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

Ntinos Krampis
Asst. Professor
J. Craig Venter Institute
kkrampis@jcvi.org

http://www.jcvi.org/cms/about/bios/kkrampis/

J. Craig Venter™
INSTITUTE

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

J. Craig Venter™ INSTITUTE

A little bit of background information...

- Konstantinos (Ntinos) 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@jcvi.org*

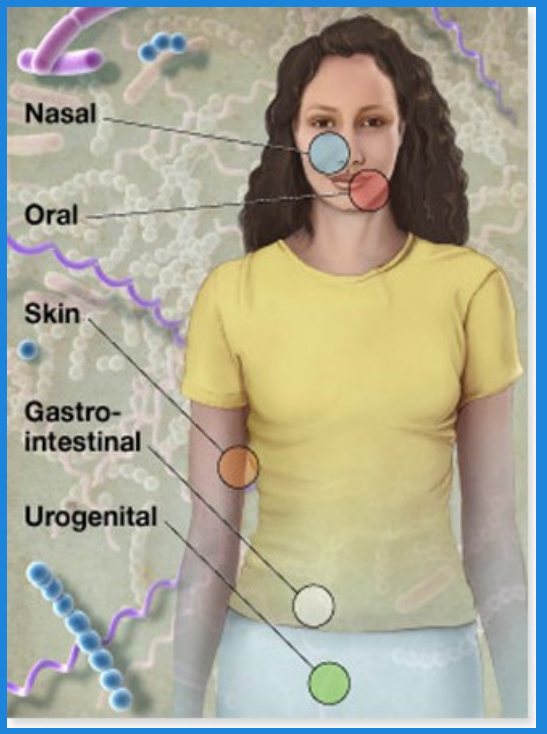
J. Craig Venter™ INSTITUTE

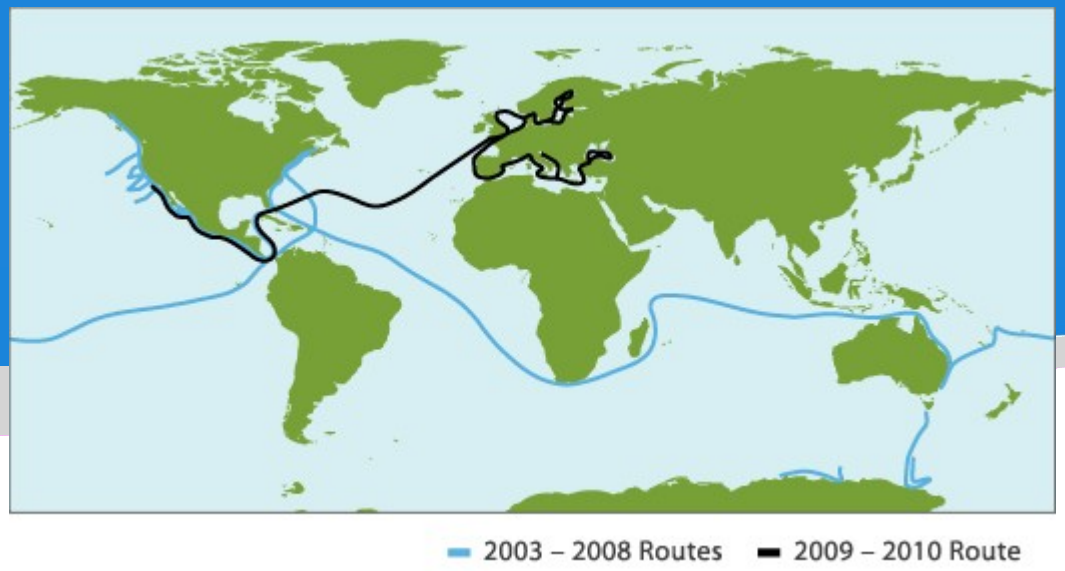
J. Craig Venter Institute (JCVI)

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





J. Craig Venter™
INSTITUTE

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

J. Craig Venter™
INSTITUTE

Low-cost sequencing instruments

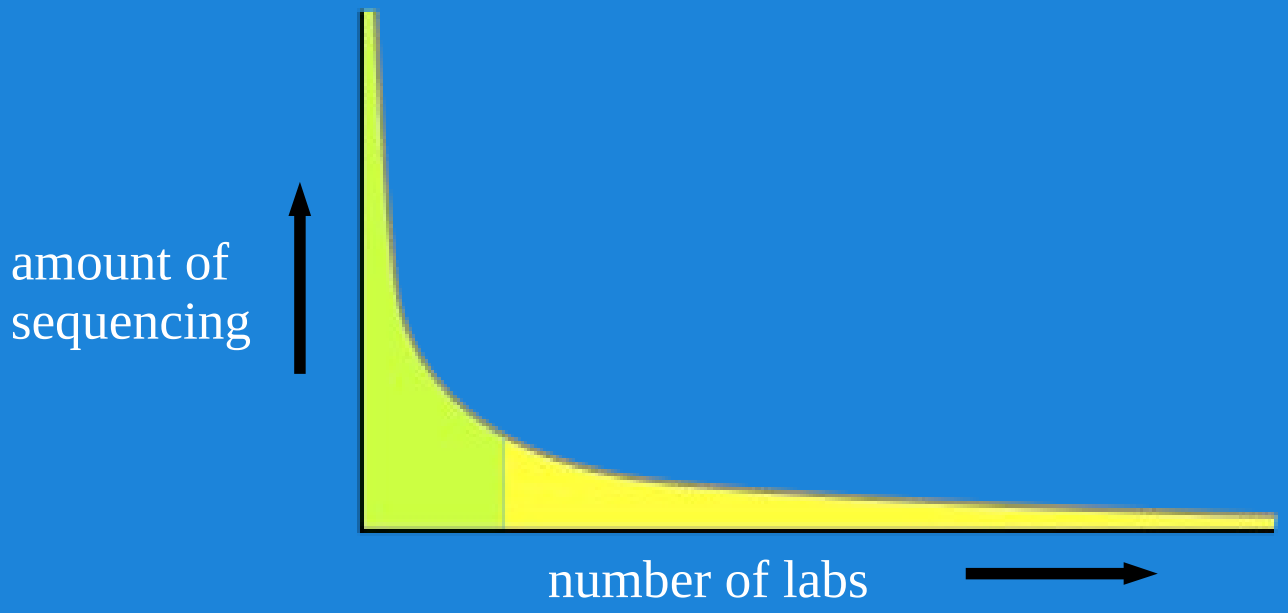
- small-factor sequencers available: GS Junior by 454, MiSeq by Illumina
- bacterial, viral, small fungal genomes, sequencing for variant discovery
- sequencing as a standard technique in molecular biology and genetics
- RNAseq (instead of microarrays) and ChiPseq (instead of yeast 2-hybrid)

http://www.gsjunior.com/

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

J. Craig Venter™ INSTITUTE

More small laboratories doing genome sequencing



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

J. Craig Venter™
INSTITUTE

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.

J. Craig Venter INSTITUTE

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

J. Craig Venter™ INSTITUTE

Large sequencing centers offering bioinformatics analysis services ?

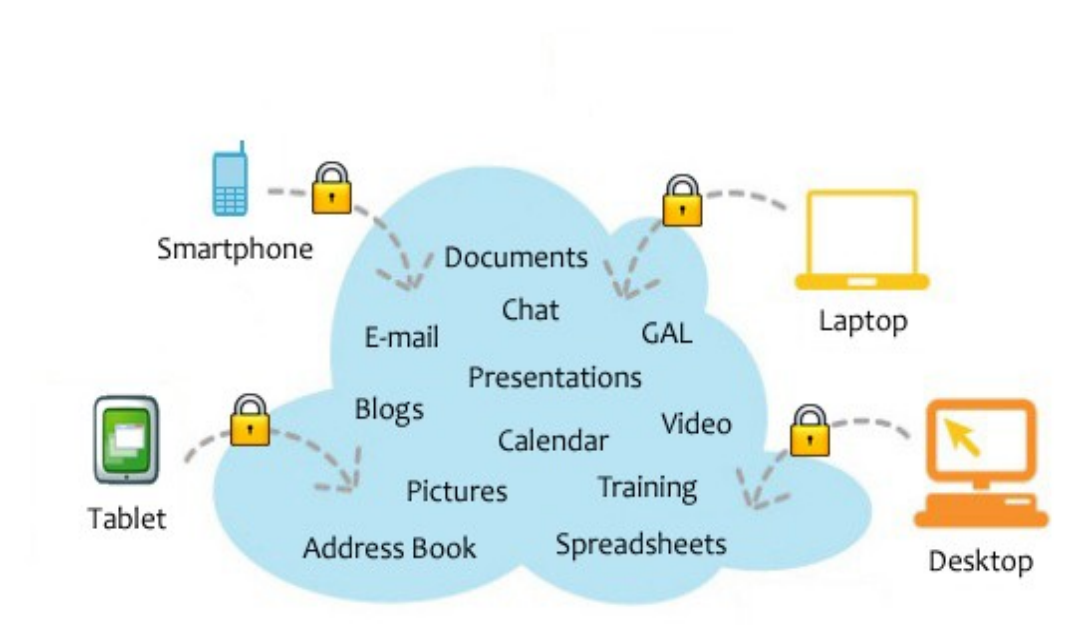
- Bioinformatic Resource Centers (BRC)
- bioinformatic analysis coupled with sequencing of an organism
- mostly provide data browsing and few analysis tools to the public
- cannot serve the bioinformatic needs of every small lab acquiring a sequencing instrument
- need end-to-end solutions, users submit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



J. Craig Venter INSTITUTE

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/

J. Craig Venter™
INSTITUTE

How does cloud computing work ?

