

I have all this data – now what?

Introduction to Bioinformatics Software and Computing Infrastructures

Timothy Stockwell (JCVI)

Vivien Dugan (NIAID/NIH)

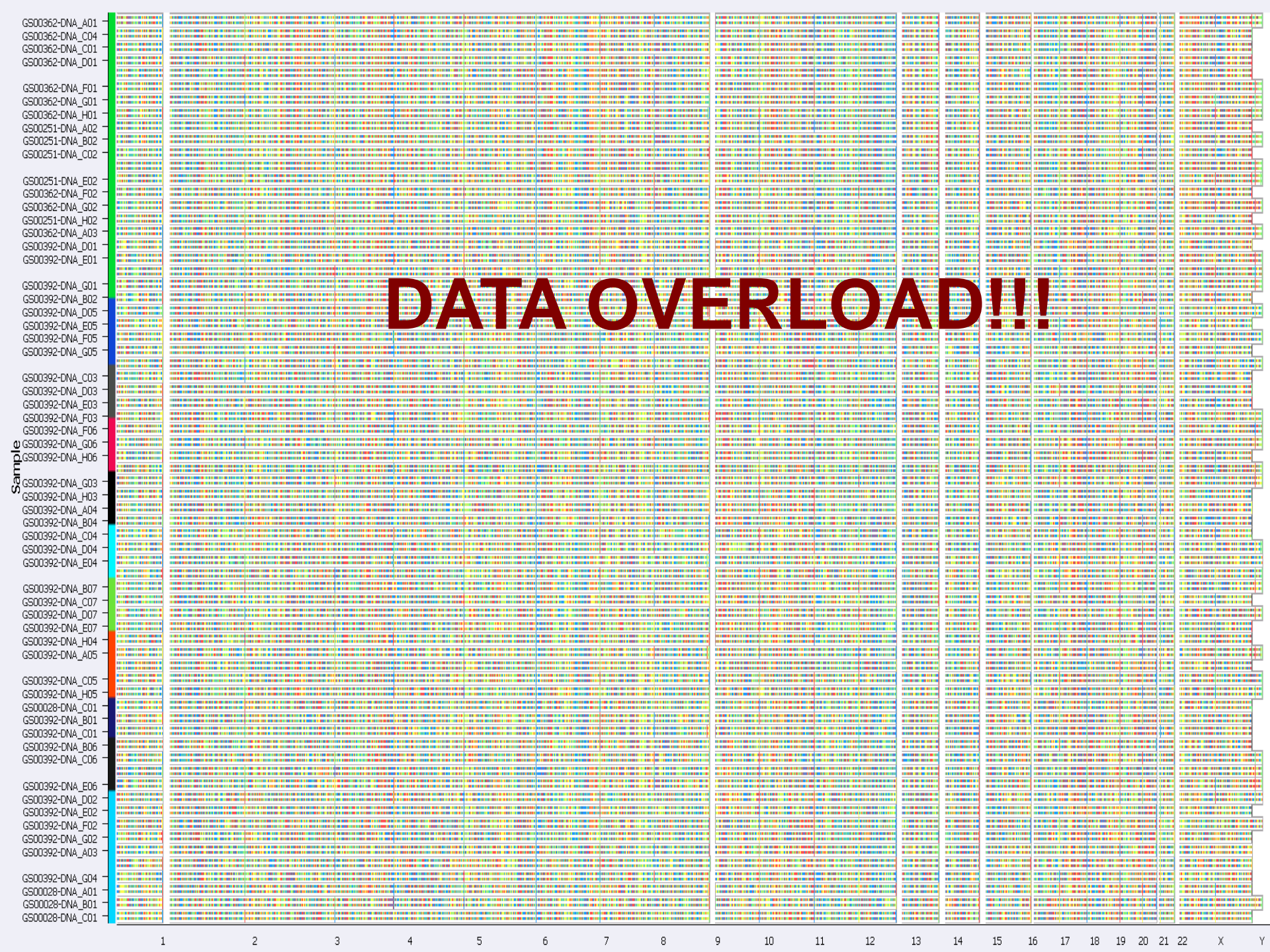


Bioinformatics

- **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.....







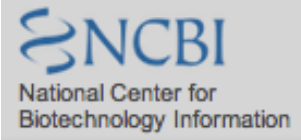


Genomics Resources

- 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>

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>

BLAST®

Bioinformatics Resource Centers (BRCs)

Welcome To VectorBase!

