I have all this data – now what? Introduction to Bioinformatics Software and Computing Infrastructures

Timothy Stockwell (JCVI)

Vivien Dugan (NIAID/NIH)

NIH National Institute of Allergy and Infectious Diseases

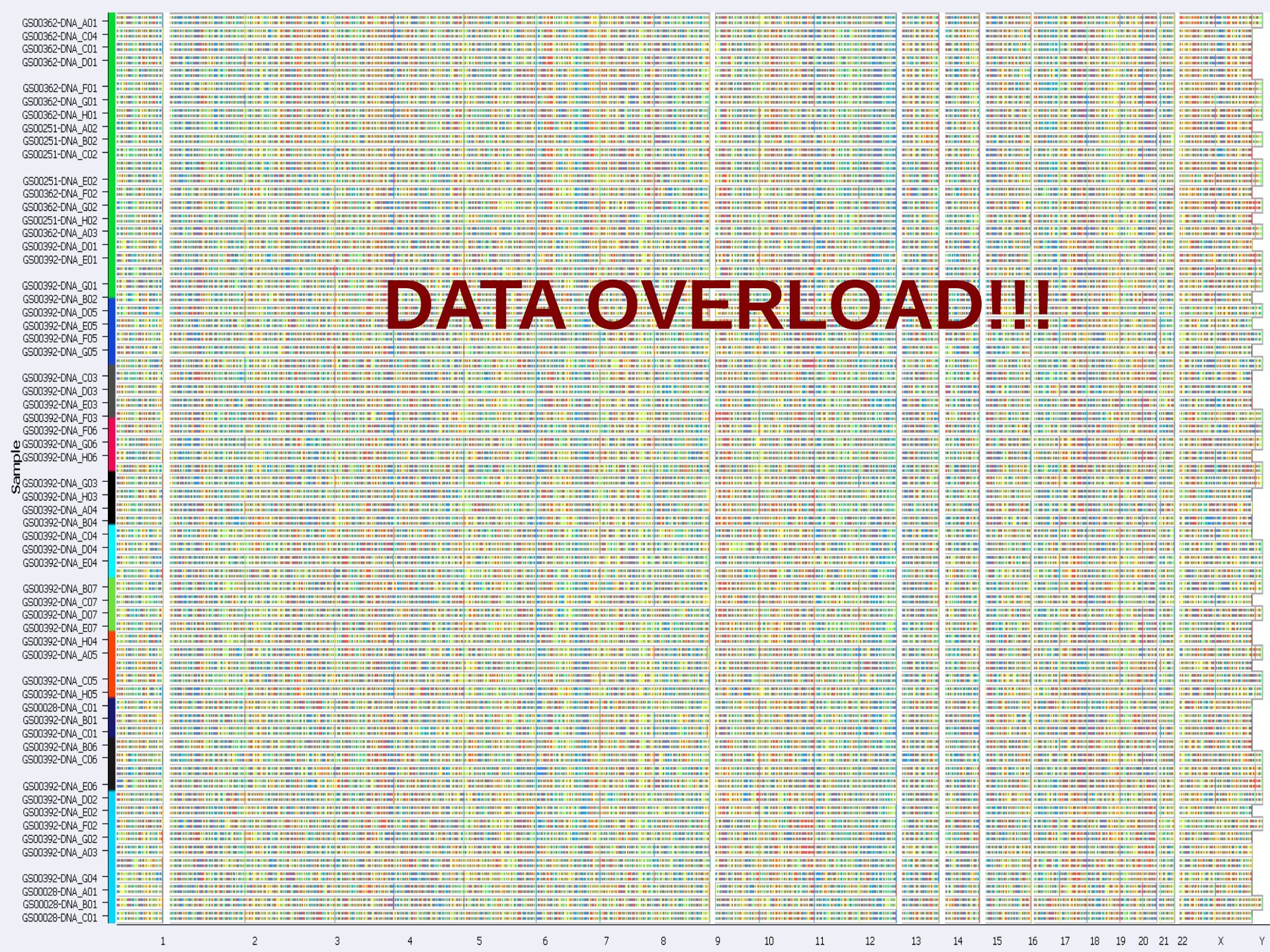
J. Craig Venter INSTITUTE

Bioinformatics

- Bioinformatics
  - Deals with methods for managing and analyzing biological data.
  - Focus on using software to generate useful biological knowledge.
- Bioinformatics for Sequencing and Genomics
  - Sample tracking, LIMS (lab process tracking).
  - Sequencing and assembly
  - Genome annotation (structural and functional)
  - Cross-genome comparisons

My NGS run finished…..


GS00362-DNA_A01
GS00362-DNA_C04
GS00362-DNA_C01
GS00362-DNA_D01
GS00362-DNA_F01
GS00362-DNA_G01
GS00362-DNA_H01
GS00251-DNA_A02
GS00251-DNA_B02
GS00251-DNA_C02
GS00251-DNA_E02
GS00362-DNA_F02
GS00362-DNA_G02
GS00251-DNA_H02
GS00362-DNA_A03
GS00392-DNA_D01
GS00392-DNA_E01
GS00392-DNA_G01
GS00392-DNA_B02
GS00392-DNA_D05
GS00392-DNA_E05
GS00392-DNA_F05
GS00392-DNA_G05
GS00392-DNA_C03
GS00392-DNA_D03
GS00392-DNA_E03
GS00392-DNA_F03
GS00392-DNA_F06
GS00392-DNA_G06
GS00392-DNA_H06
GS00392-DNA_G03
GS00392-DNA_H03
GS00392-DNA_A04
GS00392-DNA_B04
GS00392-DNA_C04
GS00392-DNA_D04
GS00392-DNA_E04
GS00392-DNA_B07
GS00392-DNA_C07
GS00392-DNA_D07
GS00392-DNA_E07
GS00392-DNA_H04
GS00392-DNA_A05
GS00392-DNA_C05
GS00392-DNA_H05
GS00028-DNA_C01
GS00392-DNA_B01
GS00392-DNA_C01
GS00392-DNA_B06
GS00392-DNA_C06
GS00392-DNA_E06
GS00392-DNA_D02
GS00392-DNA_E02
GS00392-DNA_F02
GS00392-DNA_G02
GS00392-DNA_A03
GS00392-DNA_G04
GS00028-DNA_A01
GS00028-DNA_B01
GS00028-DNA_C01
Sample
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y


GS00362-DNA_A01
GS00362-DNA_C04
GS00362-DNA_C01
GS00362-DNA_D01
GS00362-DNA_F01
GS00362-DNA_G01
GS00362-DNA_H01
GS00251-DNA_A02
GS00251-DNA_B02
GS00251-DNA_C02
GS00251-DNA_E02
GS00362-DNA_F02
GS00362-DNA_G02
GS00251-DNA_H02
GS00362-DNA_A03
GS00392-DNA_D01
GS00392-DNA_E01
GS00392-DNA_G01
GS00392-DNA_B02
GS00392-DNA_D05
GS00392-DNA_E05
GS00392-DNA_F05
GS00392-DNA_G05
GS00392-DNA_C03
GS00392-DNA_D03
GS00392-DNA_E03
GS00392-DNA_F03
GS00392-DNA_F06
GS00392-DNA_G06
GS00392-DNA_H06
GS00392-DNA_G03
GS00392-DNA_H03
GS00392-DNA_A04
GS00392-DNA_B04
GS00392-DNA_C04
GS00392-DNA_D04
GS00392-DNA_E04
GS00392-DNA_B07
GS00392-DNA_C07
GS00392-DNA_D07
GS00392-DNA_E07
GS00392-DNA_H04
GS00392-DNA_A05
GS00392-DNA_C05
GS00392-DNA_H05
GS00028-DNA_C01
GS00392-DNA_B01
GS00392-DNA_C01
GS00392-DNA_B06
GS00392-DNA_C06
GS00392-DNA_E06
GS00392-DNA_D02
GS00392-DNA_E02
GS00392-DNA_F02
GS00392-DNA_G02
GS00392-DNA_A03
GS00392-DNA_G04
GS00028-DNA_A01
GS00028-DNA_B01
GS00028-DNA_C01
Sample
DATA OVERLOAD!!!
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y