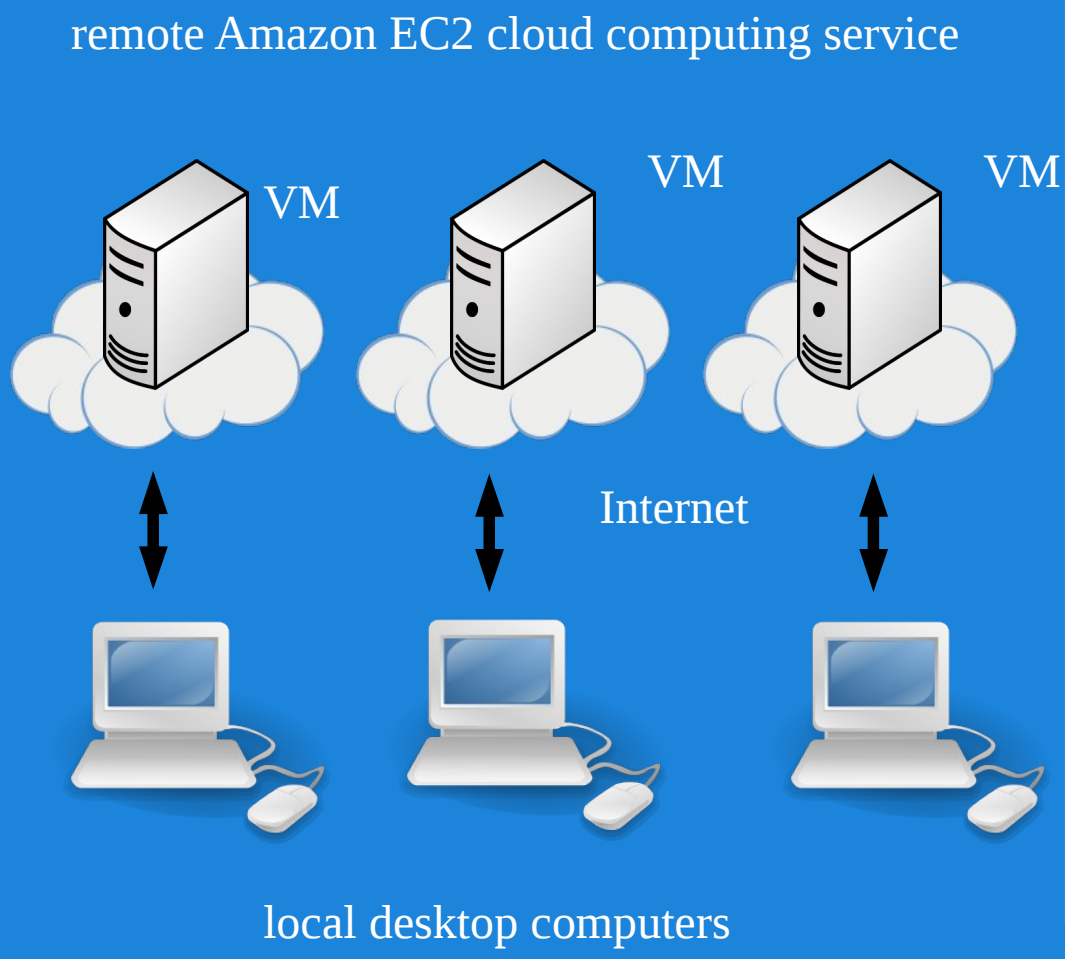
- Cloud computing evolved from Virtualization technology
- operating system, bioinformatics software and data, are installed on a Virtual Machine (VM)
- a VM is a full-featured Unix server, in a single, executable binary file
- no need to compile source code, set up configuration files, software installation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

Virtual Machine

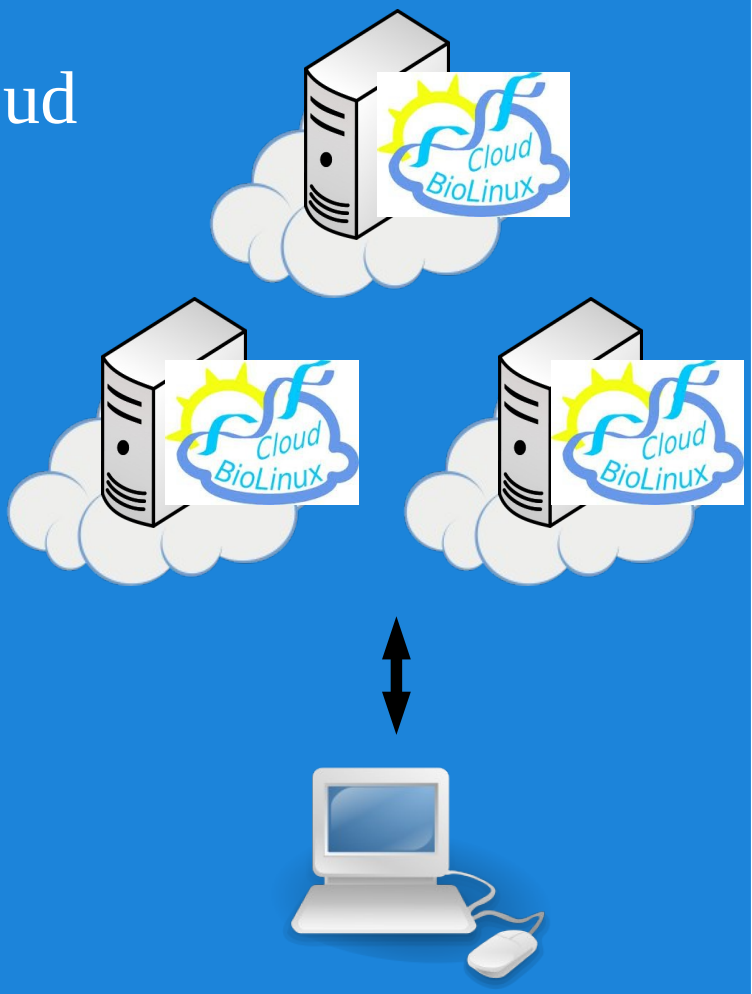
J. Craig Venter™ INSTITUTE

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



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

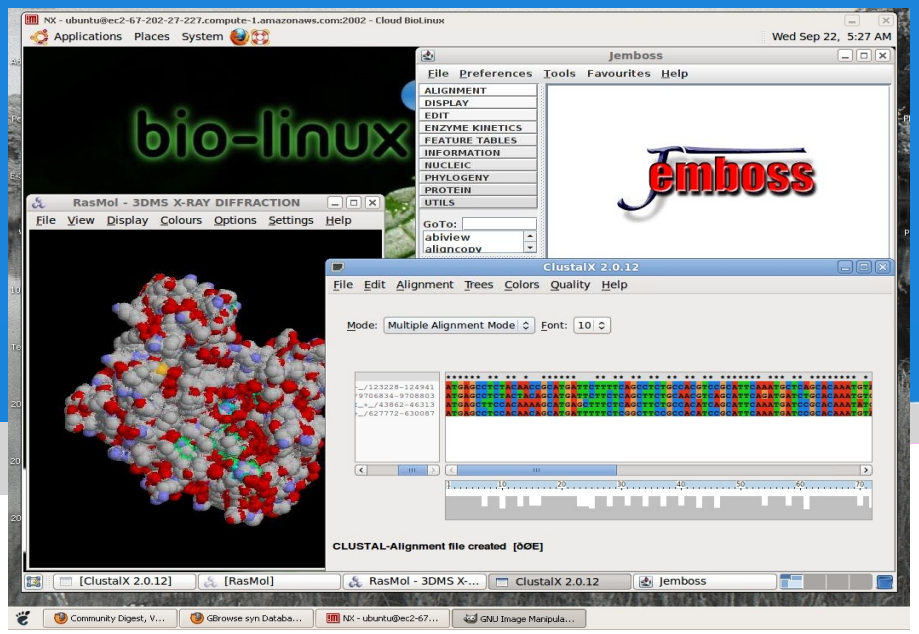
J. Craig Venter™ INSTITUTE

Cloud BioLinux for Bioinformatics

- how the Cloud BioLinux project came to be, what it can offers 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...

