

Cloud Computing Solutions for Genomics Across Geographic, Institutional and Economic Barriers

Ntinios Krampis
Asst. Professor
J. Craig Venter Institute
kkrampis@jcvl.org

<http://www.jcvl.org/cms/about/bios/kkrampis/>

Workshop Schedule

Morning Session: Background Presentations and Prep

- 9:00 – 9:45 Introduction to Cloud Computing for Bioinformatics
- 9:45 – 10:00 Questions and Answers
- 10:00 – 10:30 Using Cloud BioLinux on the Amazon EC2 Cloud
- 11:00 – 12:00 Preparation: install Cloud Virtual Machines on laptops

Afternoon Session: Hands on Session

- 1:30 – 3:30 Bioinformatic Analysis using Cloud BioLinux
- 3:30 – 5:00 Customized Bioinformatics Solutions for Participants

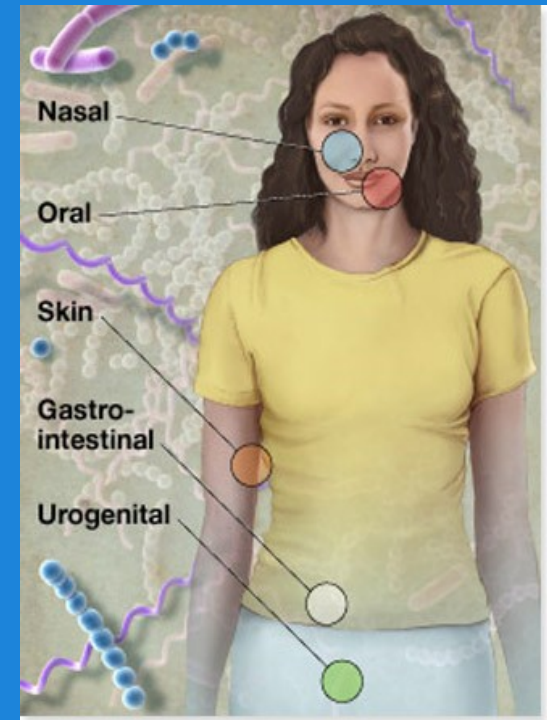
A little bit of background information...

- Konstantinos (Ntinou) Krampis, started working at J. Craig Venter Inst. (JCVI) in 2009
- Background training in Molecular Biology, PhD in Bioinformatics
- Research: cloud and high-performance computing, genome assembly
- Projects: Cloud BioLinux (cloudbiolinux.org)
- Taught Cloud BioLinux workshop at Univ. of Limpopo last May
- Slides available at <http://www.slideshare.com/>
- Email me for slides, meeting, questions: kkrampis@jcvl.org

J. Craig Venter Institute (JCVI)

Large-scale genome sequencing and bioinformatics computing

- Human Microbiome Project (HMP): genome sequencing of microbes living in and on the human body
- Global Ocean Sampling (GOS) survey: genome sequencing of microbes sampled from oceans around the world



— 2003 – 2008 Routes — 2009 – 2010 Route

JCVI: sequencing and computing infrastructure

- core sequencing laboratory: 454, Solexa, HiSeq, IonTorrent on the way
- dedicated bioinformatics department (57 bioinformaticians)
- large-scale computations, ~1000 node Sun Grid Engine (SGE) cluster



Low-cost sequencing instruments

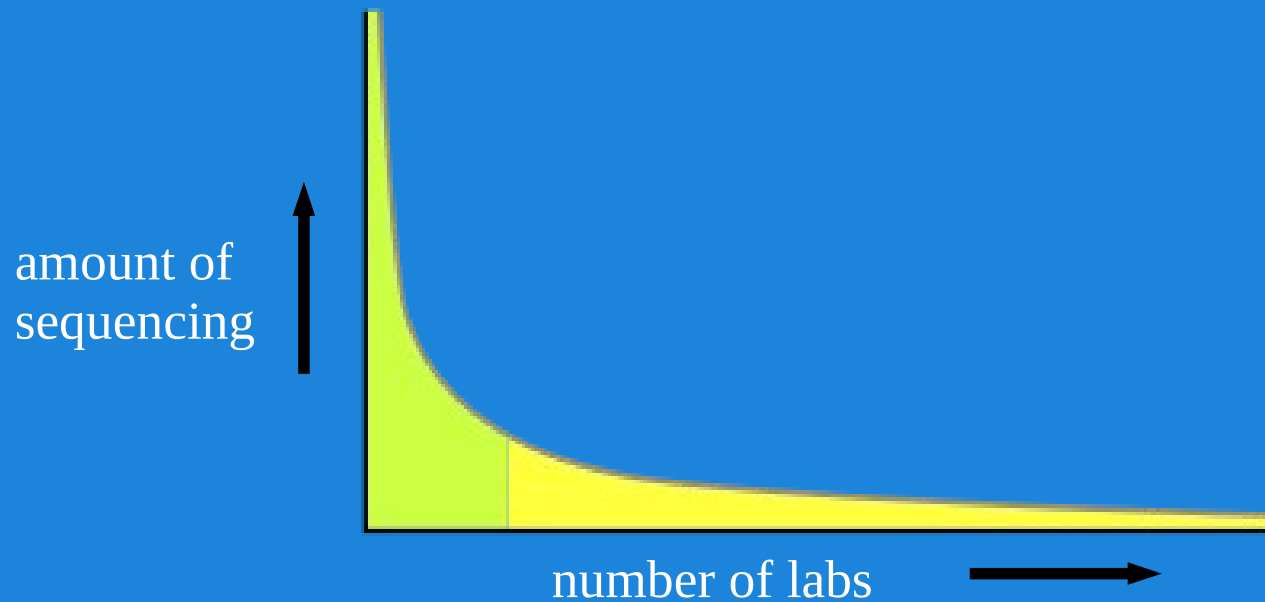
- small-factor sequencers available: GS Junior by 454, MiSeq by Illumina
- bacterial, viral, small fungal genomes, sequencing for variant discovery
- sequencing as a standard technique in molecular biology and genetics
- RNAseq (instead of microarrays) and ChiPseq (instead of yeast 2-hybrid)



<http://www.gsjunior.com/>

<http://www.illumina.com/systems/miseq.ilmn>

More small laboratories doing genome sequencing



acquiring the sequence data is only the first step...

Sequencing instruments shipped with minimal computational capacity

- Problem 1: sequencing data analysis requires high-performance and expensive computing hardware, for example: genome assembly, BLAST, genome annotation
- Problem 2: much of bioinformatics software are difficult to install by biologists, need technical expertise with operating systems, compiling source code etc.



Each lab building their own informatics infrastructure ?

