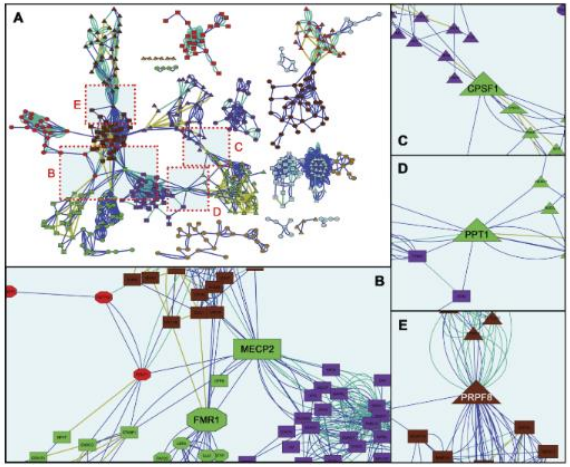
Importance of Cyberinfrastructure for Scientific Discovery

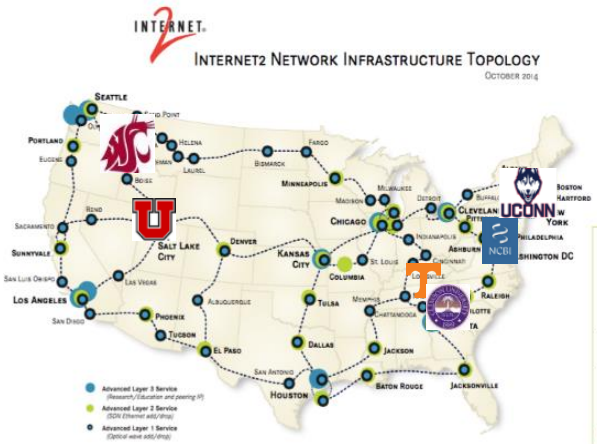
F. Alex Feltus, Ph.D.

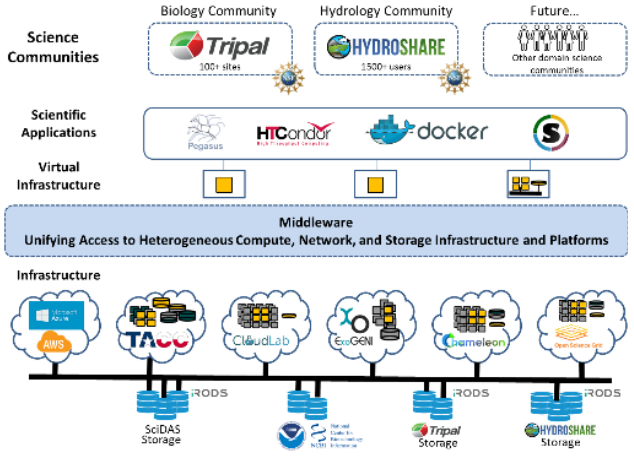
Clemson Dept. of Genetics & Biochemistry (Professor)
CU-MUSC Biomedical Data Science & Informatics Program (Member)
CU Center for Human Genetics (Member)
Allele Systems LLC (CEO)
PraxisAI (Founder)
Internet2 Board of Trustees (Member)

ffeltus@clemson.edu

Rethinking NSF's Computational Ecosystem for 21st Century Science and Engineering
July 22 2019 @ 8.40am



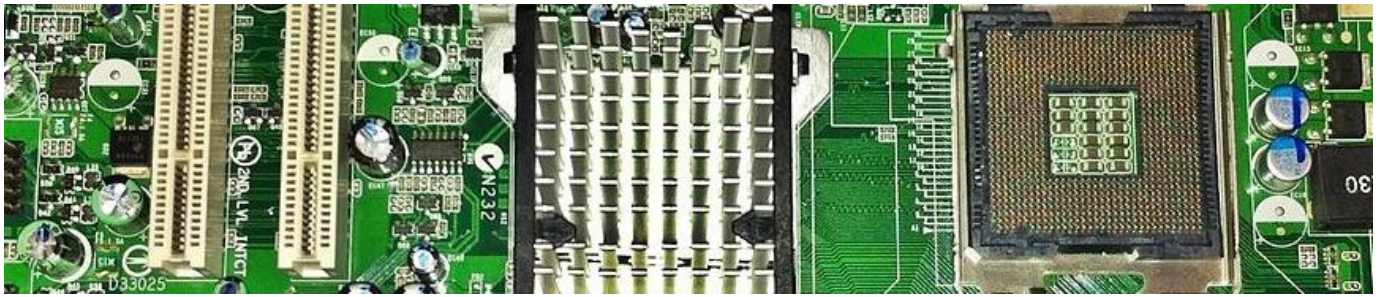




My BIOLOGY Lab = 1/3 Animal; 1/3 Plant; 1/3 Computational

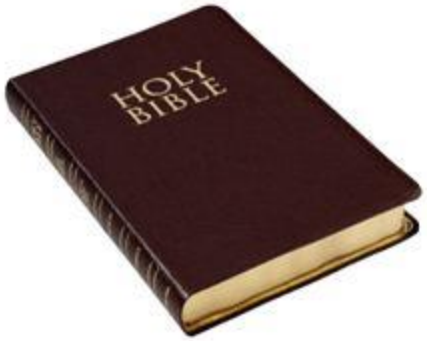
Vertebrates

Angiosperms



Bioinformatics/ Cyberinfrastructure

Information in Bible vs Human Genome

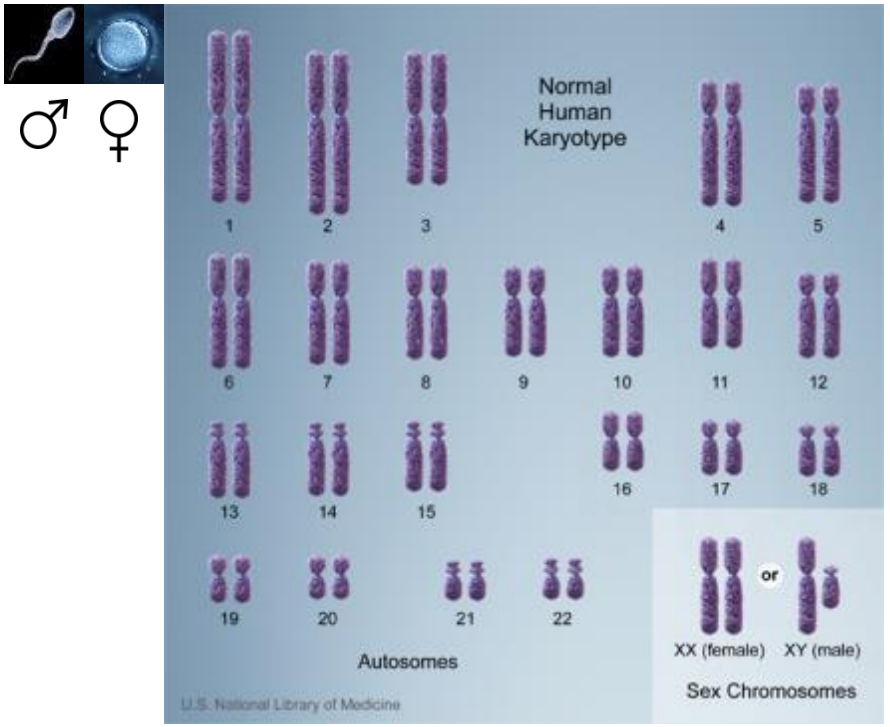


774,746 Words
26 letter alphabet

John 3:16: For God so loved the world that he gave his one and only Son, that whoever believes in him shall not perish but have eternal life.

John 3:16: For God so loved the world that he gave his one and only Sun, that whoever believes in him shall not perish but have eternal life.

One letter change = big difference!



58,721 Words (Genes)
4 letter alphabet (ATGC)

AATGGAGCCACATAACACATTCAAACTTACTTGCAAAATAT

AATGGAGCCACATAACACATGCAAACTTACTTGCAAAATAT

One letter change = higher risk for breast cancer!

What is a genome? What is a gene?

Human Genome "Facts"

