bioRxiv preprint doi: https://doi.org/10.1101/209734. this version posted October 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

miCloud: a plug and play, on-premises bioinformatics cloud, providing seamless integration with Illumina genome sequencers.

Baekdoo Kim[1], Thahmina Ali[1], Konstantinos Krampis[1,2], Changsu Dong[1], Bobby Laungani[1], Claudia Wultsch[3] and Carlos Lijeron[1].**

[1]Weill Cornell Medicine - Hunter College Belfer Research Building, New York, NY; [2]Department of Biological Sciences, Hunter College of The City University of New York, NY; [3]American Museum of Natural History, New York, NY.

[**]Corresponding author: kk104@hunter.cuny.edu

Benchtop genome sequencers such as the Illumina MiSeq or MiniSeq [1], [2] are revolutionizing genomics research for smaller, independent laboratories, by enabling access to low-cost Next Generation Sequencing (NGS) technology in-house. These benchtop genome sequencing instruments require only standard laboratory equipment, in addition to minimal time for sample preparation. However, post-sequencing bioinformatics data analysis still presents a significant bottleneck, for research laboratories lacking specialized software and technical data analysis skills on their teams. While bioinformatics computes clouds providing solutions following a Software as a Service (SaaS) are available ([3]–[6], review in [7]), currently, there are only a few options which are user-friendly for non-experts while at the same time are also low-cost or free. One primary example is Illumina BaseSpace [8] that is very easy to access by non-experts, and also offers an integrated solution where data are streamed directly from the MiSeq sequencing instrument to the cloud. Once the data is on the BaseSpace cloud, users can access a range of bioinformatics applications with pre-installed algorithms through an intuitive web interface. Nonetheless, BaseSpace can be a costly solution as a yearly subscription depending on whether the user is associated with an academic or private institution, ranges in price from $999 - $4,999. Additional “iCredits” [9] might need to be purchased for frequent users that exhaust the base credit allowance as part of the subscription. Considering the reduction of computer hardware cost in recent years, a multi-core Intel Xeon server with 64 GigaByte (GB) of memory and multiple TeraByte (TB) of storage is priced less than the yearly subscription to Basespace [10], and similarly when compared to renting compute cycles from providers such as Amazon Web Services (AWS) [11]. Furthermore, the current generation of laptops usually come with 6-10 GigaBytes (GB) of memory and 1 TeraByte (TB) of storage, providing enough computational capacity to analyze data from small NGS experiments [12] that include only a few samples.

We developed miCloud, a bioinformatics platform for NGS data analysis, as a solution to fill the gap between the low-cost, widely available computational resources and lack of user-friendly bioinformatics software. Laboratories lacking NGS data analysis expertise can easily perform analysis of data generated from in-house sequencing instruments or external service providers using this platform. The miCloud is highly modular and is based on Docker virtual machine containers [13] with components (e.g., user interface, file manager,pre-configured data analysis

bioRxiv preprint doi: https://doi.org/10.1101/209734. this version posted October 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

pipelines) encapsulated in separate containers (**Fig 1a**). These virtual machine containers are automatically instantiated and interconnected upon installation of the miCloud, which is as simple as running a script that users can download from our code repository ([14], **Suppl.**). As a result, users have access to an on-premises cloud application that leverages their local computational infrastructure and can access the miCloud graphical user interface (**Fig 1b**) via a web browser on a personal computer. The interface provides built-in functionality for users to run and monitor a set of preconfigured bioinformatics analysis pipelines, in addition to import, export, and management of NGS sequencing data through a visual file explorer (**Fig 1c**). By default, there are three pipelines ready to execute with the miCloud upon installation, two for single and paired-end ChIP-Seq data, in addition to one more for paired-end RNA-Seq data (**Suppl.**). Each is a multi-step pipeline processing the input data through a set of bioinformatics algorithms, which provide beginning to end data analysis and generate differential gene expression charts, tables of p-values for chromatin peaks found in the genome, and lists of SNP polymorphisms. The pipelines have been implemented following standard published protocols for RNA-seq and CHIP-seq [15], [16]. Additionally, we have integrated the Visual Omics Explorer (VOE, [17]) with the miCloud, to provide users with access to rich, interactive visualizations and publication-ready graphics from the pipeline outputs.

