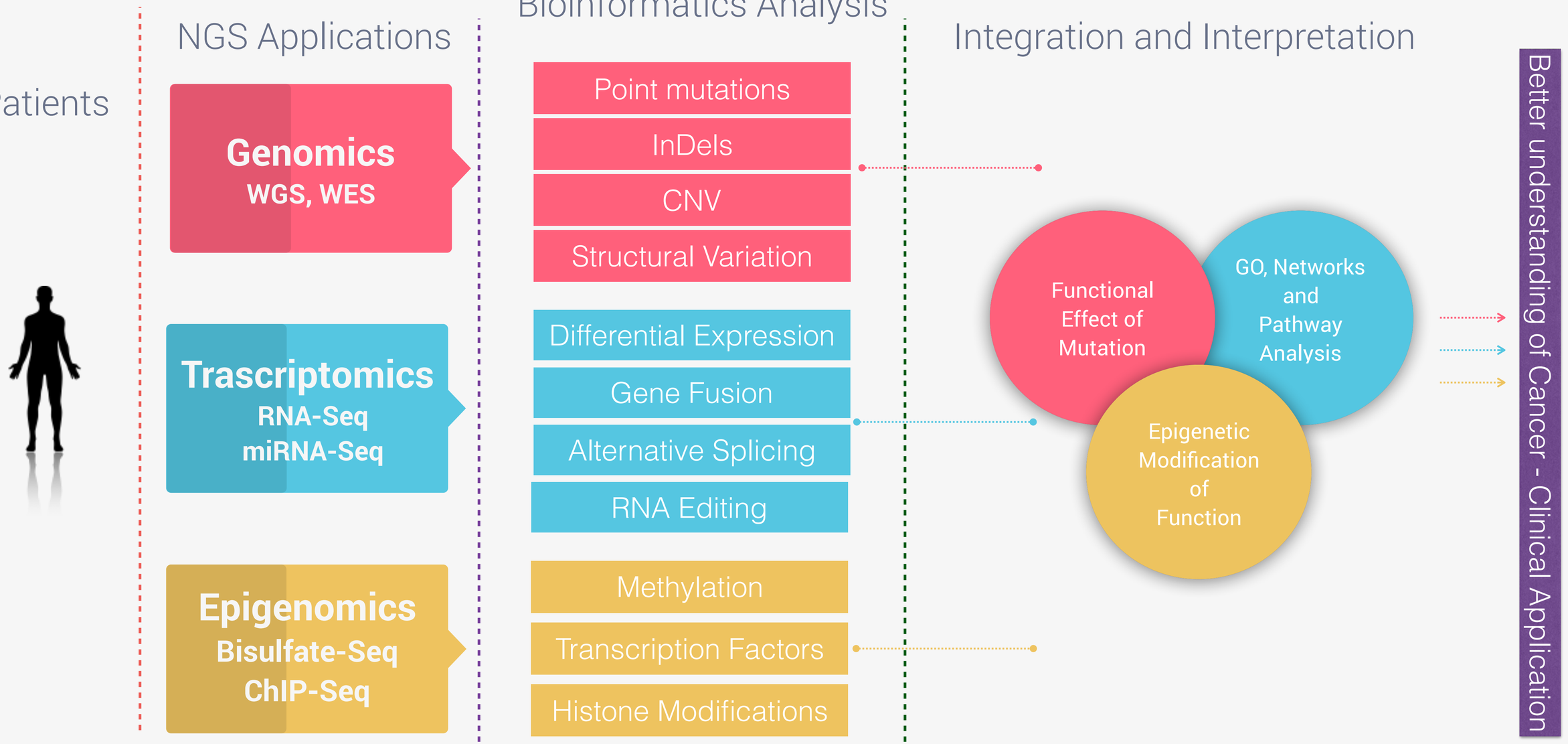
Mining Large-scale Cancer Genomics Data Using Cloud-based Bioinformatics Approaches

A very short introduction to Cloud Computing

Jorge Andrade, Ph.D.