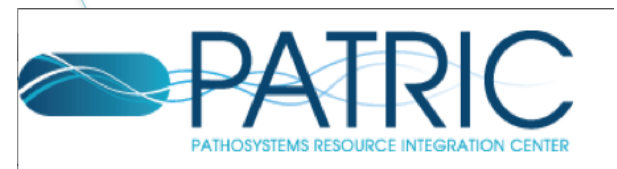
.....
An NIAID Bioinformatics Resource Center for
Invertebrate Vectors of Human Pathogens.



<https://vectorbase.org/>



<http://eupathdb.org/>



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



<http://www.pathogenportal.org/>



<http://www.viprbrc.org/>



<http://www.fludb.org/>

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



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	<i>JCVI Primer Designer</i> 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 <i>JLIMS</i> , sample tracking in JIRA	454, Illumina	454, Illumina
Assembly and Finishing	CLC bio de novo and mapping assemblers and consed for finishing	Newbler, <i>CABOG</i> , velvet, CLC bio de novo and mapping assemblers	<i>CABOG</i> , CLC bio, Newbler
Annotation Tools	<i>VIGOR</i>	<i>Prokaryotic Annotation Pipeline</i> , <i>MGAT</i> , <i>MANATEE</i>	<i>PASA</i> , <i>MANATEE</i> , <i>EVM</i> , <i>GSAC</i> , Trinity, RnNotator, Tuxedo Package,
Comparative Genomics	<i>ANDES Tools</i> , Phylogenetic Tools	<i>Prokaryotic Pan-Genome Pipeline</i> , <i>PanOCT</i> , <i>SYBIL</i>	GBrowse , OrthoMCL, Artemis, <i>SYBIL</i>

Italics indicate software developed at JCVI.

The Sequencing App Store

Chad Nussbaum
Broad Institute

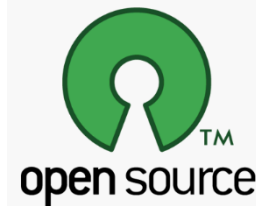
Just a sampling...



Examples: Genome Software Tools

- **ALL PATHS:** Assembly of Short Reads
- Using RNA-Seq for eukaryotic genome annotation
 - Inchworm RNA-Seq Assembler
 - Available at <http://inchworm.sf.net>
 - enhanced PASA: incorporate RNA-Seq 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>

RNA-Seq De novo Assembly



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>



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.

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 Sequencing Alignment/Map format files, <http://samtools.sourceforge.net/>

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

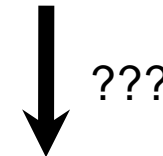
- Your institute's data center – typically, an IT department controls the hardware, the operating system, and the software installed
 - IT administrators install individual tools and programs – they might have to compile them from source code – very flexible, but often very expensive.
 - Need enough 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



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



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)

