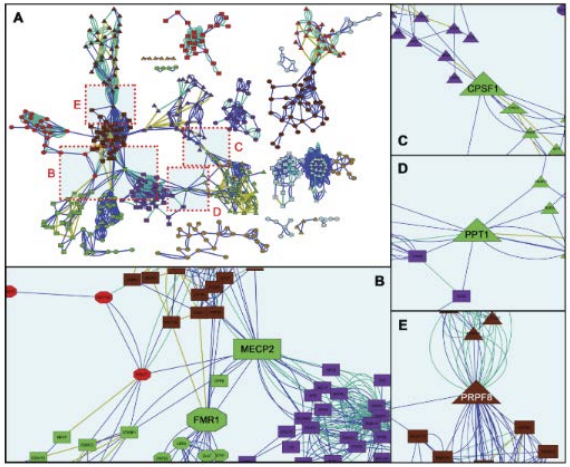
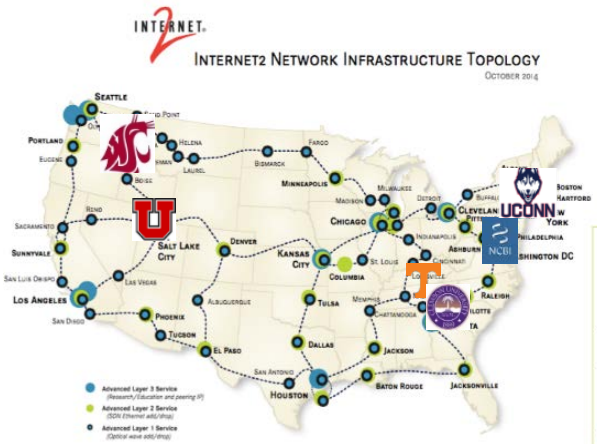
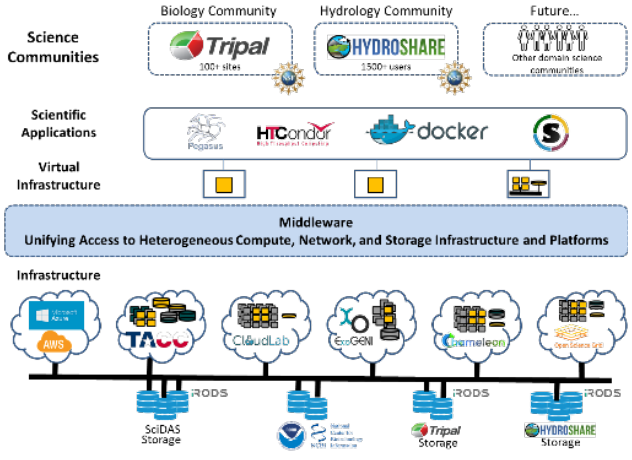
Mining Huge Collections of Genomics Datasets for Genes Controlling Complex Traits from Humans to Legumes

F. Alex Feltus, Ph.D.
Clemson Dept. of Genetics & Biochemistry (Associate Professor)
Allele Systems LLC (CEO)
Internet2 Board of Trustees (Member)
ffeltus@clemson.edu

OSG All Hands Meeting: 21 March 2018 @ 11am



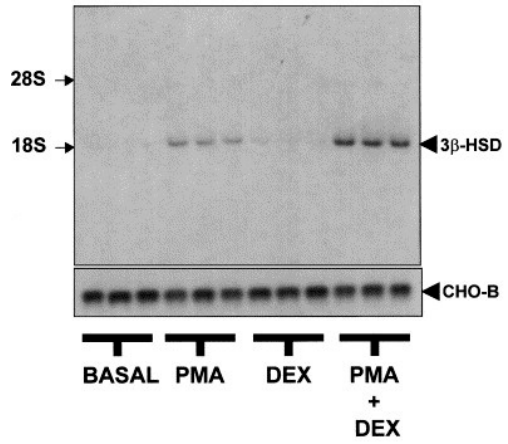




Core Principle of My Lab

Embrace Biological Complexity!
Holism > Reductionism

2x12 matrix

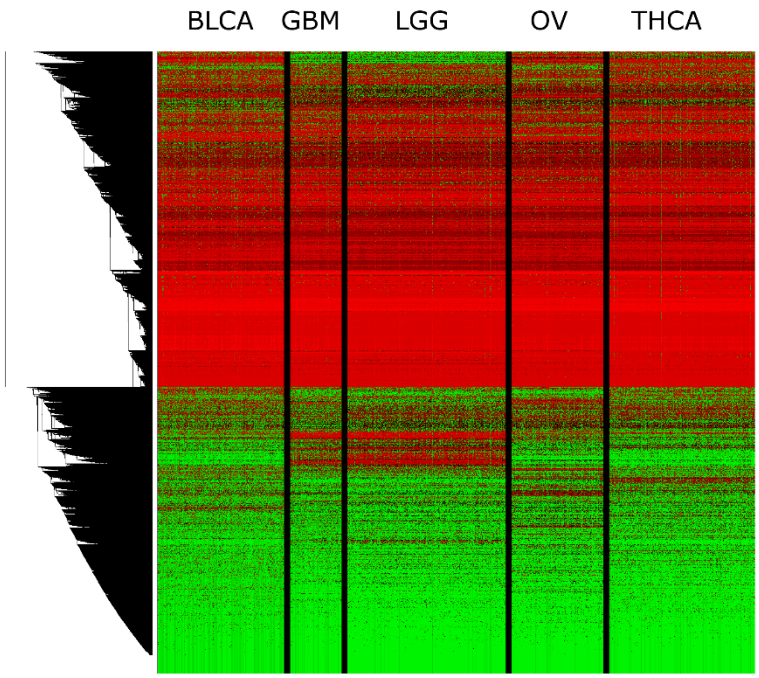


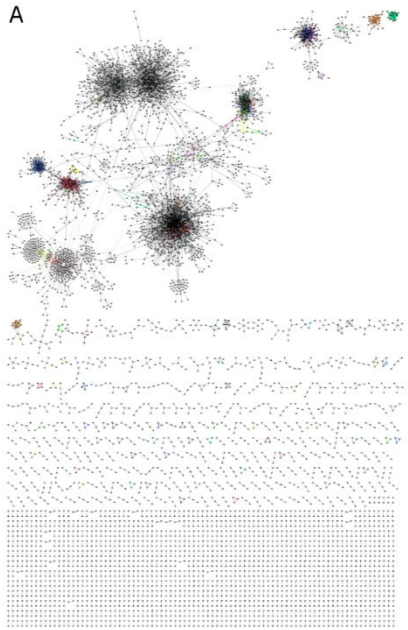
The Journal of Steroid Biochemistry and Molecular Biology
Volume 82, Issue 1, September 2002, Pages 55-63

Glucocorticoids enhance activation of the human type II 3β-hydroxysteroid dehydrogenase/Δ5–Δ4 isomerase gene

F.Alex Feltus [a, b], Stephanie Cote [c], Jacques Simard [c], Sebastien Gingras [d], William J. Kovacs [e, f], Wendell E. Nicholson [e, f], Barbara J. Clark [g], Michael H. Melner [a, b]

2016x73599 matrix





Discovering Condition-Specific Gene Co-Expression Patterns Using Gaussian Mixture Models: A Cancer Case Study

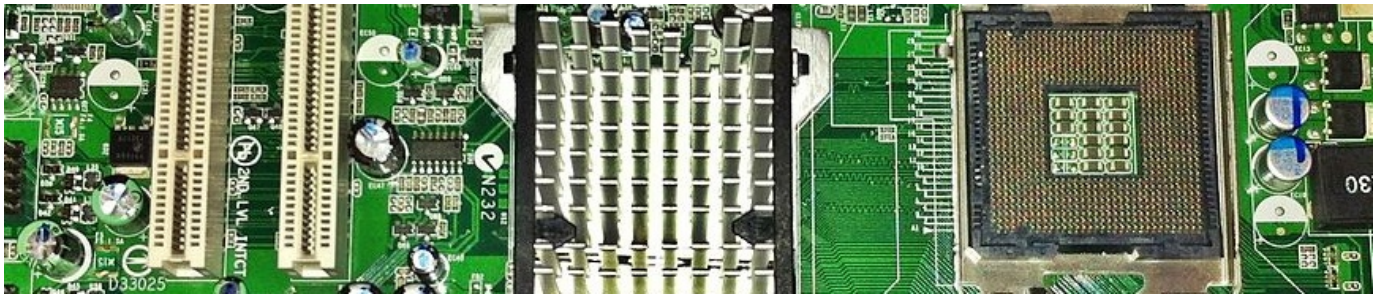
Stephen P. Ficklin, Leland J. Dunwoodie, William L. Poehlman, Christopher Watson, Kimberly E. Roche & F. Alex Feltus

Scientific Reports **7**, Article number: 8617 (2017)
Received: 12 April 2017
Accepted: 21 July 2017

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

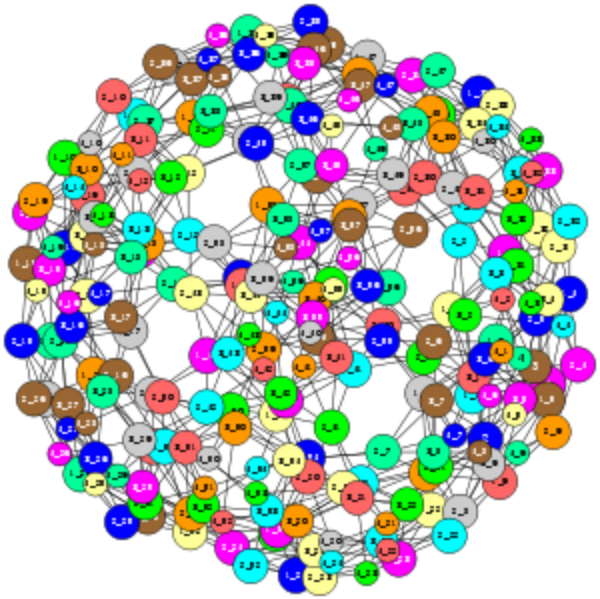
Vertebrates

Angiosperms



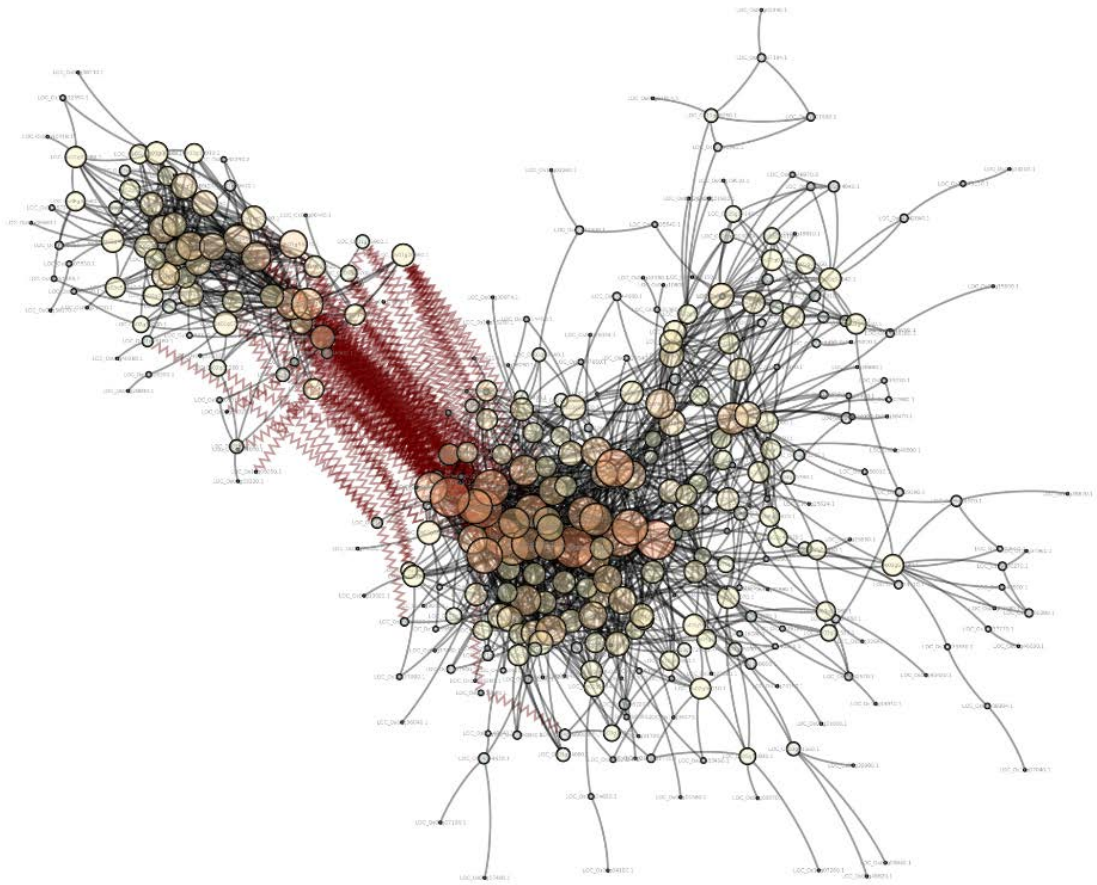
Bioinformatics/ Cyberinfrastructure

Gene Interaction Graphs:



NCBI: 4RHV Structure

Gene Co-Expression Networks (GCN)



- A.K.A Relevance Networks
- Network:
  - A graph
  - Qualitative model
- Nodes: gene products
- Edges: correlated expression
  - Positively correlated
  - Negatively correlated

Slide courtesy of Stephen Ficklin

My Lab's Core Workflow: Make GCNs From "all" RNAseq Data for a Species

0. Move public RNA datasets from NCBI & NIH. Mix with private data.

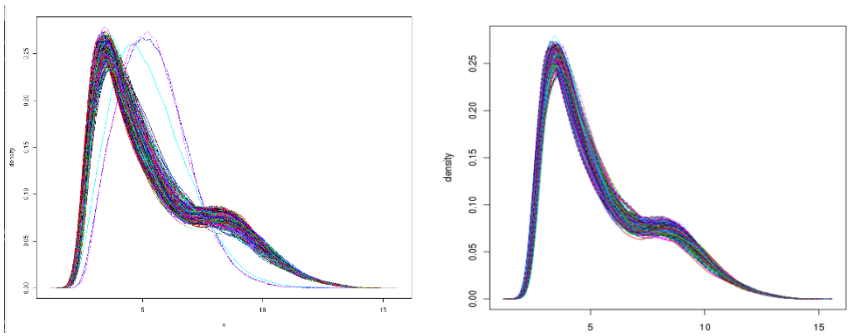
INTERNET2 · PRP PACIFIC RESEARCH PLATFORM · globus · aspera · iRODS

1. *n X m* Gene Expression Matrix (GEM) Construction.