- small labs need additional funds to build computing clusters
- funds for bioinformaticians and software developers to maintain the clusters and software
- duplication of effort across labs
- sub-optimal utilization of the hardware due to small amounts of sequencing



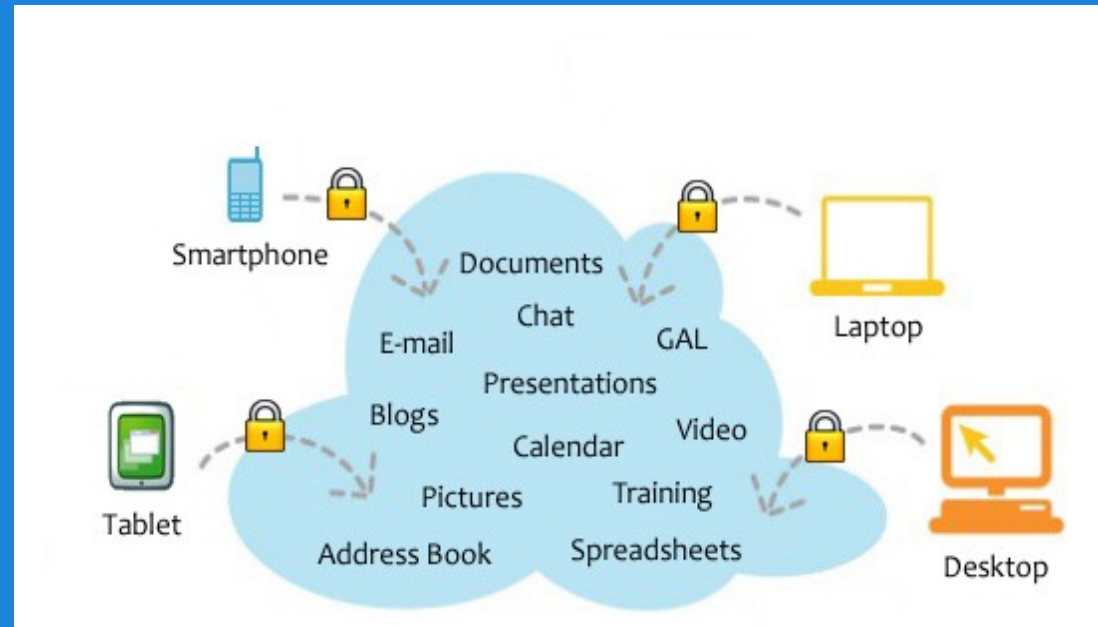
Large sequencing centers offering bioinformatics analysis services ?

- Bioinformatic Resource Centers (BRC)
- bioinformatic analysis coupled with sequencing of an organism
- mostly provide data browsing and few analysis tools to the public
- cannot serve the bioinformatic needs of every small lab acquiring a sequencing instrument
- need end-to-end solutions, users submit sequence data and get final annotation



Solving Problem 1: using high-performance computing hardware available on the cloud

- cloud computing : high performance computers and data storage, remotely accessible through the Internet
- we are all using the cloud: Gmail, Google Docs, FaceBook; you store and access data on a remote computer
- cloud computers rented pay-as-you-go by service providers such as Amazon Elastic Compute Cloud (EC2)



The Amazon EC2 cloud computing service

- a subsidiary company of Amazon.com, rents computing pay-as-you go
- cloud computers cost \$0.085 - \$2 per hr (max 64GB memory and 8 processors)
- used by companies that need additional computers without investing on hardware
- physical locations US East / West regions, EU, Singapore, Japan researchers
- democratizes access to computing resources outside of institutional, economic or national boundaries



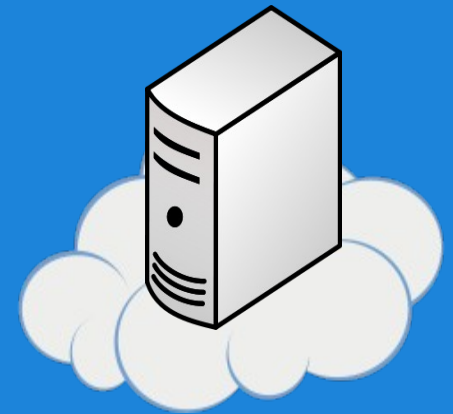
<http://aws.amazon.com>

750 hours free for new users, sign up here:
<http://aws.amazon.com/free/>

How does cloud computing work ?

- Cloud computing evolved from Virtualization technology
- operating system, bioinformatics software and data, are installed on a Virtual Machine (VM)
- a VM is a full-featured Unix server, in a single, executable binary file
- no need to compile source code, set up configuration files, software installation dependencies
- why Virtualization: simplify IT maintenance

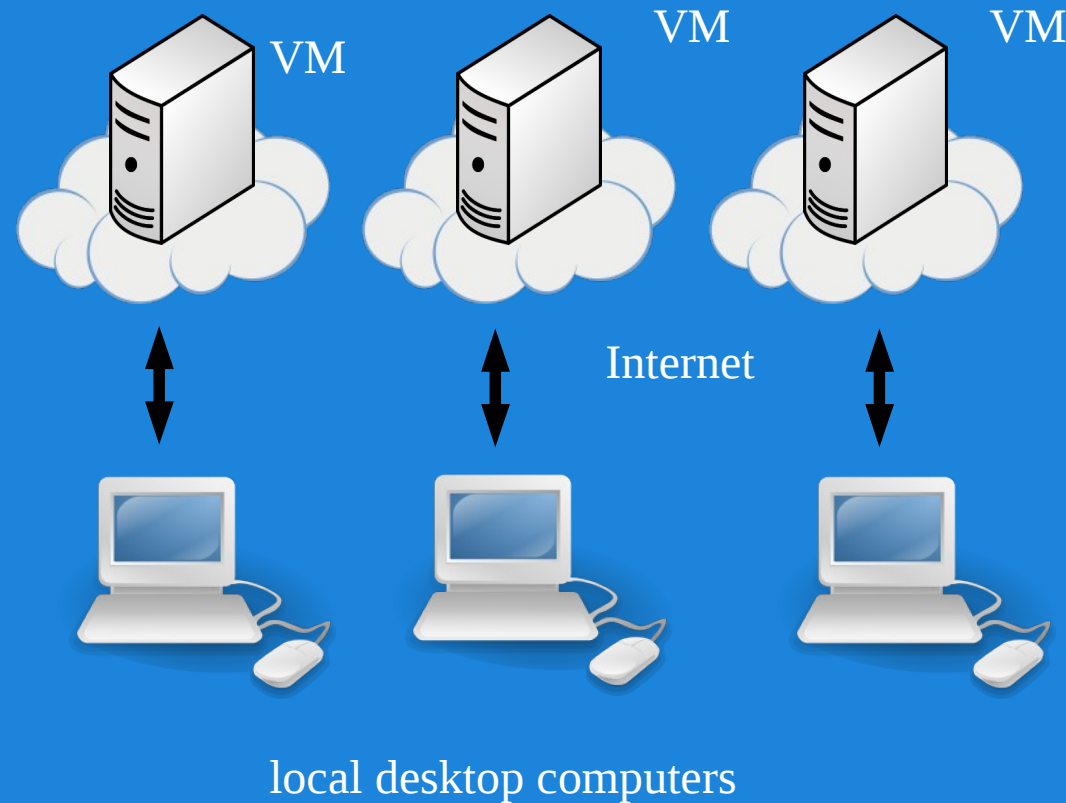
Virtual Machine



How does cloud computing work ?

