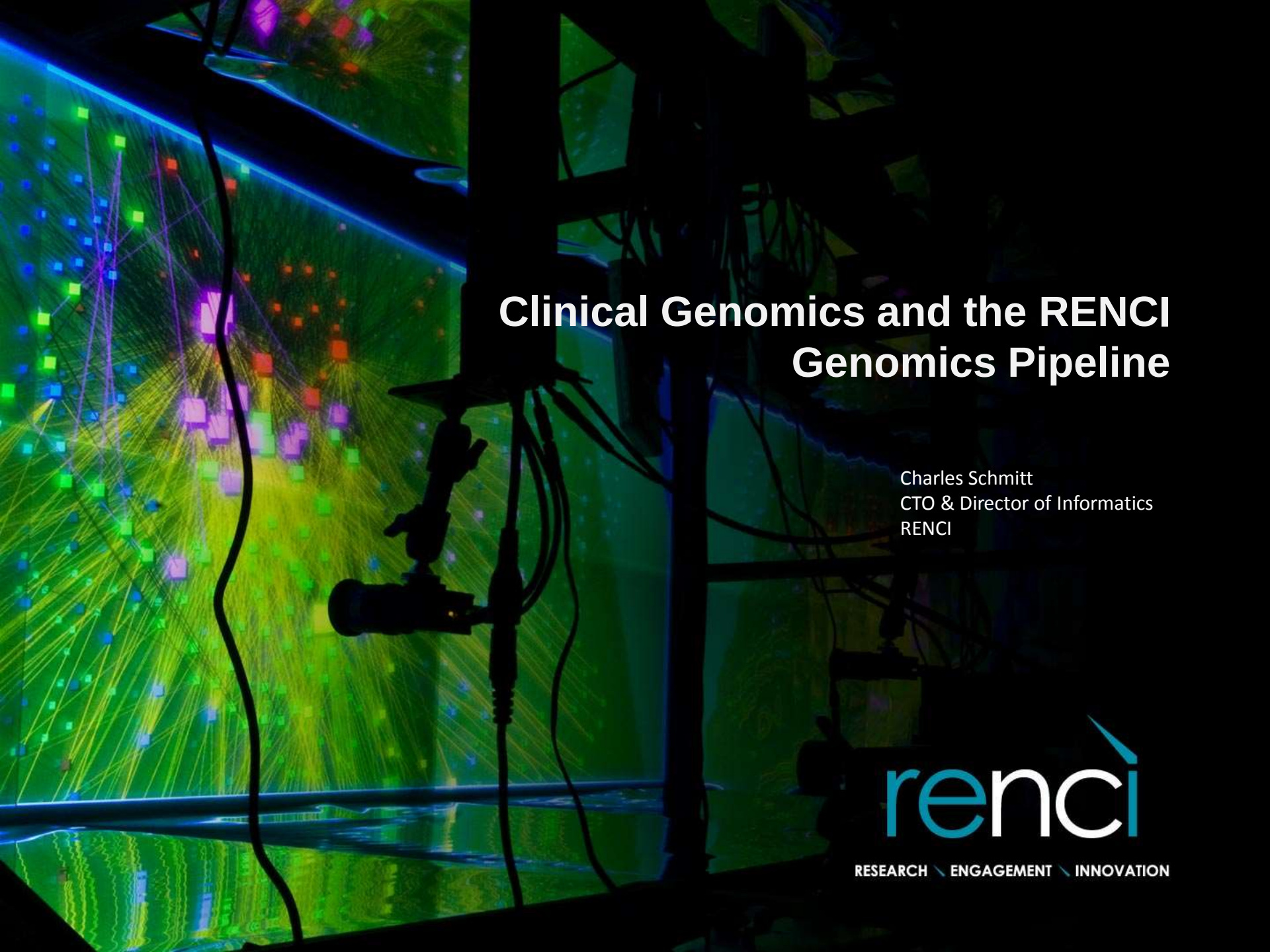


Clinical Genomics and the RENCI Genomics Pipeline

Charles Schmitt
CTO & Director of Informatics
RENCI

renci

RESEARCH \ ENGAGEMENT \ INNOVATION

Acknowledgements

- *Biomedical Informatics team at RENCI led by Dr. Kirk Wilhelmsen*
- *DICE team at UNC and UCSD*
- *Networking team at RENCI and Duke*
- *Data sciences team at RENCI*
- *UNC Dept of Genetics, Research Computing, Lineberger Comprehensive Cancer Center, NC Tracs Institute, Center for Bioinformatics, Institute for Pharmacogenetics and Personalized Treatment*
- Presented work funded in part by multiple grants from NIH, NSF, and internal support

renci

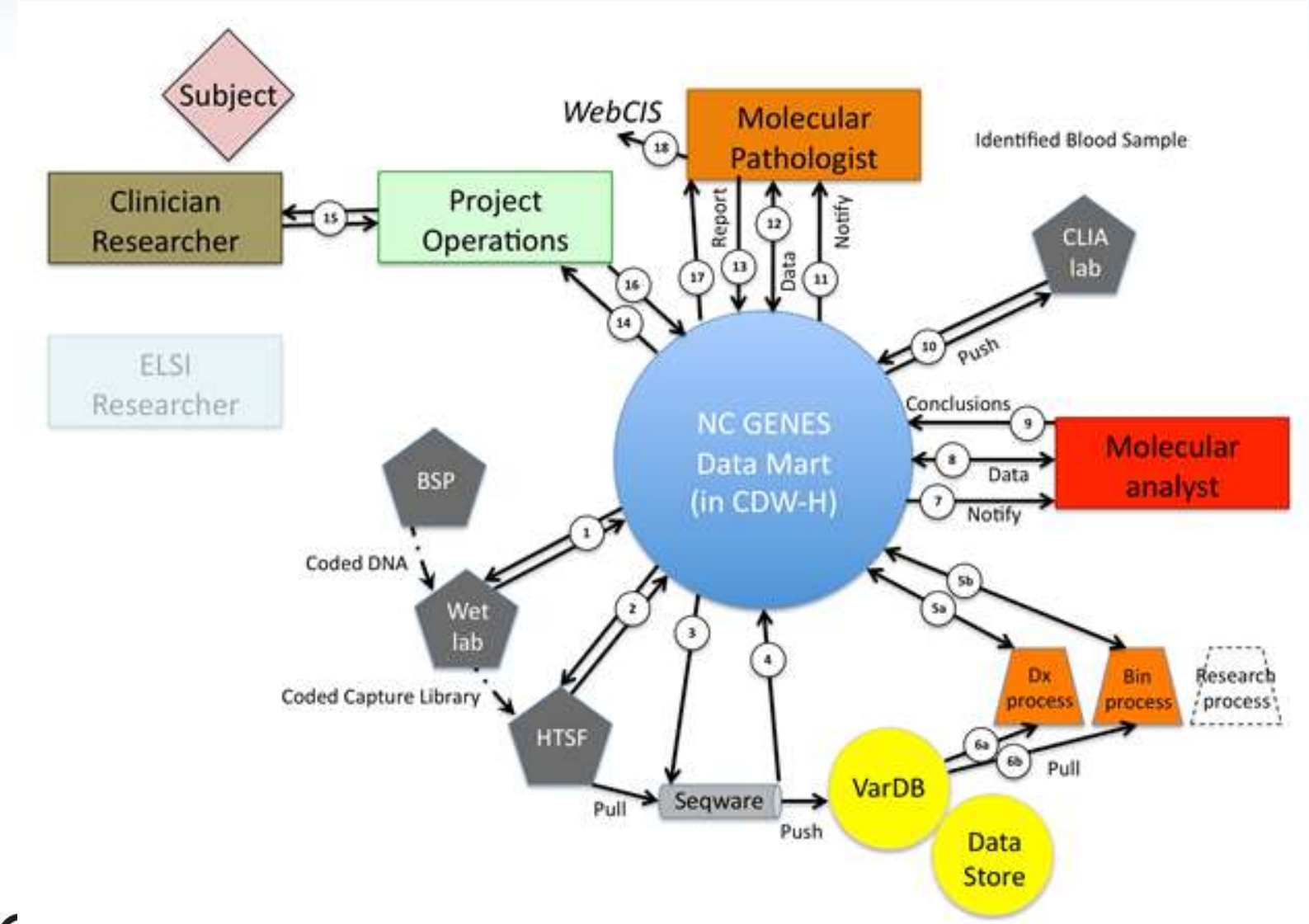
RENCI-UNC Genomics

Disease Type	Common	Mendelian
Variant Effect	Risk	Causal
Genomic Domain	Basic Research	Clinical Application
RENCI/UNC NIH Funded Projects	NIDA Sequencing[1] ~5,500 samples Alzheimer's Analysis	NCGENES[2] ~2000 samples NC NEXUS[3] ~400 samples ClinVar[4]