Clemson Palmetto Cluster

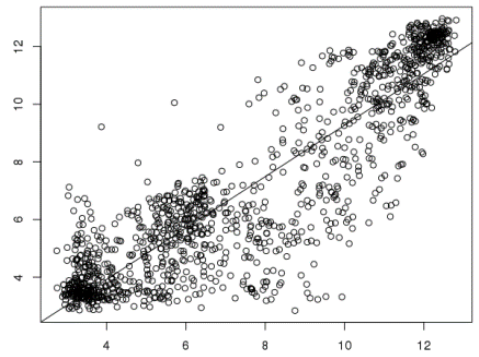
GeneID	ERR712353	ERR712355	ERR712362	ERR712366
GRMZM2G059856_T01.v6a	9.40	20.25	4.65	0.00
GRMZM5G811273_T02.v6a	0.00	0.38	0.00	0.45
GRMZM5G811273_T01.v6a	5.84	0.00	0.00	9.06
GRMZM2G059778_T01.v6a	0.45	1.65	0.00	0.88
GRMZM2G394757_T01.v6a	1.31	1.38	1.57	0.00
GRMZM2G093399_T01.v6a	9.00	1.08	1.00	0.65
GRMZM5G809743_T01.v6a	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

2. Normalization, Outlier removal



Clemson Palmetto Cluster

3. Pair-wise Correlation Analysis

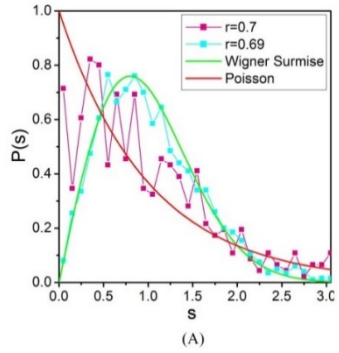


	GENE001	GENE002	GENE003	GENE004	GENE005	GENE006	GENE007	GENE008	GENE009	GENE010
GENE001	1.00									
GENE002	0.41	1.00								
GENE003	0.45	0.39	1.00							
GENE004	0.66	0.44	0.36	1.00						
GENE005	0.91	0.70	0.51	0.33	1.00					
GENE006	0.20	0.25	0.11	0.75	0.97	1.00				
GENE007	0.38	0.73	0.34	0.73	0.38	0.95	1.00			
GENE008	0.75	0.44	0.23	0.90	0.23	0.54	0.37	1.00		
GENE009	0.55	0.72	0.64	0.00	0.18	0.75	0.91	0.48	1.00	
GENE010	0.77	0.30	0.10	0.90	0.16	0.50	0.83	0.91	0.91	1.00

SciDAS · XSEDE

n x *n* similarity matrix
(n * (n-1)) / 2 comparisons

4. Significance Thresholding



Random Matrix Theory

Clemson Palmetto Cluster

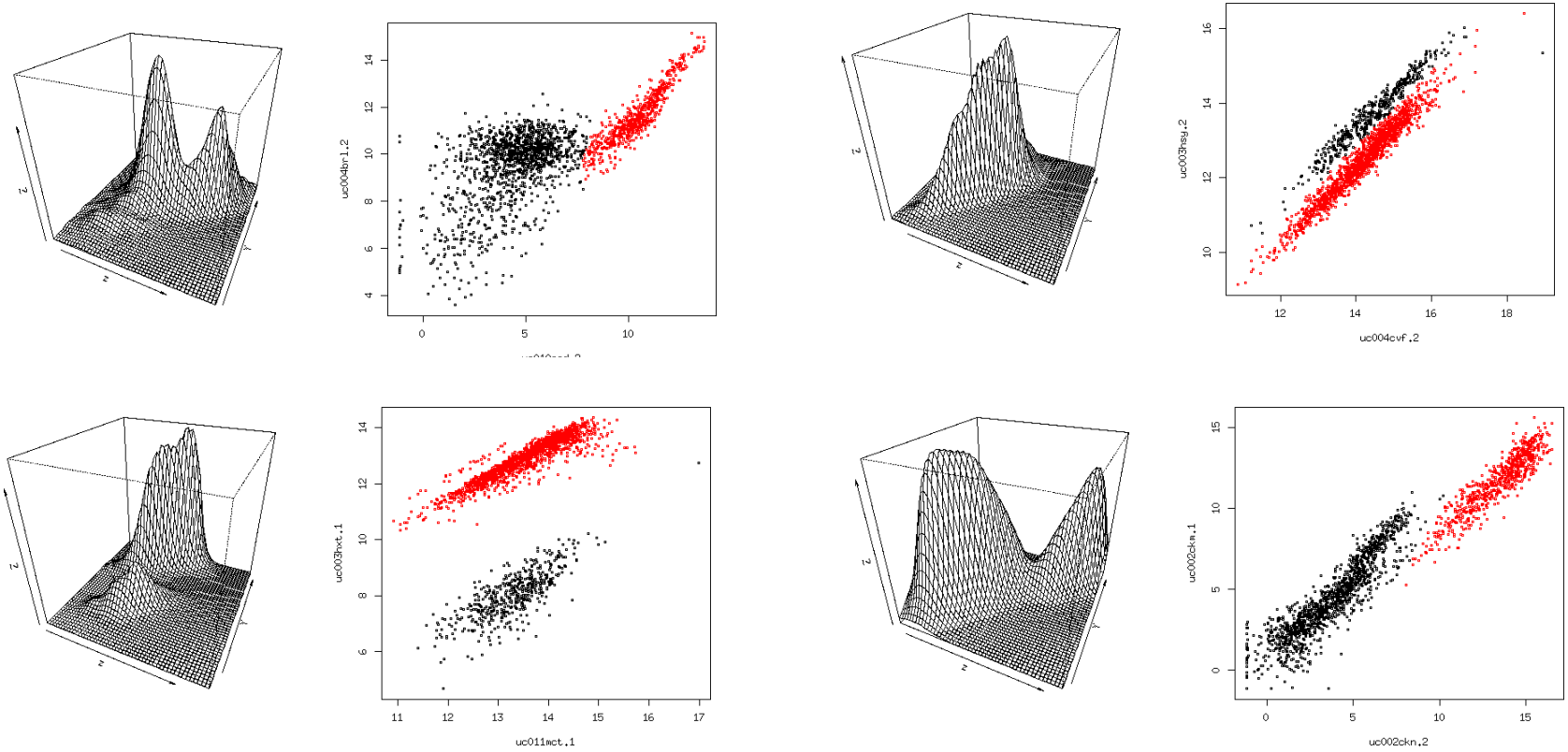
5. Gene Coexpression Network (GCN) Extraction

Clemson Palmetto Cluster

Current Approach: Gaussian Mixture Models (GMMs)

https://github.com/SystemsGenetics/KINC

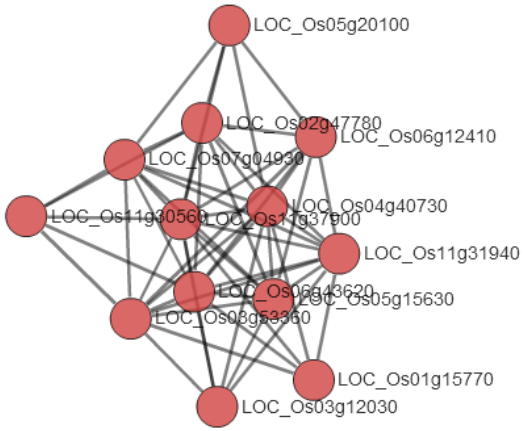
- Model data using a mixture of Gaussian distributions
- Identifies clusters in the data
- Clusters undergo separate correlation analysis.
- RMT-based significance thresholding.

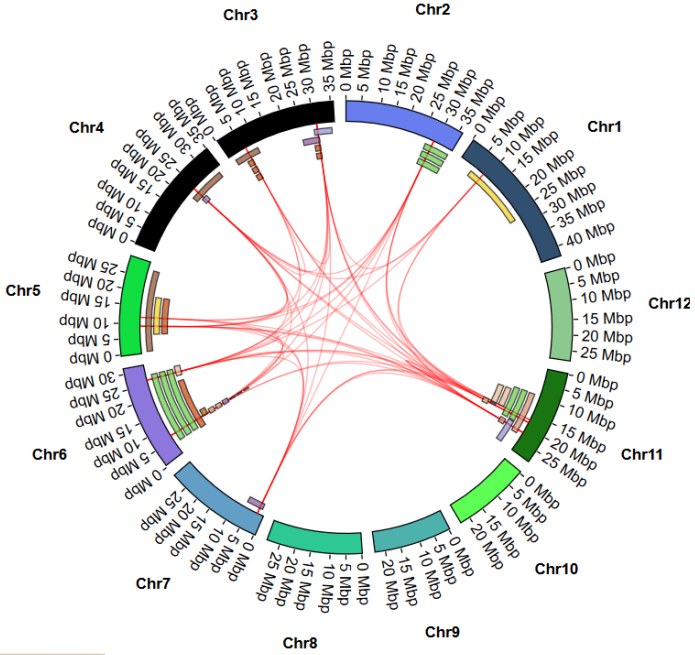


Slide courtesy of Stephen Ficklin

Genes Interact in Modules (complexity shards)