GS00362-DNA_A01
GS00362-DNA_C04
GS00362-DNA_C01
GS00362-DNA_D01
GS00362-DNA_F01
GS00362-DNA_G01
GS00362-DNA_H01
GS00251-DNA_A02
GS00251-DNA_B02
GS00251-DNA_C02
GS00251-DNA_E02
GS00362-DNA_F02
GS00362-DNA_G02
GS00251-DNA_H02
GS00362-DNA_A03
GS00392-DNA_D01
GS00392-DNA_E01
GS00392-DNA_G01
GS00392-DNA_B02
GS00392-DNA_D05
GS00392-DNA_E05
GS00392-DNA_F05
GS00392-DNA_G05
GS00392-DNA_C03
GS00392-DNA_D03
GS00392-DNA_E03
GS00392-DNA_F03
GS00392-DNA_F06
GS00392-DNA_G06
GS00392-DNA_H06
GS00392-DNA_G03
GS00392-DNA_H03
GS00392-DNA_A04
GS00392-DNA_B04
GS00392-DNA_C04
GS00392-DNA_D04
GS00392-DNA_E04
GS00392-DNA_B07
GS00392-DNA_C07
GS00392-DNA_D07
GS00392-DNA_E07
GS00392-DNA_H04
GS00392-DNA_A05
GS00392-DNA_C05
GS00392-DNA_H05
GS00028-DNA_C01
GS00392-DNA_B01
GS00392-DNA_C01
GS00392-DNA_B06
GS00392-DNA_C06
GS00392-DNA_E06
GS00392-DNA_D02
GS00392-DNA_E02
GS00392-DNA_F02
GS00392-DNA_G02
GS00392-DNA_A03
GS00392-DNA_G04
GS00028-DNA_A01
GS00028-DNA_B01
GS00028-DNA_C01
Sample
DATA OVERLOAD!!!
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y


GS00362-DNA_A01
GS00362-DNA_C04
GS00362-DNA_C01
GS00362-DNA_D01
GS00362-DNA_F01
GS00362-DNA_G01
GS00362-DNA_H01
GS00251-DNA_A02
GS00251-DNA_B02
GS00251-DNA_C02
GS00251-DNA_E02
GS00362-DNA_F02
GS00362-DNA_G02
GS00251-DNA_H02
GS00362-DNA_A03
GS00392-DNA_D01
GS00392-DNA_E01
GS00392-DNA_G01
GS00392-DNA_B02
GS00392-DNA_D05
GS00392-DNA_E05
GS00392-DNA_F05
GS00392-DNA_G05
GS00392-DNA_C03
GS00392-DNA_D03
GS00392-DNA_E03
GS00392-DNA_F03
GS00392-DNA_F06
GS00392-DNA_G06
GS00392-DNA_H06
GS00392-DNA_G03
GS00392-DNA_H03
GS00392-DNA_A04
GS00392-DNA_B04
GS00392-DNA_C04
GS00392-DNA_D04
GS00392-DNA_E04
GS00392-DNA_B07
GS00392-DNA_C07
GS00392-DNA_D07
GS00392-DNA_E07
GS00392-DNA_H04
GS00392-DNA_A05
GS00392-DNA_C05
GS00392-DNA_H05
GS00028-DNA_C01
GS00392-DNA_B01
GS00392-DNA_C01
GS00392-DNA_B06
GS00392-DNA_C06
GS00392-DNA_E06
GS00392-DNA_D02
GS00392-DNA_E02
GS00392-DNA_F02
GS00392-DNA_G02
GS00392-DNA_A03
GS00392-DNA_G04
GS00028-DNA_A01
GS00028-DNA_B01
GS00028-DNA_C01
Sample
DATA OVERLOAD!!!
HELP
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y

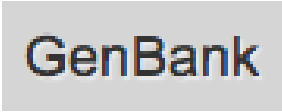
Genomics Resources

- General resources
- Genomics resources
- Bioinformatics resources
- Pathogen-specific resources

USA NIH National Center for Biomedical Information (NCBI)

- **NCBI** – home page, http://www.ncbi.nlm.nih.gov

- **GenBank** – genetic sequence database
  http://www.ncbi.nlm.nih.gov/genbank



- **PubMed** – database of citations and links to over 22 million biomedical articles
  http://www.ncbi.nlm.nih.gov/pubmed

- **BLAST** – **B**asic **L**ocal **A**lignment **S**earch **T**ool – to search for similar biological sequences
  http://blast.ncbi.nlm.nih.gov/Blast.cgi

Bioinformatics Resource Centers (BRCs)

https://vectorbase.org/

EuPathDB Version 2.13 11 Jan 12
Eukaryotic Pathogen Database Resources

http://eupathdb.org/

PATRIC
PATHOSYSTEMS RESOURCE INTEGRATION CENTER

http://patric.vbi.vt.edu/

PATHOGENPORTAL
THE BIOINFORMATICS RESOURCE CENTERS PORTAL

http://www.pathogenportal.org/

ViPR
Virus Pathogen Resource

http://www.viprbrc.org/

IRD Influenza Research Database

http://www.fludb.org/

NIH National Institute of Allergy and Infectious Diseases

J. Craig Venter INSTITUTE

NIAID Bioinformatics Resource Centers Bioinformatic Services

- Community-based Database & Bioinformatics Resource Centers
- Partnerships with Infectious Diseases Research and Public Health communities
- Genomic, omics, experimental, & clinical metadata available with analysis tools
  - PATRIC RAST: Free prokaryotic genome annotation
  - RNA-Seq analysis : Free RNA-seq data processing and pipeline for
- Data submission to GenBank/NCBI
- Data Analysis tools and workbenches
- Training with workshops globally

NIAID Genomic Sequencing Centers for Infectious Diseases

J. Craig Venter INSTITUTE

UNIVERSITY of MARYLAND SCHOOL OF MEDICINE INSTITUTE FOR GENOME SCIENCES

BROAD INSTITUTE

- Sample Processing Method Develop
- High Throughput Sequencing Pipelines
- Metagenomics Transcriptomics
- Bioinformatics Tools Data Analysis Pipelines
- Genomics Bioinformatics Training

GSC Bioinformatics

	Viral Genomes	Prokaryotic Genomes	Eukaryotic Genomes
Library Prep	*JCVI Primer Designer* and PCR, SISPA barcoding of each sample, vendor barcoding of SISPA Sample Pools	Whole Genome Shotgun (WGS) libraries of multiple insert sizes	WGS libraries of multiple insert sizes, RNA-seq libraries
Sequencing	454, Illumina, IonTorrent – library tracking in *JLIMS*, sample tracking in JIRA	454, Illumina	454, Illumina
Assembly and Finishing	CLC bio de novo and mapping assemblers and consed for finishing	Newbler, *CABOG*, velvet, CLC bio de novo and mapping assemblers	*CABOG,* CLC bio, Newbler
Annotation Tools	*VIGOR*	*Prokaryotic Annotation Pipeline, MGAT, MANATEE*	*PASA, MANATEE, EVM, GSAC,* Trinity, RnNotator, Tuxedo Package,
Comparative Genomics	*ANDES Tools*, Phylogenetic Tools	*Prokaryotic Pan-Genome Pipeline, PanOCT, SYBIL*	GBrowse , OrthoMCL, Artemis, *SYBIL*

Italics indicate software developed at JCVI.

NIH National Institute of Allergy and Infectious Diseases

The Sequencing App Store

Chad Nussbaum
Broad Institute

Just a sampling...

