

Mining Large-scale Cancer Genomics Data Using Cloud-based Bioinformatics Approaches



A very short introduction to Cloud Computing

Jorge Andrade, Ph.D.

Why Cloud-based Bioinformatics

Patients



NGS Applications

Genomics
WGS, WES

Transcriptomics
RNA-Seq
miRNA-Seq

Epigenomics
Bisulfate-Seq
ChIP-Seq

Bioinformatics Analysis

Point mutations

InDels

CNV

Structural Variation

Differential Expression

Gene Fusion

Alternative Splicing

RNA Editing

Methylation

Transcription Factors

Histone Modifications

Integration and Interpretation

Functional
Effect of
Mutation

GO, Networks
and
Pathway
Analysis

Epigenetic
Modification
of
Function

Better understanding of Cancer - Clinical Application

Next Generation Sequencing = Big Data

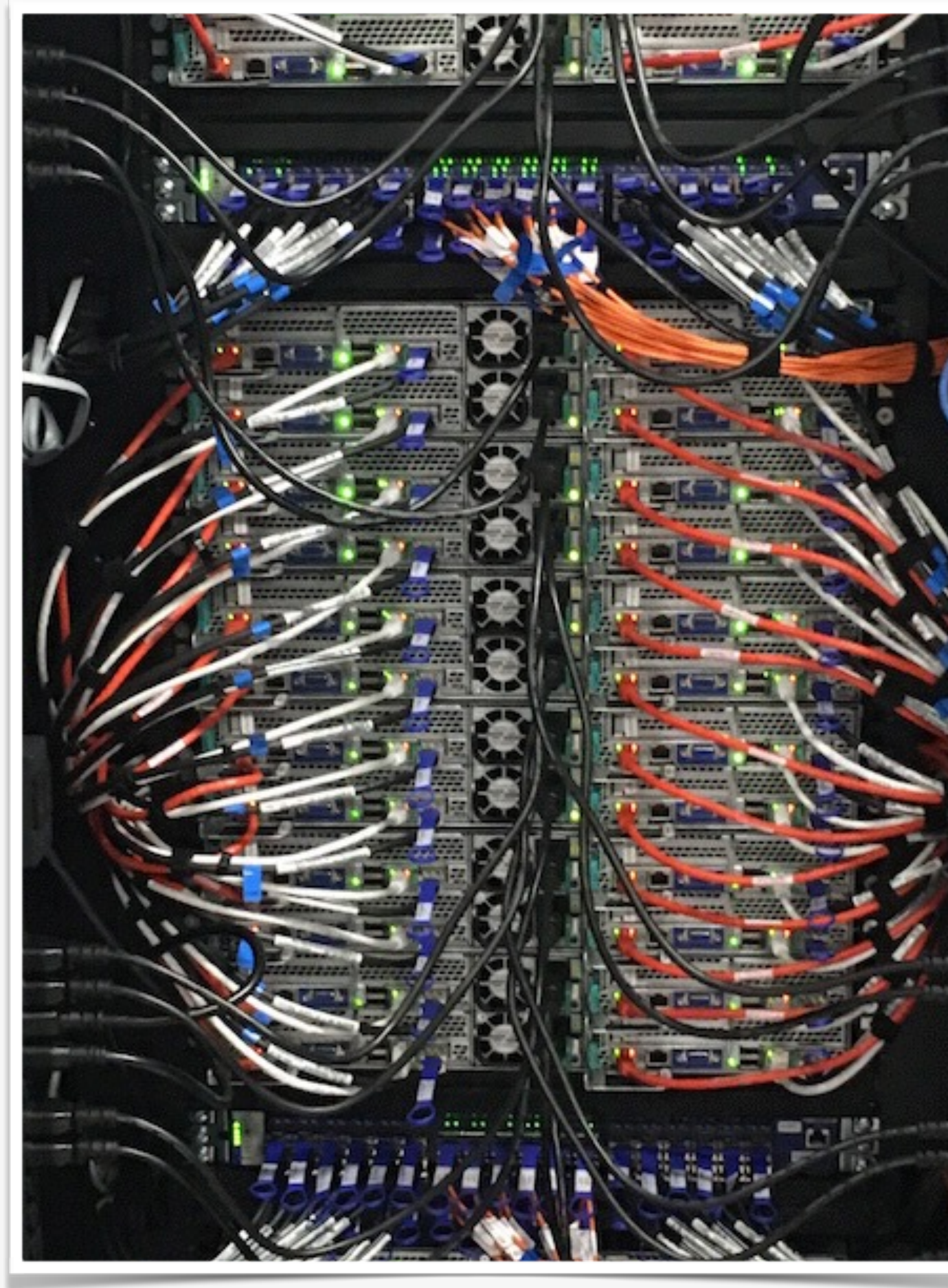
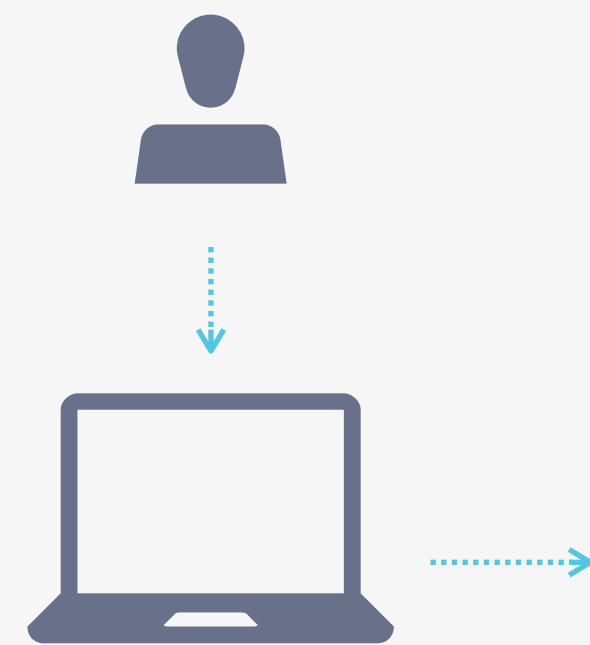