13 rice genes overlapping 1000-seed weight QTLs





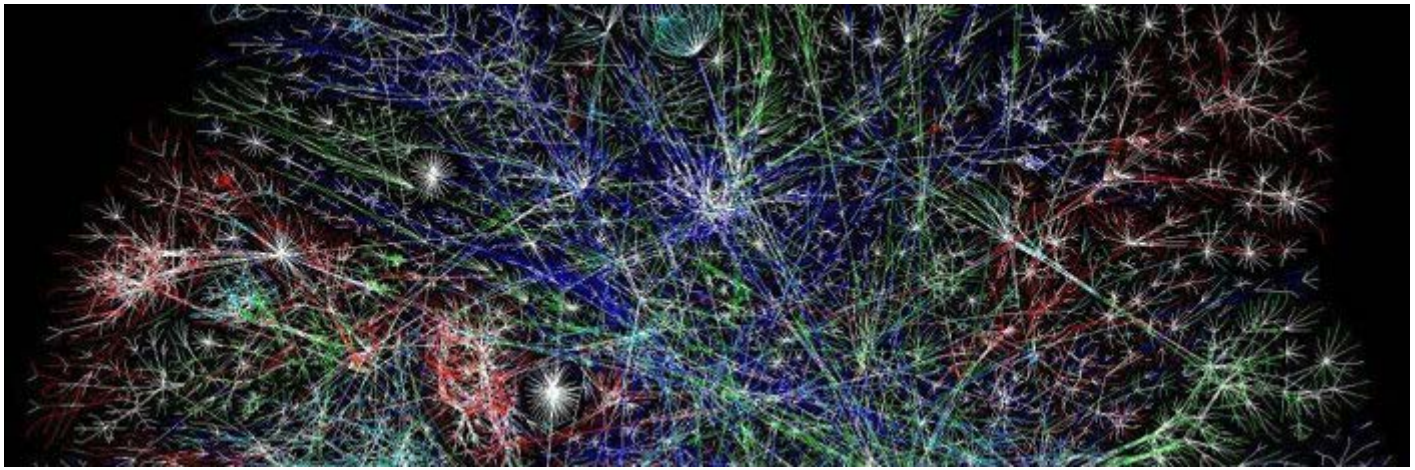
sysbio.genome.clemson.edu

Term	Definition	Fisher's p-value	Bonferroni	FDR	Nodes with Term
IPR012392	Very-long-chain 3-ketoacyl-CoA synthase	2.60e-5	4.67e-4	2.60e-5	LOC_Os03g12030, LOC_Os11g37900
IPR013601	FAE1/Type III polyketide synthase-like protein	3.81e-5	6.87e-4	4.04e-5	LOC_Os11g37900, LOC_Os03g12030
IPR012328	Chalcone/stilbene synthase, C-terminal	1.01e-4	1.82e-3	1.14e-4	LOC_Os03g12030, LOC_Os11g37900
IPR016038	Thiolase-like, subgroup	2.25e-4	4.05e-3	2.70e-4	LOC_Os03g12030, LOC_Os11g37900
IPR016039	Thiolase-like	2.52e-4	4.53e-3	3.23e-4	LOC_Os11g37900, LOC_Os03g12030
IPR002347	Glucose/ribitol dehydrogenase	4.77e-4	8.59e-3	6.61e-4	LOC_Os11g30560, LOC_Os04g40730
IPR002198	Short-chain dehydrogenase/reductase SDR	6.48e-4	1.17e-2	9.72e-4	LOC_Os04g40730, LOC_Os11g30560
IPR001087	Lipase, GDSL	7.61e-4	1.37e-2	1.24e-3	LOC_Os11g31940, LOC_Os06g12410
GO:0000003	reproduction	1.13e-3	1.36e-2	3.39e-3	LOC_Os06g12410, LOC_Os11g31940
GO:0005576	extracellular region	2.00e-3	2.40e-2	8.01e-3	LOC_Os11g31940, LOC_Os06g12410
IPR004254	Hly-III-related	2.50e-3	4.50e-2	5.63e-3	LOC_Os06g43620
IPR006634	TRAM/LAG1/CLN8 homology domain	4.06e-3	7.31e-2	1.04e-2	LOC_Os01g15770
IPR006459	Uncharacterised protein family UPF0497, trans-membrane plant subgroup	7.49e-3	1.35e-1	2.25e-2	LOC_Os05g15630
IPR016040	NAD(P)-binding domain	8.60e-3	1.55e-1	3.10e-2	LOC_Os04g40730, LOC_Os11g30560
IPR002123	Phospholipid/glycerol acyltransferase	9.04e-3	1.63e-1	4.07e-2	LOC_Os05g20100

CU PhD

Stephen P. Ficklin and **F. Alex Feltus**. *A Systems-Genetics Approach and Data Mining Tool For the Discovery of Genes Underlying Complex Tra in Oryza Sativa.* PloS ONE 8(7): e68551, 2013.

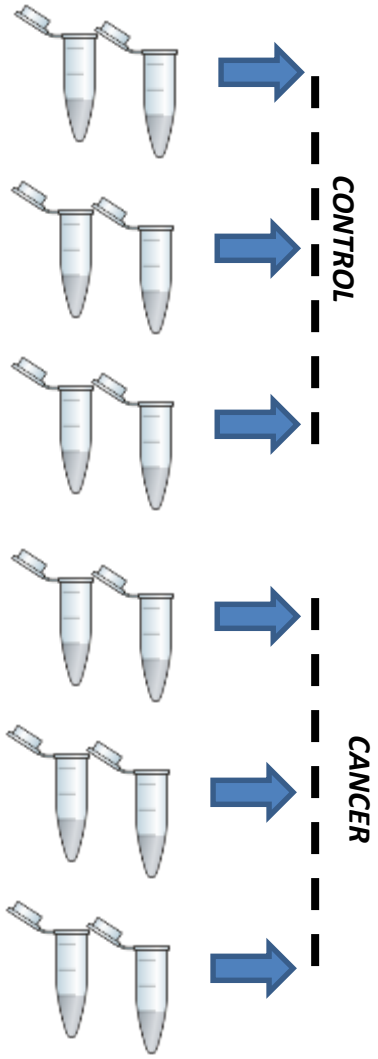
Bioinformatics ❤ Cyberinfrastructure



Bioinformatics is at the interface between biological measurement and result

Molecular Biology

BIOINFORMATICS



Patient RNA/DNA

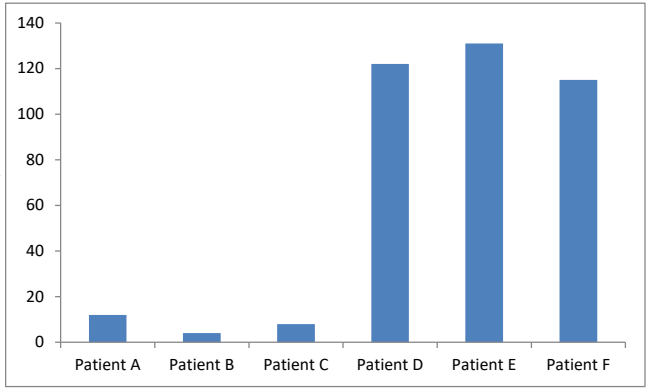
1/200 million records

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

```
00000001 tcaatgtagacagacgtcttttgaggttgtatccgctgc 00000039
>>>>>>>> ||||||||||||||||||||||||||||||||||||||| >>>>>>>>
41251797 tcaatgtagacagacgtcttttgaggttgtatccgctgc 41251835
```

DNA Sequencer

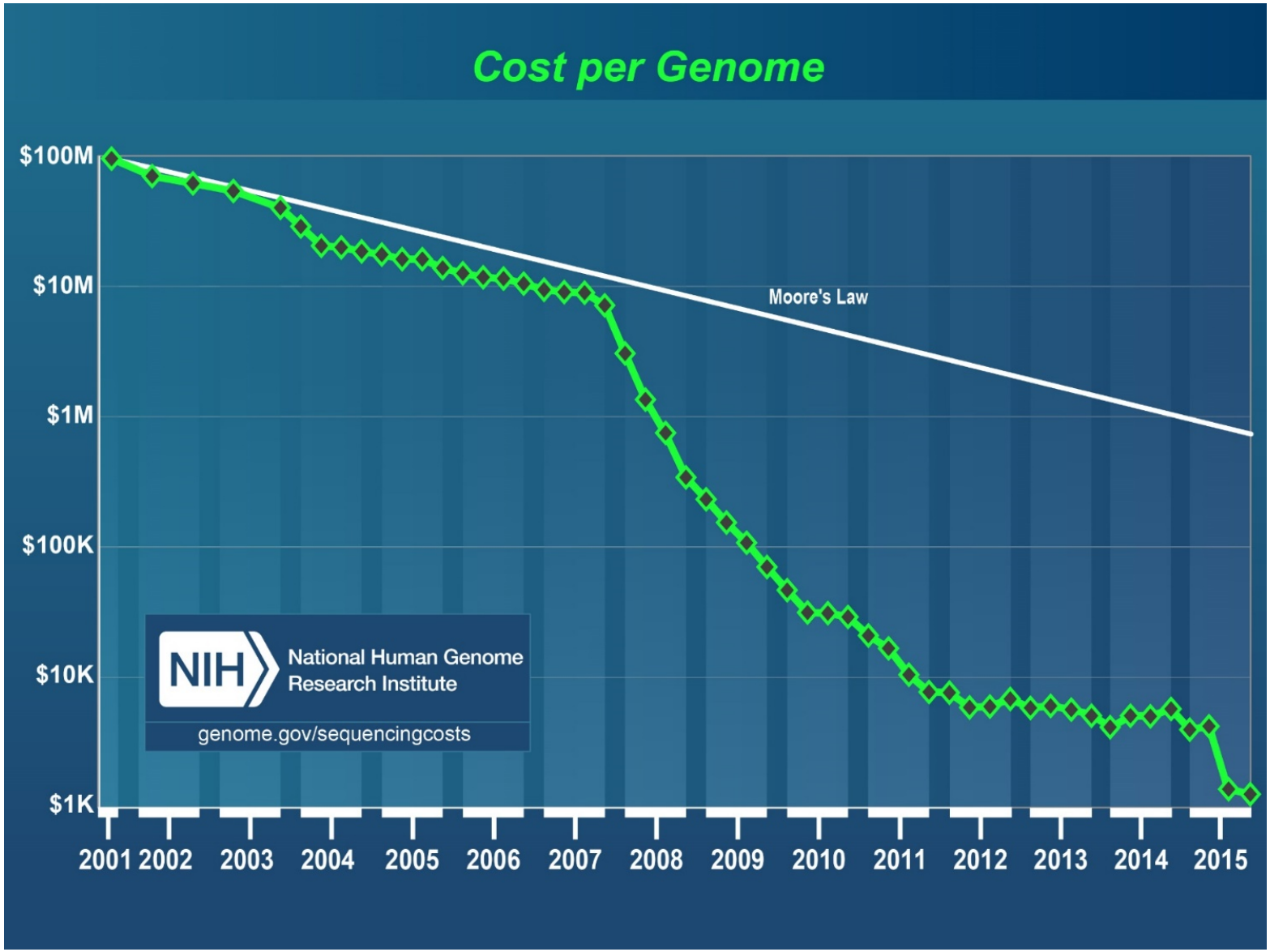
Supercomputer



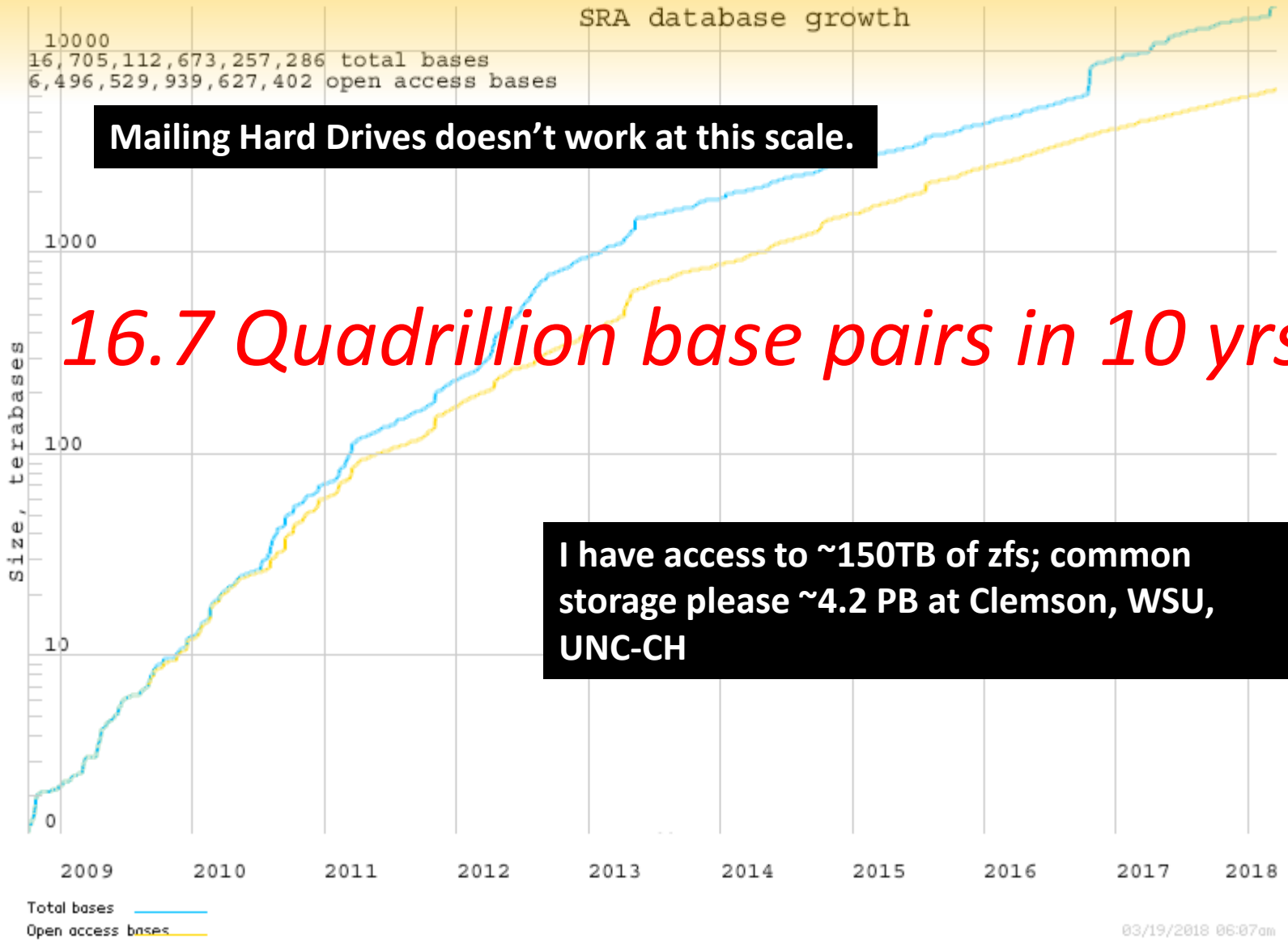
Excel Based Epiphany!