J. Craig Venter INSTITUTE

Creating Cloud Biolinux

tinyurl.com/BioLinux-NEBC

+

amazon web services

=

Cloud BioLinux

- 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:

HARVARD SCHOOL OF PUBLIC HEALTH

Galaxy

debian

http://www.cloudbiolinux.org

J. Craig Venter INSTITUTE

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

J. Craig Venter™
INSTITUTE

Research at JCVI with Cloud BioLinux

- Funded by NIAID until 2013, focus on Viral sequencing-to-annotation data pipelines

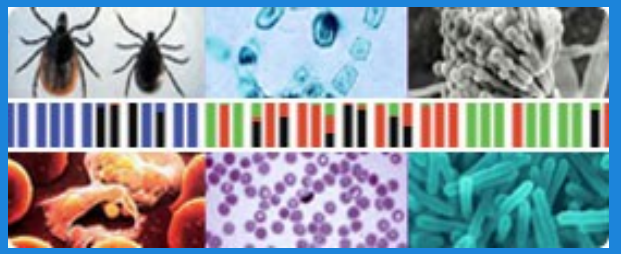
- bioinformatics data analysis pipelines have complex dependencies: operating system, software libraries, reference databases etc.

- 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



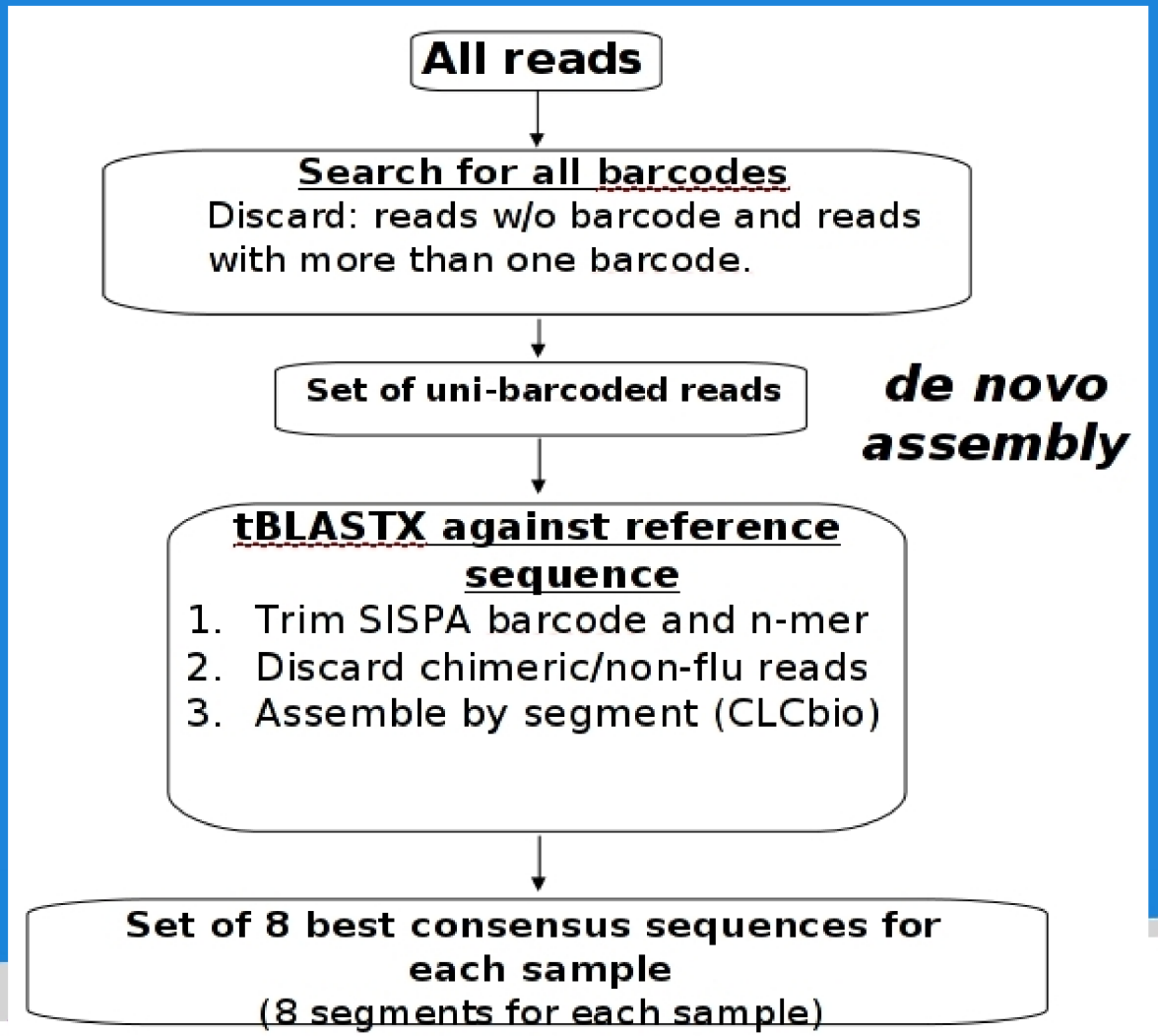
JCVI - GSC

National Institute of Allergy and Infectious Diseases
Leading research to understand, treat, and prevent infectious, immunologic, and allergic diseases.