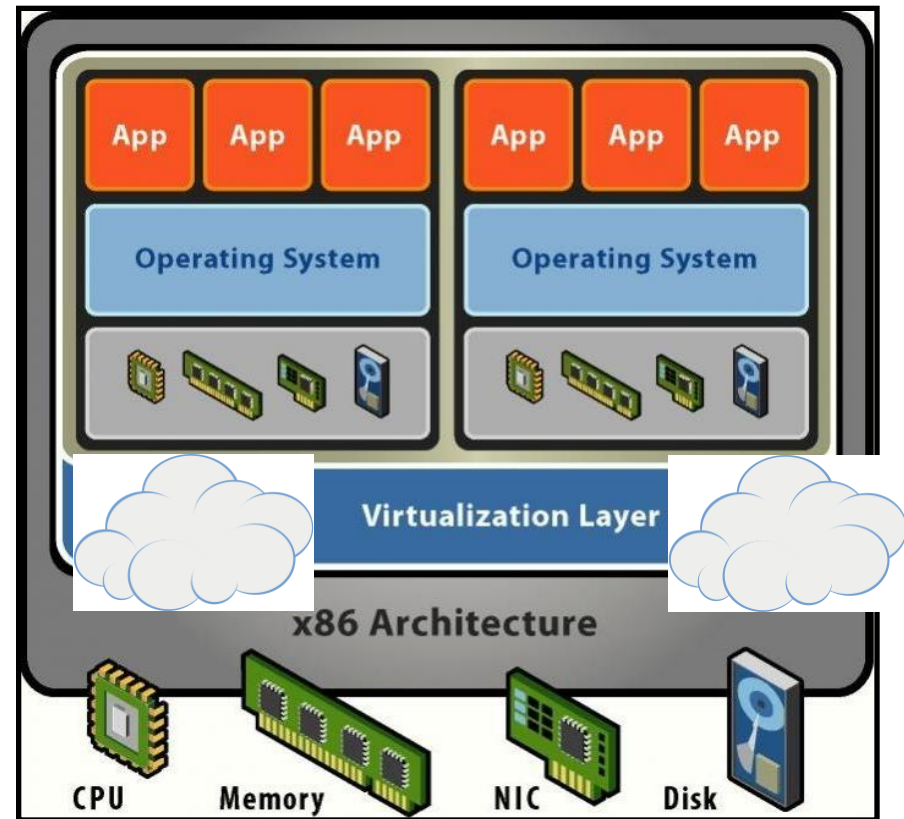


World-wide data centers

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

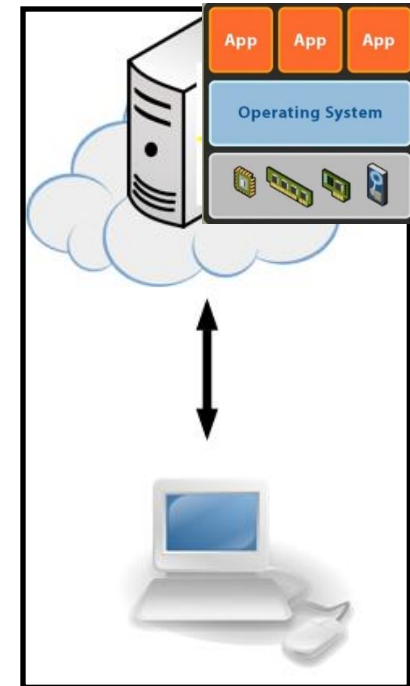
Problem B: bioinformatics tools need software engineering expertise

- 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



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



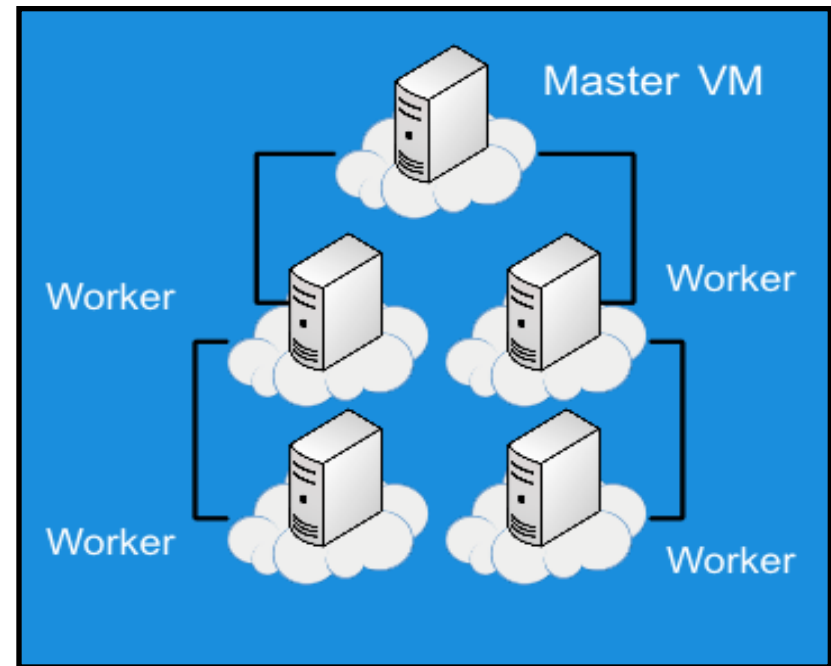
[Krampis K](#), Booth T, Chapman B, Tiwari B, Bicak M, Field D, Nelson K

Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community.

BMC Bioinformatics. 2012 Mar 19;13(1):42.

Compute Clusters on the Cloud

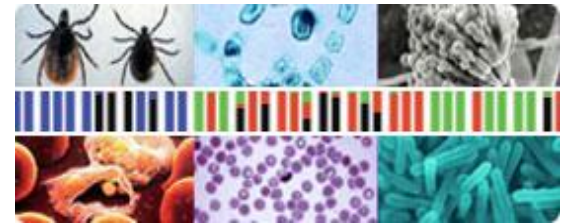
- Cloud BioLinux creates Sun Grid Engine (SGE) clusters on the Amazon Cloud
- Cloud BioLinux + Cloudman scripts boot SGE worker VMs
- Multiple genome runs in parallel: each worker runs one genome.