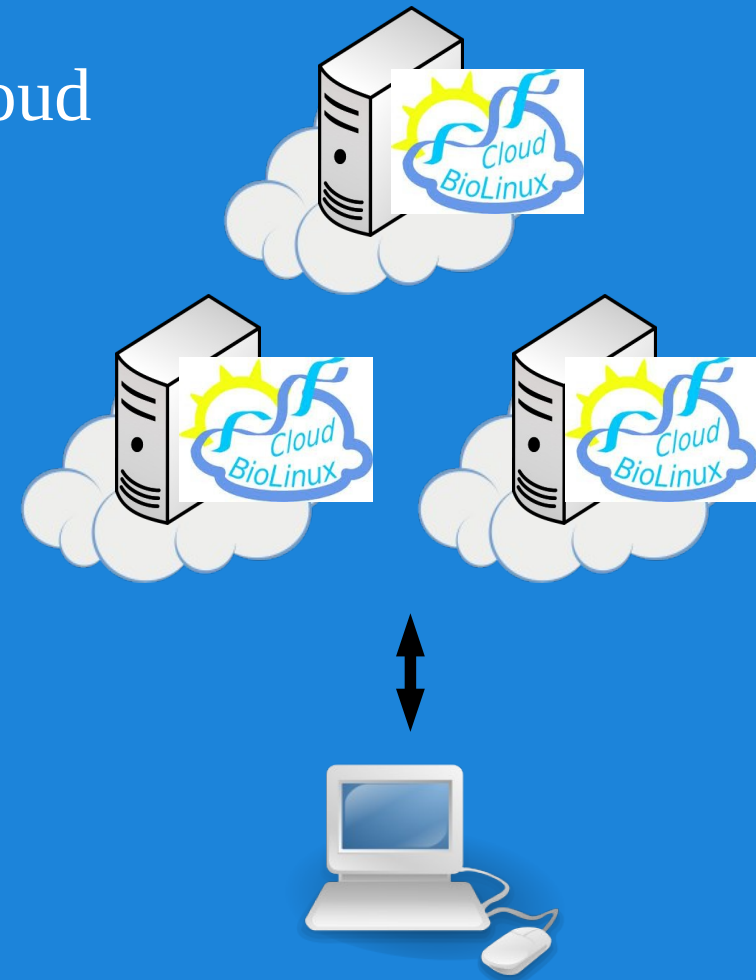
- a VM is uploaded on the cloud service; runs by renting computing capacity from Amazon EC2
- bioinformatics software can be executed from anywhere in the world through a desktop computer with Internet access
- removes need for local computer clusters at each laboratory
- alternatively if you have a cluster locally it can run on a private cloud

remote Amazon EC2 cloud computing service



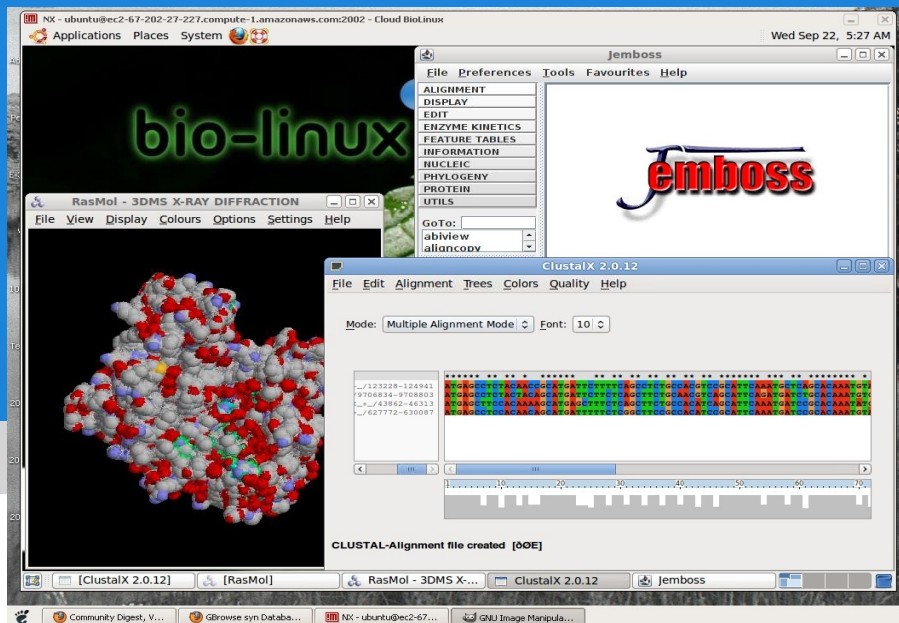
Solving problem 2: pre-installed and configured bioinformatics software on cloud Virtual Machines

- Cloud BioLinux: a publicly accessible Virtual Machine (VM) on the Amazon EC2 cloud
- 100+ pre-configured and installed bioinformatics software tools
- sequence analysis, genome assembly, annotation, phylogeny, molecular modeling, gene expression
- a researcher can initiate a practically unlimited number of VMs for large-scale data analysis and access them using a local desktop computer



Cloud BioLinux for Bioinformatics

- how the Cloud BioLinux project came to be, what it can offer to small labs for genome sequence analysis
- where and how do I run Cloud BioLinux, especially if I am not a computer expert
- besides end-users, bioinformatics developers are provided a framework for modifying and sharing VM configurations and data



Before we go on, a short break for questions...

Creating Cloud Biolinux



tinyurl.com/BioLinux-NEBC

+



=



<http://www.cloudbiolinux.org>

- JCVI bioinformatics cloud computing research
- NEBC BioLinux software repository
- community effort at BOSC 2009 – 11
- initially: a VM on Amazon EC2 with the tools copied and installed from the NEBC repository
- now: framework for creating customized cloud VMs
- major contributors:



Research at JCVI with Cloud BioLinux

- Eucalyptus private cloud currently installed at JCVI, OpenStack on the way
- open-source cloud platforms, fully compatible with Amazon EC2 (identical API)
- easy to set up on a local computer cluster, comes with Ubuntu server (UEC)
- develop VMs in-house with complex bioinformatics pipelines pre-installed and upload to Amazon EC2 for public access



Research at JCVI with Cloud BioLinux

- Funded by NIAID until 2013, focus on Viral sequencing-to-annotation data pipelines
- bioinformatics data analysis pipelines have complex dependencies: operating system, software libraries, reference databases etc.
- approach: pre-install pipelines and all dependencies in a single binary VM file using a private cloud
- upload VM on Amazon EC2: pipelines ready to execute, no need to purchase hardware
- benefits small laboratories that lack resources
- if you own a cluster: download and run VM on your private Eucalyptus or Openstack cloud