J. Craig Venter™ INSTITUTE

JCVI's Viral Genome Pipelines

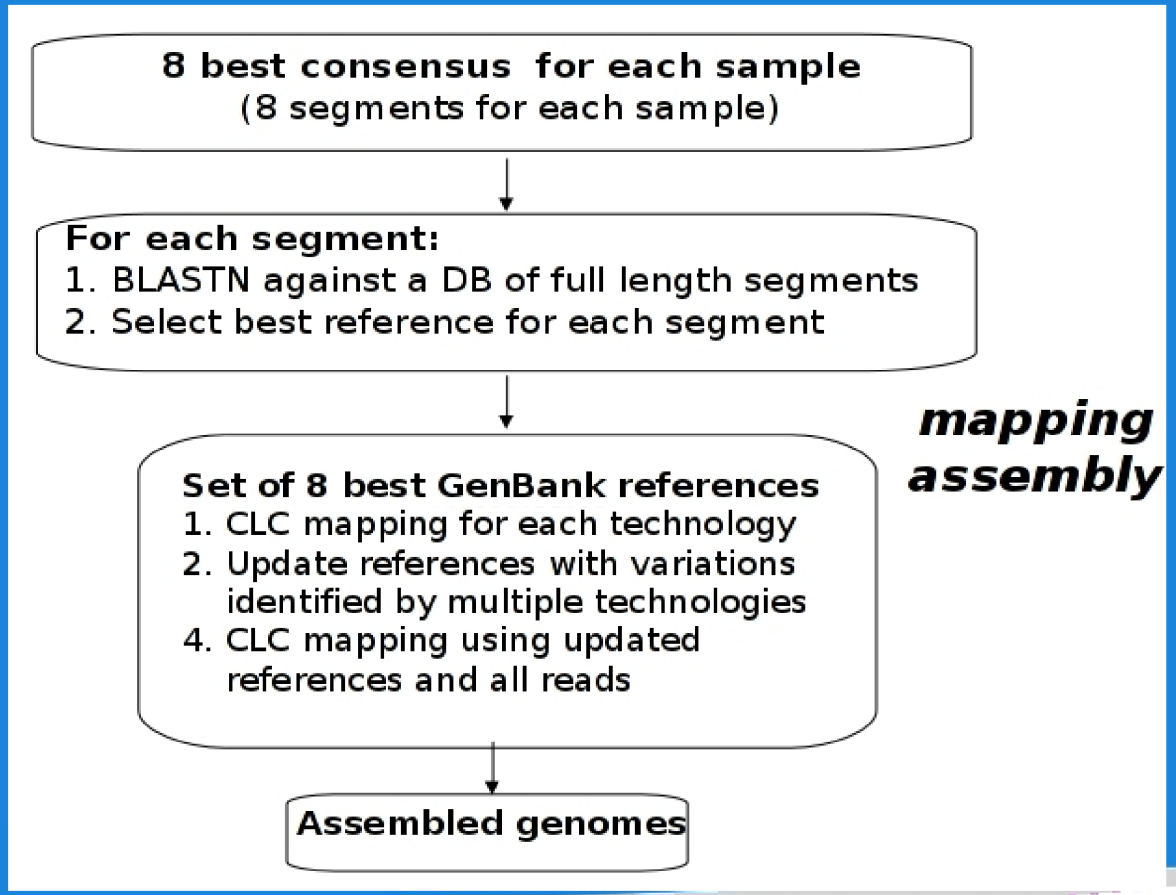
Phase I, sequencing and assembly



J. Craig Venter INSTITUTE

JCVI's Viral Genome Pipelines

Phase I, sequencing and assembly



J. Craig Venter INSTITUTE

JCVI's Viral Genome Pipelines

Phase II, annotation

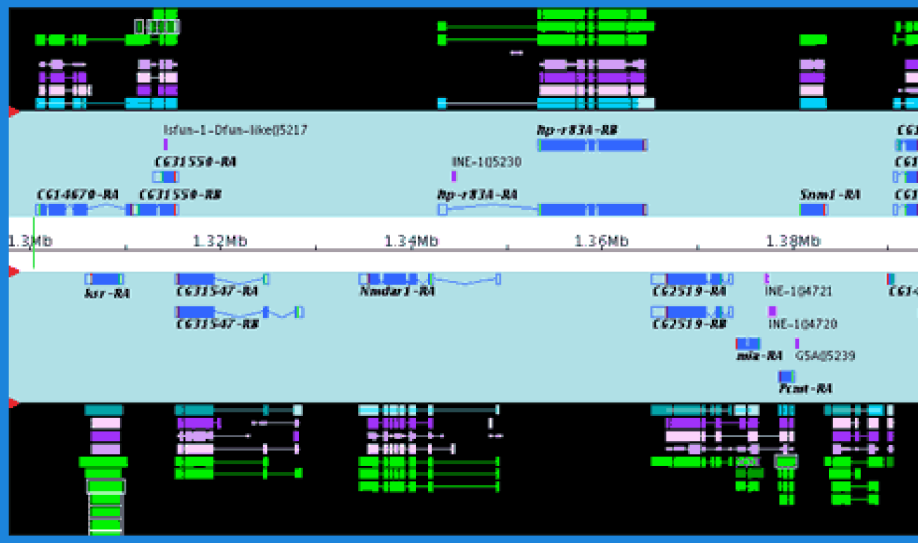
- assembly as input to Viral Genome ORF Reader (VIGOR) pipeline
- detect coding regions, frame shifts, overlapping and embedded genes
- used for annotating the influenza virus, rotavirus, rhinovirus, coronavirus and subtypes

J. Craig Venter™
INSTITUTE

JCVI's Viral Genome Pipelines

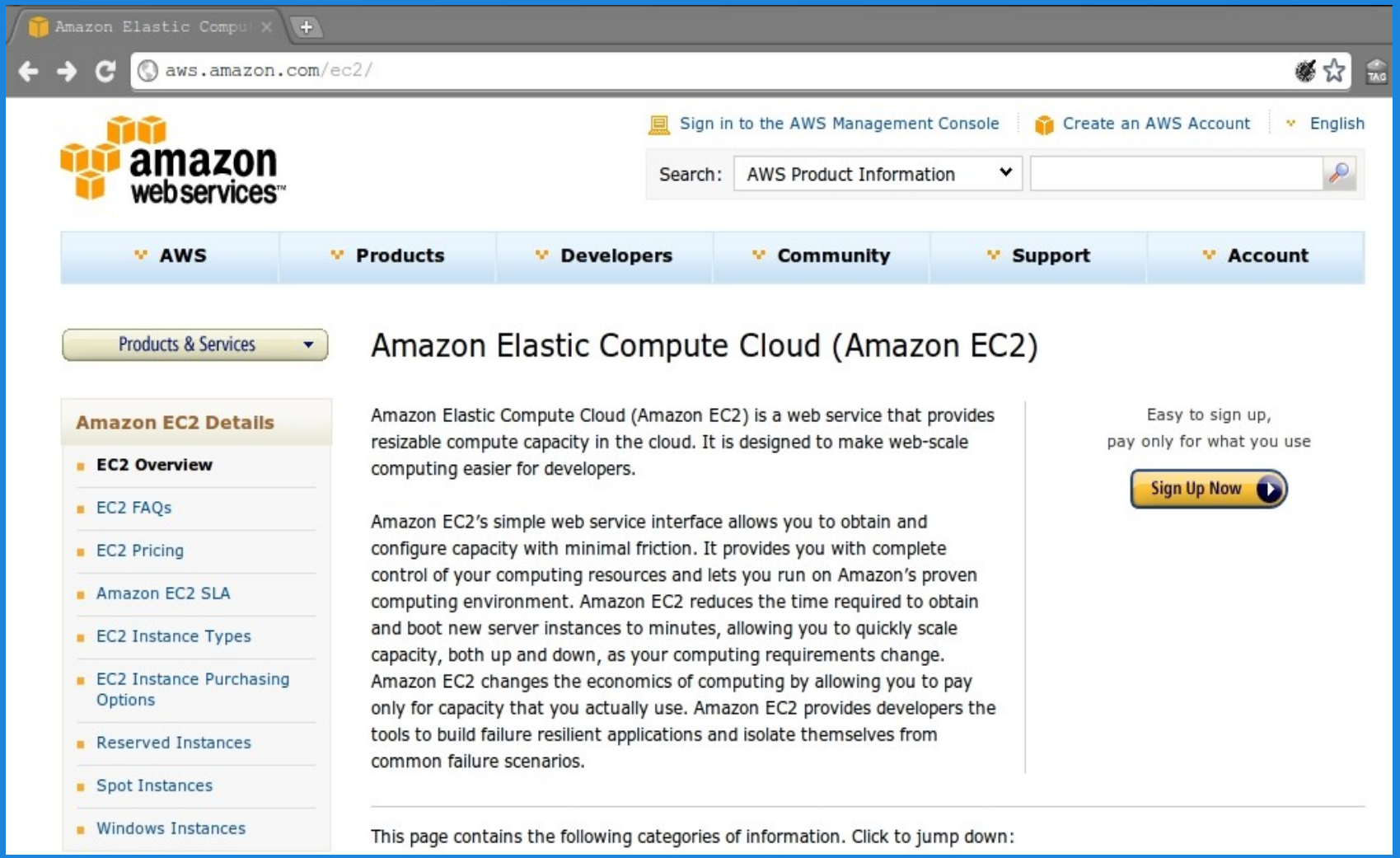
Phase III, visualization and community annotation

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



J. Craig Venter INSTITUTE

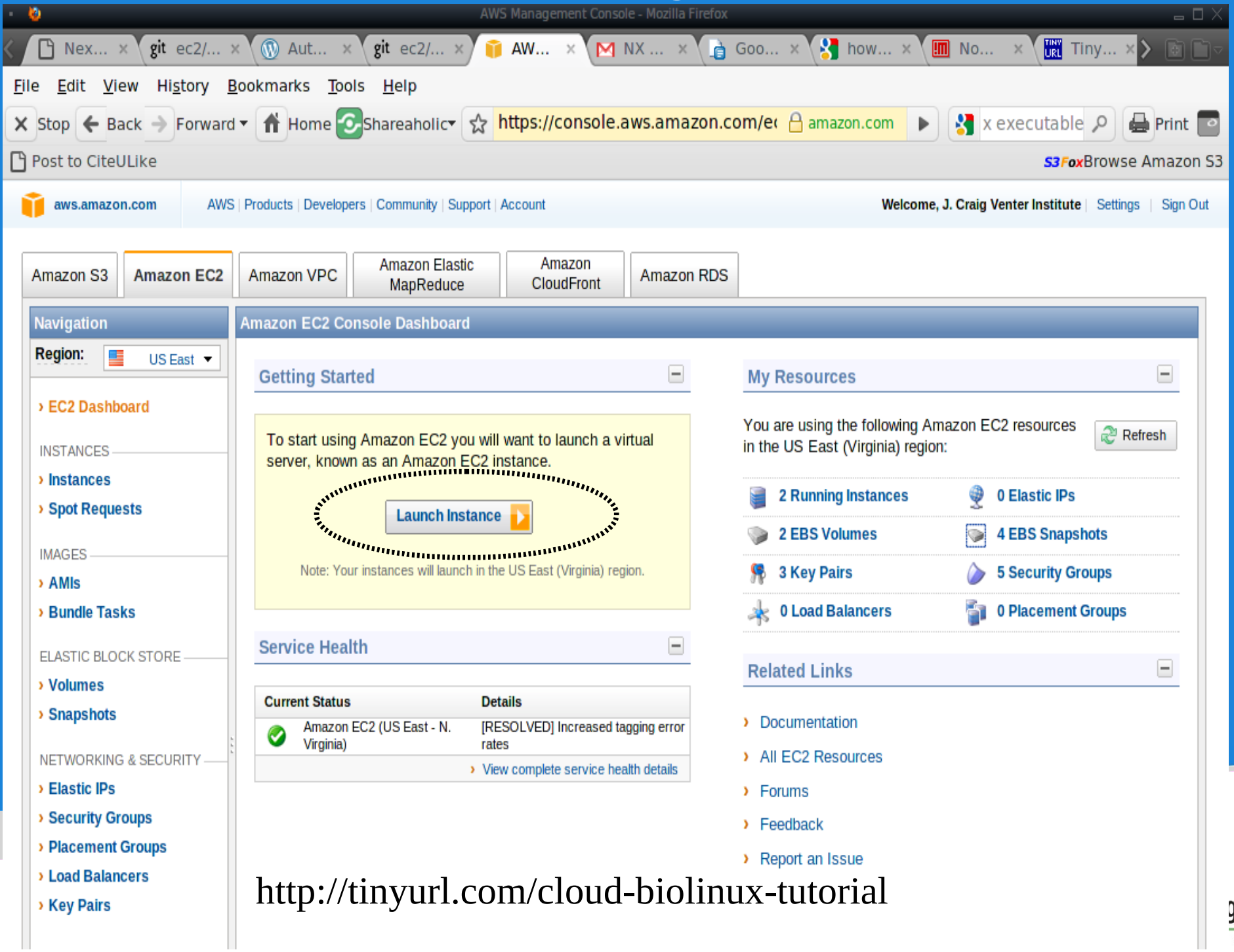
Running Cloud BioLinux on the Amazon EC2 cloud