CATGACGTGCGGACACCCAGAATTGTCTTGAGGATGTAAGATCTAACCTCACTGCCGGGGAGGCTCATAC
CTGGGGCTTTACTGATGTCATACCGCTTGACGGGGATAGAATGACGGTGCCGCTGCTGCTGCTGCTCGAAGCA
ATTTTCTGAAGTTACAGACTCGATTAAGAATCGGACTGCCGCTGGGCGGAGAGACATGCGTGGTAGTCA
TTTTTGGACGCTCAAGGACTCAAGGGAAATAGTTGGGGGGAGCTTACAGCTTCAATTCCTCAAGGTCGAGA
CGATAAATCAACTACTGGTTTCGGCTAATAGGTCACGTTTATGTGAATAGAGGGGAACCGCTCCAAAT
CCCTGGGTGTTCTATGATAAGCTCCTGCTTATAACACGGGGCGTTAGGTTAAATGACTCTCTATCTTATGGTG
ATCCAGGCGCGGCTAATTCGCTGCTGTTAATGTTTATACCAATCACTCAGATCAGATTAGATCAAGGATGCGG
AGCCAGTGCAGGGCTGCTGCTGCTGTTGTCGACGCTCATGTTACTCTGGAATCTACCTGCCCTCCCTCAAC
GGTTAAGGGGTGTGATGACGATGCGAGTATACATCGGCTGCGACTACAGTGGTGGATGACGCTGCTGCTGCT
TCGCGGTTTCGGGCGCTAGTTGAGTGCATACCAACCGGTGGCAAGTAGCAAGAGACCTACCTGGGTCACTT
AGACACCTCACTTAATAGTCTTACGGGGAAATACCTTTACCACTCAATGCTCCAAATATATCTGACCGCTT
CAATGATATGCGCCACAGAAATAGGTTCTCAAGTATGCGATACGCGCGCGCGGCTCCAGCTACGCTCAGGAC
GACGATAGAGGCTATTGTGTAATTCAGGCTCAGCATTCATGACCTTTCTGTTGTAATTTGCTGAATGCA
TCTGCTCGGTAAAGTCTGGGGGGCAACGAAATATCCGATTTCTGCTCTACGGGTCACAAATGAGAAATGCT
TGGCGGTGATGCTGAGTTAAATTAATTCAGGCTACGGTAACTTGTAGTGAAGTCAAGAAATCAGGGAATC
ACGGGTTCGCTACAGATGAATGAATTAACACGCAACCTCATGCGCCATTGGGCTGGGACCGCAGATCA
AAAGTGGCAGATTAGGAGTGTCTGATCAGGTTAGCAGGTGGACTGTATCCACAGCGCATCAAACTTCAATAAAT
CCAAAGGCTTGTAGTGGCTTACGACCGCTGAACAGTGGGCGCATGTTAGGTTAGTACAACTTTCCCGCTT
AGGTGCGACATGGGGCAGTTAGCTGCGCTATATCCCTTGACACGTTCAATAAGAGGGGCTCACAGCGCGC
TTTTTAATTAGGATGCGGACCCCATCTTGTAACTGTATGTTCTATAGATATTTCTTCAGGAGTAATAGGACA
AGCTGACACCGAGGGTCAACAAATTTCTACTATCACCCTGGAACGCTGCTTTGCAAGAACCACTGGG
CTTAGATTCGGCTTCACTAGCTAGTGGGGCGAGTATCATAGATCAGGCTAGCAAAACGAGCTGAGTCTA
CACACGAGTTGTAAACACTTGATTGCTATCTGATACGCAAGGATCTCTACATCAAGACTACGGGG
ATCTGGATCGAGTCAGAAATACGAGTTAATGCAATTTACGATAGCGGTGAACACGCTGCCATGGGTGCGT
AGACCGTATGCAAGTGGGGGCGCTATTGTCAGCAACGCGTGGAGTATACAGAAATGCTCTTCTACGAGTA
AGGAGCTCGGTCCCAATGCGCAAAAGGAATAAGATTCAAACTGCGCATGCTCCCTCGCGCGGTGCA
CTATTATCCATCGAAGTTGAACCTACTTCTCGGCTTATGCTGCTCCTCAACGATATGCTTATGAATGCGATG
CGGCTGTGGATCTTAAGGGCCACATCTTAATTCGACGATCAACGATGCGCTTTCTGCTGGTACAATGAGT
ACTAAGTTATCCAGTCAAGTTTGACCGGACTCGTATGACATGTTGACGTAACCGGGAGGAATGCAAGAA
CTGTTTCAAGGCTCTGCTTGGTATCACTCAATATTCAGACGAGCAAGTGGCAAAATTTGTCGCGCTCTC
CTAGGATTTCAGCAACCGTGTAACTATGCACTAAGGATAACTAGCGCAGGGGGGACATAGTCCCGGAGCT
AAGACTACCTATAGATTCCTTGGAGCGGGACATGACAGCGGTTAGCACACAAATATCGGGATGCTTGA
GGTATTATTAGCAGCAATAAAGGACTTGACAGAGACTTATTAGAATTCACAAACAGGATCATATCATGG
GTGTTGGGTGGGCAAGTCCCGAAGCTCGGCCAAAGATTCCCATGGACCGTCTGGTCTGTTAGCGGTGAC
GCTGCTGCTGTTCCGGGTACCATAGATAGACTGAGATTGCTCAAAAAATTTGGGCGAAATAGAGGGGCTCCT
TGTAGAAATCCAGACTGGGAATTAAGGCTTTCCACTATCTGAGCGCTAATACATCAACAAATGCTCTACT
CGAATCCGAGTAGGCAATTACAACCTGGTTCAGATCACTGGTTAATCAGGGATGCTTCTAAGATTATCTTG
CCCCAGCGGACGCTCTTCAAGGGGCGATTTTGGACTTCAGATACGCTAGAATTAAGGGTCTCTTACAC
TCTGCGGGCTGCGAGGACCTAGAGCTTGGCGCTACTTGTCTCAGTCAATACGCGCGAAGCGCTGGGCA
CGTGACCTTAAGTGCAGAGCGAGTGAATTTGGAGCTAATATGGGTGAATAGAGACTTATATCATCAGGG



