Starting Bioinformatics from Zero as a Biologist

Presented by
Jessica Chen, Andrea (Ray) Etter, Peter Cook
Sponsored by IEH Laboratories & Consulting
Organized by the Developing Food Safety Professionals PDG
& Student PDG

Webinar Housekeeping

For best viewing of the presentation material, please click on ‘maximize’ in the upper right corner of the ‘Slide’ window, then ‘restore’ to return to normal view.

Audio is being transmitted over the computer so please have your speakers ‘on’ and volume turned up in order to hear. A telephone connection is not available.

Questions should be submitted to the presenters during the presentation via the **Questions section** at the right of the screen.

Webinar Housekeeping

It is important to note that all opinions and statements are those of the individual making the presentation and not necessarily the opinion or view of IAFP

This webinar is being recorded and will be available for access by IAFP members at **www.foodprotection.org** within one week.

Andrea (Ray) Etter completed her PhD at Purdue University under Haley Oliver, where she taught herself bioinformatics in order to investigate the role of heat stress in a recent salmonellosis outbreak. She will start as an assistant professor in the Department of Nutrition and Food Sciences at the University of Vermont in January 2019, where she plans to continue to use bioinformatics to understand phenotypic characteristics of outbreak-associated foodborne pathogens.

Jessica Chen is a microbiologist (IHRC. Inc.) supporting the National Antimicrobial Resistance Monitoring System at the Centers for Disease Control and Prevention. Her research focuses on understanding the molecular epidemiology and evolution of drug-resistant foodborne pathogens. Prior to joining NARMS, Jessica was a postdoc at the University of British Columbia where she conducted comparative genomics research involving *Listeria monocytogenes* and Shiga-toxigenic *E. coli*. Jessica received her PhD in Animal Science with an emphasis on microbial food safety from Texas Tech University in 2013. Jessica is the chair of the Developing Food Safety Professionals PDG and secretary for the Georgia Association for Food Protection.

Peter Cook is currently a post-doctoral researcher in the Center for Food Safety at the University of Georgia. He graduated in 2017 with a PhD in Animal Science and a focus in food safety from Texas Tech University, and will begin a Bioinformatics Fellowship with the Center for Disease Control and Prevention, Atlanta, GA in 2018. His dissertation work involved the comparison of virulence-attenuated and fully virulent *Listeria monocytogenes* using whole genome sequencing and transcriptomics, and his current work involves microbiome analysis and fungal sequencing.

Special thanks to Lee Katz, Bioinformatician in the Enteric Diseases Laboratory Branch at the Centers for Disease control and Prevention for his assistance in the development of this webinar.

Webinar Overview:

- Figuring out what your institution has in terms of resource
- Finding resources outside your institution
- Discussion of tools requiring little/no programming experience
- What should be in your very basic toolkit
- What to do when you’re stuck
- Discussion of file inputs and program outputs
- Open Source vs Closed Source
- CLI vs GUI
- Career paths in bioinformatics

Resources for Learning Bioinformatics

Andrea Etter

Step one (a): What am I trying to do?

Genomics?

- Assemble and compare DNA sequences
  - Gene presence/absence
  - SNPs in shared genes

Transcriptomics?

- Assemble RNA sequences and compare gene expression

Metagenomics?

- Compare communities of bacteria in different samples

Step one (b): What does my institution have?

Why start with your institution?

- Cheap
- Easiest to attend
- Best options for follow-up help
  - Code
  - Models
  - Stats
  - Troubleshooting?

Step one: What does my institution have?

Possible options:

- Seminars
- Classes (RNA-seq analysis, Intro to R, intro computer science classes, etc)
  - Semester long or short courses
- Workshops
- Computational resources
  - High performance computers for large analysis
  - Can offer good introductory courses and example code
- Help sessions/coffee hours
- Helpful faculty, staff, postdocs, etc

How do I find these resources?

- Departmental/program emails
- Institution website keyword search (“bioinformatics”, “RNA-seq”, “genomics”, “R”)
- Class listings search
- Ask sequencing core personnel about options
- Ask students/faculty/staff who’ve done similar projects

What Institutional options are best?

Personally---

- Seminars for overview
- 1-2 day workshops for code & hands-on practice
- Coffee hours or one-on-one when you’re stuck
- Classes to learn programming languages

Step two: resources outside your institution

MOOCs-

- Coursera: University of Michigan's *Python for Everyone* course
  - Many, many other options--overviews and in depth courses

Workshops - IAFP, ASM, meetings.cshl.edu, bioinformatics.ca, evomics.org

Other options:

- Code academy
- Rosalind
- University of Washington short courses each summer

What should I choose?

Institution: may be free, easy access, etc.

- Time constrained

Moocs: Free option or pay $50/month for graded assignments and certificates.

- MOOCs vary in content, style, and quality

Rosalind: Free, computer game style.

Code Academy: Course-style. $199 for intensive python course, but other languages available

Bottom Line: experiment and find what works for you!

I’m learning to write some code but I’m not a wizard just yet.

What can I do in the meantime?

Jessica Chen

Limited Programming Experience? You have options!

NCBI Genome Workbench - Windows/Mac/Linux FREE

Align, view, edit sequences. Build and view phylogenetic trees.

Center for Genomic Epidemiology - Web-based FREE

WGS assemblies, subtyping, phylogenetic trees

Galaxy - Web-based FREE or Installed on a Mac or Linux computer.

Thousands of tools for RNA/DNA Seq, Metagenomics etc.

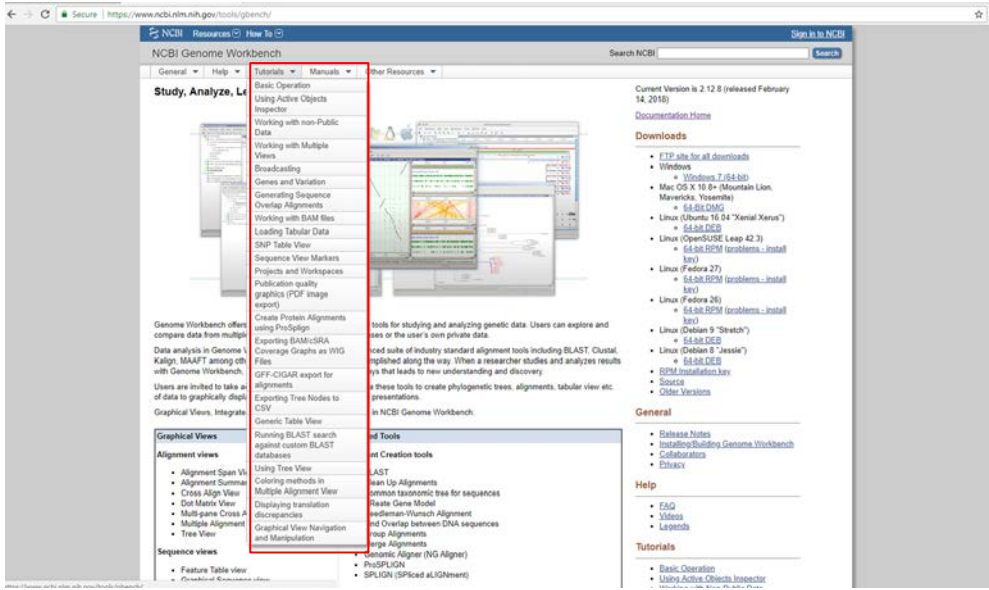
CLC Genomics Workbench, Bionumerics, Geneious - $$$

