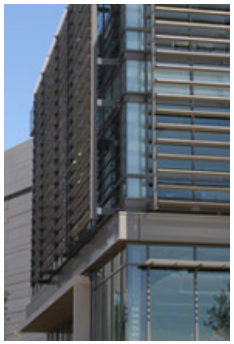
Data Management? I'm a Biologist!

Richard LeDuc, Ph.D.
National Center for Genome Analysis Support

5/1/2012

INDIANA UNIVERSITY

Acknowledgements



Le-Shin Wu,
Bill Barnett,
Robert Henschel,
Matthias Lieber (ZIH Dresden),
Phil Nistra,
and a cast of thousands

GREAT LAKES BIOENERGY RESEARCH CENTER

Yury Bukhman, James McCurdy, Adam Halstead, Enhai Xie, Irene Ong (Area 3), Mary Lipton (PNNL), Kathryn Richmond (Enabling Technologies) and others

Drs. Neil Kelleher, Paul Thomas, and Andy Forbes ProSight Development Team (past and present)

- Leonid Zamdborg
- Shannee Babai
- Bryon Early
- Ian Spauling
- Kevin Glowacz
- Eric Bluhm
- Vinayak Viswanathan
- Yong-Bin Kim
- Ryan Fellers
- Tom Januszyk
- Brian Cis
- Chris Strouse
- Seyoung Sohn
- Greg Taylor
- Joe Sola
- Lee Bynum
- Andrew Birck

Proteomics Core
Reid Townsend
Petra Gilmore
Cheryl Lichti
James Malone
Alan Davis

Michael Gross (NCRR Mass Spec)
Henry Rohrs (NCRR Mass Spec)
Ron Bose (Oncology)
Jeffry Hiken (Genetics)

Limbrick Laboratory
David Limbrick
Diego Morales

Holtzman Laboratory
David Holtzman
Rick Perrin
Jacqueline Payton
Chengjie Xiong (Biostatistics)

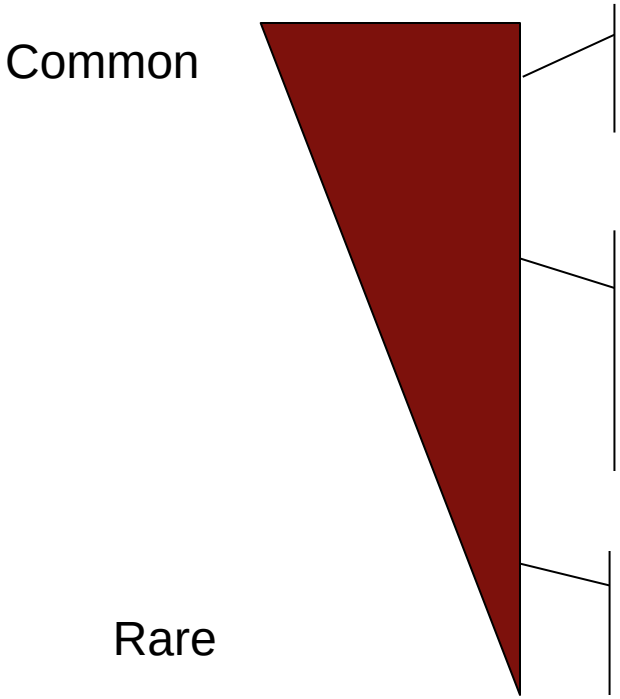
All the other numerous members of the KRG who have contributed insights over the years.

Overview

- What every researcher should know about bioinformatics (with 5 key points)
- Examples in the form of “vignettes”
- How NCGAS solves a specific bioinformatic/data management problem

Key Point #1: Different Computational Skill Levels

Computational Skills



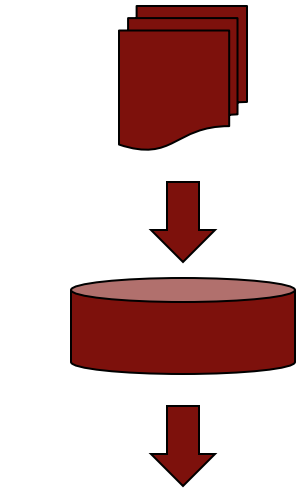
Office applications such as spreadsheets and word processors, e-mail, and web browsers.

Shell script execution, basic time sharing, scripting or programming in one or a few languages, basic statistical programming in R or SAS.