Examples: Genome Software Tools

- **ALL PATHS**: Assembly of Short Reads

RNA-Seq De novo Assembly

- Using RNA-Seq for eukaryotic genome annota
  - Inchworm RNA-Seq Assembler
    - Available at http://inchworm.sf.net
  - enhanced PASA: incorporate RNA-Se(annotations.
    - Available at http://pasa.sf.net

- Automated accurate prokaryotic genome annotation
  - DASH: Draft genome Annotation by Strength of Homology evidence.
    - soon to be released at http://genedash.sf.net

Some More Virus Resources

- **ICTV** – International Committee on Taxonomy of Viruses, nomenclature of viruses, http://ictvonline.org/index.asp

- **ViralZone** – provides general information on most viruses, including molecular and epidemiological information, virion and genome diagrams, and links to other information, http://viralzone.expasy.org/

Some More Bacteria Resources

- **ICSP** – International Committee on Systematics of Prokaryotes, nomenclature of prokaryotes, http://www.the-icsp.org/

- **JGI** – US Department of Energy's Joint Genome Institute, focuses on genomics for energy and the environment, http://www.jgi.doe.gov/

- **CMR** – JCVI's Comprehensive Microbial Resource, has tools for managing, searching and comparing microbial genomes, however, no longer updated, http://cmr.jcvi.org/cgi-bin/CMR/CmrHomePage.cgi

NIH National Institute of Allergy and Infectious Diseases

JCVI CMR

J. Craig Venter INSTITUTE

Computational Resources

- JIRA Sample Tracking
- Open Source Software Tools
- Bioinformatics libraries for Software Development
- Computing Infrastructures

JIRA Sample Tracking

- JCVI track thousands of Samples from arrival to publication. To keep track of all of this information a customized version of Atlassian's JIRA is used. JIRA was originally created for software support but contains a wide array of customization options and interfaces.
- The JIRA Sample Tracking System mirrors the Viral pipelines combination of high throughput and high flexibility. Large groups of samples can be updated as a single operation; through JIRA's GUI or via command line scripts. Each sample can also be treated individually when special processing is required and then return to being bulk updated.
- A sample's status is only part of its meta-data. The sample tracking system is integrated with the 'Viral Genome Database'. Gathering information from the system and synchronizing changes to its status.
- Reporting is key to using the information gathered by the system. It has been integrated into several highly tailored status reports and the JCVI website.

J. Craig Venter™
INSTITUTE

Open Source Software Tools

- General Packages for Manipulating Data
  - **Galaxy** – web-based platform for biomedical research, contains many of the tools below, http://galaxyproject.org/
  - **EMBOSS** - The European Molecular Biology Open Software Suite, has something for everything, as long as you have UNIX, http://emboss.sourceforge.net/
- Mapping/Aligning Reads to a Reference
  - **Bowtie** – short read aligner, http://bowtie-bio.sourceforge.net/index.shtml
  - **BWA** - Burrows-Wheeler Aligner for short reads, http://bio-bwa.sourceforge.net/
  - **GATK** – toolkit for analyzing resequencing data, http://www.broadinstitute.org/gatk/
  - **SAMtools** – tools for processing Sequence Alignment/Map format files, http://samtools.sourceforge.net/

J. Craig Venter™ INSTITUTE

More Open Source Software Tools

- De novo Assembly Software
  - **Phred, phrap, consed** – for base calling Sanger chromatograms, de novo assembly, and reviewing/editing the results, http://www.phrap.org/phredphrapconsed.html (free for academic and non profit)
  - **Velvet** – a de novo assembler for short reads, http://www.ebi.ac.uk/~zerbino/velvet/
- Mapping AND De Novo Assembly
  - **SOAP** – Short Oligonucleotide Analysis Package, http://soap.genomics.org.cn/
  - **Mira-assembler** – Sequence assembler and mapper for whole genome shotgun and EST / RNASeq sequencing data, http://mira-assembler.sourceforge.net/

Bioinformatics Libraries to help when developing software

- **BioPerl** (Perl modules) – http://www.bioperl.org
- **BioJava** (Java tools) – http://www.biojava.org
- **Biopython** (Python tools) – http://www.biopython.org
- **BioRuby** (Ruby classes) – http://bioruby.open-bio.org
- **BioPHP** (PHP code) – http://www.biophp.org
- **BioConductor** (R tools) – http://www.bioconductor.org

Computing Infrastructures for Bioinformatics – option 1

- Your own computer – you control the hardware, the operating system, and the software installed
  - Install individual tools and programs yourself – might have to compile them from source code – very flexible, you can keep the versions up-to-date, the most amount of work
  - Install an integrated bioinformatics tool suite – examples include BioEdit and Ugene (free) or CLC Bio (not free) – less flexible, you get the whole suite, with the tools/versions that are included, but less work
  - Install a complete Linux workstation, OS and tools – BioLinux - http://envgen.nox.ac.uk/tools/bio-linux

Computing Infrastructures for Bioinformatics – option 2

- Your institute's data center – typically, an IT department controls the hardware, the operating system, and the software installed
  - IT administrators install individual tools and programs – they might have to compile them from source code – very flexible, but often very expensive.
  - Need enough data processing demand to justify both the hardware and the IT staff to keep everything operating smoothly
  - Good for large sequencing centers and other organizations with large computing requirements

Computing Infrastructures for Bioinformatics – option 3

- The cloud – pay by the hour computing at a data center accessible via the internet
  - Provides virtual machines and storage at data centers located around the world
  - An option when bioinformatics needs “come and go”, but when they come, there is too much processing for a single computer to handle

Sequencers shipped without clusters

- *Problem A* : sequence analysis requires computational capacity
- genome assembly, BLAST, gene finders - annotation
- *Problem B*: bioinformatics tools need software engineering expertise
- unix/linux operating systems, maintaining software libraries, compiling source code

???

J. Craig Venter INSTITUTE

Each lab builds a cluster ?

- need additional funds to buy the hardware
- funds for personnel to maintain the cluster and software
- duplication of effort across labs
- sub-optimal utilization of the hardware
- few sequencing runs per year

J. Craig Venter™ INSTITUTE

Problem A : sequence analysis requires computational capacity

- Amazon Elastic Compute Cloud (EC2), pay-by-the-hour computing
- cloud servers cost $0.085 - $2 per hour
- max capacity 64GB RAM / 8 CPU (can boot hundreds of servers)

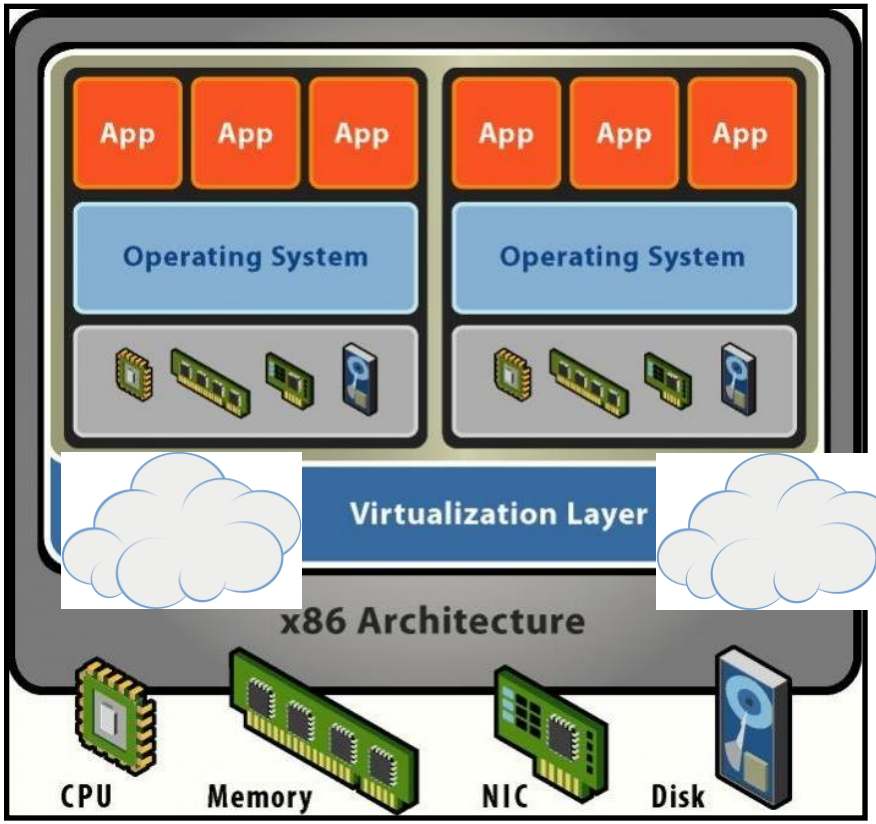
amazon web services™ World-wide data centers