97,197 Sample instances, 239,915 workflow runs, 40 pipelines (20 active) across these and other projects in support of UNC HTSF

(1) National Institute on Drug Abuse– funded NIDASeq, "Deep Sequencing Studies for Cannabis and Stimulant Dependence" (Dr. Kirk Wilhelmsen, PI),
(2) National Human Genome Resource Institute–funded NCGENES, "North Carolina Clinical Genomic Evaluation by NextGen Exome Sequencing" (Dr. James Evans, PI)
(3) National Institute of Child Health and Development–funded NC Nexus, "North Carolina Newborn Exome Sequencing and Newborn Screening Disorders" (Dr. Cynthia Powell, PI)
(4) National Human Genome Resource Institute–funded ClinGen (Dr. Jonathan Berg PI)

Genomics Pipeline – High Level



Clinical Side

UNC SCHOOL OF MEDICINE

Good day, Chris Bizon.
Your roles and studies:
NCGENES Binning Technicians,NCGENES Analysis Pipeline Technicians,NCGENES Administrators, HRC Study Ophthalmology Study,NCGENES Study.

genetics NCGENES WorkFlow Manager

Home | Workflows | Administration | Participants | Analysis | CLIA | ELSI | GenPhenAnnot | Help | Log out

Analysis results

For participant: NCG_00014
Diagnostic type: Cardiomyopathy

Donor selection: NCG_00014
Dx filter selection: No filter
Gene filter selection: No filter
Please select an analysis result type: Diagnostic | Load | View coverage data by gene | View coverage data by exon | Download Bam files

Class	Calculated Class	HGNC gene	Type	HGVS genome	Variant Effect name	Max allele freq	Tag	ACC num	Depth	QD	Read pos rank sum	Frac reads with alt	Allele	Strand score	Ref depth	Alt depth	Genotype qual	rs ID	Tier	Inheritance	Dx type	Notes
VUS	A	ACADVL	snp	NC_000017.10:g.7127757G>A	missense	0.012199	DM	CM980187	58	35.34		0	1	0	0	98	99	rs2230180	15	AR	Cardiomyopathy	1 notes. View/Add
VUS	C	ALMS1	snp	NC_000002.11:g.73027601A>G	missense	0			69	12.56	0.315	0	0	0	39	30	99		15	AR	Cardiomyopathy	1 notes. View/Add
LB	D	ANKRD1	del	NC_000010.10:g.92672409_92672412delTAAA	UTR-3	0.011046			38	23.26	0.472		0	0	18	20	99		1	AD	Cardiomyopathy	1 notes. View/Add
LB	D	BAG3	del	NC_000010.10:g.121439814_121439817delNNCAA	intron	0.00588077			26	21.46	0.249		0	0	14	12	99		1	AD	Cardiomyopathy	1 notes. View/Add
VUS	D	VCL	snp	NC_000010.10:g.75834530C>T	synonymous	0.002762			43	15.29	-0.109	0	0	3.513	21	22	99	rs56284452	1	AD	Cardiomyopathy	1 notes. View/Add
LB	D	MYL2	snp	NC_000012.11:g.111353634A>G	intron	0.00813			39	17.16	-0.688	0	1	0	19	20	99	rs2231259	1	AD	Cardiomyopathy	1 notes. View/Add
LB	D	MYL2	snp	NC_000012.11:g.111358229C>T	intron	0.00813			33	18.36	-0.021	0	0	2.88	19	14	99	rs116376364	1	AD	Cardiomyopathy	1 notes. View/Add
LB	D	MYL3	snp	NC_000012.11:g.120555877T>C	intron	0.006898			14	8.2	0.133	0	0	0	9	5	99	rs11832528	1	AD	Cardiomyopathy	1 notes. View/Add
LB	D	DSG2	snp	NC_000018.9:g.29116284G>A	missense	0.004715			114	12.19	1.163	0	0	2.446	62	52	99	rs2230235	1	AD	Cardiomyopathy	1 notes. View/Add
LB	D	TTN	snp	NC_000002.11:g.179598965T>G	UTR-5	0.034553			51	17.75	-0.094	0	1	0	22	29	99	rs72619787	1	AD	Cardiomyopathy	1 notes. View/Add

Class	Transcript	Loc type	Strand	Intron exon dist	Variant Effect	HGVS cds	HGVS transcript	HGVS protein
A	NM_001033859.1	exon	+	-6	missense	NM_001033859.1:c.1534G>A	NM_001033859.1:g.16850>A	NP_001028031.1:p.Val512Ile
A	NM_000018.2	exon	+	-6	missense	NM_000018.2:c.1600G>A	NM_000018.2:g.17510>A	NP_000009.1:p.Val534Ile

Go to page: 1 Show rows: 10 1-10 of 110

Clinical Side

Variant Effect:

Class	Calculated Class	HGNC gene	Type	HGVS genomic	Variant Effect Rank	Max allele freq
VUS	A	ACADVL	snp	NC_000017.10:g.7127707G>A	missense	0.012195

Known Disease:

Tag	ACC num
DM	CM990107

Variant Notes

HGVS Genomic: NC_000017.10:g.7127707G>A

Gene: ACADVL

Date	Note	Author
07/01/2012 15:32:05	18 of 37 children (67%) had severe dilated cardiomyopathy whith ACADVL mutations (1999, PubMed 10077518). In 7 children only one mutation was found after sequencing all exons. No info on ACADVL E1685K mutation (Glu to Lys)	Cecile Skrzynia

Other Annotation:

rs ID	Tier	Inheritance	Dx type	Notes
rs2230180	1S	AR	Cardiomyopathy	1 notes. (View/Add)

Transcript-Based Information:

Class	transcr	Loc type	strand	Intron exon dist	Variant Effect	HGVS cds	HGVS transcript	HGVS protein
A	NM_001033859.1	exon	+	-6	missense	NM_001033859.1:c.1534G>A	NM_001033859.1:g.1685G>A	NP_001029031.1:p.Glu512Lys
A	NM_000018.2	exon	+	-6	missense	NM_000018.2:c.1600G>A	NM_000018.2:g.1751G>A	NP_000009.1:p.Glu534Lys

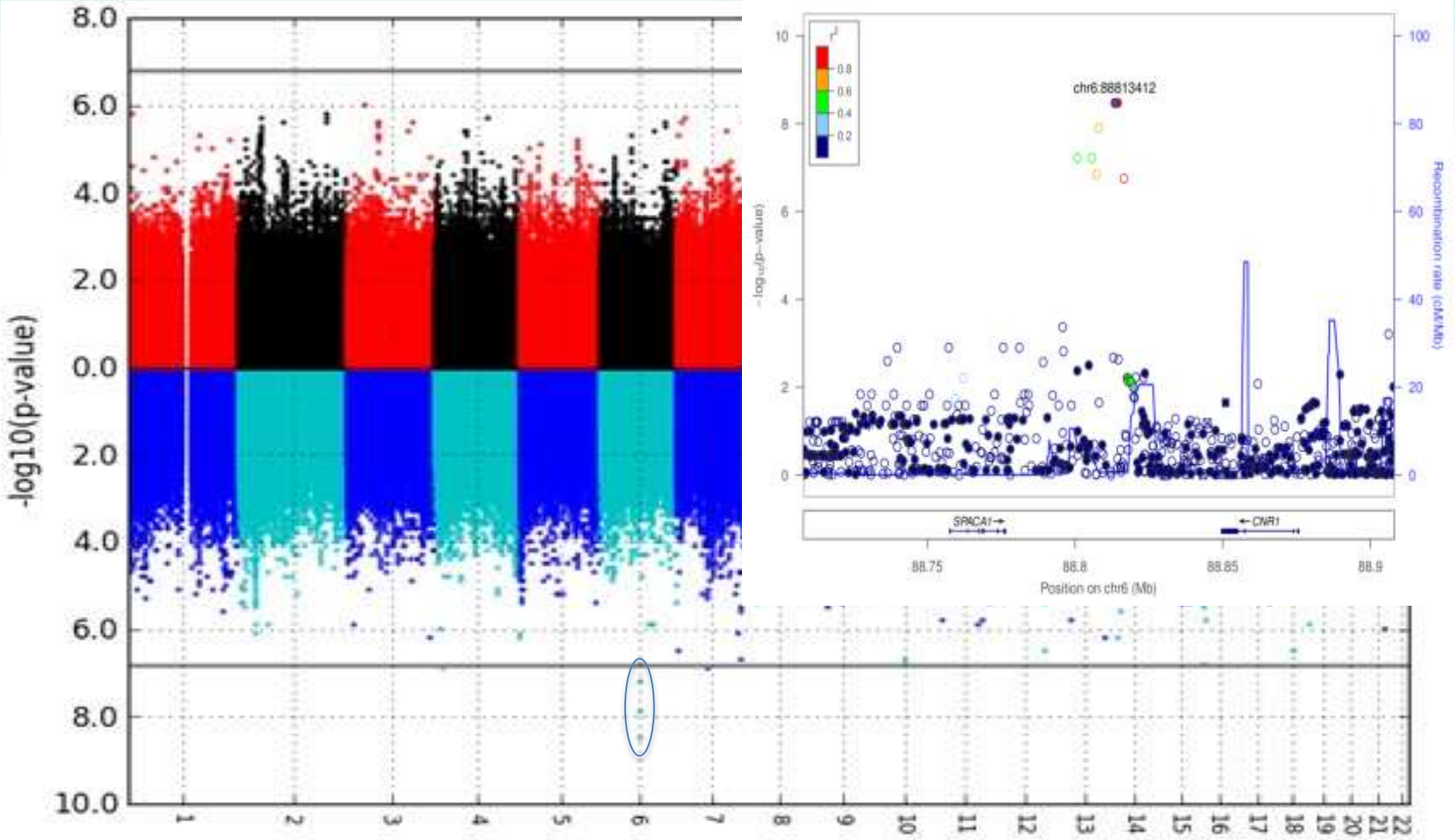
Journal of Neurology
March 2014, Volume 261, Issue 3, pp 622-624