JCVI - GSC



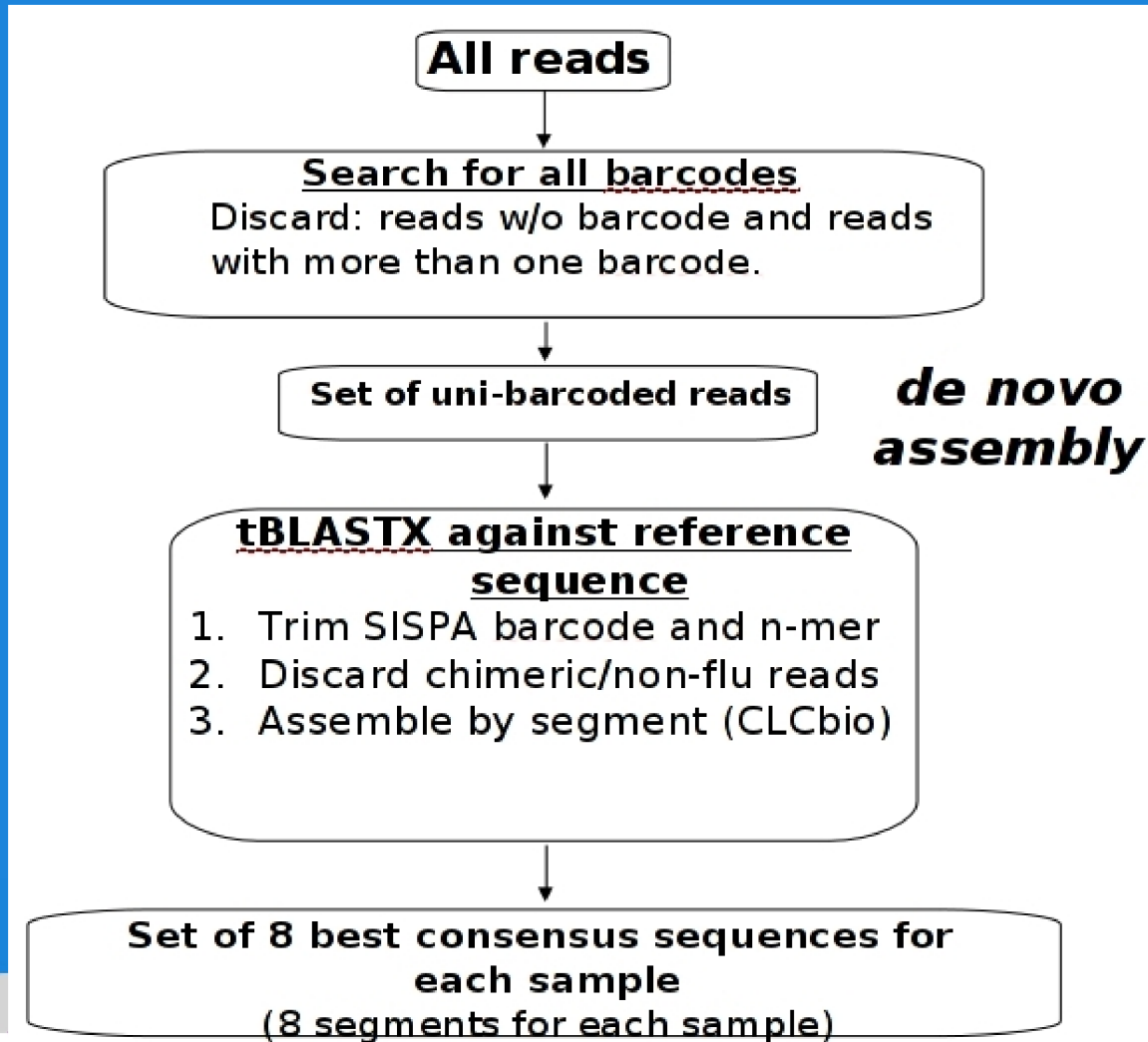
National Institute of Allergy and Infectious Diseases

Leading research to understand, treat, and prevent infectious, immunologic, and allergic diseases.