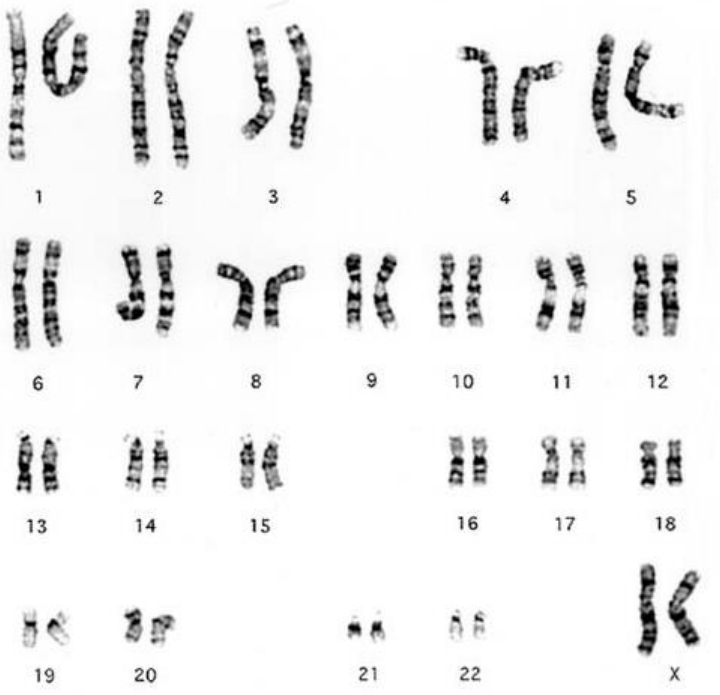
Initiation instructions for development

(3.2 billion*2) ATGCN

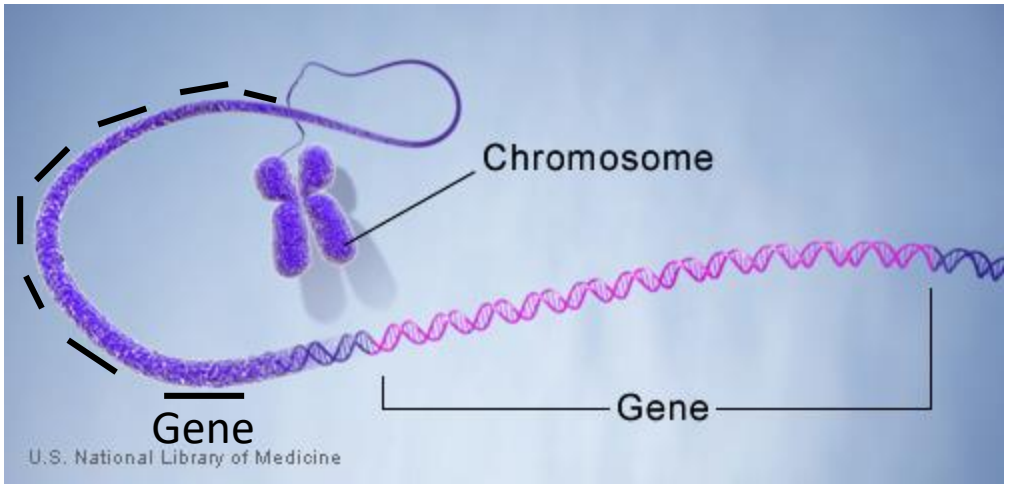
46 total chromosomes

24 unique chromosomes

24 strings (Chr1 = 250 million ATGCN)



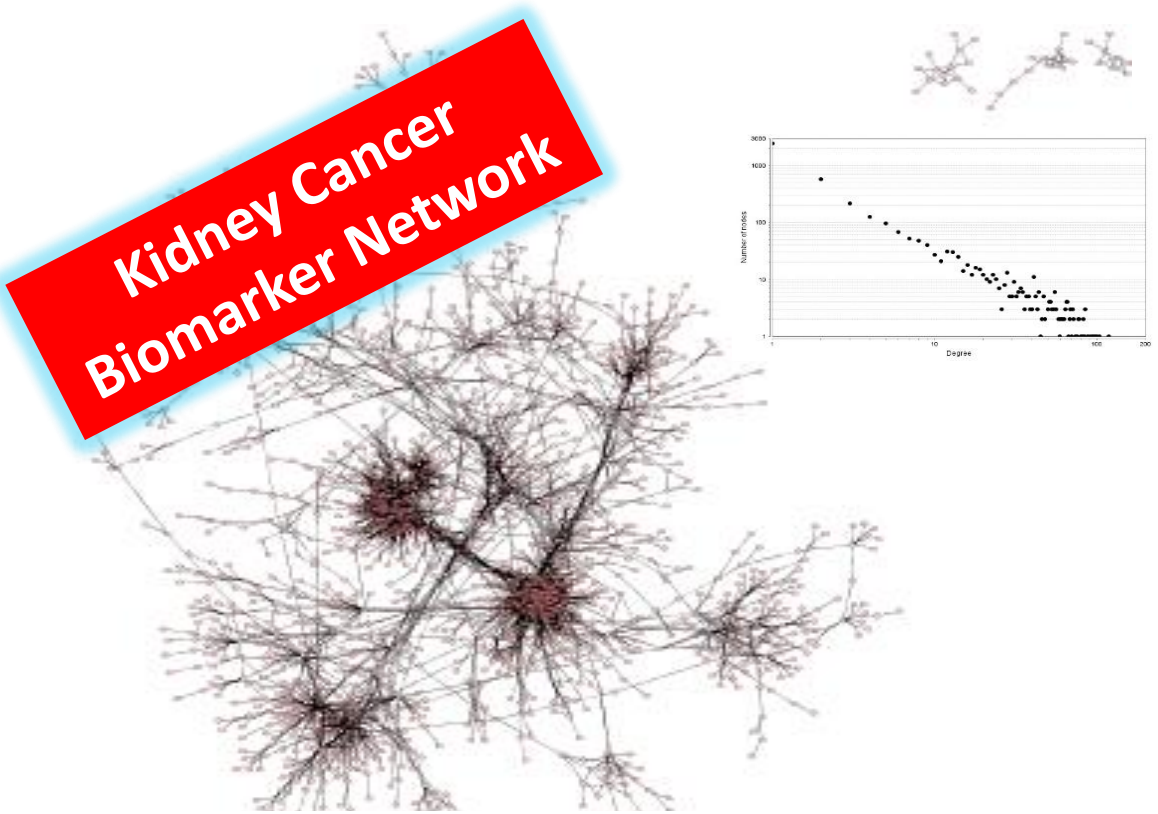
Courtesy of Dr. K. Phelan, Greenwood Genetic Center.
Noncommercial, educational use only.

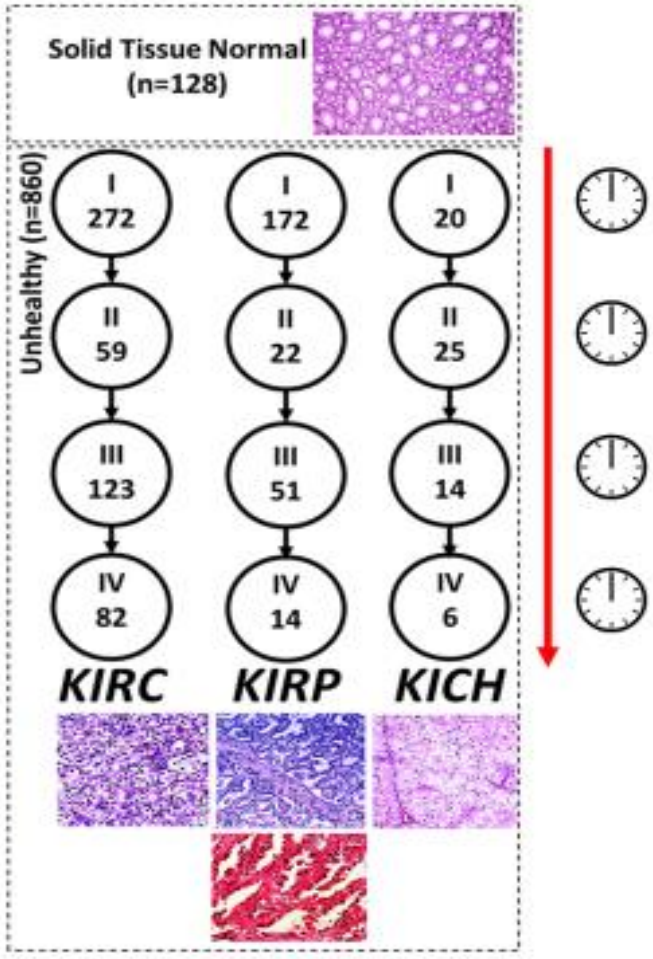


1-2TBs/human =
Electronic Medical Record

We use advanced cyberinfrastructure to move and process terabytes of DNA sequence to build biological networks

Kidney Cancer Biomarker Network





Article | OPEN | Published: 27 February 2019

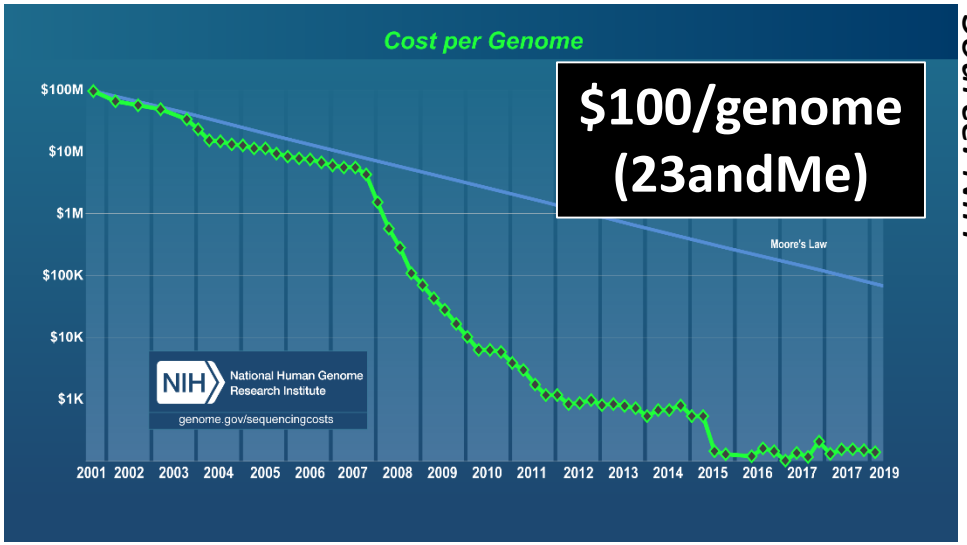
Linking Binary Gene Relationships to Drivers of Renal Cell Carcinoma Reveals Convergent Function in Alternate Tumor Progression Paths