	 MiniSeq System	 MiSeq Series	 NextSeq Series	 HiSeq Series	 HiSeq X Series*
Key Methods	Amplicon, targeted RNA, small RNA, and targeted gene panel sequencing.	Small genome, amplicon, and targeted gene panel sequencing.	Everyday exome, transcriptome, and targeted resequencing.	Production-scale genome, exome, transcriptome sequencing, and more.	Population- and production-scale whole-genome sequencing.
Maximum Output	7.5 Gb	15 Gb	120 Gb	1500 Gb	1800 Gb
Maximum Reads per Run	25 million	25 million [†]	400 million	5 billion	6 billion
Maximum Read Length	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp
Run Time	4–24 hours	4–55 hours	12–30 hours	<1–3.5 days (HiSeq 3000/HiSeq 4000) 7 hours–6 days (HiSeq 2500)	<3 days
Benchtop Sequencer	Yes	Yes	Yes	No	No
System Versions	<ul style="list-style-type: none">MiniSeq System for low-throughput targeted DNA and RNA sequencing	<ul style="list-style-type: none">MiSeq System for targeted and small genome sequencingMiSeq FGx System for forensic genomicsMiSeqDx System for molecular diagnostics	<ul style="list-style-type: none">NextSeq 500 System for everyday genomicsNextSeq 550 System for both sequencing and cytogenomic arrays	<ul style="list-style-type: none">HiSeq 3000/HiSeq 4000 Systems for production-scale genomicsHiSeq 2500 Systems for large-scale genomics	<ul style="list-style-type: none">HiSeq X Five System for production-scale whole-genome sequencingHiSeq X Ten System for population-scale whole-genome sequencing