Advanced file systems, batch computing, object oriented programming, …

You do the same things in HP and personal computing

- Run applications
- Store data in files
- Transfer files over networks



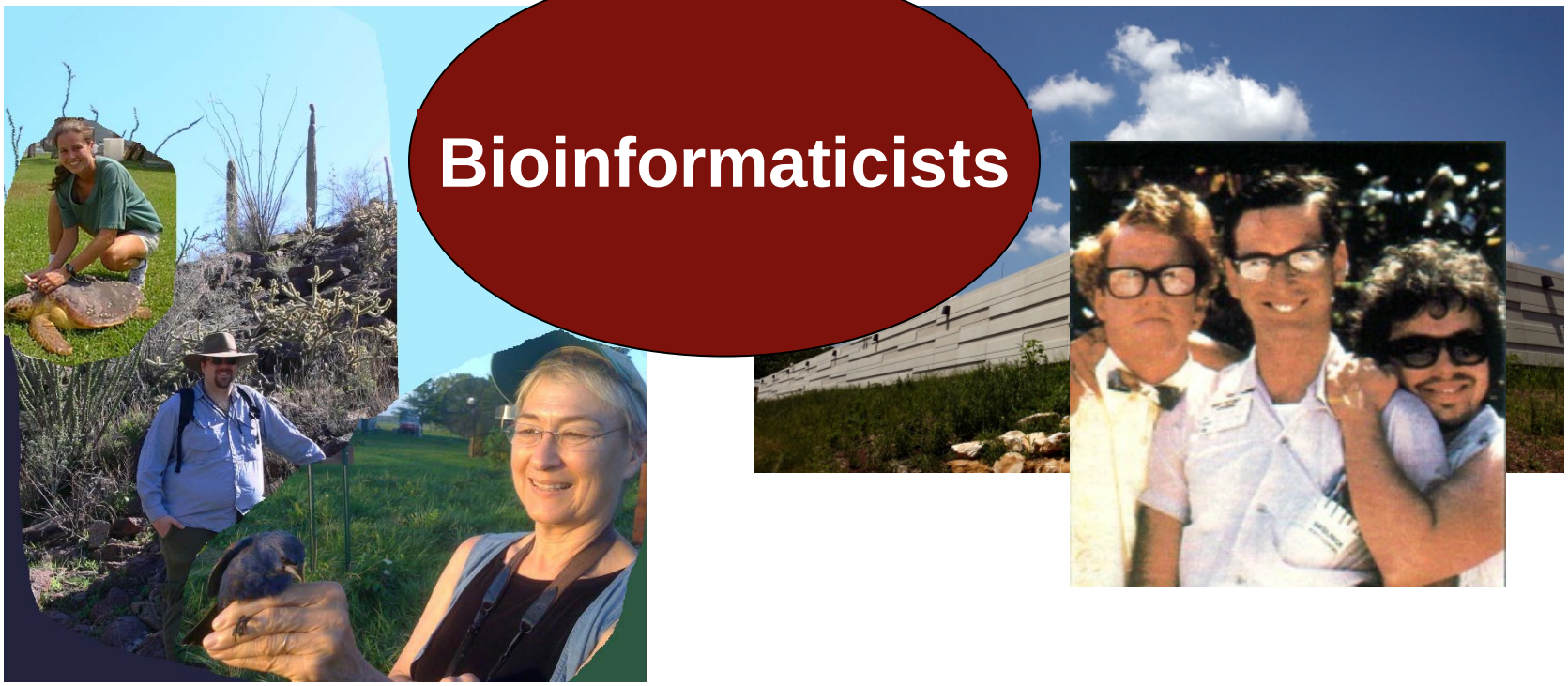
But…

No one is looking out for your overall user experience

Key Point #2: Bioinformatics spans (at least) two cultures

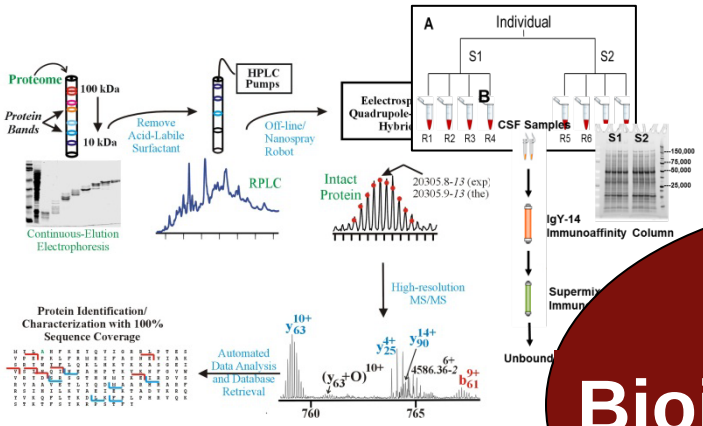
Academic

IT Professional

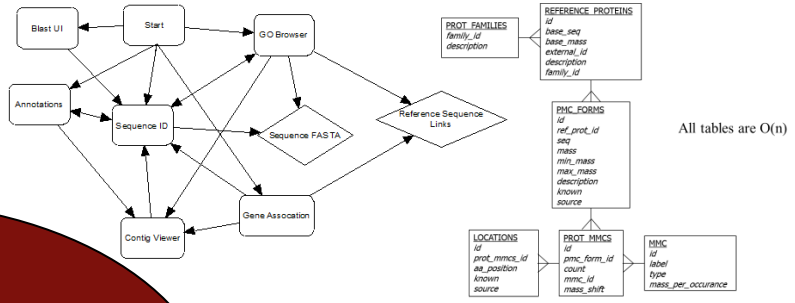


Key Point #3: Lots of Different Jobs

Experimental Biology

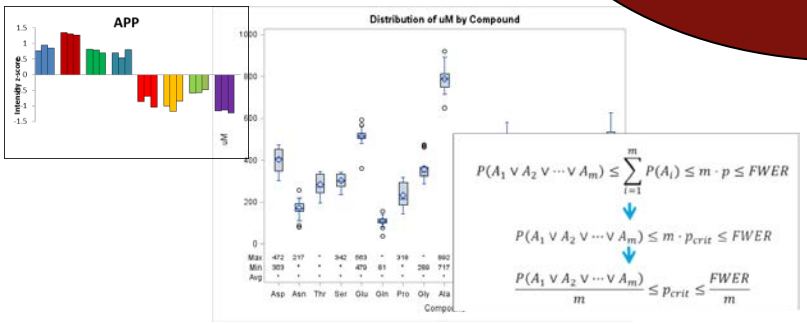


Information Technology

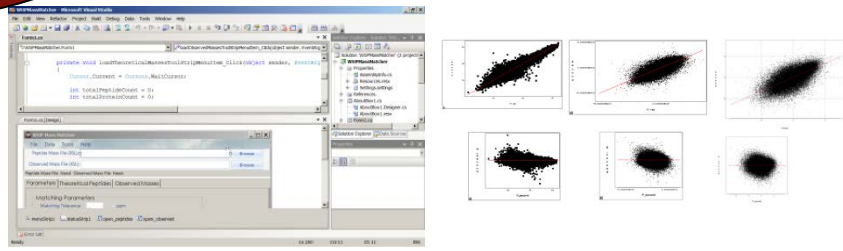


Bioinformaticists

Statistics



Computer Science

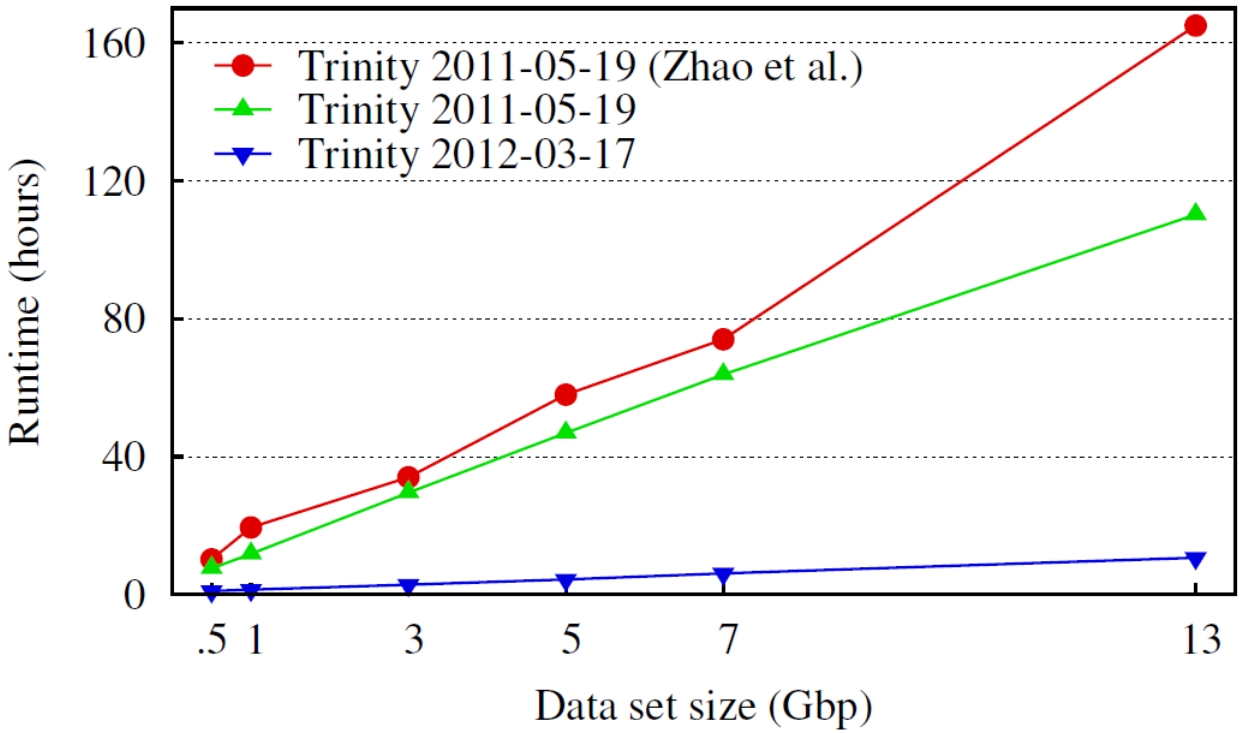


Key to Bioinformaticians and Related Types

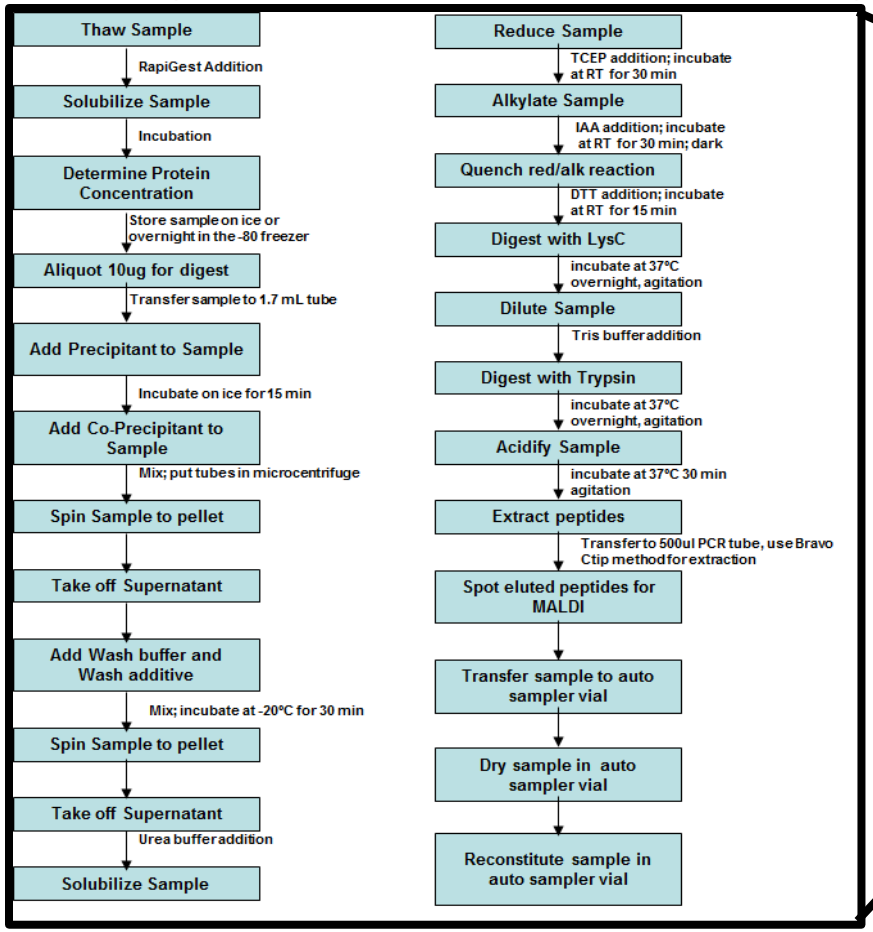
1.	a. Writes computer code	2
	b. More interested in computation the result	4
2.	a. Writes in SAS or R (maybe SPSS), also designs experiments and tests statistical significance	**Statistician**