Date: 08 Feb 2014

GCH1 heterozygous mutation identified by whole-exome sequencing as a treatable condition in a patient presenting with progressive spastic paraplegia

Zheng Fan, Robert Greenwood, Ana C. G. Felix, Yael Shiloh-Malawsky, Michael Tennison, Myra Roche, Kristy Crooks, Karen Weck, Kirk Wilhelmsen, Jonathan Berg, James Evans

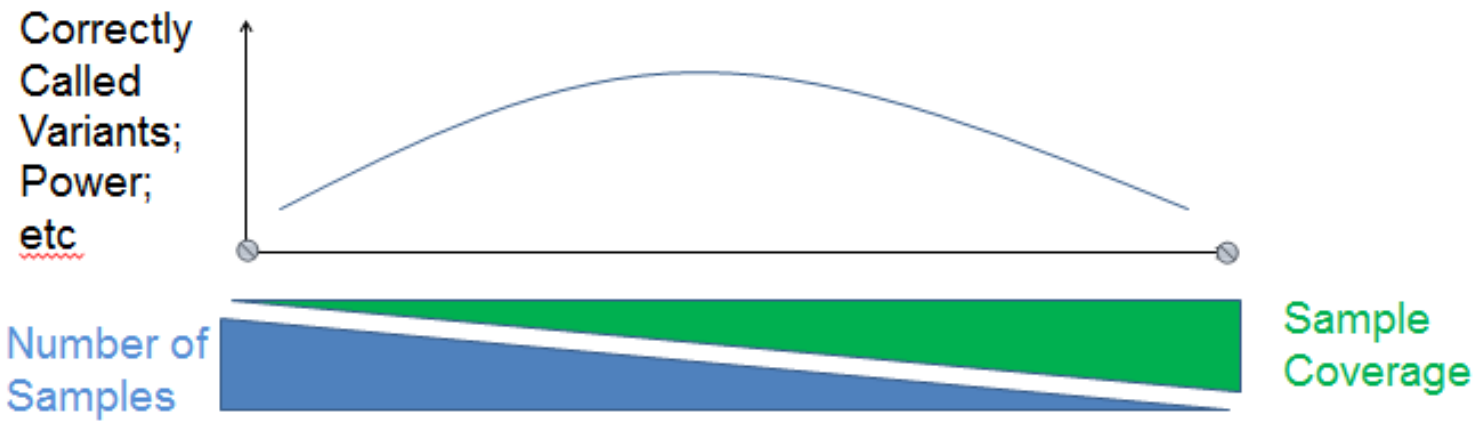
- Patient was “living” with Dx spastic paraplegia
  - Mostly wheelchair bound for 15 years
- Genetically as Dopa-responsive dystonia
  - Rx with dopamine replacement >> Clinically normal in weeks
- ~40% of patients now have an answer for something that had gone undiagnosed



Genotype association with LAC-Lifetime Amphetamine or Cocaine use

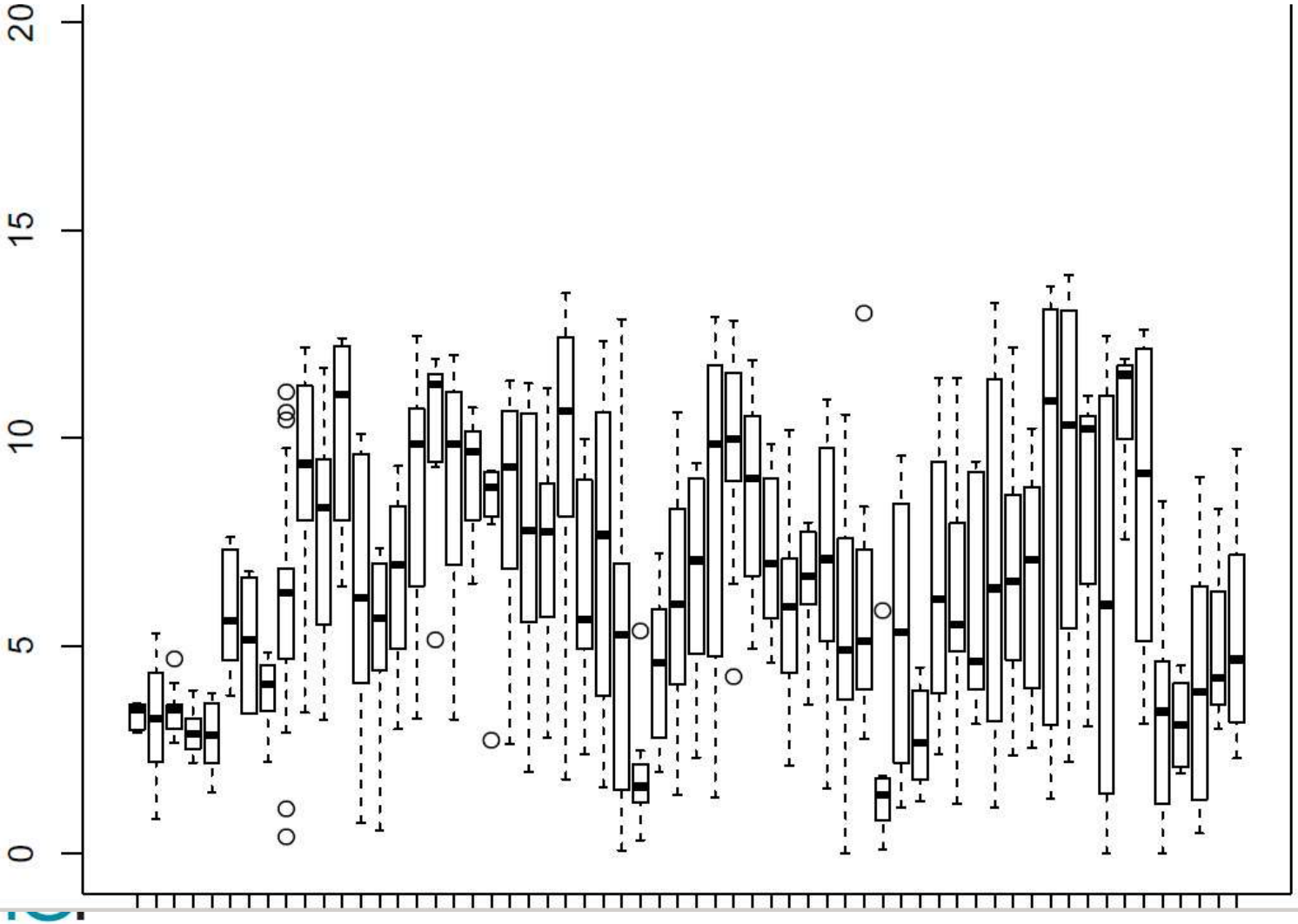
Low Coverage WGS

For a given sequencing budget, a balance must be struck between sequencing many samples at low coverage and sequencing a few samples at high coverage.