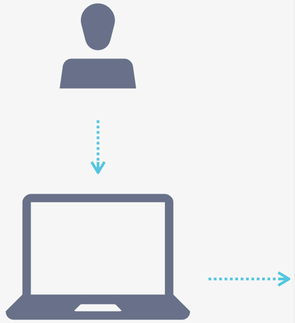
Why Cloud-based Bioinformatics

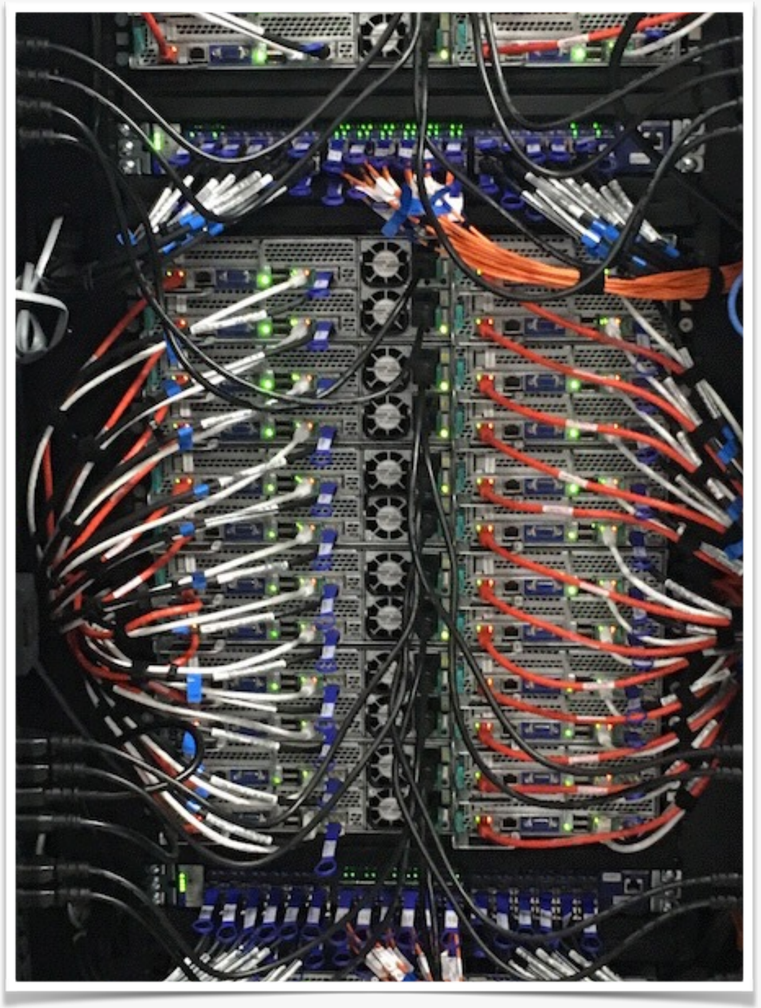


Next Generation Sequencing = Big Data

	 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	• MiniSeq System for low-throughput targeted DNA and RNA sequencing	• MiSeq System for targeted and small genome sequencing • MiSeq FGx System for forensic genomics • MiSeqDx System for molecular diagnostics	• NextSeq 500 System for everyday genomics • NextSeq 550 System for both sequencing and cytogenomic arrays	• HiSeq 3000/HiSeq 4000 Systems for production-scale genomics • HiSeq 2500 Systems for large-scale genomics	• HiSeq X Five System for production-scale whole-genome sequencing • HiSeq X Ten System for population-scale whole-genome sequencing

High Performance Computing Cluster (HPC)