J. Craig Venter™
I N S T I T U T E

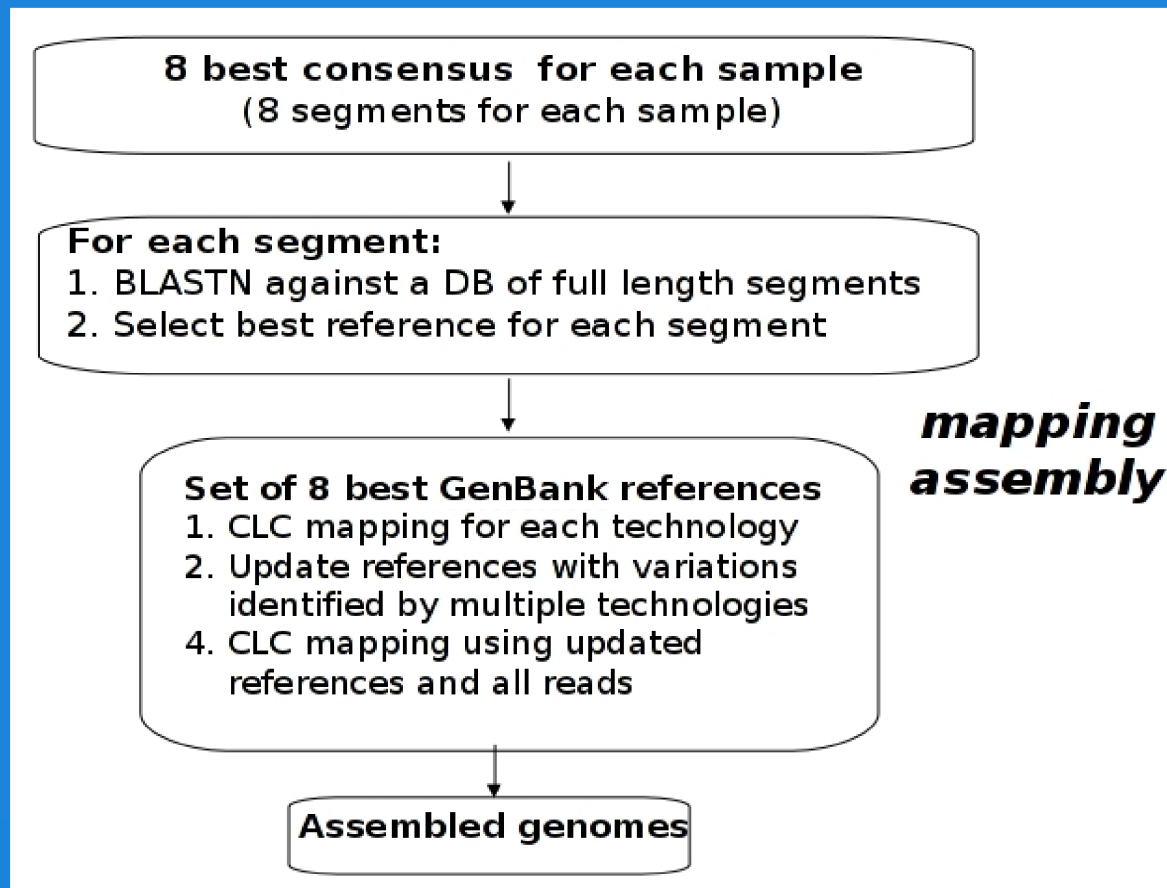
JCVI's Viral Genome Pipelines

Phase I, sequencing and assembly



JCVI's Viral Genome Pipelines

Phase I, sequencing and assembly



JCVI's Viral Genome Pipelines

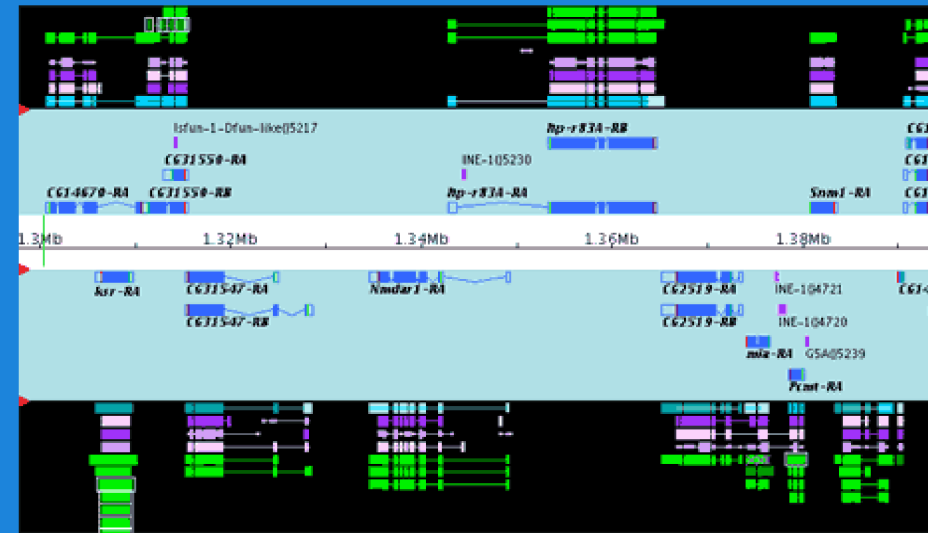
Phase II, annotation

- ▶ assembly as input to Viral Genome ORF Reader (VIGOR) pipeline
- ▶ detect coding regions, frame shifts, overlapping and embedded genes
- ▶ used for annotating the influenza virus, rotavirus, rhinovirus, coronavirus and subtypes

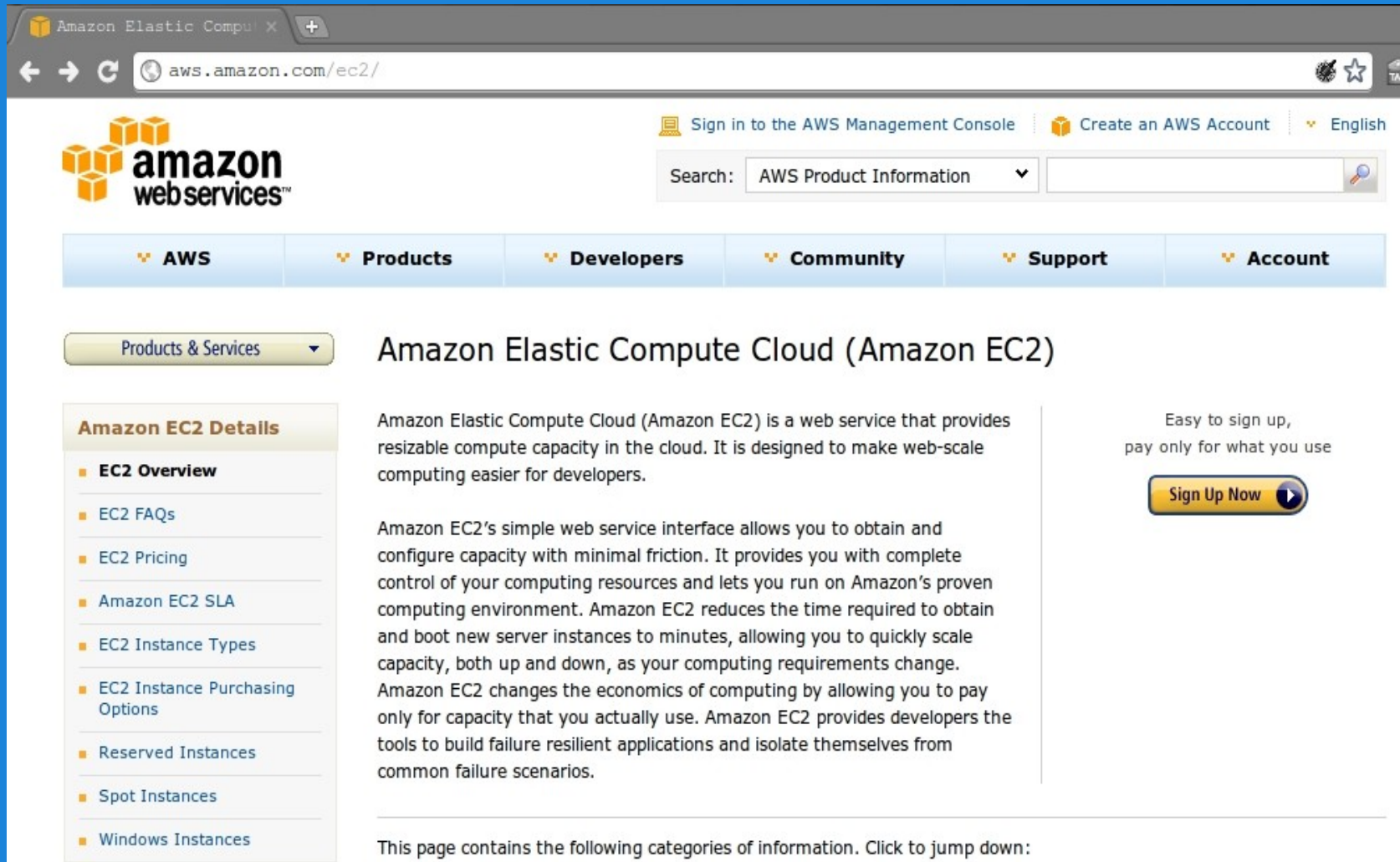
JCVI's Viral Genome Pipelines

Phase III, visualization and community annotation

- ▶ Apollo genome annotation viewer and editor
- ▶ used for collaborative annotation of the *Drosophila* genome
- ▶ fully integrated annotation tracks (ORFs, homologous genes, features etc.)



Running Cloud BioLinux on the Amazon EC2 cloud



The screenshot shows the Amazon Elastic Compute Cloud (Amazon EC2) product page. The browser address bar displays `aws.amazon.com/ec2/`. The page features the Amazon Web Services logo, navigation links for AWS, Products, Developers, Community, Support, and Account, and a search bar. The main content area is titled "Amazon Elastic Compute Cloud (Amazon EC2)" and includes a "Sign Up Now" button. A sidebar on the left lists "Amazon EC2 Details" with links to Overview, FAQs, Pricing, SLA, Instance Types, Purchasing Options, Reserved Instances, Spot Instances, and Windows Instances.

Products & Services

Amazon Elastic Compute Cloud (Amazon EC2)

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud. It is designed to make web-scale computing easier for developers.

Amazon EC2's simple web service interface allows you to obtain and configure capacity with minimal friction. It provides you with complete control of your computing resources and lets you run on Amazon's proven computing environment. Amazon EC2 reduces the time required to obtain and boot new server instances to minutes, allowing you to quickly scale capacity, both up and down, as your computing requirements change. Amazon EC2 changes the economics of computing by allowing you to pay only for capacity that you actually use. Amazon EC2 provides developers the tools to build failure resilient applications and isolate themselves from common failure scenarios.

Easy to sign up,
pay only for what you use

[Sign Up Now](#)

This page contains the following categories of information. Click to jump down:

- Amazon EC2 Details
 - EC2 Overview
 - EC2 FAQs
 - EC2 Pricing
 - Amazon EC2 SLA
 - EC2 Instance Types
 - EC2 Instance Purchasing Options
 - Reserved Instances
 - Spot Instances
 - Windows Instances

Account on the Amazon EC2 cloud <http://aws.amazon.com/ec2>

Launch Cloud BioLinux through the EC2 cloud console

The screenshot displays the AWS Management Console interface for the Amazon EC2 console dashboard. The browser address bar shows the URL <https://console.aws.amazon.com/ec2>. The page header includes the AWS logo, navigation links (Products, Developers, Community, Support, Account), and a welcome message for J. Craig Venter Institute.