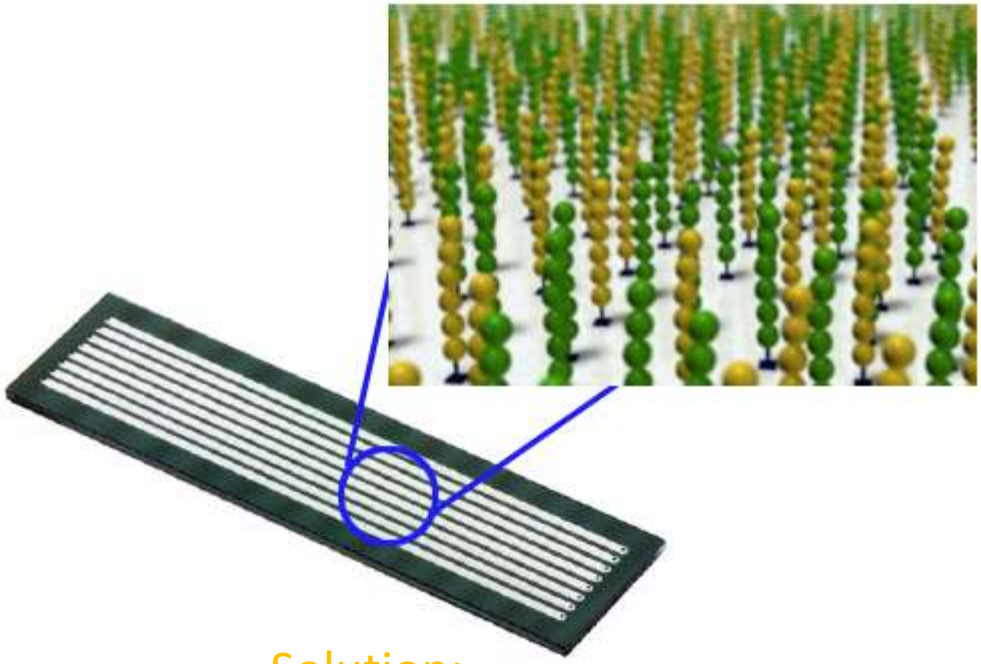


renci

Sequence Coverage Variation



Optimization of Sequence Yield



- Cluster density can be empirically optimized based on library properties
- Cannot be predicted for physical properties of a library

Solution:

Make Deep Multiplex Library

Optimize stoichiometry and cluster density

Sequence many lanes of pooled library at optimum cluster density

->13X/lane

renci

Imputing genotypes

ATCGATCGATCAG- reference

ATCGGTCGATCAG– patient
TCGGTNNNTCAG
GTCGGTCAG
ATCGGTCGGTCA
ATCGGTCGGTC

Population Evidence

ATCGGTCGGTCAG- patient 2
ATCGGTCGGTCAG– patient 3
ATCGGTCGGTCAG– patient 4
ATCGGTCGGTCAG– patient 5
ATCGATCGATCAG– patient 6
ATCGATCGATCAG– patient 7
ATCGATCGATCAG– patient 8

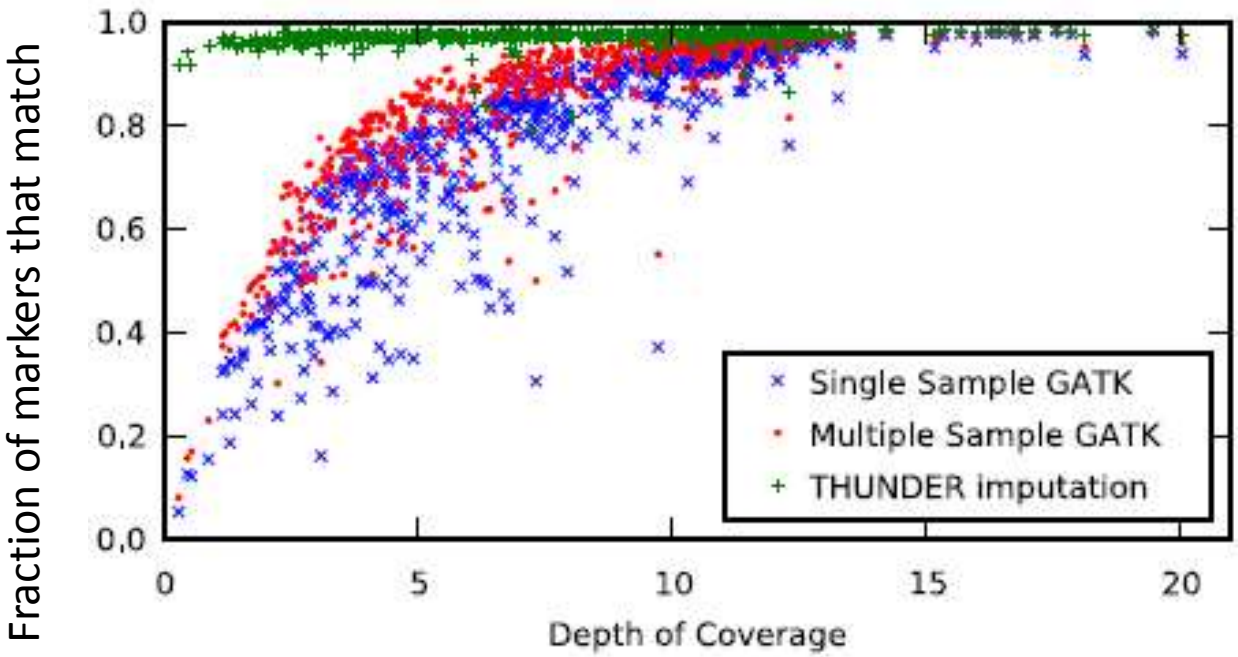
Hidden Markov Models for cross-genome statistical correlations

Imputation on 708 samples takes over 200,000 CPU hours to complete, or 22 CPU years

Scaling Time: O(N^3)

Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. Low-coverage sequencing: Implications for design of complex trait association studies. *Genome Res.* 2011 Jun;21(6):940-51

Test of the Imputation Approach for Low Coverage



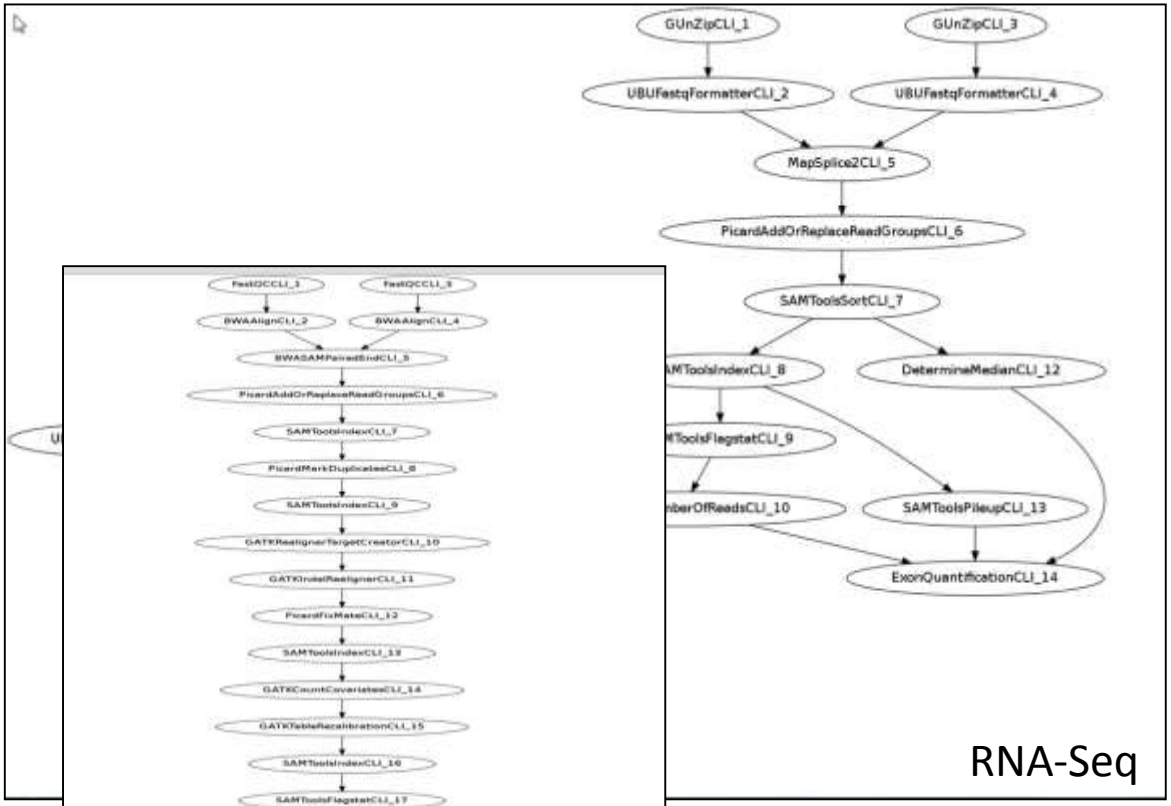
- Based on variant concordance, 5X LD-aware calling is equivalent to 15X multi-sample calling.
- LD-aware calling allows a 3X increase in the sample size for the same sequencing budget.