750 hours free for new users: aws.amazon.com/free/
free compute for teaching: aws.amazon.com/grants/

J. Craig Venter™ INSTITUTE

Problem B: bioinformatics tools need software engineering expertise

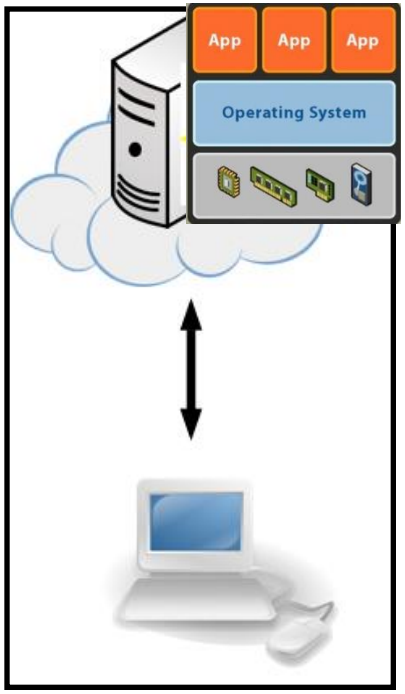
- OS, software, data, pre-installed in Virtual Machine (VM)
- VM is a full-featured server in a binary, downloadable file
- avoid compiling source code, or other software dependencies



J. Craig Venter INSTITUTE

Solving Problems A & B : Cloud BioLinux

- Cloud BioLinux: publicly accessible VM on EC2
- 100+ pre-installed bioinformatics tools
- remote desktop for non-command line experts
- comes with Galaxy - CloudMan

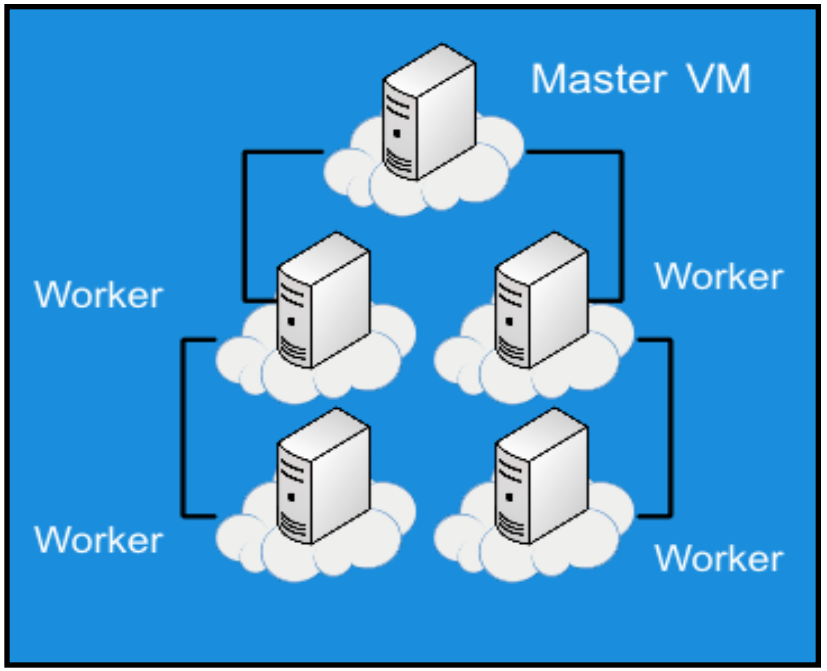


Krampis K, Booth T, Chapman B, Tiwari B, Bicak M, Field D, Nelson K

Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community.
BMC Bioinformatics. 2012 Mar 19; 13:42

Compute Clusters on the Cloud

- Cloud BioLinux creates Sun Grid Engine (SGE) clusters on the Amazon Cloud
- Cloud BioLinux + Cloudman scripts boot SGE worker VMs
- Multiple genome runs in parallel: each worker runs one genome.



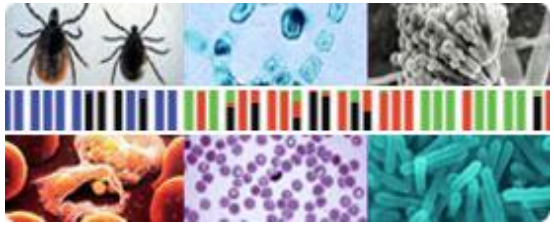
Afgan, E., Chapman, B. et al. (2012). Using Cloud Computing Infrastructure with CloudBioLinux, CloudMan, and Galaxy. *Current Protocols in Bioinformatics*, 11-9.

J. Craig Venter INSTITUTE

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 Virtual Machine (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

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

J. Craig Venter™ INSTITUTE

JCVI GSCID Cloud project for Viral and Prokaryotic genome annotation pipelines

- Enabled public access to JCVI's GSCID data analysis pipelines for Viral and Prokaryotic genomes, using Virtual Machine (VM) servers on the Amazon EC2 Cloud
- VM servers are available for private Eucalyptus and OpenStack Clouds (http://www.eucalyptus.org, http://www.openstack.org), or desktop computers using the VirtualBox software (http://www.virtualbox.org).
- Eucalyptus private cloud has been deployed at JCVI and will become publicly accessible to the community in the fall of 2013.
- The research findings from the GSC Cloud project have been summarized in a set of manuscripts written in the first quarter of 2013 (see references below)
- *References:*
- *Alin V, Anton G, Krampis K, Jing W, Lam N, Mazmuder R, Simoyan S, Hayley D. (2013) High performance integrated virtual environment (hive) for next generation sequencing analysis. BioIT-Wold Conference and Expo, Boston, MA*
- *Krampis K, Sarangi V, Sutton G. (2013) Six Questions and Answers Defining Cloud Computing for Digital, Sequencing-Based Biological Research. BMC Bioinformatics; (in peer review).*
- *Kumari P, Krampis K. (2013) Advantages of distributed and parallel algorithms that leverage Cloud Computing platforms for large-scale genome assembly. BMC Research Reports; (in peer review).*
- *Krampis K, Inman J, Richter A, Sanka R, Tovichgrechko A, Sutton G, Stockwell T. (2013) Viral and Prokaryotic genome data analysis using cloud computing platforms. BMC Bioinformatics; (in preparation).*

J. Craig Venter INSTITUTE

From sequencer to the cloud

Galaxy

credit: basespace.illumina.com

J. Craig Venter INSTITUTE

Questions?