The main content area is divided into several sections:

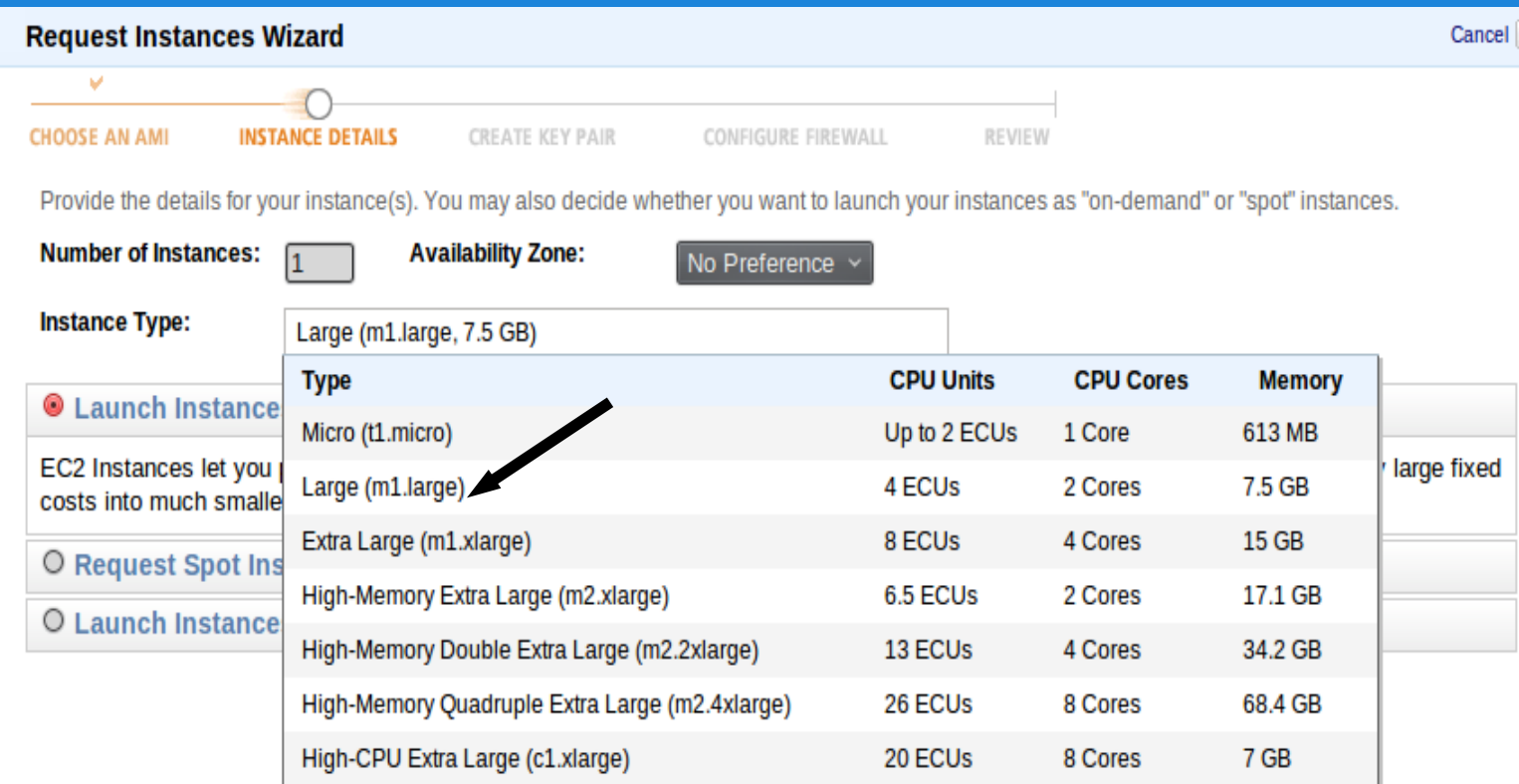
- Navigation:** A sidebar on the left with a region dropdown set to "US East" and a list of navigation items including EC2 Dashboard, INSTANCES (Instances, Spot Requests), IMAGES (AMIs, Bundle Tasks), ELASTIC BLOCK STORE (Volumes, Snapshots), and NETWORKING & SECURITY (Elastic IPs, Security Groups, Placement Groups, Load Balancers, Key Pairs).
- Amazon EC2 Console Dashboard:**
 - Getting Started:** A yellow box with the text "To start using Amazon EC2 you will want to launch a virtual server, known as an Amazon EC2 instance." and a "Launch Instance" button circled in a dashed line. A note below states: "Note: Your instances will launch in the US East (Virginia) region."
 - My Resources:** A section titled "You are using the following Amazon EC2 resources in the US East (Virginia) region:" with a "Refresh" button. It lists: 2 Running Instances, 0 Elastic IPs, 2 EBS Volumes, 4 EBS Snapshots, 3 Key Pairs, 5 Security Groups, 0 Load Balancers, and 0 Placement Groups.
 - Service Health:** A section titled "Service Health" with a table showing the current status of Amazon EC2 (US East - N. Virginia) as "RESOLVED" with the message "[RESOLVED] Increased tagging error rates". A link "View complete service health details" is provided.
 - Related Links:** A section with links to Documentation, All EC2 Resources, Forums, Feedback, and Report an Issue.

<http://tinyurl.com/cloud-biolinux-tutorial>

Cloud BioLinux launch wizard: steps 1 & 2



1. go to the “Community AMIs” tab, specify the Cloud BioLinux VM identifier (most recent update: cloudbiolinux.org)



2. select computational capacity

Cloud BioLinux launch wizard: step 3

Request Instances Wizard

CHOOSE AN AMI **INSTANCE DETAILS** CREATE KEY PAIR CONFIGURE FIREWALL REVIEW

Number of Instances: 1


Availability Zone: No Preference

Advanced Instance Options

Here you can choose a specific kernel or RAM disk to use with your instances. You can also choose to enable CloudWatch data that will be available from your instances once they launch.

Kernel ID: Use Default **RAM Disk ID:** Use Default

Monitoring: Enable CloudWatch detailed monitoring for this instance (additional charges will apply)

User Data: 

base64 encoded

Termination Protection: Prevention against accidental termination.

Shutdown Behavior: Stop Choose the behavior when the instance is shutdown from within the instance.