NCBI Genome Workbench

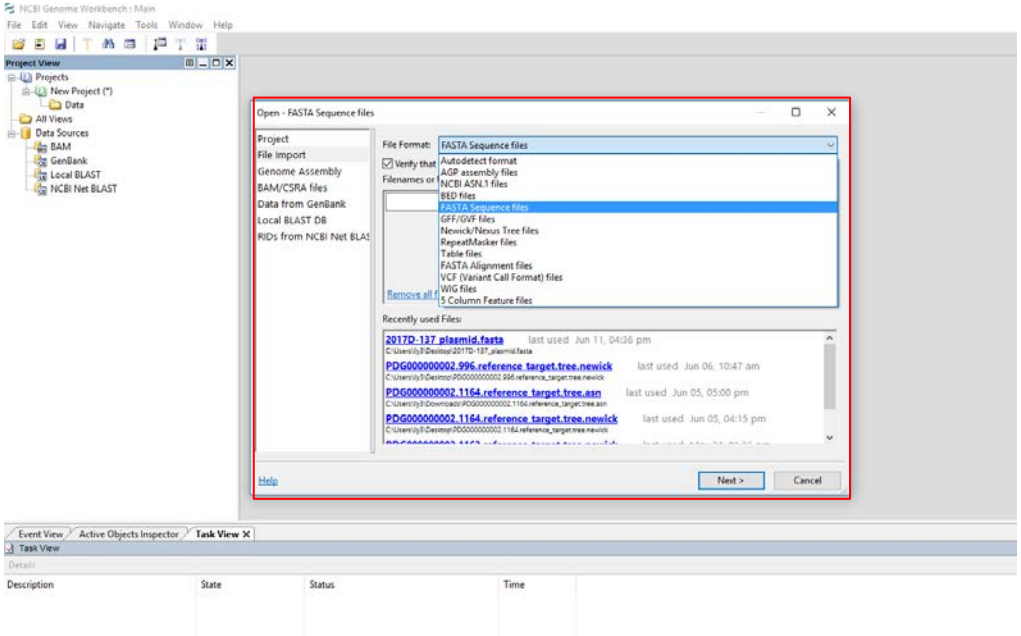
https://www.ncbi.nlm.nih.gov/tools/gbench/

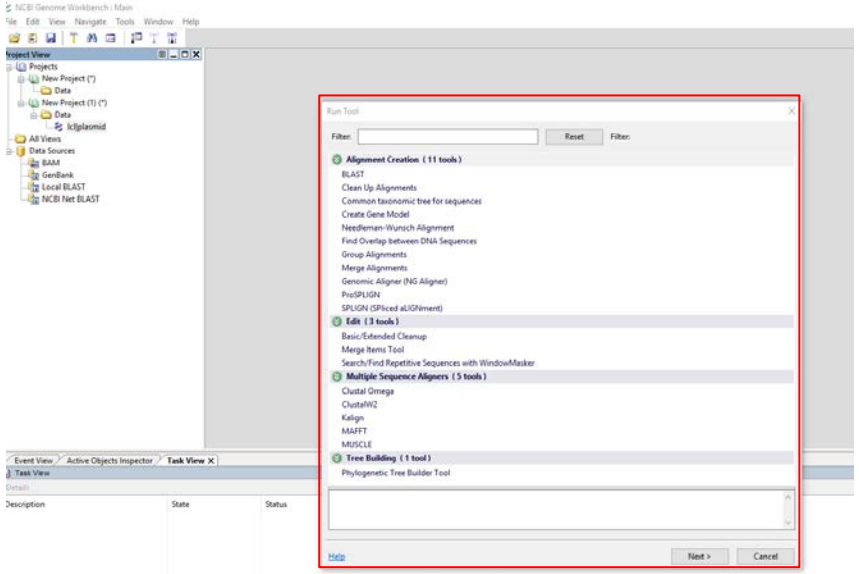
- Common sequence analysis tools
  - Alignment
  - Trees
  - BLAST

See tutorials for detailed instructions.