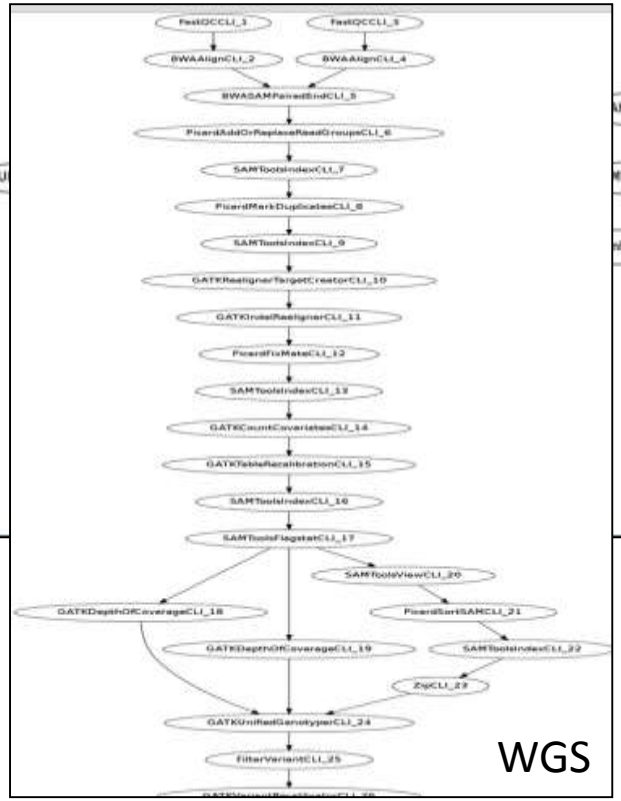
renci

Computational workflows in Genomics

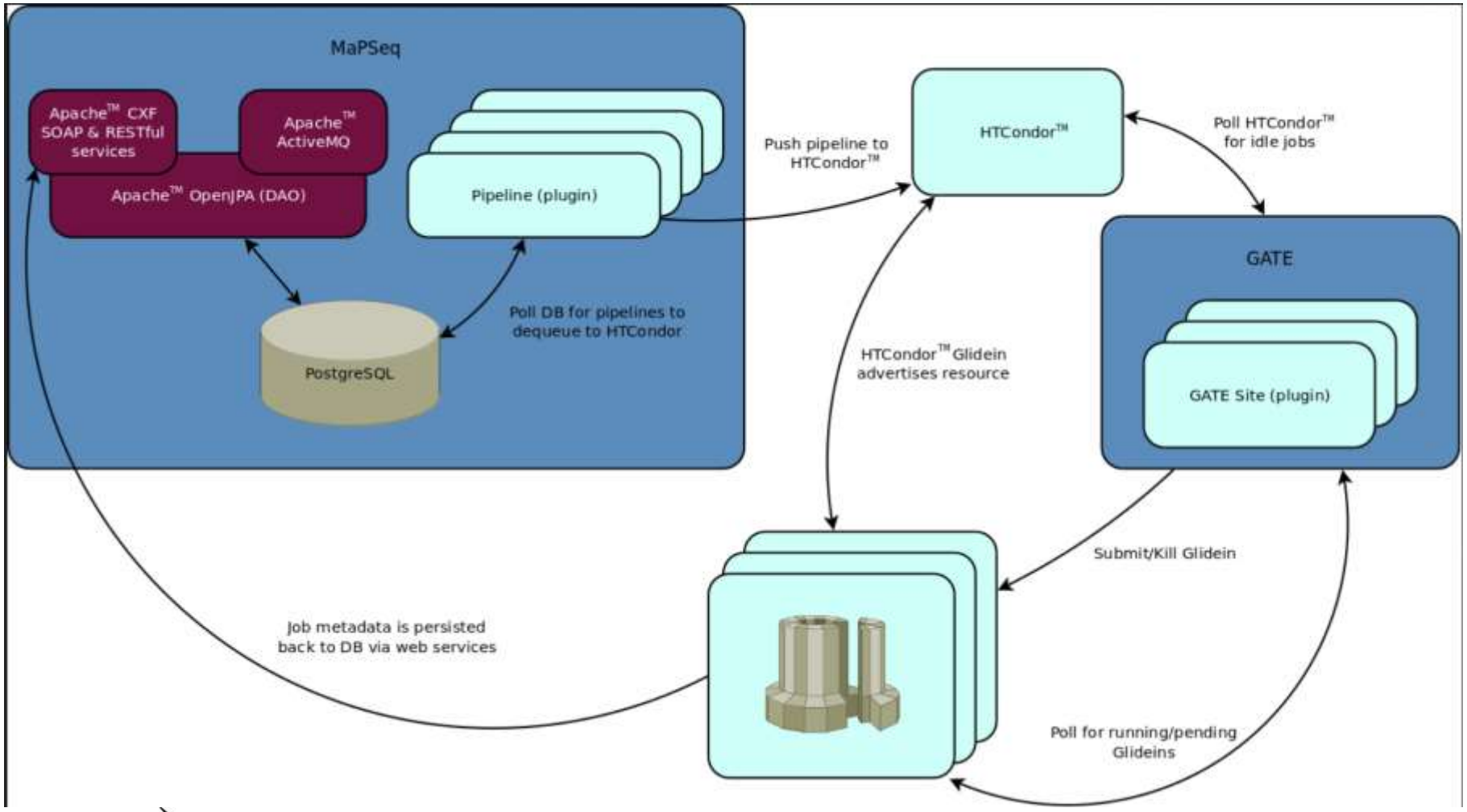
- 4 versions as we scaled:
  - single machine
  - cluster based
  - 2 multi-cluster/grid

- MapSeq – Current solution, Condor-based computational workflow





Analysis Pipeline - MapSeq



renci

Jason Reilly

Analysis Pipeline - MapSeq

Local Computational Clusters Currently Accessed by MaPSeq/GATE:

- KillDevil (ITS Research Computing): • Dell blade-based Linux Cluster • 604 Compute Nodes: 48GB RAM • 68 Compute Nodes: 96GB RAM • 2 Compute Nodes: 1TB RAM • 32 GPU compute nodes with 64 Nvidia Tesla GPU cards

- Kure (ITS Research Computing): • 2.2 PB Isilon system • HP blade-based Linux Cluster • 136 Compute Nodes: 48GB RAM • 80 Compute Nodes: 72GB RAM • 2 Compute Nodes: 96GB RAM • 3 Compute Nodes: 192GB RAM • Infiniband 4x QDR

- BlueRidge (RENCI) • Dell blade-based Linux Cluster • 128 Dell PowerEdge m610 blades (1024 cores total) • 32 Dell PowerEdge m610 blades (384 cores total) • 2 NVidia Teslas s1070-500 • 2 Dell PowerEdge R910 4 x 2.00Ghz Intel Nehalum-EX, 8 core, 1TB 1066Mhz memory

- Topsail (Genetics) 400 node dual quad-core Intel Clovertown 2.33 gigahertz processors with 12 gigabytes

MapSeq features

- Open source
  - https://github.com/jdr0887/MaPSeq-Distribution
- Multiple clusters are accessed opportunistically
- Minimal user intervention after system configuration
- Modularity of pipelines allows for reuse
- Pipeline can be tailored, modified, and updated as needed
- Pipeline workflows can be revised and deployed by clients, thus minimizing administrator burden