[< Back](#) [Continue >](#)

3. specify a password for login to Cloud BioLinux in the “User Data” box

new History Bookmarks Tools Help

Back Forward Home Shareaholic <https://console.aws.amazon.com/ec2> amazon.com x executable Print

ULike S3FoxBrowse Amazon S3

on.com AWS Products Developers Community Support Account Welcome, J. Craig Venter Institute Settings Sign Out

Amazon EC2 Amazon VPC Amazon Elastic MapReduce Amazon CloudFront Amazon RDS

My Instances

Launch Instance Instance Actions Reserved Instances Show/Hide Refresh Help

Viewing: All Instances All Instance Types 1 to 4 of 4 Instances

Instance	AMI ID	Root Dev	Type	Status	Lifecycle	Public D	Security	Key Pair	Moni
i-49201823	ami-6011e409	ebs	m1.large	pending	normal		default	jcvl_key1	disab
<input checked="" type="checkbox"/> i-f7340c9d	ami-6011e409	ebs	m1.large	running	normal	ec2-184-	default	jcvl_key1	disab
<input type="checkbox"/> i-795b6313	ami-6816e301	ebs	m1.large	terminated	normal				
<input type="checkbox"/> i-f9330b93	ami-6011e409	ebs	m1.large	terminated	normal				

1 EC2 Instance selected

EC2 Instance: i-f7340c9d

Description Monitoring Tags

AMI ID:	ami-6011e409	Zone:	us-east-1d
Security Groups:	default	Type:	m1.large
Status:	running	Owner:	767506454313
VPC ID:	-	Subnet ID:	-
Virtualization:	paravirtual	Placement Group:	
Reservation:	r-6ec5cf05	RAM Disk ID:	-
Platform:	-	Key Pair Name:	jcvl_key1
Kernel ID:	aki-0b4aa462	Monitoring:	disabled
AMI Launch Index:	0	Elastic IP:	-
Root Device:	/dev/sda1	Root Device Type:	ebs
Block Devices:	/dev/sda1=vol-68cce401:attached:2010-09-22T22:57:42.000Z:true		
Lifecycle:	normal		
Public DNS:	ec2-184-73-27-151.compute-1.amazonaws.com		
Private DNS:	ip-10-245-207-16.ec2.internal		
Private IP Address:	10.245.207.16		

remote desktop client

NOMACHINE

General Advanced Services Environment

Server

Host Port

Remember my password

Desktop

MODEM ISDN ADSL WAN LAN

Display

W H

Use custom settings

- Applications
- Places
- System
- Accessories >
- Bioinformatics >
- Graphics >
- Internet >
- Programming >
- Science >
- Sound & Video >
- System Tools >
- Ubuntu Software Center
- CLC Sequence Viewer
- clustalw
- clustalx
- Cn3D
- dendroscope
- entrez
- Exchanger
- fastDNAmI
- Forester ATV
- gap4
- jalview
- Jemboss
- maxdLoad2
- mesquite
- Mr Bayes Multi
- oligoarray
- omegamap
- pfaat
- pregap4
- sequin
- sqint
- taverna
- TaxInspector
- tetra
- treeview
- trev

Jemboss

File Preferences Tools Favourites Help

ALIGNMENT
DISPLAY
EDIT
ENZYME KINETICS
FEATURE TABLES
INFORMATION
NUCLEIC
PHYLOGENY
PROTEIN
UTILS

GoTo:

- abiview
- aligncopy
- aligncopypair
- allversusall
- antigenic
- backtranambig
- backtranseq
- banana
- biosed
- btwisted
- cai
- cathparse
- chaos
- charge
- checktrans
- chips
- cirdna
- codcmp
- codcopy
- coderet
- compseq
- consambig

Keyword Search

AND OR

(No Current Jobs)



Distributing Data Analysis Results with Cloud BioLinux

The screenshot displays a desktop environment on a Cloud BioLinux instance. The window title bar reads "NX - ubuntu@ec2-67-202-27-227.compute-1.amazonaws.com:2002 - Cloud BioLinux". The system tray shows the date and time: "Wed Sep 22, 5:27 AM".

Open applications include:

- Jemboss**: A menu-driven interface with options like ALIGNMENT, DISPLAY, EDIT, ENZYME KINETICS, FEATURE TABLES, INFORMATION, NUCLEIC, PHYLOGENY, PROTEIN, and UTILS. The main display shows the "Jemboss" logo.
- RasMol - 3DMS X-RAY DIFFRACTION**: A 3D molecular model viewer showing a protein structure with atoms represented by colored spheres (red, white, blue, yellow, green).
- ClustalX 2.0.12**: A sequence alignment tool. The main window shows a multiple sequence alignment of DNA sequences. The alignment is displayed in a grid format with a scale bar at the bottom. The text "CLUSTAL-Alignment file created [d0E]" is visible at the bottom of the window.

The desktop background features the "bio-linux" logo in a stylized green font. The taskbar at the bottom shows icons for ClustalX 2.0.12, RasMol, and Jemboss.

Distributing Data Analysis Results with Cloud BioLinux

Whole System Snapshot Exchange

- how difficult is to share bioinformatics work on your computer with a collaborator ?
- capture the state of the computing system (OS + software), data, analysis results
- make VM snapshots: executable, binary file replica of the original VM
- distribute a VM snapshot with pre-installed software and data so collaborators can replicate, re-run, add to your data analysis
- a snapshot can be shared directly on the Amazon cloud, downloaded on a private cloud or run on desktop using virtualization software

Cloud BioLinux: whole system snapshot exchange

The screenshot shows the AWS Management Console interface. At the top, there are navigation links for 'aws.amazon.com', 'AWS', 'Products', 'Developers', 'Community', 'Support', and 'Account'. On the right, it says 'Welcome, J. Craig Venter Institute' with links for 'Settings' and 'Sign Out'. Below this, there are tabs for 'Amazon S3', 'Amazon EC2', 'Amazon VPC', 'Amazon Elastic MapReduce', 'Amazon CloudFront', and 'Amazon RDS'. The 'Amazon EC2' tab is selected.

The main content area is titled 'My Instances'. It includes a 'Launch Instance' button, 'Instance Actions' dropdown, and 'Reserved Instances' dropdown. There are also 'Show/Hide', 'Refresh', and 'Help' buttons. The 'Viewing:' dropdown is set to 'All Instances'. A search bar is present. Below this is a table of instances with columns: 'Instance', 'Root Dev', 'Type', 'Status', 'Lifecycle', 'Public D', 'Security', 'Key Pair', and 'Moni'. The table contains four rows of instance data. The 'Instance Actions' menu is open over the second row, with 'Create Image (EBS AMI)' highlighted. An arrow points to the 'running' status of the second instance.