Import files and run tools with easy to use menus



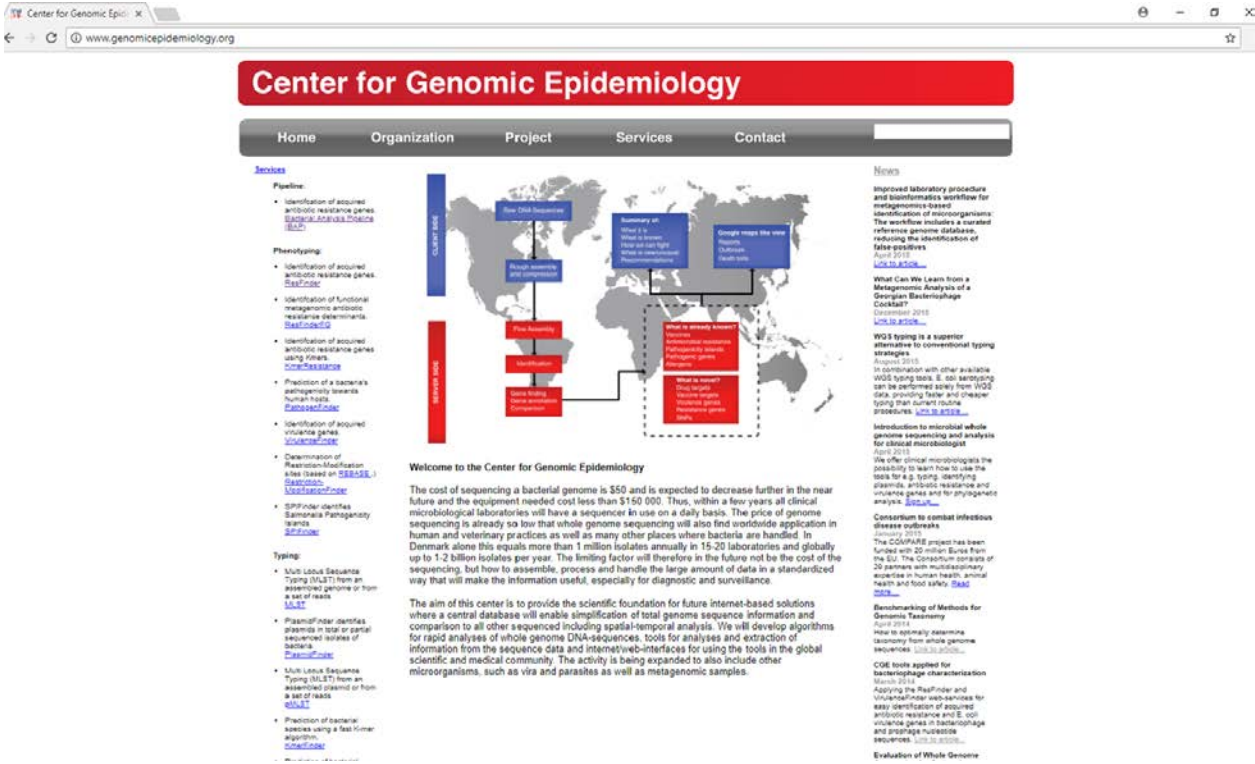


Center for Genomic Epidemiology Tools

www.genomicepidemiology.org

Commonly used tools for bacterial genome characterization

- Assembly
- MLST
- Resistance/Virulence/ Plasmid gene detection
- *In-silico* serotyping
- Whole genome phylogenies



Using CGE Tools

Center for Genomic Epidemiology

Welcome **lly3**

Home | Services | Instructions | Output | Overview of genes | Article abstract

ResFinder 3.0

ResFinder identifies acquired antimicrobial resistance genes and/or find chromosomal mutations in total or partial sequenced isolates of bacteria.

The database is curated by:
Valeria Bortolaia
(click to contact)

View the version history of this server.

Chromosomal point mutations

Resistance caused by mutations
Select species
Salmonella

Show unknown mutations
Show only known mutations
Show all mutations, known and unknown

Acquired antimicrobial resistance genes

Select Antimicrobial configuration
Select multiple items, with Ctrl-Click (or Cmd-Click on Mac) - by default all databases are selected
Aminoglycoside
Beta-lactam
Colistin
Fluoroquinolone
Fosfomycin
Fusidic Acid

Select threshold for %ID
90 %

Select type of your reads
Assembled Genome/Contigs*
Assembled Genome/Contigs*
454 - single end reads
454 - paired end reads
Illumina - single end reads
Illumina - paired end reads
Ion Torrent
SOLiD - single end reads
SOLiD - paired end reads
SOLiD - mate pair reads

of the web adress is https and not just http. Fix it by clicking here.

Size	Progress	Status

Upload | Remove

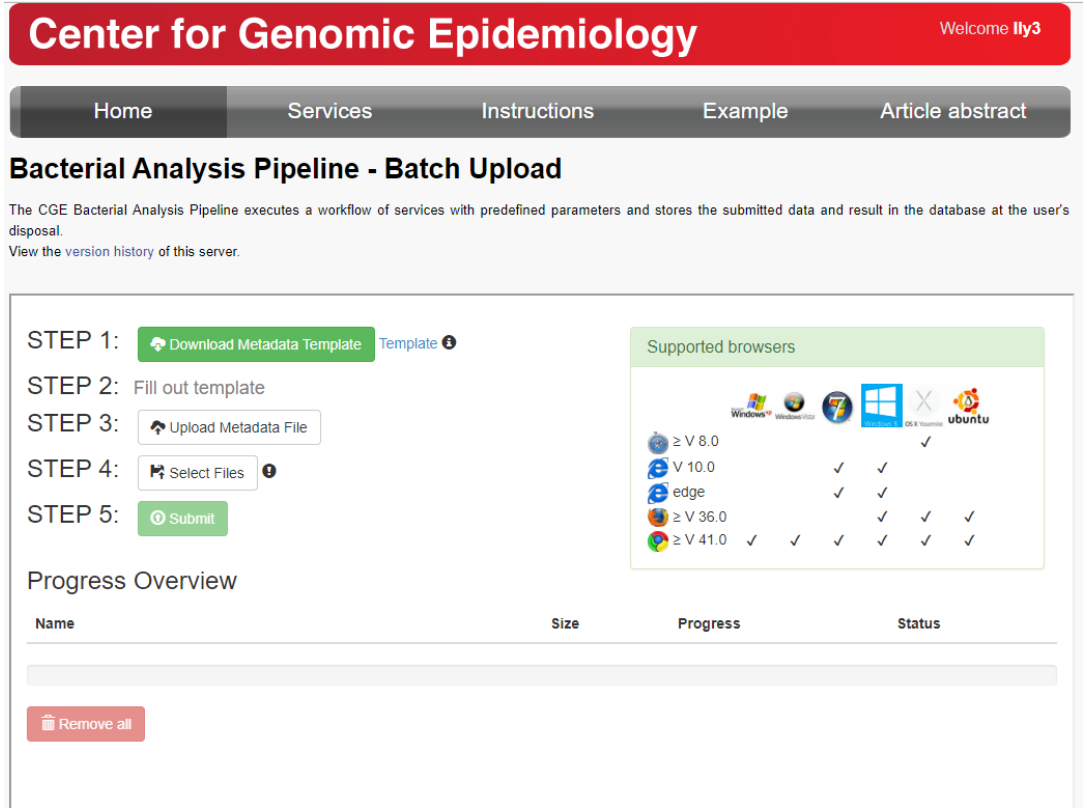
Confidentiality:
The sequences are kept confidential and will be deleted after 48 hours.