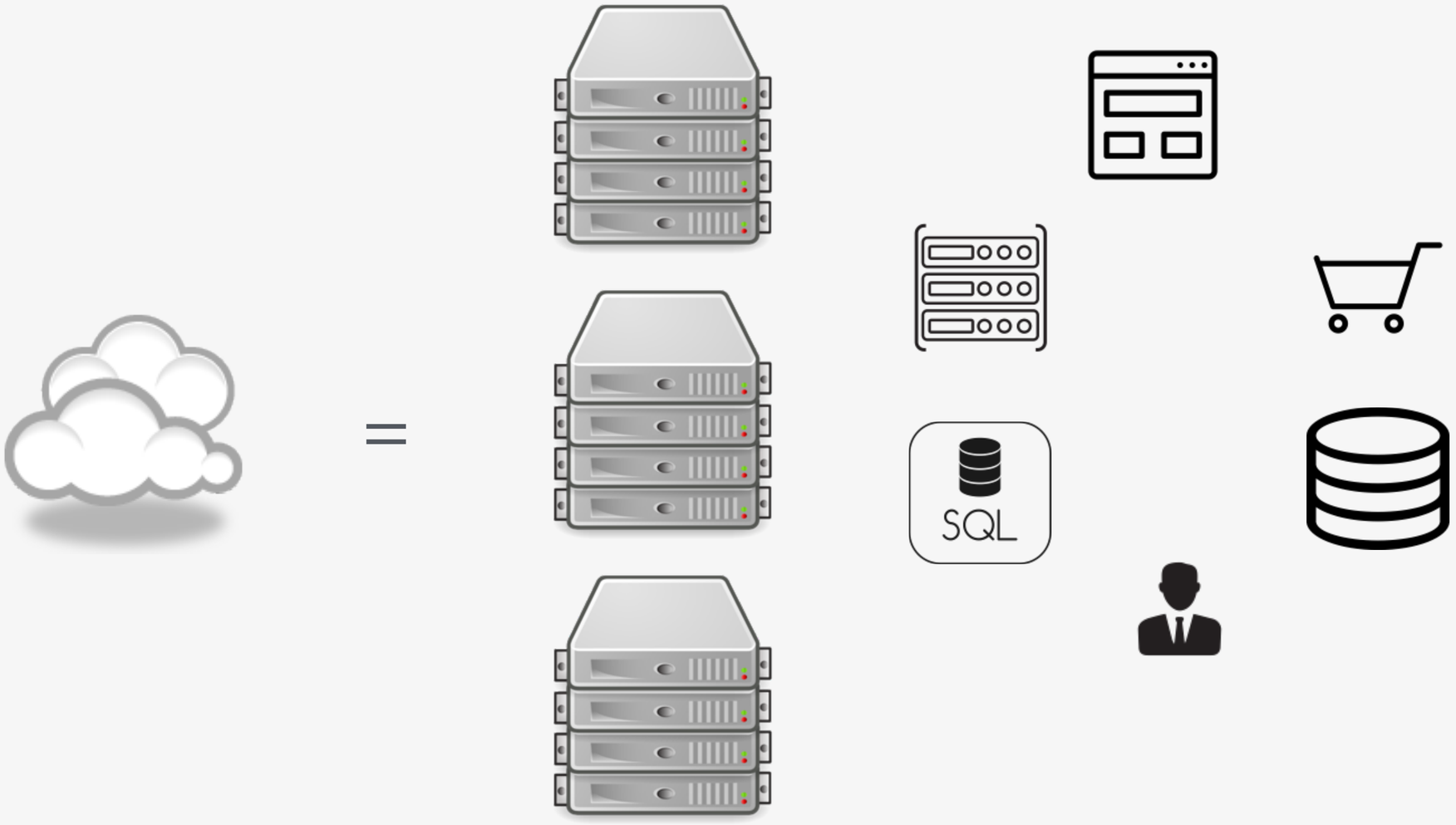


CRI's HPC cluster

- 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

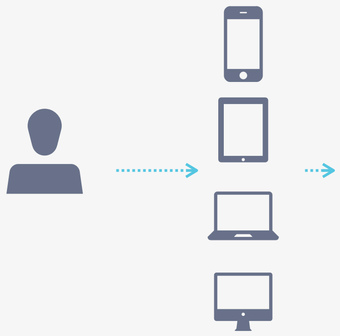
The Cloud

So what is the Cloud?



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

Cloud Computing Services



Internet is required

SaaS

Application Application Application

PaaS Self-Service Interface

PaaS

Authentication, Authorization, Access Control, Security

OS

Operative Systems

Network

Storage

...