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

J. Craig Venter INSTITUTE

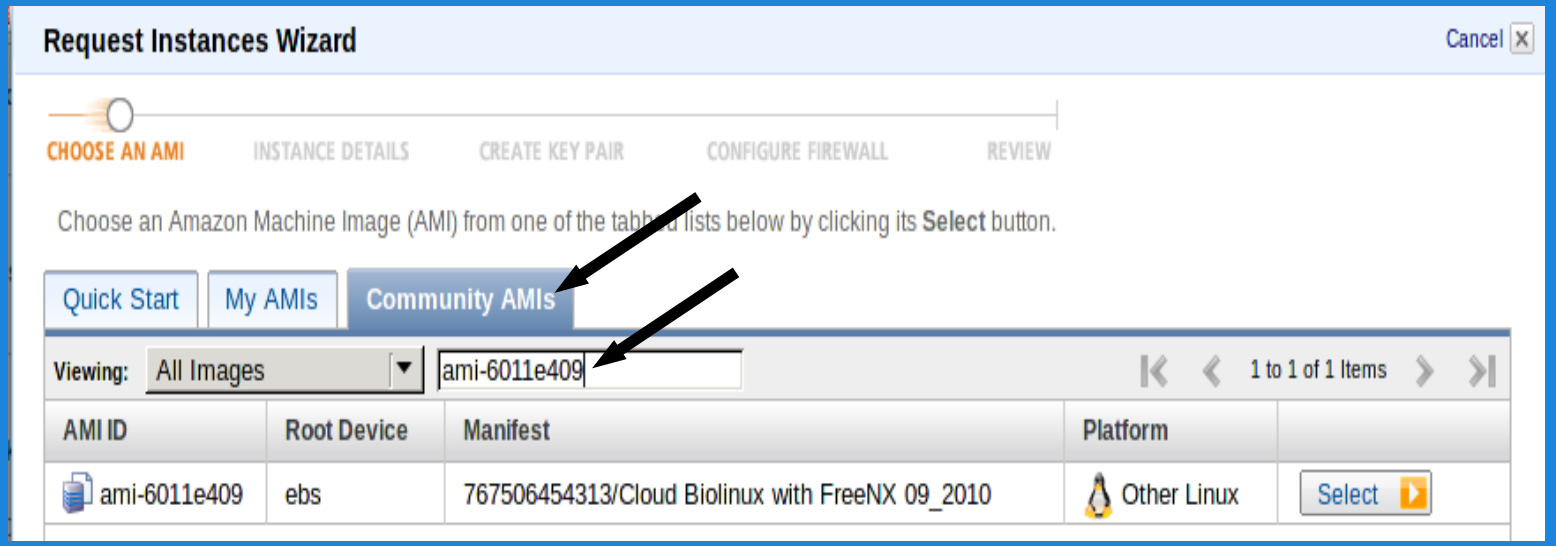
Launch Cloud BioLinux through the EC2 cloud console



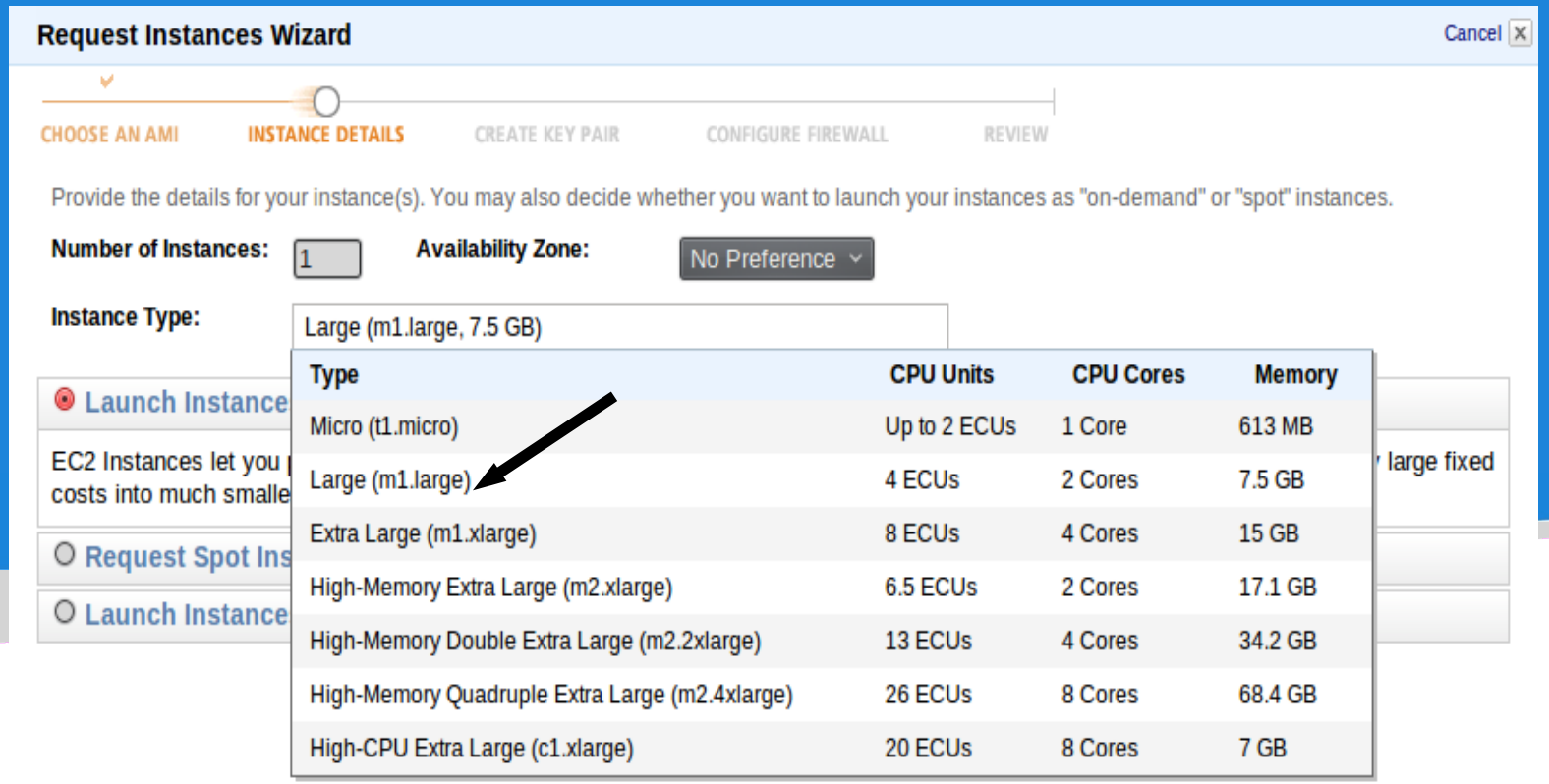
http://tinyurl.com/cloud-biolinux-tutorial