J. Craig Venter™
INSTITUTE

Acknowledgments

- Cloud BioLinux community: *Brad Chapman, Enis Afgan,Tim Booth, Mesude Bicak, Dawn Field*
- JCVI group and collaborators: *Alex Richter, Ravi Sanka, Andrey Tovichgrechko, Karen Nelson, Bill Nierman, JCVI IT.*
- NIAID and for funding: *Maria Giovani, Punam Mathur*

cloudbiolinux.org

groups.google.com/group/cloudbiolinux

tinyurl.com/cloudboot1

tinyurl.com/cloudboot2

kkrampis@jcvi.org

slideshare.com/agbiotec

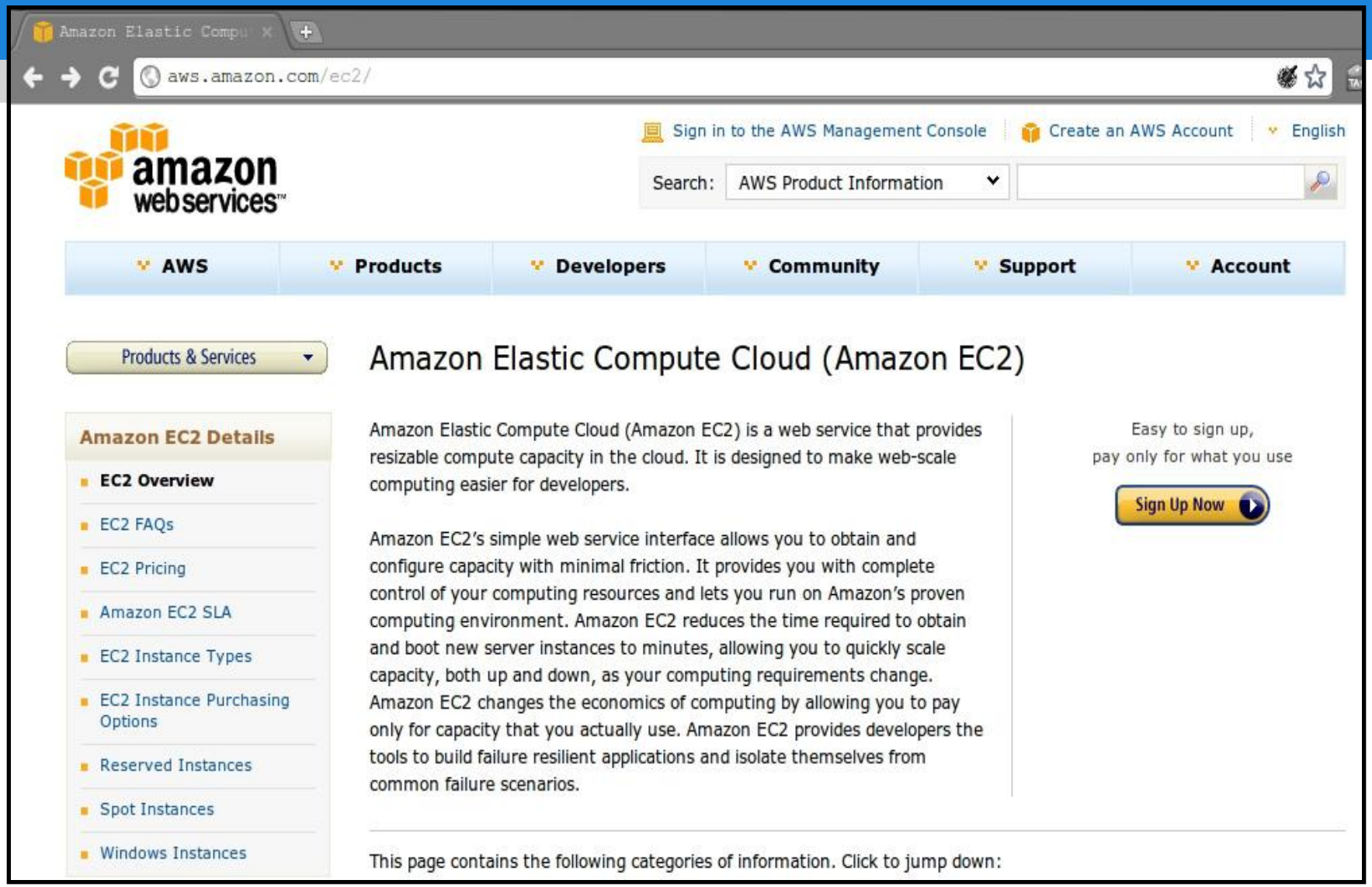
Thank you !

J. Craig Venter INSTITUTE

More Cloud research at JCVI: breaking bioinformatics software silos

- open-source Clouds, fully compatible with Amazon
- one Cloud BioLinux VM with pre-installed pipelines to run across all
- just copy and boot the VM - no pipeline / tools modification required
- collaborators have choice of Amazon, private cloud, or desktop

