# Cloud BioLinux: Pre-configured and On-demand Bioinformatics Computing for the Genomics Community

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

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

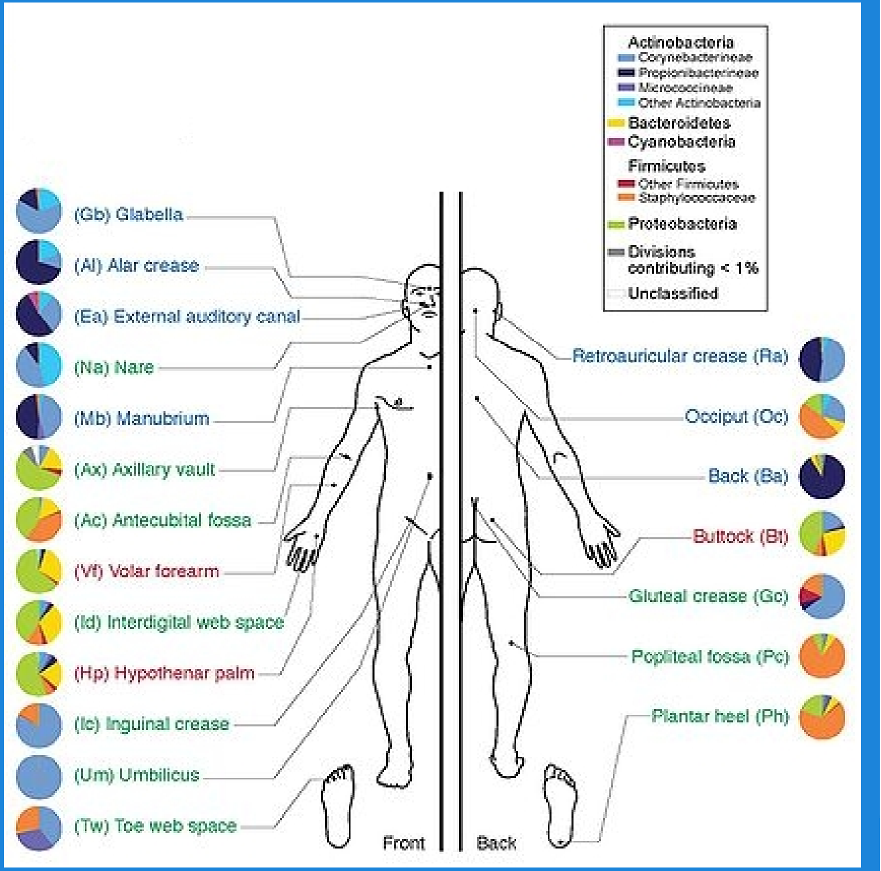
J. Craig Venter™
INSTITUTE

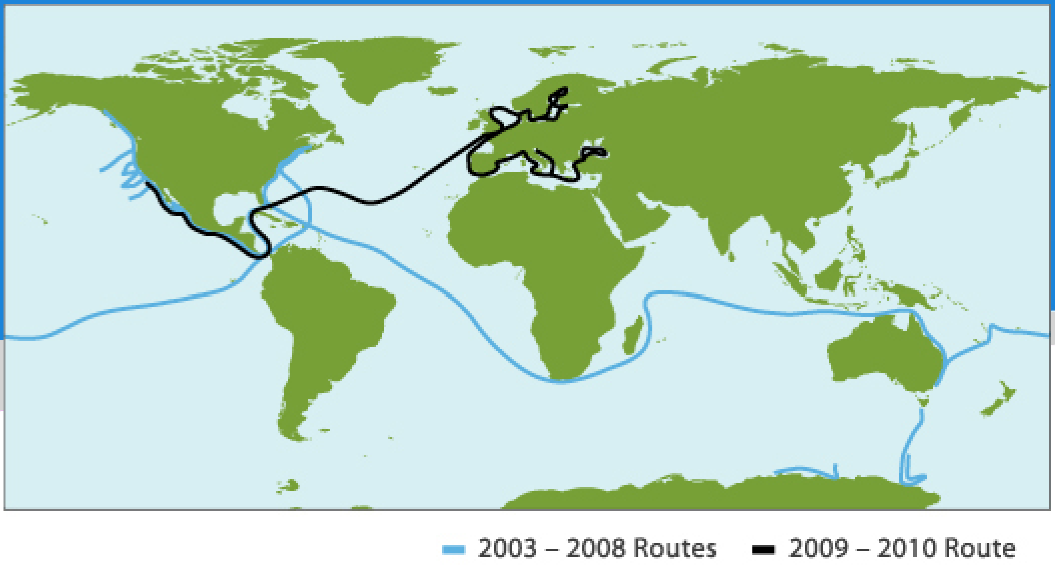
# J. Craig Venter Institute (JCVI)

## Large-scale genome sequencing and bioinformatics computing

- Human Microbiome Project (HMP): sequencing and assembly of 1000 reference microbe genomes from the human body

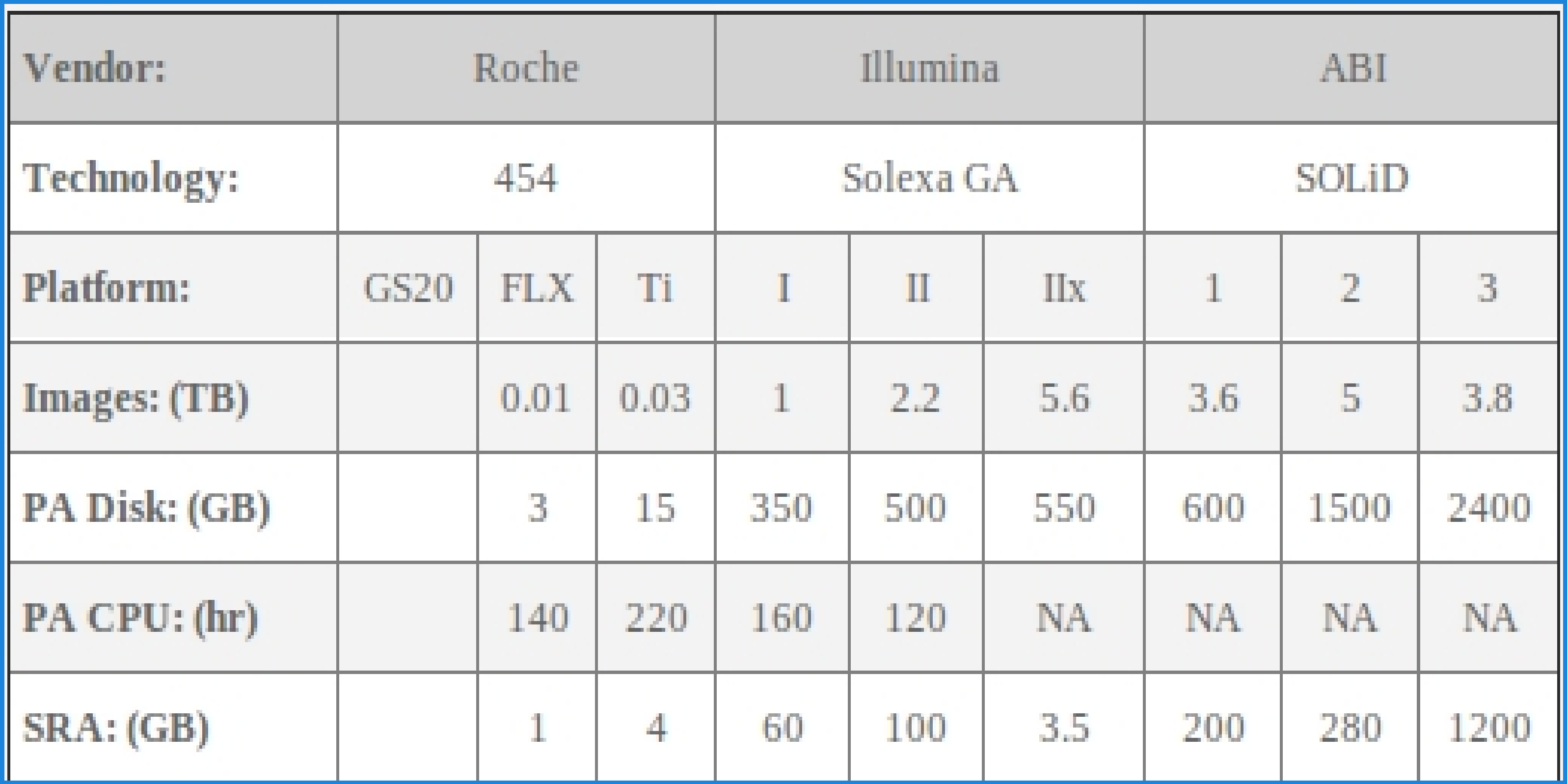
- Global Ocean Sampling (GOS) survey: metagenomic sequencing of microbes sampled from oceans around the world





# JCVI: sequencing and computing infrastructure

sequencing laboratory: 454, Solexa, HiSeq, and IonTorrent on the way



| Vendor: | Roche | | | Illumina | | | ABI | | |
|---|---|---|---|---|---|---|---|---|---|
| Technology: | 454 | | | Solexa GA | | | SOLiD | | |
| Platform: | GS20 | FLX | Ti | I | II | IIx | 1 | 2 | 3 |
| Images: (TB) | | 0.01 | 0.03 | 1 | 2.2 | 5.6 | 3.6 | 5 | 3.8 |
| PA Disk: (GB) | | 3 | 15 | 350 | 500 | 550 | 600 | 1500 | 2400 |
| PA CPU: (hr) | | 140 | 220 | 160 | 120 | NA | NA | NA | NA |
| SRA: (GB) | | 1 | 4 | 60 | 100 | 3.5 | 200 | 280 | 1200 |

Source: Dave Dooling,
http://www.politigenomics.com/next-generation-sequencing-informatics

J. Craig Venter™ INSTITUTE

# JCVI: sequencing and computing infrastructure

- large-scale sequencing needs large-scale informatics

- workhorse : ~1000 node Sun Grid Engine (SGE) cluster

- research in data processing and software development model with Hadoop / MapRecuce and a small private cloud

- bioinformatics department (57 bioinformaticians + software developers)

J. Craig Venter™ INSTITUTE

# A new paradigm:
# Low-cost, bench-top sequencers

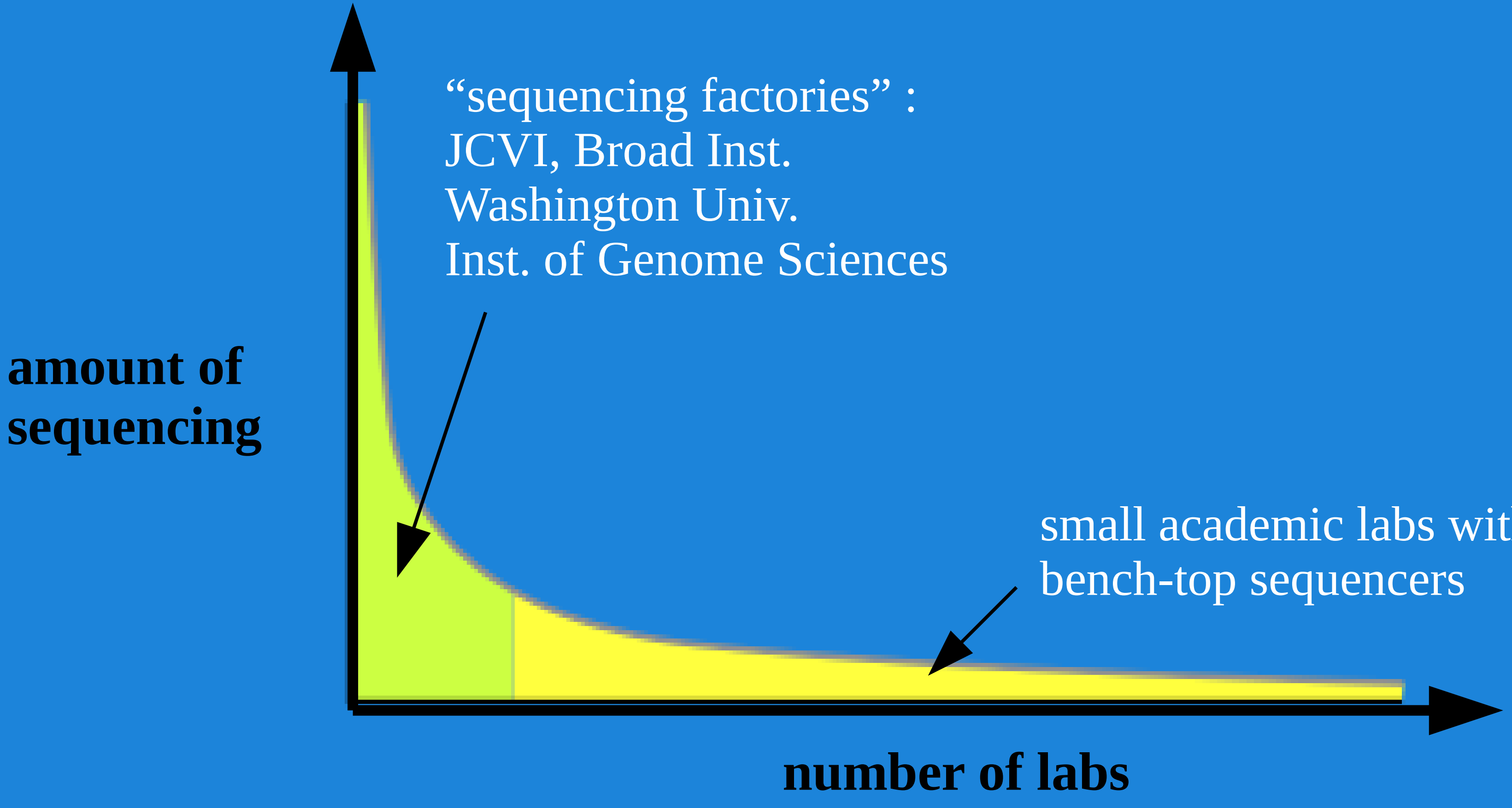
- small-scale sequencers available: GS Junior by 454, MiSeq by Illumina
- complete sequencing of bacterial, viral, small fungal genomes
- RNAseq (gene expression), ChiPseq (protein interactions), gene variant discovery
- sequencing as a standard technique in basic genetics research - like PCR ?