RNA/DNA Differences = Biomarkers!

DNA Sequencing Costs Dropping



Genomics is a Big Data Discipline



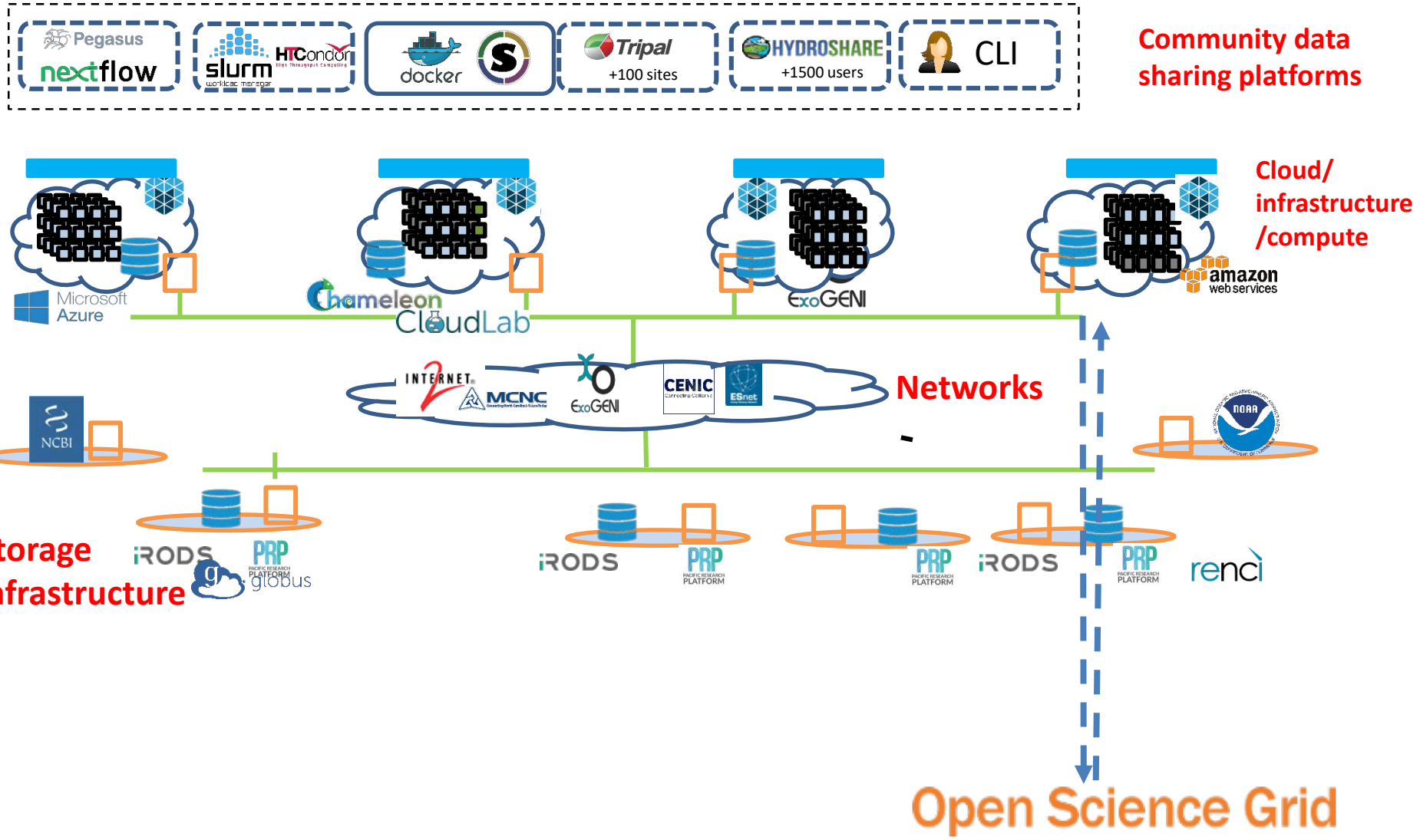
Mailing Hard Drives doesn't work at this scale.

16.7 Quadrillion base pairs in 10 yrs!

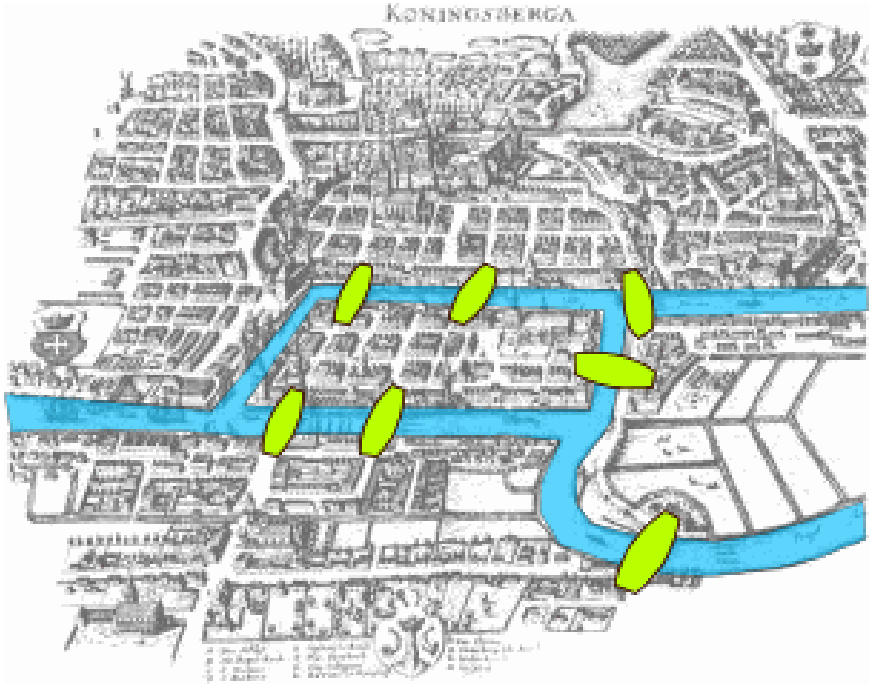
I have access to ~150TB of zfs; common storage please ~4.2 PB at Clemson, WSU, UNC-CH

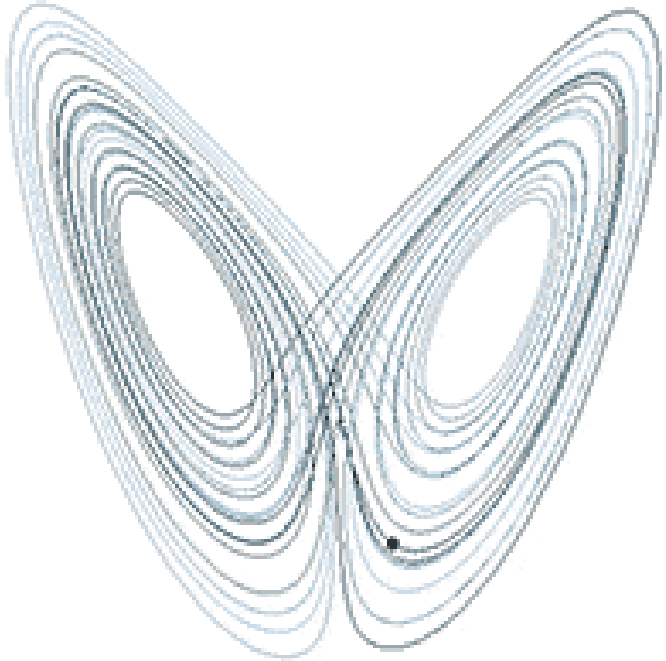
http://www.ncbi.nlm.nih.gov/Traces/sra/

SciDAS Ecosystem: CI, clouds and community platforms

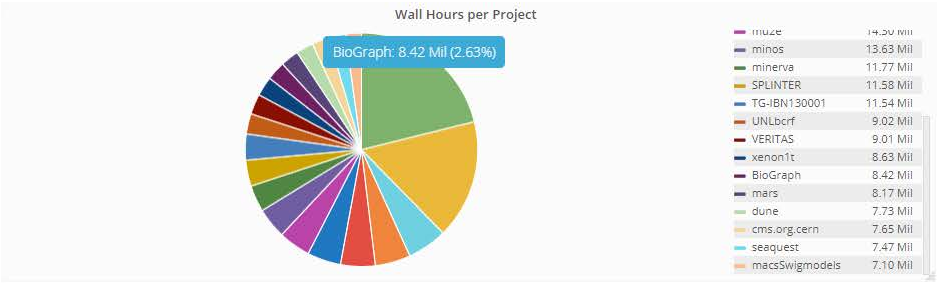


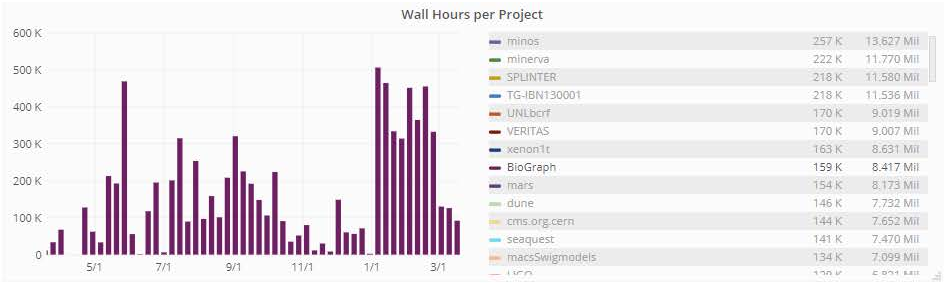
The OSG “Biograph” Project Aggregates and Processes Huge Datasets to Mine for Biological Solutions

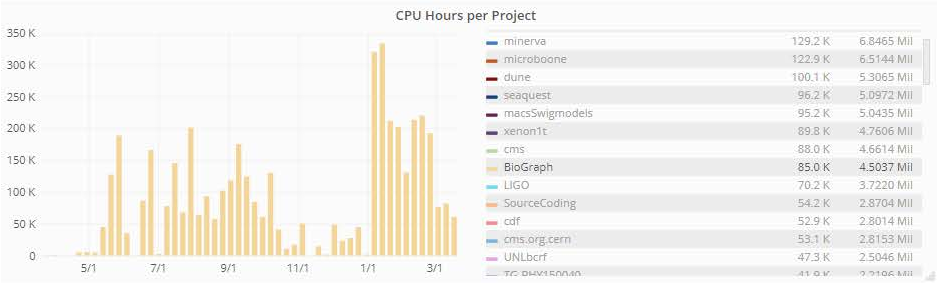


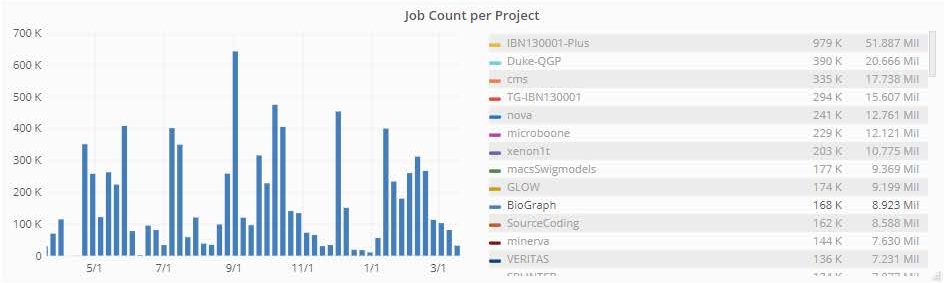


OSG Project “BioGraph” Usage: Exa-thanks to OSG!