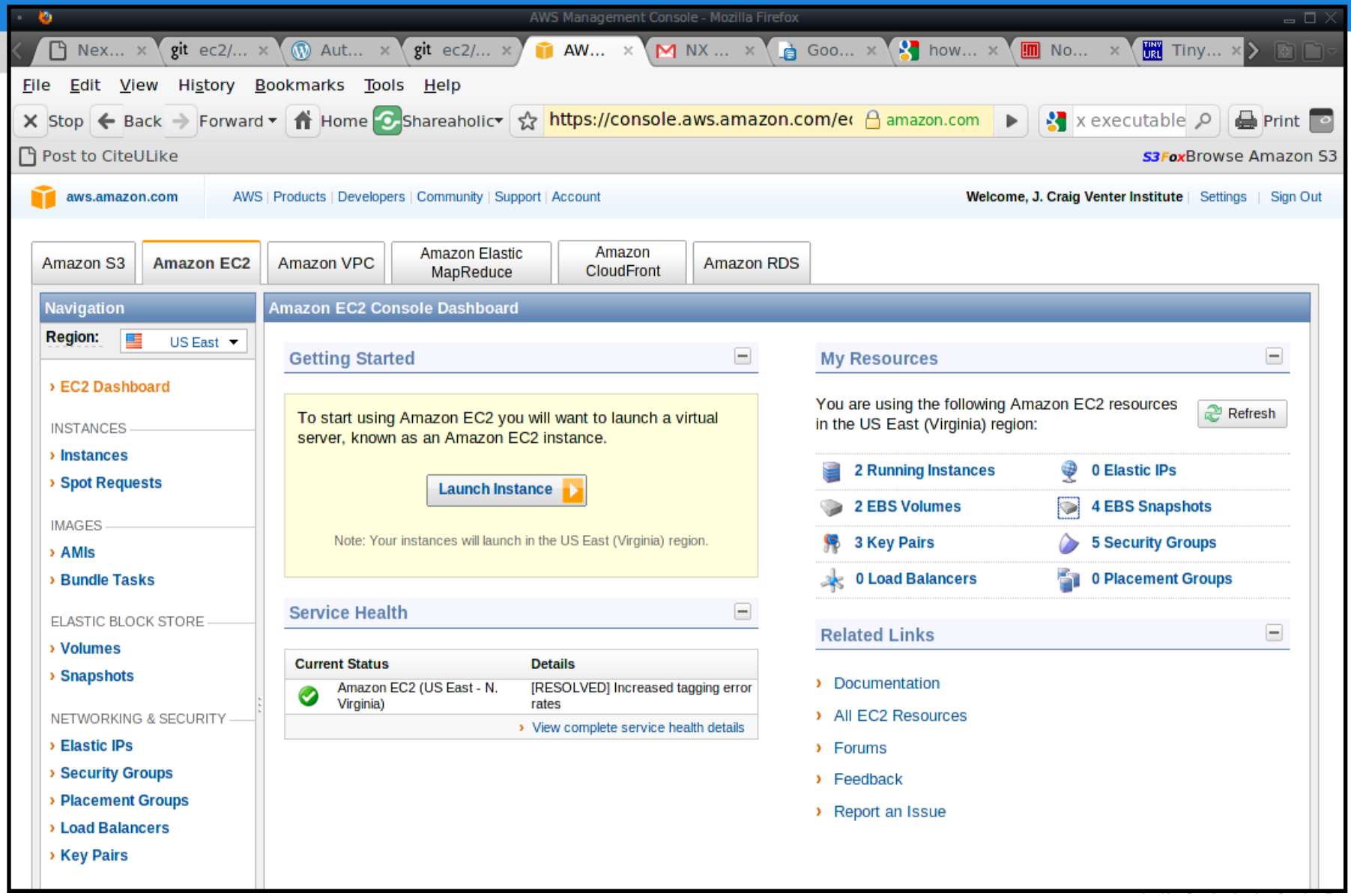
Accessing Cloud BioLinux



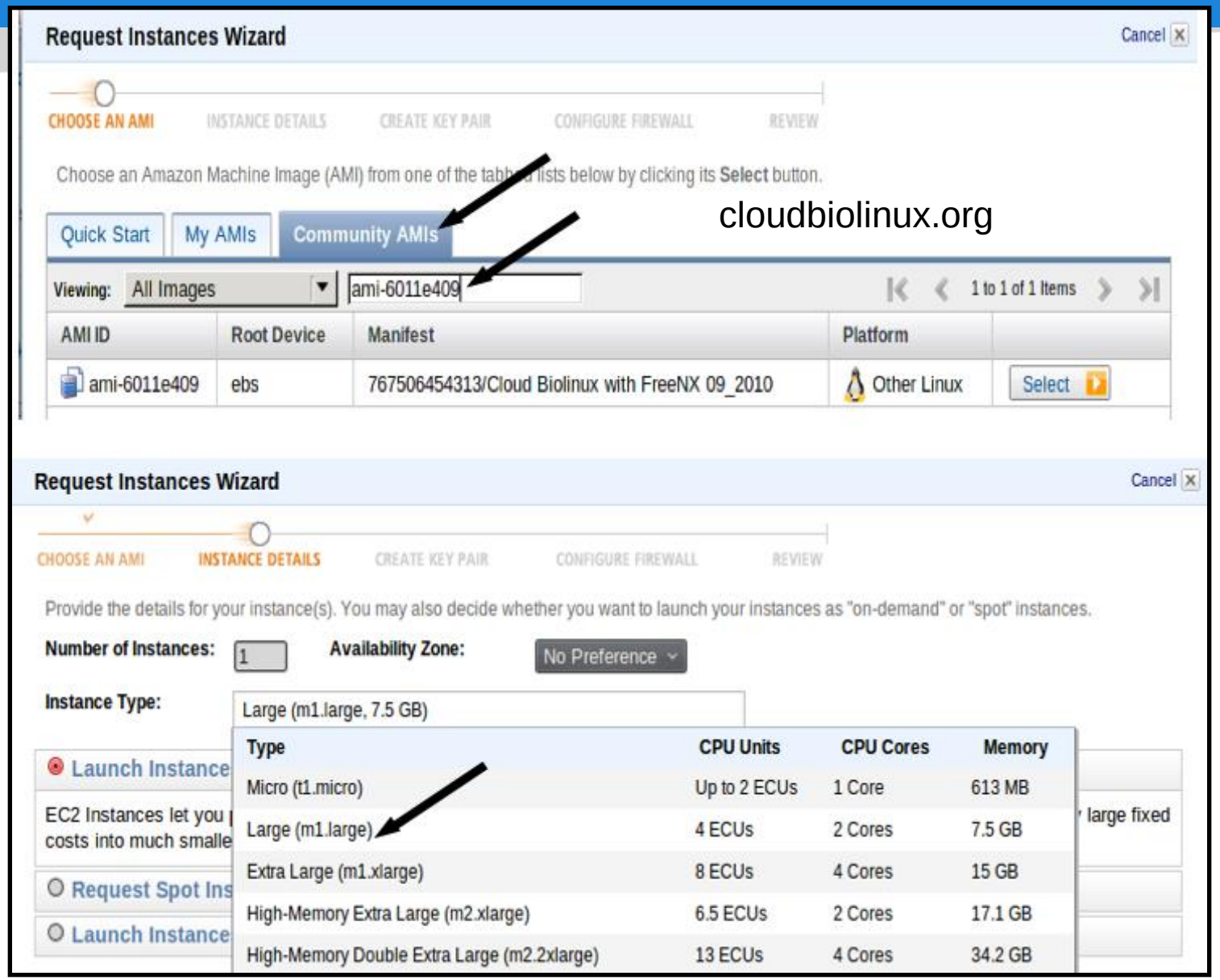
aws.amazon.com/console

J. Craig Venter INSTITUTE

Starting a VM: EC2 cloud console