http://www.gsjunior.com/

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

J. Craig Venter™ INSTITUTE

# Will small academic labs lead by individual PIs become the long tail of sequencing ?



"sequencing factories" :
JCVI, Broad Inst.
Washington Univ.
Inst. of Genome Sciences

small academic labs with
bench-top sequencers

J. Craig Venter™ INSTITUTE

# Sequencers shipped with minimal computational capacity

- Problem 1: sequence analysis requires  plenty of computational capacity

  For example: genome assembly, BLAST and genome annotation

- Problem 2: bioinformatics tools need expertise with unix/linux operating systems, software libraries, compiling source code etc.

  Difficult to install and use for biologists

???

# Each lab with a sequencer building an informatics infrastructure ?

- difficult for individual PIs to get additional funds to build clusters
- funds for personnel to maintain the clusters and software
- duplication of effort across labs
- sub-optimal utilization of the hardware
- few sequencing runs per year

# Solution ? Large sequencing centers offering bioinformatics services

- Bioinformatic Resource Centers (BRCs) by NIAID

- bioinformatic services usually coupled with sequencing of a genome

- provide data access and on-line tools

- cannot provide bioinformatic support for every PI in a lab acquiring a sequencing instrument

- need end-to-end solutions, users submit sequence data and get final annotation

ViPR
Virus Pathogen Resource

PATRIC
PATHOSYSTEMS RESOURCE INTEGRATION CENTER

EuPathDB
Eukaryotic Pathogen Database Resources

NIAID

# Solving Problem 1: sequence analysis requires computational capacity

- computational capacity on-demand without investment on hardware
- Amazon Elastic Compute Cloud (EC2), pay-by-the-hour computing
- cloud servers cost $0.085 - $2 per hour
- max capacity per server 64GB RAM / 8 CPU (but a PI can run thousands of servers)
- access to computing resources without institutional, economic or national boundaries

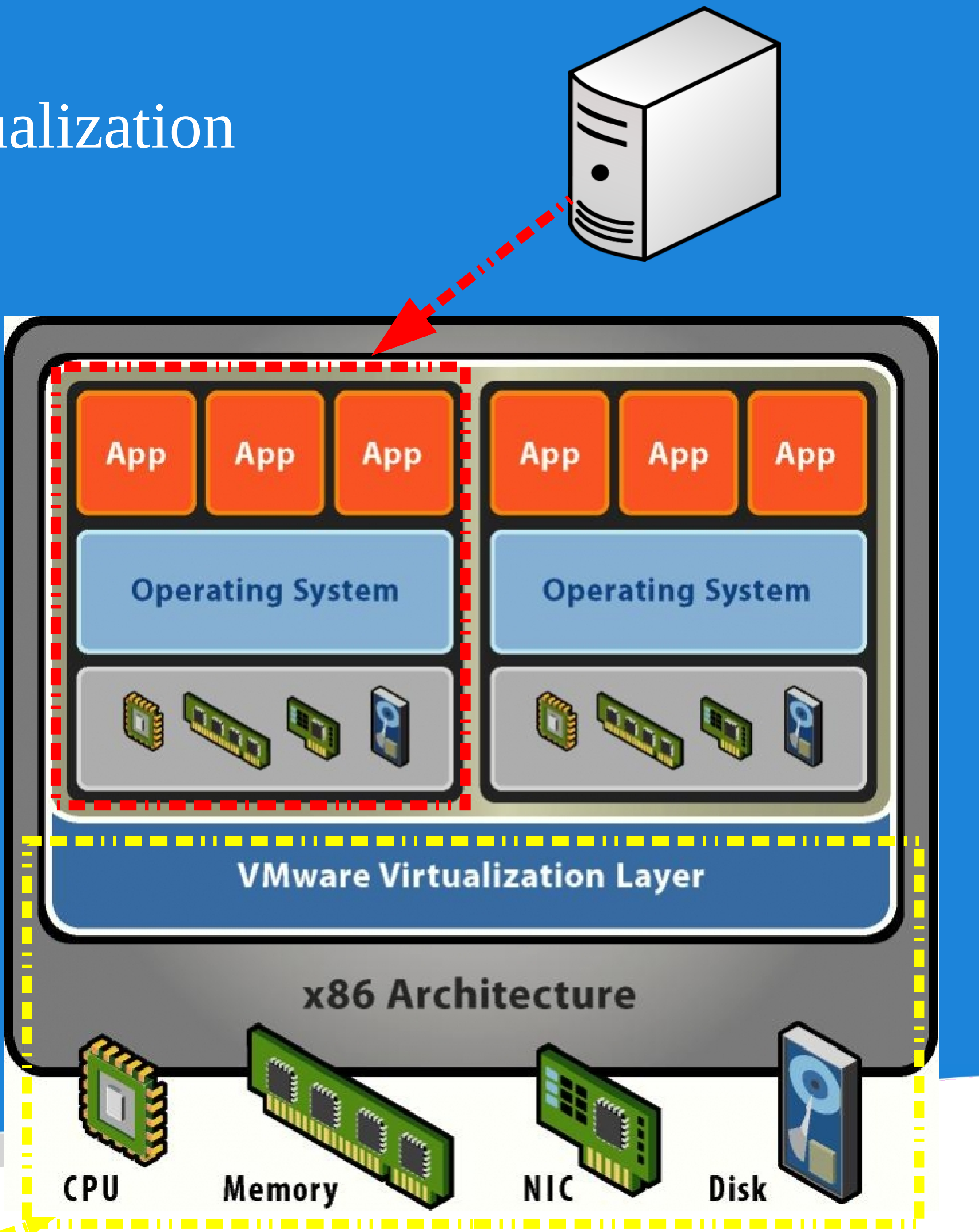
amazon webservices™

750 hours free for new users:
http://aws.amazon.com/free/

'enter™
INSTITUTE

# Cloud Computing and Virtualization

- operating system, bioinformatics software and data, are pre-installed on a Virtual Machine (VM)
- a VM is a full-featured unix/linux server, in the form of a single, executable binary file
- the cloud provides the physical computational resources and virtualization layer to run the VM



amazon web services™

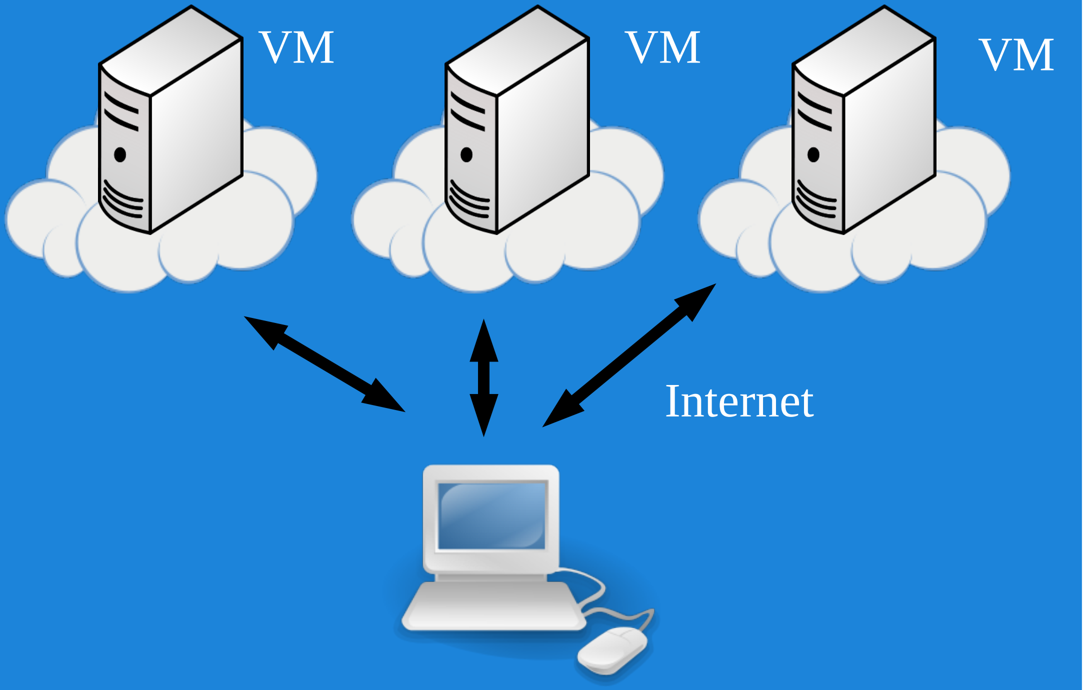
Credit: VMware Inc.

J. Craig Venter™ INSTITUTE

# Solving Problem 2: bioinformatics tools need software engineering expertise

- a VM with pre-installed bioinformatics software publicly accessible on the cloud
- no need to compile source code, set-up configuration files, or other software dependencies
- PIs rent computational capacity to run the VM
- bioinformatics software can be accessed from anywhere in the world via a local computer with Internet access
- no need for sequencing informatics infrastructure at each laboratory

Amazon EC2 cloud



local desktop computer at laboratory

J. Craig Venter™ INSTITUTE

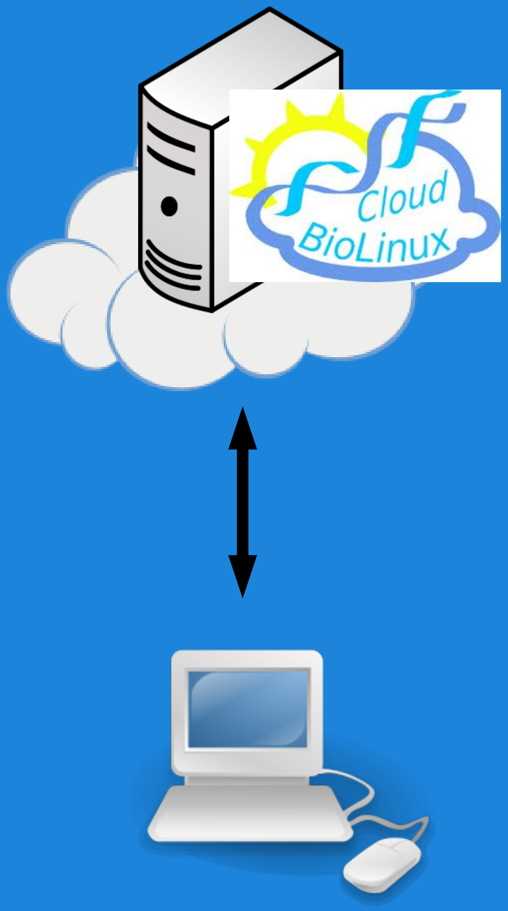
# Solving Problems 1 & 2: the Cloud BioLinux project

- Cloud BioLinux: a publicly accessible Virtual Machine (VM) on the Amazon EC2 cloud

- 100+ pre-installed bioinformatics tools on the VM with a graphical interface for non-technical users

