

# 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)

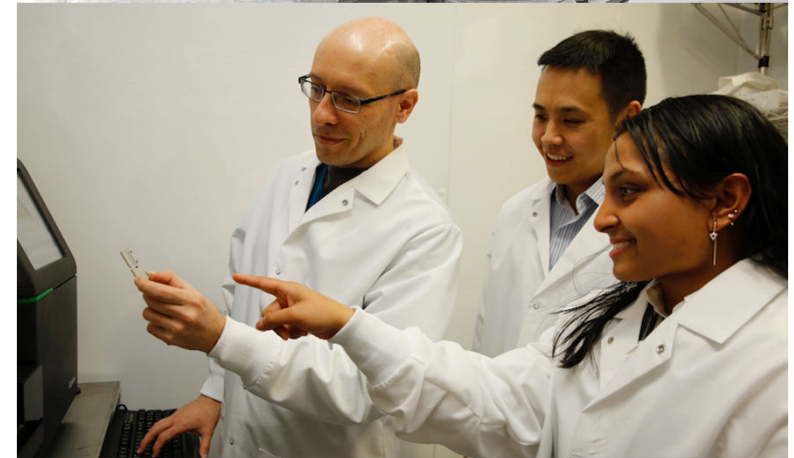
Director of Bioinformatics  
Center for Translational and Basic Research (CTBR-CUNY)

Faculty, Institute of Computational Biomedicine  
Weill Cornell Medical College



# Bioinformatics Core Infrastructures Lab (BCIL)

- 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).

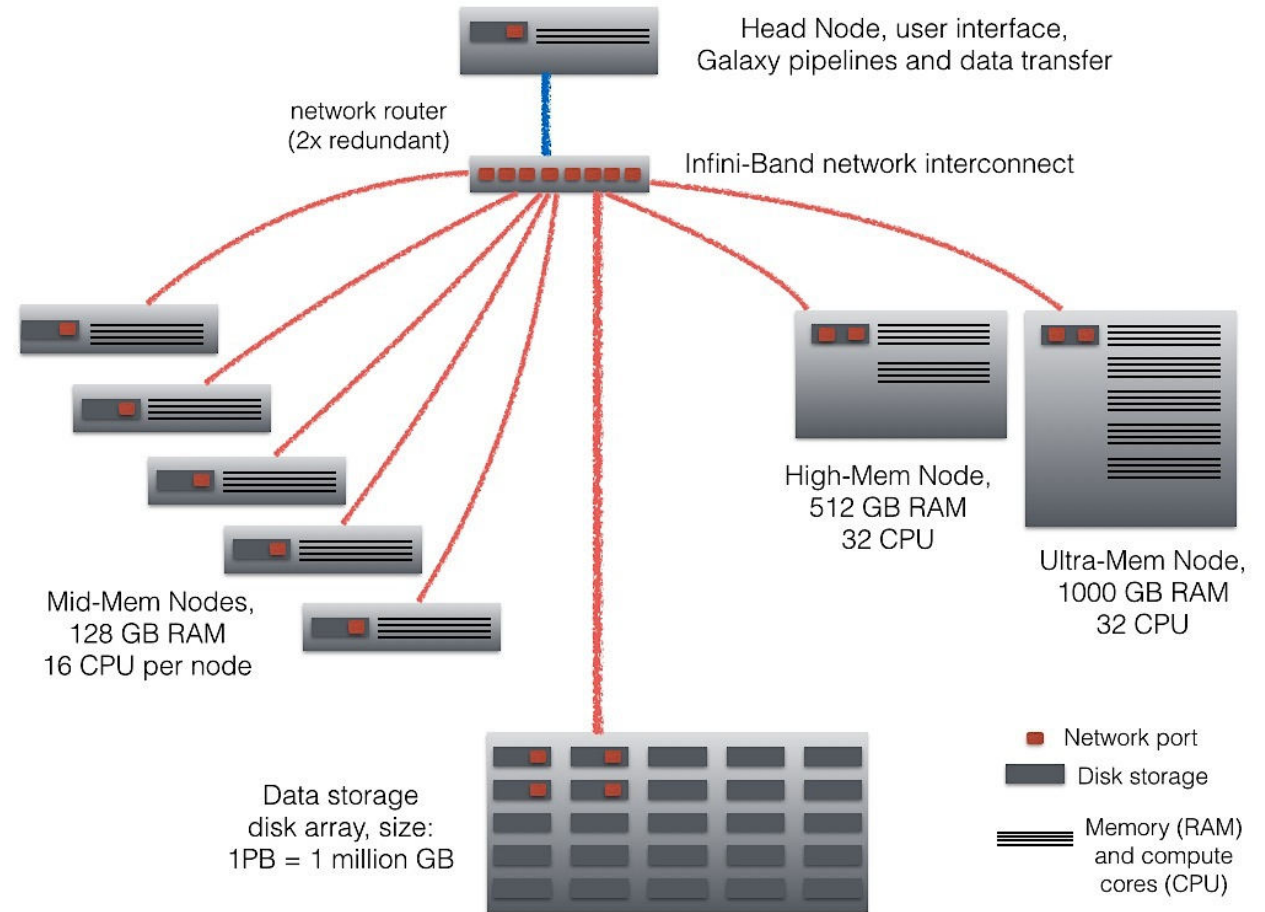


**Weill Cornell**  
**Medicine**



# BCIL Infrastructure and Computational Research

- 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.

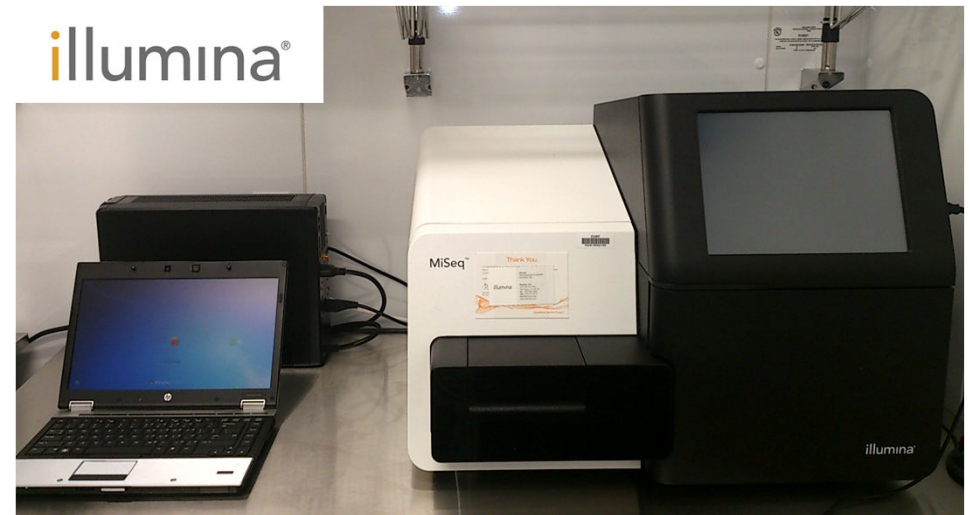
- Expensive: \$300-\$600K or more initial investment per sequencer.
- Dedicated teams of laboratory technicians within a core sequencing facility.
- Investment in computational infrastructure and bioinformatics personnel.