When the required computational capacity is available, users can start multiple replicas of each container through the miCloud interface, to run copies of the pipeline in parallel and process multiple datasets (**Suppl.**). The miCloud platform builds on technology we have previously developed through the Bio-Docklets project [18], and for the pipeline implementations, we have used the Galaxy workflow engine which runs inside the Docker containers. The BioBlend software library is also integrated with the miCloud, enabling seamless connection to the Galaxy Application Programming Interface (API). Furthermore, access to the API is used internally by the miCloud for controlling the execution of the pipelines within the Docker containers (**Fig 1c**), in addition to managing data inputs and outputs from the containers to the local filesystem. With these technologies as the foundation, we provide a fully extensible platform enabling developers to implement new Docker containers with bioinformatics pipelines and make them easily accessible for users through the miCloud interface. Developers can also leverage our platform to provide access to NGS data analysis containers publicly available from the Galaxy community [19] or Docker Store [20].

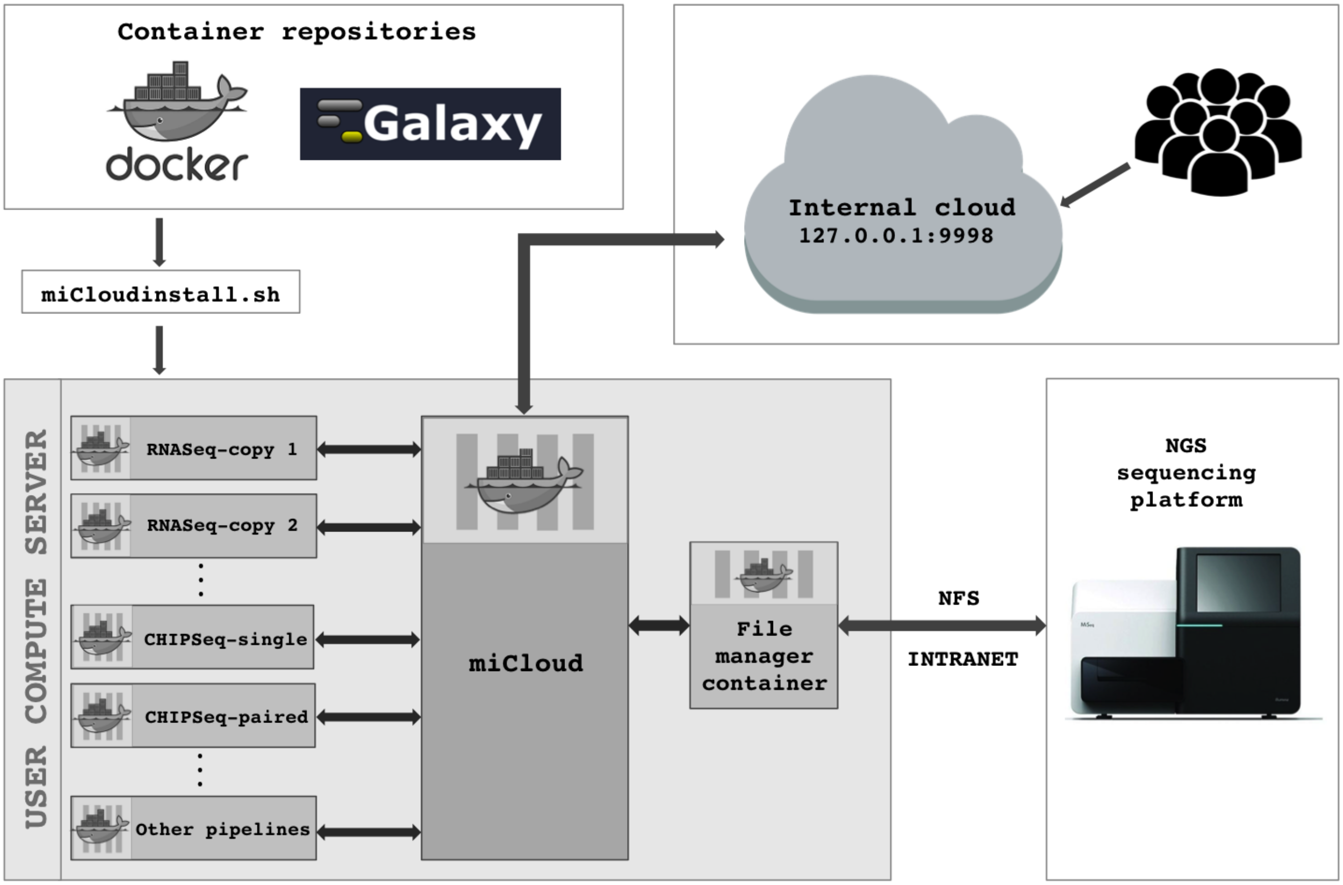
While Illumina sequencing instruments include a complete Windows desktop computer onboard that is used for the instrument control software, the operating system and data storage capacity of the onboard computer is not suitable for running the data-intensive pipelines such as RNA-seq and CHIP-seq. However, the operating system's desktop can be accessed through the touch-screen monitor on the instrument [21], where users can simply create a Windows Network File System (NFS, [22]) folder with password protection, to share the sequencing read data (**Suppl.**) on the local network. The miCloud file manager was implemented so that it includes build-in NFS connectivity, allowing users to easily connect to the Illumina instrument and transfer the data over the local network on the computer or server where the miCloud containers are running.