- sequence analysis, genome assembly, annotation, phylogeny, molecular modeling, gene expression

- a researcher can initiate a practically unlimited number of Cloud BioLinux VMs for large-scale data analysis



**Krampis K.**, Booth T., Chapman B., Tiwari B., Bicak M., Field D. and Nelson K.E. *(2012) BMC Bioinformatics (in review)*, "Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community"

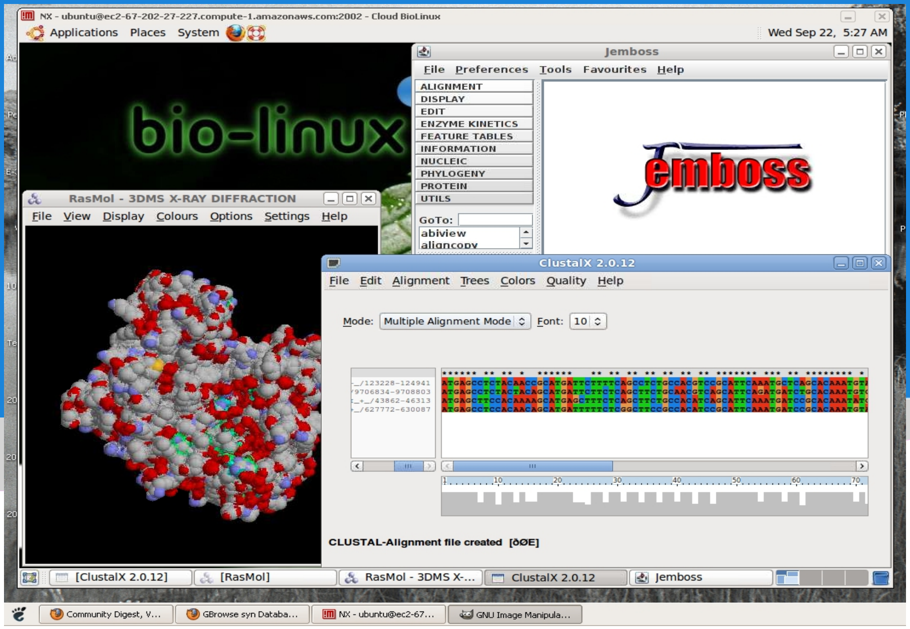
J. Craig Venter™ INSTITUTE

# 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, how bioinformatics developers are provided a framework for modifying and sharing VM configurations and data



Mode: Multiple Alignment Mode
Font: 10

```
_/123228-124941    ATGAGCCTCTACAACCGCATGATTCTTTTCAGCCTCTGCCACGTCCGCATTCAAATGCTCAGCACAAATGT
/9706834-9708803   ATGAGCCTCTACTACAGCATGATTCTTCTCAGCTTCTGCAACGTCAGCATTCAGATGATCTGCACAAATGT
_+_/43862-46313    ATGAGCTTCCACAAAAGCATGAGCTTTCTCAGCTTCTGCCACATCAGCATTCAAATGATCCGCACAAATAT
_/627772-630087    ATGAGCCTCCACAACAGCATGATTTTTCTCGGCTTCCGCCACATCCGCATTCAAATGATCCGCACAAATGT
```

CLUSTAL-Alignment file created [ðØE]

http://www.cloudbiolinux.org

http://tinyurl.com/BioLinux-NEBC

J. Craig Venter™
INSTITUTE

# The making of Cloud Biolinux

tinyurl.com/BioLinux-NEBC

+

amazon web services™

=

Cloud BioLinux

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: developer's framework for creating a customized cloud VM for bioinformatics

- main contributors:

HARVARD SCHOOL OF PUBLIC HEALTH

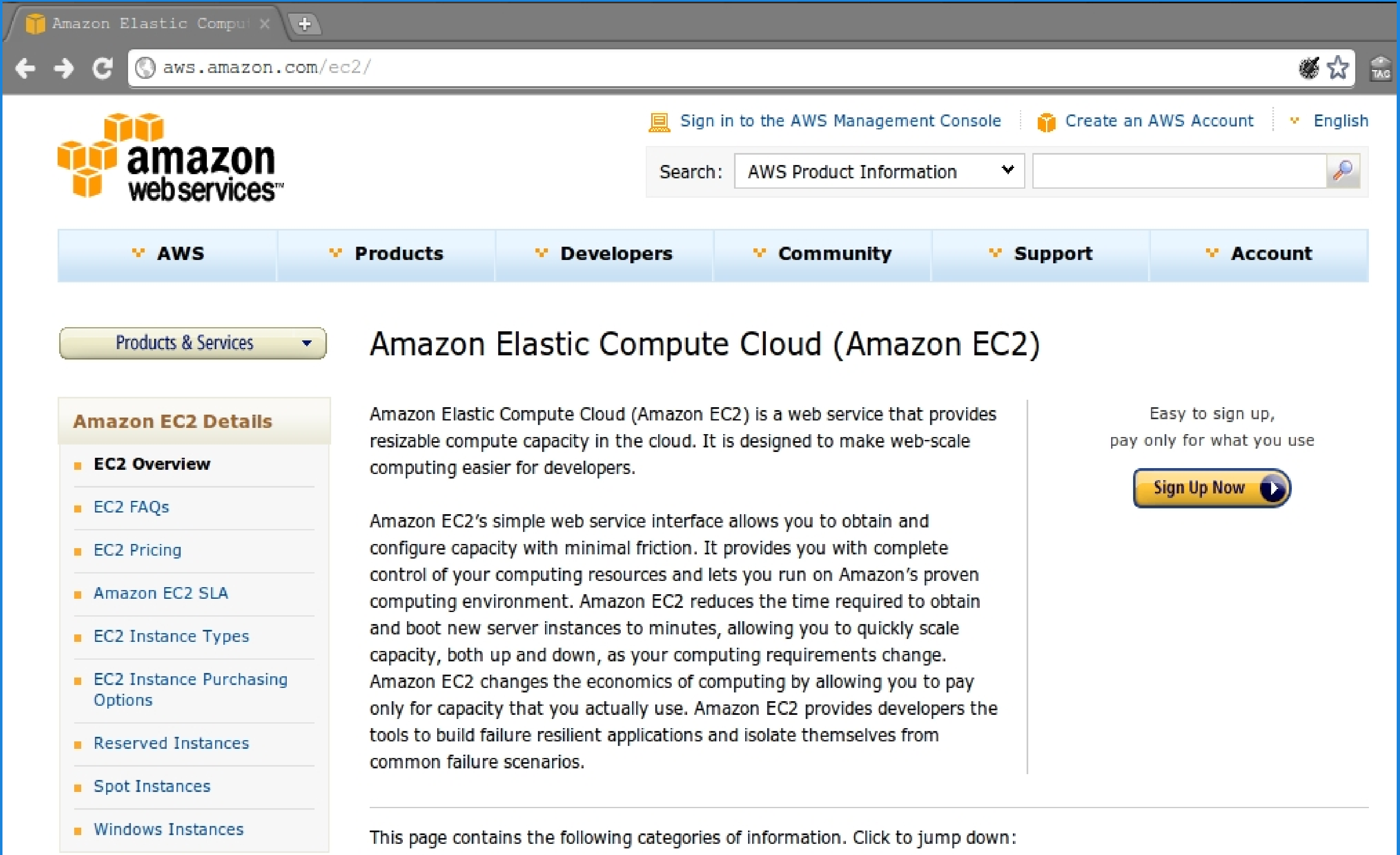
Galaxy

debian

NEBC

J. Craig Venter™ INSTITUTE

# Accessing Cloud BioLinux



## Amazon Elastic Compute Cloud (Amazon EC2)

**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

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.

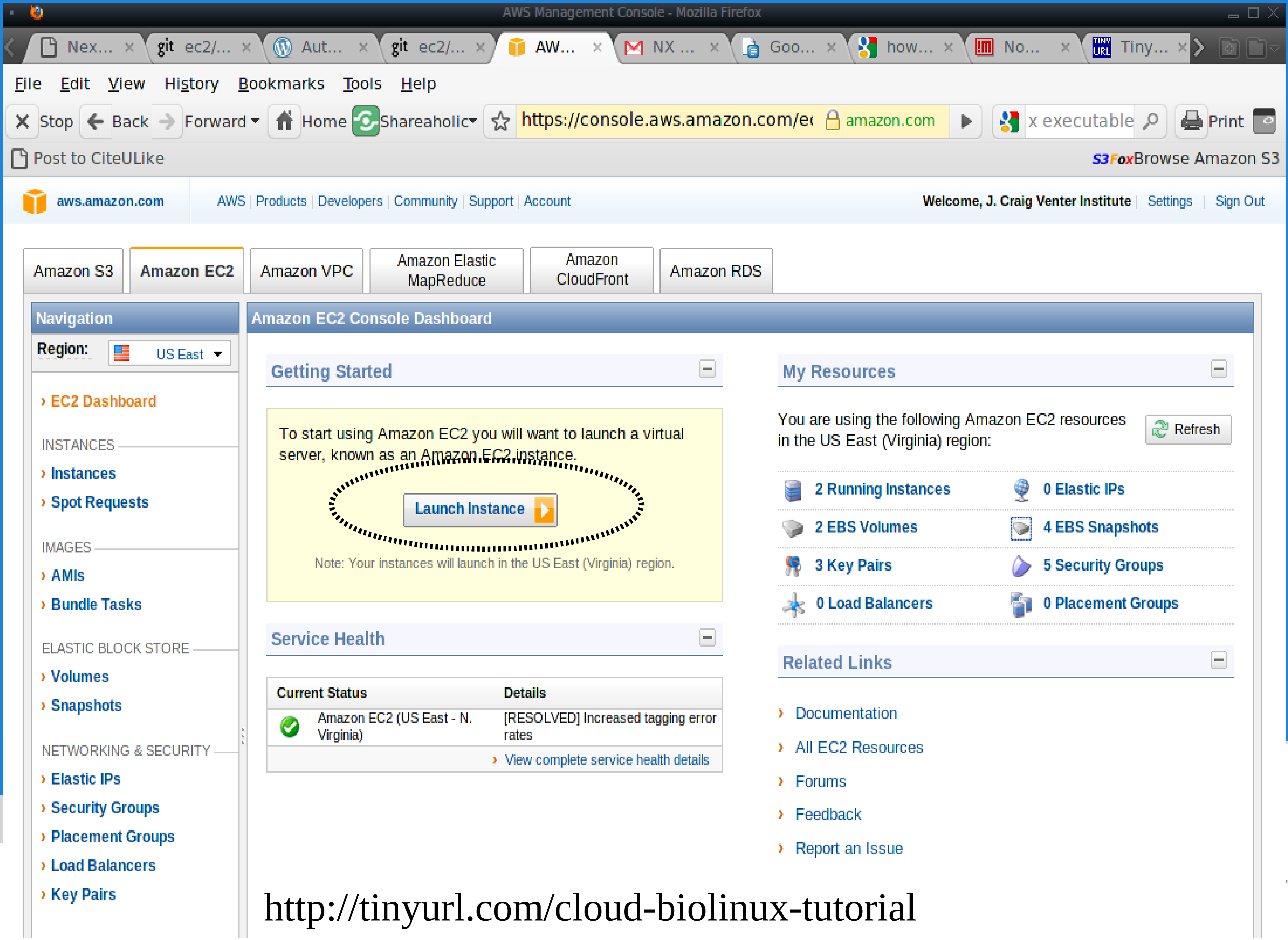
Easy to sign up,
pay only for what you use

Sign Up Now

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

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

# Launch Cloud BioLinux through the EC2 cloud console



Amazon EC2 Console Dashboard

Getting Started

To start using Amazon EC2 you will want to launch a virtual server, known as an Amazon EC2 instance.

Note: Your instances will launch in the US East (Virginia) region.