IaaS

IaaS Self-Service Interface

VM

Virtual Machine

VS

Virtual Storage

VN

Virtual Network

Administration services

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

Advantages and Disadvantages

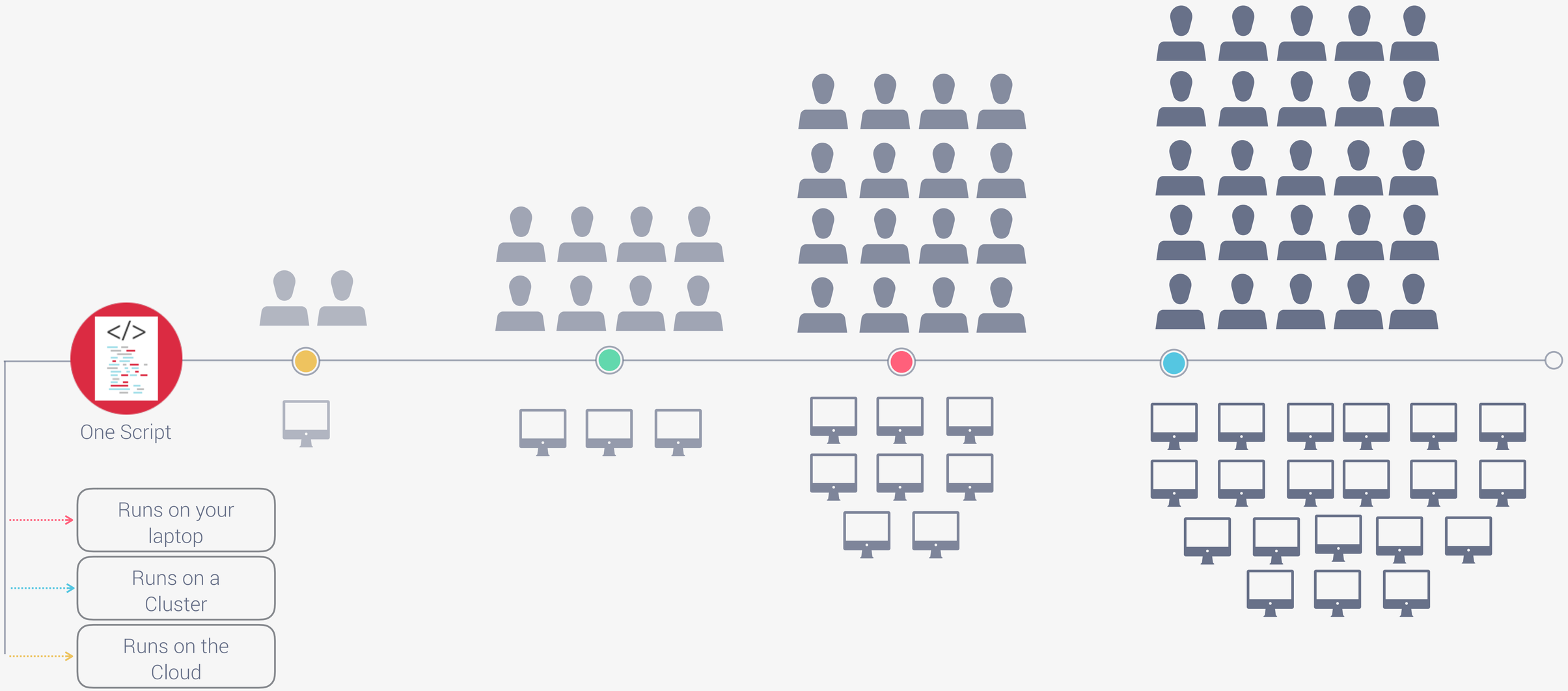
Advantages

- 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



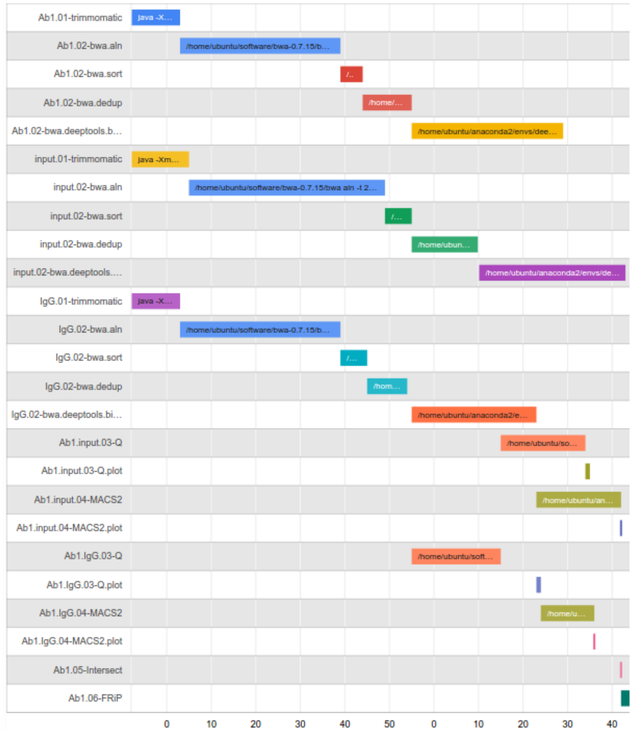
BigDataScript

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.IgG.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