bioRxiv preprint doi: https://doi.org/10.1101/209734. this version posted October 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The miCloud provides a bioinformatics platform for NGS data analysis that can be deployed without any technical expertise, enabling laboratories with an Illumina desktop genome sequencer, to seamlessly integrate the instrument with a fully-featured, on-site data analysis cloud. Bioinformatics developers can leverage the plethora of publicly available Docker containers with preconfigured NGS data analysis pipelines, to deploy complete data analysis solutions through miCloud. With this approach, non-bioinformatics experts have easy access to a fully-featured solution for execution of complex bioinformatics pipelines, where the underlying software complexity is fully abstracted through an intuitive interface.

Figure 1.

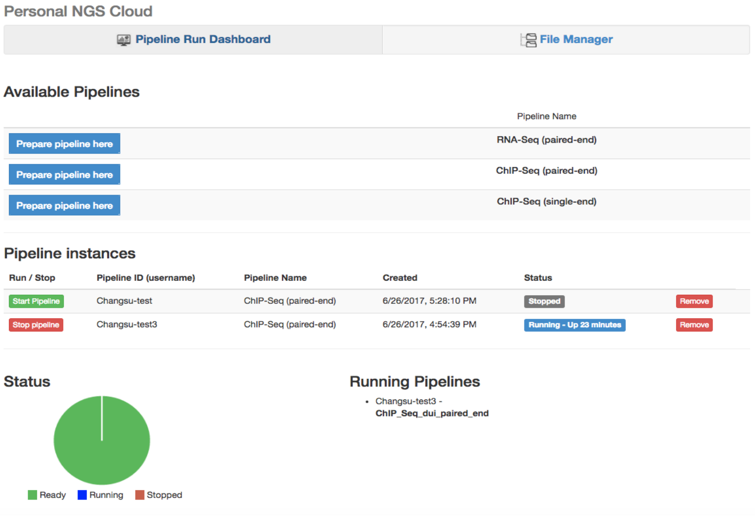
A. miCloud modular design with the different components in Docker containers.



bioRxiv preprint doi: https://doi.org/10.1101/209734. this version posted October 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 1 (continued).

B. miCloud interface.



Personal NGS Cloud

Available Pipelines

	Pipeline Name
Prepare pipeline here	RNA-Seq (paired-end)
Prepare pipeline here	ChIP-Seq (paired-end)
Prepare pipeline here	ChIP-Seq (single-end)

Pipeline instances

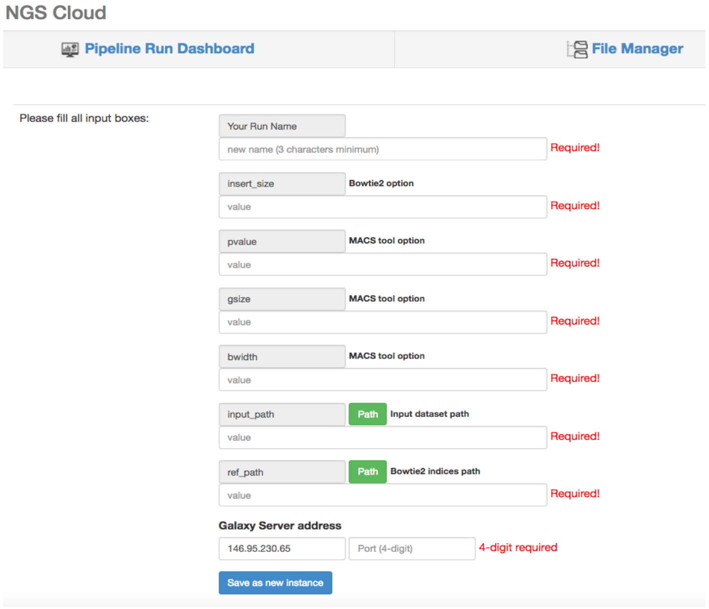
Run / Stop	Pipeline ID (username)	Pipeline Name	Created	Status	
Start Pipeline	Changsu-test	ChIP-Seq (paired-end)	6/26/2017, 5:28:10 PM	Stopped	Remove
Stop pipeline	Changsu-test3	ChIP-Seq (paired-end)	6/26/2017, 4:54:39 PM	Running - Up 23 minutes	Remove