Service Health

| Current Status | Details |
|---|---|
| ✓ Amazon EC2 (US East - N. Virginia) | [RESOLVED] Increased tagging error rates |

› View complete service health details

My Resources

You are using the following Amazon EC2 resources in the US East (Virginia) region:

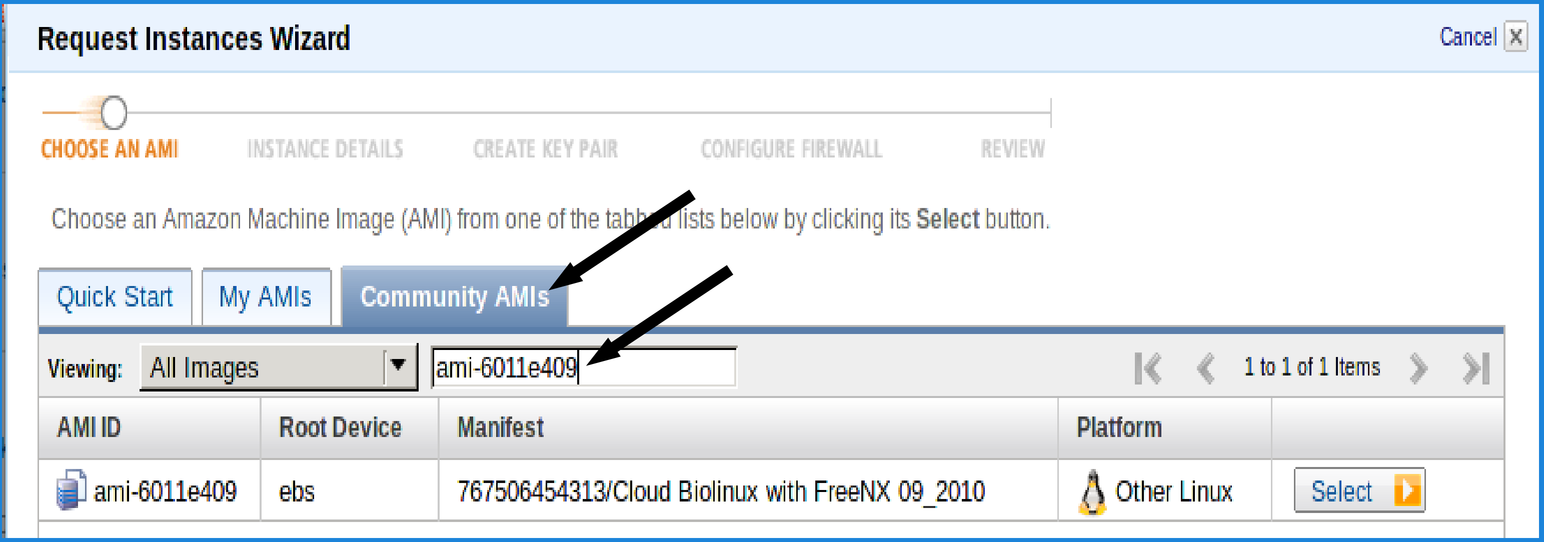
| | |
|---|---|
| 2 Running Instances | 0 Elastic IPs |
| 2 EBS Volumes | 4 EBS Snapshots |
| 3 Key Pairs | 5 Security Groups |
| 0 Load Balancers | 0 Placement Groups |

Related Links

- Documentation
- All EC2 Resources
- Forums
- Feedback
- Report an Issue

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

# Cloud BioLinux and VM launch wizard



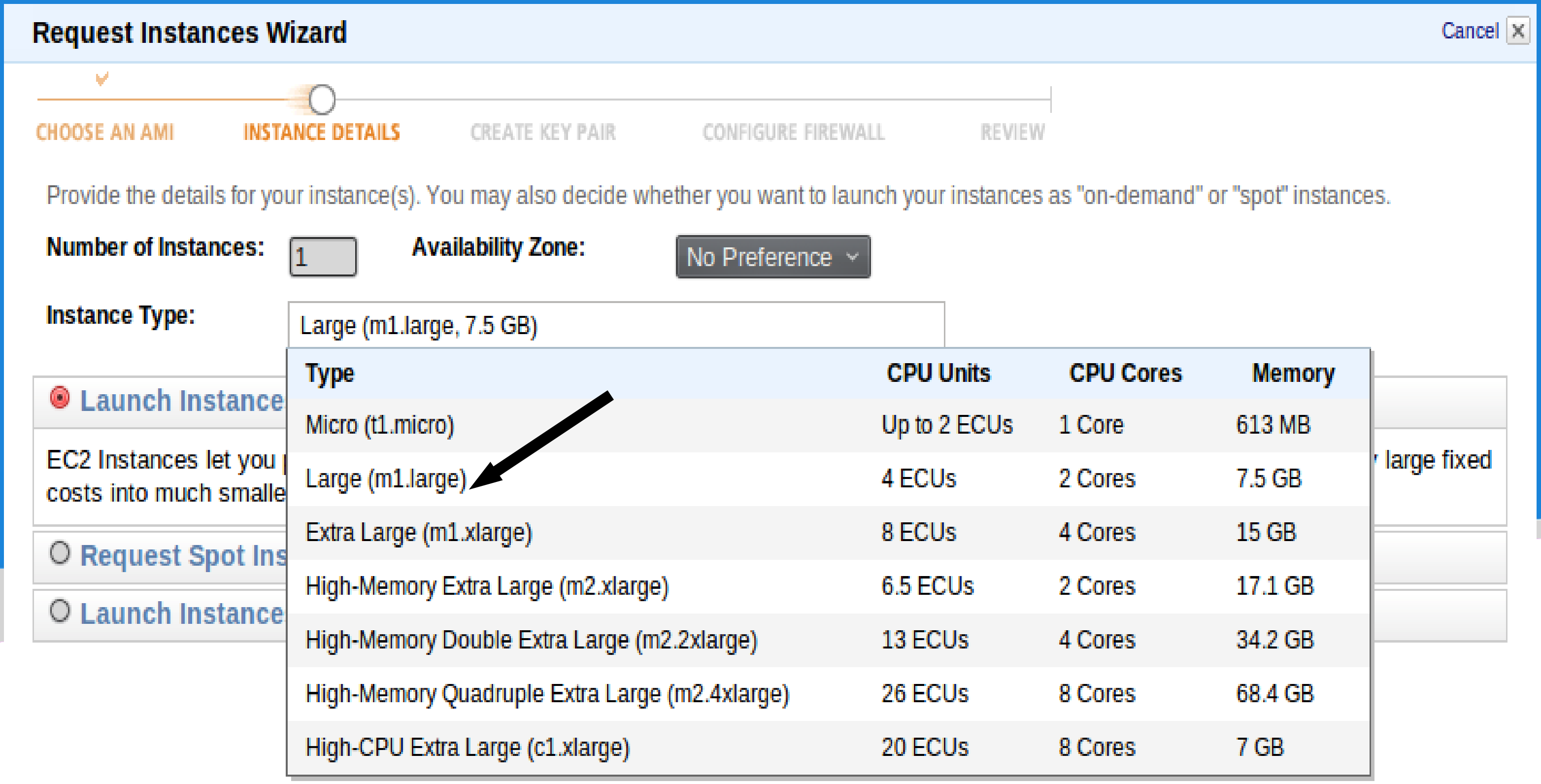
**Request Instances Wizard**

Choose an Amazon Machine Image (AMI) from one of the tabbed lists below by clicking its **Select** button.

Viewing: All Images ▾ ami-6011e409

| AMI ID | Root Device | Manifest | Platform | |
|---|---|---|---|---|
| ami-6011e409 | ebs | 767506454313/Cloud Biolinux with FreeNX 09_2010 | Other Linux | Select |

Community AMIs, search for Cloud BioLinux VM identifier (most recent update: cloudbiolinux.org)



**Request Instances Wizard**

Provide the details for your instance(s). You may also decide whether you want to launch your instances as "on-demand" or "spot" instances.

**Number of Instances:** 1 **Availability Zone:** No Preference ▾

**Instance Type:** Large (m1.large, 7.5 GB)

| Type | CPU Units | CPU Cores | Memory |
|---|---|---|---|
| Micro (t1.micro) | Up to 2 ECUs | 1 Core | 613 MB |
| Large (m1.large) | 4 ECUs | 2 Cores | 7.5 GB |
| Extra Large (m1.xlarge) | 8 ECUs | 4 Cores | 15 GB |
| High-Memory Extra Large (m2.xlarge) | 6.5 ECUs | 2 Cores | 17.1 GB |
| High-Memory Double Extra Large (m2.2xlarge) | 13 ECUs | 4 Cores | 34.2 GB |
| High-Memory Quadruple Extra Large (m2.4xlarge) | 26 ECUs | 8 Cores | 68.4 GB |
| High-CPU Extra Large (c1.xlarge) | 20 ECUs | 8 Cores | 7 GB |

● Launch Instance

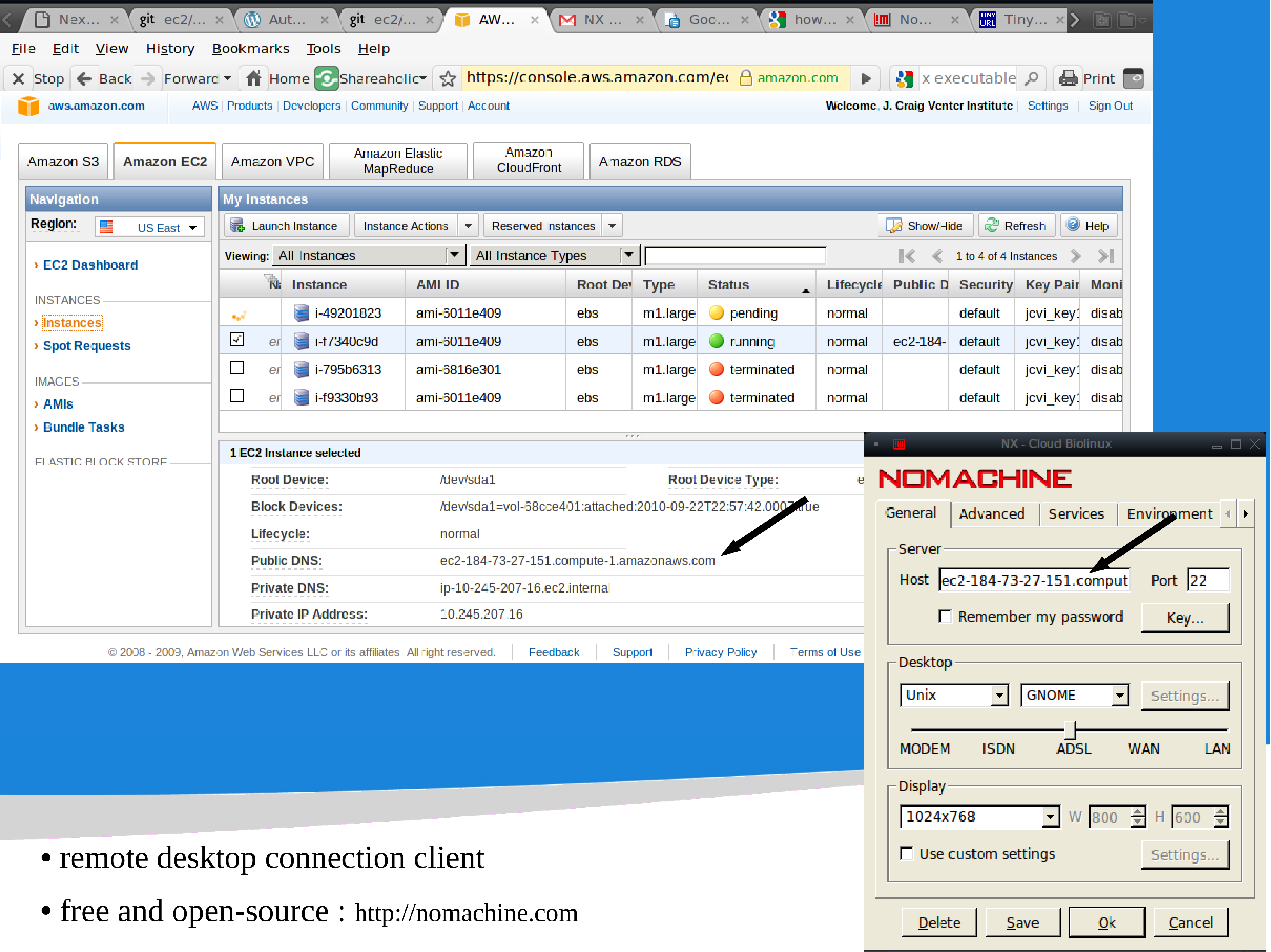
EC2 Instances let you p … large fixed costs into much smalle

○ Request Spot Ins

○ Launch Instance

select computational capacity for the VM

J. Craig Venter™ INSTITUTE



|  | Na | Instance | AMI ID | Root Dev | Type | Status | Lifecycle | Public D | Security | Key Pair | Moni |
|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  | i-49201823 | ami-6011e409 | ebs | m1.large | pending | normal |  | default | jcvi_key1 | disab |
| ☑ | er | i-f7340c9d | ami-6011e409 | ebs | m1.large | running | normal | ec2-184- | default | jcvi_key1 | disab |
| □ | er | i-795b6313 | ami-6816e301 | ebs | m1.large | terminated | normal |  | default | jcvi_key1 | disab |
| □ | er | i-f9330b93 | ami-6011e409 | ebs | m1.large | terminated | normal |  | default | jcvi_key1 | disab |

1 EC2 Instance selected

Root Device: /dev/sda1 Root Device Type: e

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

Server

Host ec2-184-73-27-151.comput Port 22

□ Remember my password Key...