	b. Focused on writing code	3
3.	a. More interested in output than code	**Research Programmer**
	b. More interested in code efficiency	**Software Engineer**
4.	a. Holds an academic position; also interested in algorithms etc.	**Computer Scientist**
	b. Is an Information Technology professional	5
5.	a. Normalizes tables in sleep	**Data Base Administrator**
	b. Uses vi because it is "intuitive"	**Systems Administrator**

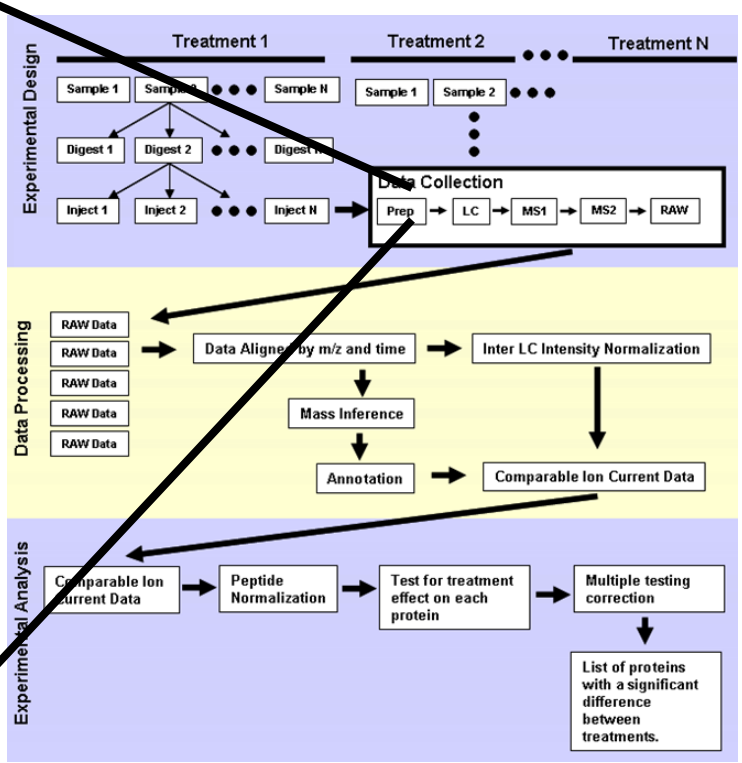

Commands
Results
Human Protein Reference Database
Human Proteomics Initiative
Swiss-Prot/TrEMBL
Alternate Splicing Database
PTM Prediction Algorithms
RESID DB of PTMs
ProSight Warehouse
Exon/Intron Prediction Algorithms
Non Redundant Database
SNP Prediction Algorithms
dbSNP
Known PTMs
FASTA Files
ProSight Retriever
Results

Vignette: Not all programming is equal

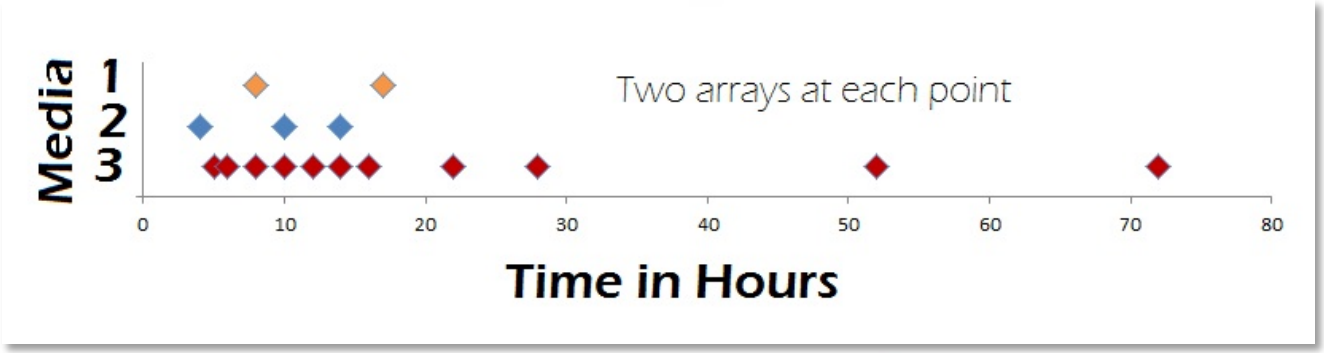


Key Point #4: Every detail is critical!