# 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 for sequencing run.

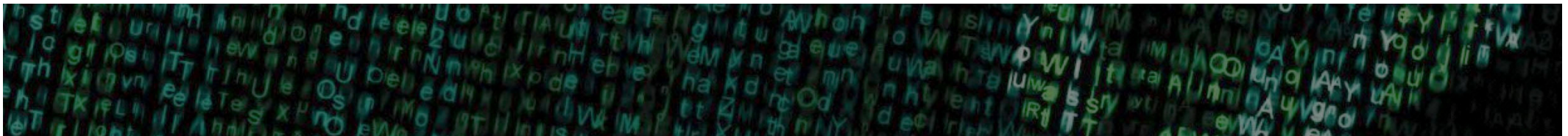




# Bioinformatics is the bottleneck.

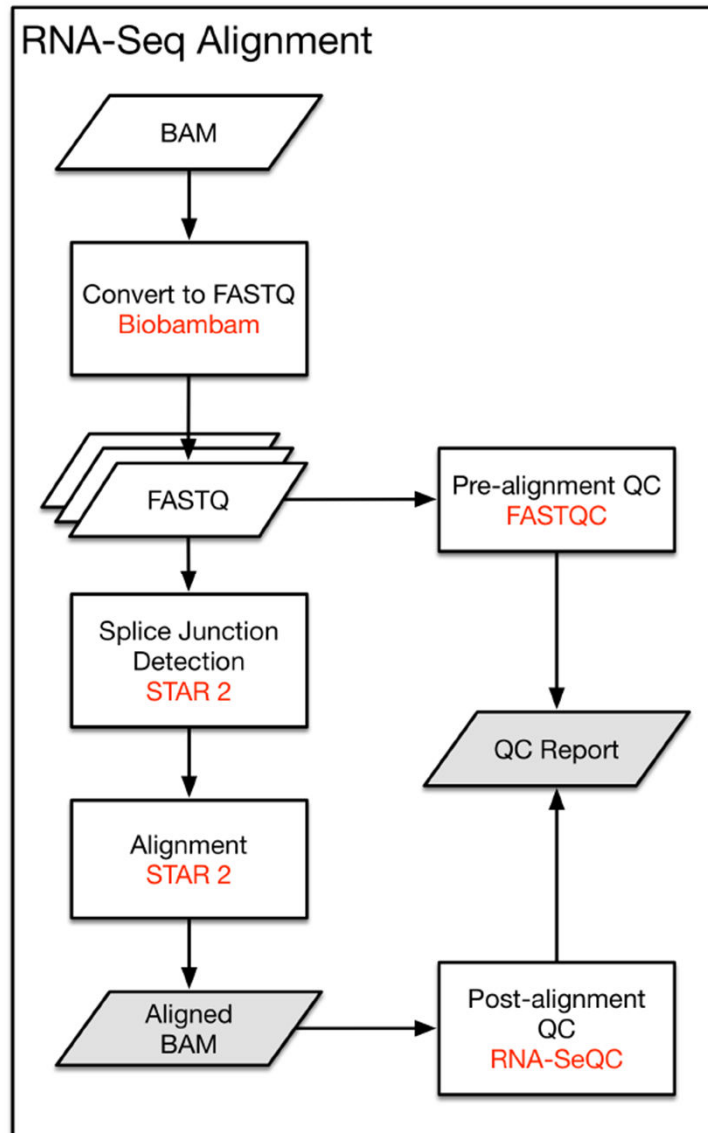
- **Software complexity**: 5-10 or more algorithms in each bioinformatics data analysis pipeline, complex software dependencies, code and data libraries.
- **Usability**: Linux command-line expertise, managing large-scale input-output datasets, coordinating data flow between software components in a 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.

Kompis R & Wulfinck C. (2013) *Methods in Next-Generation Sequencing*, 2(1). "A Review of Cloud Bioinformatics Solutions for Next-Gen Sequencing Data Analysis and Research".





# Usability and software complexity.



From: harmonized pipelines, ENCODE, TCGA

## A. Strand-specific RNA-seq

1. At Step 1, supply the option '--library-type' to TopHat to enable strand-specific processing of the reads. TopHat will map the reads for each sample to the reference genome and will attach meta-data to each alignment that Cufflinks and Cuffdiff can use for more accurate assembly and quantification. The --library-type option requires an argument that specifies which strand-specific protocol was used to generate the reads. See Table 1 for help in choosing a library type.

```

$ tophat -p 8 -G genes.gtf -o C1_R1_thout --library-type=fr-firststrand \
  genome C1_R1_1.fq C1_R1_2.fq
$ tophat -p 8 -G genes.gtf -o C1_R2_thout --library-type=fr-firststrand \
  genome C1_R2_1.fq C1_R2_2.fq
$ tophat -p 8 -G genes.gtf -o C1_R3_thout --library-type=fr-firststrand \
  genome C1_R3_1.fq C1_R3_2.fq
$ tophat -p 8 -G genes.gtf -o C2_R1_thout --library-type=fr-firststrand \
  genome C2_R1_1.fq C2_R1_2.fq
$ tophat -p 8 -G genes.gtf -o C2_R2_thout --library-type=fr-firststrand \
  genome C2_R2_1.fq C2_R2_2.fq
$ tophat -p 8 -G genes.gtf -o C2_R3_thout --library-type=fr-firststrand \
  genome C2_R3_1.fq C1_R3_2.fq
  
```

## B. Quantification of reference annotation only (no gene/transcript discovery)

1. At Step 1, supply the option '--no-novel-juncs' to TopHat to map the reads for each sample to the reference genome, with novel splice discovery disabled:

```

$ tophat -p 8 -G genes.gtf -o C1_R1_thout --no-novel-juncs genome C1_R1_1.fq C1_R1_2.fq
$ tophat -p 8 -G genes.gtf -o C1_R2_thout --no-novel-juncs genome C1_R2_1.fq C1_R2_2.fq
$ tophat -p 8 -G genes.gtf -o C1_R3_thout --no-novel-juncs genome C1_R3_1.fq C1_R3_2.fq
$ tophat -p 8 -G genes.gtf -o C2_R1_thout --no-novel-juncs genome C2_R1_1.fq C1_R1_2.fq
$ tophat -p 8 -G genes.gtf -o C2_R2_thout --no-novel-juncs genome C2_R2_1.fq C1_R2_2.fq
$ tophat -p 8 -G genes.gtf -o C2_R3_thout --no-novel-juncs genome C2_R3_1.fq C1_R3_2.fq
  
```

2. Skip PROCEDURE Steps 2-4.

3. Run Cuffdiff using the reference transcriptome along with the BAM files from TopHat for each replicate:

```

$ cuffdiff -o diff_out -b genome.fa -p 8 -u genes.gtf \
  ./C1_R1_thout/accepted_hits.bam, ./C1_R2_thout/accepted_hits.bam, ./C1_R3_thout/accepted_hits.
  bam \
  ./C2_R1_thout/accepted_hits.bam, ./C2_R3_thout/accepted_hits.bam, ./C2_R2_thout/accepted_hits.
  bam
  
