazon.com's e-commerce web portals. Services most applicable to bioinformatics computing offered by the Amazon Cloud include the Amazon Elastic Compute (EC2, [16]), Elastic Block Store (EBS, [45]) and Simple Storage Service (S3, [47]), while a complete list of the available options can be found at [46]. These services provide respectively compute cycles through a high-performance virtualization layer where users can run Virtual Machine (VM) servers, virtual hard drives and data storage, while none is tied to any specific operating system, or programming model. Amazon Web Services also provides competitive pricing for these options, where a large capacity VM 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, and storage costs $0.01 US per GigaByte (GB) used per month (for a complete list see [17]). In addition, a large software developer community with a lot of expertise has been build during the past few years around Amazon Web Services discussion forums [48], and based on our experience questions regarding the different services or requests for technical advice posted there, are always answered in a day or less.

Doing bioinformatics analysis on the Cloud is not limited only to the Amazon EC2 platform, since researchers with access to a local computing cluster at their home institution will have the option to download snapshots of a VM server and run it on open source Clouds, such as Eucalyptus [39] or OpenStack [40]. While OpenStack is the official Cloud of the Ubuntu Linux operating system [49] and can be included as default option on a compute cluster set-up to run with this Linux distribution, it can also be installed on clusters running other Linux versions [50]. On the other hand, Eucalyptus is available as

software package for a range of Linux distributions [51]. These two open-source Cloud platforms are replicas of Amazon EC2 offering identical Application Programming Interfaces (API), meaning that applications developed for one platform will work seamlessly across all three. Furthermore, by transferring VM servers across different installations of these Clouds, researchers are not required to configure software since everything comes pre-installed and ready-to-execute in the VM, allowing to leverage available bioinformatics tools and data analysis pipelines with minimal effort [21].

For researchers that do not want to lease computing time from the Amazon Cloud, or do not have access to a compute cluster where the open-source Cloud platforms can be installed, OpenStack is available through the government-funded Magellan Cloud [52], in addition to a number of academic computing centers in both the US and other countries [53]. Another option for users is to to run the VM server on a desktop computer, using virtualization software such as VirtualBox [54], that is also open-source and can be installed with a single step on Windows, Mac or Linux desktop computers. The Cloud BioLinux project for example, provides on its website VM server downloads for all these platforms [55].

Alternative Clouds include Google App Engine, Microsoft Azure, GoGrid [17] and FlexiScale [18].... from Vivek thesis.

Data storage using the Cloud provides the advantage that large-scale sequencing and other datasets can be easily exchanged among collaborators globally. The Amazon S3 for example [47] offers access to data centers across several geographical regions including the US East and West regions, European Union (Ireland), South America (Brazil), and Asia Pacific (both in Japan and Singapore). For data upload a researcher can choose the closest region in order to minimize data transfer latency over the internet, then initiated replication across the different locations through the Cloud service [56], and allowing collaborators worldwide to retrieve the data from their geographically closest Cloud data center.

Finally, researchers have the potential to access computational and data storage resources at no cost through the Amazon Web Services educational grants [57], that provide free computing and data storage resources for teaching courses in any field that involves the use of cloud computing. Furthermore, the Amazon Cloud has established the public datasets program [58] that hosts a variety of large-scale public datasets that are of wide interest to the different scientific communities. For example, genomic datasets hosted through that program include the 1000 human genomes data [**?**], [59], the NCBI flu genomes database [60], and Ensembl human genome annotation data [61] are a few examples. Researchers can access, copy, and perform computation on these datasets using VM servers on Amazon Cloud, and just pay for the compute resources they use.

Bioinformatics Computing and Science as A Service on the Cloud ?

Science as a Service (ScaaS) for computational analysis of scientific datasets, can be defined along similar lines with that of Software as a Service (SaaS, [62]), where a provider offers pre-installed and configured software on remote data centers. In principle, users of ScaaS do not need to provision any hardware other than a desktop computer and a network connection, and simply access software through a web browser or a client application. Furthermore, there is no requirement to purchase expensive hardware for large-scale computations or have expertise for performing complex software installations, as these are handled by the SaaS provider. This approach can lead to significant savings for both cost of hardware and labor especially in the case of specialized applications involving only occasional large-scale computations.

A Science as a Service (ScaaS) model for computing could be beneficial for the bioinformatics community, where currently there is significant fragmentation in the ways researchers access software or datasets created from publicly funded research [63]. Examples range from individual-run, small laboratories that set-up websites providing online access to bioinformatic tools, to centralized web portals developed by large institutions such as NCBI where researchers can access mainstream applications such as BLAST [**?**], and finally software available for download just as source code along with datasets on FTP sites. A significant disadvantage is that computer servers offered by small laboratories have fixed computational capacity that poses a limit on the size of datasets that can be processed, while large institutions still are restricted for the computational resources that they can offer to the public and cannot possibly provide access to all specialized bioinformatic tools. In the case of software available only for download, provisioning computational infrastructure and acquiring the technical expertise to compile software and perform complex installation procedures can be a burden for scientists, especially in cases where access to high performance computing servers for specialized bioinformatics tasks such as genome assembly is required.