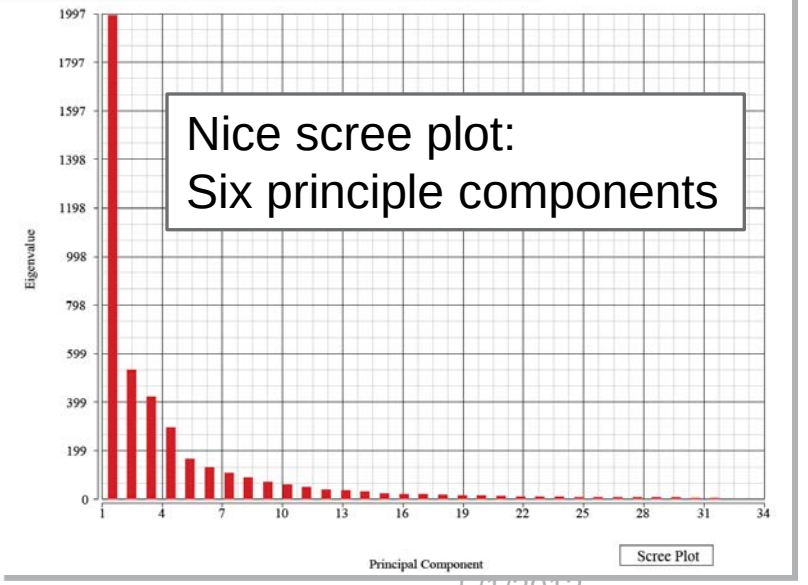




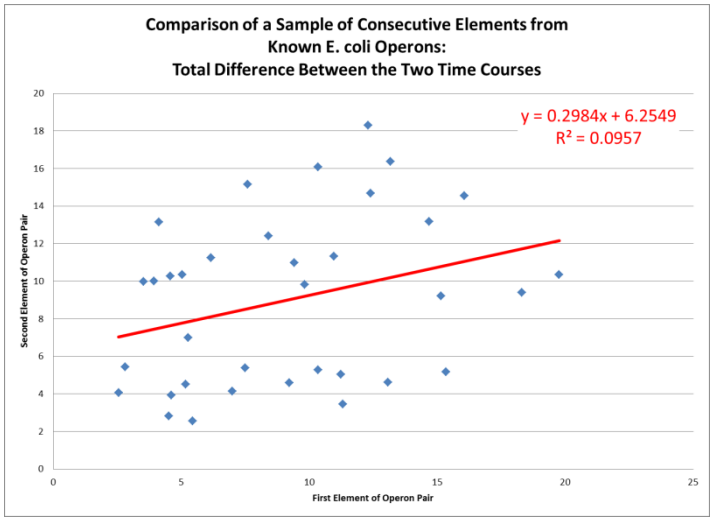
Vignette: A bad experiment is "bad"



- Researcher unclear on how to interpret results.
- GSEA over all *E. coli* pathways fails to find any significant results (after FDR correction).
- Compare expression levels of adjacent genes on known operons.

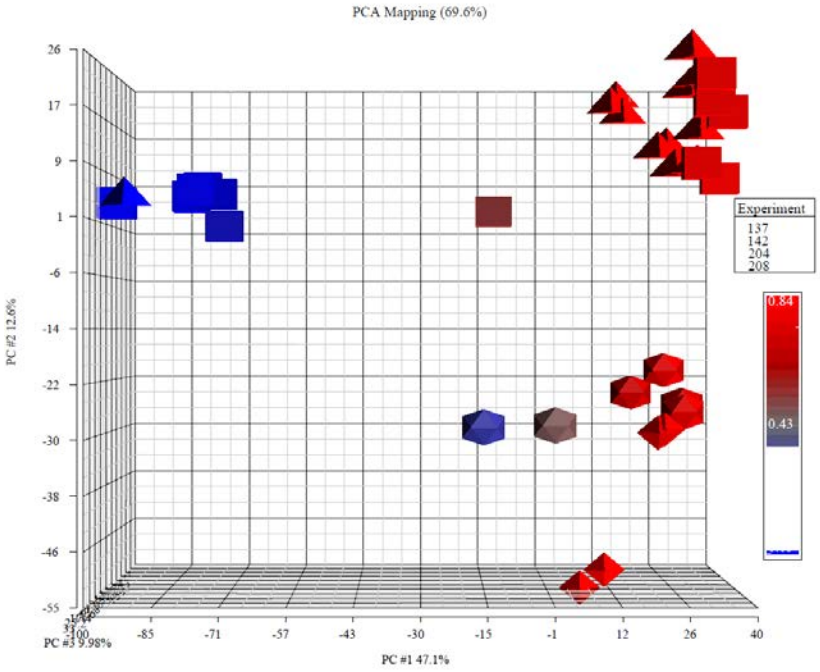


First PC is Technical Noise



Pilot study with 35 pairs

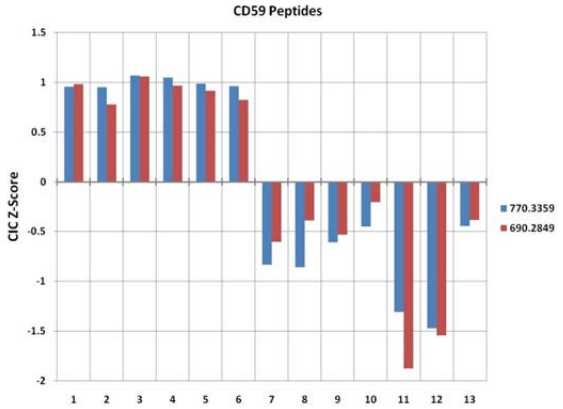
Biology still trumps mathematical formalism and wishful thinking.