William L. Poehlman, James J. Hsieh & F. Alex Feltus

Scientific Reports **9**, Article number: 2899 (2019) | Download Citation

A Lot Has Happened in Biology in the Last Decade

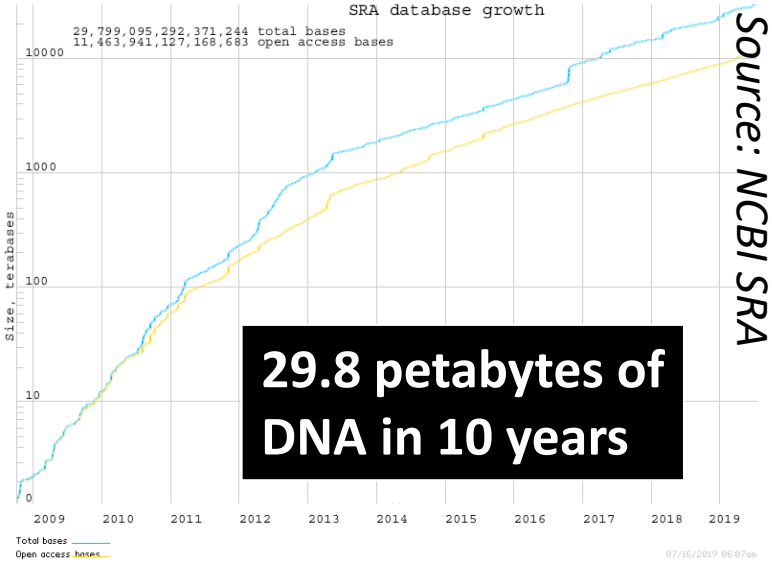
DNA Sequencing Cost is Dropping



$100/genome (23andMe)

Source: NIH

DNA Databases are Swelling

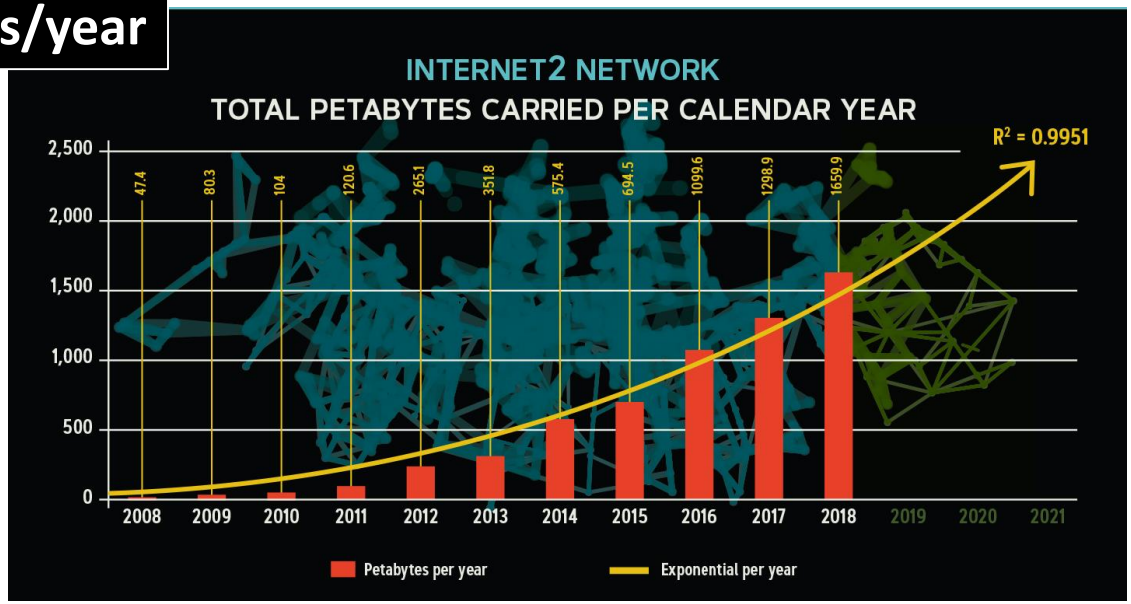


29.8 petabytes of DNA in 10 years

Source: NCBI SRA

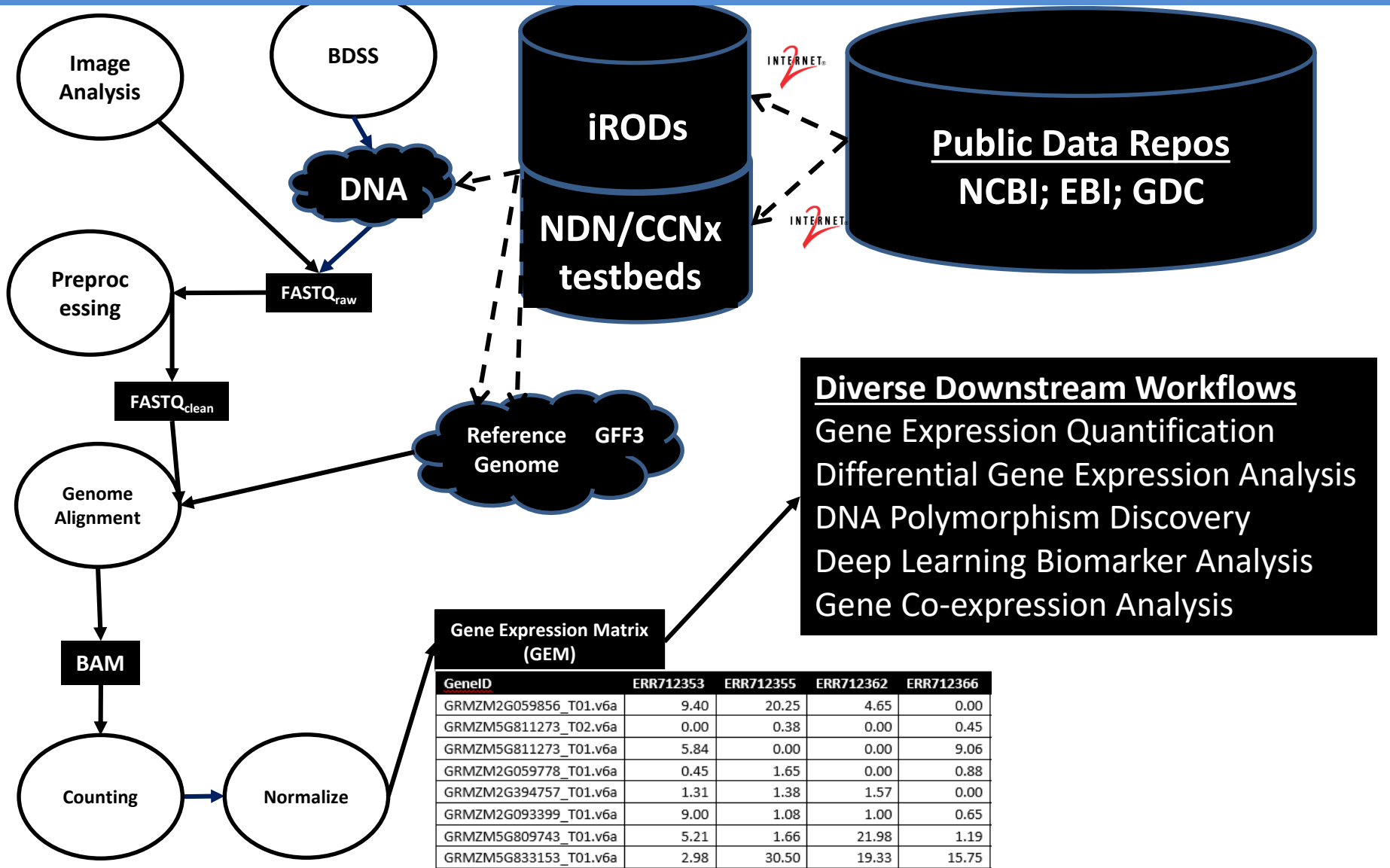
Data Networks are Flowing

1.6 exabytes/year



Source: Internet2

Computational Workflows Are Part of the Modern Biology Lab



GeneID	ERR712353	ERR712355	ERR712362	ERR712366
GRMZM2G059856_T01.v6a	9.40	20.25	4.65	0.00
GRMZM5G811273_T02.v6a	0.00	0.38	0.00	0.45
GRMZM5G811273_T01.v6a	5.84	0.00	0.00	9.06
GRMZM2G059778_T01.v6a	0.45	1.65	0.00	0.88
GRMZM2G394757_T01.v6a	1.31	1.38	1.57	0.00
GRMZM2G093399_T01.v6a	9.00	1.08	1.00	0.65
GRMZM5G809743_T01.v6a	5.21	1.66	21.98	1.19
GRMZM5G833153_T01.v6a	2.98	30.50	19.33	15.75
AC177838.2_FGT012.v6a	0.29	0.24	0.00	0.00