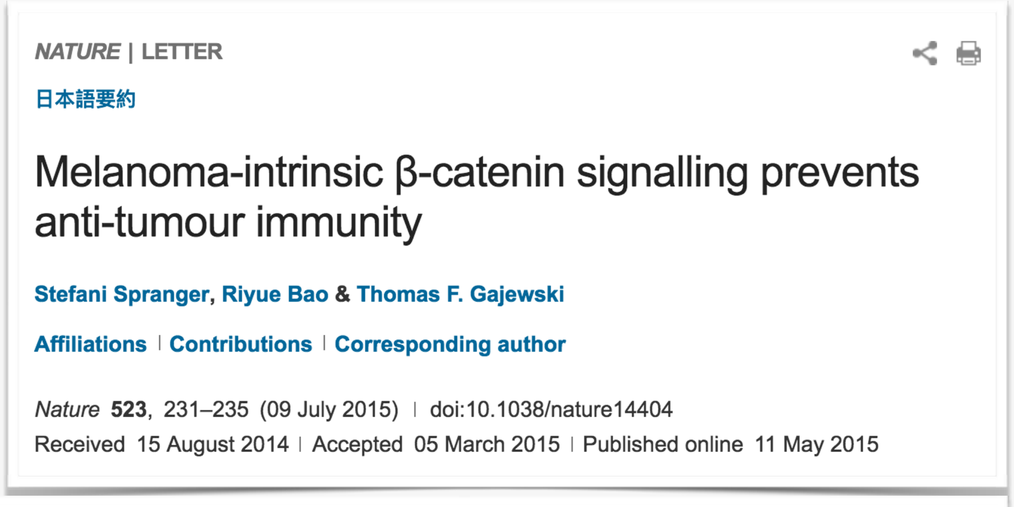
Task		Execution		Time		Dependencies		Task program, Errors, StdOut / StdErr
Num	1	**OK**	true	**Start**	2016-10-26 19:14:52	**Input files**	/home/ubuntu/data/chipseq/subset/input_files/Ab1.subset.fq.gz	
ID	task.AMIA-ChIP-Seq-Pipeline.Ab1.01_trimmomatic.line_138.id_1	**Exit Code**	0	**End**	2016-10-26 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	
Name	Ab1.01-trimmomatic	**Retries**		**Elapsed**	00:00:11	**Dependencies**		
Thread	thread_Root	**State**	FINISHED	**Timeout**	1 day			
PID	31953	**Dep.**	OK	**Wall Timeout**	1 day			
		Cpus	2					
		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	
ID	task.AMIA-ChIP-Seq-Pipeline.Ab1.02_bwa.aln.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.flt.bam	
Name	Ab1.02-bwa.aln	**Retries**		**Elapsed**	00:00:35	**Dependencies**	Ab1.01-trimmomatic Ab1.01-trimmomatic Ab1.01-trimmomatic Ab1.01-trimmomatic Ab1.01-trimmomatic Ab1.01-trimmomatic Ab1.01-trimmomatic Ab1.01-trimmomatic	
Thread	thread_Root	**State**	FINISHED	**Timeout**	1 day			
PID	32096	**Dep.**	OK	**Wall Timeout**	1 day			
		Cpus	2					
		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.flt.bam	
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.flt.sort.bam	
Name	Ab1.02-bwa.sort	**Retries**		**Elapsed**	00:00:05	**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 Ab1.02-bwa.aln Ab1.02-bwa.aln	
Thread	thread_Root	**State**	FINISHED	**Timeout**	1 day			
PID	428	**Dep.**	OK	**Wall Timeout**	1 day			
		Cpus	2					
		Mem	4.0 GB					