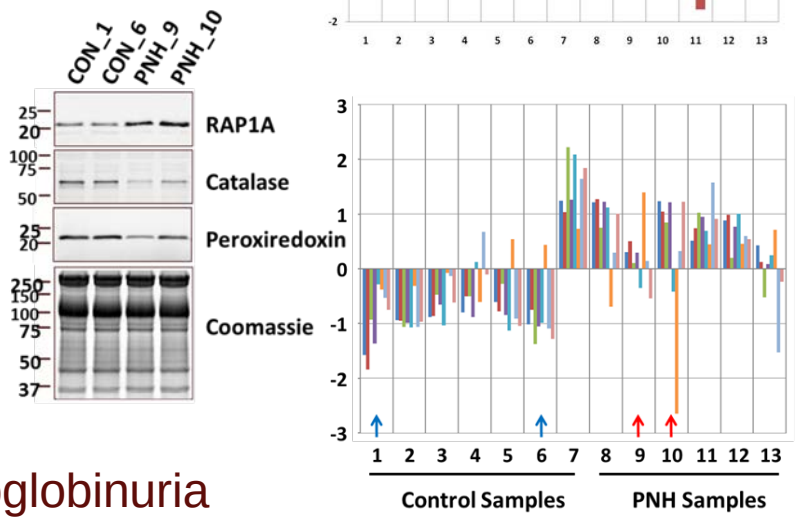


- Shape shows media.
- Color shows intra-operon correlation.
- Over all known operons.

Key Point #5: ‘omics experiments are lots of parallel small experiments

- Microarray
- Shotgun proteomics
- RNA-Seq
- High Throughput Screening
- Even Bird Songs



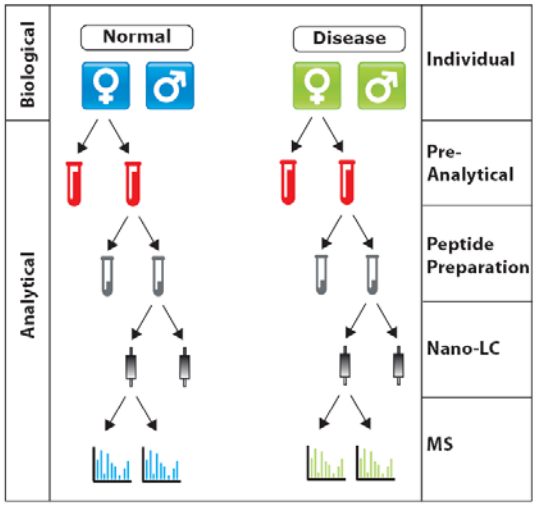


Human RBC Ghosts
Paroxysmal nocturnal hemoglobinuria

Vignette: Experimental Design for Shotgun Proteomics

Infant CSF

$$I_{ijkl} = \mu + a_i + d_{j(i)} + r_{k(ij)} + e_{ijkl}$$



Where

i=1 or 2 and represents the two preparations of human CSF,
j = 1 to 3 for each digestion within a given preparation,
k = 1 to 3 for each injection (or run) within each digestion
l = 1 to the number of peptides for the given protein.

Under this model, let

$$a_i \sim iid\ \mathrm{N}(0, \sigma_{\mathrm{P}}^2) \qquad r_{k(ij)} \sim iid\ \mathrm{N}(0, \sigma_{\mathrm{R}}^2)$$

$$d_{j(i)} \sim iid\ \mathrm{N}(0, \sigma_{\mathrm{D}}^2) \qquad e_{ijkl} \sim iid\ \mathrm{N}(0, \sigma_{\mathrm{e}}^2)$$

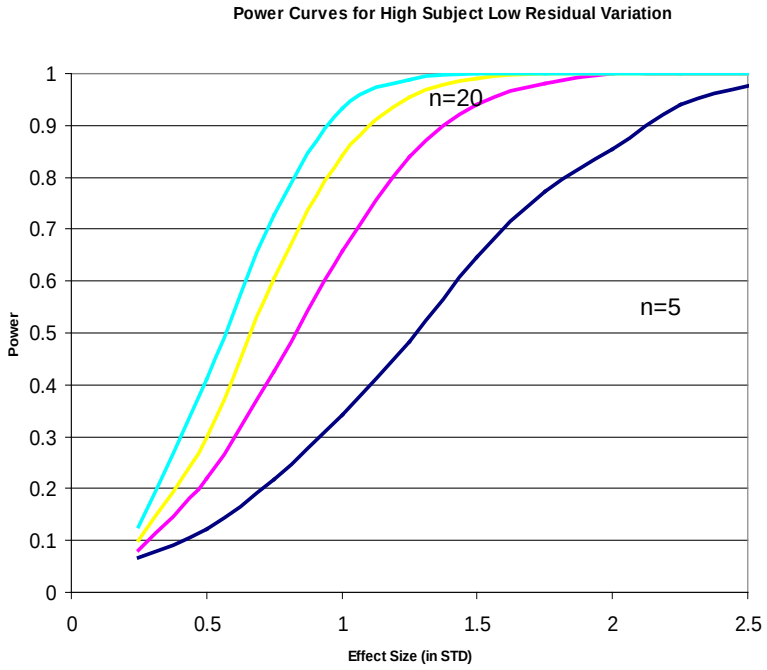
a_i is the effect for the i^{th} random preparation

$d_{j(i)}$ is the effect for the j^{th} random digestion from the i^{th} preperation

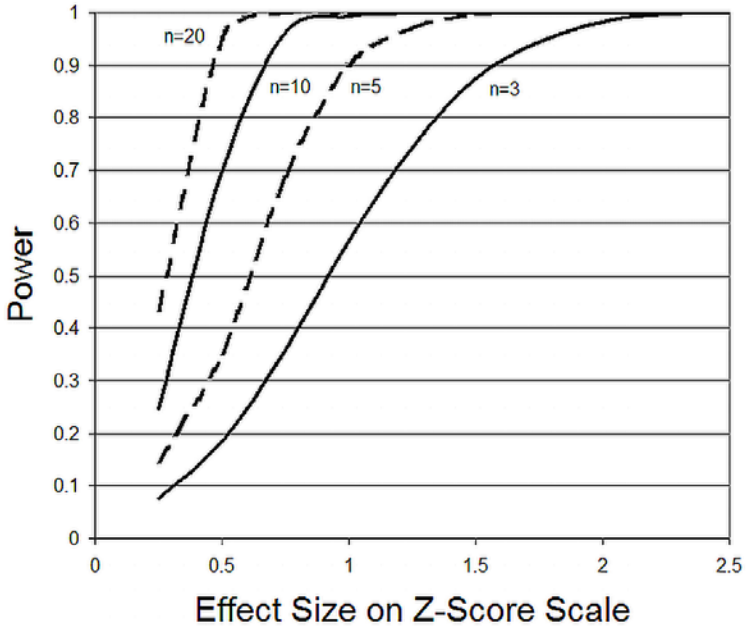
$r_{k(ij)}$ is the effect for the k^{th} random run from the j^{th} random digestion from the i^{th} preperation