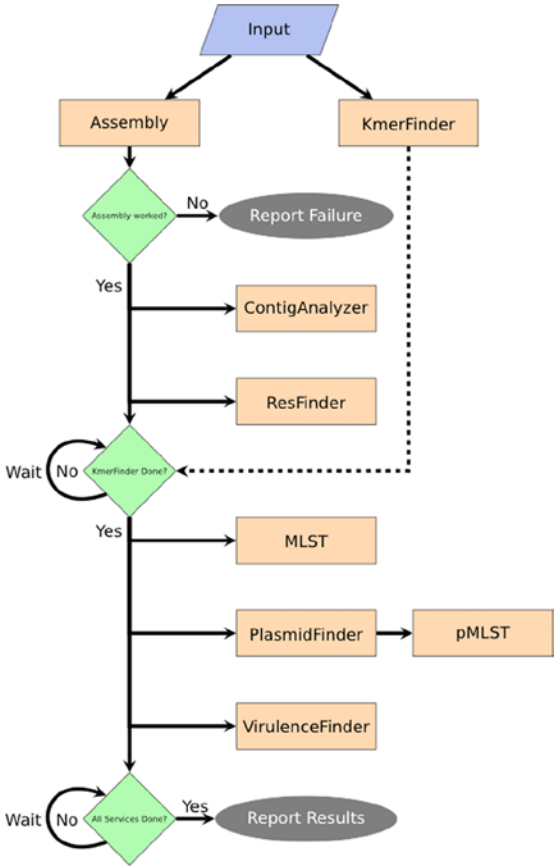
Database Updates (Acquired antimicrobial resistance)

The ResFinder database download site

- 21-Mar-2018 Major updates in tetracycline db
- 21-Mar-2018 mcr-4.2_MG026621, mcr-5.2_ MG384740 and mcr-7.1_ MG267386 added to colistin db
- 19-Feb-2018 Major updates and corrections in quinolone and colistin database
- 30-Nov-2017 General updates and corrections
- 14-Sep-2017 Fosfomycin database was updated with 20 genes
- 25-Aug-2017 Colistin database was updated with the genes mcr-4 and mcr-5
- 25-Aug-2017 Beta-lactam database was updated with the gene blaOXA-427
- 25-Aug-2017 Beta-lactam database entries for blaSHV-5, blaSHV-12 and blaSHV-129 were corrected
- 25-Aug-2017 Sulphonamide database entries for sul3 were corrected
- 04-Jul-2017 Colistin database was updated with the gene mcr-3

Register for batch uploading and use of the Bacterial Analysis Pipeline





Thomsen et al. 2016

All CGE databases and scripts are available on BitBucket

genomicepidemiology

Atlassian, Inc. [US] | https://bitbucket.org/genomicepidemiology/

Genomic Epidemiology

- Overview
- Projects
- Snippets
- Members

Genomic Epidemiology

Repositories

Language

Find repositories

Repository	Project	Last updated	Builds
apache-webserver	Web	36 minutes ago	
kma	CGE	6 hours ago	
resfinder_db	Databases	10 hours ago	
mlst_db	Databases	2 days ago	
cge_core_module	CGE	5 days ago	
MGmapper	CGE	2018-06-05	
MLST	CGE	2018-06-04	
ENAUploader	CGE	2018-06-01	
MetaPhylogeny_paper	CGE	2018-05-31	
cgMLSTFinder	CGE	2018-05-30	
Evergreen	CGE	2018-05-28	
PointFinder	CGE	2018-05-23	
pointfinder_db	Databases	2018-05-23	

Recent activity

genomicepidemiology/mlst-2.0
Repository deleted
Camilla Hundahl Johnsen · 2018-05-24

genomicepidemiology/mlst-2.0_db
Repository deleted
Camilla Hundahl Johnsen · 2018-05-24

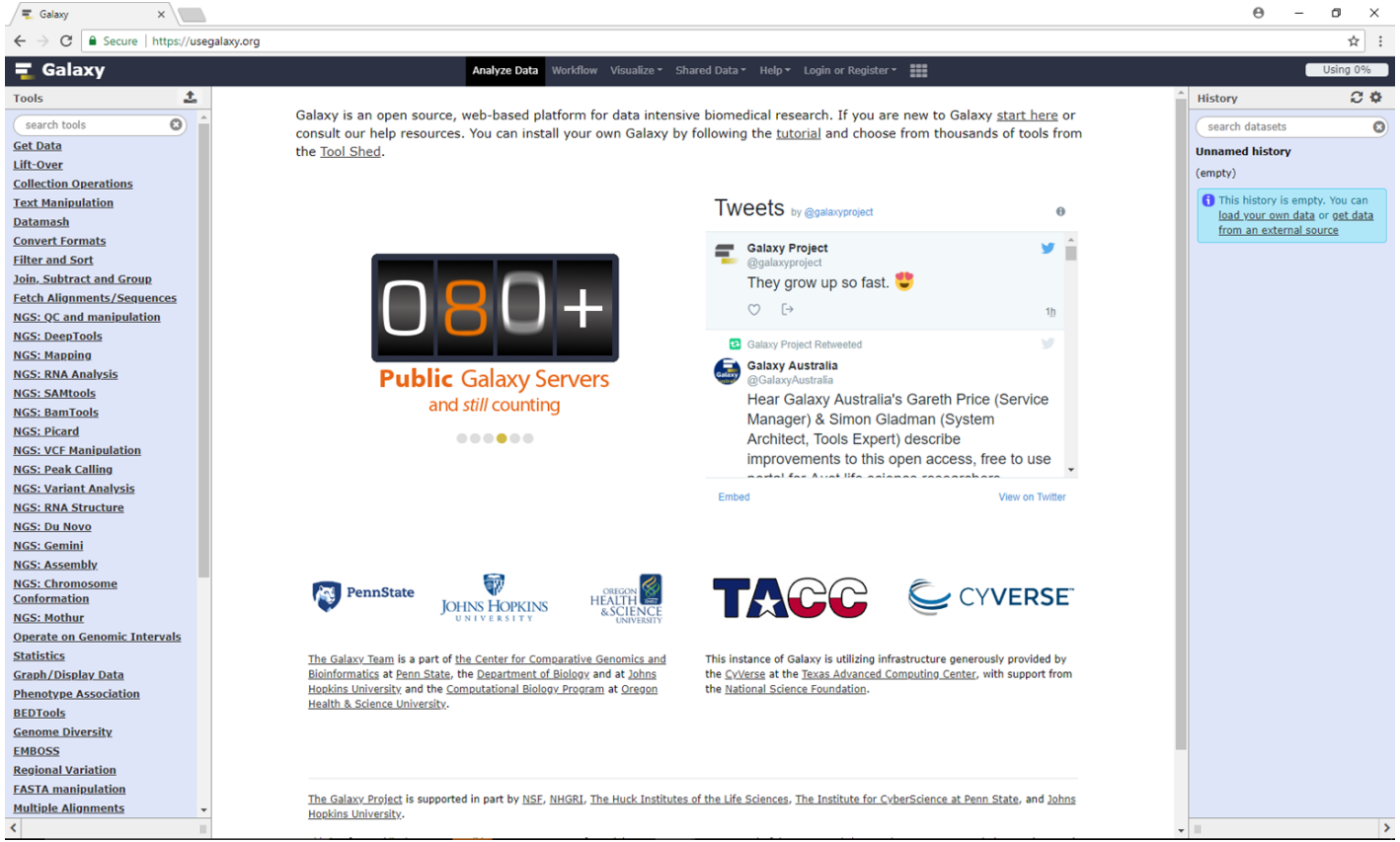
genomicepidemiology/mlst-2.0_db
Repository deleted
Camilla Hundahl Johnsen · 2018-05-01

genomicepidemiology/PointFinder
Repository deleted
Rosa Lundbye Allesøe · 2017-12-07