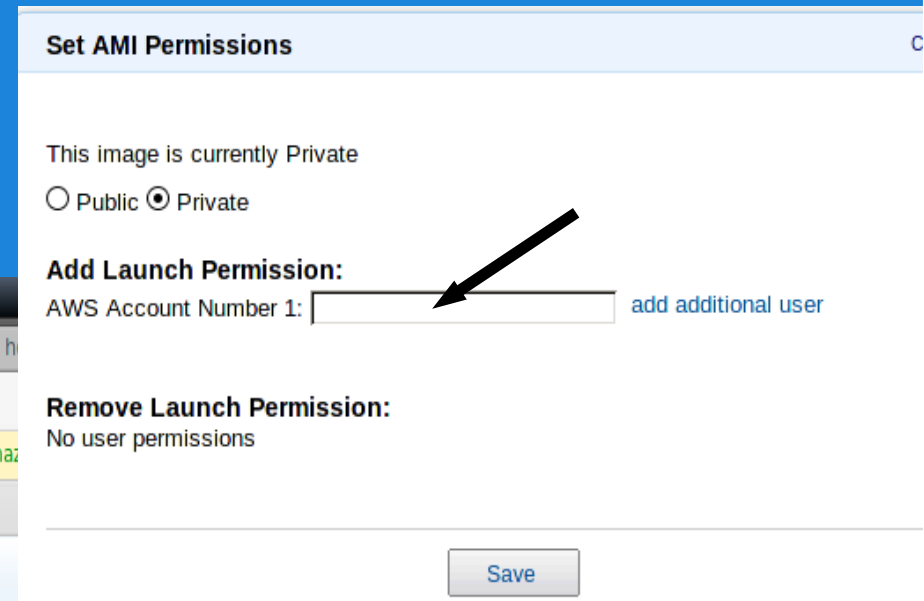
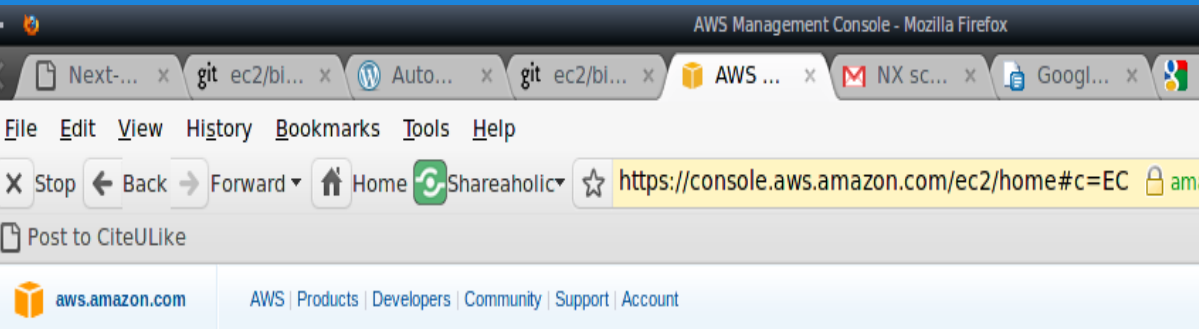
Instance	Root Dev	Type	Status	Lifecycle	Public D	Security	Key Pair	Moni
i-4920...	ebs	m1.large	running	normal	ec2-67-2...	default	jcvi_key...	disab
i-f7340...	ebs	m1.large	running	normal	ec2-184-	default	jcvi_key...	disab
i-795b...	ebs	m1.large	terminated	normal		default	jcvi_key...	disab
i-f9330...	ebs	m1.large	terminated	normal		default	jcvi_key...	disab

storage cost: 0.10\$ / GB / month

Cloud BioLinux: whole system snapshot exchange

authorize access to the VM: public or for certain users

other researchers can access the VM with all the software, data, analysis results directly on the cloud



Amazon S3 Amazon EC2 Amazon VPC Amazon Elastic MapReduce Amazon CloudFront Amazon RDS

Navigation

Region: US East

- EC2 Dashboard
- INSTANCES
 - Instances
 - Spot Requests
- IMAGES
 - AMIs
 - Bundle Tasks
- ELASTIC BLOCK STORE
 - Volumes
 - Snapshots

Amazon Machine Images

Launch Spot Request Register New AMI De-register Permissions Show/Hide Refresh Help

Viewing: Owned By Me All Platforms 1 to 7 of 7 AMIs

Name	AMI ID	Source	Owner	Visibility	Status	Platf
<input type="checkbox"/>	ami-6011e409	767506454313/Cloud Biolinux with FreeNX 09_2010	767506454313	Public	available	C
<input checked="" type="checkbox"/>	ami-6816e301	767506454313/Cloud Biolinux 6.0 with FreeNX remote desktop	767506454313	Private	available	C
<input type="checkbox"/>	ami-6953b200	jcvicloudbiolinux/JCVI-Cloud-BioLinux.m	767506454313	Public	available	C
<input type="checkbox"/>	ami-7c27cd15	767506454313/VICSTEST	767506454313	Private	available	C
<input type="checkbox"/>	ami-8815e0e1	767506454313/CloudBiolinux-09-2010	767506454313	Public	available	C
<input type="checkbox"/>	ami-8816e3e1	767506454313/CloudBiolinux-FreeNX-092010	767506454313	Private	failed	C
<input type="checkbox"/>	ami-fa16e393	767506454313/CloudBiolinux-09-2010 FreeNX	767506454313	Private	failed	C

Launch Instance Request Spot Instance Add/Edit Tags Edit Permissions De-register AMI



*5 min of questions, and then
5 more min to close the session....*

Cloud BioLinux for Software Developers

- Issue 1: for researchers with sensitive data a public cloud might not be an option
- Problem 1: moving VMs across clouds is not trivial, need low level operations

- Issue 2: bioinformatic specializations (ex. sequencing, phylogeny, protein structure)
- Problem 2: one VM to fit all becomes over-sized

- Cloud BioLinux VM deployment framework

Cloud BioLinux for Software Developers

- framework to describe software components in cloud VM / image
- based on python Fabric automated deployment tool
- software components listed in simple text files
- edit the files to mix and match software according to your community needs
- community members use files to share descriptions of customized systems
- start with a bare-bones VM on Amazon EC2 or Eucalyptus private cloud
- Fabric scripts download and install specified software



Free, available from: <https://github.com/chapmanb/cloudbiolinux>

100644 | 39 lines (38 sloc) | 0.668 kb

```
1 ---
2 # Top level configuration file that specifies w
3 # should be installed. New sections that are ad
4 # files should go here. Comment out any groups
5 # installed.
6 packages:
7   - desktop
8   - programming
9   - distributed
10  - amazon
11  - python
12  - r
13  - ruby
14  - perl
15  - java
16  - erlang
17  - haskell
18  - databases
19  - math
20  - viz
21  - web
22  - bio_general
23  - bio_search
24  - bio_alignment
25  - bio_nextgen
26  - bio_sequencing
27  - bio_annotation
28  - bio_microarray
29  - bio_visualization
30  - bio_utils
31  - phylogeny
```

software domains in Cloud BioLinux:

Genome sequencing, *de novo* assembly, annotation, phylogeny, molecular structures, gene expression analysis

high-level configuration describing software groups for each group individual bioinformatics tools

```
516 - apache2
517 bio_general:
518   - emboss
519   - emboss-data
520   - emboss-lib
521   - primer3
522   - readseq
523   - bio-linux-taverna
524   - bio-linux-xcut
525 bio_search:
526   - blast2
527   - hmmer
528   - ncbi-tools-bin
529   - bio-linux-blast+
```



Team

- ▶ Ntino Krampis - PI, Cloud BioLinux
- ▶ Andrey - coPI, NSF-Microsoft Cloud Research grant
- ▶ Tim Stockwell - Investigator Viral Group at JCVI
- ▶ Ramana Madupu - Investigator, Genomic Annotation experience
- ▶ Tim Prindle - Soft. Engineer, 20+ years
- ▶ Alex Richter - Soft. Engineer, 15+ years
- ▶ Systems Administrator - JCVI's Eucalyptus cloud and VMs
- ▶ **Funding:**
- ▶ Maria Giovanni - NIAID
- ▶ Punam Mathur - NIAID
- ▶ Karen Nelson - JCVI

Acknowledgments & Credits

Brad Chapman - development of the Fabric scripts, website

Tim Booth, Mesude Bicak, Dawn Field - BioLinux 6.0 development

Enis Afgan - Cloudman and Cloud BioLinux integration

Members of the Cloud Biolinux community:

<http://groups.google.com/group/cloudbiolinux>

And again our contacts:

kkrampis@jcvl.org

<http://www.cloudbiolinux.org>

Thank you !