High Performance Computing Cluster (HPC)



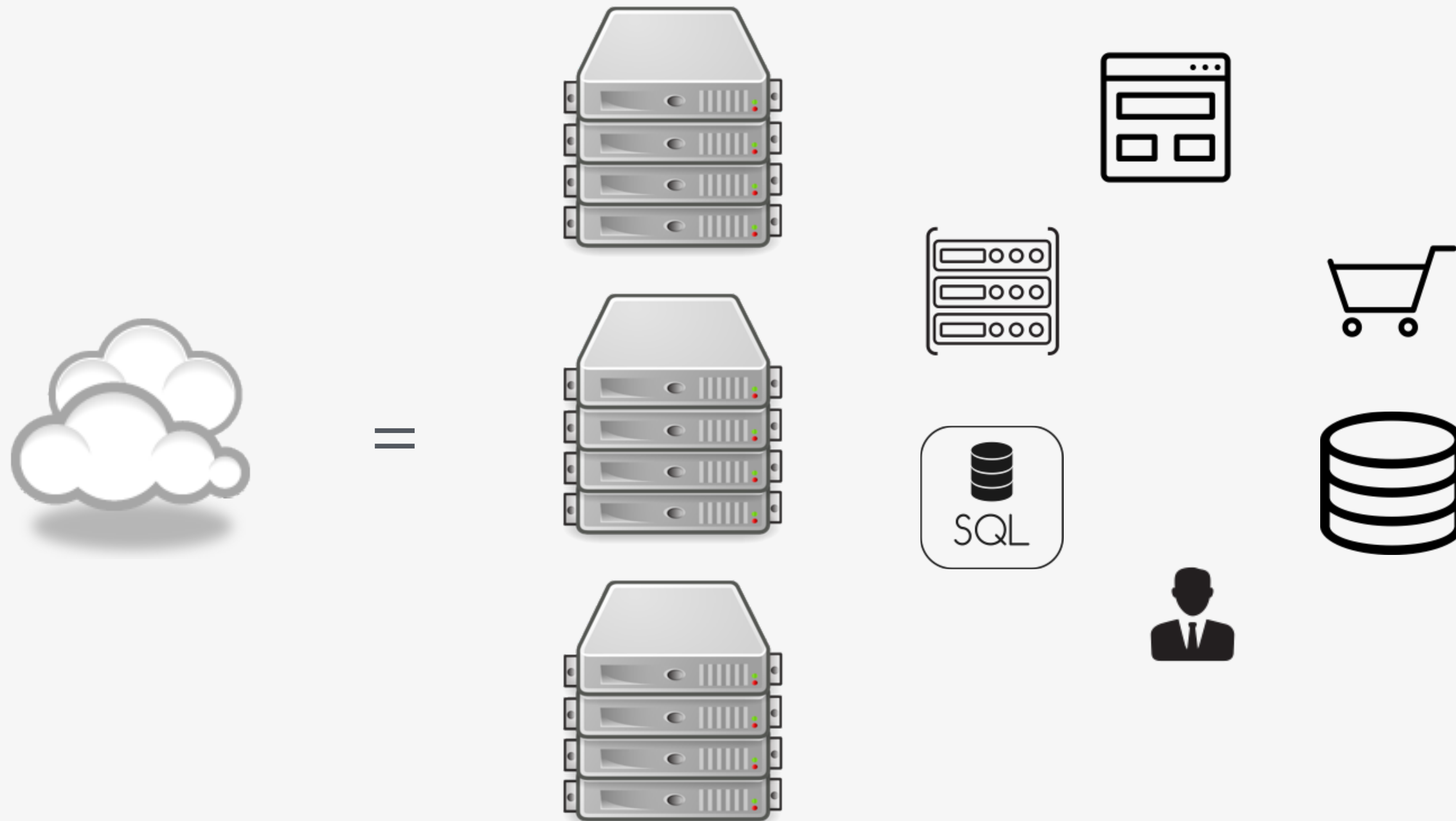
CRI's HPC cluster

- 120 compute nodes with 3 tiers of memory (3360 total cores)
- 6 accelerator nodes with NVidia GPUs and Intel Xeon Phi Coprocessors
- 350 TB scratch space
- 1.83 petabytes of storage
- Secured and HIPAA-compliant

The Cloud



So what is the Cloud?



Computing and software resources that are delivered on demand, as service.

Cloud Computing Services

SaaS



Application



Application



Application

PaaS Self-Service Interface

PaaS



Authentication, Authorization,
Access Control, Security

OS

Operative Systems



Network



Storage

...

Administration
services

- Configuration
- Deployment
- Scaling
- Users Admin
- Etc.