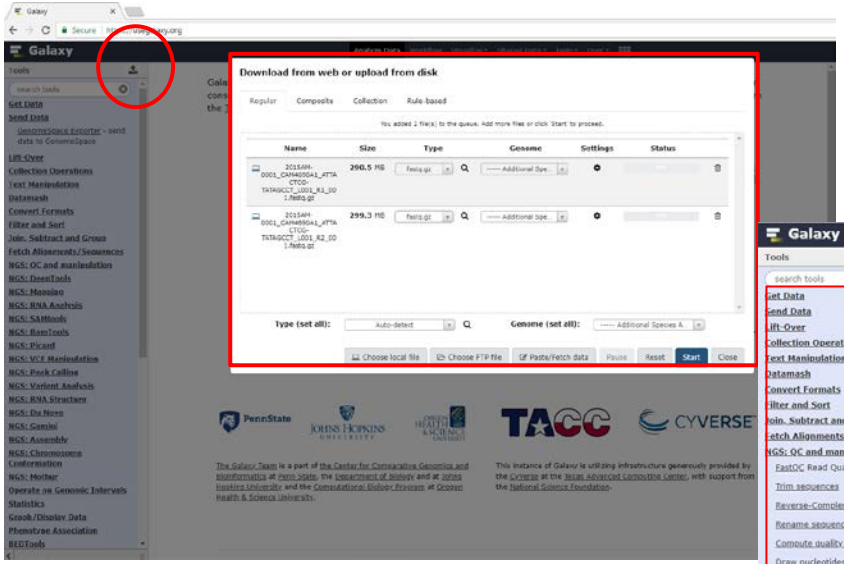
genomicepidemiology/KMA
Repository deleted
ptlcc · 2017-10-19

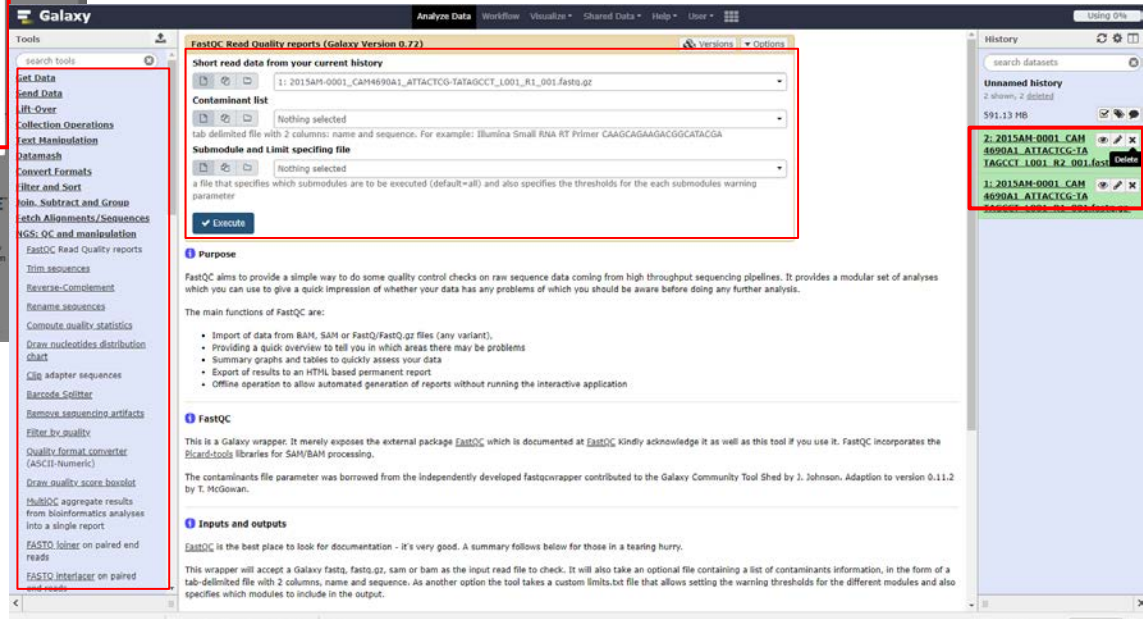
Galaxy

www.usegalaxy.org



Option 1: Use the free public server
Just upload your own data & use the tools available

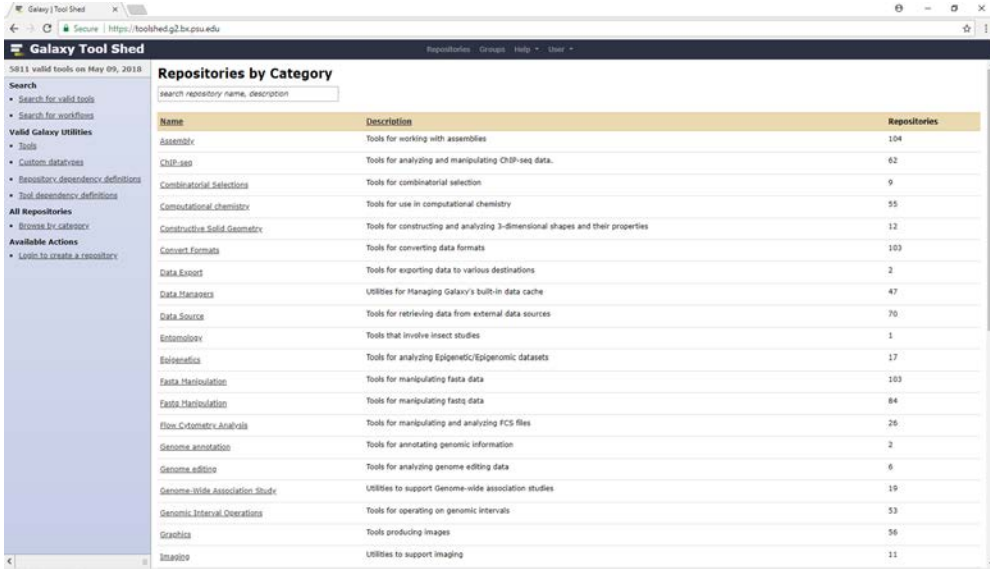




Can chain together multiple programs
to produce workflows

Option 2 (more complicated): Install galaxy and run your analysis locally.