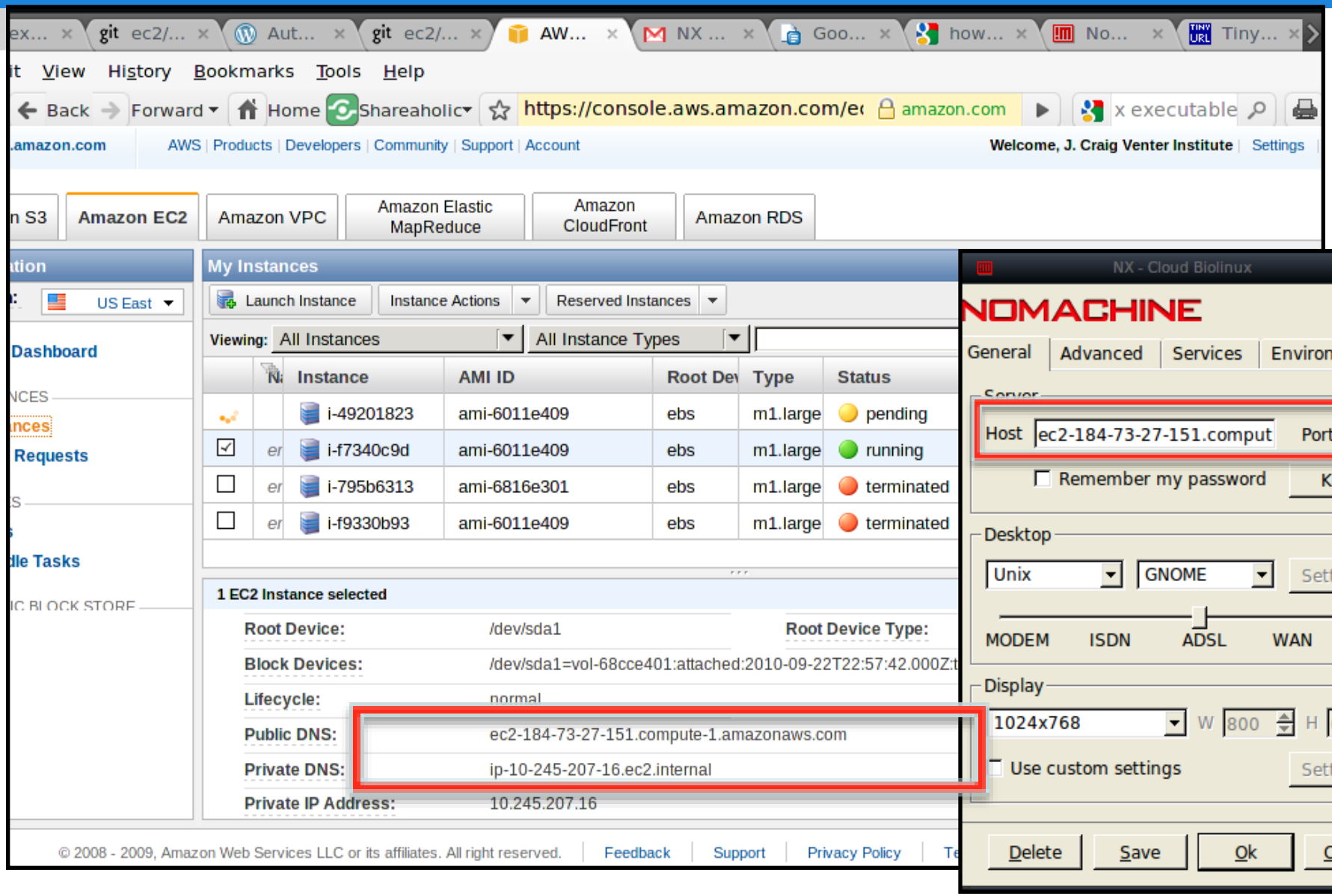
Amazon EC2 VM launch wizard



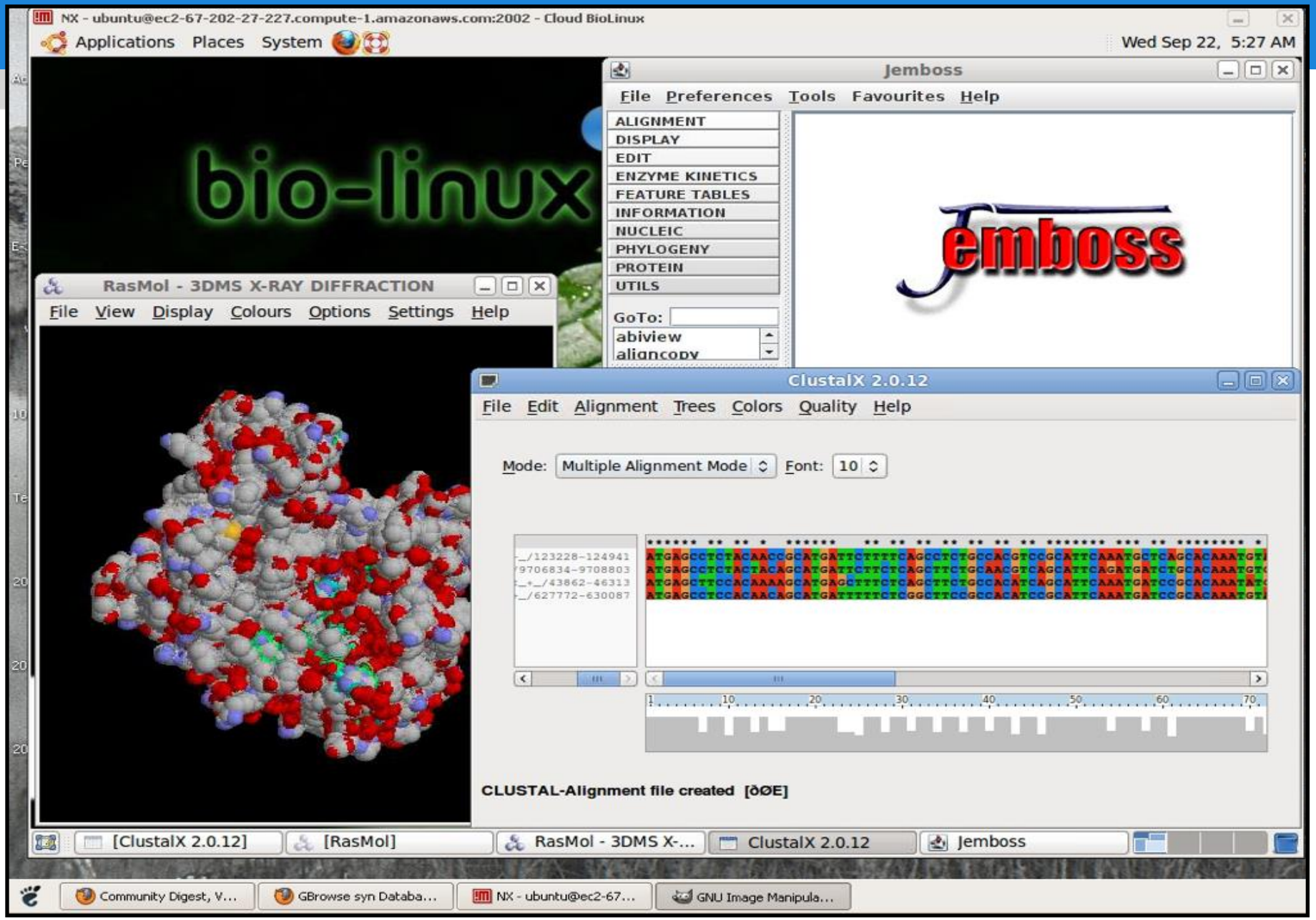
J. Craig Venter INSTITUTE

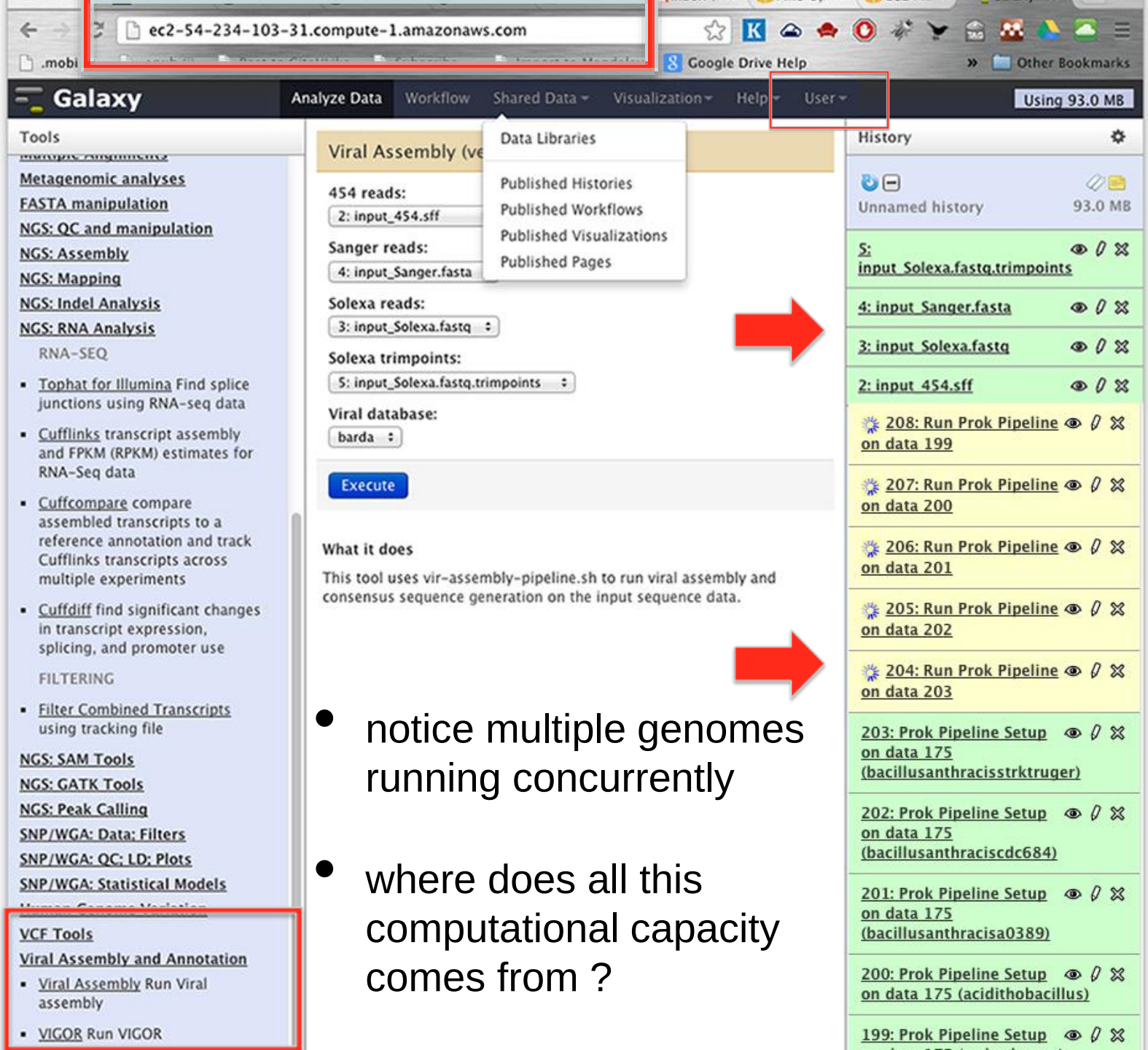
Cloud BioLinux remote desktop connection