```

## C. Quantification without a reference annotation

1. Map the reads for each sample to the reference genome:

```

$ tophat -p 8 -o C1_R1_thout genome C1_R1_1.fq C1_R1_2.fq
$ tophat -p 8 -o C1_R2_thout genome C1_R2_1.fq C1_R2_2.fq
$ tophat -p 8 -o C1_R3_thout genome C1_R3_1.fq C1_R3_2.fq
$ tophat -p 8 -o C2_R1_thout genome C2_R1_1.fq C1_R1_2.fq
$ tophat -p 8 -o C2_R2_thout genome C2_R2_1.fq C1_R2_2.fq
$ tophat -p 8 -o C2_R3_thout genome C2_R3_1.fq C1_R3_2.fq
  
```

2. Perform PROCEDURE Steps 2 and 3.

3. Run Cuffmerge on all your assemblies to create a single merged transcriptome annotation:

```
cuffmerge -s genome.fa -p 8 assemblies.txt
```

## D. Analysis of single-ended sequencing experiments

1. At Step 1, simply supply the single FASTQ file for each replicate to TopHat to map the reads for each sample to the reference genome:

```

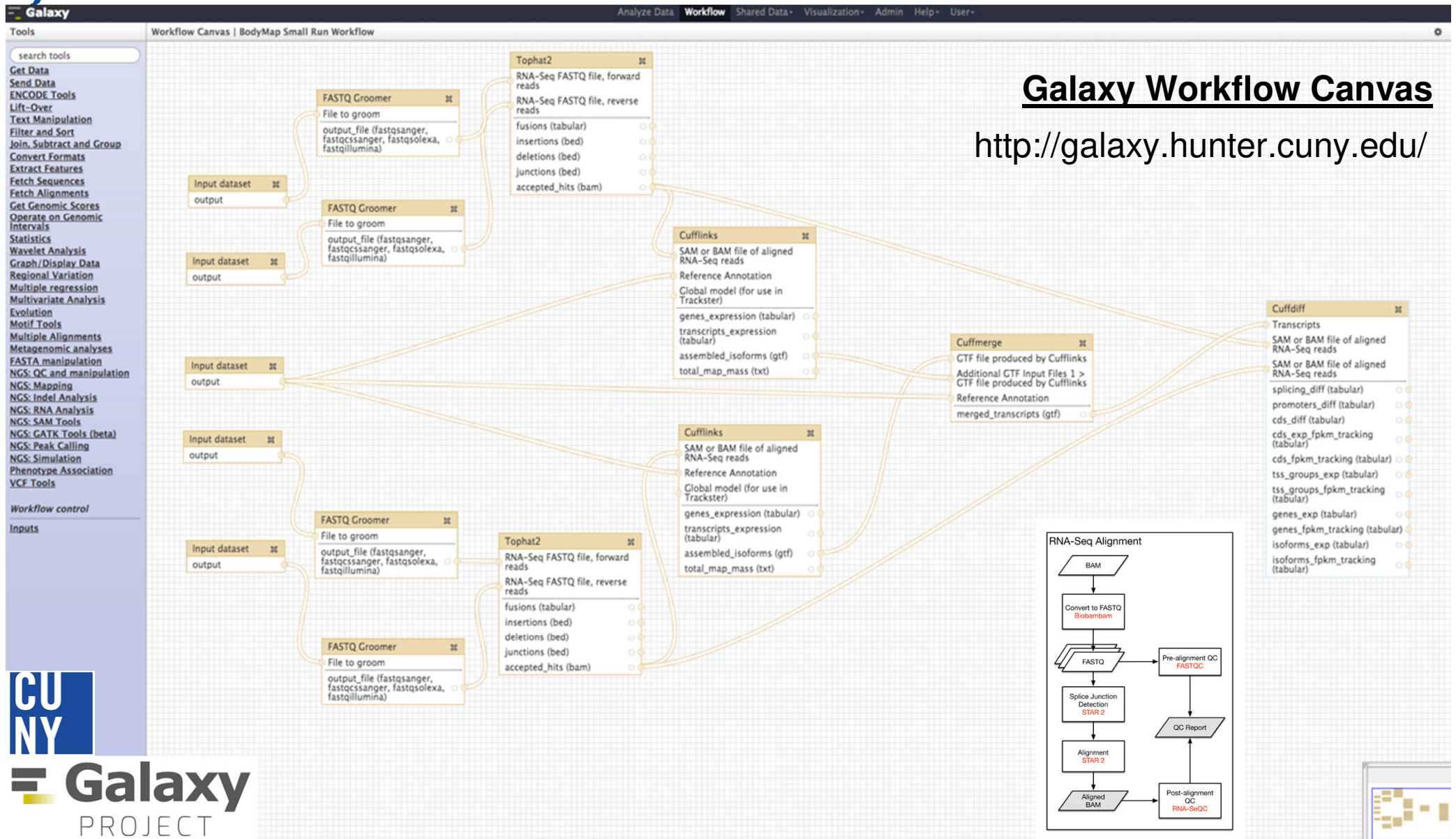
$ tophat -p 8 -G genes.gtf -o C1_R1_thout genome C1_R1.fq
$ tophat -p 8 -G genes.gtf -o C1_R2_thout genome C1_R2.fq
$ tophat -p 8 -G genes.gtf -o C1_R3_thout genome C1_R3.fq
$ tophat -p 8 -G genes.gtf -o C2_R1_thout genome C2_R1.fq
$ tophat -p 8 -G genes.gtf -o C2_R2_thout genome C2_R2.fq
$ tophat -p 8 -G genes.gtf -o C2_R3_thout genome C2_R3.fq
  
```

2. Perform PROCEDURE Steps 2-18.

From: Trapnell et al. Nature Protocols 7, 562-578 (2012)



# Bioinformatics pipelines without the command line.

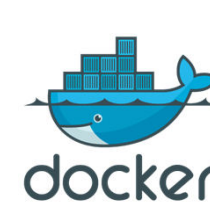
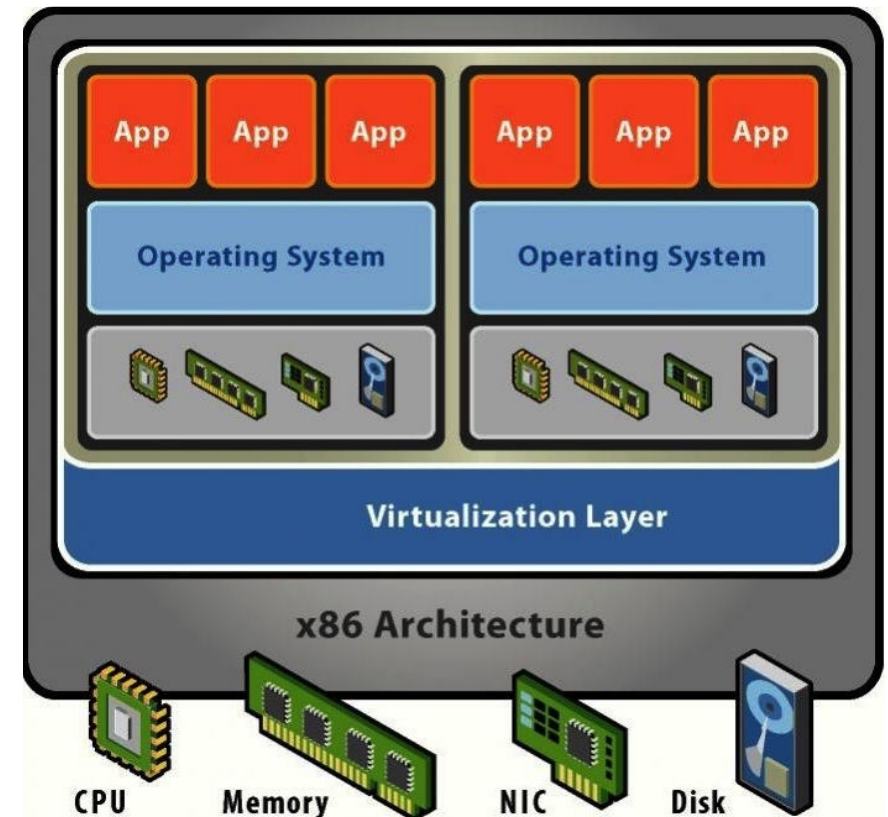






