MiCloud and BioDocklets: A Plug-n-Play, on-premises Bioinformatics Cloud, Providing Seamless Execution of NGS Pipelines.

Ntino Krampis, Ph.D.
Associate Professor, Biological Sciences
Hunter College, City University of New York (CUNY)

Director of Bioinformatics
Center for Translational and Basic Research (CTBR-CUNY)

Faculty, Institute of Computational Biomedicine
Weill Cornell Medical College
• Bioinformatics for Next Generation Sequencing (NGS).

• NGS analysis pipelines for QC, RNA-seq, Hi-C, metagenomics, variant discovery, genome assembly.

• Integrative analysis of variation, expression, chromatin and epigenetic data from TCGA, Encode, 4DN.

• Meta-barcoding for conservation and biodiversity monitoring using environmental DNA (eDNA).
- 500 CPU, 3 TB aggregate memory, 2 PB storage.
- Scalability: Kubernetes, NextFlow, Docker Swarm.
- Cross-platform bioinformatics through Docker containers.
- Visualization of genomic data on cloud databases using HTML5 / D3.js and in-browser computing.
Next Generation Sequencing is expensive.

- Expensive: $300-$600K or more initial investment per sequencer.
- Dedicated teams of laboratory technicians within a core sequencing facility.
- Investment in computational infrastructure and bioinformatics personnel.
Next Generation Sequencing can be affordable.

- MiSeq ($90K), MiniSeq ($50K), iSeq ($20K), Oxford Nanopore MinION ($1K).

- MiSeq: Small genomes, 16S metagenomics and barcoding, human transcriptomes and exons, deep sequencing of gene panels.

- MiSeq: $400 for library prep, $400-$1000 reagents for sequencing run.
Bioinformatics is the bottleneck.

- **Software complexity**: 5-10 or more algorithms in each bioinformatics data analysis pipeline, complex software dependencies, code and data libraries.

- **Usability**: Linux command-line expertise, managing large-scale input-output datasets, coordinating data flow between software components in a pipeline.

- **Provenance**: distribute working copy of the pipelines, track software versions.

- **Computing infrastructure**: large-scale computing capacity on a cluster or the cloud.

- **Output analysis and interpretation**: tabular output and static visualizations.

Usability and software complexity.

From: harmonized pipelines, ENCODE, TCGA

Bioinformatics pipelines without the command line.

Galaxy Workflow Canvas
http://galaxy.hunter.cuny.edu/
Provenance and computing infrastructure.

- Operating system, bioinformatics pipelines, and supporting data, pre-installed in Virtual Machine (VM) Container.

- The VM Container is a complete Linux computer server in a single binary file.

- Runs independently of underlying hardware through virtualization (Amazon, VirtualBox, Docker, Vmware).

- Cloud BioLinux: the first public bioinformatics VM on the Amazon cloud in 2010.


https://www.docker.com/
Running VM containers on the Amazon global cloud.

- Amazon Elastic Compute Cloud (EC2), rent on-demand VM container servers: up to $13 per hour depending on capacity.
- Max capacity 2TB RAM / 128CPU (can run hundreds of these).
- Data storage $0.1 per GB per month, or archival for $0.01.

http://aws.amazon.com/
Pipelines on Cloud BioLinux VM, build once, run on multiple platforms.

- Improving usability, reducing complexity.
- Provenance: collaborators can receive software and data, also used in publications.
- Can seamlessly run on local or remote clouds, and desktops / lab servers.

Internal cloud / cluster.
Bioinformatics pipeline output visualization.

- Web 1.0 technology, multi-tier, complicated stack.
- Static visualizations, not portable to smartphone user interfaces.
- Centralized databases, dependent on provider to provider maintenance and scalability.

New data visualization paradigms.

- Data-Drive Documents (D3) Javascript.
- Web 2.0, distributed databases, Application Programming Interfaces (APIs).
- Web browser computes the visualization instead of centralized web application (remember SETI @ home?).

https://d3js.org
Visual Omics Explorer (VOE), Web 2.0 for bioinformatics.

• Runs purely within the web browser: http://bcil.github.io/VOE/

• Import data from Google Genomics API, DropBox, Google Drive, FTP, local data.

• GFF, BED, PhyloXML, tabular etc

• Javascript - D3 / HTML5 multi-threaded (“parallel”) computing.

• Works well on smartphones and tablets: https://tinyurl.com/omics-explorer

VOE: demonstration of mobile interface.

tutorials: http://tinyurl.com/bioit-cuny
BioDocklets: integrated bioinformatics solution.

- Pre-configured pipelines in Galaxy, integrated with VOE and Docker UI.
- Run on multiple platforms, modify pipelines or build new using the Galaxy interface.
- VOE output is HTML / D3.js loaded with the data from the pipeline output.
- Docker UI abstracts the multi-step pipeline in a single page / command.


Galaxy

VOE

Docker UI
MiCloud: on-premises, scalable bioinformatics cloud for single step execution of complex bioinformatics pipelines.

- Abstracting multi-step data analysis to a simple interface.
- Run and monitor multiple pipeline instances in parallel.
- Data output automatically loaded in VOE visualizations.

https://github.com/kevana/ui-for-docker
MiCloud and BioDocklets: a plug and play, on-premises bioinformatics cloud for seamless execution of NGS pipelines.

BioDocklets from virtual machine repositories.

http://hub.docker.com

MiCloud "master" virtual machine.

Cloud or local cluster, kubectl or Docker daemon.

File Manager virtual machine.

NGS sequencing platform.

MiCloud deployment scripts.

Docker UI.

<table>
<thead>
<tr>
<th>Publication</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>miR-1207-3p regulates the androgen receptor in prostate cancer via FNDC1/fibronectin.</td>
</tr>
<tr>
<td></td>
<td>Non-synonymous variations in cancer and their effects on the human proteome: workflow for NGS data bio-curation and proteome-wide analysis of TCGA data.</td>
</tr>
<tr>
<td></td>
<td>In Vitro Mutational and Bioinformatics Analysis of the M71 Odorant Receptor and Its Superfamily.</td>
</tr>
<tr>
<td></td>
<td>CensuScope: census-based, rapid and accurate metagenome taxonomic profiling.</td>
</tr>
<tr>
<td></td>
<td>Fast functional annotation of metagenomic shotgun data by DNA alignment to a microbial gene catalog.</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
BioDocklets: integrated bioinformatics solution, pre-configured pipelines, visualization, runs local or the cloud.

MiCloud: leverages BioDocklets for deployment of a local cloud with intuitive UI for large-scale data analysis.

Seamless NGS data processing, from sequencer output to smartphone visualizations, enabling genomic data science.

Thank you! Follow up: kk104@hunter.cuny.edu