HTML Report

CRI experience

RNA-seq & Clinical Data Integration

RNAseq & ChIP-Seq Integration



NATURE | LETTER

日本語要約

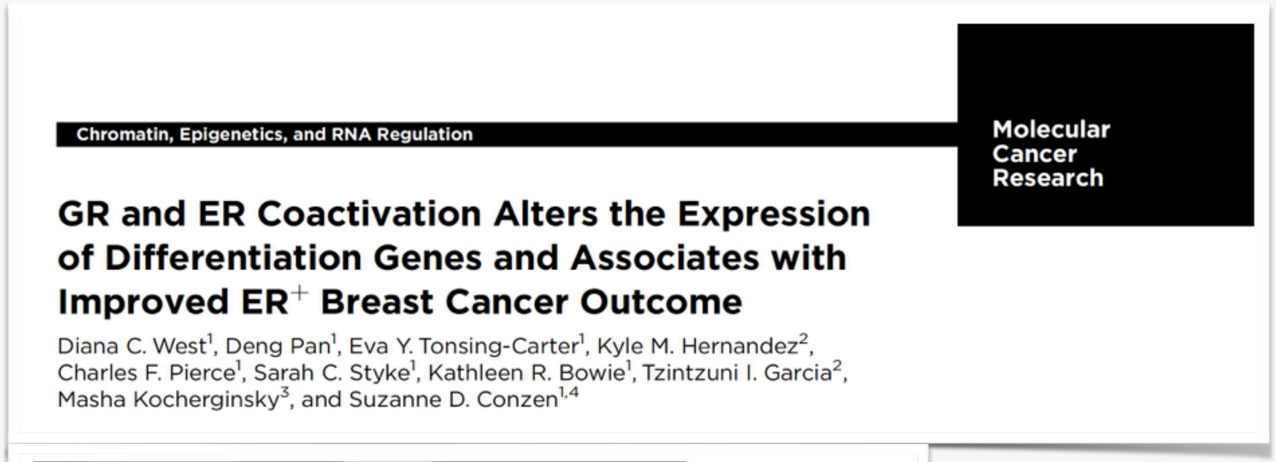
Melanoma-intrinsic β-catenin signalling prevents anti-tumour immunity

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

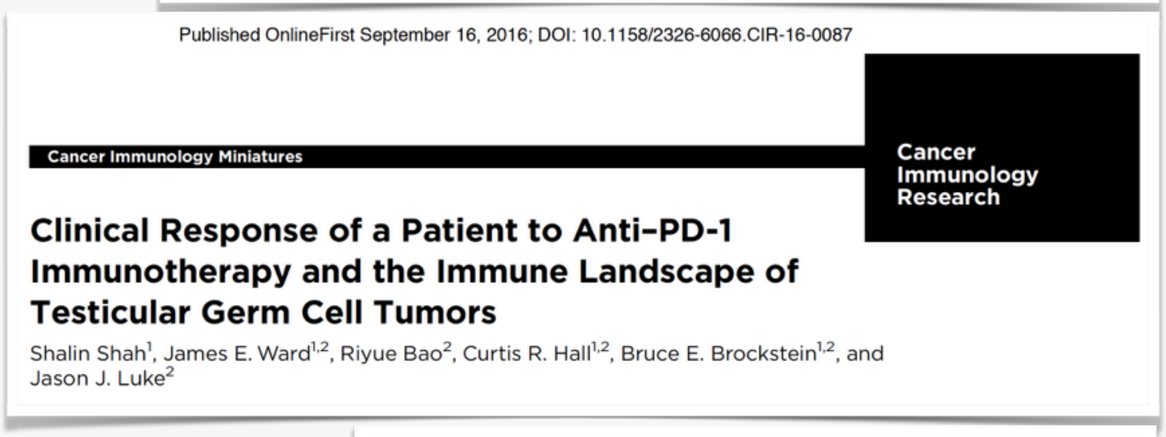


Chromatin, Epigenetics, and RNA Regulation

Molecular Cancer Research

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

Diana C. West[1], Deng Pan[1], Eva Y. Tonsing-Carter[1], Kyle M. Hernandez[2], Charles F. Pierce[1], Sarah C. Styke[1], Kathleen R. Bowie[1], Tzintzuni I. Garcia[2], Masha Kocherginsky[3], and Suzanne D. Conzen[1,4]