# Provenance and computing infrastructure.

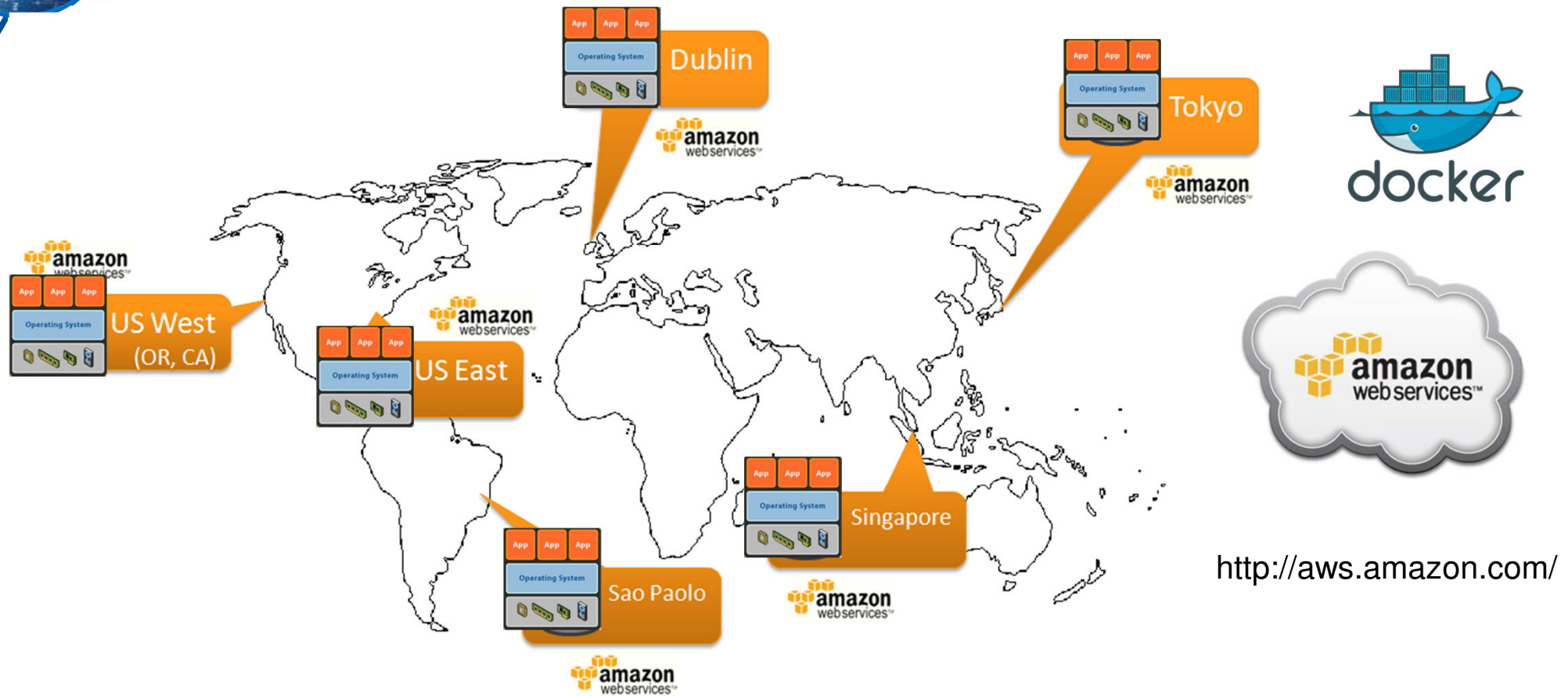
- Operating system, bioinformatics pipelines, and supporting data, pre-installed in Virtual Machine (VM) Container.
- The VM Container is a complete Linux computer server in a single binary file.
- Runs independently of underlying hardware through virtualization (Amazon, VirtualBox, Docker, Vmware).
- Cloud BioLinux: the first public bioinformatics VM on the Amazon cloud in 2010.



**Krampis K.** et al. (2012). Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community. BMC Bioinformatics, 13(1), 1-8.

<https://www.docker.com/>

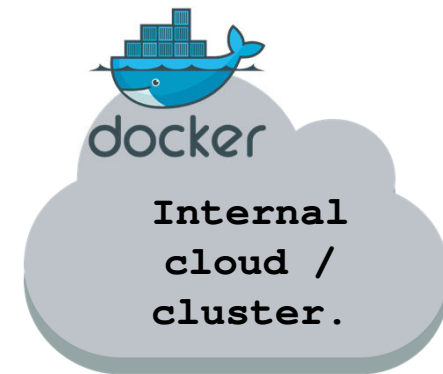
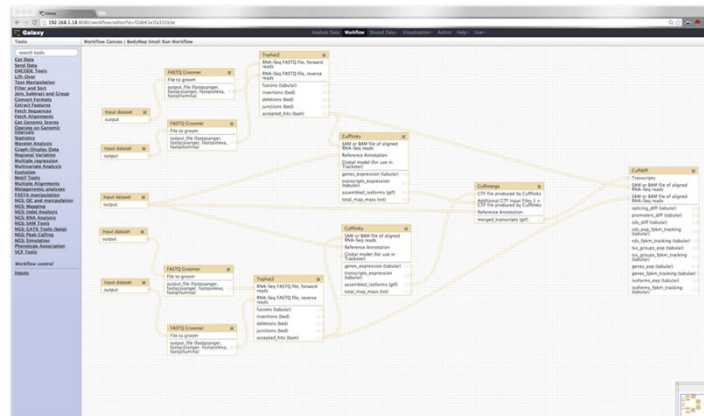
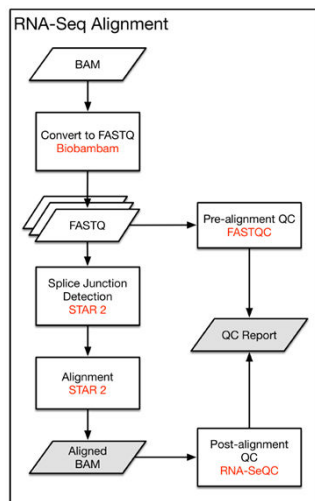
# Running VM containers on the Amazon global cloud.