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

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™
I N S T I T U T E

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).*

From sequencer to the cloud

Processing

Free alignment and variant processing.



Storage

1 FREE TB

FREE

1 Terabyte Storage for Illumina data

+ 1 TB

PLUS ONE

+1 Terabyte Storage
\$250 USD/Month or
\$2,000 USD/Year

+ 10 TB

PLUS TEN

+10 Terabytes Storage
\$1,500 USD/Month or
\$12,000 USD/Year

Apps

Integrated applications.
App price set by vendor.



credit: basespace.illumina.com

Questions?

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@jcv.org

slideshare.com/agbiotec

Thank you !

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

The screenshot shows the AWS Management Console interface. At the top, there's a navigation bar with the Amazon Web Services logo, a search bar containing "AWS Product Information", and links for "Sign in to the AWS Management Console", "Create an AWS Account", and "English". Below the navigation bar is a horizontal menu with tabs for "AWS", "Products", "Developers", "Community", "Support", and "Account". The main content area is titled "Amazon Elastic Compute Cloud (Amazon EC2)". On the left, there's a sidebar with "Amazon EC2 Details" and a list of links: "EC2 Overview", "EC2 FAQs", "EC2 Pricing", "Amazon EC2 SLA", "EC2 Instance Types", "EC2 Instance Purchasing Options", "Reserved Instances", "Spot Instances", and "Windows Instances". The main text describes Amazon EC2 as a web service providing resizable compute capacity in the cloud. A "Sign Up Now" button is visible on the right side of the page.

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:

aws.amazon.com/console

Starting a VM: EC2 cloud console

The screenshot displays the AWS Management Console in a Mozilla Firefox browser. The browser's address bar shows the URL <https://console.aws.amazon.com/ec2>. The page title is "AWS Management Console - Mozilla Firefox". The browser's tab bar shows several open tabs, including "Nex...", "git ec2/...", "Aut...", "git ec2/...", "AW...", "NX ...", "Goo...", "how...", "No...", and "Tiny...".

The console interface features a navigation menu on the left with the following categories and items:

- Region: US East
- EC2 Dashboard
 - INSTANCES
 - Instances
 - Spot Requests
 - IMAGES
 - AMIs
 - Bundle Tasks
 - ELASTIC BLOCK STORE
 - Volumes
 - Snapshots
 - NETWORKING & SECURITY
 - Elastic IPs
 - Security Groups
 - Placement Groups
 - Load Balancers
 - Key Pairs

The main content area is titled "Amazon EC2 Console Dashboard" and contains the following sections:

- 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. Below it, a note states: "Note: Your instances will launch in the US East (Virginia) region."
- Service Health**: A table showing the current status of Amazon EC2 in the US East - N. Virginia region.
- My Resources**: A summary of resources in the US East (Virginia) region, including 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. A "Refresh" button is available.
- Related Links**: A list of links including Documentation, All EC2 Resources, Forums, Feedback, and Report an Issue.

Current Status	Details
Amazon EC2 (US East - N. Virginia)	[RESOLVED] Increased tagging error rates View complete service health details

Amazon EC2 VM launch wizard

Request Instances Wizard Cancel

CHOOSE AN AMI INSTANCE DETAILS CREATE KEY PAIR CONFIGURE FIREWALL REVIEW

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

Quick Start My AMIs **Community AMIs**

Viewing: All Images ami-6011e409 1 to 1 of 1 Items

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

cloudbiolinux.org

Request Instances Wizard Cancel

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

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:

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

Launch Instance

EC2 Instances let you break down costs into much smaller pieces.

Request Spot Instance

Launch Instance

Cloud BioLinux remote desktop connection

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

The image shows a screenshot of the AWS Management Console and a remote desktop connection window. The AWS console displays a list of EC2 instances and details for a selected instance. The remote desktop window shows the connection settings for the instance.

AWS Management Console - My Instances

Instance	AMI ID	Root Dev	Type	Status
i-49201823	ami-6011e409	ebs	m1.large	pending
i-f7340c9d	ami-6011e409	ebs	m1.large	running
i-795b6313	ami-6816e301	ebs	m1.large	terminated
i-f9330b93	ami-6011e409	ebs	m1.large	terminated

1 EC2 Instance selected

Root Device:	/dev/sda1	Root Device Type:	
Block Devices:	/dev/sda1=vol-68cce401:attached:2010-09-22T22:57:42.000Z		
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 Connection Window (NX - Cloud Biolinux)

Host: `ec2-184-73-27-151.compute-1.amazonaws.com` Port: `22`

Remember my password: Key:

Desktop: Unix, GNOME, Settings...