IaaS

IaaS Self-Service Interface

VM

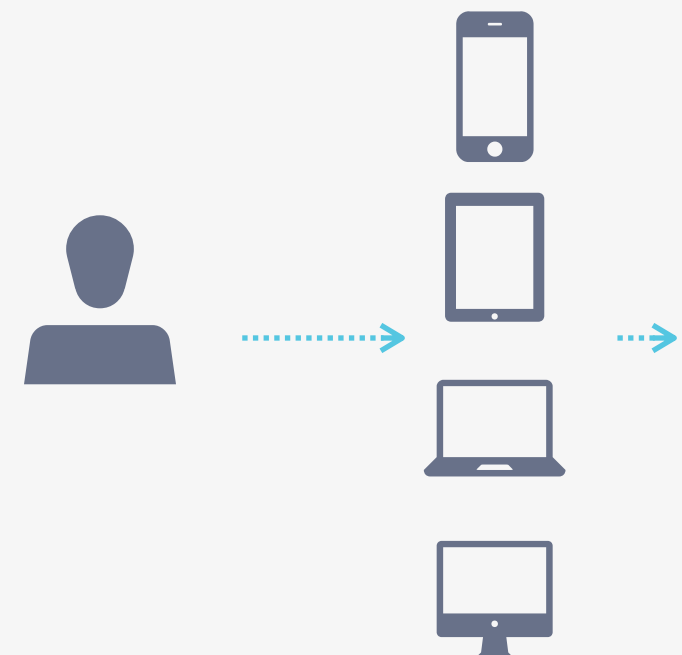
Virtual Machine

VS

Virtual Storage

VN

Virtual Network



Internet is required

Advantages and Disadvantages



Advantages

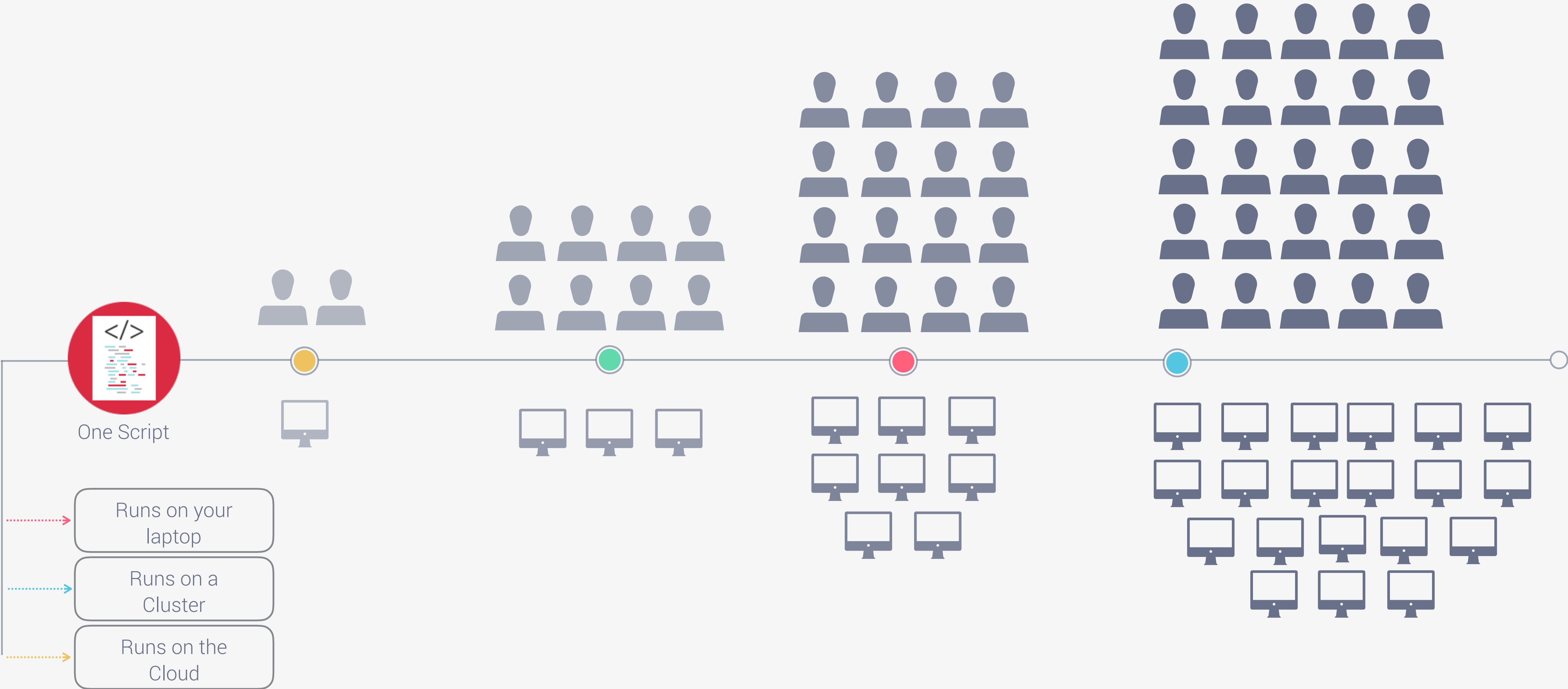
- Pay as you go: Can be less expensive compared to buying installing and maintaining your own
- Flexibility: Theoretical infinite scalability
- Easy Access: Can be used from any computer or device with an Internet connection
- Easy updates: Updates occur across the service



Disadvantages

- Security Concerns
- Terms of Service
- Privacy Policies

BigDataScript



BigDataScript

BigDataScript report: AMIA-ChIP-Seq-Pipeline.bds

Script file	/home/ubuntu/dev/chipseq/2016-cri-amia-workshop/pipelines/AMIA-ChIP-Seq-Pipeline.bds
Program ID	AMIA-ChIP-Seq-Pipeline.bds.20161026_191451_414
Start time	2016-10-26 19:14:51
Run time	00:01:53.305
Tasks executed	25
Tasks failed	0
Tasks failed names	
Arguments*	[-configFile, /home/ubuntu/dev/chipseq/2016-cri-amia-workshop/pipelines/config/chipseq.cfg, -contrastFiles, /home/ubuntu/dev/chipseq/2016-cri-amia-workshop/pipelines/config/contrast.Ab1.input.cfg, /home/ubuntu/dev/chipseq/2016-cri-amia-workshop/pipelines/config/contrast.Ab1.lgG.cfg, -runDir, /home/ubuntu/dev/chipseq/2016-cri-amia-workshop/data/sample_run/run, -logDir, /home/ubuntu/dev/chipseq/2016-cri-amia-workshop/data/sample_run/logs]
System*	local
Cpus*	1
Exit value	0

* Values in global scope when program finished execution.