g Venter
ITUTE

Cloud BioLinux launch wizard: steps 1 & 2



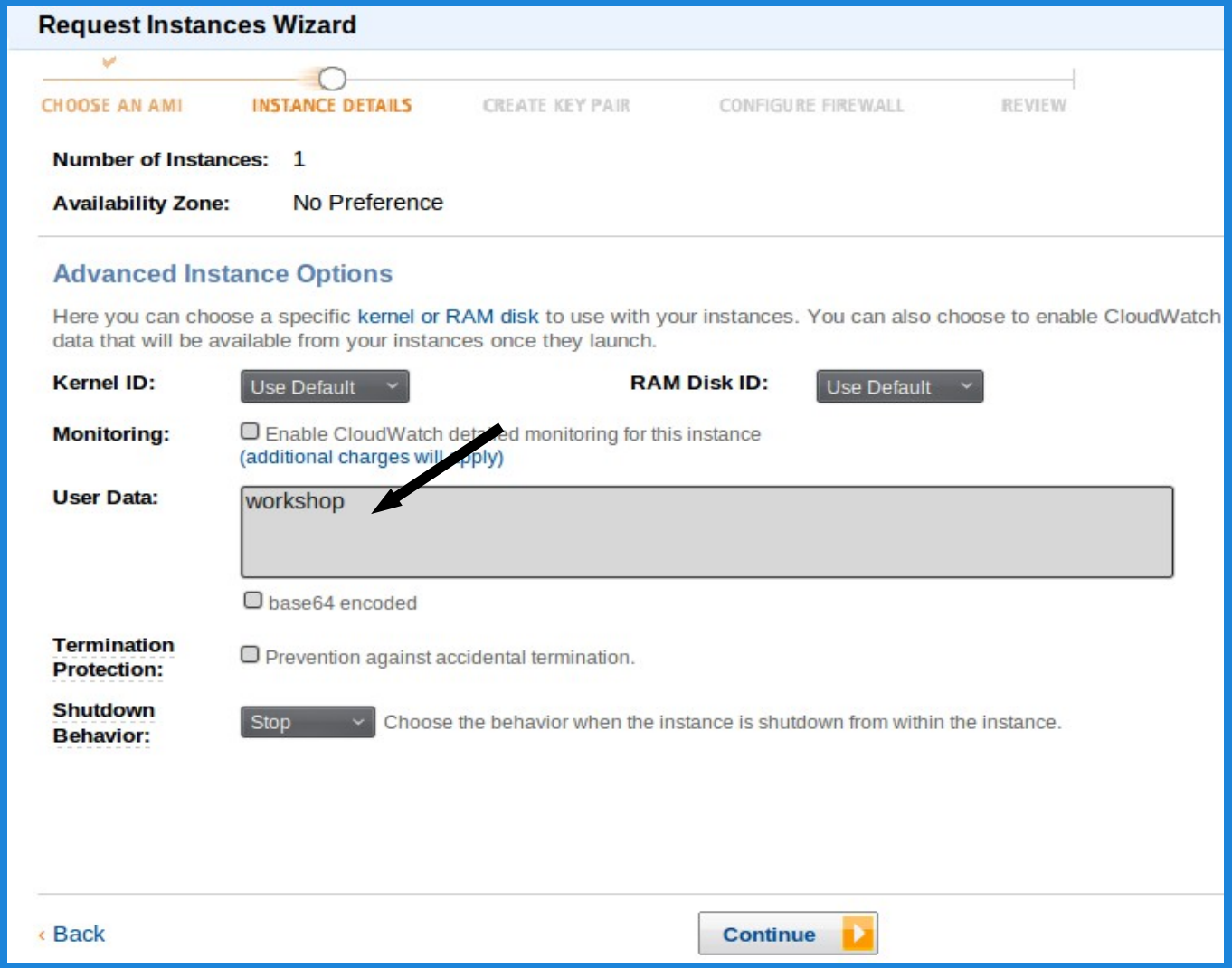
1. go to the "Community AMIs" tab, specify the Cloud BioLinux VM identifier (most recent update: cloudbiolinux.org)