- Available for Linux/Mac Users
- https://galaxyproject.org/admin/get-galaxy/
- Have access to additional tools in the Galaxy Tool Shed
  - Contains thousands of NGS analysis tools!



Option 3: Institutional Galaxy Servers

Your institution may have its own instance of Galaxy already set up.

FDA has a Galaxy instance with its own tools:

- https://www.galaxytrakr.org/root/login?redirect=%2F

What should be in your basic toolkit and how to troubleshoot problems.

What should be in your toolkit and why?

Initially: Command line experience in *nix(Unix/Linux).

- You can download/install/run programs from others.
- Can run batch analysis, or chain several commands together to run sequentially.
- Many tutorials available - here's a good one. http://www.ee.surrey.ac.uk/Teaching/Unix/

Later: Some sort of programming experience in a language you feel comfortable learning (Python, Perl, R, etc.)

- Extends what can be done directly on the command line.
- Can write a script to run a specified analysis workflow.
  - Useful if you are running the same types of analysis repeatedly.

What *is* Linux Exactly?

Linux is a free and open-source computer operating system.

OK, but why should I use it?

- Many bioinformatics software run exclusively on linux.
- Multi-tasking - can specify a process to run on each file in a way that's not always possible on a PC.
- Customization - able to modify and customize processes in a way that may not be possible when using GUI-based software.

HALP! I’m Stuck.

Googling the error can often yield helpful information

- Chances are you're not the first person to encounter this problem.

Bioinformatics web forums

- https://www.biostars.org/
- http://seqanswers.com
- https://stackoverflow.com/ (general programming)
- https://www.researchgate.net/ (questions)

Senior graduate students, postdocs, or collaborators can be a great resource

Introduction to the Bioinformatic Environment

Peter Cook

Computer Architecture

- OS
  - Operating System
    - Requirements depend on the software you are going to run
- CPU
  - Central Processing Unit
    - Requires at least 1
- RAM
  - Random Access Memory
    - Requires a minimum of 4 Gigabytes
- HDD
  - Hard Disc Drive
    - 256 Gb is a good place to start

Software Selection

- Open Source (Free)
  - Requires compiling
  - Dependencies (which must be installed as well)
  - Can require depreciated versions of other software
  - No personal IT help, but large online community
  - Often CLI only
- Commercial Software
  - Can only perform the analyses provided
  - Can be restricted in algorithm selection
  - Exporting data in a common format

Computer Languages

- Computer languages
  - building software
    - C, C++, C#, Python, Bash, R
- Command line statements
  - Define the program, Select input files, Set output files
- Scripts
  - Conserved or stored pieces of code
  - Generalize command line statements to be more useful
- Pipelines
  - Coagulated scripts and software to quick transfer of data from one software to another.

GUI vs CLI

- GUI (commonly pronounced as “Gooey”)
  - Graphical User Interface
    - Point and click
- CLI (also called a “terminal”)
  - Command Line Interface
    - Using the keyboard

How to access the Unix/Linux CLI

On your PC:

- Mac:
  - Open Terminal
- Windows:
  - Windows 10 and Up - Linux can be run by installing the Windows Subsystem for Linux
    - https://docs.microsoft.com/en-us/windows/wsl/install-win10
  - Older versions of Windows - through a virtual box

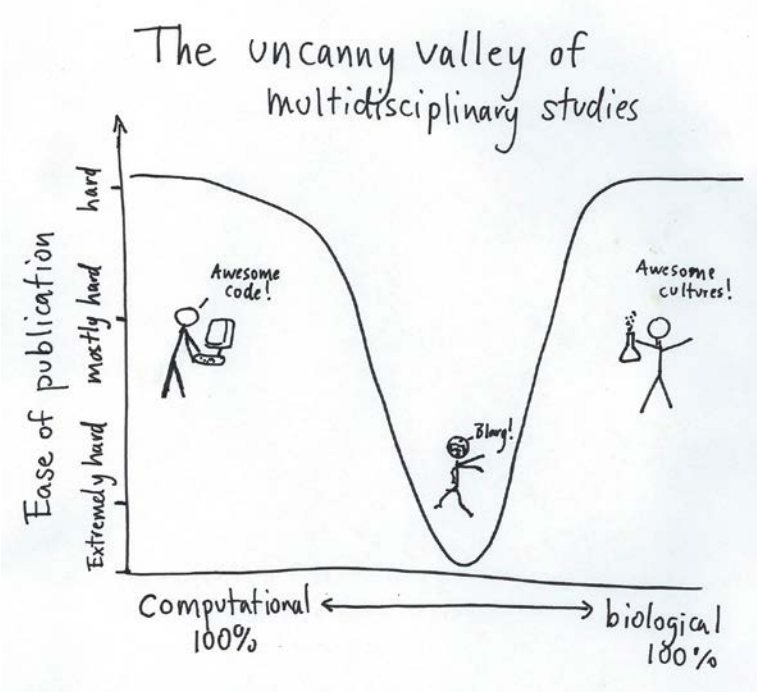
Remote access an external computer:

- Secure Shell interface (ssh terminal)

Bioinformaticians

- Compare and contrast volumes of data that cannot be compared by hand
  - Use prebuilt programs to analyze data
    - CLI or GUI
  - Write programs to fill the need

- Interpret the results, and understand the effect “assumptions” can have



http://jasonya.com/wp/the-uncanny-valley-of-multidisciplinary-studies/

Algorithms

- Defined set of instructions
- History (light)
  - Used for thousands of years (before computers)
  - Computers are just more efficient at implementation
- “long way” vs “short way”
  - Brute Force (check every possibility)
  - Algorithms (check only the possibilities that are necessary)

Algorithmic Example – Searching Comparisons

Number of Values in the list	Unsorted searching	Sorted with brute force searching	Sorted with Binary Searching
15	15	~8	4
1,000	1,000	~500	10
1,000,000	1,000,000	~500,000	20

Algorithmic Example – Searching Comparisons

Number of Values in the list	Unsorted searching	Sorted with brute force searching	Sorted with Binary Searching
15	15	~8	4
1,000	1,000	~500	10
1,000,000	1,000,000	~500,000	20