Diverse Downstream Workflows

Gene Expression Quantification
Differential Gene Expression Analysis
DNA Polymorphism Discovery
Deep Learning Biomarker Analysis
Gene Co-expression Analysis

nextflow Pegasus Galaxy

Source: Feltus Lab

We Run Workflows at the Tera/Petascale on Multiple Systems

Clemson Palmetto Cluster

2021 compute nodes (23072 cores) 814.4 TFlops.
100Gbit Internet2; Condo Model

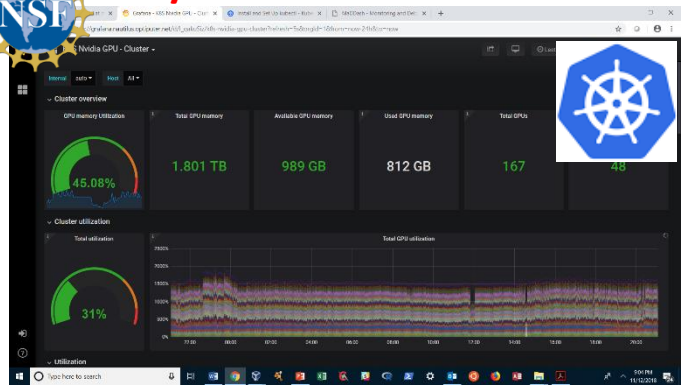
Open Science Grid

CENTER FOR HIGH THROUGHPUT COMPUTING
chtc.cs.wisc.edu

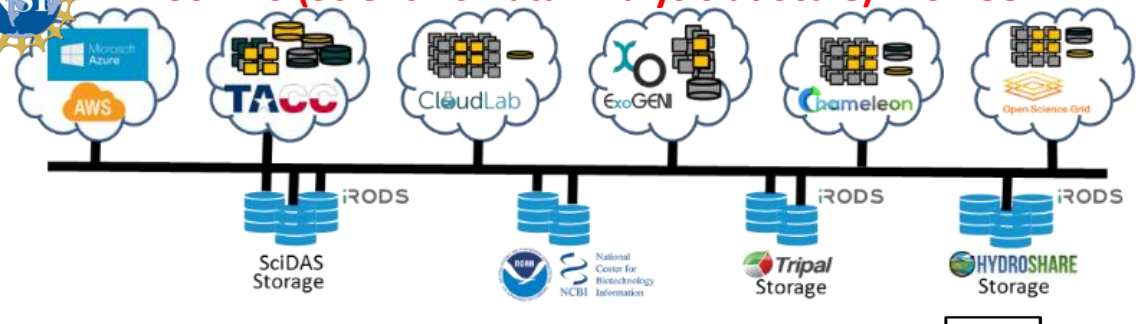
In 2017 on OSG …

8.43 Million Wall Hours
4.50 Million CPU Hours
8.92 Million Jobs
16.6 Million Transfers
4.07 PB

PRP/NRP Kubernetes Cluster



SciDAS (Scientific Data Analysis at Scale): NSF CC*



SciDAS

SciDAS dynamically creates an elastic distributed cloud compute system.

Google Cloud Platform

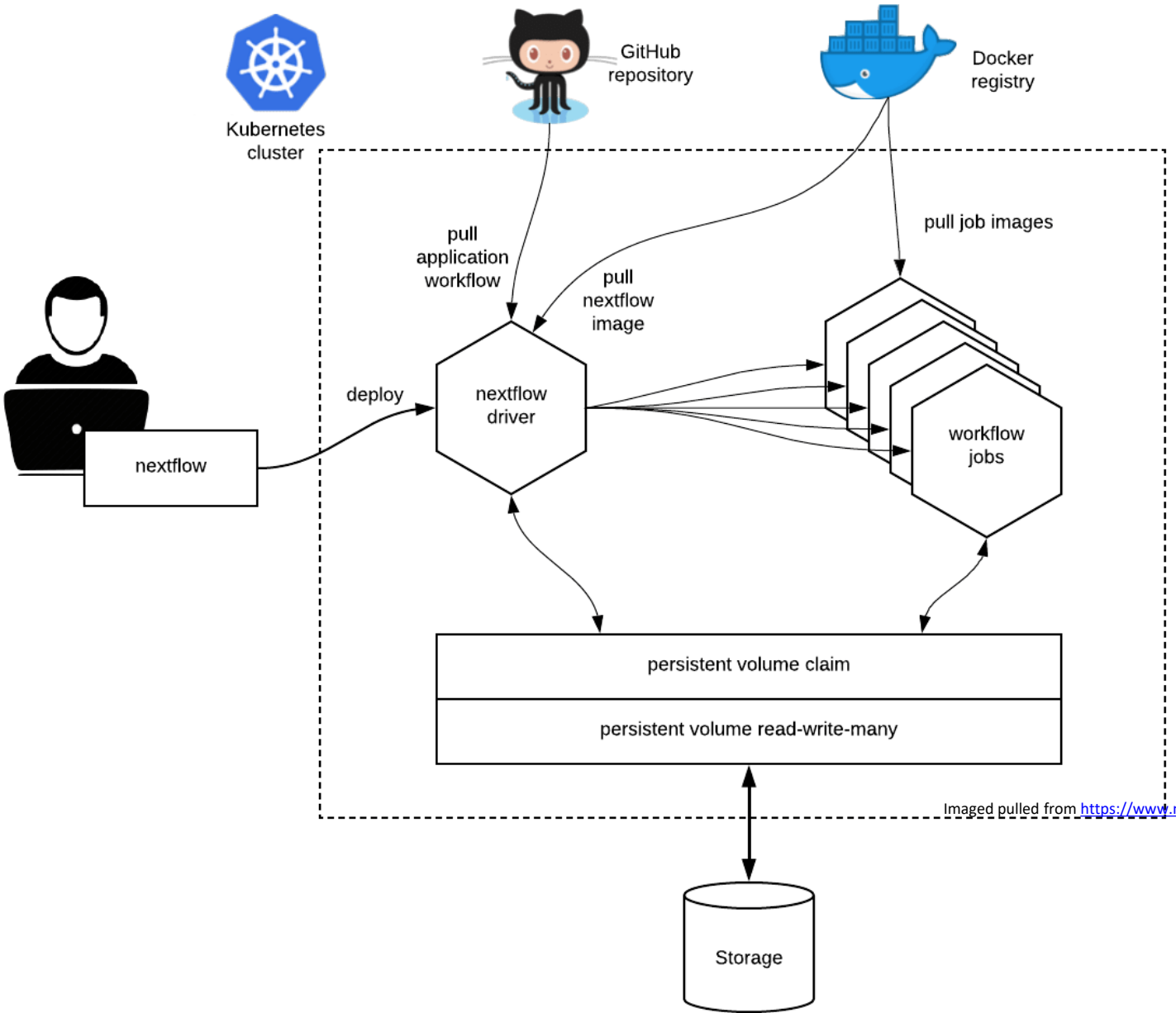
Cisco Hybrid Cloud

cisco

Google Cloud

Washington State University
renci
HYDROSHARE
Tripal
Chameleon
ExoGENI
Open Science Grid
NCBI National Center for Biotechnology Information
INTERNET2

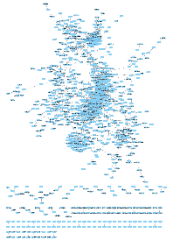
Kubernetes + NextFlow + Applications in Containers

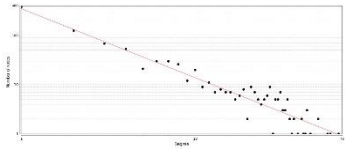


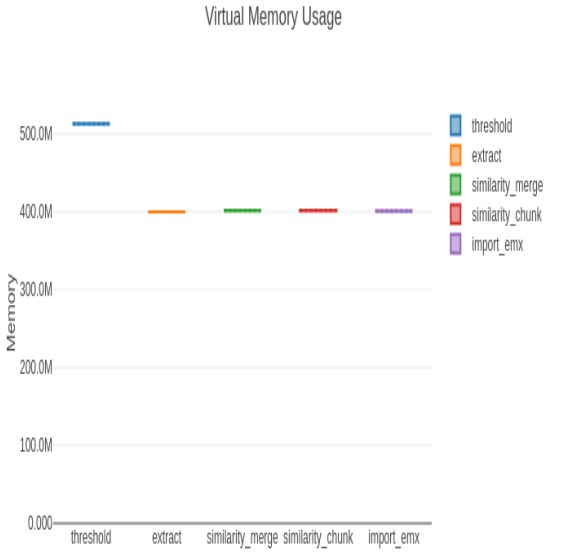
Imaged pulled from https://www.nextflow.io/docs/latest/kubernetes.html