e_{ijkl} is the residuals

Human versus Mouse Power

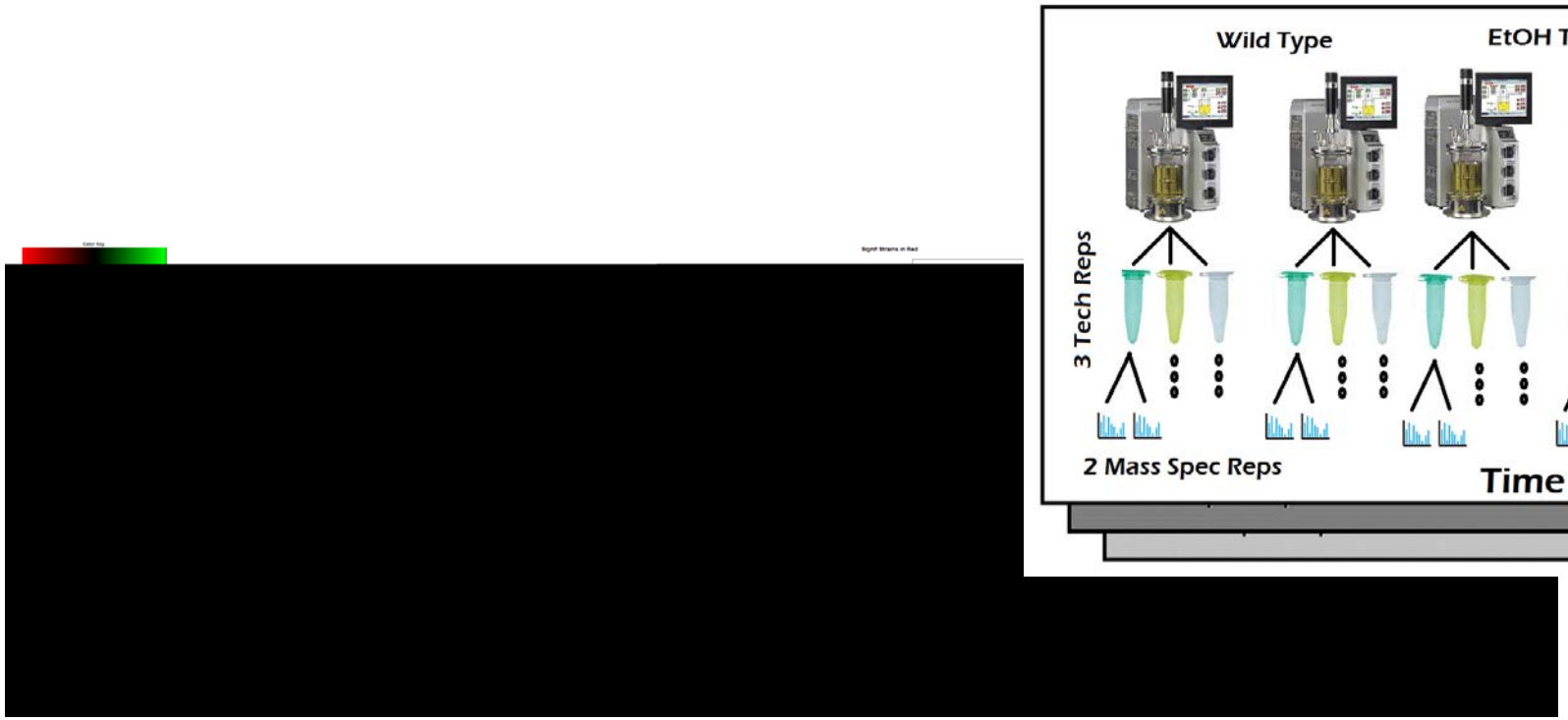


Human Subjects



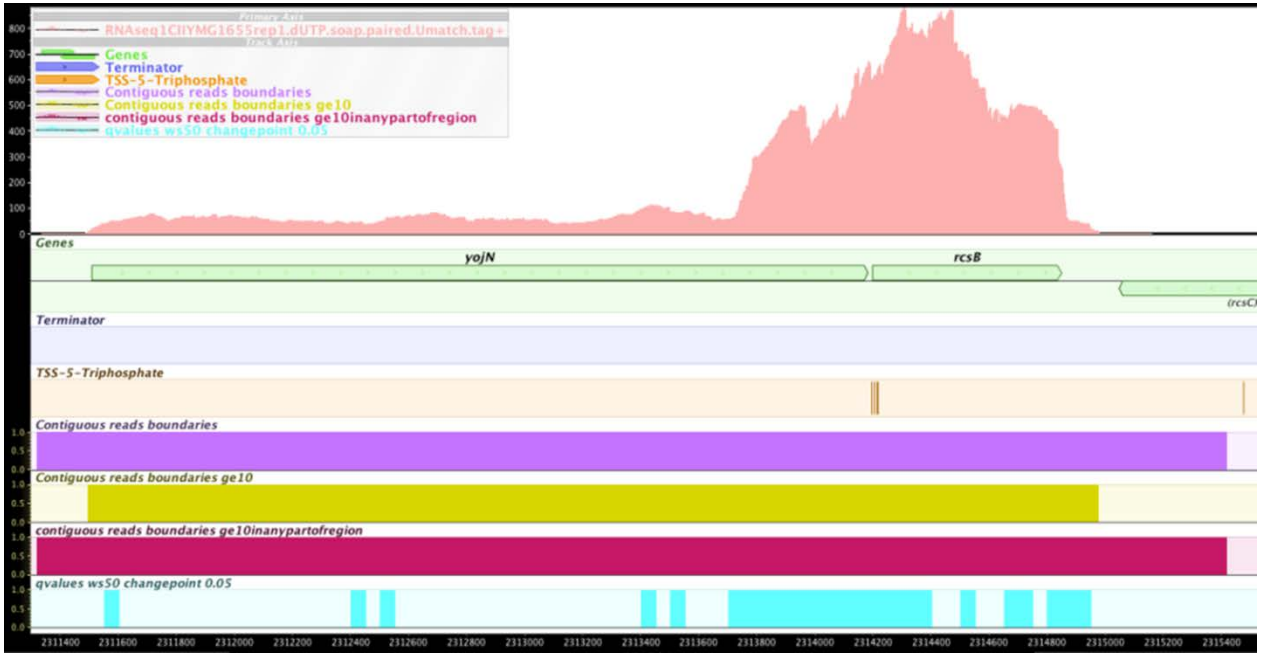
Inbreed Mice

E. coli Shotgun Proteomics



Notice the sample preparation-level variation.