Desktop

Unix GNOME Settings...

MODEM ISDN ADSL WAN LAN

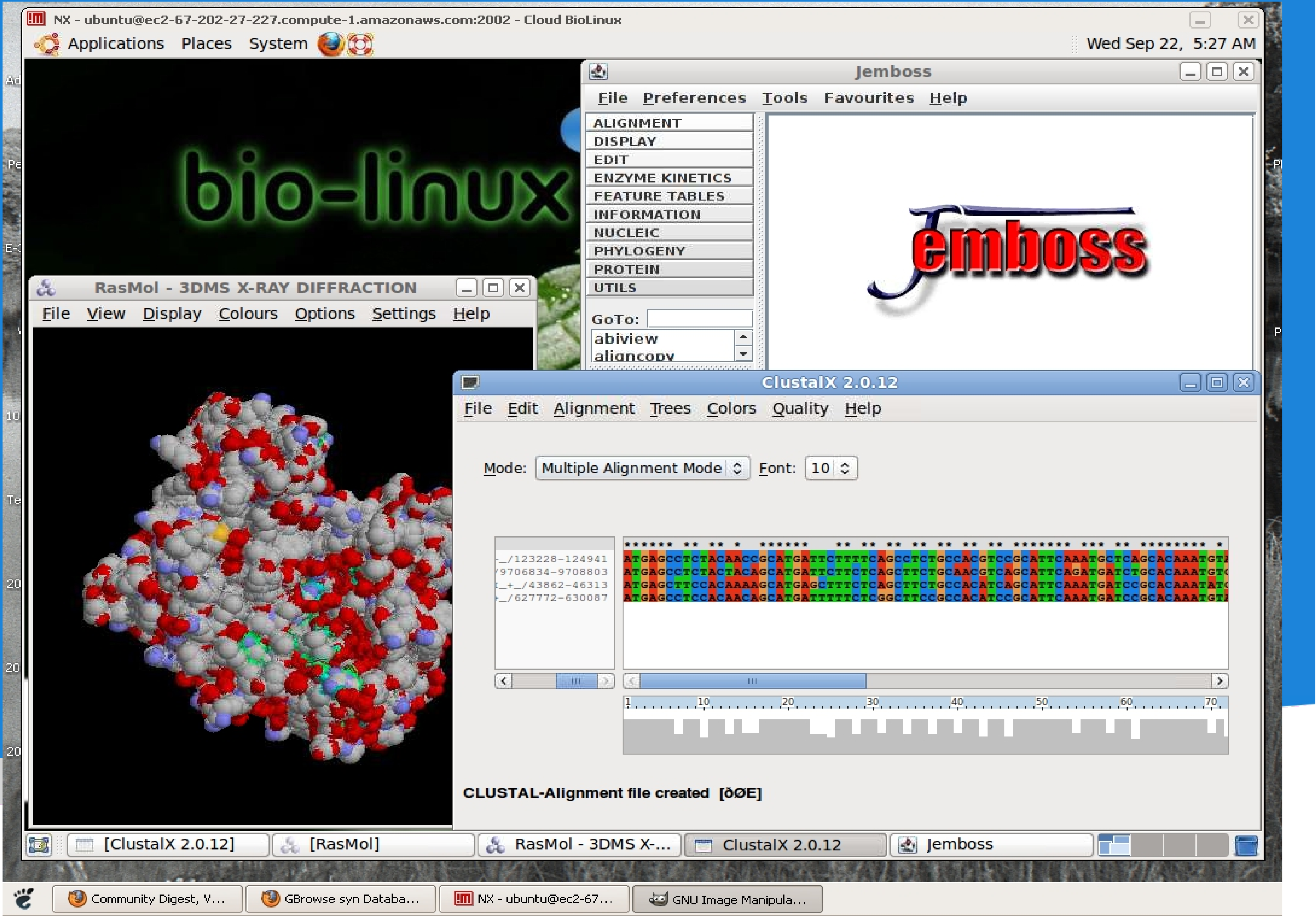
Display

1024x768 W 800 H 600

□ Use custom settings Settings...

- remote desktop connection client
- free and open-source : http://nomachine.com

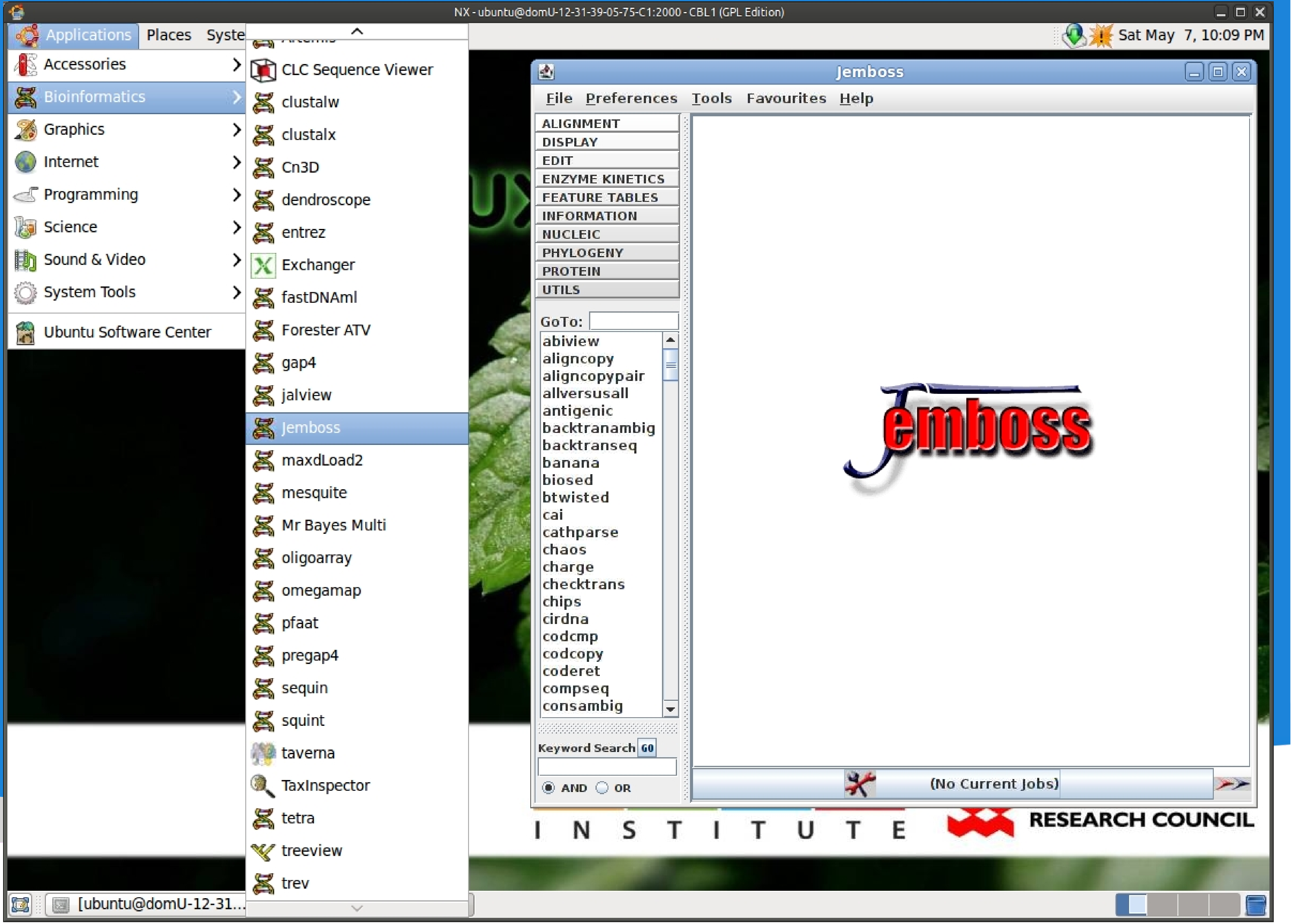
# Cloud BioLinux with remote desktop connection



Mode: Multiple Alignment Mode  Font: 10

```
                  ****** ** ** * ******   ** ** ** ** ** ** ** ******* *** ** ********* *
-_/123228-124941  ATGAGCCTCTACAACCGCATGATTCTTTTCAGCCTCTGCCACGTCCGCATTCAAATGCTCAGCACAAATGT
/9706834-9708803  ATGAGCCTCTACTACAGCATGATTCTTCTCAGCTTCTGCAACGTCAGCATTCAGATGATCTGCACAAATGT
_+_/43862-46313   ATGAGCTTCCACAAAAGCATGAGCTTTCTCAGCTTCTGCCACATCAGCATTCAAATGATCCGCACAAATAT
-_/627772-630087  ATGAGCCTCCACAACAGCATGATTTTTCTCGGCTTCCGCCACATCCGCATTCAAATGATCCGCACAAATGT
```

CLUSTAL-Alignment file created [ðØE]


NX - ubuntu@domU-12-31-39-05-75-C1:2000 - CBL1 (GPL Edition)
Applications
Places
Syste
Sat May 7, 10:09 PM
Accessories
Bioinformatics
Graphics
Internet
Programming
Science
Sound & Video
System Tools
Ubuntu Software Center
Artemis
CLC Sequence Viewer
clustalw
clustalx
Cn3D
dendroscope
entrez
Exchanger
fastDNAml
Forester ATV
gap4
jalview
Jemboss
maxdLoad2
mesquite
Mr Bayes Multi
oligoarray
omegamap
pfaat
pregap4
sequin
squint
taverna
TaxInspector
tetra
treeview
trev
[ubuntu@domU-12-31...
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
GO
● AND
○ OR
Jemboss
(No Current Jobs)
I N S T I T U T E
RESEARCH COUNCIL

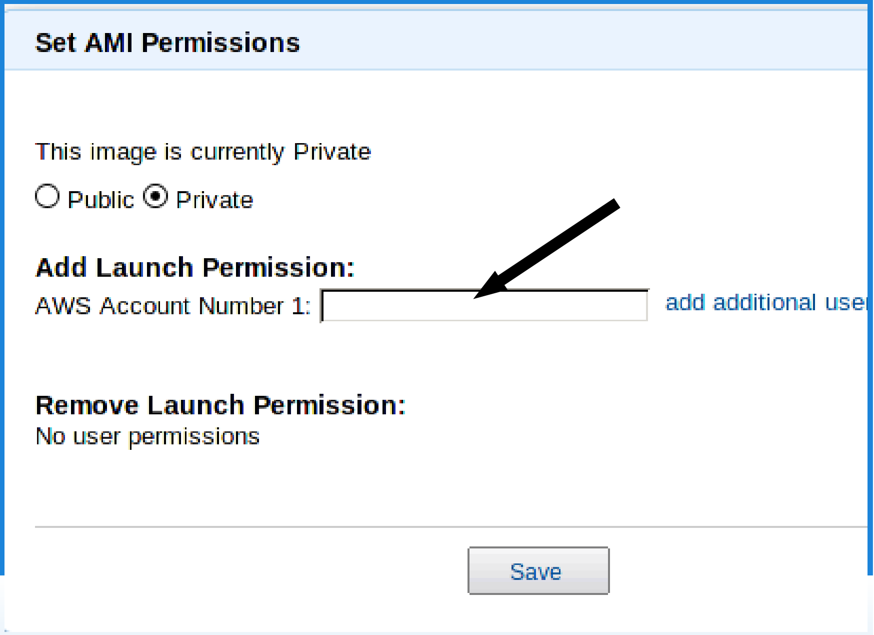
# Cloud BioLinux:

## sharing data & results with VM snapshots

- access rights to the “snapshot” VM: public or for specific user

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

- storage cost: 0.10$ / GB / month



**Set AMI Permissions**

This image is currently Private

○ Public ● Private

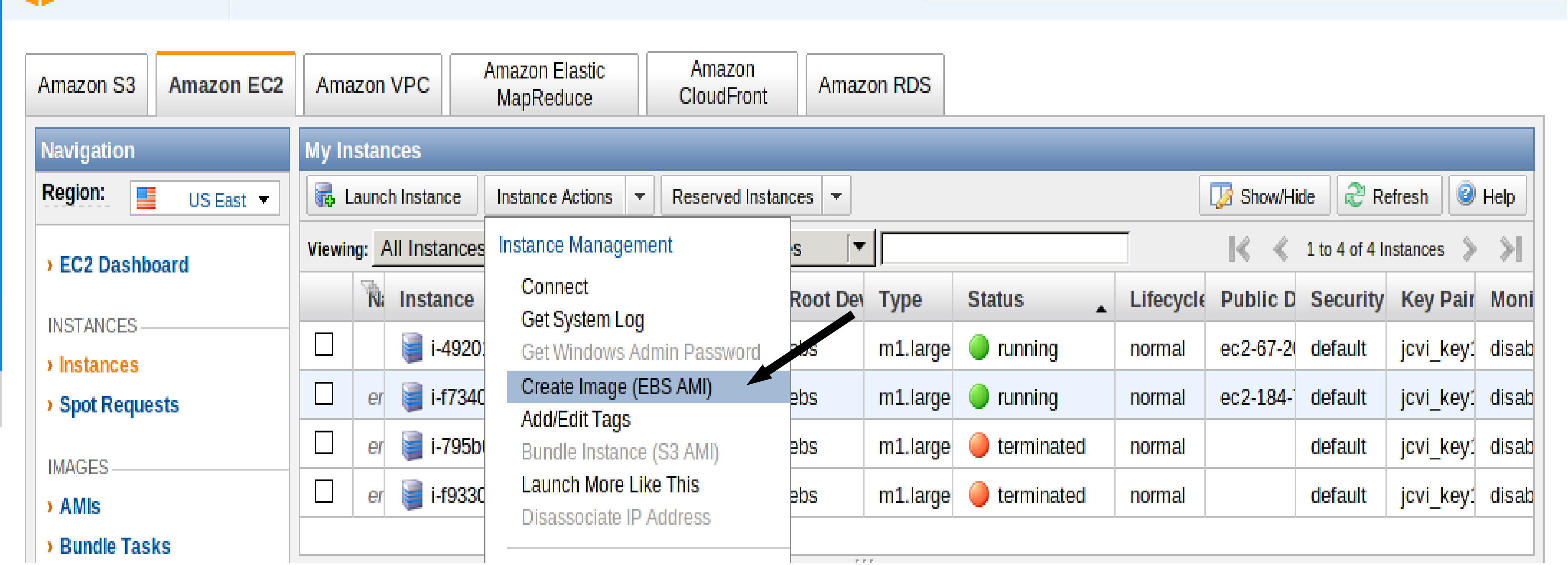
**Add Launch Permission:**

AWS Account Number 1: __________ add additional use

**Remove Launch Permission:**

No user permissions

aws.amazon.com AWS | Products | Developers | Community | Support | Account



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

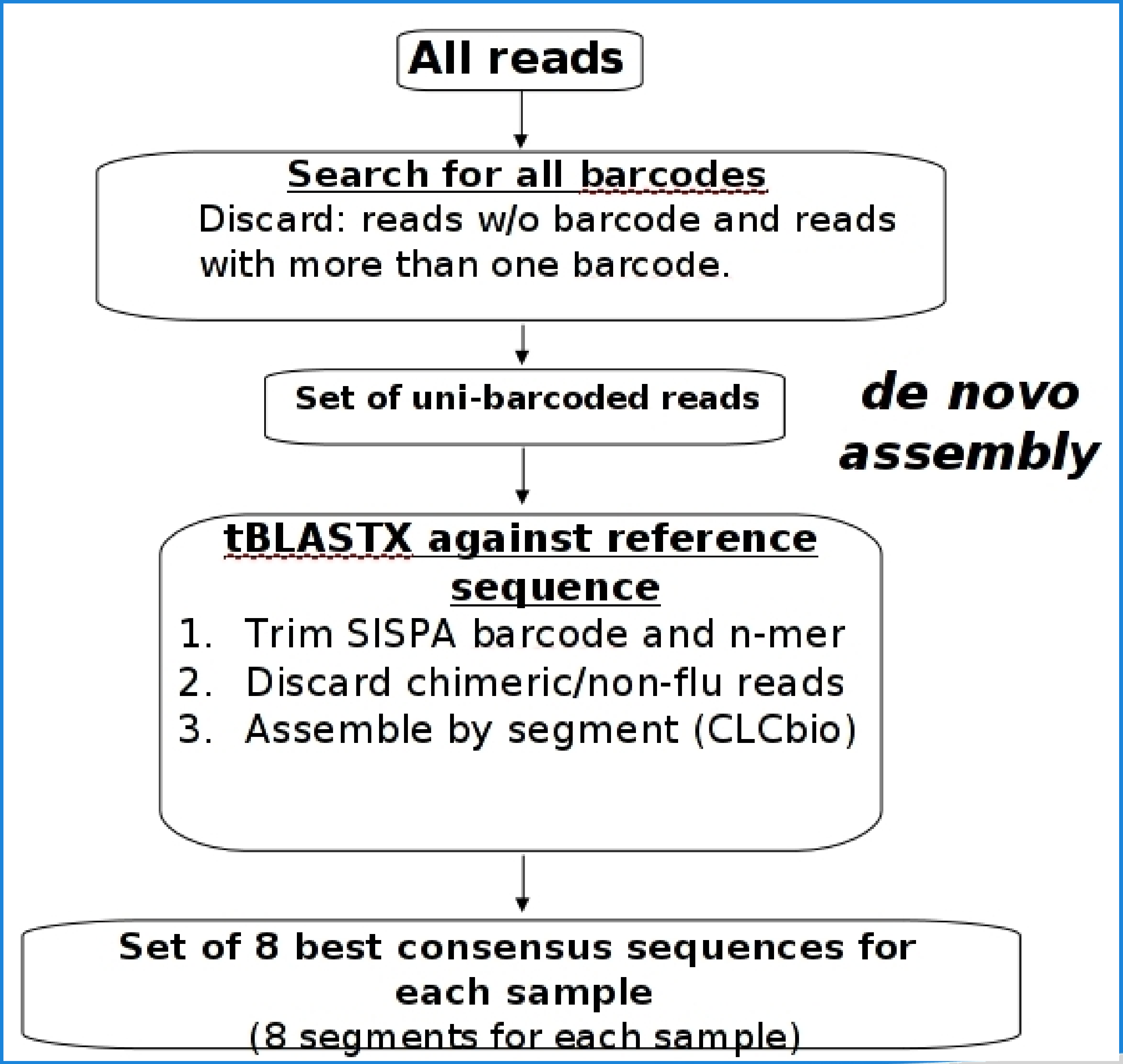
# Research at JCVI with Cloud BioLinux

- bioinformatics data analysis pipelines are complex
- approach: pre-install pipelines and all their dependencies within a VM
- make VM available on Amazon EC2
- we use private clouds, Eucalyptus & OpenStack
- open-source cloud platforms, fully compatible with Amazon EC2 (identical API)
- easy to set up on a local computer cluster, comes with Ubuntu Linux server edition
- also can run on your laptop with VirtualBox

instructions and VM at http://www.cloudbiolinux.org

# JCVI's Viral Genome Sequencing Pipelines

## Phase I-a
## Sequencing & Assembly



**All reads**

**Search for all barcodes**
Discard: reads w/o barcode and reads with more than one barcode.

**Set of uni-barcoded reads**

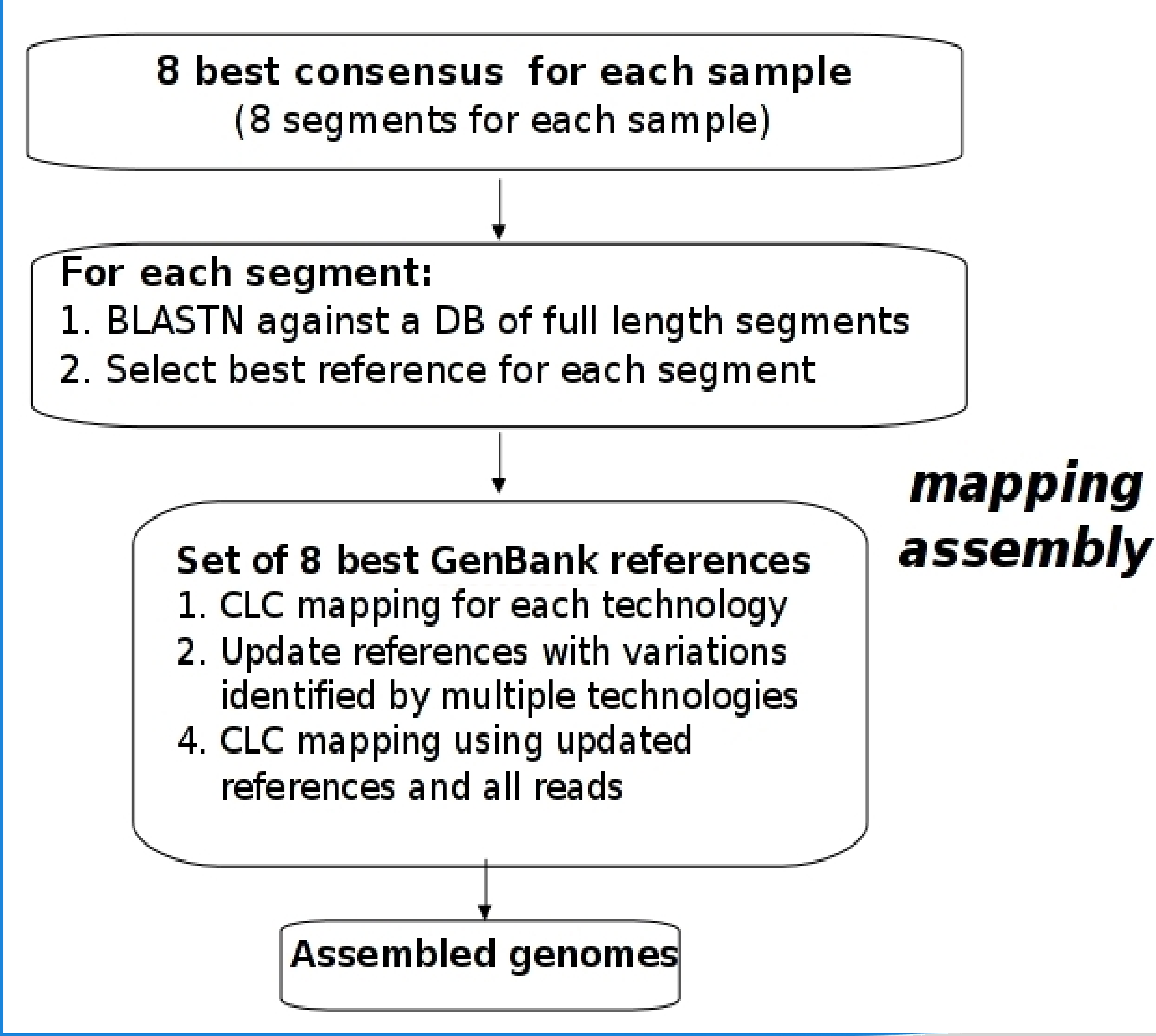
***de novo* assembly**

**tBLASTX against reference sequence**

1. Trim SISPA barcode and n-mer
2. Discard chimeric/non-flu reads
3. Assemble by segment (CLCbio)

**Set of 8 best consensus sequences for each sample**
(8 segments for each sample)

Credit: Tim Stockwell, JCVI Viral Informatics



8 best consensus for each sample
(8 segments for each sample)

For each segment:
1. BLASTN against a DB of full length segments
2. Select best reference for each segment

Set of 8 best GenBank references
1. CLC mapping for each technology
2. Update references with variations identified by multiple technologies
4. CLC mapping using updated references and all reads