2. select computational capacity

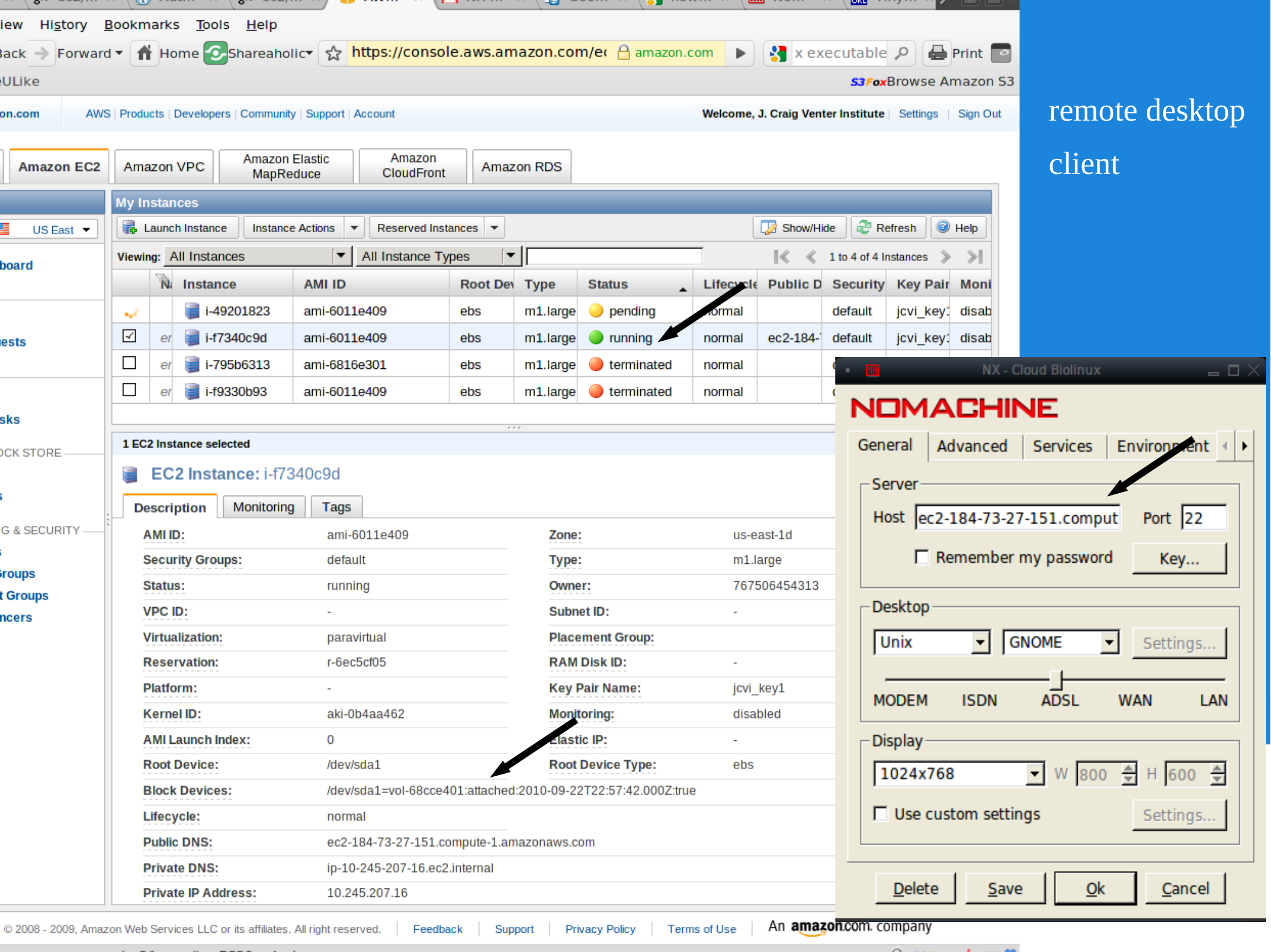
J. Craig Venter INSTITUTE

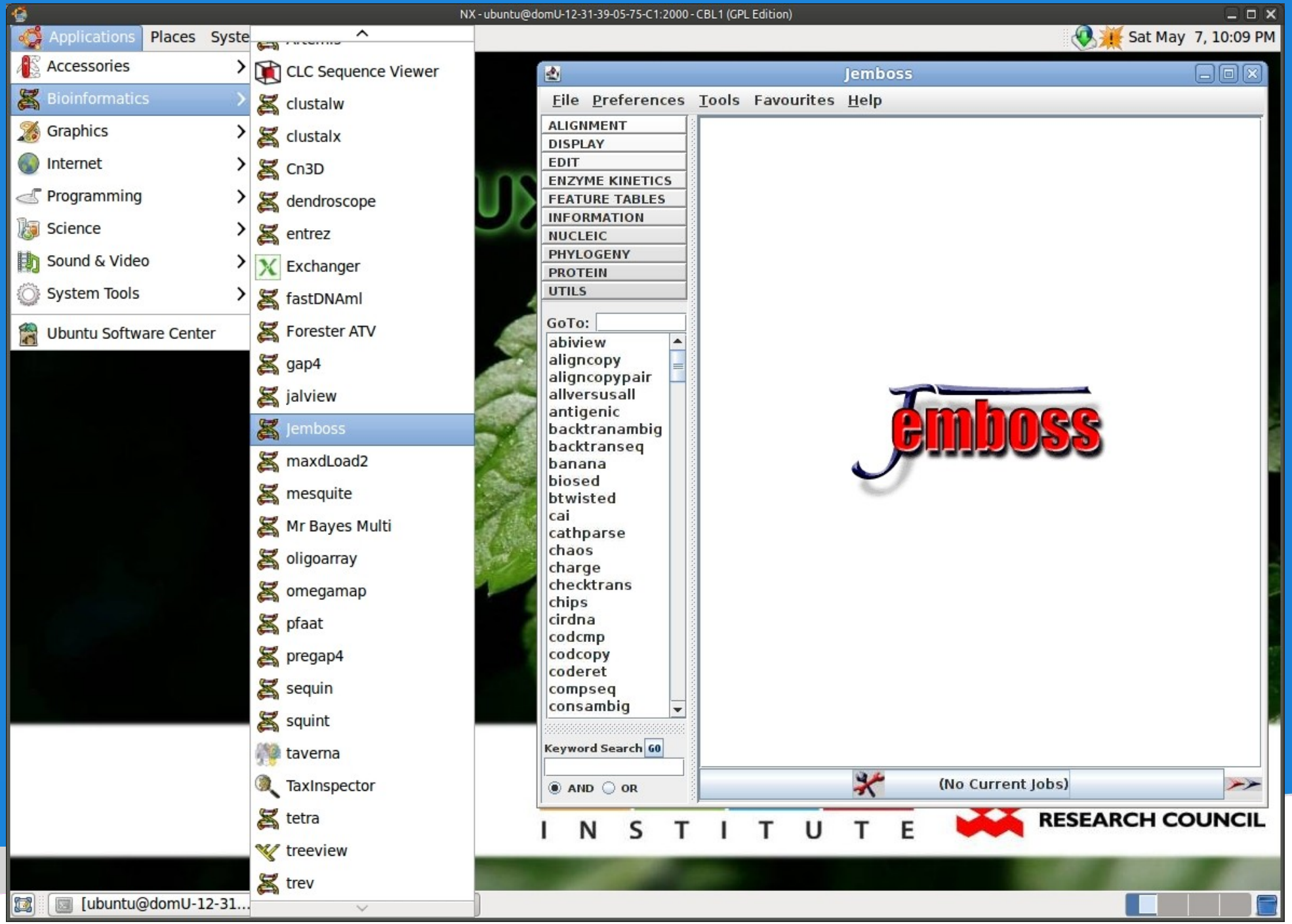
Cloud BioLinux launch wizard: step 3



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

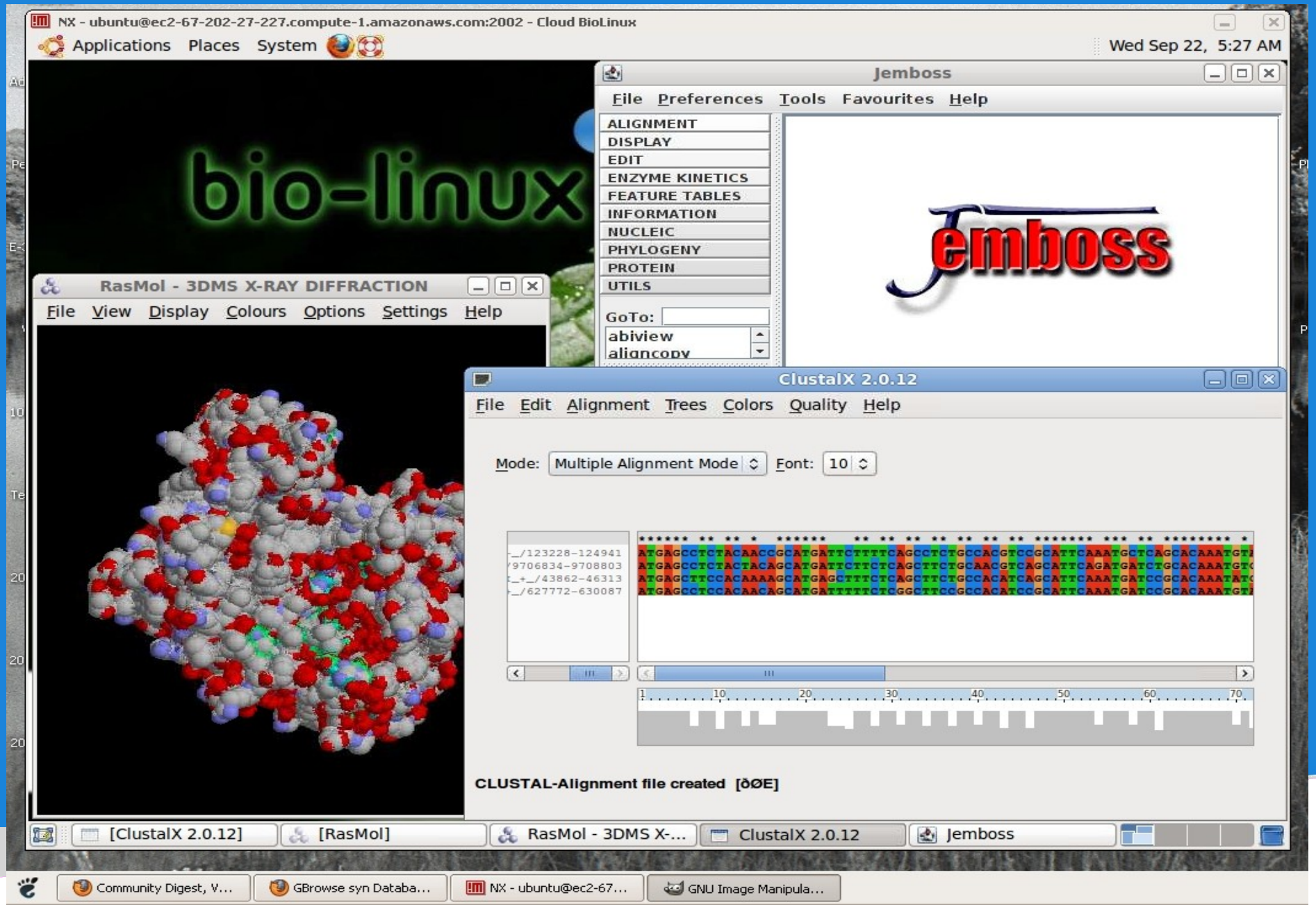
J. Craig Venter INSTITUTE


remote desktop client
History Bookmarks Tools Help
Back Forward Home Shareaholic https://console.aws.amazon.com/ec amazon.com x executable Print
S3FoxBrowse Amazon S3
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
i-49201823 ami-6011e409 ebs m1.large pending
i-f7340c9d ami-6011e409 ebs m1.large running normal ec2-184- default jcvi_key1
i-795b6313 ami-6816e301 ebs m1.large terminated normal
i-f9330b93 ami-6011e409 ebs m1.large terminated normal
1 EC2 Instance selected
EC2 Instance: i-f7340c9d
Description
Monitoring
Tags
AMI ID: ami-6011e409
Security Groups: default
Status: running
VPC ID: -
Virtualization: paravirtual
Reservation: r-6ec5cf05
Platform: -
Kernel ID: aki-0b4aa462
AMI Launch Index: 0
Root Device: /dev/sda1
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
Zone: us-east-1d
Type: m1.large
Owner: 767506454313
Subnet ID: -
Placement Group:
RAM Disk ID: -
Key Pair Name: jcvi_key1
Monitoring: disabled
Elastic IP: -
Root Device Type: ebs
© 2008 - 2009, Amazon Web Services LLC or its affiliates. All right reserved.
Feedback
Support
Privacy Policy
Terms of Use
An amazon.com company
NX - Cloud Biolinux
NOMACHINE
General
Advanced
Services
Environment
Server
Host ec2-184-73-27-151.comput
Port 22
Remember my password
Key...
Desktop
Unix
GNOME
Settings...
MODEM ISDN ADSL WAN LAN
Display
1024x768
W 800
H 600
Use custom settings
Settings...
Delete
Save
Ok
Cancel



J. Craig Venter™
INSTITUTE

Distributing Data Analysis Results with Cloud BioLinux



J. Craig Venter INSTITUTE

Distributing Data Analysis Results with Cloud BioLinux

Whole System Snapshot Exchange

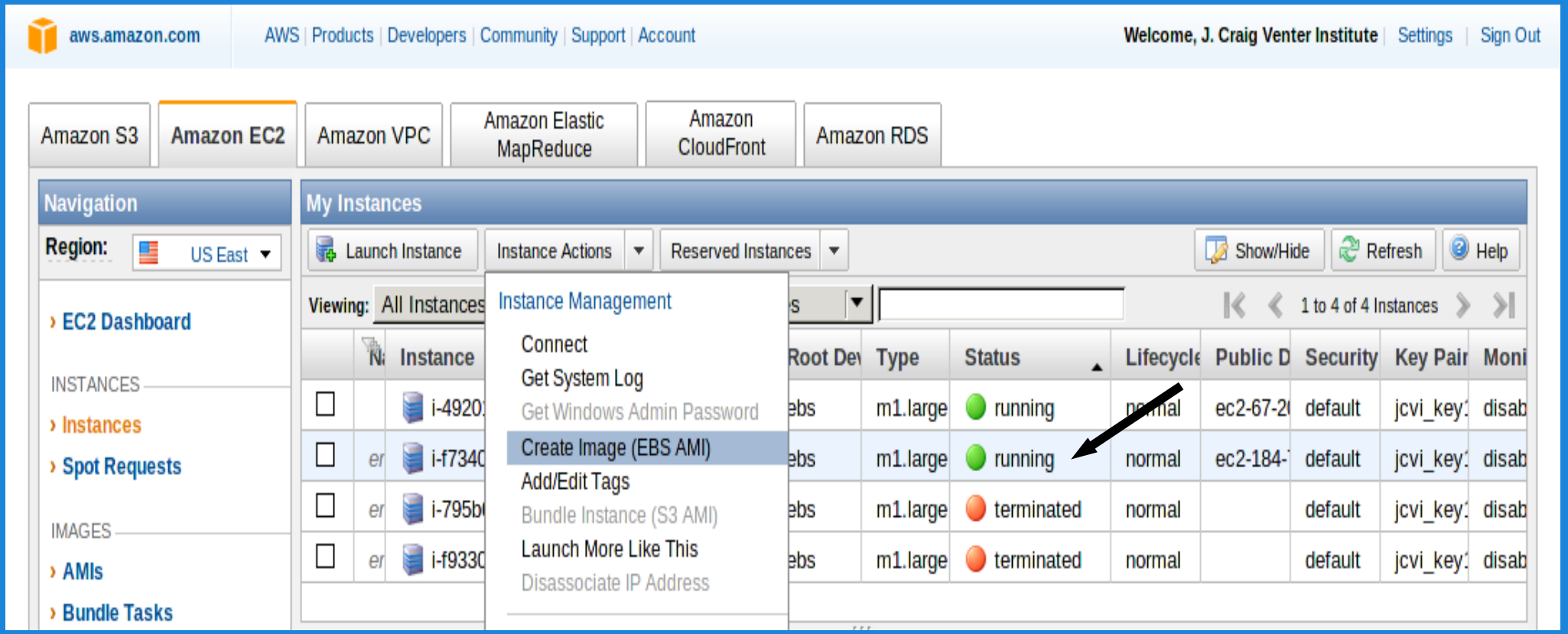
- how difficult is to share bioinformatics work on your computer with a collaborator ?
- capture the state of the computing system (OS + software), data, analysis results
- make VM snapshots: executable, binary file replica of the original VM