Both public and commercial offerings of ScaaS sequence analysis tool suites on the Cloud have become available in recent years, including Illumina BaseSpace [64], Broad Institute's GenomeSpace [65], SeqWare [66]. On the commercial side, DNAnexus [67] 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.

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, [68]). The Amazon Elastic Compute Cloud (EC2) is one of the most popular providers with this

model and has become the standard for IaaS service providers. This Cloud uses Virtual Machines (VM) servers as the basic unit for computational resource allocation, come in different capacities and costs, giving users the option to access different portions [] of the underlying physical servers. The term for the Amazon VMs is specifically Amazons Machine Images (AMIs), which run on top of a virtualization layer, but user interaction is no different than accessing a physical server over the network with its own virtual operating system, CPU and memory depending on the VM capacity selected. The network latency for inter-communication of nodes of cluster instantiated using EC2 VMs, was found to be higher [69] than that of high-speed, specialized cluster network interconnects such as Myrinet or Infiniband ([70], [71]). Nonetheless, users can rent at higher cost specialized VMs that are connected within the Cloud with high-speed network [] and which have shown promising results in studies so far []. Overall, for applications other than those relying heavily on the use of message passing protocols such as MPI [] that require specialized networking within the cluster, the on-demand accessibility of the Amazon EC2 Cloud service can provide a viable alternative to dedicated clusters, directly comparable with commodity hardware built clusters often found in scientific labs.

The options for data storage on the Amazon Cloud are multiple, with first being the transient virtual hard-drive available as the root file system available when a user is running a VM. A second option includes additional Elastic Block Store (EBS, [45]) hard drives up to 1 TeraByte (TB), multiples of which can be attached to each VM. While these are fully POSIX-compliant [] data volumes that persist after VM shutdown and can be re-attached to a new VM booted at a later time, data stored on EBS only reside at a single Amazon data center. Alternatively, disaster- proof, triple replicated across different data centers storage options is the Amazon Simple Storage Service (S3, []). Note that in this case triple replication is not synonymous with backup in order to prevent accidental deletion by the user, but instead refers to protection from physical disaster and permanent loss when data reside at a single data center. Amazon S3 provides storage at low cost, high scalability for concurrent reading of the same data file from multiple applications running on EC2 VMs on the same cloud or any server on the internet, in addition to highly-available, disaster-proof service where data are replicated across three different data centers. The S3 storage model deviates from that of a POSIX-compliant hard drives, but instead are data objects with unique Uniform Resource Location (URL) identifiers across the Amazon Cloud platform and the wider internet. In more detail, S3 is organized over a two-level namespace with top-level folders called "buckets", and while each Amazon account may have up to 100 S3 buckets each bucket can store an unlimited number of data objects. While each file on S3 can be accessed directly using its URL, for querying the service and

in order to retrieve the list of available files two Application Programming Interfaces (API) are available. The two APIs supported are Representational State Transfer (REST,) and the Simple Object Access Protocol (SOAP) is a protocol specification for exchanging structured information in the implementation of Web Services in computer networks) [W3C, Soap Version 1.2, June 2003, http://www.w3.org/TR/soap/], REST (: emphasizes scalability of component interactions, generality of interfaces, independent deployment of components, and intermediary components to reduce interaction latency) [Fielding, R. T. 2000. Architectural Styles and the Design of Network-Based Software Architectures, PhD Dissertation, University of California, Irvine, 2000.], BitTorrent [BitTorrent. http://www.bittorrent.com].
The user is assigned an identity key and a private key when they register for the Amazon's Web Services, using which one can access the S3 account. The security provided by S3 is dependent on these identity key and the private key.
Using BitTorrent S3 can provide tracker and seed functionality which can save bandwidth when multiple concurrent clients are demanding the same set of objects. S3 has attracted a large user base due to its simple charging scheme, unlimited storage capacity, open protocols, and simple API for easy integration with applications. But the current S3 design needs to be improved before it can provide durable storage for scientific community[24]. A storage infrastructure which aims at data intensive scientific community must provide data durability, data availability, access performance, usability, support for security and privacy and low cost. S3 show 100 percent durability,for concurrent performance and remote access performance with high competency, but the cost of storing data for experiments like DZero[20] can cost upto $1.02 million for a year$[24]. $Storage cost can be reduced by storing 'cold' data (rarely used data) on low cost storage and maintaining on$
$availability, low -$
$latency storage. Another way is to only store raw data and derive the rest from the raw data. This is called the Reduced Redundancy S$
$//aws.amazon.com/s3$]. $It allows users to cut down on the costs by storing non -$
$critical, reproducible data at lower levels of redundancy than Amazon S3's standard storage S3 doesn't provide any checkpointing or ba$
$end running on Amazon's EC2 service. The front -$
$end would be responsible for individual account management, fine - grained trust decisions, and billing.$[24]

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).