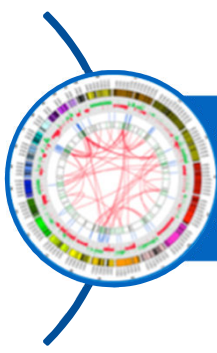
- Amazon Elastic Compute Cloud (EC2), rent on-demand VM container servers: up to \$13 per hour depending on capacity.
- Max capacity 2TB RAM / 128CPU (can run hundreds of these).
- Data storage \$0.1 per GB per month, or archival for \$0.01

# 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.

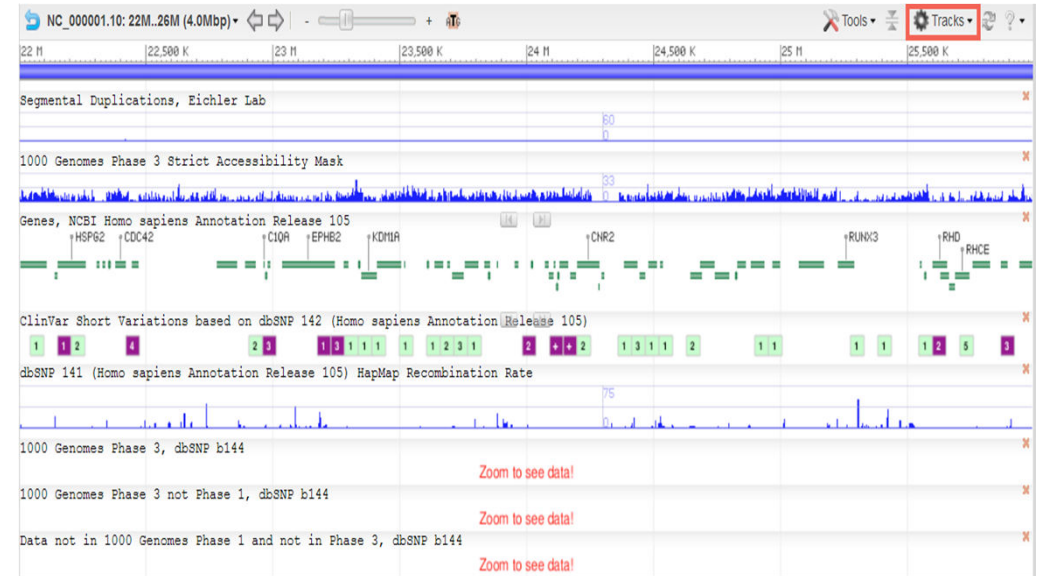






# Bioinformatics pipeline output visualization .

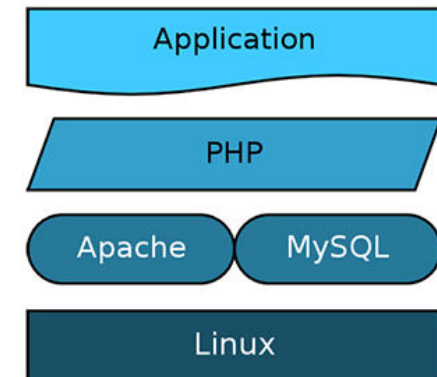
Galaxy						
Tools						
TopHat2 Gapped-read mapper for RNA-seq data	port_name	tss_id	locus	length	coverage	FPKM
NGS: RNA Analysis						FPKM_conf_I
FILTERING	03	-	NC_004460.2:0-801	801	16.3301	92.4725
Cuffdiff find significant changes in transcript expression, splicing, and promoter use	04	-	NC_004460.2:2079-2544	465	17.4998	97.1707
Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data	06	-	NC_004460.2:803-1997	1194	33.5451	187.647
Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments	07	-	NC_004460.2:2789-4112	1323	0	0
Cuffmerge merge together several Cufflinks assemblies	08	-	NC_004460.2:6708-7542	834	39.9904	221.182
NGS: Simulation	09	-	NC_004460.2:7586-8600	1014	37.1129	210.797
Phenotype Association	13	-	NC_004460.2:8666-9896	1230	5.72867	31.6875
Assembly	14	-	NC_004460.2:9943-10612	669	5.73133	32.4949
Alignment	15	-	NC_004460.2:4162-6577	2415	38.5902	215.84
Workflows	16	-	NC_004460.2:15293-15905	612	74.2108	466.728
All workflows	18	-	NC_004460.2:16007-17084	1077	34.7048	192.674
	19	-	NC_004460.2:17133-18126	993	34.2126	189.459
	22	-	NC_004460.2:18129-19212	1083	42.5602	246.133
	20	-	NC_004460.2:19346-19949	603	63.9312	391.555
	24	-	NC_004460.2:20023-21172	1149	17.4843	98.0768
	23	-	NC_004460.2:23122-23401	279	49.3228	281.411
	25	-	NC_004460.2:21424-22870	1446	22.2061	128.155
	26	-	NC_004460.2:24834-25038	204	52.9505	320.179
	31	-	NC_004460.2:23693-24731	1038	11.9555	65.845
		-	NC_004460.2:11012-12656	1644	38.1737	215.53
		-	NC_004460.2:12649-13966	1317	34.4836	194.696
		-	NC_004460.2:13962-15186	1224	43.7778	247.172
		-	NC_004460.2:25061-26012	951	39.3544	223.758
		-	NC_004460.2:26008-26515	507	50.1444	285.108
		-	NC_004460.2:31023-32295	1272	42.7182	247.605

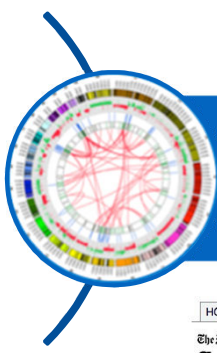


<https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/>

- 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.

## LAMP stack configuration





# New data visualization paradigms.

HOME PAGE TODAY'S PAPER VIDEO MOST POPULAR U.S. Edition ▼ kk104... ▼ Help

The New York Times

**Sunday Review** | The Opinion Pages

Search All NYTimes.com

Go

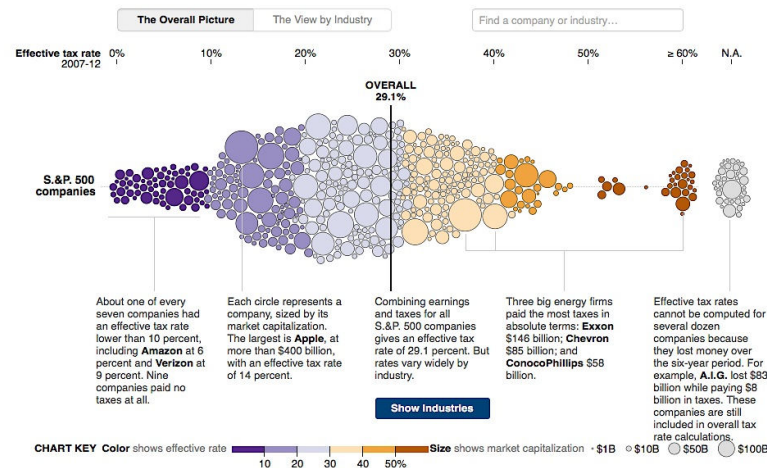
WORLD U.S. N.Y. / REGION BUSINESS TECHNOLOGY SCIENCE HEALTH SPORTS OPINION ARTS STYLE TRAVEL JOBS REAL ESTATE AUTOS

Published: May 25, 2013

FACEBOOK TWITTER GOOGLE+ EMAIL SHARE

## Across U.S. Companies, Tax Rates Vary Greatly

Last week, in a Congressional hearing, Apple got grilled for its low-tax strategy. But not every business can copy that approach. Here is a look at what S&P 500 companies paid in corporate income taxes — federal, state, local and foreign — from 2007 to 2012, according to S&P Capital IQ. [Related Article »](#)