Assembled genomes

*mapping assembly*

## JCVI's Viral Genome Sequencing Pipelines

## Phase I-b Sequencing & Assembly

Credit: Tim Stockwell, JCVI Viral Informatics

J. Craig Venter™ INSTITUTE

# JCVI's Viral Genome Sequencing Pipelines

## Phase II
## Annotation

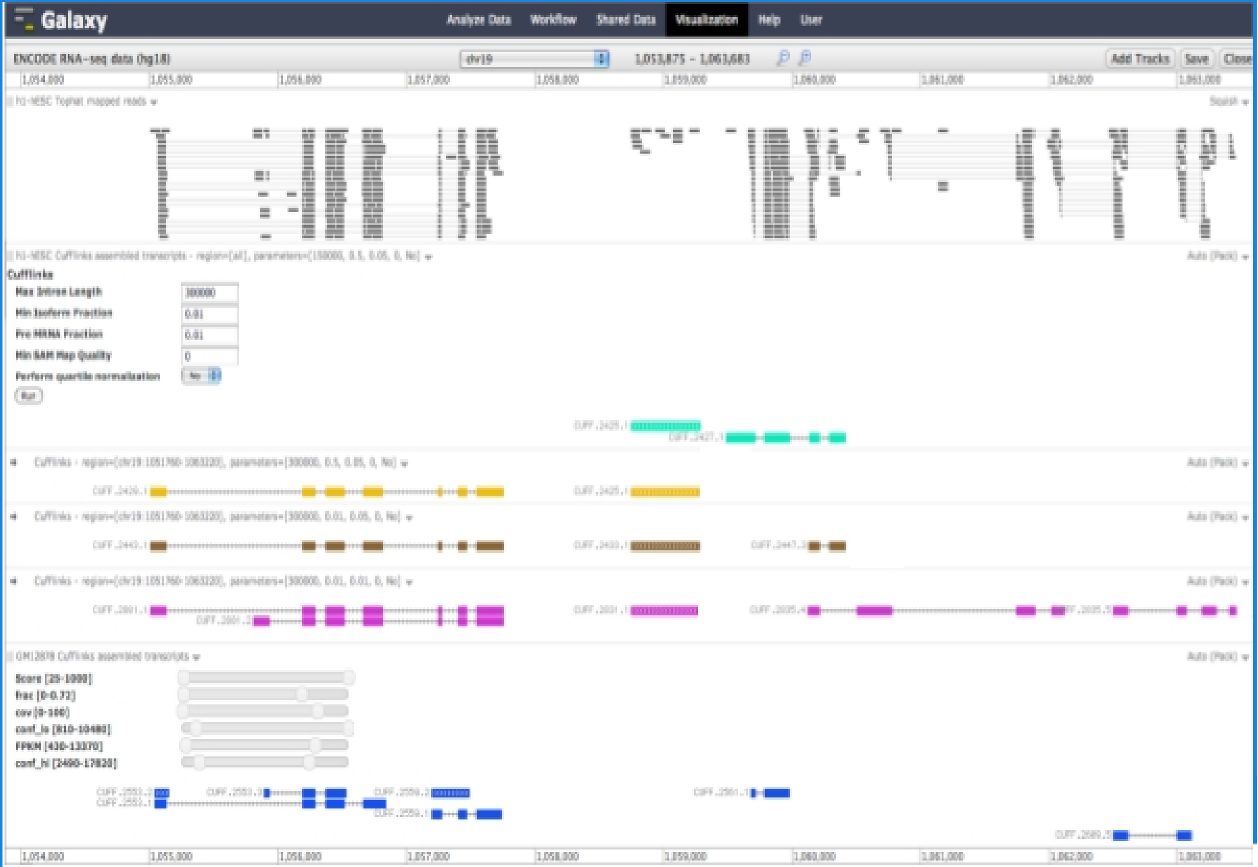
- Assembled genomes as input to Viral Genome ORF Reader (VIGOR)

Wang et al. BMC Bioinformatics 2010, 11:451

- detect coding regions, frame shifts, overlapping and embedded genes

- successfully used for annotating the influenza virus, rotavirus, rhinovirus, coronavirus and subtypes

## Phase III
## Annotation Visualization & Editing


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
ENCODE RNA-seq data (hg18)
chr19
1,053,875 - 1,063,683
Add Tracks
Save
Close
1,054,000
1,055,000
1,056,000
1,057,000
1,058,000
1,059,000
1,060,000
1,061,000
1,062,000
1,063,000
h1-hESC Tophat mapped reads
Squish
h1-hESC Cufflinks assembled transcripts - region=[all], parameters=[150000, 0.5, 0.05, 0, No]
Auto (Pack)
CUFF.2425.1
CUFF.2427.1
Cufflinks - region=[chr19:1051760-1063220], parameters=[300000, 0.5, 0.05, 0, No]
CUFF.2420.1
CUFF.2425.1
Cufflinks - region=[chr19:1051760-1063220], parameters=[300000, 0.01, 0.05, 0, No]
CUFF.2443.1
CUFF.2433.1
CUFF.2447.1
Cufflinks - region=[chr19:1051760-1063220], parameters=[300000, 0.01, 0.01, 0, No]
CUFF.2001.1
CUFF.2001.2
CUFF.2031.1
CUFF.2035.4
CUFF.2035.5
GM12878 Cufflinks assembled transcripts
CUFF.2553.2
CUFF.2553.1
CUFF.2553.3
CUFF.2559.2
CUFF.2559.1
CUFF.2561.1
CUFF.2589.5
Run

Cufflinks
Max Intron Length 300000
Min Isoform Fraction 0.01
Pre MRNA Fraction 0.01
Min SAM Map Quality 0
Perform quartile normalization No

Score [25-1000]
frac [0-0.72]
cov [0-100]
conf_lo [810-10480]
FPKM [430-13370]
conf_hi [2490-17820]

# Research at JCVI with Cloud BioLinux

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

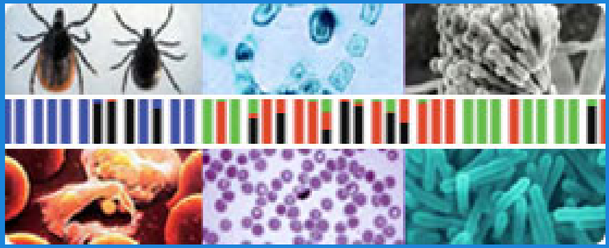
- approach: pre-install pipelines and all their software dependencies in a VM

- export VM on Amazon EC2: pipelines ready to execute, no need to purchase hardware

- users simply need a web browser

- benefits small laboratories that lack resources or expertise

- 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

# Scalable Data Analysis with Cloud BioLinux

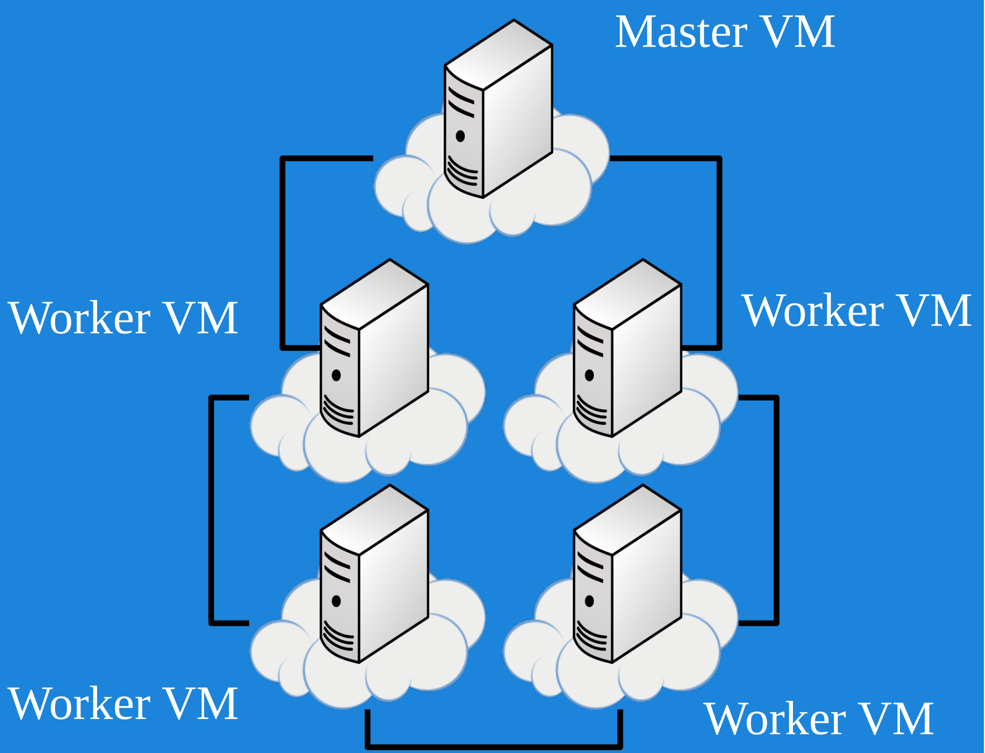
- Sun / Oracle Grid Engine (GE) cluster: computational task scheduling

- Cloud BioLinux VM, dual role: Master or Worker

- Master VM coordinates distribution of computational tasks, Workers runs the computes

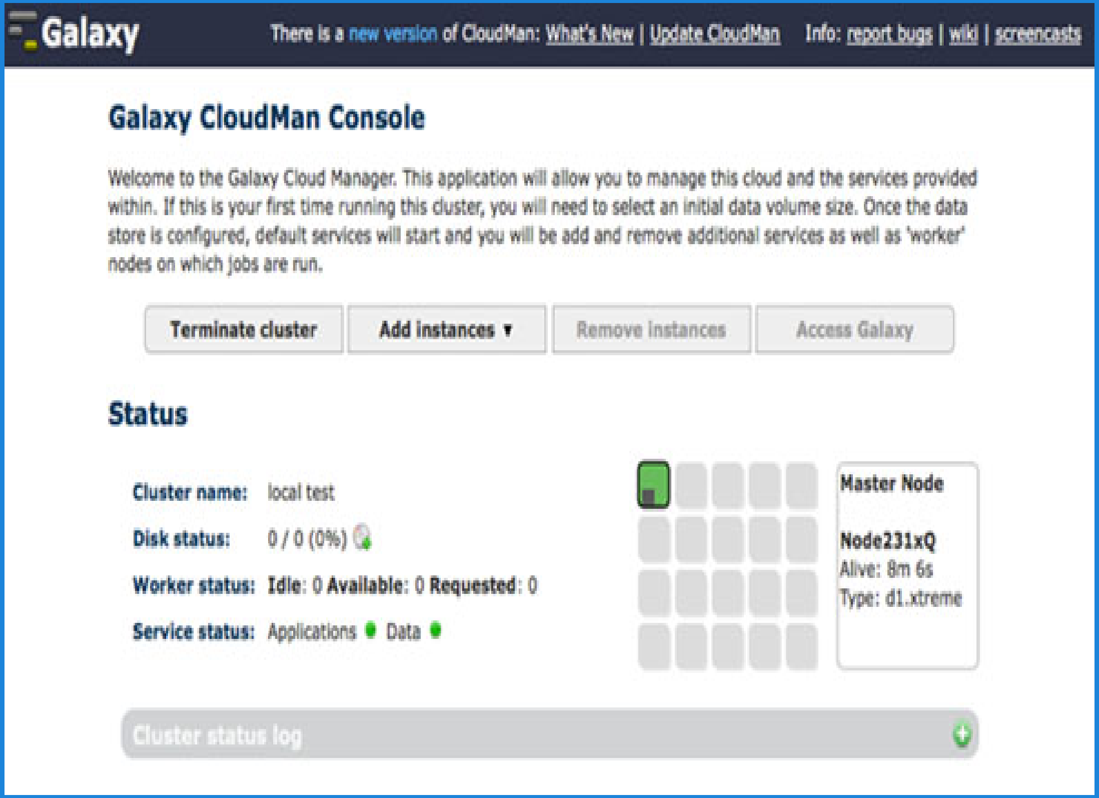
- The Master VM contains all code needed to start Workers and assemble a virtual cluster on the cloud

- Currently works on Amazon EC2



J. Craig Venter™ INSTITUTE

# Scalable Data Analysis with Cloud BioLinux


Galaxy
There is a new version of CloudMan: What's New | Update CloudMan
Info: report bugs | wiki | screencasts
Terminate cluster
Add instances ▾
Remove instances
Access Galaxy
Cluster status log

## Galaxy CloudMan Console

Welcome to the Galaxy Cloud Manager. This application will allow you to manage this cloud and the services provided within. If this is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, default services will start and you will be add and remove additional services as well as 'worker' nodes on which jobs are run.