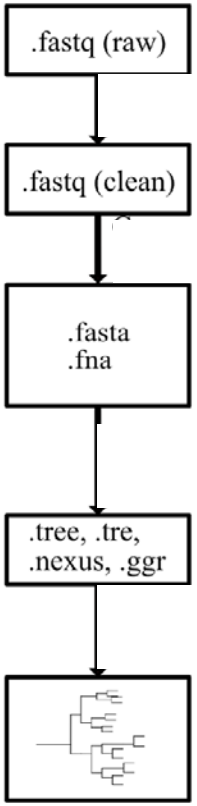
Files

- File name
- Store of information
  - Text or bits(binary)
- File ending
  - .txt, .fasta, .gbk, .zip, .gz
- Compressed
  - .zip, .gz
- Multiple files
  - .phi, .phs, .psa

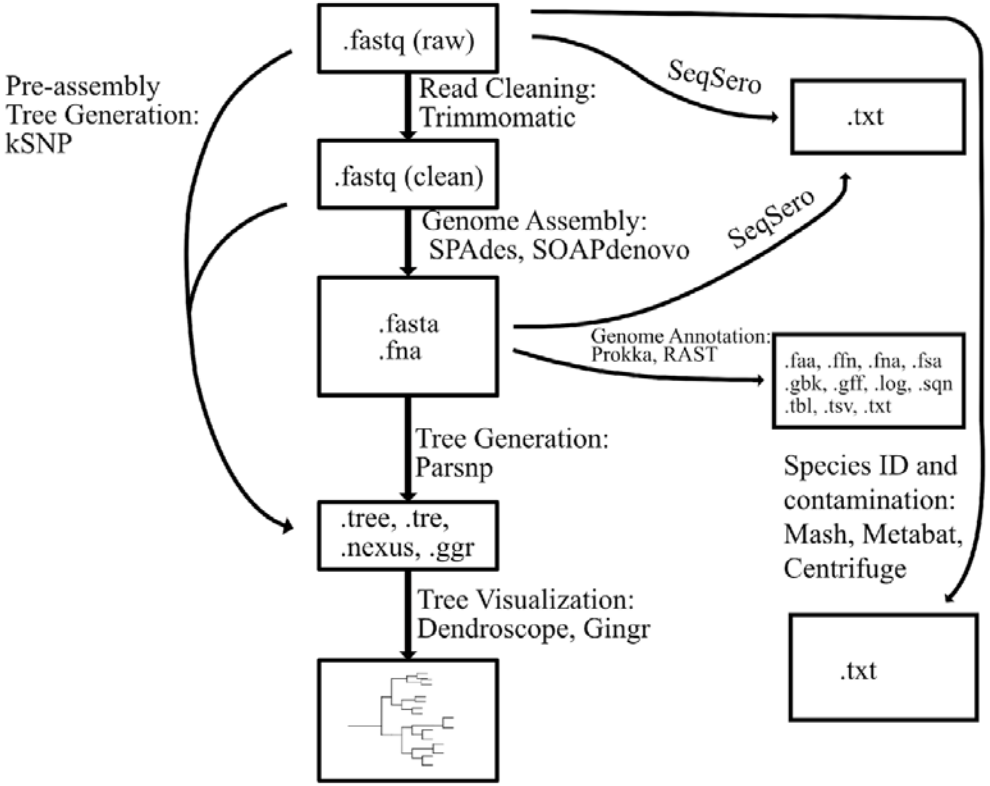
Bioinformatic File Formats

- Standardized Formats vs non-standardized
- File Formats and the “extension”
  - File extensions represent the data
    - Modifying the ”extensions”
- Knowing the structure based on the data type
  - FAST (.fasta, .fna, .faa, fastq)
  - Compressed files (.gz, .bgz2)
  - Phylogenetic trees (.ggr, .tre)

Pipeline Development and Analysis



Pipeline Development and Analysis



Careers in bioinformatics

Slides by Lee Katz, presented by Peter Cook and Jess Chen

"A day in the life"

Quora

Search for questions, people, and topics

Bioinformatics

What is a typical day for a bioinformatician?

https://www.quora.com

I am a bioinformatician pursuing my doctorate second year. This is how I can describe bioinformatician's life :

- Wake up and prepare for lab
- While in lab get your caffeine boost and kickstart the day
- Look at the computer for last nights running programs. So, it crashed. Run the program again
- Check for the space on the cluster and run some more codes.
- Read the literature for the project and develop some ideas. Now, think of how to write them into a code else look for some online programs or tools to get to your solution.
- Post your questions on websites like stack overflow. They are the best for improving your programming skills.
- Perform huge data analysis so keep tricks ready to automate the process.
- Explain to molecular people that all this is actual work who seem to think it's just some computer "thing"
- By the end of the day close eyes and see the pixelated screen in your eyes.

So it's a lot of computer love along with biological sense.

I forgot that we also have lunch in between just like everyone else :)

A typical career in bioinformatics and public health

- Outbreak analysis
- Applied research
  - Comparative genomics
  - Functional analysis
- Pipeline (workflow) development
- Working with others
  - Projects between teammates
  - Projects with collaborators
  - Training others
  - Conferences

- Hopefully not too much of these although some is necessary
  - Systems administration
  - Purchasing
  - IT advice
  - "Hey--you're good at computers. Can you fix this?"

Oak Ridge

https://orise.orau.gov/cdc/

ORISE fellowship

Within the last five years of your degree

https://www.aphl.org/fellowships/Pages/Bioinforma

Bioinformatics Fellowship

Sign In OR Create an Account

APHL | ASSOCIATION OF PUBLIC HEALTH LABORATORIES

Search... Search

Search for Training and Resources

Our Value — About APHL & Our Membership

Our Work — Programs, Publications & Services

Your Resources — Member, Funding, Emergency & Contact Information

Your Development — Training, Conferences & Careers

I Want To

From APHL Follow

APHL | FELLOWSHIPS | APHL-CDC BIOINFORMATICS FELLOWSHIP

APHL-CDC Bioinformatics Fellowship

- APHL-CDC Antimicrobial Resistance Fellowship
- APHL-CDC Bioinformatics Fellowship
- APHL-CDC Environmental Public Health Lab Fellowship
- APHL-CDC Infectious Diseases Laboratory Fellowship
- Ronald H. Laessig Memorial Newborn Screening Fellowship