HOME PAGE TODAY'S PAPER VIDEO MOST POPULAR U.S. Edition ▼ kk104... ▼ Help

The New York Times

**Politics**

Search All NYTimes.com

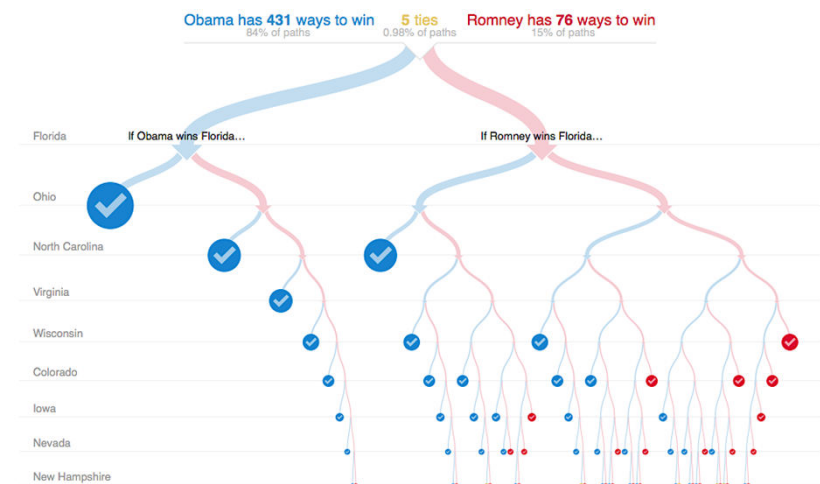
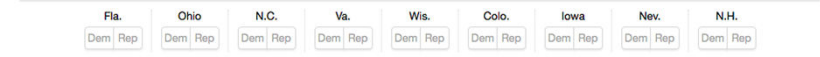
Go

Published: November 2, 2012

FACEBOOK TWITTER GOOGLE+ EMAIL SHARE

## 512 Paths to the White House

Select a winner in the most competitive states below to see all the paths to victory available for either candidate.



- 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 ?).

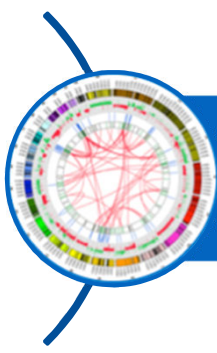
<https://d3js.org>

API.DATA.GOV

Open Data Network

**D3** Data-Driven





# Visual Omics Explorer (VOE), Web 2.0 for bioinformatics.

- Runs purely within in 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>



Google  
Genomics

illumina®

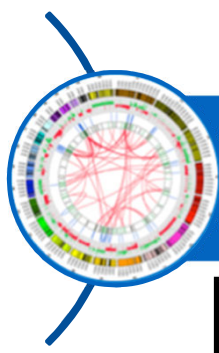


ANDROID

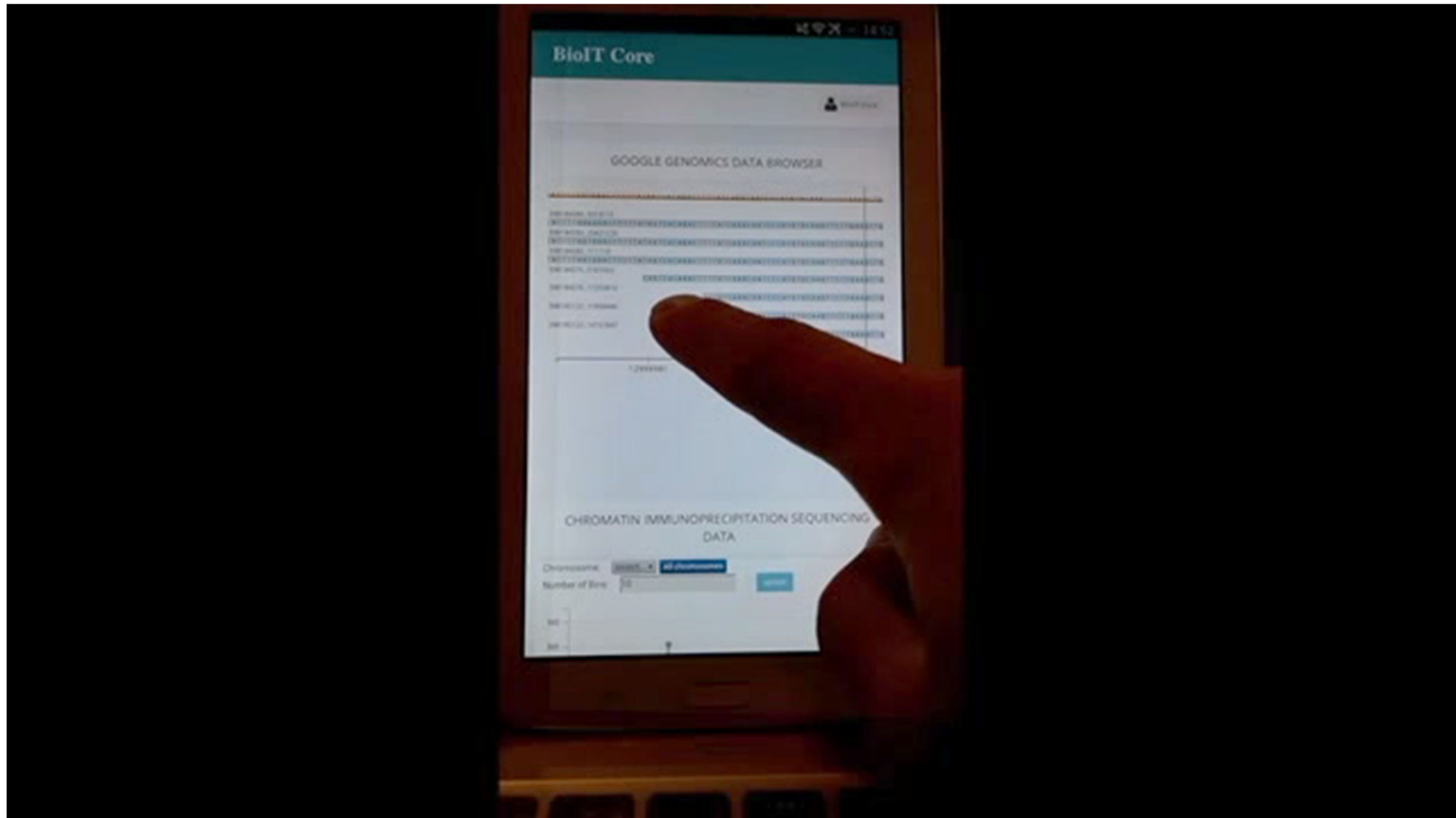


Visual Omics Explorer (VOE): a cross-platform portal for interactive data visualization.  
Kim, B., Ali, T., Hosmer, S. and **Krampis, K.** *Bioinformatics* (2016) 32 (13): 2050-2052.