Status

Running Pipelines

- Changsu-test3 - **ChIP_Seq_dui_paired_end**

Dashboard with pre-configured RNA-seq and CHIP-seq pipelines.



NGS Cloud

Please fill all input boxes:

Your Run Name
new name (3 characters minimum) Required!

insert_size Bowtie2 option
value Required!

pvalue MACS tool option
value Required!

gsize MACS tool option
value Required!

bwidth MACS tool option
value Required!

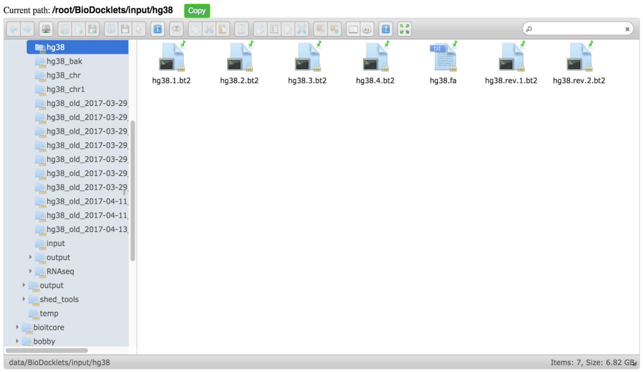
input_path Path Input dataset path
value Required!

ref_path Path Bowtie2 indices path
value Required!

Galaxy Server address
146.95.230.65 Port (4-digit) 4-digit required

Pipeline run setup interface, where users enter the parameters for the different algorithms of a pipeline in each box.

C. miCloud file manager enabling users to browser local network file locations.



Current path: /root/BioDocklets/input/hg38

hg38
hg38_bak
hg38_chr
hg38_chr1
hg38_old_2017-03-29
hg38_old_2017-03-29
hg38_old_2017-03-29
hg38_old_2017-03-29
hg38_old_2017-03-29
hg38_old_2017-03-29
hg38_old_2017-03-29
hg38_old_2017-04-11
hg38_old_2017-04-11
hg38_old_2017-04-13
input
output
RNAseq
output
shed_tools
temp
bioitcore
bobby

hg38.1.bt2 hg38.2.bt2 hg38.3.bt2 hg38.4.bt2 hg38.fa hg38.rev.1.bt2 hg38.rev.2.bt2

data/BioDocklets/input/hg38 Items: 7, Size: 6.82 GB

bioRxiv preprint doi: https://doi.org/10.1101/209734. this version posted October 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References:

[1] “MiSeq System | Focused power for targeted gene and small genome sequencing,” 10-Oct-2017. [Online]. Available: https://www.illumina.com/systems/sequencing-platforms/miseq.html. [Accessed: 10-Oct-2017].

[2] “MiniSeq Sequencing System | Small, affordable benchtop sequencer,” 10-Oct-2017. [Online]. Available: https://www.illumina.com/systems/sequencing-platforms/miniseq.html. [Accessed: 10-Oct-2017].

[3] “Seven Bridges Genomics - The biomedical data analysis company,” 10-Oct-2017. [Online]. Available: https://www.sevenbridges.com/. [Accessed: 10-Oct-2017].

[4] “Galaxy Community Server” 10-Oct-2017. [Online]. Available: https://usegalaxy.org/. [Accessed: 10-Oct-2017].

[5] “BGI Online,” 10-Oct-2017. [Online]. Available: http://www.genomics.cn/bgionline/. [Accessed: 10-Oct-2017].