Challenges - Today

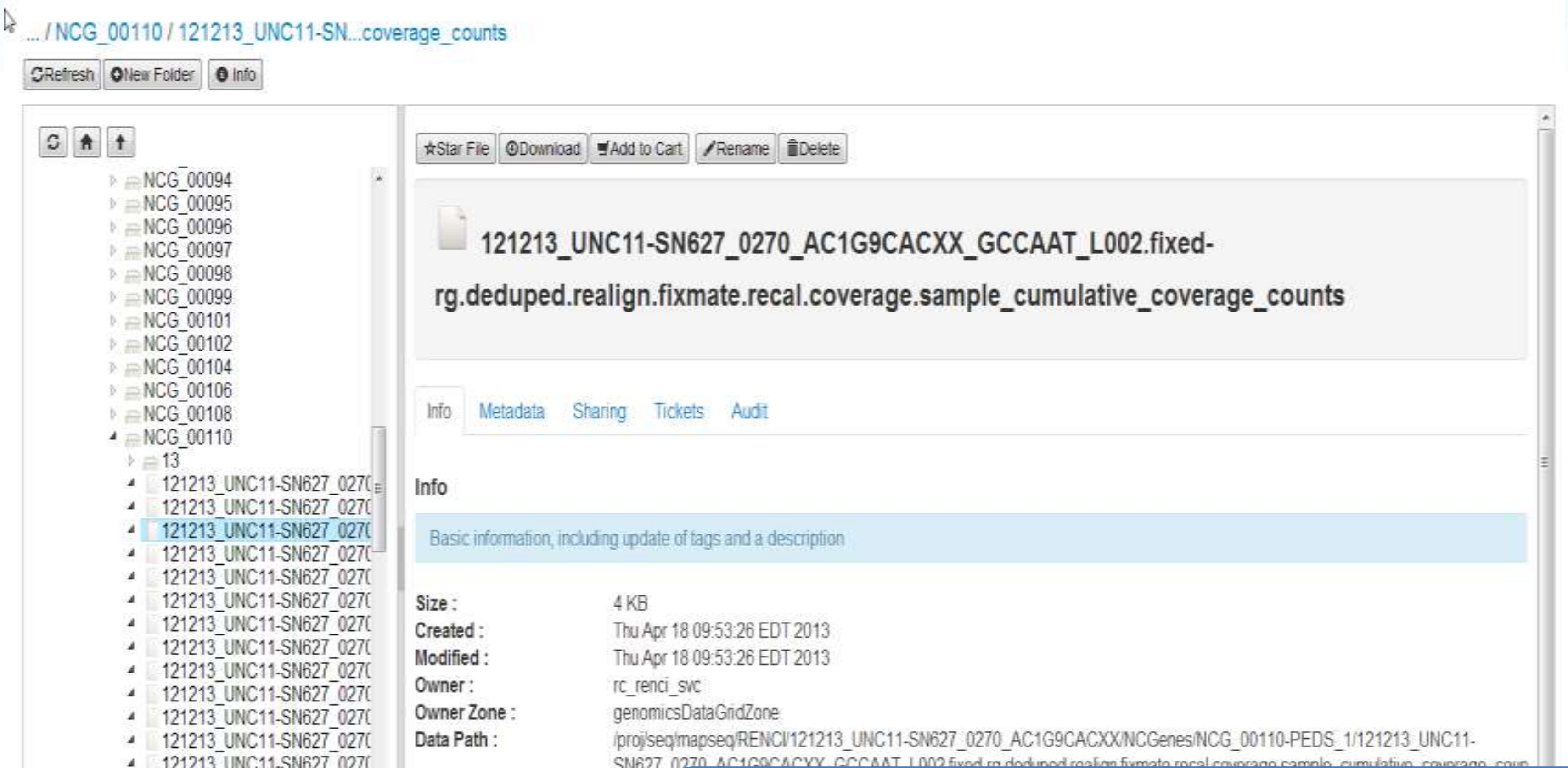
- Complexity of solutions
  - We are too weighted toward Comp Sci/IT, not Bio
- Risk
  - Too dependent on our own expertise in running complex systems
- Transportable
  - Solutions don’t transfer and aren’t easily shared
- Cost of scaling
  - We would prefer cloud like models

Challenges - Tomorrow

- Bridging islands of data
  - Autism data in Amazon
  - Environmental data at NIEHS
  - Genetic data at NCBI
- Moving data is increasingly challenged
  - Size, cost, security, regulatory, and privacy
- Inefficient use of resources/high cost of ownership
- Suboptimal collaboration structure

renci

iRODS: unified view of data



... / NCG_00110 / 121213_UNC11-SN...coverage_counts

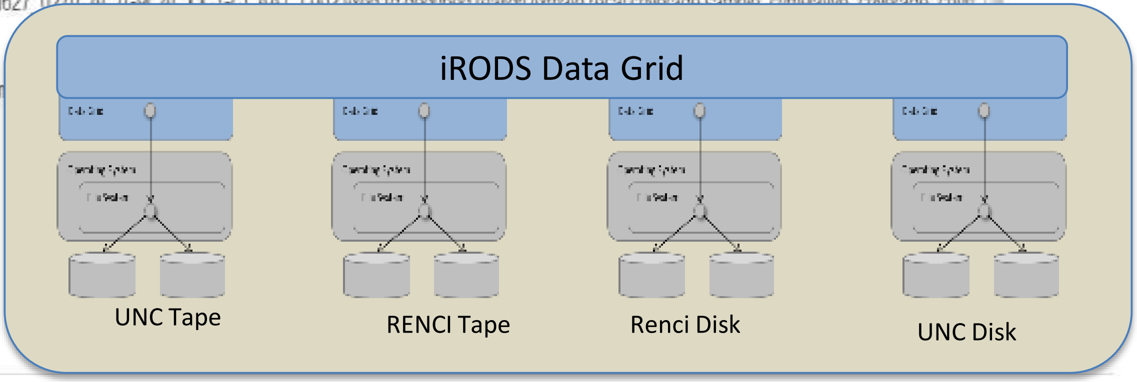
NCG_00094
NCG_00095
NCG_00096
NCG_00097
NCG_00098
NCG_00099
NCG_00101
NCG_00102
NCG_00104
NCG_00106
NCG_00108
NCG_00110
13
121213_UNC11-SN627_027
121213_UNC11-SN627_027
121213_UNC11-SN627_027
121213_UNC11-SN627_027
121213_UNC11-SN627_027
121213_UNC11-SN627_027
121213_UNC11-SN627_027
121213_UNC11-SN627_027
121213_UNC11-SN627_027
121213_UNC11-SN627_027
121213_UNC11-SN627_027
121213_UNC11-SN627_027
121213_UNC11-SN627_027
121213_UNC11-SN627_027
121213_UNC11-SN627_027
121213_UNC11-SN627_027
121213_UNC11-SN627_027
121213_UNC11-SN627_027
121213_UNC11-SN627_027
121213_UNC11-SN627_027

121213_UNC11-SN627_0270_AC1G9CACXX_GCCAAT_L002.fixed-rg.deduped.realign.fixmate.recal.coverage.sample_cumulative_coverage_counts

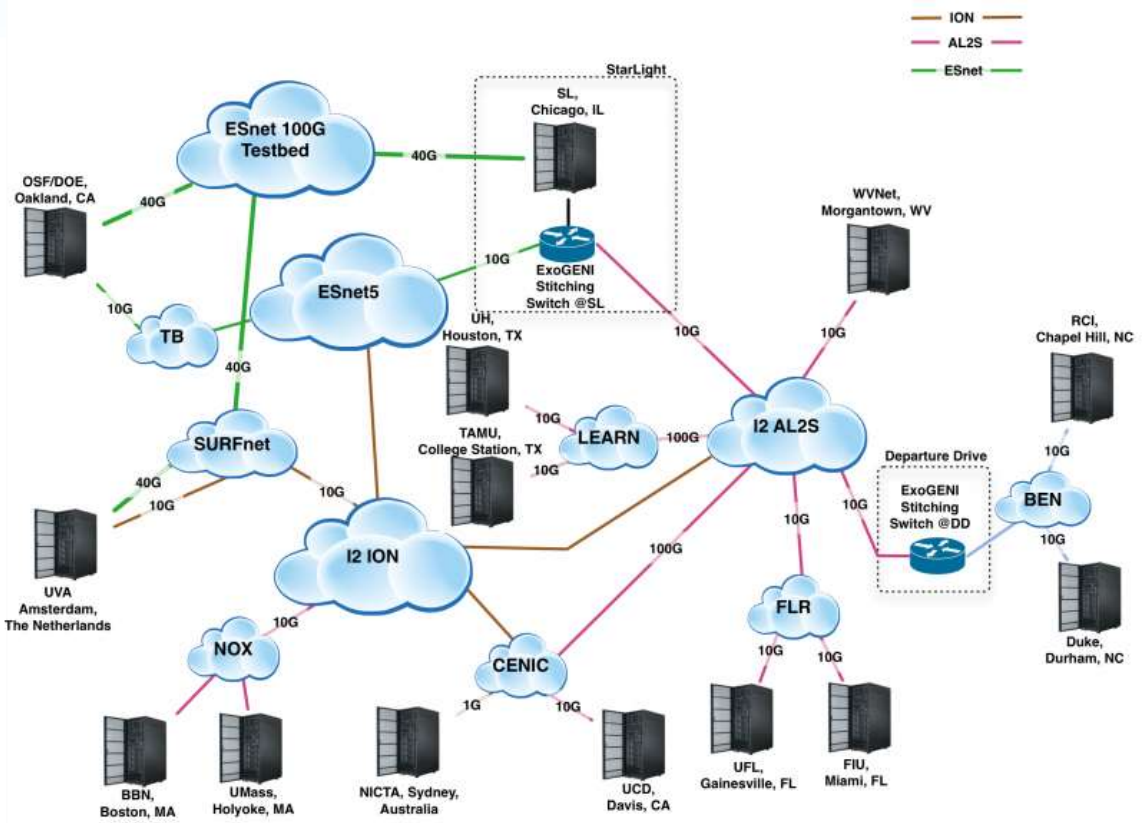
Info

Basic information, including update of tags and a description

Size : 4 KB
Created : Thu Apr 18 09:53:26 EDT 2013
Modified : Thu Apr 18 09:53:26 EDT 2013
Owner : rc_renci_svc
Owner Zone : genomicsDataGridZone
Data Path : /proj/seq/mapseq/RENCI/121213_UNC11-SN627_0270_AC1G9CACXX/NCGenes/NCG_00110-PEDS_1/121213_UNC11-SN627_0270_AC1G9CACXX_GCCAAT_L002.fixed-rg.deduped.realign.fixmate.recal.coverage.sample_cumulative_coverage_coun
Resource Group :
Checksum :
Resource : gen
Replica Number : 0
Replication Status : 1
Status :