Display: 1024x768, W 800, H 600, Use custom settings: Settings...

Buttons: Delete, Save, Ok, Cancel

Cloud BioLinux remote desktop connection

The screenshot displays a remote desktop environment for Cloud BioLinux. The desktop background features the "bio-linux" logo in green. The system tray at the top shows the date and time as "Wed Sep 22, 5:27 AM".

Open windows include:

- Jemboss**: A menu-driven application with options like ALIGNMENT, DISPLAY, EDIT, ENZYME KINETICS, FEATURE TABLES, INFORMATION, NUCLEIC, PHYLOGENY, PROTEIN, and UTILS. The "Jemboss" logo is prominently displayed.
- RasMol - 3DMS X-RAY DIFFRACTION**: A window showing a 3D molecular model of a protein structure, rendered with red, white, and blue spheres.
- ClustalX 2.0.12**: A sequence alignment tool showing a multiple sequence alignment of DNA sequences. The alignment is displayed in a grid format with a progress bar at the bottom. The text "CLUSTAL-Alignment file created [ØØE]" is visible at the bottom of the window.

The taskbar at the bottom shows several open applications: [ClustalX 2.0.12], [RasMol], RasMol - 3DMS X..., ClustalX 2.0.12, and Jemboss. The system tray at the very bottom includes icons for "Community Digest, V...", "GBrowse syn Databa...", "NX - ubuntu@ec2-67...", and "GNU Image Manipula..."

ec2-54-234-103-31.compute-1.amazonaws.com

Galaxy

Analyze Data Workflow Shared Data Visualization Help User

Using 93.0 MB

Tools

Metagenomic analyses

FASTA manipulation

NGS: QC and manipulation

NGS: Assembly

NGS: Mapping

NGS: Indel Analysis

NGS: RNA Analysis

RNA-SEQ

- Tophat for Illumina Find splice junctions using RNA-seq data
- Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- Cuffdiff find significant changes in transcript expression, splicing, and promoter use

FILTERING

- Filter Combined Transcripts using tracking file

NGS: SAM Tools

NGS: GATK Tools

NGS: Peak Calling

SNP/WGA: Data: Filters

SNP/WGA: QC: LD: Plots

SNP/WGA: Statistical Models

Human Genome Mapping

VCF Tools

Viral Assembly and Annotation

- Viral Assembly Run Viral assembly
- VIGOR Run VIGOR

Viral Assembly (ve)

454 reads:

2: input_454.sff

Sanger reads:

4: input_Sanger.fasta

Solexa reads:

3: input_Solexa.fastq

Solexa trimpoints:

5: input_Solexa.fastq.trimpoints

Viral database:

barda

Execute

Data Libraries

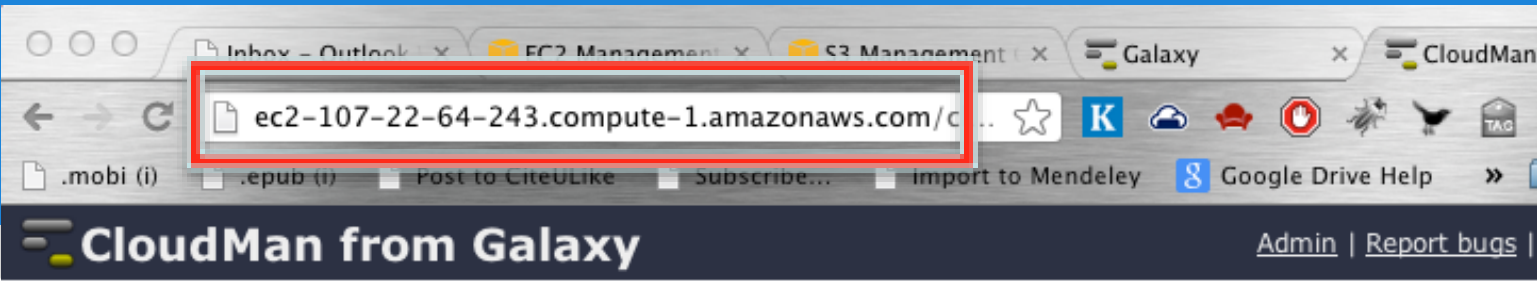
- Published Histories
- Published Workflows
- Published Visualizations
- Published Pages