- 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

J. Craig Venter™
INSTITUTE

Cloud BioLinux: whole system snapshot exchange



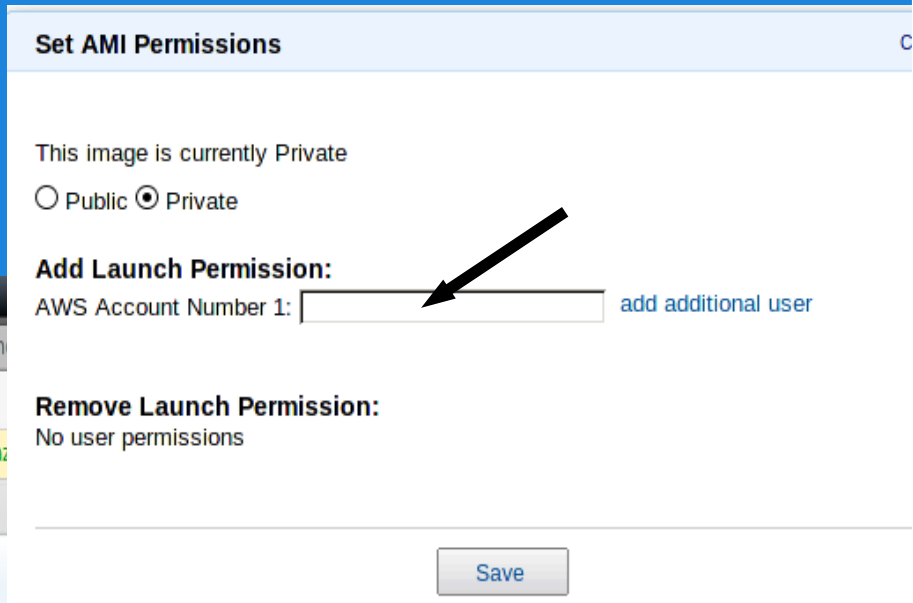
storage cost: 0.10$ / GB / month

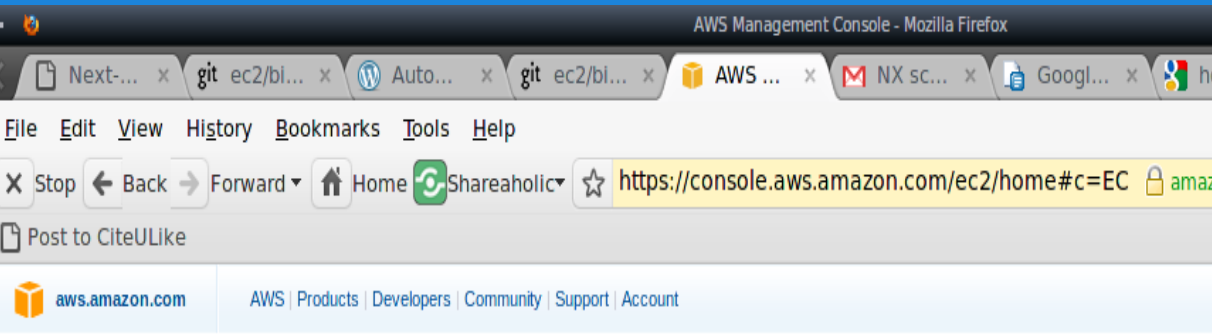
J. Craig Venter INSTITUTE

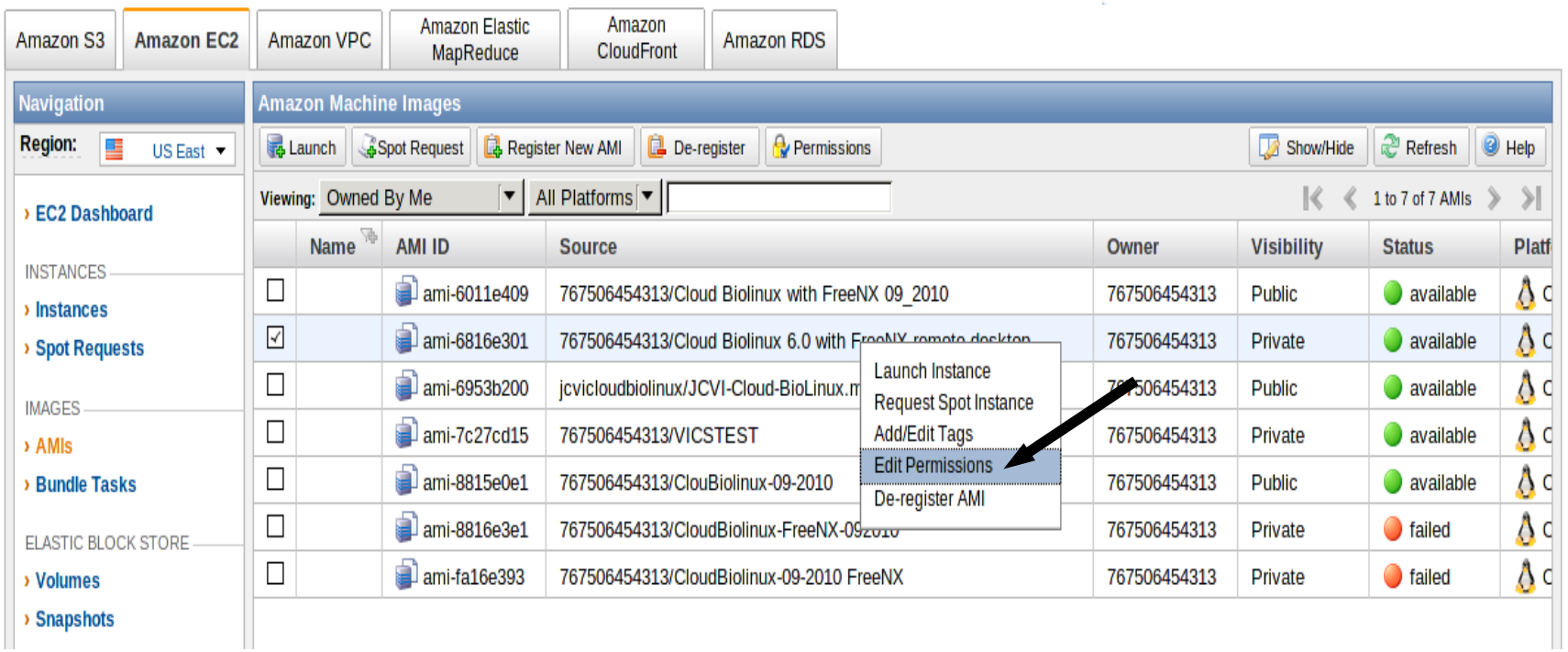
Cloud BioLinux: whole system snapshot exchange

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







J. Craig Venter INSTITUTE

5 min of questions, and then

5 more min to close the session....

J. Craig Venter™
INSTITUTE

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

J. Craig Venter™ INSTITUTE

bcbb / ec2 / biolinux / config / main.yaml

100644 | 39 lines (38 sloc) | 0.668 kb

```
---
# Top level configuration file that specifies w
# should be installed. New sections that are ad
# files should go here. Comment out any groups
# installed.
packages:
  - desktop
  - programming
  - distributed
  - amazon
  - python
  - r
  - ruby
  - perl
  - java
  - erlang
  - haskell
  - databases
  - math
  - viz
  - web
  - bio_general
  - bio_search
  - bio_alignment
  - bio_nextgen
  - bio_sequencing
  - bio_annotation
  - bio_microarray
  - bio_visualization
  - bio_utils
  - 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

bcbb / ec2 / biolinux / config / packages.yaml

```
    - apache2
bio_general:
    - emboss
    - emboss-data
    - emboss-lib
    - primer3
    - readseq
    - bio-linux-taverna
    - bio-linux-xcut
bio_search:
    - blast2
    - hmmer
    - ncbi-tools-bin
    - bio-linux-blast+
```

J. Craig Venter INSTITUTE

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

J. Craig Venter™ INSTITUTE

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@jcvi.org

http://www.cloudbiolinux.org

Thank you !

J. Craig Venter™ INSTITUTE