Timeline



Command-line Reproducibility



Task details

Task	Execution	Time	Dependencies	Task program, Errors, StdOut / StdErr
Num 1	OK true	Start 2016-10-26 19:14:52	Input files /home/ubuntu/data/chipseq/subset/input_files/Ab1.subset.fq.gz	# SYS command. line 139
ID task AMIA-ChIP-Seq-Pipeline.Ab1.01-trimmomatic.line_138.id_1	Exit Code 0	End 2016-10-26 19:15:03	Output files /home/ubuntu/dev/chipseq/2016-cri-amia-workshop/data/sample_run/run/01-trimmomatic/Ab1/Ab1.subset.trimmed.fq.gz	java -Xmx20 -jar /home/ubuntu/software/trimmomatic-0.36/trimmomatic-0.36.jar SE -phred33 /home/ubuntu/data/chipseq/subset/input_files/Ab1.subset.ILLUMINACLP:/home/ubuntu/software/trimmomatic-0.36/adapters/TruSeq3-SE.fa:2:30:10 SLIDINGWINDOW:4:15 MINLEN:36 > /home/ubuntu/dev/chipseq/2016-cri-amia-workshop/data/sample_run/run/01-trimmomatic/Ab1/Ab1.subset.trimmed.fq.gz
Name Ab1.01-trimmomatic	Retries	State FINISHED	Dependencies	
Thread thread_Root	State FINISHED	Elapsed 00:00:11		
PID 31953	Dep. OK	Timeout 1 day		
	Cpus 2	Wall Timeout 1 day		
	Mem 2.0 GB			
Num 2	OK true	Start 2016-10-26 19:15:03	Input files /home/ubuntu/dev/chipseq/2016-cri-amia-workshop/data/sample_run/run/01-trimmomatic/Ab1/Ab1.subset.trimmed.fq.gz	# SYS command. line 170
ID task AMIA-ChIP-Seq-Pipeline.Ab1.02-bwa.sort.line_169.id_2	Exit Code 0	End 2016-10-26 19:15:39	Output files /home/ubuntu/dev/chipseq/2016-cri-amia-workshop/data/sample_run/run/02-bwa/Ab1/Ab1.bwa_aln.ftt.bam	/home/ubuntu/software/bwa-0.7.15/bwa aln -t 2 -q 2 /home/ubuntu/data/reference/GRCh38.primary_assembly.chr19.genome.fa /home/ubuntu/dev/chipseq/2016-cri-amia-workshop/data/sample_run/run/01-trimmomatic/Ab1/Ab1.subset.trimmed.fq.gz
Name Ab1.02-bwa.aln	Retries	State FINISHED	Dependencies Ab1.01-trimmomatic Ab1.01-trimmomatic Ab1.01-trimmomatic Ab1.01-trimmomatic Ab1.01-trimmomatic Ab1.01-trimmomatic	/home/ubuntu/software/bwa-0.7.15/bwa samse -t 800 /home/ubuntu/dev/chipseq/2016-cri-amia-workshop/data/sample_run/run/02-bwa/Ab1/Ab1.bwa_aln.ftt.bam /dev/stdin > /home/ubuntu/dev/chipseq/2016-cri-amia-workshop/data/sample_run/run/02-bwa/Ab1/Ab1.bwa_aln.ftt.bam
Thread thread_Root	State FINISHED	Elapsed 00:00:35		
PID 32096	Dep. OK	Timeout 1 day		
	Cpus 2	Wall Timeout 1 day		
	Mem 6.0 GB			
Num 3	OK true	Start 2016-10-26 19:15:39	Input files /home/ubuntu/dev/chipseq/2016-cri-amia-workshop/data/sample_run/run/02-bwa/Ab1/Ab1.bwa_aln.ftt.bam	# SYS command. line 185
ID task AMIA-ChIP-Seq-Pipeline.Ab1.02-bwa.sort.line_184.id_3	Exit Code 0	End 2016-10-26 19:15:44	Output files /home/ubuntu/dev/chipseq/2016-cri-amia-workshop/data/sample_run/run/02-bwa/Ab1/Ab1.bwa_aln.ftt.sort.bam	/home/ubuntu/software/sambamba-0.6.4/sambamba sort -m 20M --tmpdir=/home/ubuntu/dev/chipseq/2016-cri-amia-workshop/data/sample_run/run/02-bwa/Ab1/Ab1.bwa_aln.ftt.sort.bam
Name Ab1.02-bwa.sort	Retries	State FINISHED	Dependencies Ab1.02-bwa.aln Ab1.02-bwa.aln Ab1.02-bwa.aln Ab1.02-bwa.aln Ab1.02-bwa.aln Ab1.02-bwa.aln	# SYS command. line 186
Thread thread_Root	State FINISHED	Elapsed 00:00:05		rm -rf /home/ubuntu/dev/chipseq/2016-cri-amia-workshop/data/sample_run/run/02-bwa/Ab1/Ab1.bwa_aln.ftt.sort.bam
PID 428	Dep. OK	Timeout 1 day		
	Cpus 2	Wall Timeout 1 day		
	Mem 4.0 GB			