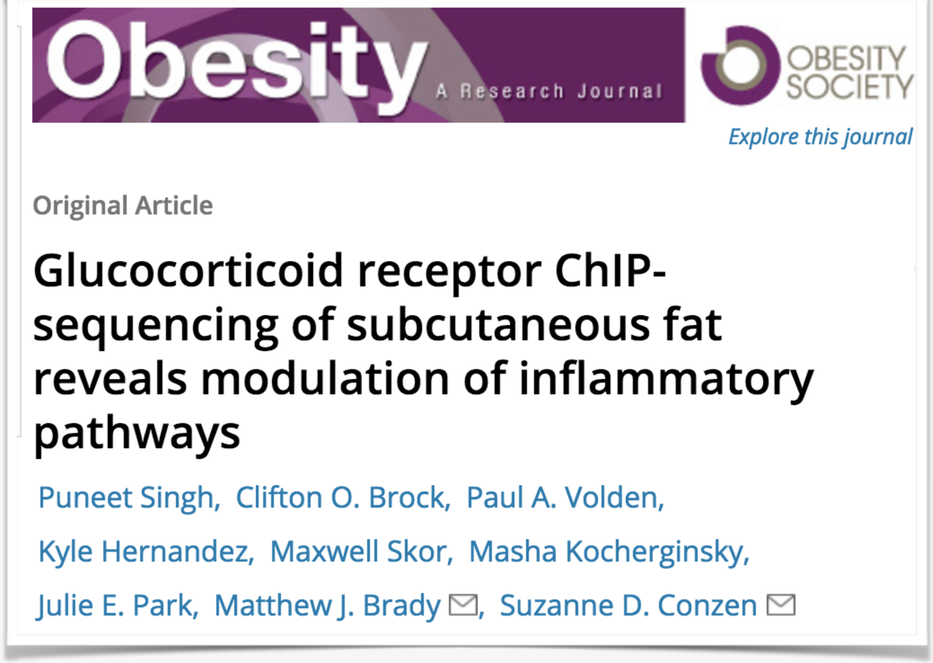
Published OnlineFirst September 16, 2016; DOI: 10.1158/2326-6066.CIR-16-0087

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 Shah[1], James E. Ward[1,2], Riyue Bao[2], Curtis R. Hall[1,2], Bruce E. Brockstein[1,2], and Jason J. Luke[2]



Obesity A Research Journal

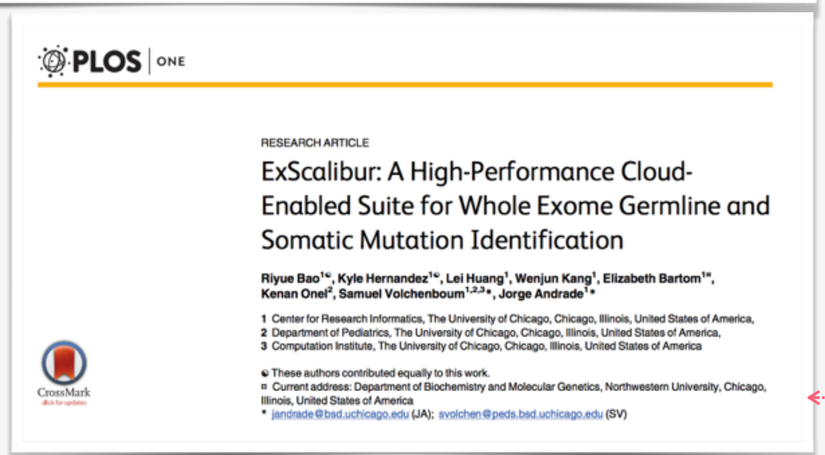
OBESITY SOCIETY

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



PLOS ONE

RESEARCH ARTICLE

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

Riyue Bao[1☯], Kyle Hernandez[1☯], Lei Huang[1], Wenjun Kang[1], Elizabeth Bartom[1¤], Kenan Onel[2], Samuel Volchenboum[1,2,3]*, Jorge Andrade[1]*

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)

Cloud-based Pipeline Development

THE UNIVERSITY OF CHICAGO
CENTER FOR RESEARCH INFORMATICS

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