A Yeast Gene Network Unit Test Built Across Three Kubernetes Clusters

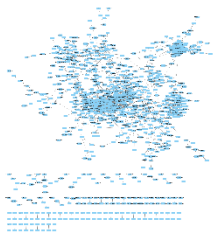
Cisco – Hybrid Cloud

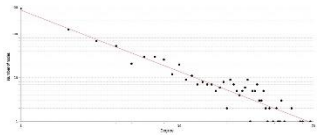


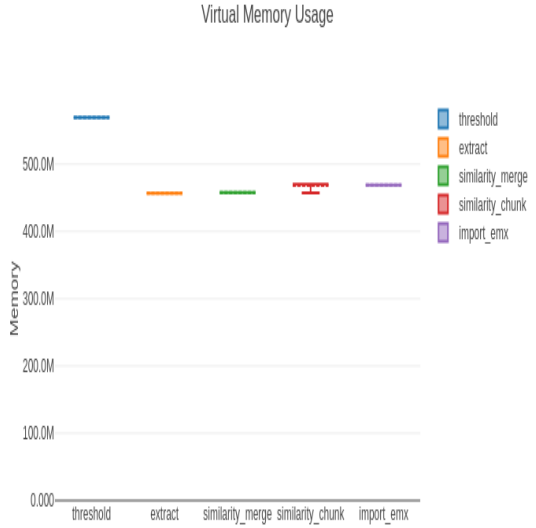




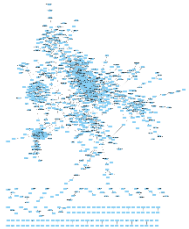
GCP - GKE

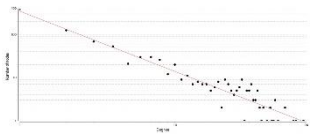


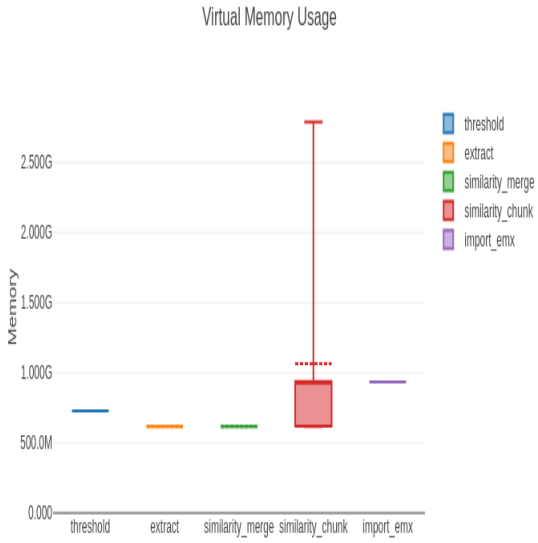




National Research Platform





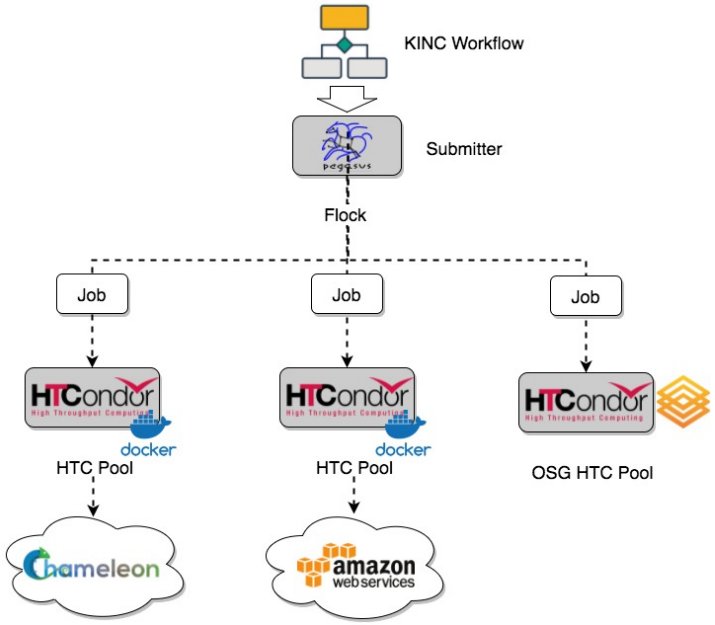


Nautilus

Example SciDAS "HTCondor" SciApp (Pegasus + HTCondor+ KINC)


```
{
  "id": "pegasus-htc",
  "containers": [
    {
      "id": "submitter",
      "image": "scidas/kinc-submitter",
      "resources": { "cpus": 2,   "mem": 4096,"disk": 10240
      },
      "cluster": "chameleon",
      "port_mappings": [{"container_port": 22, "host_port": 0,"protocol": "tcp"}],
      "args":
      [
   "-f","chameleon-master,aws-master,azure-master",
"-k", "ssh-rsa
AAAAB3NzaC1yc2EAAAADAQABAAABAQC303e2y8aUaMQ1IkHWnGFyb5XykxO
M5pLK83XFxWZMKsbYcgmkoODZ4w4COratlQPyMXSz7yaFUbYUccXjIjz8SDZf/9
c3xI0UuILOiVfb5QI/dsfssgsfvxcvfdsss321nksnvsnvlkvlksvkkdkddvlkssvn/xk+TOR
ZYK3CE3Oqu9p77nrFM7W3M5khsb5Qg/z0W1TQmVWvo5/i3QbDK6YaWhw/0
DXjfCeEtdlTVdIq1EJxMWuJnm5IptB1EtG9GBhuHq5Ct2XkUh",
        "-u", "irodsuser","-p","fdsfsfsdczxv3rr3r",
        "-h","irods-renci.scidas.org", "-z", "irodsZone"
      ]
    },
    {
      "id": "chameleon-master",
      "image": "scidas/htcondor-worker-centos7:1",
      "cluster": "chameleon",
      "resources": {
        "cpus": 48,
        "mem": 49152,
        "disk": 10240
      },
```


Input ➡ **Output**



Other SciApps

SLURM-NEXTFLOW

Kubernetes-Galaxy

Kubernetes-Nextflow

Jupyter Notebook

TOIL-CWL

INPUT: DNA SEQUENCE/QUALITY FASTQ FILE

(10 – 200 Million "DNA Reads"; $slen=36-300 characters))

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

SEQUENCE

QUALITY

http://en.wikipedia.org/wiki/FASTQ_format

Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL (null)	32	20	040		Space	64	40	100	@	@	96	60	140	`	`
1	1	001	SOH (start of heading)	33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
2	2	002	STX (start of text)	34	22	042	"	"	66	42	102	B	B	98	62	142	b	b
3	3	003	ETX (end of text)	35	23	043	#	#	67	43	103	C	C	99	63	143	c	c
4	4	004	EOT (end of transmission)	36	24	044	$	$	68	44	104	D	D	100	64	144	d	d
5	5	005	ENQ (enquiry)	37	25	045	%	%	69	45	105	E	E	101	65	145	e	e
6	6	006	ACK (acknowledge)	38	26	046	&	&	70	46	106	F	F	102	66	146	f	f
7	7	007	BEL (bell)	39	27	047	'	'	71	47	107	G	G	103	67	147	g	g
8	8	010	BS (backspace)	40	28	050	(	(	72	48	110	H	H	104	68	150	h	h
9	9	011	TAB (horizontal tab)	41	29	051	)	)	73	49	111	I	I	105	69	151	i	i
10	A	012	LF (NL line feed, new line)	42	2A	052	*	*	74	4A	112	J	J	106	6A	152	j	j
11	B	013	VT (vertical tab)	43	2B	053	+	+	75	4B	113	K	K	107	6B	153	k	k
12	C	014	FF (NP form feed, new page)	44	2C	054	,	,	76	4C	114	L	L	108	6C	154	l	l
13	D	015	CR (carriage return)	45	2D	055	-	-	77	4D	115	M	M	109	6D	155	m	m
14	E	016	SO (shift out)	46	2E	056	.	.	78	4E	116	N	N	110	6E	156	n	n
15	F	017	SI (shift in)	47	2F	057	/	/	79	4F	117	O	O	111	6F	157	o	o
16	10	020	DLE (data link escape)	48	30	060	0	0	80	50	120	P	P	112	70	160	p	p
17	11	021	DC1 (device control 1)	49	31	061	1	1	81	51	121	Q	Q	113	71	161	q	q
18	12	022	DC2 (device control 2)	50	32	062	2	2	82	52	122	R	R	114	72	162	r	r
19	13	023	DC3 (device control 3)	51	33	063	3	3	83	53	123	S	S	115	73	163	s	s
20	14	024	DC4 (device control 4)	52	34	064	4	4	84	54	124	T	T	116	74	164	t	t
21	15	025	NAK (negative acknowledge)	53	35	065	5	5	85	55	125	U	U	117	75	165	u	u
22	16	026	SYN (synchronous idle)	54	36	066	6	6	86	56	126	V	V	118	76	166	v	v
23	17	027	ETB (end of trans. block)	55	37	067	7	7	87	57	127	W	W	119	77	167	w	w
24	18	030	CAN (cancel)	56	38	070	8	8	88	58	130	X	X	120	78	170	x	x
25	19	031	EM (end of medium)	57	39	071	9	9	89	59	131	Y	Y	121	79	171	y	y
26	1A	032	SUB (substitute)	58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	z	z
27	1B	033	ESC (escape)	59	3B	073	;	;	91	5B	133	[	[	123	7B	173	{	{
28	1C	034	FS (file separator)	60	3C	074	<	<	92	5C	134	\	\	124	7C	174			|
29	1D	035	GS (group separator)	61	3D	075	=	=	93	5D	135	]	]	125	7D	175	}	}
30	1E	036	RS (record separator)	62	3E	076	>	>	94	5E	136	^	^	126	7E	176	~	~
31	1F	037	US (unit separator)	63	3F	077	?	?	95	5F	137	_	_	127	7F	177	&#127;	DEL

Source: www.LookupTables.com

http://en.wikipedia.org/wiki/Phred_quality_score
http://www.asciitable.com/

Base Call Quality (Q) score is ASCII Decimal# minus 33 (usually).
Q(I)=40:::99.999% Accuracy