History

Unnamed history 93.0 MB

- 5: input_Solexa.fastq.trimpoints
- 4: input_Sanger.fasta
- 3: input_Solexa.fastq
- 2: input_454.sff
- 208: Run Prok Pipeline on data 199
- 207: Run Prok Pipeline on data 200
- 206: Run Prok Pipeline on data 201
- 205: Run Prok Pipeline on data 202
- 204: Run Prok Pipeline on data 203
- 203: Prok Pipeline Setup on data 175 (bacillusanthracisstrktruger)
- 202: Prok Pipeline Setup on data 175 (bacillusanthraciscdc684)
- 201: Prok Pipeline Setup on data 175 (bacillusanthracisa0389)
- 200: Prok Pipeline Setup on data 175 (acidithobacillus)
- 199: Prok Pipeline Setup

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



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 



Autoscaling
Turn g

```
Cluster status log
15:41:09 - Retrieved file 'persistent_data.yaml' from bucket 'cm-8188de9adf6373de804f46d7c45ae892' to 'pd.yaml'.
15:41:09 - Master starting
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
```

gsc-cloud.herokuapp.com


J. Craig Venter[®] INSTITUTE NIAID VCR : Viral genomics Cloud Resource

administration agbiotec 1 logout about faq

questions tags users badges unanswered ask a question

search

questions tags users

All Questions  active newest hottest most voted

0 votes 0 answers 1 view **GSC project cloud team** cloud project 2 hours ago rsanka ♦♦ 1

0 votes 0 answers 2 views **Overview of of the GSC Cloud Computing Project** cloud project Feb 12 at 17:19 agbiotec ♦♦ 1

0 votes 1 answer 4 views **What is Fabric?** fabric python Jan 31 at 18:19 rsanka ♦♦ 1

0 votes 1 answer 3 views **How are new tools added to Galaxy?** galaxy viral tools installation Jan 31 at 17:52 rsanka ♦♦ 1

0 votes 1 answer 2 views **What is galaxy?** galaxy Jan 31 at 16:19 rsanka ♦♦ 1

0 votes 1 answer 7 views **How can I enable galaxy as a valid user on an Amazon instance?** amazon user galaxy permission instance Jan 31 at 16:16 rsanka ♦♦ 1

6 questions
4 answers
Most recently updated questions

Interesting tags
Ignored tags

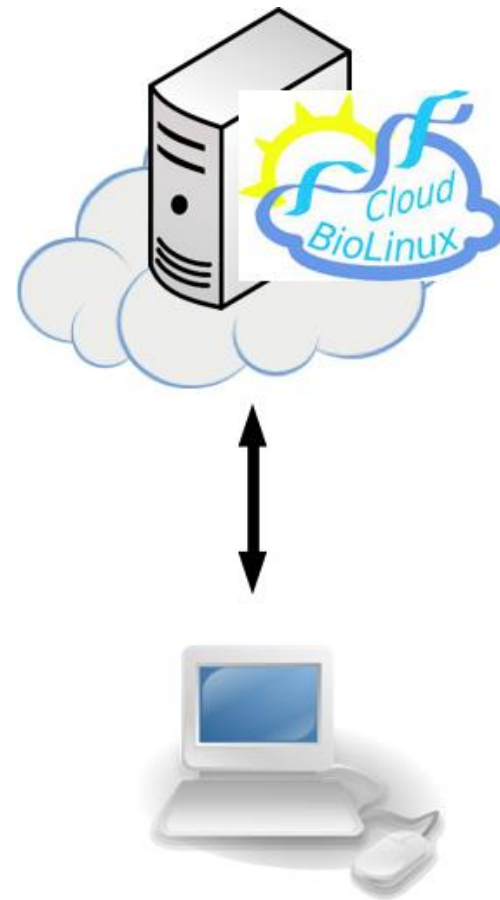
Learn more about the GSC Cloud Project.
The GSC Cloud Project is funded in whole or part with federal funds from the National Institute of Allergy and Infectious Diseases, \ National Institutes of Health, Department of Health and Human Services under contract numbers N01-AI30071 and/or HHSN272200900007C. This project is part of the To learn more about this documentation system visit the [OSQA website](#) and [OSQA wiki](#)

Recent tags
project cloud python fabric
installation tools viral instance
galaxy permission user

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"