# VOE: demonstration of mobile interface.



tutorials: <http://tinyurl.com/bioit-cuny>

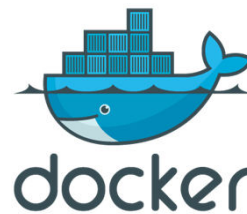
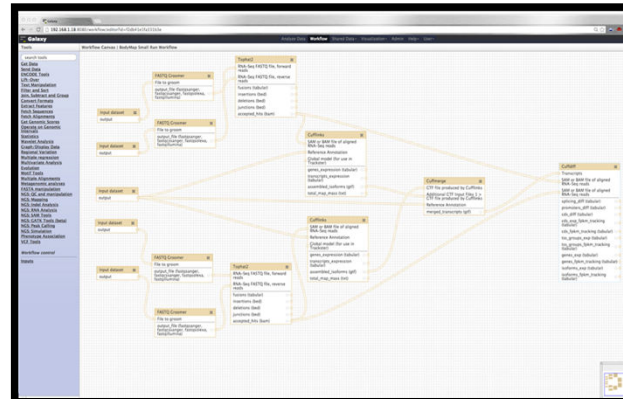




# BioDocklets: integrated bioinformatics solution.

- 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

Kim B., Ali T., Lijeron C. and **Krampis K.** (2017) "Bio-Docklets: Virtualization Containers for Portable, Scalable NGS Data Analysis". *Gigascience* 6(8):1-7

Alterovitz G., Dean D.A., .... **Krampis K.**, et al. (2017). bioRxiv, p.191783. "Enabling Precision Medicine via standard communication of NGS provenance, analysis, and results".





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

**Dashboard**

Pipeline: RNA\_Seq\_tp2\_dui\_MatInnerDistance\_200  
Username: /RNAseq\_inner\_dist\_200 [Rename](#)

[Start pipeline](#)

State: **stopped**

Environment:

Options:

mate\_std\_dev: 200

anchor\_length: 8

segment\_length: 25

[Save as new instance](#) [Cancel](#)

Internal Port: 8090/tcp

Exposed Port: 6200/tcp

Bindings: /home/dockerui/BCIL\_pipeline\_runs:/home/data

Galaxy Server: 146.95.173.35:6200

Output / Error messages: [Output / Error messages](#)

System Stats: [Stats](#)

[Remove this instance](#)

**All Pipelines**

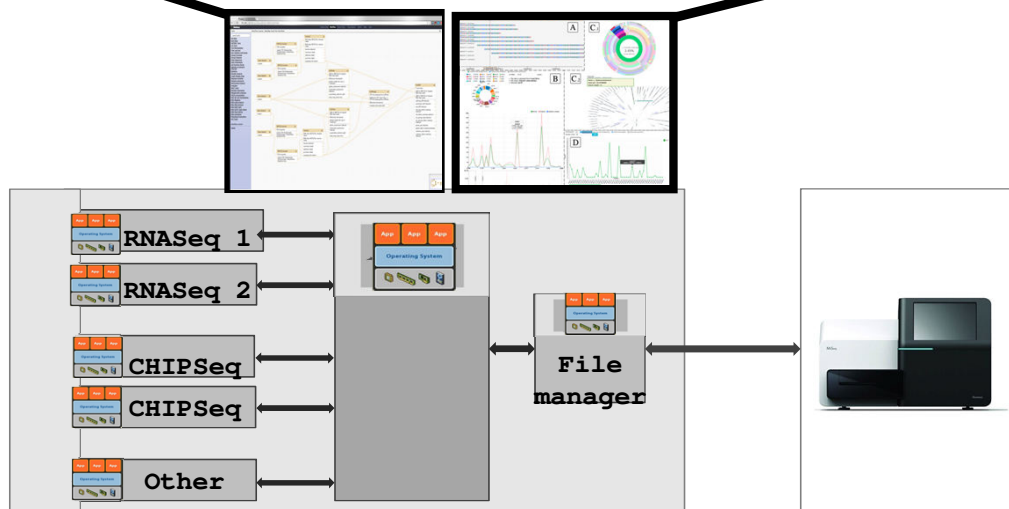
Run / Stop	Pipeline ID (username)	Pipeline Name	Started	Status
<a href="#">Pipeline detail</a>	RNAseq_inner_dist_200	RNA_Seq_tp2_dui_MatInnerDistance_200	12/5/2016, 6:34:36 PM	<a href="#">Running - Up 3 hours</a>
<a href="#">Pipeline detail</a>	RNAseq_inner_dist_170	RNA_Seq_tp2_dui_MatInnerDistance_170	12/5/2016, 6:34:35 PM	<a href="#">Running - Up 3 hours</a>
<a href="#">Pipeline detail</a>	RNAseq_inner_dist_150	RNA_Seq_tp2_dui_MatInnerDistance_150	12/5/2016, 6:34:34 PM	<a href="#">Stopped</a>
<a href="#">Pipeline detail</a>	RNAseq_inner_dist_130	RNA_Seq_tp2_dui_MatInnerDistance_130	12/5/2016, 6:34:33 PM	<a href="#">Stopped</a>
<a href="#">Pipeline detail</a>	RNAseq_inner_dist_100	RNA_Seq_tp2_dui_MatInnerDistance_100	12/5/2016, 6:34:31 PM	<a href="#">Stopped</a>
<a href="#">Pipeline detail</a>	ChIPseq_dockerui	ChIP_Seq_dui_paired_5010	12/5/2016, 6:34:27 PM	<a href="#">Stopped</a>

**Status**

**Running Pipelines**

- RNAseq\_inner\_dist\_200 - RNA\_Seq\_tp2\_dui\_MatInnerDistance\_200
- RNAseq\_inner\_dist\_170 - RNA\_Seq\_tp2\_dui\_MatInnerDistance\_170