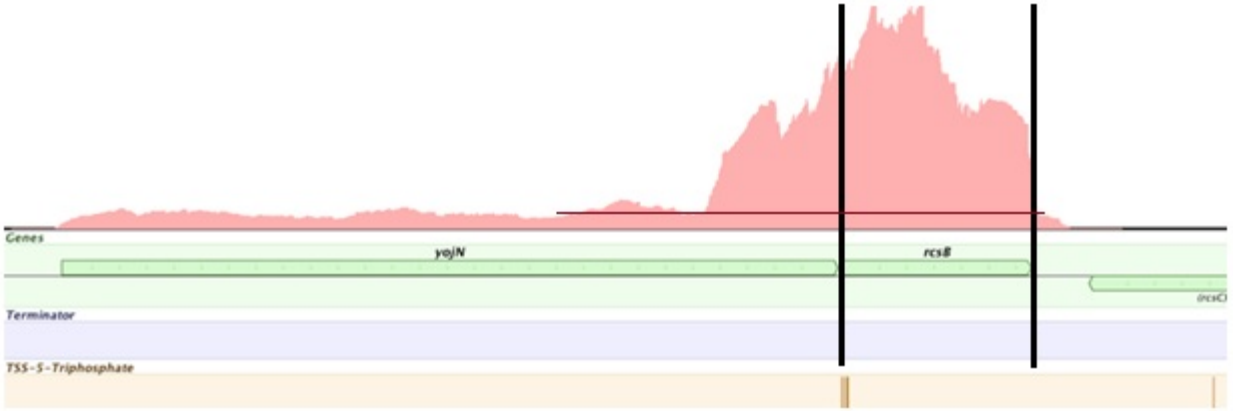
Vignette: RNA-Seq “Biases”



- RNA-Seq uses NGS to sequence transcribed RNA.
- Aligns short reads either against a reference genome, or *de novo*.
- Considered to be rapidly replacing microarray.

Courtesy of Irene Ong at the Great Lakes Bioenergy Research Center

A Biological Problem



- What is the expression of yojN?
- Clearly promoter sequences for rcsB are contained within this gene.

Key points

1. Different skill levels – no single “computer expert”
2. Multiple cultures – known what motivates your associates
3. Bioinformaticists do lots of different “jobs”
4. Any detail can potentially destroy an experiment
5. ‘omics experiments are many small experiments

National Center for Genome Analysis Support

- Funded by National Science Foundation
- Large memory clusters for *de novo* sequence assembly
- Bioinformatics consulting for biologists
- Optimized software for better efficiency
- Open for business at: http://ncgas.org

NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT
INDIANA UNIVERSITY

National Center for Genome Analysis Support

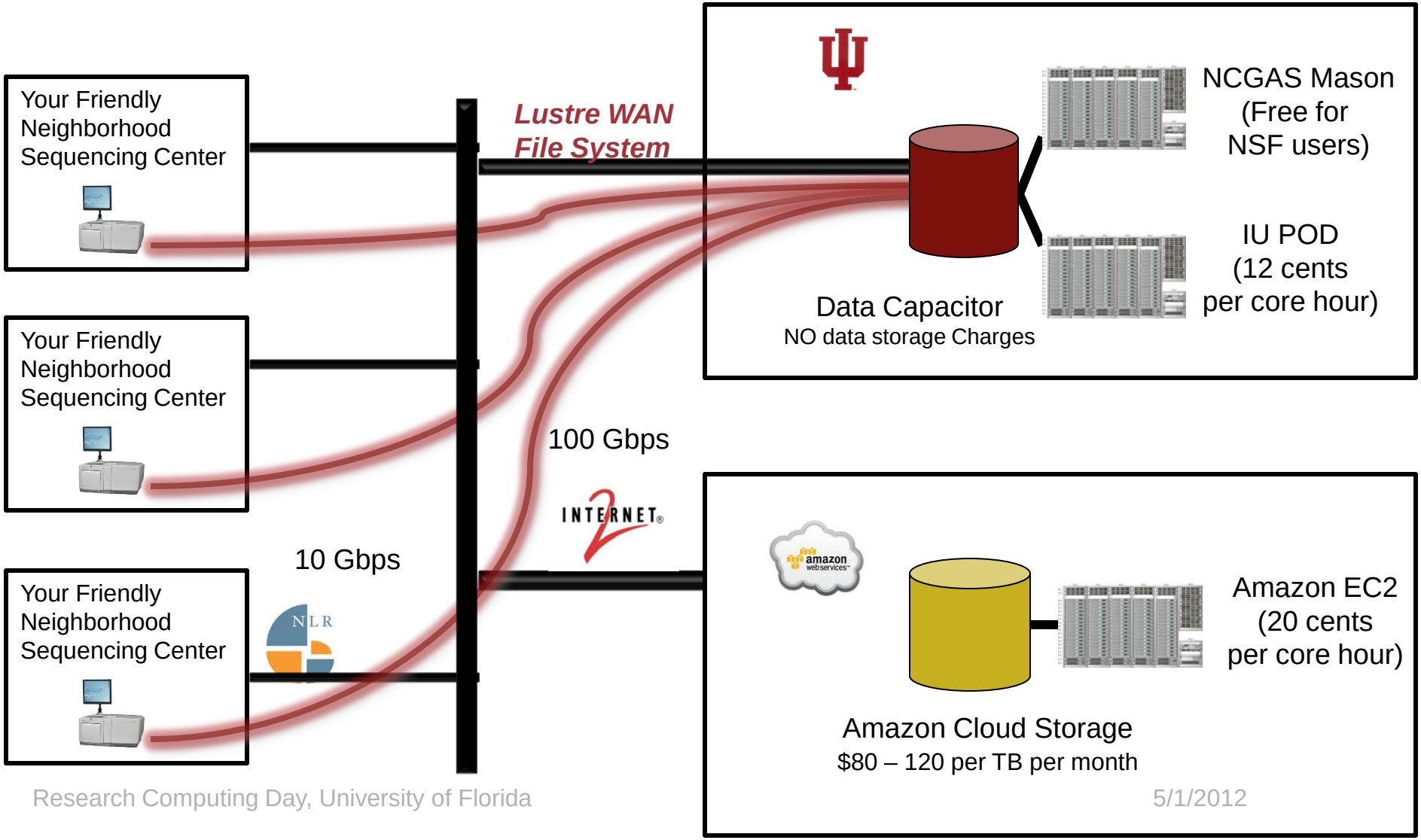
Cyberinfrastructure

- Mason large memory cluster (512 GB/node)
- Quarry cluster (16 GB/node)
- Data Capacitor (1 PB at 20 Gbps throughput)
- Research File System (RFS) for data storage
- Research Database Cluster for managing data sets.
- All interconnected with a high speed internal network (40 Gbps)