$$Q = -10 \log_{10} P$$

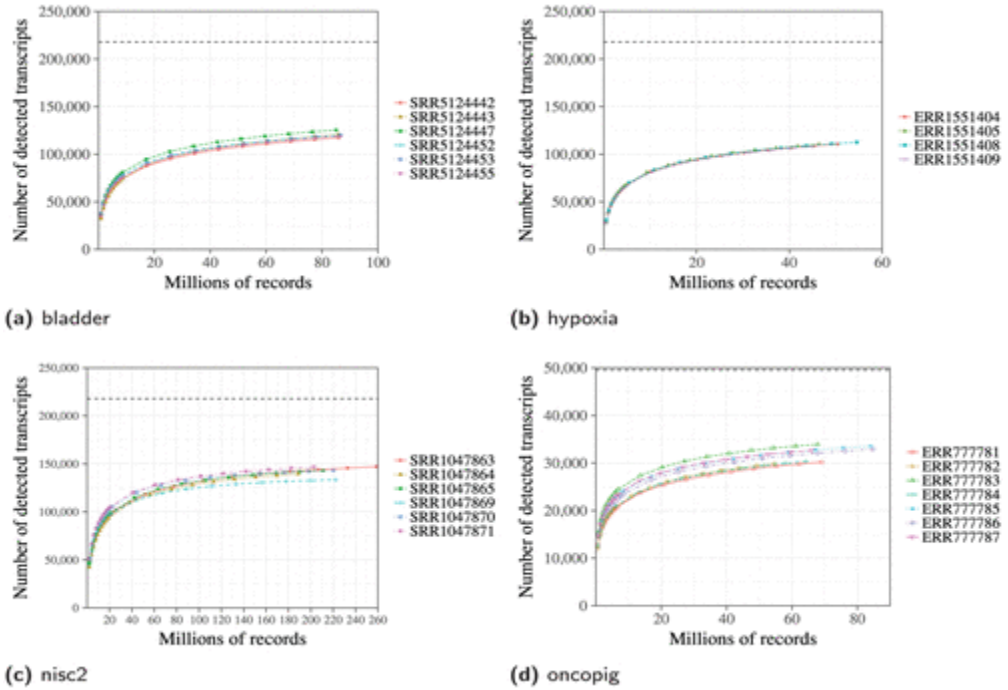
or

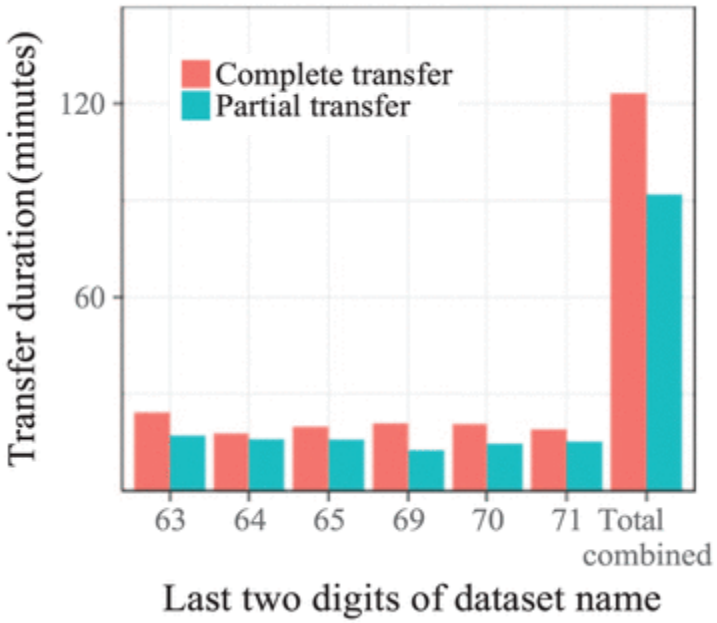
$$P = 10^{\frac{-Q}{10}}$$

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

Partial Data Transfer Reduces Time to Result





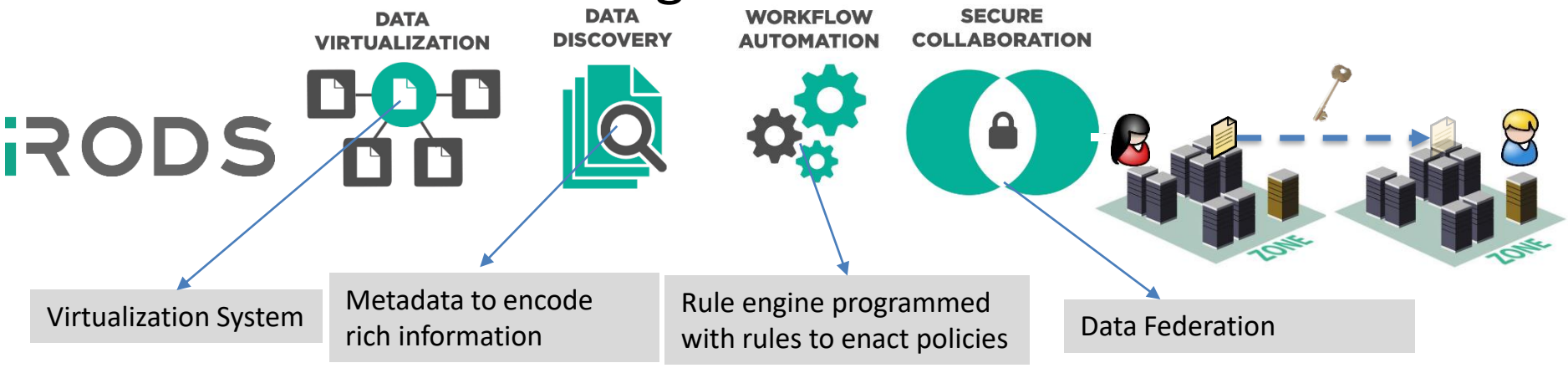
Moving Just Enough Deep Sequencing Data to Get the Job Done

Nicholas Mills[1], Ethan M Bensman[2], William L Poehlman[3], Walter B Ligon III[1] and F Alex Feltus[3]

[1]Holcombe Department of Electrical and Computer Engineering, Clemson University, Clemson, SC, USA. [2]School of Computing, Clemson University, Clemson, SC, USA. [3]Department of Genetics and Biochemistry, Clemson University, Clemson, SC, USA.

iRODS Distributed Data grid

- Integrated Rule Oriented Data System (iRODS) provides a **distributed** unified namespace over SciDAS storage infrastructure across Clemson, RENCI and WSU (4.2 petabyte Data Grid;1232 indexed genomes; GEM-GCN Storage)
- iRODS provides enable policy-driven management critical to data-sharing collaborations in SciDAS

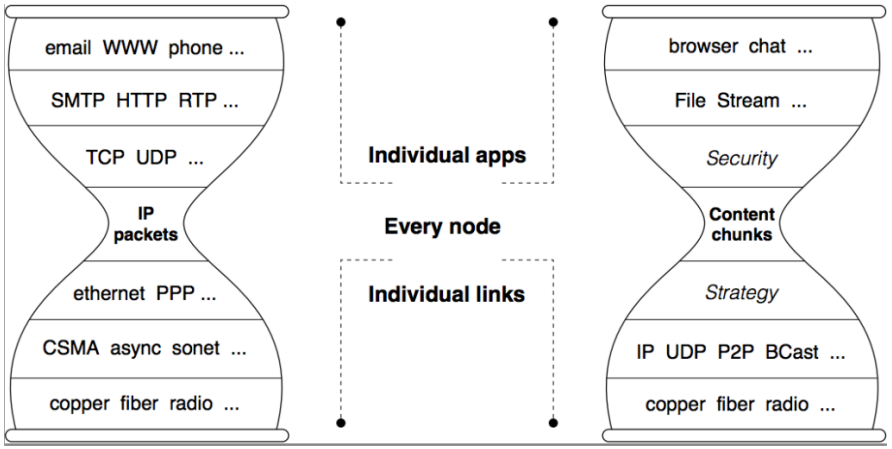


Terrell Russell, Michael Stealey, Jason Coposky, Ben Keller, Claris Castillo, Ray Idaszak, Alex Feltus. Distributing the iRODS Catalog: A Way Forward. iRODS UGM 2017 Proceedings. Page 35, 2017.

Named Defined Networking (NDN) “Data Grid”

NDN Design Principles

- **[1] Universality**: NDN should be a common network protocol for all applications and network environments.
- **[2] Data-Centricity and Data Immutability**: NDN should fetch uniquely named, immutable “data packets” requested using “interest packets”.
- **[3] Securing Data Directly**: Security should be the property of data packets, staying the same whether the packets are in motion or at rest.
- **[4] Hierarchical Naming**: Packets should carry hierarchical names to enable demultiplexing and provide structured context.
- **[5] In-Network Name Discovery**: Interests should be able use incomplete names to retrieve data packets.
- **[6] Hop-by-Hop Flow Balance**: Over each link, one interest packet should bring back no more than one data packet.



https://named-data.net

We Published 1,232 Indexed Genome Files in NDN Framework

- 1,232 Genomes pulled from iRODS-SciDAS Data Grid (NSF CC*1659300)
- >500 Gigabytes in Aggregate (tar-gz compressed)
- WSU indexed; CSU-CU moved to NDN
- Naming: 'Genus->Species->Intraspecies->Assembly->Files'
- NDN Caching near Genomics Workflows!

Hierarchical Named Data Format

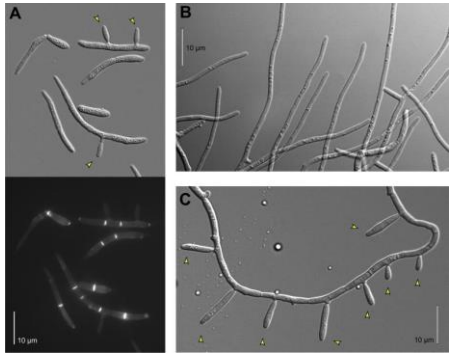
[genus][species]{[infraspecific name]}/[assembly_name] that contains these files:

> [genus]_[species]{_[infraspecific name]}-[assembly_name].{n}.ht2
> [genus]_[species]{_[infraspecific name]}-[assembly_name].gff3
> [genus]_[species]{_[infraspecific name]}-[assembly_name].fasta
> [genus]_[species]{_[infraspecific name]}-[assembly_name].gtf
> [genus]_[species]{_[infraspecific name]}-[assembly_name].Splice_sites
> [genus]_[species]{_[infraspecific > name]}-[assembly_name].meta.json

Genome Example

/scidasZone/sysbio/PynomeGenomes/Genome/Zymoseptoria_tritici/MG2:

> Zymoseptoria_tritici-MG2.1.ht2
> Zymoseptoria_tritici-MG2.2.ht2
> Zymoseptoria_tritici-MG2.3.ht2
> Zymoseptoria_tritici-MG2.4.ht2
> Zymoseptoria_tritici-MG2.5.ht2
> Zymoseptoria_tritici-MG2.6.ht2
> Zymoseptoria_tritici-MG2.7.ht2
> Zymoseptoria_tritici-MG2.8.ht2
> Zymoseptoria_tritici-MG2.fa
> Zymoseptoria_tritici-MG2.gff3
> Zymoseptoria_tritici-MG2.gtf
> Zymoseptoria_tritici-MG2.meta.json
> Zymoseptoria_tritici-MG2.Splice_sites



Fungal Genet Biol. 2015 Jun; 79: 17–23.