HTML Report



CRI experience

RNA-seq & Clinical Data Integration

NATURE | LETTER

日本語要約

Melanoma-intrinsic β -catenin signalling prevents anti-tumour immunity

Stefani Spranger, Riyue Bao & Thomas F. Gajewski

Affiliations | Contributions | Corresponding author

Nature 523, 231–235 (09 July 2015) | doi:10.1038/nature14404

Received 15 August 2014 | Accepted 05 March 2015 | Published online 11 May 2015

Published OnlineFirst September 16, 2016; DOI: 10.1158/2326-6066.CIR-16-0087

Cancer Immunology Miniatures

Cancer Immunology Research

Clinical Response of a Patient to Anti-PD-1 Immunotherapy and the Immune Landscape of Testicular Germ Cell Tumors

Shalin Shah¹, James E. Ward^{1,2}, Riyue Bao², Curtis R. Hall^{1,2}, Bruce E. Brockstein^{1,2}, and Jason J. Luke²

PLOS ONE

RESEARCH ARTICLE

ExScalibur: A High-Performance Cloud-Enabled Suite for Whole Exome Germline and Somatic Mutation Identification

Riyue Bao¹*, Kyle Hernandez¹*, Lei Huang¹, Wenjun Kang¹, Elizabeth Bartom¹*, Kenan Onel², Samuel Volchenbom^{1,2,3}*, Jorge Andrade¹*

¹ Center for Research Informatics, The University of Chicago, Chicago, Illinois, United States of America, ² Department of Pediatrics, The University of Chicago, Chicago, Illinois, United States of America, ³ Computation Institute, The University of Chicago, Chicago, Illinois, United States of America

* These authors contributed equally to this work.

Current address: Department of Biochemistry and Molecular Genetics, Northwestern University, Chicago, Illinois, United States of America

* jandrade@bsd.uchicago.edu (JA); svolchen@peds.bsd.uchicago.edu (SV)

RNAseq & ChIP-Seq Integration

Chromatin, Epigenetics, and RNA Regulation

Molecular Cancer Research

GR and ER Coactivation Alters the Expression of Differentiation Genes and Associates with Improved ER⁺ Breast Cancer Outcome

Diana C. West¹, Deng Pan¹, Eva Y. Tonsing-Carter¹, Kyle M. Hernandez², Charles F. Pierce¹, Sarah C. Styke¹, Kathleen R. Bowie¹, Tzintzuni I. Garcia², Masha Kocherginsky³, and Suzanne D. Conzen^{1,4}

Obesity A Research Journal

OBESITY SOCIETY

Explore this journal

Original Article

Glucocorticoid receptor ChIP-sequencing of subcutaneous fat reveals modulation of inflammatory pathways

Puneet Singh, Clifton O. Brock, Paul A. Volden, Kyle Hernandez, Maxwell Skor, Masha Kocherginsky, Julie E. Park, Matthew J. Brady ✉, Suzanne D. Conzen ✉

Cloud-based Pipeline Development

RNA-Seq and ChIP-Seq Tutorials: Jupyter Notebooks on Amazon Web Service (AWS) EC2 instance

<https://github.com/cribioinfo/>

[https://github.com/cribioinfo/CRI-Workshop-AMIA-2016-RNAseq/blob/master/
notebook_ext/2016-AMIA-Workshop-AWS-iPython-Guide.pdf](https://github.com/cribioinfo/CRI-Workshop-AMIA-2016-RNAseq/blob/master/notebook_ext/2016-AMIA-Workshop-AWS-iPython-Guide.pdf)