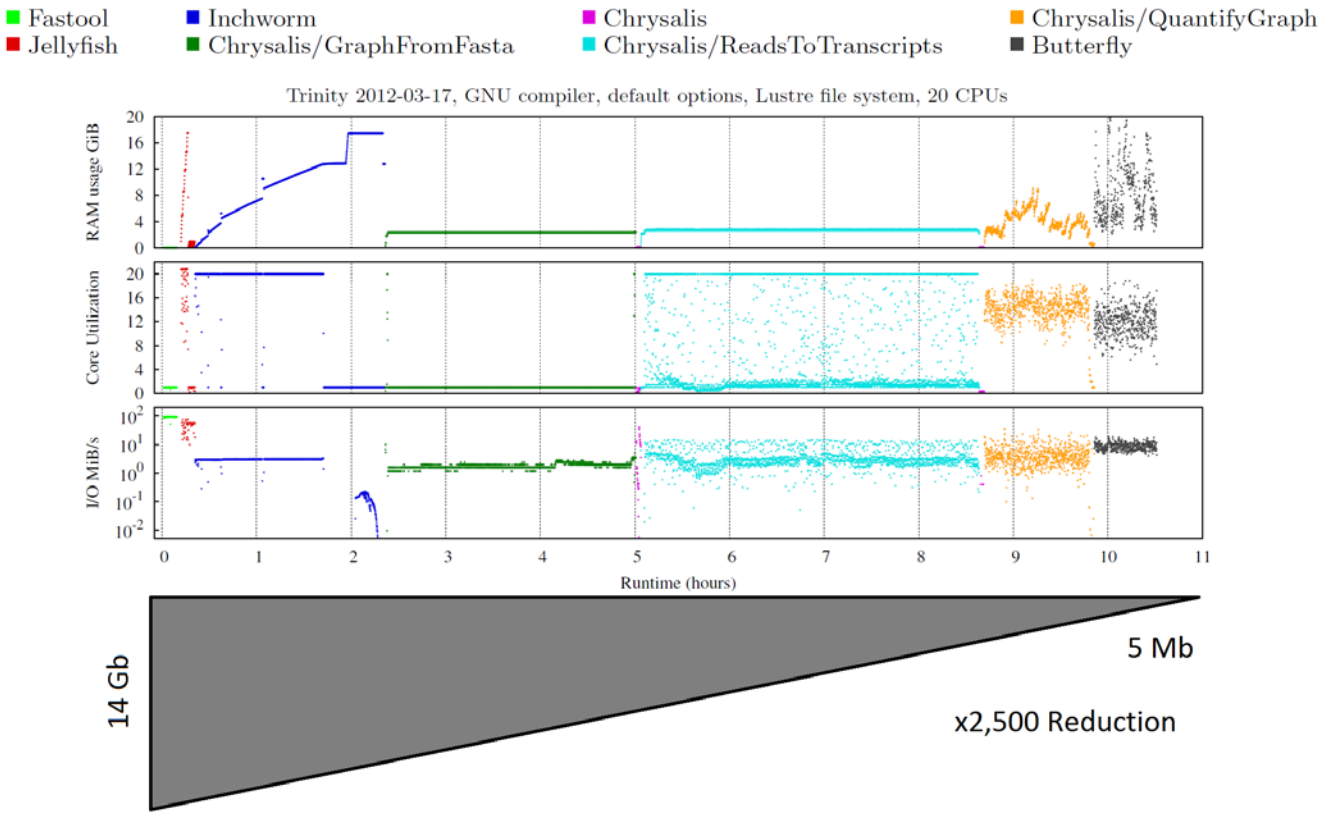
Services

- Short Consult
- Long Consult
- Intellectual Contribution
- Letter of Support
- Subcontract / Co-PI

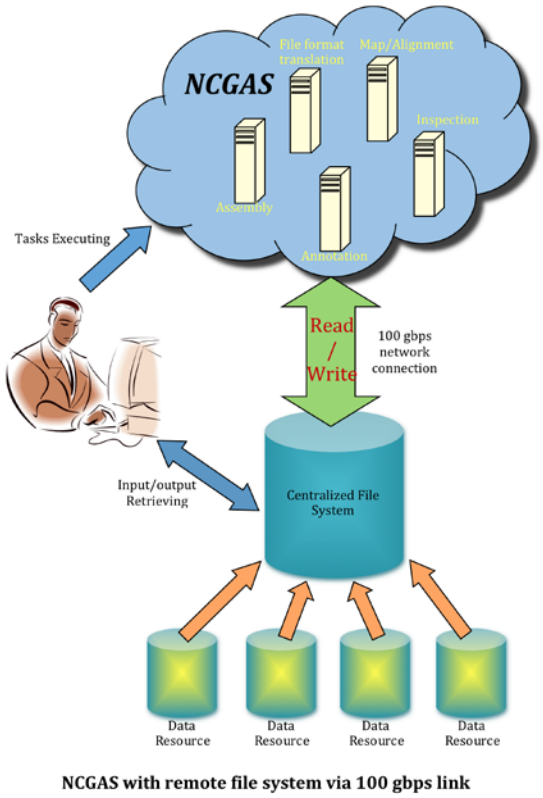
Two Options for Computation and Storage



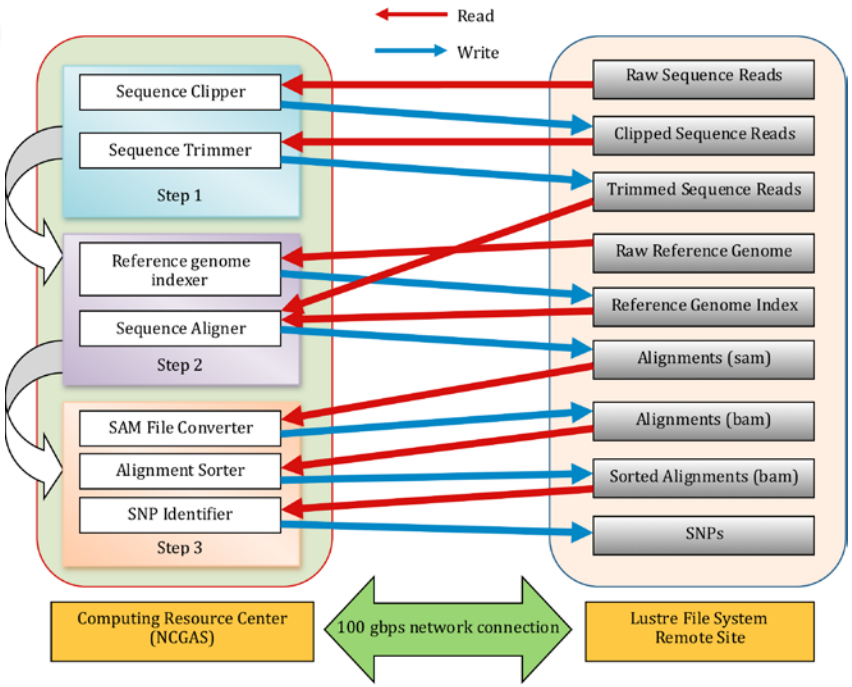
Trinity Data Reduction



Demo at Supercomputing 11



NCGAS with remote file system via 100 gbps link



- *STEP 1:* ***data pre-processing****, to evaluate and improve the quality of the input sequence*
- *STEP 2:* ***sequence alignment*** *to a known reference genome*
- *STEP 3:* ***SNP detection*** *to scan the alignment result for new polymorphisms*

Thank You

Questions?

Bill Barnett (barnettw@iu.edu)
Rich LeDuc (rleduc@iu.edu)

NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT
INDIANA UNIVERSITY

Data Management? I'm a Biologist

Whether you work in a clean wet-lab or in wet muddy boots, biologists are learning they must think about data management to reach their professional goals. With examples pulled from biofuels to translational medicine and point in between, we will look at "big picture" issues in life science data management and bioinformatics, and try and orient researchers to the bewildering complexity of computational concerns that have recently entered their disciplines.