NDN-SCI for Managing Large Scale Genomics Data

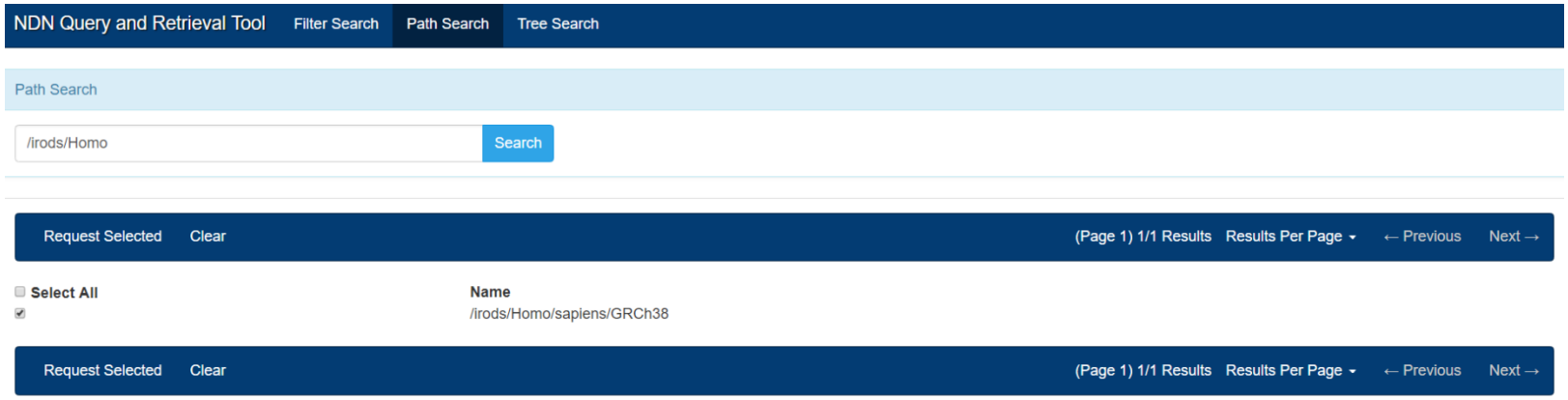
Susmit Shannigrahi
Colorado State University
susmit@cs.colostate.edu

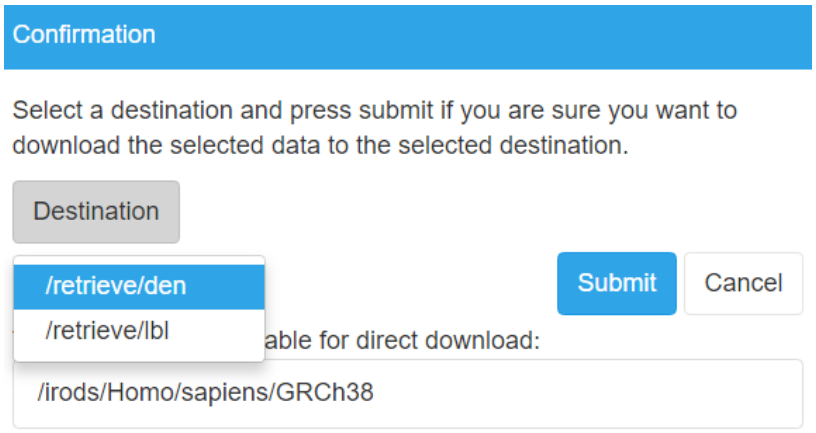
Chengyu Fan
Colorado State University
chengyu.fan@colostate.edu

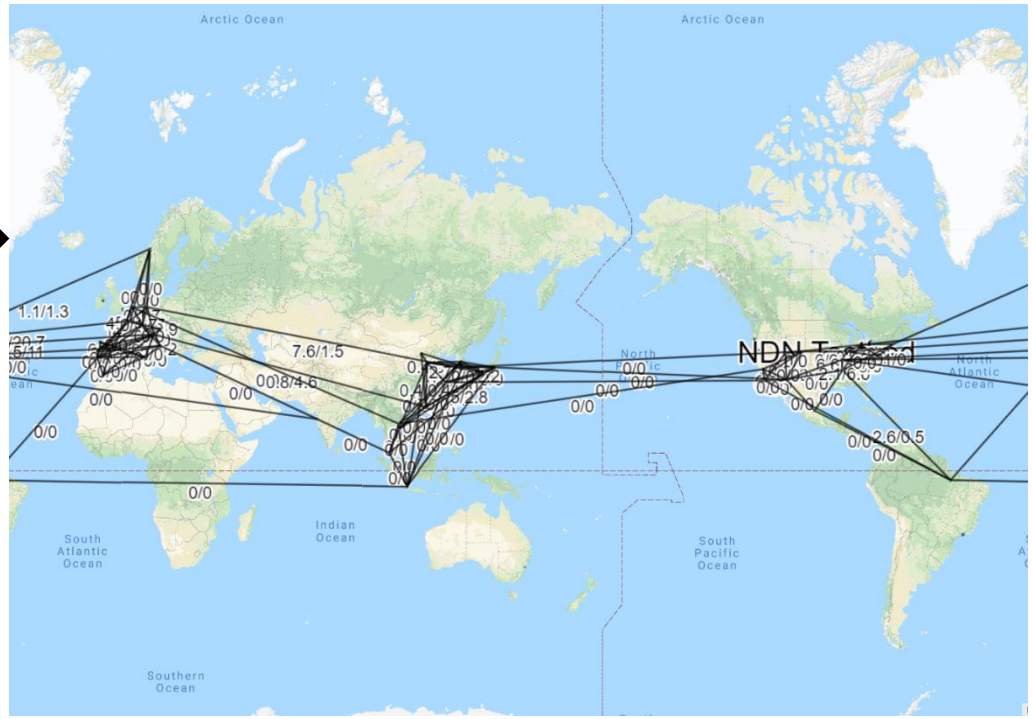
Christos Papadopoulos
Colorado State University
christos@colostate.edu

Alex Feltus
Clemson University
ffeltus@clemson.edu

NDN Data Discovery Via a Web Based UI and Moved (or copied from cache) to an Endpoint on the Network







http://atmos-sac.es.net/genome/

A Lot Will Happen in Biology in the Next 15 Years

The tea leaves say…

In 2019, most R1 biology labs are outsourcing DNA sequencing.
Terascale genomics experiments are common now.

In 2029, every university research lab will have a DNA sequencer.
R1 research labs will move to the peta-/exascale in this PhD generation.

In 2034, all pharmacies, subways, hospitals, police stations, etc will have DNA sequencers.
These IoT DNA "sensors" will generate exabytes of data in aggregate each week.

I am only talking DNA Sequencers…not CryoEM, Simulations, Medical Imaging

{ amazon web services™ INTERNET 2® CVS pharmacy® 23andMe }

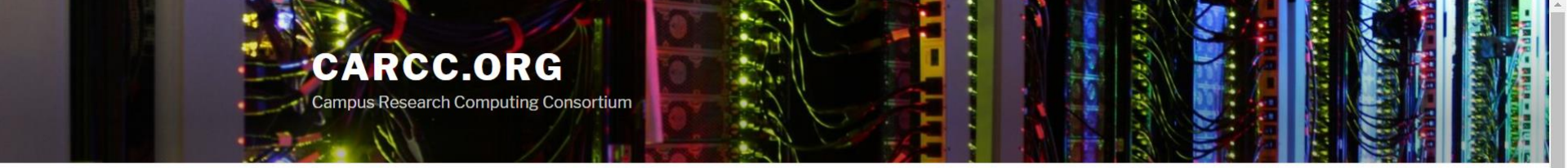
Genomics Scale Up Observations

Prediction: Giga-/Tera scale genomics experiments will move into the peta-/exa scale in this PhD generation.

Issues:::Solutions

- **Unpredictable time to compute result (queue times, queue times, queue times, broken nodes, segfaults, OOM, data geography, short walltimes) :::Software optimization; Real Parallel + Redneck Parallel Computing on GPUs/CPUs; SciDAS**
- **Not enough computational resources:::OSG, XSEDE, NRP/PRP, SLATE, Cloud**
- **Not enough in-lab ACI knowledge::: IT Engineer Lunch Dates, Governance committees, Research Facilitators, Software Carpentry, Collaborations: CS/CE/Engineering Departments/NRT**

- **Not enough storage:::Shared Data Grids; Negotiate cheaper storage with campus IT; Move to Cloud; Leverage /scratch space for intermediate files**
- **Poor use of advanced networks:::Avoid Commercial Internet; Perform data life cycle analysis and push data close to network; Data caching**
- **Data Organization:::iRODs DataGrid; Tripal Databases; Named Defined Networking**



CARCC.ORG

Campus Research Computing Consortium

Current CaRCC Working Groups (September 2018; updated May 2019)

Building Community: The People Network: This sustaining group aims to foster, build and grow an inclusive community (termed the "People Network") for campus CI, research computing and data professionals. This includes synchronous and asynchronous opportunities to leverage collective and individual expertise, with focused discussion tracks reflective of professional activities, including researcher-facing, systems-facing, software-facing, data-facing, and stakeholder/sponsor-facing.

Current co-chairs: Dana Brunson, Lauren Michael.