[6] A. Wilke, J. Wilkening, E. M. Glass, N. L. Desai, and F. Meyer, “An experience report: porting the MG-RAST rapid metagenomics analysis pipeline to the cloud: MG-RAST METAGENOMICS DATA ANALYSIS USING CLOUD RESOURCES,” *Concurrency and Computation: Practice and Experience*, vol. 23, no. 17, Dec. 2011.

[7] K. Krampis and C. Wultsch, “A Review of Cloud Computing Bioinformatics Solutions for Next-Gen Sequencing Data Analysis and Research,” *Methods in Next Generation Sequencing*, vol. 2, no. 1.

[8] “Illumina BaseSpace Sequence Hub,” 10-Oct-2017. [Online]. Available: https://basespace.illumina.com/home/index. [Accessed: 10-Oct-2017].

[9] “Illumina BaseSpace Cloud iCredits and Billing,” 10-Oct-2017. [Online]. Available: https://help.basespace.illumina.com/articles/descriptive/icredits-and-billing/. [Accessed: 10-Oct-2017].

[10] “PowerEdge T330 Tower Server | Dell United States,” 10-Oct-2017. [Online]. Available: http://www.dell.com/en-us/work/shop/cty/pdp/spd/poweredge-t330/pe_t330_1566?cid=302825&st=&gclid=CjwKCAjw3_HOBRBaEiwAvLBbosOm5up-nKbDI3NxACtj750l0d7L_xm5CXRKVcNH3kraH4vQYs4ayhoCHsgQAvD_BwE&lid=5758065&VEN1=scZ6rOQwY,112783110789,901q5c14135,c,,PE_T330_1566&VEN2=,&dgc=st&dgseg=so&acd=12309152537501410&VEN3=812104053946253783. [Accessed: 10-Oct-2017].

[11] “EC2 Instance Pricing – Amazon Web Services (AWS),” 10-Oct-2017. [Online]. Available: https://aws.amazon.com/ec2/pricing/on-demand/. [Accessed: 10-Oct-2017].

[12] “http: //www.ba.itb.cnr.it/gisel/file-aut-downloads/teaching/NGS_VM.pdf,” 10-Oct-2017. [Online]. Available: http://www.ba.itb.cnr.it/gisel/file-aut-downloads/teaching/NGS_VM.pdf.

bioRxiv preprint doi: https://doi.org/10.1101/209734. this version posted October 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[Accessed: 10-Oct-2017].

[13] “What is Docker?,” 10-Oct-2017. [Online]. Available: https://www.docker.com/what-docker. [Accessed: 10-Oct-2017].

[14] “miCloud Github source code repository.”: https://github.com/BCIL/Personal-NGS-Cloud [Accessed: 10-Oct-2017].

[15] “Galaxy CHIPseq published workflow.”: https://usegalaxy.org/u/chip-seq-helin-group/w/mmusculus-mm10-create-bam-bigwig-and-peakcalling-for-chip-seq [Accessed: 10-Oct-2017].

[16] “Galaxy RNAseq published workflow.” : https://usegalaxy.org/u/fluidigmngs/w/rnaseq-workflow [Accessed: 10-Oct-2017].

[17] B. Kim, T. Ali, S. Hosmer, and K. Krampis, “Visual Omics Explorer (VOE): a cross-platform portal for interactive data visualization,” *Bioinformatics*, vol. 32, no. 13, Jul. 2016.

[18] B. Kim, T. Ali, C. Lijeron, E. Afgan, and K. Krampis, “Bio-Docklets: virtualization containers for single-step execution of NGS pipelines.,” *GigaScience*, vol. 6, no. 8, pp. 1–7, Aug. 2017.

[19] “Virtual Appliances,” 10-Oct-2017. [Online]. Available: https://galaxyproject.org/virtual-appliances/#galaxy-virtual-appliance-directory. [Accessed: 10-Oct-2017].

[20] “Explore - Docker Store,” 10-Oct-2017. [Online]. Available: https://store.docker.com/search?q=ngs&source=community&type=image. [Accessed: 10-Oct-2017].

[21] Illumina MiSeq sequencing instrument user guide “https: //support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miseq/miseq-system-guide-15027617-01.pdf,” 10-Oct-2017. [Online]. Available: https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miseq/miseq-system-guide-15027617-01.pdf. [Accessed: 10-Oct-2017].

[22] “Network File System - Wikipedia,” 10-Oct-2017. [Online]. Available: https://en.wikipedia.org/wiki/Network_File_System. [Accessed: 10-Oct-2017].