<u>In the last year…</u>

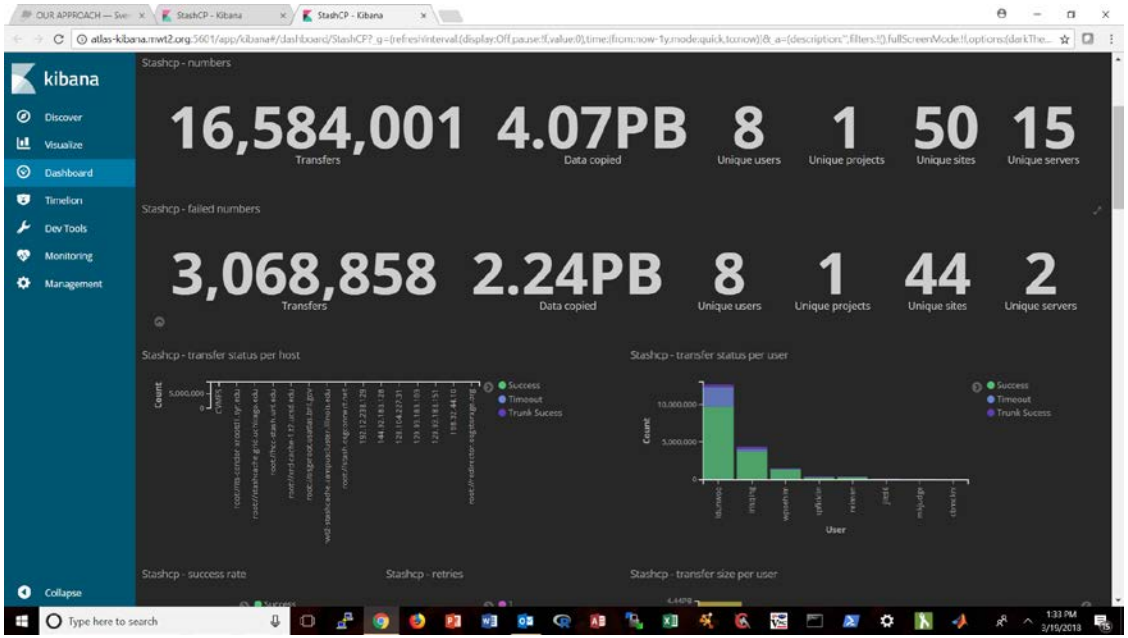
8.43 Million Wall Hours

4.50 Million CPU Hours

8.92 Million Jobs

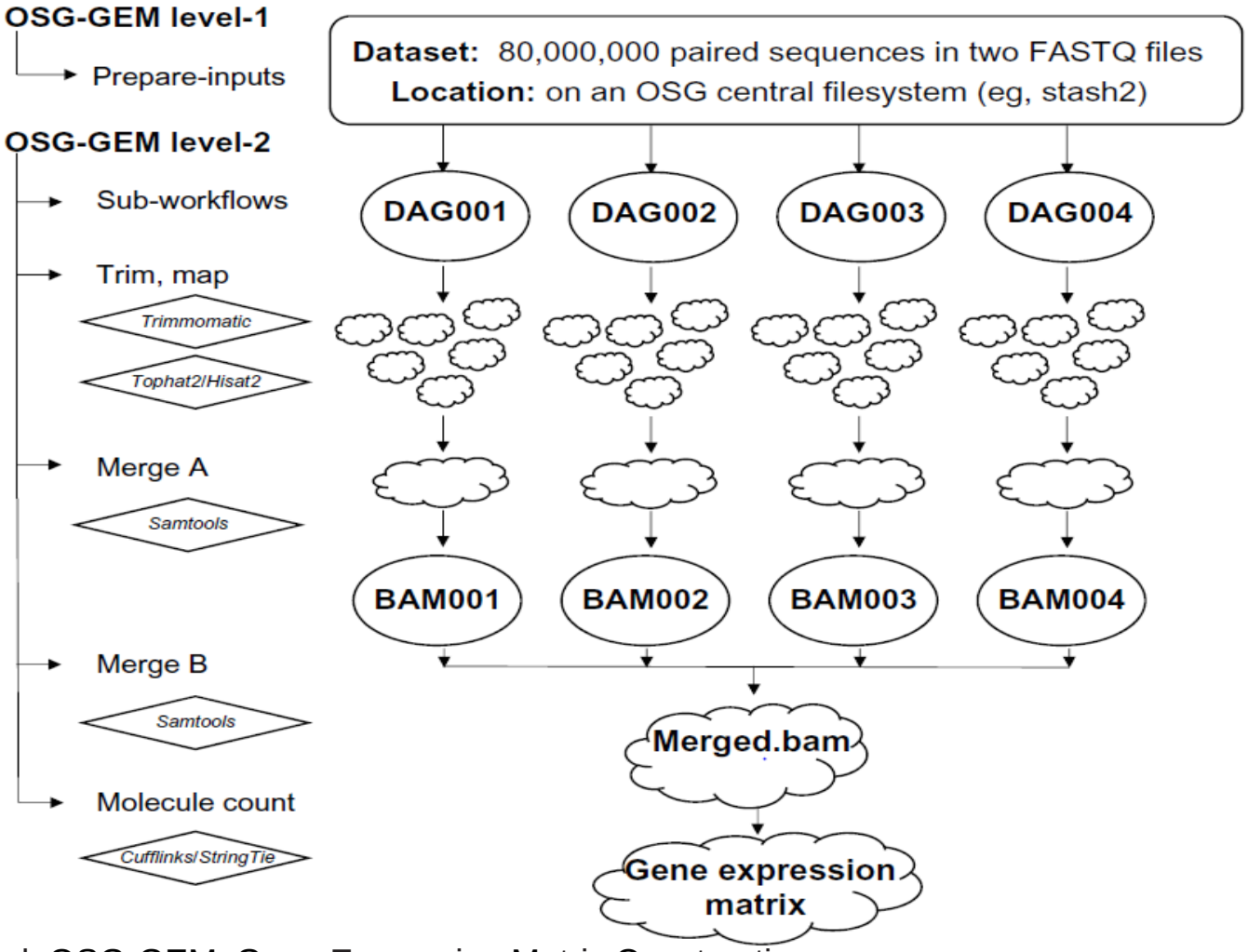
16.6 Million Transfers

4.07 PB



Open Science Grid Gene Expression Matrix Construction Workflow (OSG-GEM)

https://github.com/feltus/OSG-GEM



Poehlman et al. OSG-GEM: Gene Expression Matrix Construction Using the Open Science Grid. *Bioinformatics and Biology Insights* 2016:10 133–141 doi: 10.4137/BBI.S38193.

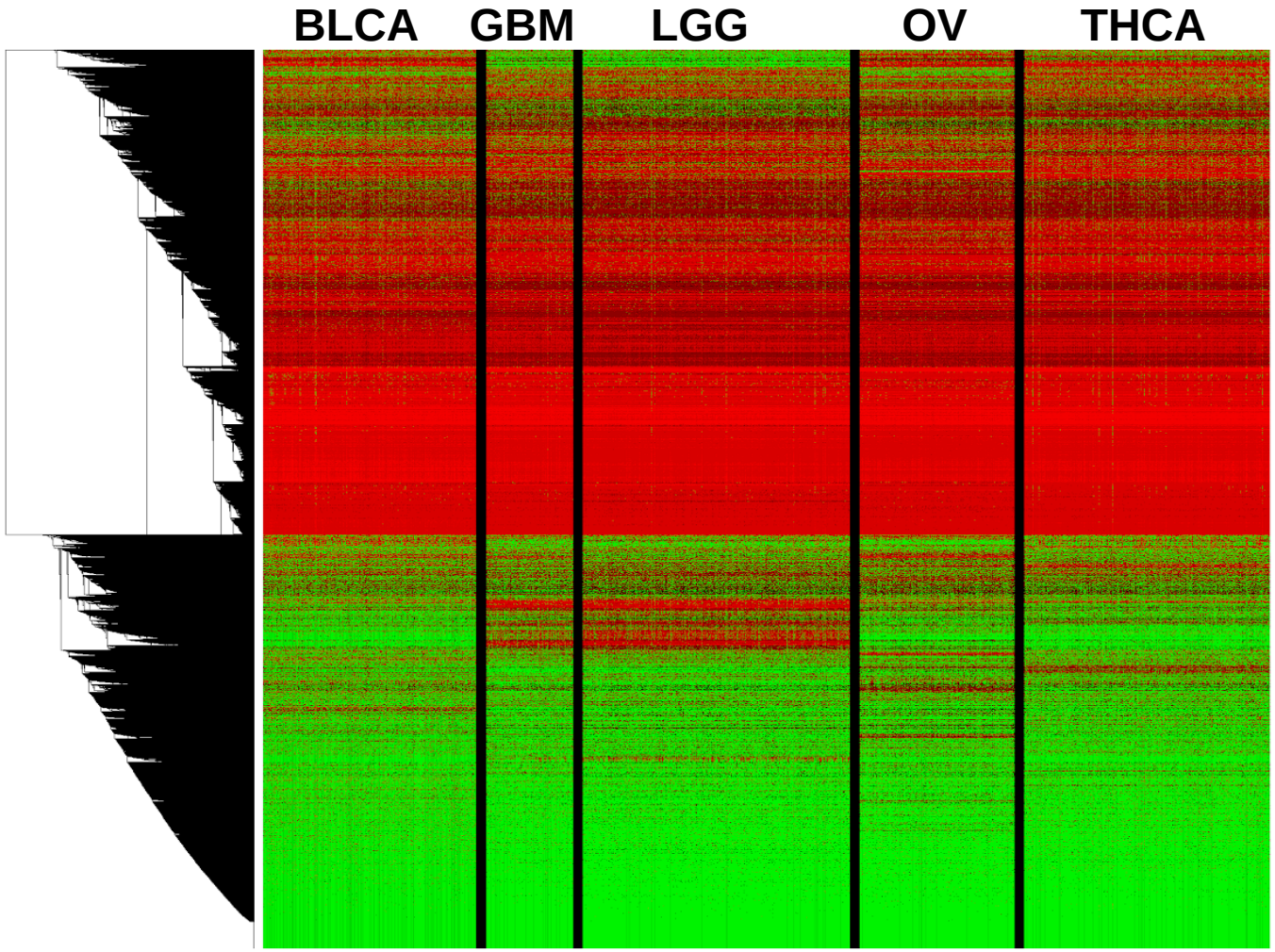
OSG-KINC: High-throughput gene co-expression network construction using the open science grid

https://github.com/feltus/OSG-KINC

1. OSG-KINC is an open source workflow that runs KINC on the Open Science Grid.
2. Builds Gene Co-expression Network (GCN) from an n X m Gene Expression Matrix GEM.
3. Instructions for Open Science Grid usage. Yeast unit test GEM included.
4. Users controls how many jobs are created. We typically run 100-200K.
5. iRODS support.

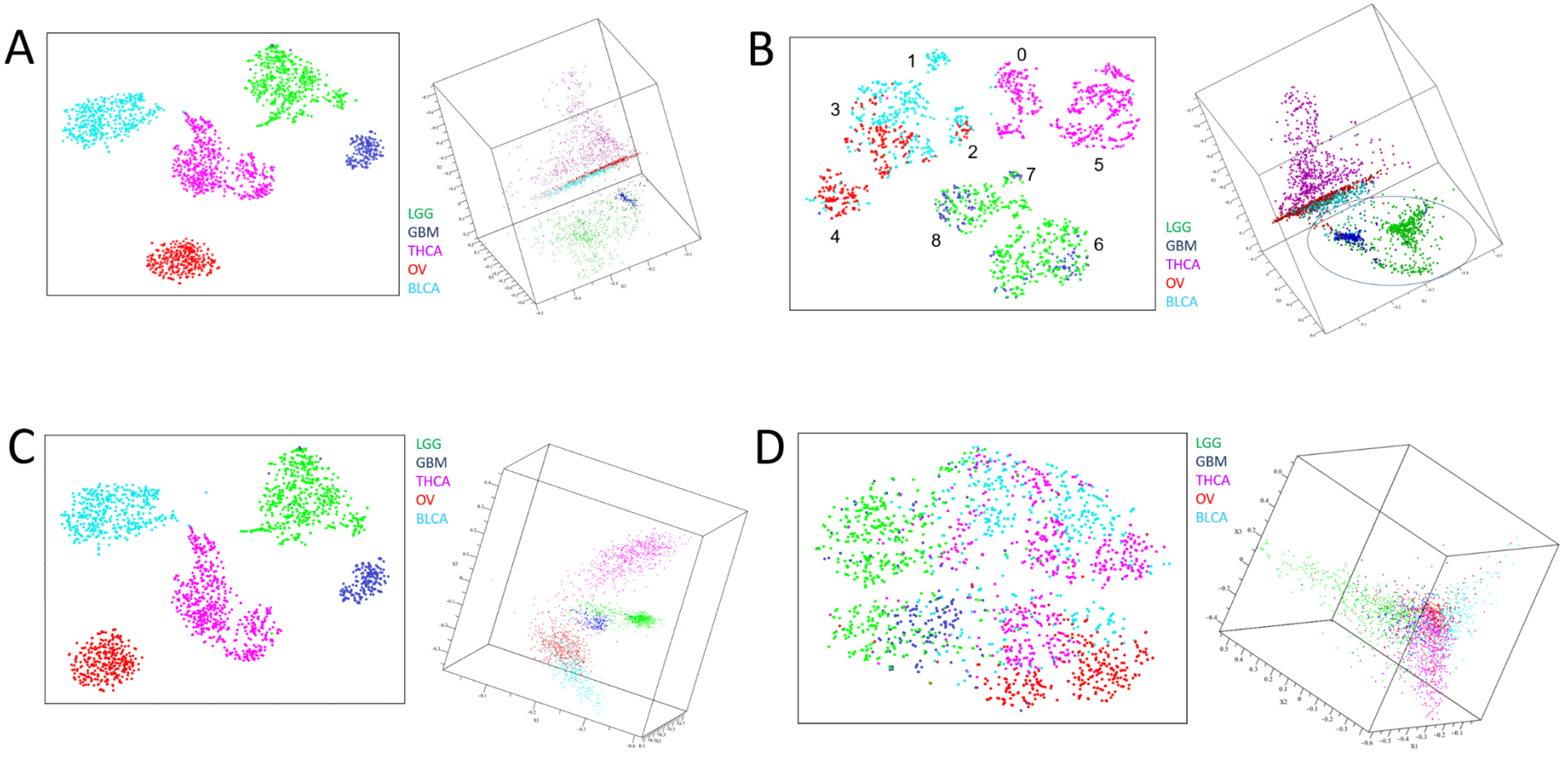
William L Poehlman, Mats Rynge, D Balamurugan, Nicholas Mills, Frank A Feltus. OSG-KINC: High-throughput gene co-expression network construction using the open science grid. Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference. 2017/11/13 (pp1827-1831).