## Status

Cluster name: local test

Disk status: 0 / 0 (0%)

Worker status: Idle: 0 Available: 0 Requested: 0

Service status: Applications ● Data ●

Master Node

Node231xQ
Alive: 8m 6s
Type: d1.xtreme

- Galaxy Cloudman: users can control size of cluster, storage through a web-browser accessible interface
- Currently in the process of porting to Eucalyptus
- Users can download a VM which can bootstrap GE clusters on their private cloud
- Elastic capacity, size of virtual cluster

Afgan et al. BMC Bioinformatics 2010 11(Suppl 12):S4

Eucalyptus

J. Craig Venter™
INSTITUTE

# Cloud BioLinux for Software Developers

- Issue 1: for researchers with sensitive data a public cloud might not be an option

  moving VMs across clouds is not trivial, need low level operations

- Issue 2: bioinformatic specializations (ex. sequencing, phylogeny, protein structure)

  over-sized VM with too much software for all specializations

- Cloud BioLinux VM deployment framework

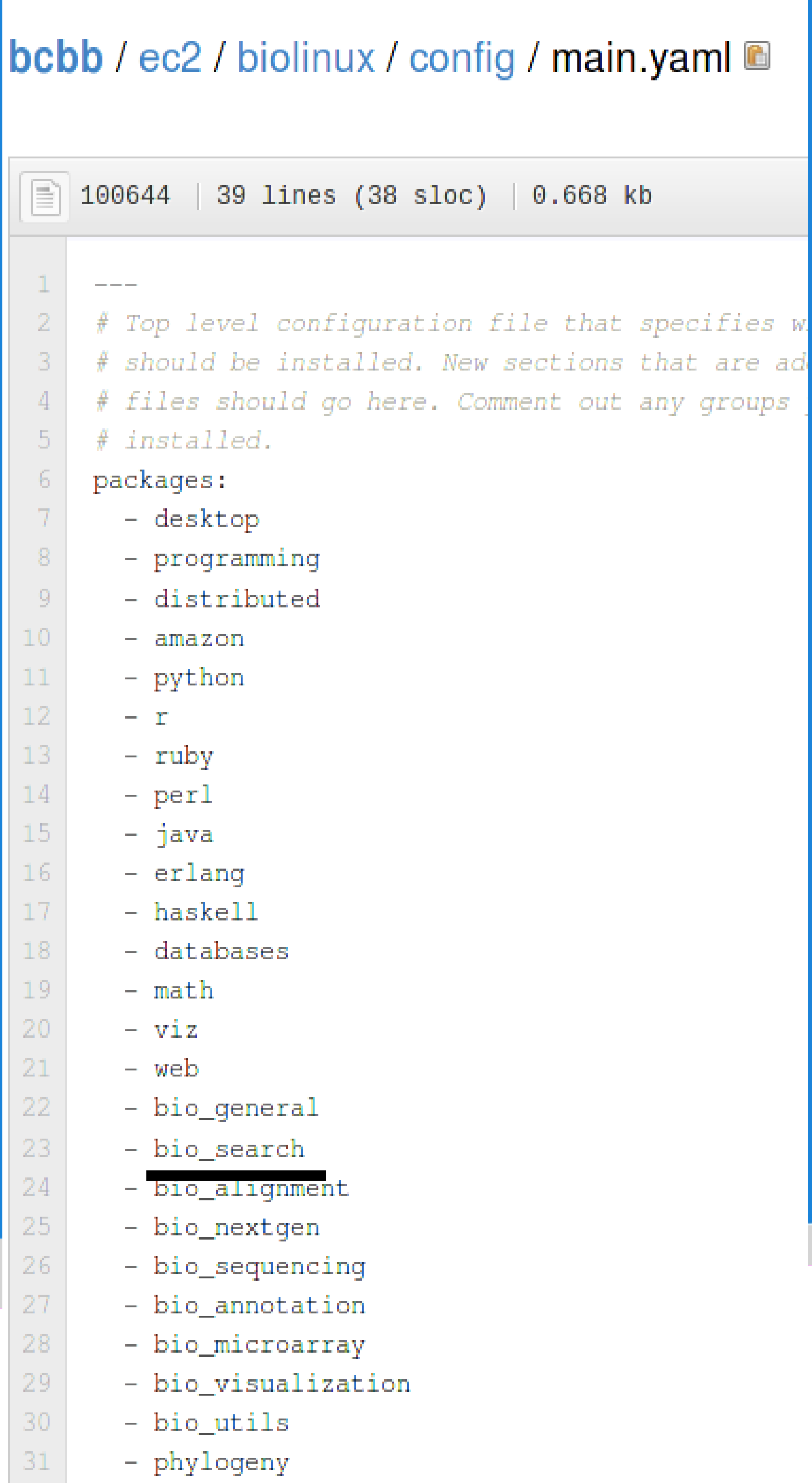
# Framework for Cloud Software Developers

- open-source framework to customize cloud Virtual Machines
- python Fabric automated deployment tool ( DevOps )
- software installed in the VM listed in simple text configuration files
- Fabric scripts automatically pull and install software from repositories
- available from: https://github.com/chapmanb/cloudbiolinux

# software domains in Cloud BioLinux:

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

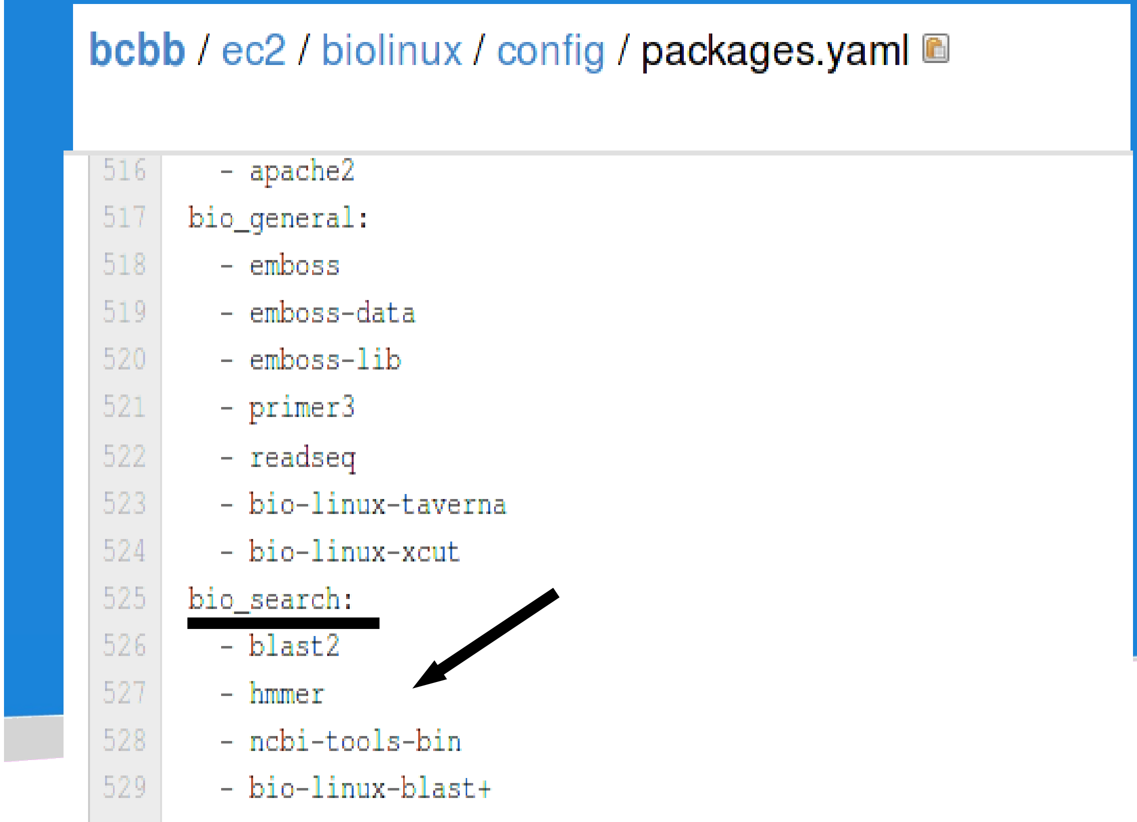
high-level configuration describing software groups for each group individual bioinformatics tools



bcbb / ec2 / biolinux / config / 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
```

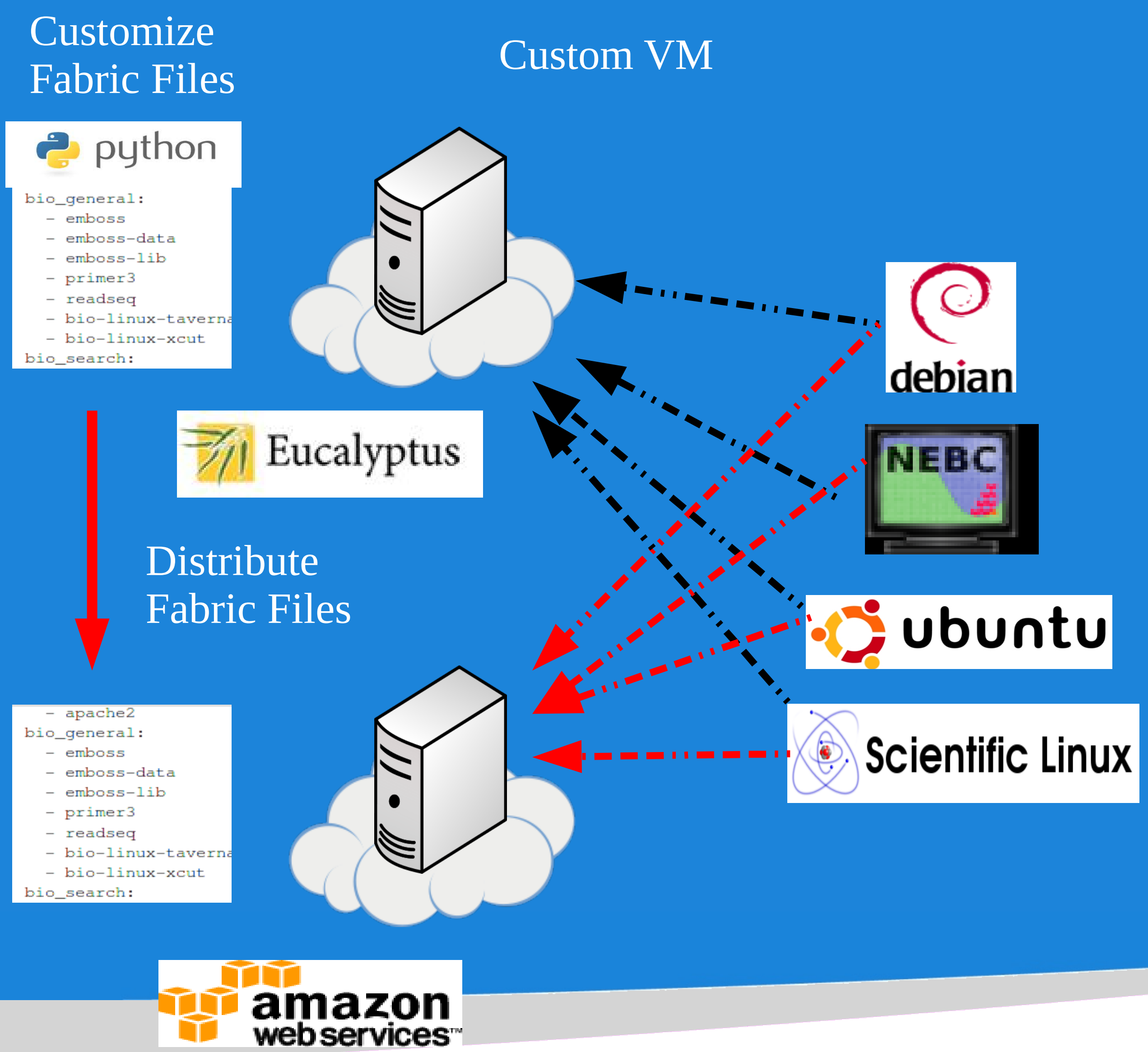


bcbb / ec2 / biolinux / config / packages.yaml

```
  - 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

# Framework for Cloud Software Developers



Customize Fabric Files

Custom VM

```
bio_general:
  - emboss
  - emboss-data
  - emboss-lib
  - primer3
  - readseq
  - bio-linux-taverna
  - bio-linux-xcut
bio_search:
```

Distribute Fabric Files

```
  - apache2
bio_general:
  - emboss
  - emboss-data
  - emboss-lib
  - primer3
  - readseq
  - bio-linux-taverna
  - bio-linux-xcut
bio_search:
```

- start a fresh VM on Amazon or private cloud
- edit Fabric files to mix and match software from repositories – customized VM
- use source code repository to share configuration files
- share configuration of VM as source code

# 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

*Alex Richter* – porting Cloudman to Eucalyptus open-source cloud

JCVI IT dept. - technology support

*Maria Giovanni, Punam Mathur* - NIAID / GSC funding

*Karen Nelson* – JCVI support for Cloud BioLinux

NIAID / OCICB – Bioinformatics Festival

*Thank you !*

*kkrampis@jcvi.org*
*http://www.cloudbiolinux.org*
*http://www.slideshare.com/agbiotec*

J. Craig Venter™
INSTITUTE