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).

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. *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. Mason C, Elemento O: **Faster sequencers, larger datasets, new challenges**. *Genome Biology* 2012.
5. Illumina I: **Technical Specifications of Illumina Sequencers**.
6. Inc AB: **The SOLiD 5500 sequencing system.**
7. Inc PB: **PacBio RS sequencing technology**.
8. Mardis E: **Next-generation DNA sequencing methods**. *Annu. Rev. Genomics Hum. Genet.* 2008, **9**:387–402.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.

14. **National Center for Biotechnology Information:**[[http://www.ncbi.nlm.nih.gov]].
15. Information NCFB: **Sequence Read Archive**.
16. Services AW: **Elastic Compute Cloud (EC2)**.
17. Services AW: **EC2 Pricing**.
18. 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.
19. Krampis K: .
20. Krampis K: .
21. 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.
22. Inc G: **Open-Source code repository**.
23. Subramaniam S: **The biology workbench—A seamless database and analysis environment for the biologist**. *Proteins* 1998, **32**:1–2.
24. Letondal C: **A Web interface generator for molecular biology programs in Unix**. *Bioinformatics* 2001, **17**:73–82.
25. Sarachu M, Colet M: **wEMBOSS: a web interface for EMBOSS**. *Bioinformatics* 2005, **21**(4):540–541.
26. 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.
27. 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.
28. 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.
29. 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.
30. *Camera 2.0: A data-centric metagenomics community infrastructure driven by scientific workflows*, IEEE 2010.
31. 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.
32. Aziz R: **Subsystems-based servers for rapid annotation of genomes and metagenomes**. *BMC Bioinformatics* 2010, **11**(Suppl 4):O2.
33. 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.
34. 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.
35. *Distributed workflow-driven analysis of large-scale biological data using biokepler*, ACM 2011.
36. team Gd: **Galaxy Bioinformatics Wiki**.
37. 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.
38. Services AW: **Amazon Cloud Console**.
39. **Eucalyptus Open Source Cloud Platform:**[[http://open.eucalyptus.com]].
40. **OpenStack Open Source Cloud Platform:**[[http://www.openstack.org]].

41. Groups G: .
42. framework CBD: .
43. *Using clouds for metagenomics: A case study*, IEEE 2009.
44. Dudley JT, Butte AJ: **In silico research in the era of cloud computing**. *Nature biotechnology* 2010, **28**(11):1181–1185.
45. **Elastic Block Store:**[[http://aws.amazon.com/ebs]].
46. **Amazon Web Services:**[[http://aws.amazon.com]].
47. **Simple Storage Service:**[[http://aws.amazon.com/s3]].
48. **Amazon Web Services Forums:**[[https://forums.aws.amazon.com]].
49. **Ubuntu Linux Operating System:**[[http://www.ubuntu.com/Cloud]].
50. **OpenStack Installation Documentation:**[[http://docs.openstack.org/essex/openstack-compute/starter/content/CentOS-de1592.html]].
51. **Eucalyptus Cloud Installation Packages:**[[http://www.eucalyptus.com/download/eucalyptus]].
52. **OpenStack at Magellan Cloud:**[[http://www.alcf.anl.gov/magellan]].
53. **OpenStack Cloud Installations:**[[http://openstack.org/user-stories/]].
54. **VirtualBox desktop virtualization software:**[[http://www.virtualbox.org]].
55. **Cloud BioLinux community site:**[[http://www.cloudbiolinux.org]].
56. **Amazon CloudFront:**[[http://aws.amazon.com/cloudfront/]].
57. **Educational Grants:**[[http://aws.amazon.com/grants/]].
58. **Community datasets hosting:**[[http://aws.amazon.com/datasets]].
59. **1000 Human Genomes dataset on the cloud:**[[http://aws.amazon.com/1000genomes/]].
60. **NCBI Flue sequence datasets:**[[http://aws.amazon.com/datasets/2419]].
61. **Ensembl Human Genome Annotation:**[[http://aws.amazon.com/datasets/3841]].
62. *Service-oriented computing: Concepts, characteristics and directions Web Information Systems Engineering, 2003. WISE 2003. Proceedings of the Fourth International Conference on*, IEEE 2003.
63. Stein LD: **The case for cloud computing in genome informatics**. *Genome Biology* 2010, **11**(5):207.
64. **Illumina BaseSpace:**[[http://basespace.illumina.com]].
65. **Broad Institute Genomespace portal:**[[http://www.genomespace.org]].
66. 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.
67. **DNANexus Inc. :**[[http://www.dnanexus.com]].
68. 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.
69. 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.
70. 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.
71. Association IT: *InfiniBand Architecture Specification: Release 1.0*. InfiniBand Trade Association 2000.