OSG is Helping us Mine The Cancer Genome Atlas for Polygenic Biomarker Sets (2,016 tumors)



BLCA=bladder cancer (427 tumors), GBM=glioblastoma multiforme (174 tumors), LGG=low grade glioma (534 tumors), OV=ovarian cancer (309 tumors), THCA=thyroid carcinoma (572 tumors).

Tumor Classification Potential Revealed by t-Distributed Stochastic Neighbor Embedding (t-SNE) and Dynamic Quantum Clustering (DQC)



Sorting Five Human Tumor Types Reveals Specific Biomarkers and Background Classification Genes

Kimberly E. Roche, Marvin Weinstein, Leland Dunwoodie, William L. Poehlman, and Frank A. Feltus (In revision)

Quantum Insights

Edge Annotated Tumor Gene Co-expression Network

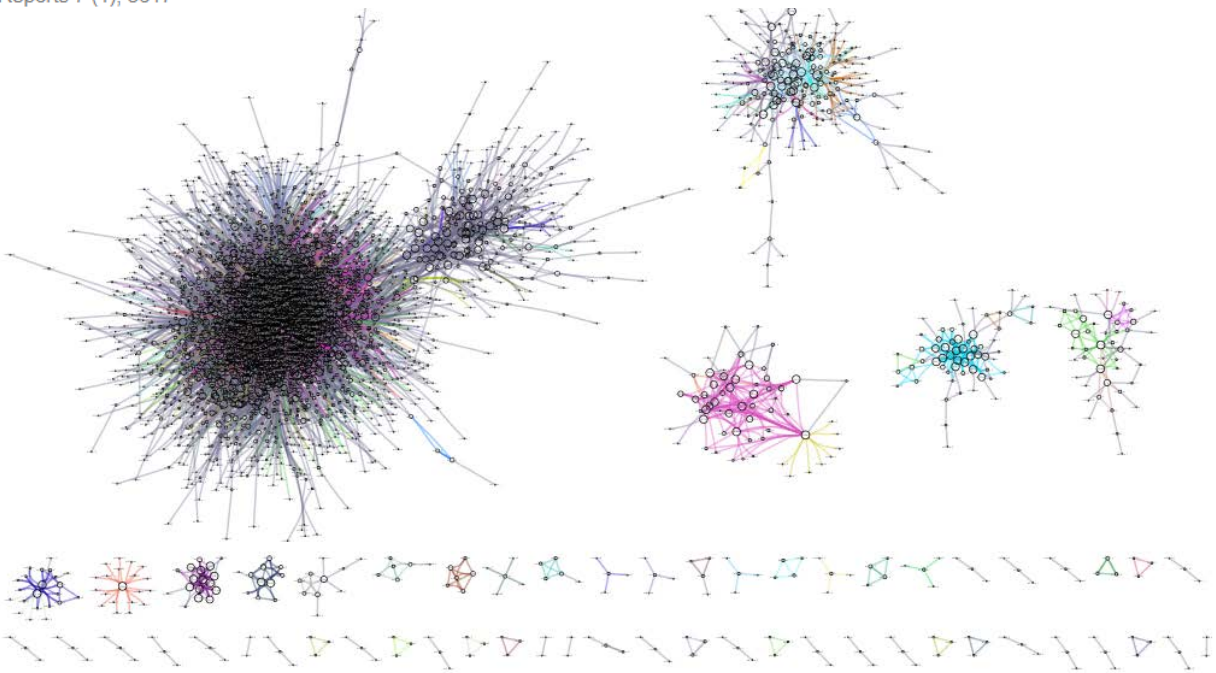
4,630 genes connected by 17,359 interactions

Discovering Condition-Specific Gene Co-Expression Patterns Using Gaussian Mixture Models: A Cancer Case Study

2017

SP Ficklin, LJ Dunwoodie, WL Poehlman, C Watson, KE Roche, FA Feltus

Scientific Reports 7 (1), 8617



Clemson Palmetto Cluster

Took Months to Process Datasets from 5 tumor Types

BLCA=bladder cancer (427 tumors),
GBM=glioblastoma multiforme (174 tumors),
LGG=low grade glioma (534 tumors),
OV=ovarian cancer (309 tumors),
THCA=thyroid carcinoma (572 tumors).

Stephen Ficklin,
Washington State University

Significant Clinical Annotation Enrichment in 375 Gene Modules

Cancer Types						
BLCA	OV	LGG	THCA	GBM		
13	15	32	9	18		
Gender						
Female	Male					
11	22					
Cancer Stage						
Stage I	Stage II	Stage III	Stage IV	Stage IVA	Stage IVC	
10	3	0	10	5	0	
*Ethnicity**						
NHL	HL	W	AA	A	NWPI	AIAN
2	3	22	0	6	0	0

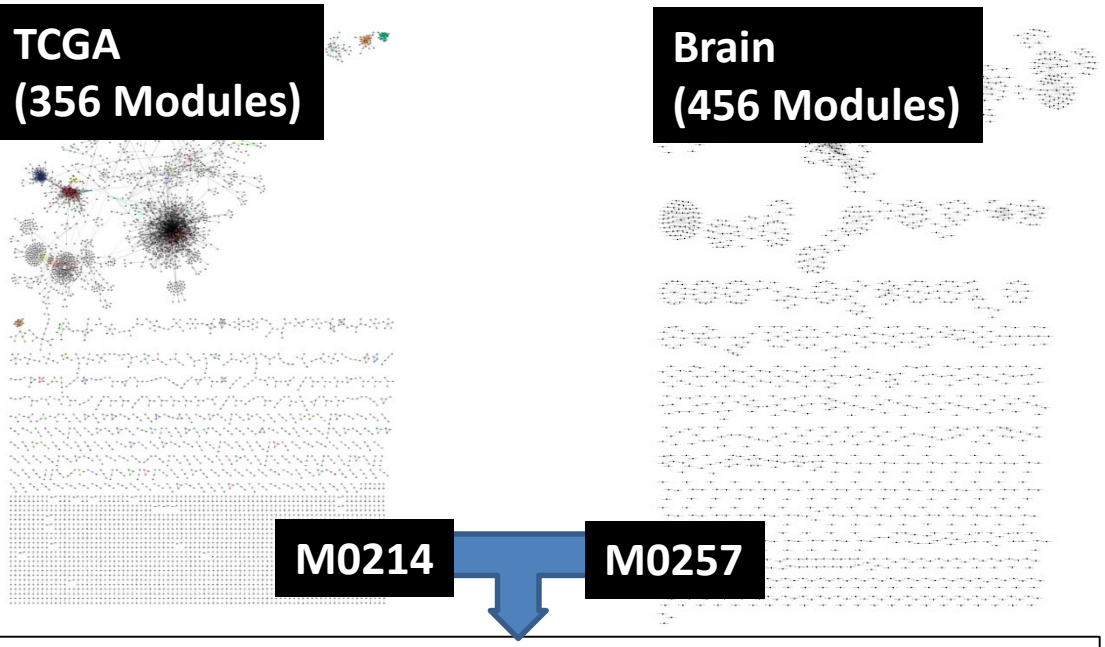
* Columns include: BLCA (bladder cancer), OV (ovarian cancer), LGG(lower grade glioma), THCA(thyroid cancer), GBM(glioblastoma), NHL (not Hispanic or Latino), HL (Hispanic or Latino), W (White), AA (African American), A (Asian), NHPI (Native Hawaiian or Pacific Islander), AIAN (American Indian, Alaska Native)

Cross-GCN Module Validation: A Glioblastoma Module

Brain (204 × 209086 GEM)
GBM (38); normal brain (138);
Brodmann's Area 9 of Parkinson's Disease patients (28)

TCGA (2016 x 73599 GEM)
BLCA=bladder cancer (427);
GBM=glioblastoma multiforme (174);
LGG=low grade glioma (534);
OV=ovarian cancer (309);
THCA=thyroid carcinoma (572)

Random (1793 × 209086 GEM)
Random human datasets(1793)



22 Genes Overlapping Between 2 GBM enriched modules:
TCGA M0214 →Brain M0257:::

ABI3, C1QA, C1QC, C3AR1, CD300A, CD86, FCER1G, FERMT3, GPR65, HAVCR2, ITGB2, LAPTM5, LY86, MYO1F, PARVG, RNASE6, SASH3, SIGLEC9, SPI1, TREM2, TYROBP, WAS

Clemson Palmetto Cluster

www.impactjournals.com/oncotarget/ Oncotarget, Advance Publications 2018

Discovery and validation of a glioblastoma co-expressed gene module

Leland J. Dunwoodie[1], William L. Poehlman[1], Stephen P. Ficklin[2] and Frank Alexander Feltus[1]

https://doi.org/10.18632/oncotarget.24228

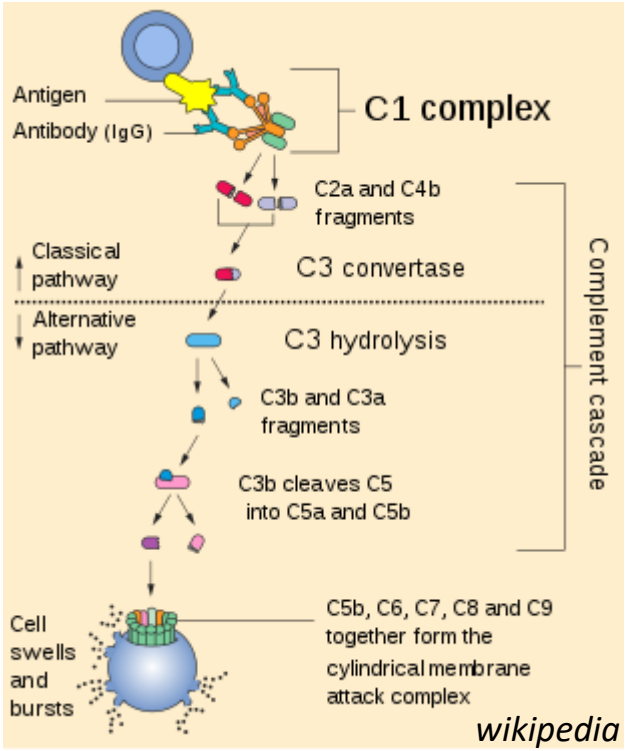
Glioblastoma Specific Module Contains Complement Immune Function

Gene Symbol	Gene Name	hg38 Ensembl ID
LAPTM5	lysosomal protein transmembrane 5	ENST00000294507
C1QA	complement C1q A chain	ENST00000374642
FCER1G	Fc fragment of IgE receptor Ig	ENST00000367992
C1QC	complement C1q C chain	ENST00000374639
CD86	CD86 molecule	ENST00000330540
HAVCR2 (TIM-3)	hepatitis A virus cellular receptor 2	ENST00000307851
LY86	lymphocyte antigen 86	ENST00000379953
TREM2	triggering receptor expressed on myeloid cells 2	ENST00000373113
FERMT3	fermitin family member 3	ENST00000345728
SPI1	Spi-1 proto-oncogene	ENST00000378538
C3AR1	complement C3a receptor 1	ENST00000307637
GPR65	G protein-coupled receptor 65	ENST00000267549
RNASE6	ribonuclease A family member k6	ENST00000304677
ABI3	ABI family member 3	ENST00000225941
CD300A	CD300a molecule	ENST00000360141
TYROBP	TYRO protein tyrosine kinase binding protein	ENST00000262629
SIGLEC9	sialic acid binding Ig like lectin 9	ENST00000250360
MYO1F	myosin IF	ENST00000613525
ITGB2	integrin subunit beta 2	ENST00000397852
PARVG	parvin gamma	ENST00000356909
WAS	Wiskott-Aldrich syndrome	ENST00000376701
SASH3	SAM and SH3 domain containing 3	ENST00000356892