<https://github.com/kevana/ui-for-docker>

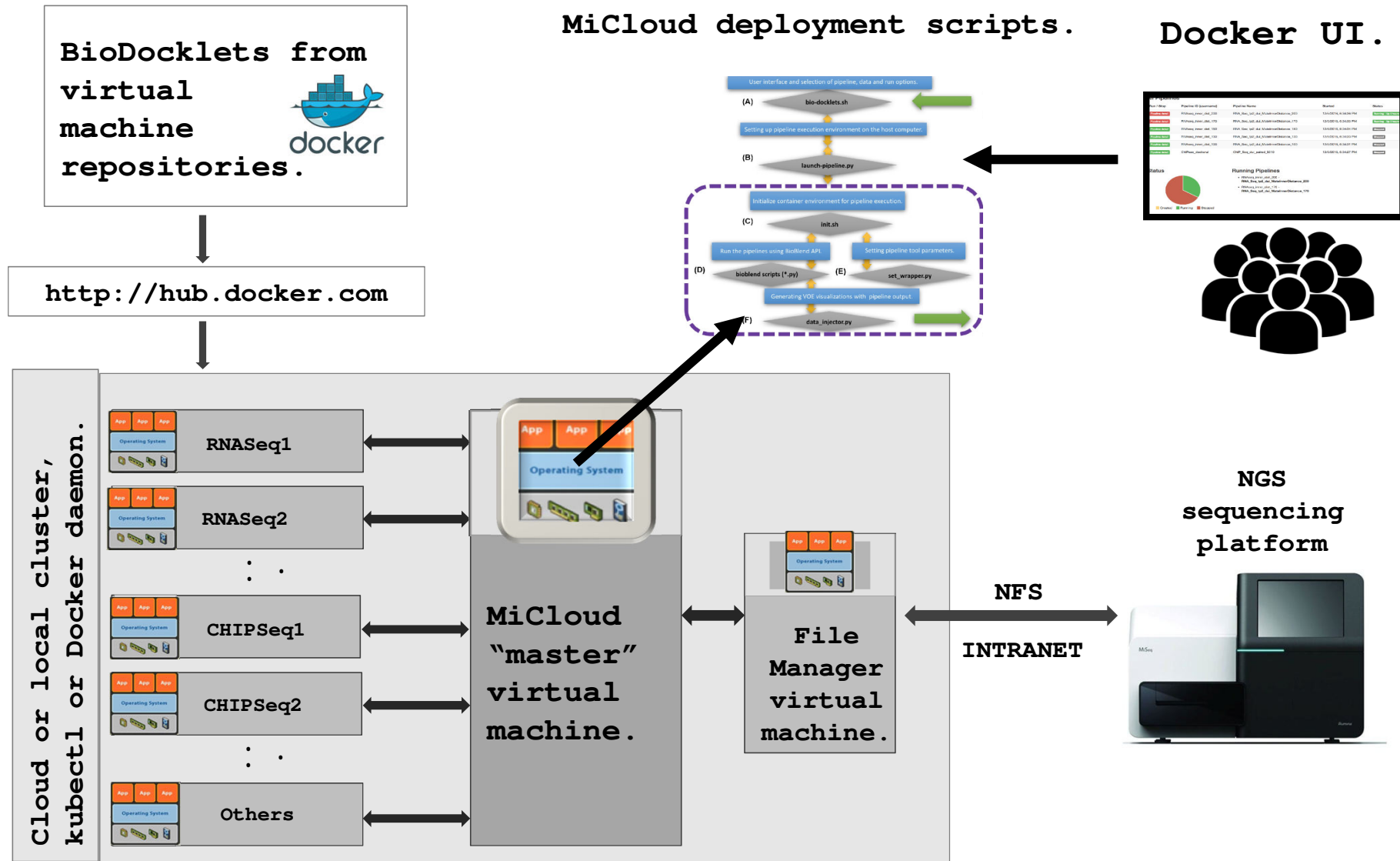


- Abstracting multi-step data analysis to a simple interface.
- Run and monitor multiple pipeline instances in parallel.
- Data output automatically loaded in VOE visualizations.





# MiCloud and BioDocklets: a plug and play, on-premises bioinformatics cloud for seamless execution of NGS pipelines.





## Selected MiCloud and BioDocklets publications.

### **miR-1207-3p regulates the androgen receptor in prostate cancer via FNDC1/fibronectin.**

Das DK, Naidoo M, Ilboudo A, Park JY, Ali T, **Krampis K** et al. Cell Research. 2016 Nov 1;348(2):190-200.

### **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.**

Cole C, **Krampis K** et al. BMC Bioinformatics. 2014 Jan 27;15(1):1.

### **In Vitro Mutational and Bioinformatics Analysis of the M71 Odorant Receptor and Its Superfamily.**

Bubnell J., Jamet S., Tomoiaga D., D'Hulst C., **Krampis K.** and Feinstein P. (2015) PloS ONE, 10(10):0141712.

### **Fibronectin and androgen receptor expression data in prostate cancer obtained from a RNA-sequencing bioinformatics analysis.**

Das DK, Ali T, **Krampis K** and Ogunwobi OO. Data in Brief. 2017 11:131-135.

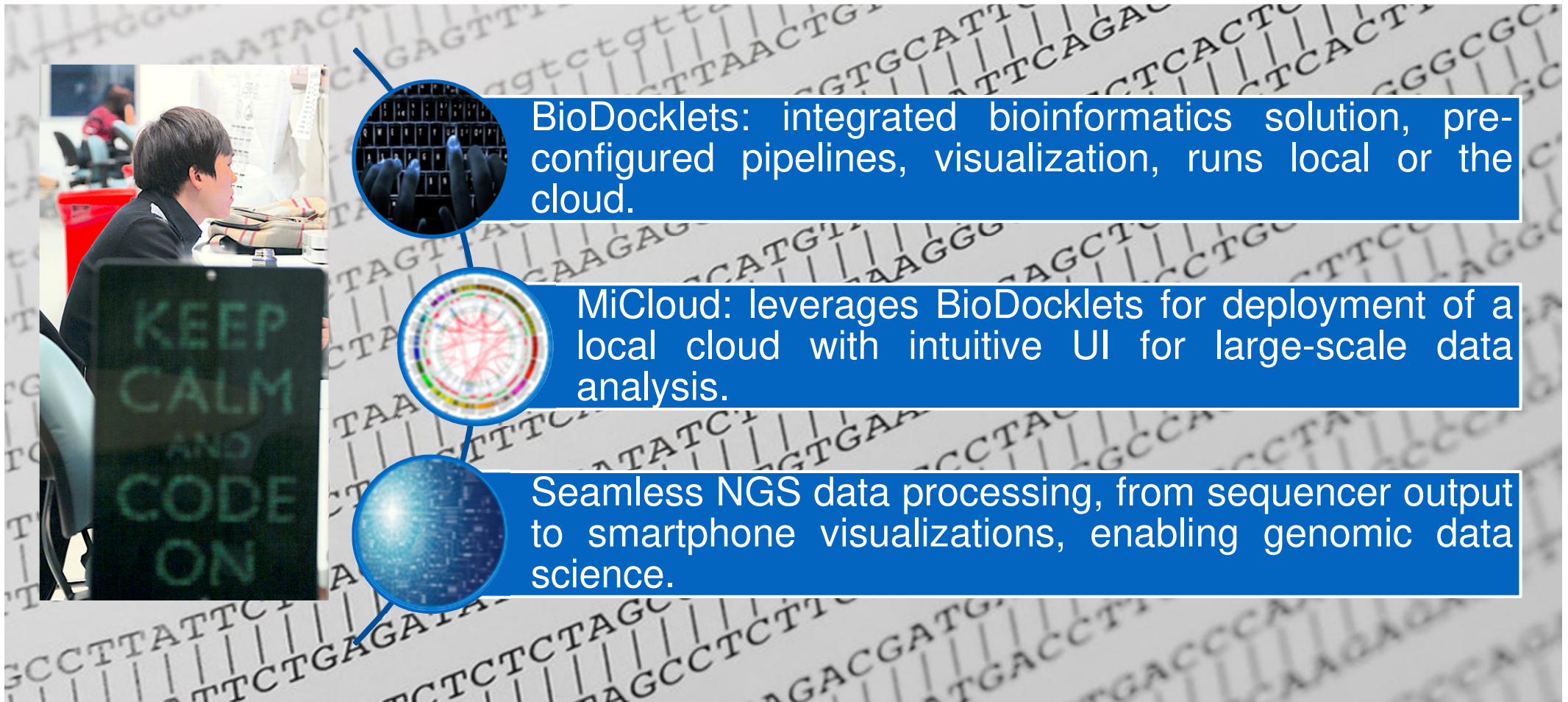
### **Fast functional annotation of metagenomic shotgun data by DNA alignment to a microbial gene catalog.**


Brown S, Hao Y, Chen H, **Krampis K** et al. BioRxiv 2017 March 2015 10:1101


### **CensuScope: census-based, rapid and accurate metagenome taxonomic profiling.**


Shamsaddini A, Pan Y, Johnson WE, **Krampis K** et al. Census-based rapid and accurate metagenome taxonomic profiling. BMC Genomics. 2014 Oct

# Summary



 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](mailto:kk104@hunter.cuny.edu)**