Video screencasts : tinyurl.com/bootcloud1 tinyurl.com/bootcloud2



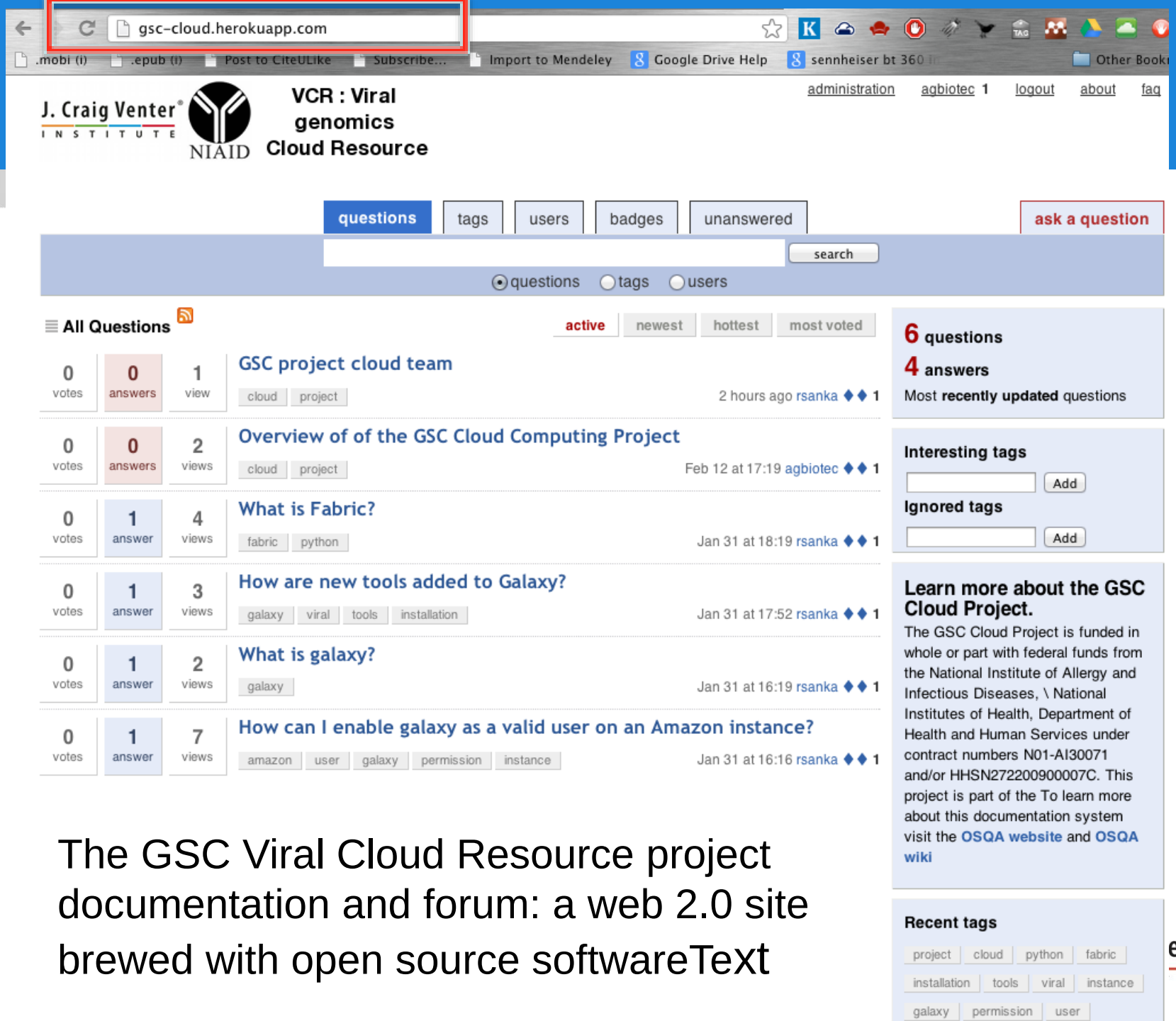
Cloud BioLinux remote desktop connection





- notice multiple genomes running concurrently
- where does all this computational capacity comes from ?

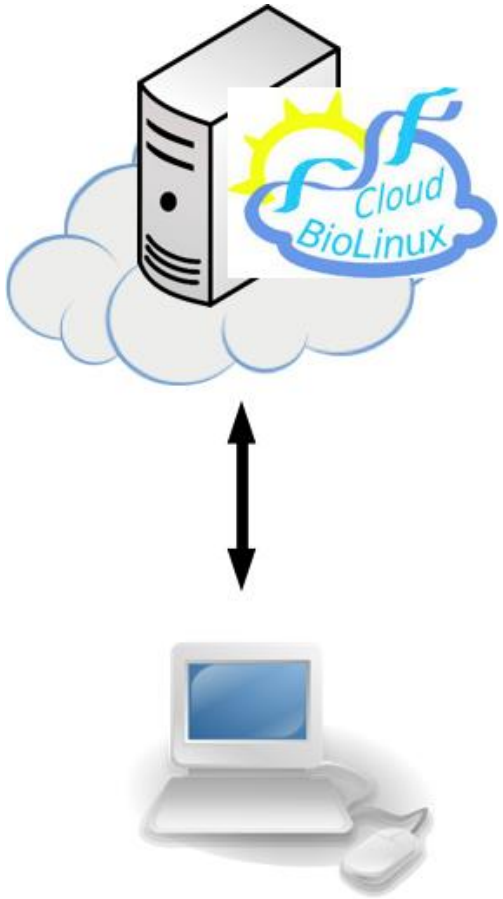

Galaxy
CloudMan
ec2-107-22-64-243.compute-1.amazonaws.com/c
.mobi (i)
Google Drive Help
CloudMan from Galaxy
Admin | Report bugs |
CloudMan Console
Welcome to CloudMan. This application allows you to manage this instance cloud cluster and the services within. Your previous data store has been reconnected. Once the cluster has initialized, use the controls manage services provided by the application.
Terminate cluster
Add nodes ▼
Remove nodes ▼
Access Galaxy
Status
Cluster name: JCVI Prok Pipelines 3.2
Disk status: 122G / 500G (25%)
Worker status: Idle: 0 Available: 5 Requested: 5
Service status: Applications Data
Autoscalin
Turn
Cluster status log
15:41:09 - Retrieved file 'persistent_data.yaml' from bucket 'cm-8188de9adf6373de804f46d7c45ae892' to 'pd.yaml'.
15:41:09 - Master starting
eviously existing cluster of type Galaxy
15:41:19 - SGE prerequisites OK; starting the service
15:41:26 - Configuring SGE...
15:41:49 - Successfully mounted file system /mnt/galaxyTools from /dev/xvdg1
kkrampis — screen — 115×22



The GSC Viral Cloud Resource project documentation and forum: a web 2.0 site brewed with open source softwareText

NIAID-JCVI Viral Genomic Pipelines Cloud Project

- Cloud BioLinux: public Virtual Machine (VM) on the Amazon EC2 cloud
- end-to-end viral genomics on the cloud, submit reads and get visualized annotation
- approach: pre-install all pipeline software on the VM, access through a web browser
- benefits underrepresented labs without computational infrastructure or expertise



Krampis K., Booth T., Chapman B., Tiwari B., Field D. and Nelson K.E. (2012) BMC Bioinformatics 13:42, "Cloud BioLinux: pre-configured and on-demand computing for the genomics community"

J. Craig Venter INSTITUTE