CI-professionalization: This sustaining group will further develop and disseminate frameworks and approaches to guide conversations between Human Resources leaders and research computing and data leaders around attracting, retaining, diversifying, and developing cyberinfrastructure and research computing and data talent. This includes boosting awareness of the value of a CI career. An initial draft of a "Research Computing and Data Professionals Job Elements and Career Guide" was developed by this group and people involved in a January 2018 CI Professionalization workshop.

"Stakeholders and Value Proposition": This working group made the original the stakeholders and value proposition document more broad and generic (i.e. less CaRCC focused). **The new draft is here** and we welcome feedback (either to the chairs or help@carcc.org). A goal of the working group is to establish a platform to support this as an interactive document. Additional short term goals are to prioritize among stakeholder groups for outreach and validation of the value propositions followed by field tests by CaRCC members on our respective campuses.

Current co-chairs: Andy Sherman, Barr von Oehsen.

CaRCC-logistics – "Logistics, Communications, and Common Infrastructure for CaRCC: Tools, Platforms, Processes, and Roles". This new "working group" aims to define the internal working structure, logistics and communications strategy for CaRCC. Initial aims are to evaluate and select tools, platforms and processes to facilitate CaRCC activities. These activities include enabling elections, communications, dissemination of CaRCC results as living and evolvable documents, overseeing working groups and timelines, membership considerations, meeting logistics, and defining committee structures and roles.

Current co-chairs: Dana Brunson, Gail Krovitz, Lauren Michael.

CaRCC-future-of-research-computing-data – "What is the future of research computing and data?". This new "working group" aims to define a vision and roadmap of expectations for where research computing and data is heading across levels from campuses, to regions, and the nation. Across all levels of compute and data we are witnessing continuous and–if not appropriately planned for — non-sustainable growth in demand for services, both in terms of physical resources and trained people/expertise. We as a community need to have clear plans and vision which can be shared with the larger community to increase awareness and to highlight risk, reward, implications, and challenges.

Current co-chairs: Jackie Milhans, Gwen Jacobs.

Ecosystem-of-research-computing-and-data – A workshop bringing elements of the community together was held in April 2019: An update regarding this will be available soon. The continuing aim of this "working group" is to follow-up on activities from the workshop and to plan next steps and for a panel presentation at PEARC19. A non-exclusive list of various research computing communities that were brought together include: CaRCC, Campus Champions, CI Engineers, CASC, among others– to better understand the vision and plans of the different communities, their synergies and overlaps, and to foster better communication, collaboration, and coordination. The intent is to help socialize and ultimately sustain the larger CI and research computing and data ecosystem.

Current co-chairs: Dana Brunson, Gail Krovitz, and Jim Wilgenbusch

Research computing and data (or Research IT) maturity model – A workshop bringing elements of the community together was held in December 2018, and a follow-on group of volunteers is continuing to develop the maturity model. Look for workshops on this at PEARC19 and EDUCAUSE in the Fall of 2019. The continuing aim of this working group is to develop a workbook or spreadsheet that allows organizations to rate or evaluate their maturity levels in research computing and data across the facings (researcher-facing, systems-facing, …) along multiple dimensions. This working group originated out of a "research IT" committee of Internet2, added EDUCASE, and in the Fall of 2018 CaRCC.

Geographically Distributed Interdisciplinary Science is Super Fun!

Feltus Lab
Yuqing Hang (<PhD, G&B)
Benafsh Husain (<PhD, BDSI)
Allison Hickman (<PhD, G&B)
Ben Shealy (<PhD, ECE)
Benafsh Husain (<PhD, BDSI)
Yueyao Gao (<PhD, G&B)
Cole Younginer (<PhD, ECE)
Mohammed Aburidi (<PhD, BDSI)
Cole McKnight (Cloud Architect)
Jordan Little (<BSc, G&B)
Ethan Bensman (<BSc, G&B)
Reed Bender (<BSc,BioE)
Cameron Ogle (<BSc, CS)
Recent alumni
Rachel Eimen (Bsc, ECE)
Courtney Shearer (BSc, CS)
Melissa Judge (<BSc, BioE)
Will Poehlman (PhD, G&B)
Colin Targonski (MSc, , ECE)
Leland Dunwoodie (<MD, G&B)
Olivia Feltus (<BSc, Intern)
Nick Watts (Progarrmer)
Zach Gerstner (<MS, BDSI)
**Jack Fletcher (<Bsc, REU)*
**Kim Roche (<PhD, CCIT, G&B)*
**Brittany Rosener (BSc, G&B)*
**Michael Sullivan (<BSc, G&B)*
**Henry Randall (<Bsc,BioE)*
**Keerti Kosana(<BSc, CS)*

@ Clemson
Melissa Smith (ECE)
KC Wang (ECE/CCIT)
Walt Ligon (ECE)
Nick Mills (ECE)
Jon Calhoun (ECE)
Brian Dean (CS)
Marc Birtwistle (ChemE)
Julia Frugoli (G&B)
Suchitra Chavan (G&B)
Elsie Schnabel (G&B)
Susan Duckett (AVS)
Jessi Britt (AVS)
Markus Miller (AVS)
Stephen Kresovich (PES)
Zach Brenton (PES)
Corey Ferrier (CCIT)
Jim Pepin (CCIT)
Clemson Networking (CCIT)
Clemson CITI (CCIT)

@ Earth
Stephen Ficklin (WSU)
Josh Burns (WSU)
Tyler Biggs (WSU)
Dorrie Main (WSU)
Sook Jung (WSU)
Joe Breen (Utah)
Jill Wegrzyn (UCONN)
Meg Staton (UTK)
Jim Bottum (Internet2)
John Moore (Internet2)
Ana Hunsinger (Internet2)
Marvin Weinstein (Quantum Insights LLC)
Ken Matusow (Synergity)
Karan Sapra (nVidia)

@ Earth (cont.)
Mats Rynge (USC-OSG)
Bala Desinghu (U Chicago-OSG)
Andrew Paterson (UGA)
Claris Castillo (RENCI)
Ray Idaszak (RENCI)
Paul Ruth (RENCI)
Michael Stealy (RENCI)
Fan Jiang (RENCI)
Mert Cevik (RENCI)
Ananya Mukherghee (RENCI)
Emily Casanova (USC-GHS)
Manual Casanova (USC-GHS)
Alex Bowers (Columbia U.)
Josh Vandenbrink (Louisiana Tech)
Ann Loraine (UNCC)
Colleen Doherty (NCSU)
John Graham (UCSD)
Wallace Chase (REANNZ)
Christos Papadopoulos (CSU)
Susmit Shannigrahi (Tennessee Tech)
Chengyu Fan (CSU)

USDA

DEPARTMENT OF ENERGY · UNITED STATES OF AMERICA ·

NSF

Many many more

renci

UCONN

NCBI

UNCG

UNC CHARLOTTE

UNIVERSITY OF CALIFORNIA SAN DIEGO

Thank You Funding Agencies!!!!!

- ***"CC*Data: National Cyberinfrastructure for Scientific Data Analysis at Scale (SciDAS)***

Source: NSF-CC* [1659300] (A. Feltus PI)

- ***"MCA-PGR: Spatial and Temporal Resolution of mRNA Profiles During Early Nodule Development."***

Source: NSF-PGRP [1444461] (J. Frugoli PI)

- ***"RCN: Advancing Research and Education Through a National Network of Campus Research Computing Infrastructures - The CaRC Consortium"***

Source: NSF [1620695] (J Bottum PI > A. Feltus PI)

- ***"MRI: Acquisition of a Cyberinstrument for Interdisciplinary Computational Science and Engineering."***

Source: NSF-MRI [1725573] (A. Apon PI)

<u>Historical</u>

- *"Tripal Gateway: Platform for Next-Generation Data Analysis and Sharing."*

Source: NSF-DIBBS [1443040] (S. Ficklin, PI)

- *"BIGDATA: F: DKM: Collaborative Research: PXFS: ParalleX Based Transformative I/O System for Big Data"*

Source: NSF-BIGDATA [1447771] (W. Ligon PI)

- *"Genomic and Breeding Foundations for Bioenergy Sorghum Hybrids."*

Source: Plant Feedstock Genomics for Bioenergy [DE-FOA-000041] (S Kresovich, PI).

- *"Big Data Visualization REU".*

Source: National Science Foundation [1359223](V Byrd, PI)

- *"MRI: Acquisition of a High Performance Computing Instrument for Collaborative Data-Enabled Science."* Source: National Science Foundation [1228312] (A Apon, PI)

- *"CC-NIE Integration: Clemson-NextNet"*

Source: National Science Foundation [1245936] (KC Wang, PI)

- *"Building non-model species genome curation communities."*

Source: National Evolutionary Synthesis Center (NESCent) (A Papanicolaou, PI)

- *"Big Data Analysis Tools for Agricultural Genomics*."

Source: Clemson University Experiment Station (USDA Hatch Project) [SC-1700492] (A. Feltus PI).

YYYYY

Change DNA and Change The Person: Coffee Example

TGTGGGCACAGGAC

TGTGGGCACAGGAC

The sequence of DNA contains information on how an organism responds to the world.

TGTGGGCCCAGGAC

TGTGGGCACAGGAC

TGTGGGCCCAGGAC

TGTGGGCCCAGGAC

	Genotype	What It Means
	AA	Fast caffeine metabolizer: drinking coffee didn't increase subjects' heart attack risk
	AC	Slow caffeine metabolizer: drinking coffee increased subjects' heart attack risk.
	CC	Slow caffeine metabolizer: drinking coffee increased subjects' heart attack risk.

https://www.23andme.com/health/Caffeine-Metabolism/