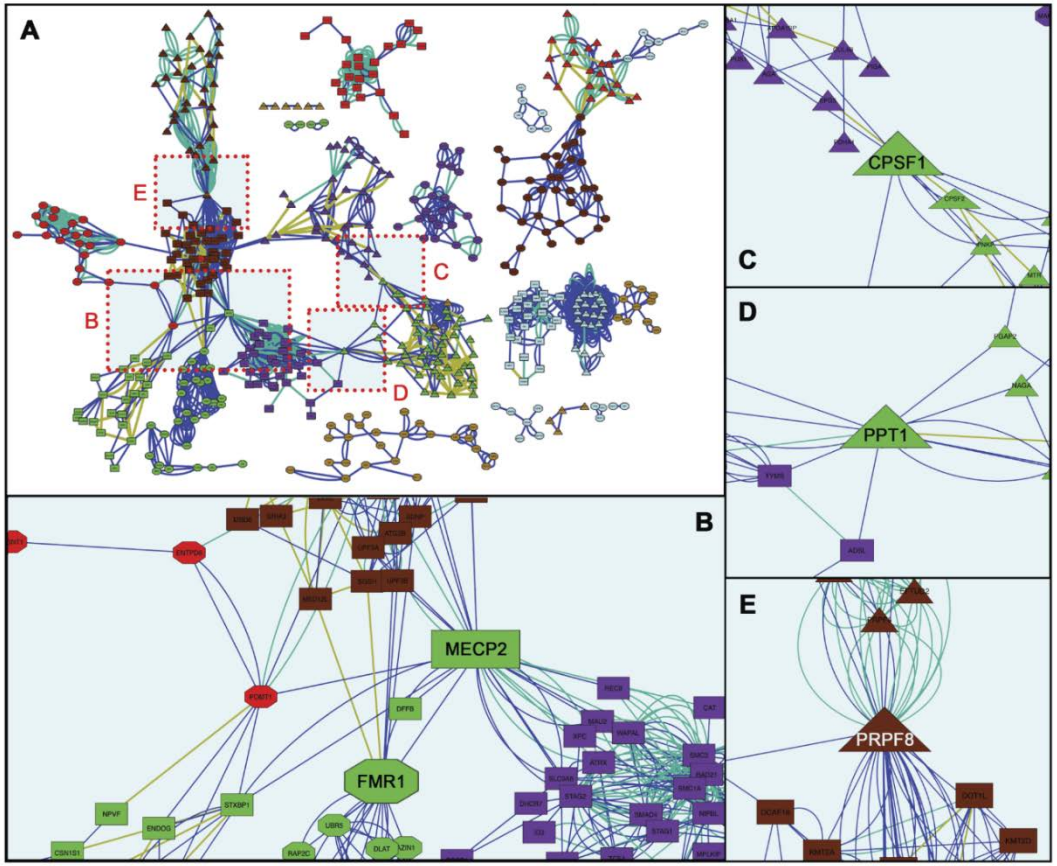
Some Enriched Functions in the Module

KEGG	hsa05322	Systemic lupus erythematosus
MIM	120575	COMPLEMENT COMPONENT 1, q SUBCOMPONENT, C CHAIN
PFAM	PF00386	C1q is a subunit of the C1 enzyme complex that activates the serum complement system.
PFAM	PF01391	Members of this family belong to the collagen superfamily.
PFAM	PF07686	This domain is found in antibodies as well as neural protein P0 and CTL4 amongst others.
REACTOME	R-HSA-173623	Classical antibody-mediated complement activation
REACTOME	R-HSA-198933	Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell
REACTOME	R-HSA-166663	Initial triggering of complement

(adj. $p < 0.001$)



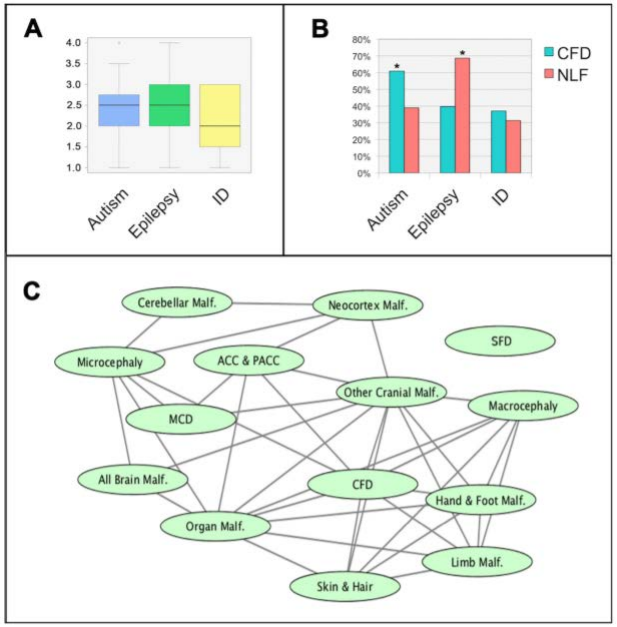
OSG is Helping us Understand How Intellectual Disability (ID) Genes Interact in Multiple Phenotype Contexts



Node Shapes:
- autism group
- epilepsy group
- ID group

Node Colors:
- CFD
- CFD/NLF
- NLF
- SFD
- SFD/NLF
- NONE

Edge Colors:
- Co-expression
- Genetic interaction
- Physical interaction



Abbreviations: intellectual disability (ID); complex facial dysmorphisms (CFD); simple facial dysmorphisms (SFD); neurodegenerative-like features (NLF); multiple congenital anomalies (MCA); upper motor neuron disease (UMND); multiple movement disorders (MMD); protein-protein interaction (PPI)

Emily Casanova, Greenville Health System

Widespread Genotype-Phenotype Correlations in Intellectual Disability

Emily L Casanova, Zachary Gerstner, Julia L. Sharp, Manuel F. Casanova, F. Alex Feltus (2018) bioRxiv; in review

OSG is helping us find genes in beans that help plants make their own fertilizer via bacterial symbiosis

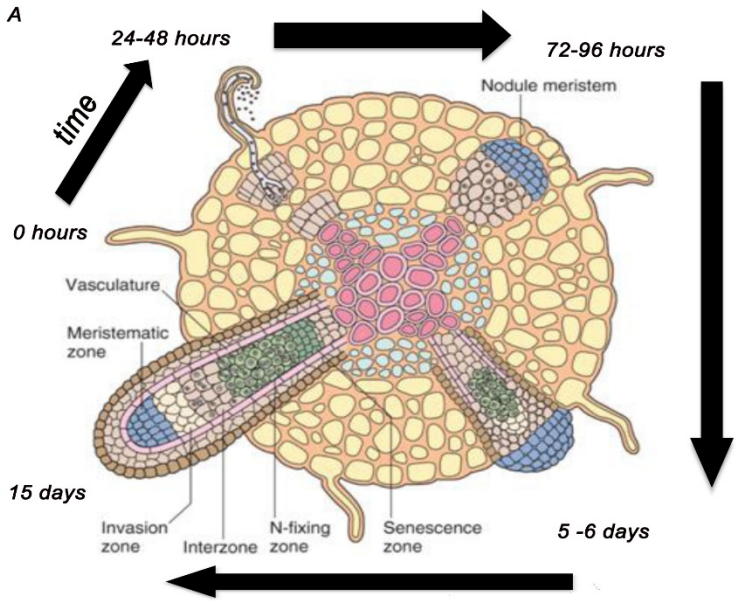
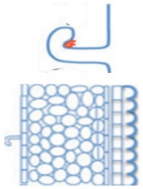
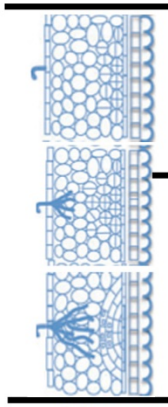
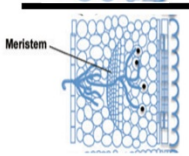
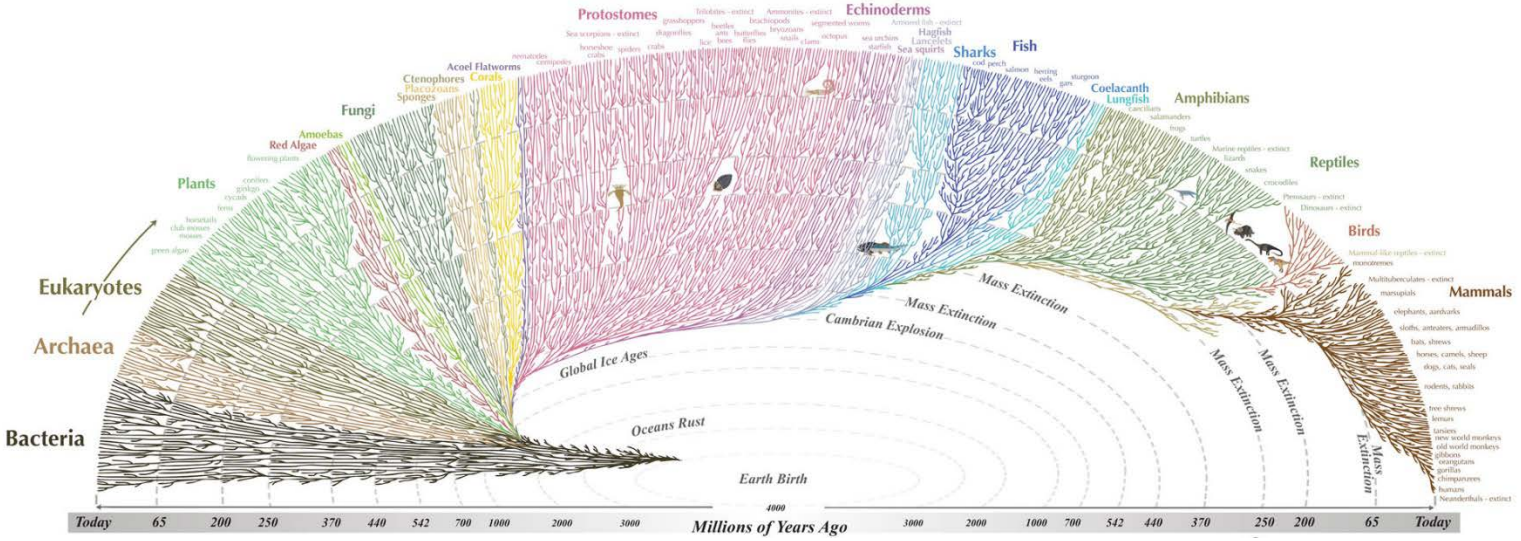


diagram*	time post innoculation	known infection events/ known diffentially regulated genes	tissues	replicates & controls	total libraries
	0	none/ none	epidermis outer cortex inner cortex (x) inner cortex (nx) pericycle phloem xylem	3 biological replicates	21 + 2 (2 whole root libraries inoculated and mock)
	12 hours	root hair tip assymetric growth; deformation 1st cell divisions begin inner cortex/ RIP1, ENOD40	epidermis outer cortex inner cortex (x) inner cortex (nx) pericycle phloem xylem	3 biological replicates (inoculated) & 3 mock inoculation controls	42 +2 (2 whole root libraries inoculated and mock)
	24 hours	root hair curling; bacterial microcolony formation; cell division/ ENOD11, ENOD20	epidermis outer cortex inner cortex (x) inner cortex (nx) pericycle phloem xylem	3 biological replicates (inoculated) & 3 mock inoculation controls	42 +2 (2 whole root libraries inoculated and mock)
	48 hours	cell division infection thread formation & branching/ CYC2, MtN6	epidermis outer cortex inner cortex (x) inner cortex (nx) pericycle phloem xylem	3 biological replicates (inoculated) & 3 mock inoculation controls	42 +2 (2 whole root libraries inoculated and mock)
	72 hours	meristem established single infection thread enters nodule CLE12, CLE13	epidermis outer cortex inner cortex (x) inner cortex (nx) pericycle phloem xylem	3 biological replicates (inoculated) & 3 mock inoculation controls	42 +2 (2 whole root libraries inoculated and mock)