ExoGENI NIaaS



Each rack is a small networked cloud:

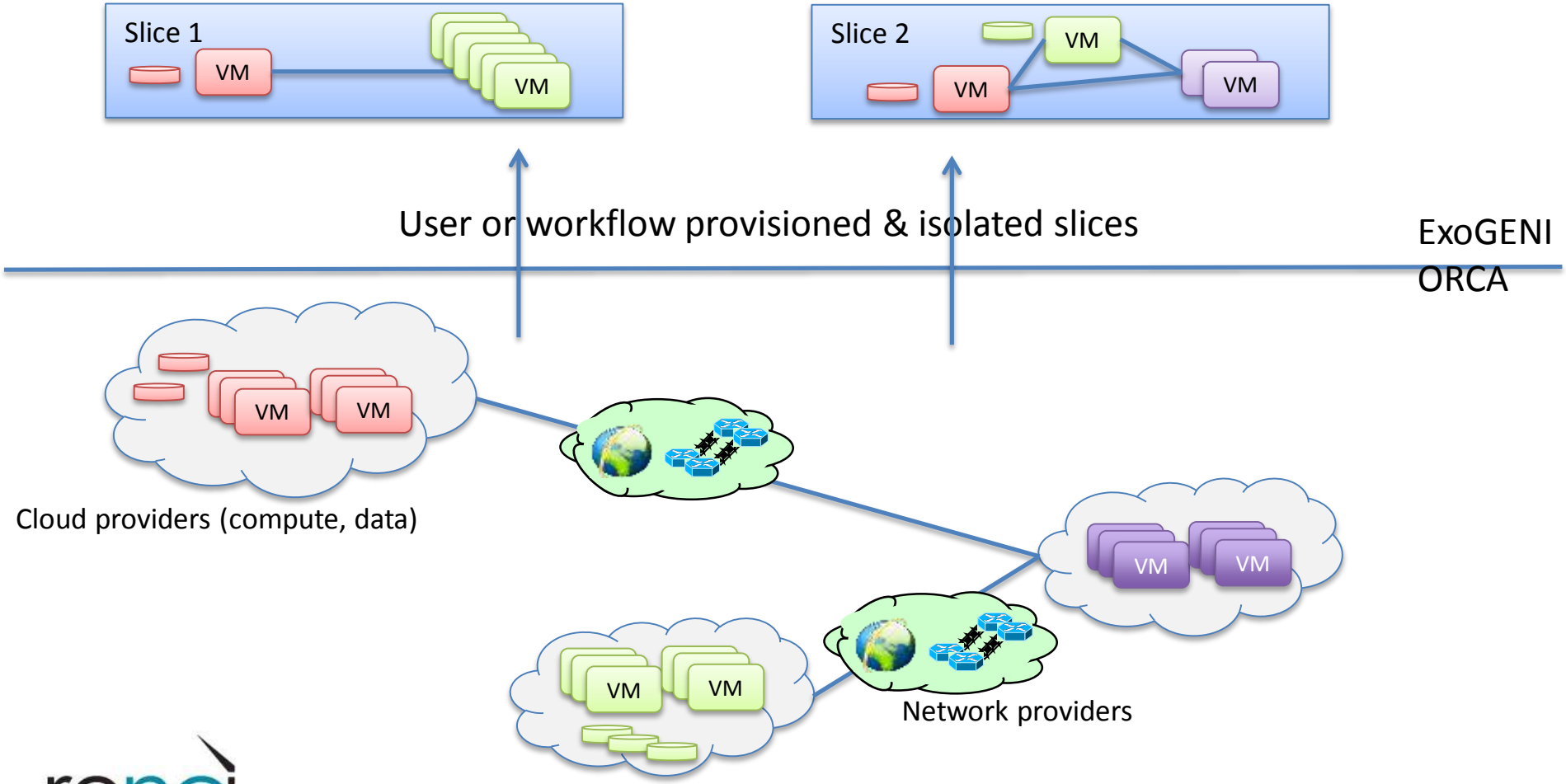
- OpenStack-based with NEuca extensions
  - www.networkedclouds.net
- EC2 node sizes (m1.small, m1.large etc)
- xCAT for baremetal node provisioning
- Sliverable storage

Resources from different racks can be interconnected on demand

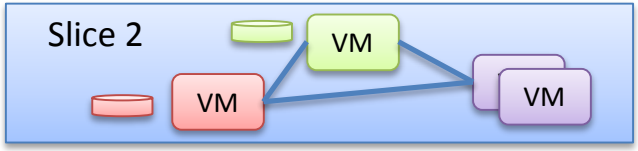
- ***PI: Ilia Baldine, RENCI***

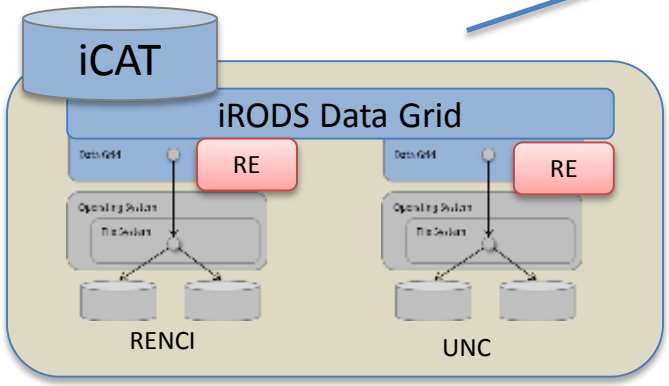
http://wiki.exogeni.net

Dynamic infrastructure



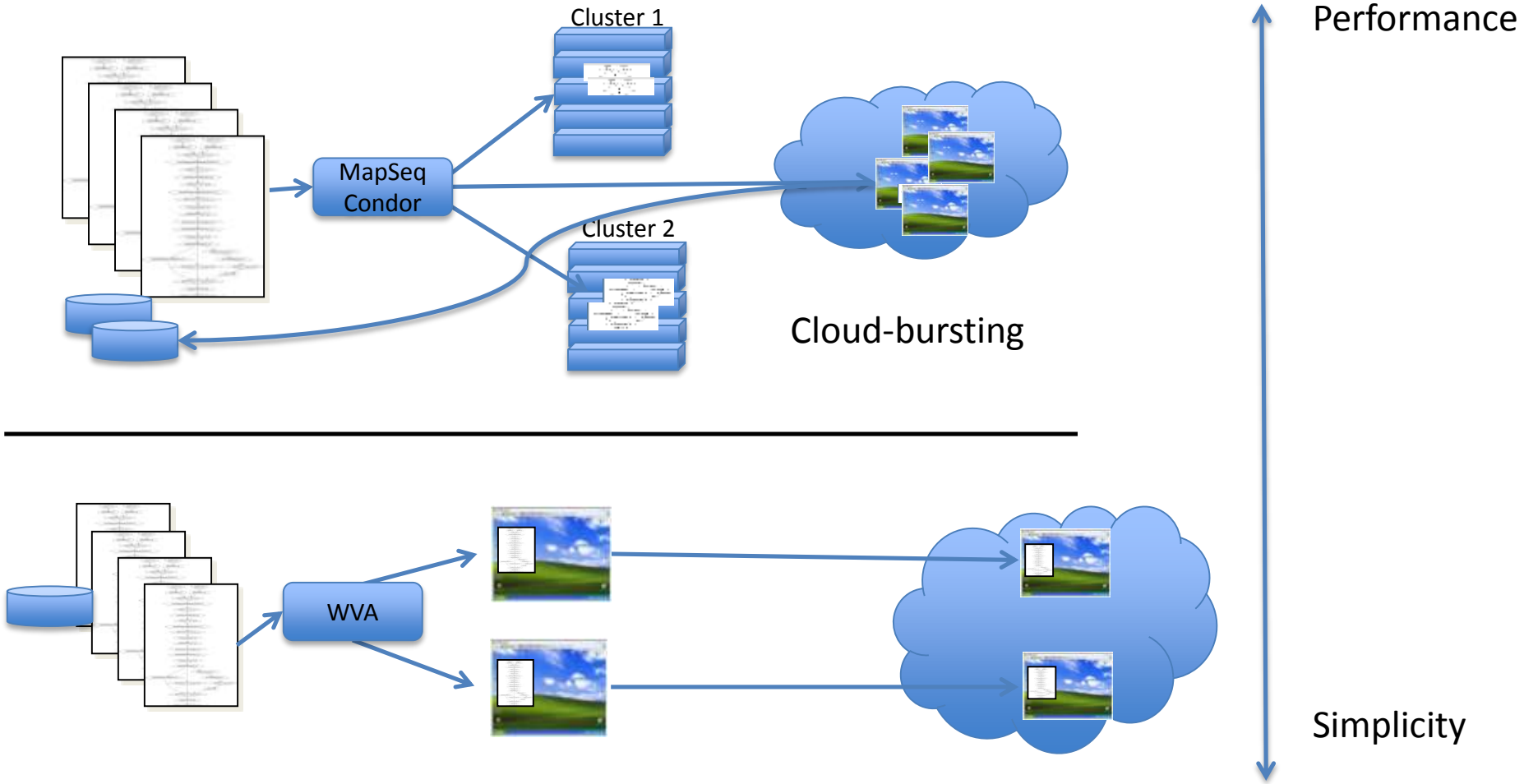
iRODS + ExoGENI: data management + virtual infrastructure