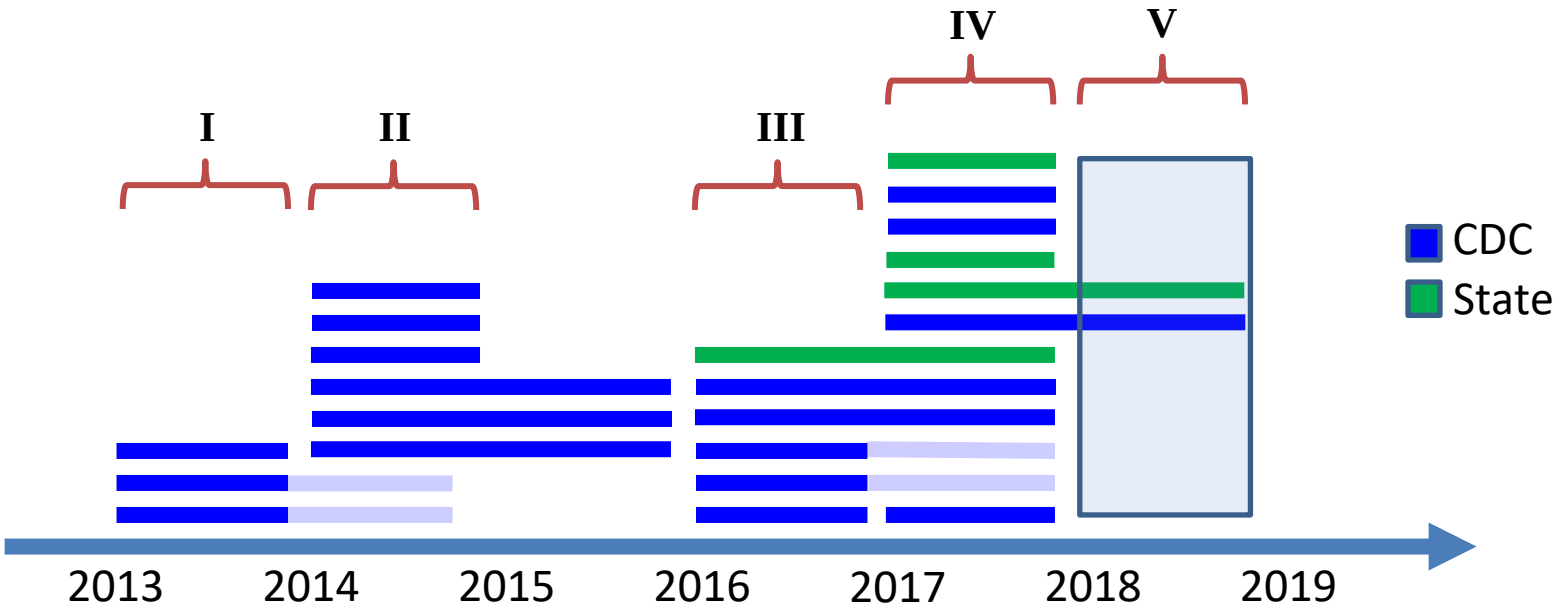
The APHL-CDC Bioinformatics Fellowship aims to train and prepare bioinformaticians to apply their expertise within public health and design tools to aid existing public health personnel in the use of bioinformatics. The Bioinformatics Fellowships provide post-master's and post-doctoral level professionals the opportunity to apply their skills to a range of important and emerging public health problems, while gaining experience in their fields. Whether your specialty is metagenomics, algorithm/software development, microbial genomics, or another research area, we have a place for you!

Program specifics

The program is a one-year full-time working fellowship for master's- and doctoral- level bioinformaticians. Postdoctoral fellows may extend for an additional year, provided funding is available. Fellows are placed in state, local and federal (CDC) public health laboratories throughout the US and collaborate on a wide range of important and emerging public health problems. All fellows participate in an orientation session. Once in their host laboratories, fellows are supervised by an experienced mentor and work on real-world infectious disease projects. Fellows will collaborate with public health laboratorians, epidemiologists, and other subject matter experts to synthesize and correlate data into actionable public health information as part of ongoing AMD projects. In addition to their project specific work, fellows will participate in distance-based training and learning activities to achieve proficiency in select public health laboratory core competencies.

https://www.aphl.org/fellowships/pages/bioinformatics.aspx

CDC-APHL Bioinformatics Fellows



- Currently finalizing recruitment/selection for CLASS V (Up to 10 new Fellows).
- For Classes I-III, 11 out of 15 have remained in public health. (Includes PhDs).

State and local health departments

- ❑ AMD program supporting the rollout of a growing number of pilot projects
  - State, county and local public health laboratories.
- ❑ All 50 state public health laboratories currently have basic NGS capabilities,
  - Funding, reagent and/or infrastructure support from CDC, FDA GenomeTrakr, and other sources.

Local careers websites

- Public health positions
  - PHEC, Public Health Employment Connection apps.sph.emory.edu/PHEC
  - APHL, careers.aphl.org
- Bioinformatics positions
  - ISCB, www.iscb.org/iscb-careers-job-database
  - www.bioinformatics.org/jobs
  - www.biostars.org/t/jobs

Careers in the Federal Government

- https://www.usajobs.gov/
  - Federal positions
  - Both permanent and temporary
- Contracting companies
  - The government can pay money to a third party companies to staff on-site contractors to complete projects
  - The contracting company being used may differ across agencies

Careers in industry

Third party testing labs and food companies

- Some have adopted sequencing technologies and have a need for bioinformatics expertise

Hospitals or non-profit organizations

- Analyzing health-related data

Careers in Industry

Biosciences/biotech companies

- Software development
- Competitive, may require bioinformatics/CS degrees or expert knowledge of programming languages

Careers in Academia

Bioinformatics staff

- Most universities now have sequencing cores & associated staff
- Staff scientist positions in large research centers
- May or may not require bioinformatics degrees
- Bio experience an asset

Faculty in bioinformatics/CS/biological systems modeling

- Focus on developing new bioinformatics tools and working with biologists
- Often require CS or bioinformatics degree

Careers in Academia

Faculty in food science and other biological sciences

- Can apply bioinformatics principles and tools to solve applied problems

Additional Resources

Google drive folder with helpful information:

- Linux cheat sheets, terminology cheat sheets, helpful papers and more:

https://drive.google.com/open?id=1BlCbje6QsSWqm6TjU3h5MCbLKoVoPoSM

Questions?

Unanswered Questions?

Answers will be posted to IAFP website or you can contact--

Jess Chen: lly3@cdc.gov

Andrea Etter: ajray2011@gmail.com

Peter Cook: peter.cook@uga.edu

Lee Katz: gzu2@cdc.gov