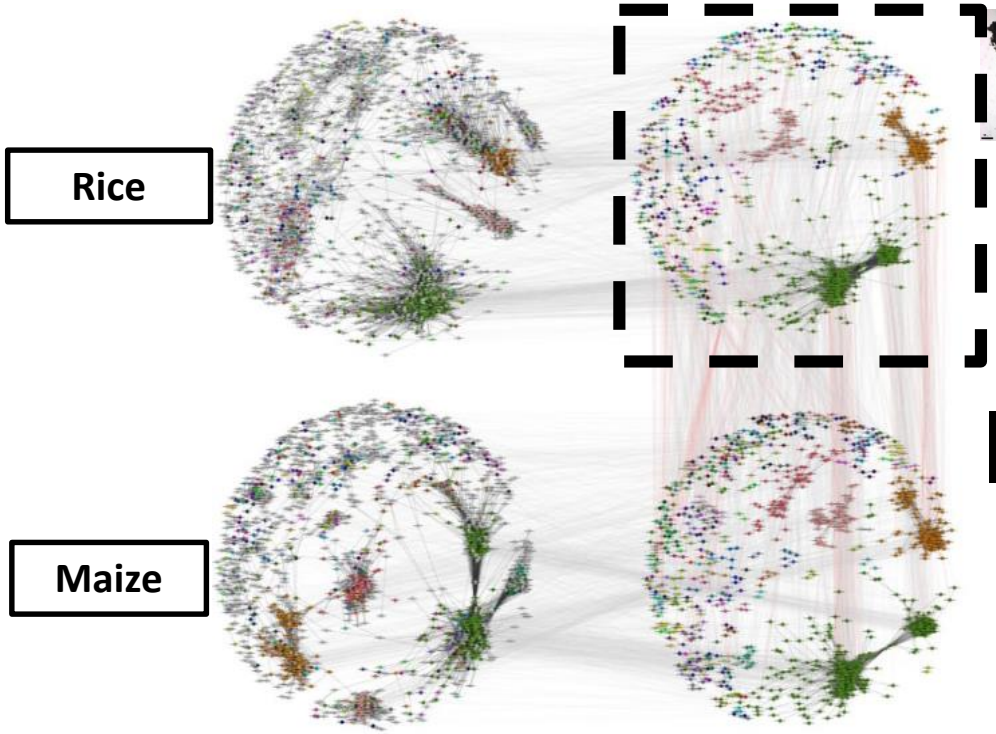
Julia Frugoli, Clemson Genetics & Biochemistry

lasernode.org

OSG is helping us reconstruct the ancestral gene interaction networks for 100s of species



https://www.evogeneao.com/learn/tree-of-life



Ancestral Paleogenomic Fossil Interactions (60-80 million years old)

Stephen Ficklin, Washington State University

Summary

1. **OSG has allowed me to scale up my science. We are just getting started.**
2. **OSG-GEM, OSG-KINC Pegasus workflows are in Github and open source!**
3. **The BioGraph project is using OSG to**
   - **Identify gene interactions in plants and animals on a massive scale (in progress)**
   - **Characterize genes that are specific to the tumor subtypes (e.g. glioblastoma 22-gene module).**
4. **OSG is helping us flock out of the SciDAS cloud onto OSG. All SciDAS infrastructure will be open source.**

OSG Rulz!

Geographically Distributed Interdisciplinary Science is Super Fun!

Feltus Lab

Will Poehlman (<PhD, G&B)
Yuqing Hang (<PhD, G&B)
Benafsh Husain (<PhD, BDSI)
Leland Dunwoodie (<BSc, G&B)
Rachel Eimen (<Bsc, ECE)
Henry Randall (<Bsc, Bioengineering)
Courtney Shearer (<BSc, CS)
Cole McKnight (<Bsc, CS)
Michael Sullivan (<BSc, G&B)
Jordan Little (<BSc, G&B)
Melissa Judge (<BSc, Bioengineering)
Keerti Kosana(<BSc, CS)
**Allison Hickman (G&B)*
**Olivia Feltus (<BSc, Intern)*
**Nick Watts (Programmer, CCIT)*
**Zach Gerstner (<BSc, Microbiology)*
**Jack Fletcher (<Bsc, REU)*
**Kim Roche (CCIT, G&B)*
**Brittany Rosener (<BSc, G&B)*
**Recent alumni*

@ Clemson

Karan Sapra (ECE)
Melissa Smith (ECE)
Ben Shealy (ECE)
Colin Targonski (ECE)
KC Wang (ECE/CCIT)
Walt Ligon (ECE)
Nick Mills (ECE)
Brian Dean (CS)
Jim Bottum (ECE/Internet2)
Brian Atkinson (ECE)
Susan Duckett (AVS)
Jessi Britt (AVS)
Markus Miller (AVS)
Stephen Kresovich (PES)
Zach Brenton (G&B)
Julia Frugoli (G&B)
Suchitra Chavan (G&B)
Elsie Schnabel *G&B)
Wallace Chase (CCIT)
Becky Ligon (CCIT)
Randy Martin (CCIT)
Corey Ferrier (CCIT)
Jim Pepein (CCIT)
Wallace Chase (CCIT)
Clemson Networking (CCIT)
Many many more

@ Earth

Stephen Ficklin (WSU)
Marvin Weinstein (Quantum Insights)
Ken Matusow (Synergity)
Don Preuss (Starfish Storage)
Joe Breen (Utah)
Jill Wegrzyn (UCONN)
Meg Staton (UT-Knoxville)
Dorrie Main (WSU)
Sook Jung (WSU)
Josh Burns (WSU)
Tyler Biggs (WSU)
Tim Gilmanov (IU)
Maciej Brodowicz (IU)
Daniel Kogler (IU)
Alireza Kheirkhahan (LSU)
Adrian Serio (LSU)
Hartmut Kaiser (LSU)
Chris Branton (Drury)
Florence Hudson (Internet2)
Josh Levine (ASU)
Mats Rynge (USC-OSG)
Bala Desinghu (U Chicago-OSG)
Andrew Paterson (UGA)
Claris Castillo (RENCI)
Ray Idaszak (RENCI)
Paul Ruth (RENCI)
Hong Yi (RENCI)
Anirban Mandal(RENCI)
Michael Stealy (RENCI)
Fan Jiang (RENCI)
Mert Cevik (RENCI)
Emily Casanova (USC-GHS)
Manual Casanova (USC-GHS)
Alex Bowers (Columbia U.)
Josh Vandenbrink (Ole Miss)
Ann Loraine (UNCC)
Colleen Doherty (NCSU)
John Graham (UCSD)
Many many more

Thank You Funding Agencies!!!!!

- ***"CC*Data: National Cyberinfrastructure for Scientific Data Analysis at Scale (SciDAS)***
NSF-CC* [1659300] (A. Feltus PI)
- ***"Tripal Gateway: Platform for Next-Generation Data Analysis and Sharing."***
Source: NSF-DIBBS [1443040] (S. Ficklin, PI)
- ***"MCA-PGR: Spatial and Temporal Resolution of mRNA Profiles During Early Nodule Development."***
Source: NSF-PGRP [1444461] (J. Frugoli 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] (Feltus, PI).

Genomics Scale Up Observations

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

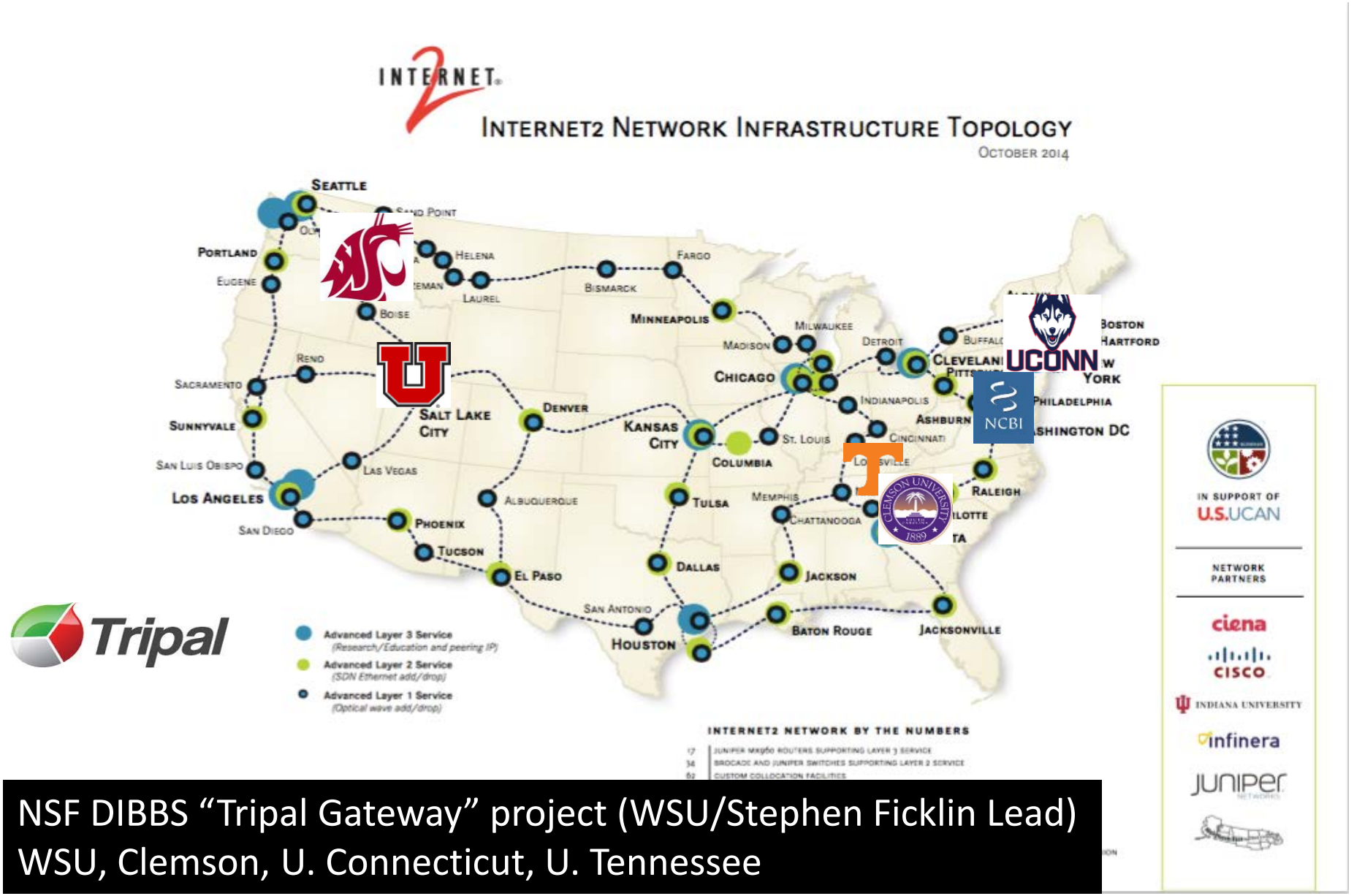
Salient Issues:::Solutions (sorted by importance)

- **Not enough storage:::Negotiate cheaper storage with campus IT (Library?) and the Cloud**
- **Not enough computational resources:::OSG, XSEDE, PRP, SLATE, negotiated Cloud credits**
- **Not enough in-lab ACI::: IT Engineer Lunch Dates, Governance committees, Research Facilitators, Software Carpentry, Collaborations: CS/CE/Engineering Departments/NRT**
- **Poor use of advanced networks:::Perform data life cycle analysis and push data close to network -- Ask IT what is possible :)**
- **Unpredictable time to compute result: queue times, queue times, queue times, broken nodes, segfaults, OOM, data geography, short walltimes:::Software optimization; Real Parallel and Redneck Parallel Computing on GPUs/CPUs; SciDAS**
- **Data Organization:::iRODs DataGrid; Tripal Databases; Named Defined Networking**

Most important: Don't ever give up.
We need to feed the hungry and heal sick kids!

Wisegeek.org
www.smartpractice.com

Research Data Transfer Networks: Internet2



NSF DIBBS “Tripal Gateway” project (WSU/Stephen Ficklin Lead)
WSU, Clemson, U. Connecticut, U. Tennessee

I2 Topology courtesy of Florence Hudson;

Tripal Databases Are Now Internet2 & Galaxy Workflow Enabled

Banana Genome Hub
http://banana-genome.cirad.fr/

Cacao Genome Database
http://www.cacaogenomedb.org/

KnowPulse
http://knowpulse2.usask.ca/portal/

Legume Information System
http://legumeinfo.org/

Citrus Genome Database
http://www.citrusgenomedb.org/

Cool Season Food Legume Database
http://www.coolseasonfoodlegume.org

Medicago truncatula Genome Database
http://medicago.jcvi.org/MTGD/?q=home

Musa Germplasm Information System
https://www.croprdiversity.org/mgis/

CottonGen
http://www.cottongen.org/

GeneNet Engine
http://xylbio.genome.clemson.edu/

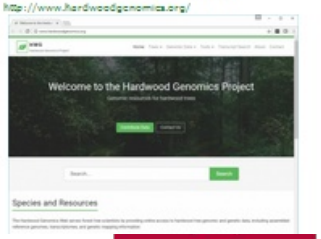
PeanutBase
http://peanutbase.org/

Planosphere
https://planosphere.stowers.org/

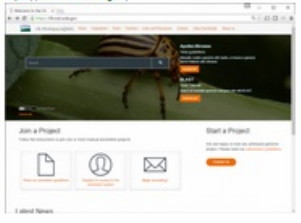
Genome Database for Rosaceae
http://www.rosaceae.org

Genome Database for Vaccinium
http://www.vaccinium.org/

Hardwood Genomics Project
http://www.hardwoodgenomics.org/



I5K Workspace
https://i5k.nal.usda.gov/



COMPATIBLE WITH...

Tripal directly supports or can be installed side-by-side with these popular tools:

CHADO

JBrowse

Galaxy PROJECT

Pathway Tools

Apollo

GBrowse

Tripal

- Over 100 Tripal Installs
- Multiple Bio-Communities
- Open Source v3.0 @Tripal.info

F INFORMATICS AND COMPUTING