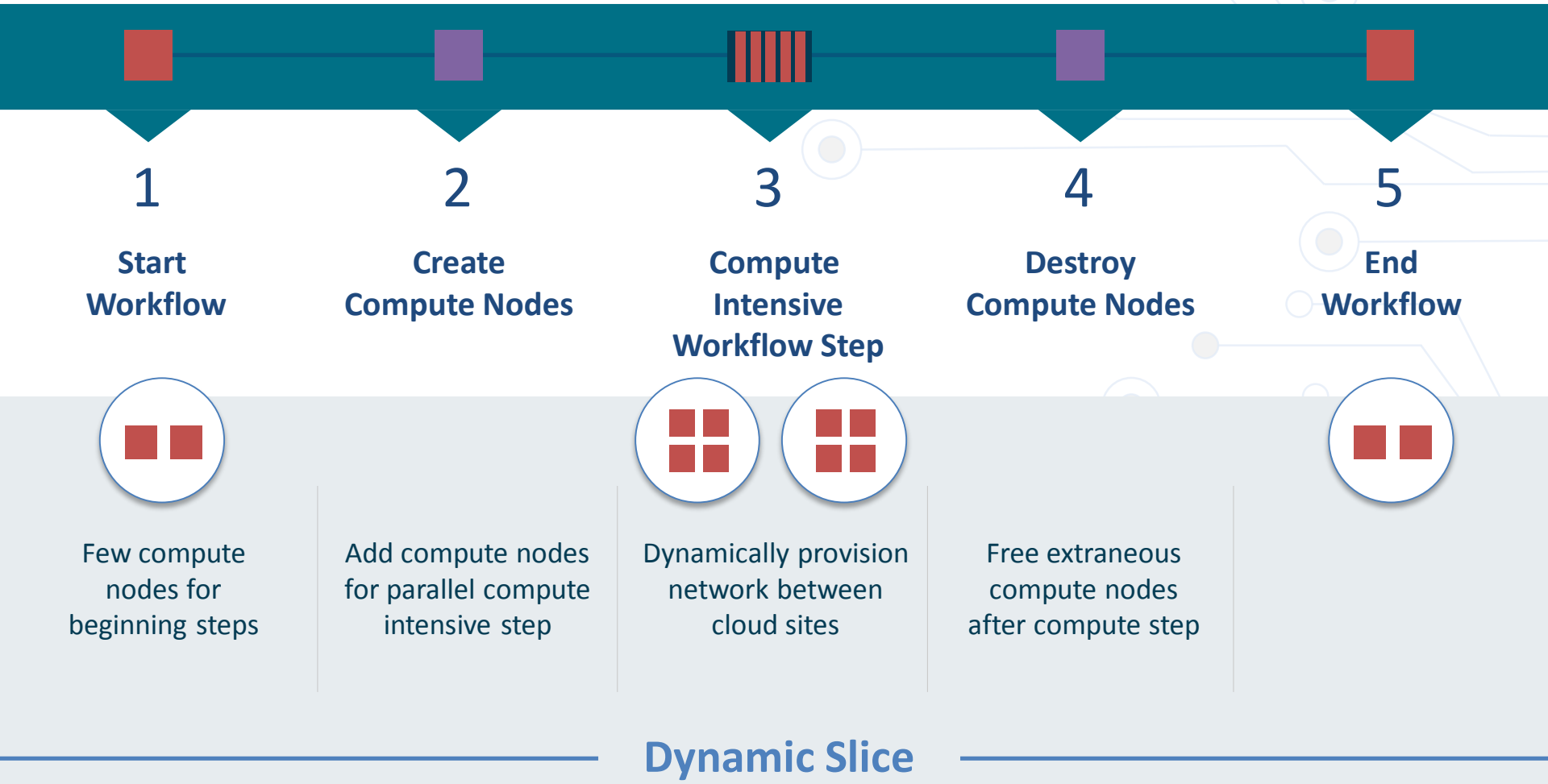
- Layer 2 connection within the slice
- Metadata control
  - Lab X can compute on Project Y data in the cloud
  - User X can move data from Study A to the cloud
  - Data from Study W cannot remain on cloud resources

Genomic workflows in the cloud



Cloud-based, 1 VM-1 workflow

ADAMANT– Pegasus/ExoGENI Dynamic Workflows



RADII: Bridging the gap between data and infrastructure

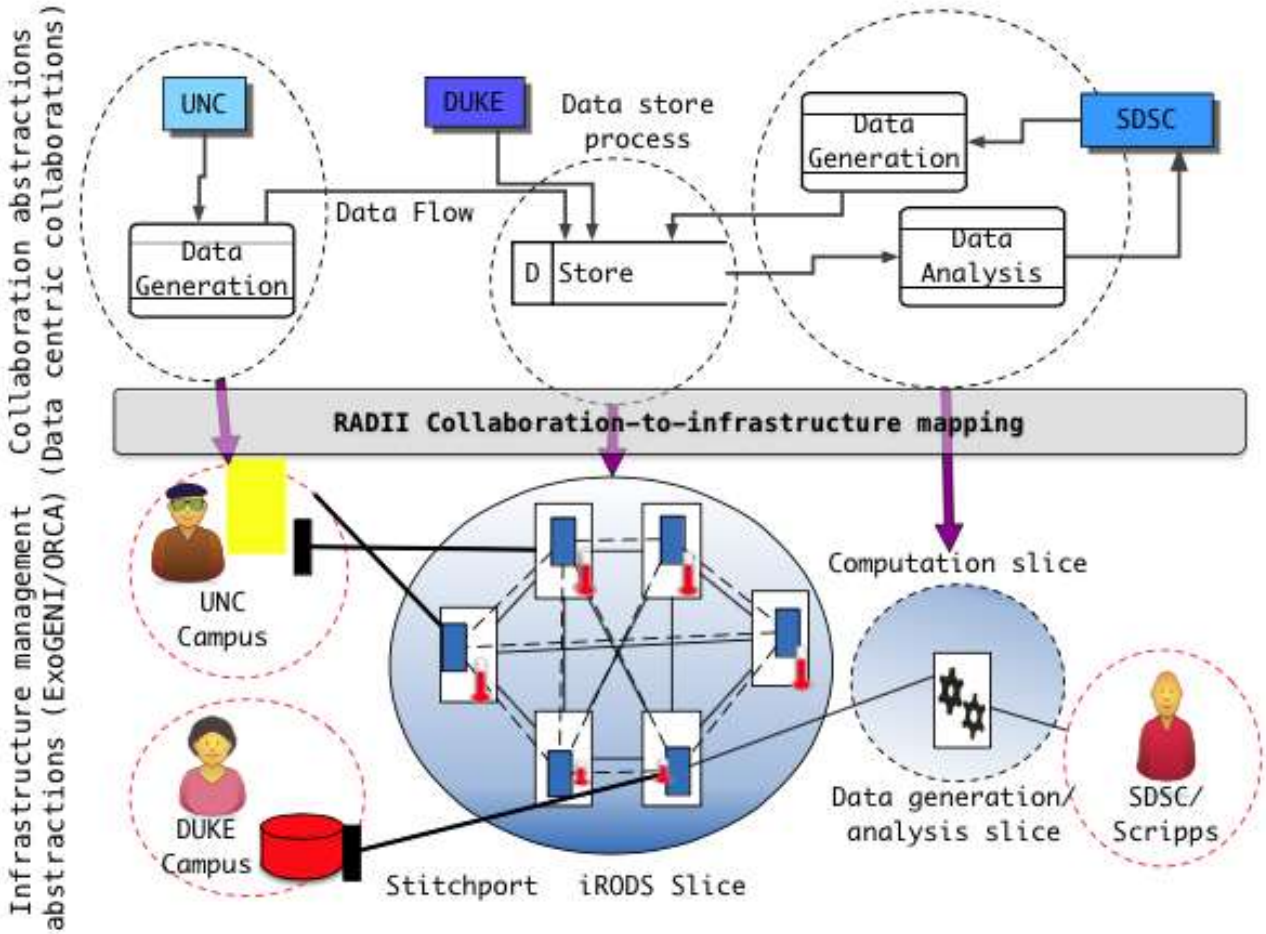


Figure 2: High level diagram of RADII's capabilities

SC15 RENCI Booth for Demo