

High Performance Computing serving Life Science Research Needs

Craig A. Stewart, Ph.D.

Executive Director, Pervasive Technology Institute
Associate Dean, Research Technologies
Indiana University

3 July 24



**RESEARCH
TECHNOLOGIES**

INDIANA UNIVERSITY
University Information Technology Services



**PERVASIVE TECHNOLOGY
INSTITUTE**

INDIANA UNIVERSITY

License Terms

- Please cite as: Stewart, C.A. 2014. High Performance Computing serving Life Science Research Needs. Presentation. Presented at Universitaet zu Koeln, Koeln, Germany, 3 July 2014. ###____
- Items indicated with a © are under copyright and used here with permission. Such items may not be reused without permission from the holder of copyright except where license terms noted on a slide permit reuse.
- Except where otherwise noted, contents of this presentation are copyright 2014 by the Trustees of Indiana University.
- This document is released under the Creative Commons Attribution 3.0 Unported license (<http://creativecommons.org/licenses/by/3.0/>). This license includes the following terms: You are free to share – to copy, distribute and transmit the work and to remix – to adapt the work under the following conditions: attribution – you must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work). For any reuse or distribution, you must make clear to others the license terms of this work.



Agenda

- Background about myself and about IU
- Life science support for genome analysis
- Life sciences support for clinical and translational research
- Alzheimer's disease
- A few closing thoughts about life sciences and cyberinfrastructure and eScience in the future



Key Events in my Professional History

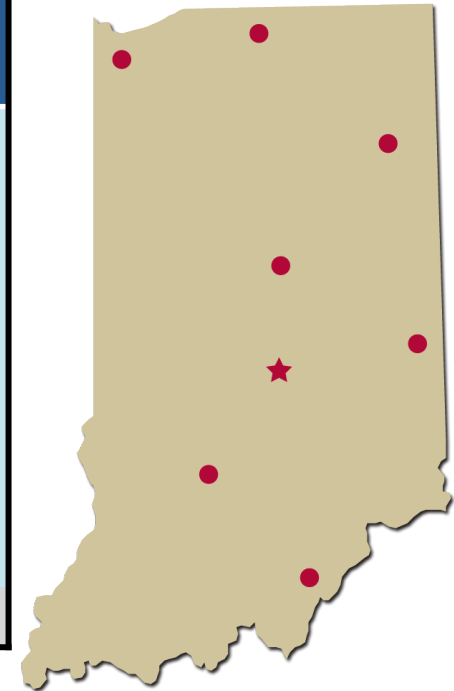
1981	Graduated with BA in biology and mathematics from Wittenberg University (Springfield, OH). Started as graduate student at Indiana University in biology.
1982	Met Marion Krefeldt (in Bremerhaven geboren)
1984	Switched from being teaching assistant in biology to assistant consultant with Bloomington Academic Computing Services, starting with Lotus 1-2-3 Key Disks.
1985	Full-time appointment at BACS Information Center (Service Desk).
1986	Manager, Business Computing Facilities (IU School of Business), finished Ph.D. in Biology
1991	Manager, Center for Statistical and Mathematical Computing (UCS).
1995	Manager, University Computing Services Support Center.
1996-7	Senior Manager, Assistant Director, Acting Director, Director research and academic computing
1997	Michael McRobbie arrived at IU from the supercomputing center at ANU to become IU's first full VP for IT and CIO and reorganized IT organization into University Information Technology Services.
1997	US Dept. of Commerce imposes a 4X tariff on purchase of Japanese supercomputers within the US.
2005	April Fool's Day: Promoted to Associate Vice President for Research and Academic Computing and COO of Pervasive Technology Labs
2008	Associate Dean for Research Technologies, Executive Director of Pervasive Technology Institute.
	<i>Key point: I have been around a long time – from when IU was unimportant in IT to when IU was sued by Metallica to having the #23 system on the Top500 list. Long enough to see technological and cultural change happen at IU, lead some of it, and learn from all of it</i>





IU – Founded in 1820

Campus	Academic appointees	Nonacademic Staff	Undergrad Students	Grad. & Prof. Students
IUB	2,942	5,379	32,371	9,762
IUPUI	3,895	4,449	22,271	8,180
IU Northwest	425	243	5,636	548
IU South Bend	542	305	7,860	630
IU East	267	159	4,052	134
IP Fort Wayne	N/A	N/A	N/A	N/A
IU Kokomo	191	138	3,581	138
IU Southeast	498	243	6,203	701
Totals	8,760	10,916	81,974	20,093



1,200 degree programs

IU community: 121,743 people total

1.2 million credit hours per semester

Two core research/education campuses, six regional campuses

Tuition and mandatory fees per year: \$10,209 FY 13/14 for IUB Undergrads





IU Budget Category	2012/2013 Budget
Unrestricted	\$2,155,174,476
Restricted	\$640,532,854
Auxiliary	\$403,026,761
Total	\$3,198,734,091



Indiana University Health

IU Health Patient Metrics – 2012/2013	
Admissions	143,219
Outpatient visits	2,244,320
Staffed Beds	3,326

- No engineering
- No agricultural research
- No Veterinary school



Office of the VP for Information Technology

Staffing and Budget

Category	FTEs	Distinct Individuals
Academic	11	11
Student Academic	2	8
Appointed Professional Staff	967	977
Hourly Staff	126	505
Total	1,106	1,501

Budget

~\$120 M US / year

Of this, roughly \$13 M US / year is from grants and contracts, primarily federal research grants and contracts



What is RT's mission?

The mission of the Research Technologies division of UITS is to develop, deliver and support advanced technology solutions that improve the productivity of and enable new possibilities in research, scholarly endeavors, and creative activity at Indiana University and beyond; and to complement this with education and technology translation activities to improve the quality of life of people in Indiana, the nation, and the world.

We are a mission- and value-driven organization. We are not a technology-driven organization.

We identify needs, identify possibilities, and discover new ways to meet those needs, realize those possibilities, and create new ones. In so doing, we create, deploy, and support technology. **We are a technology-driving organization.**

Roughly 30% of personnel are funded by external agencies



A language issue

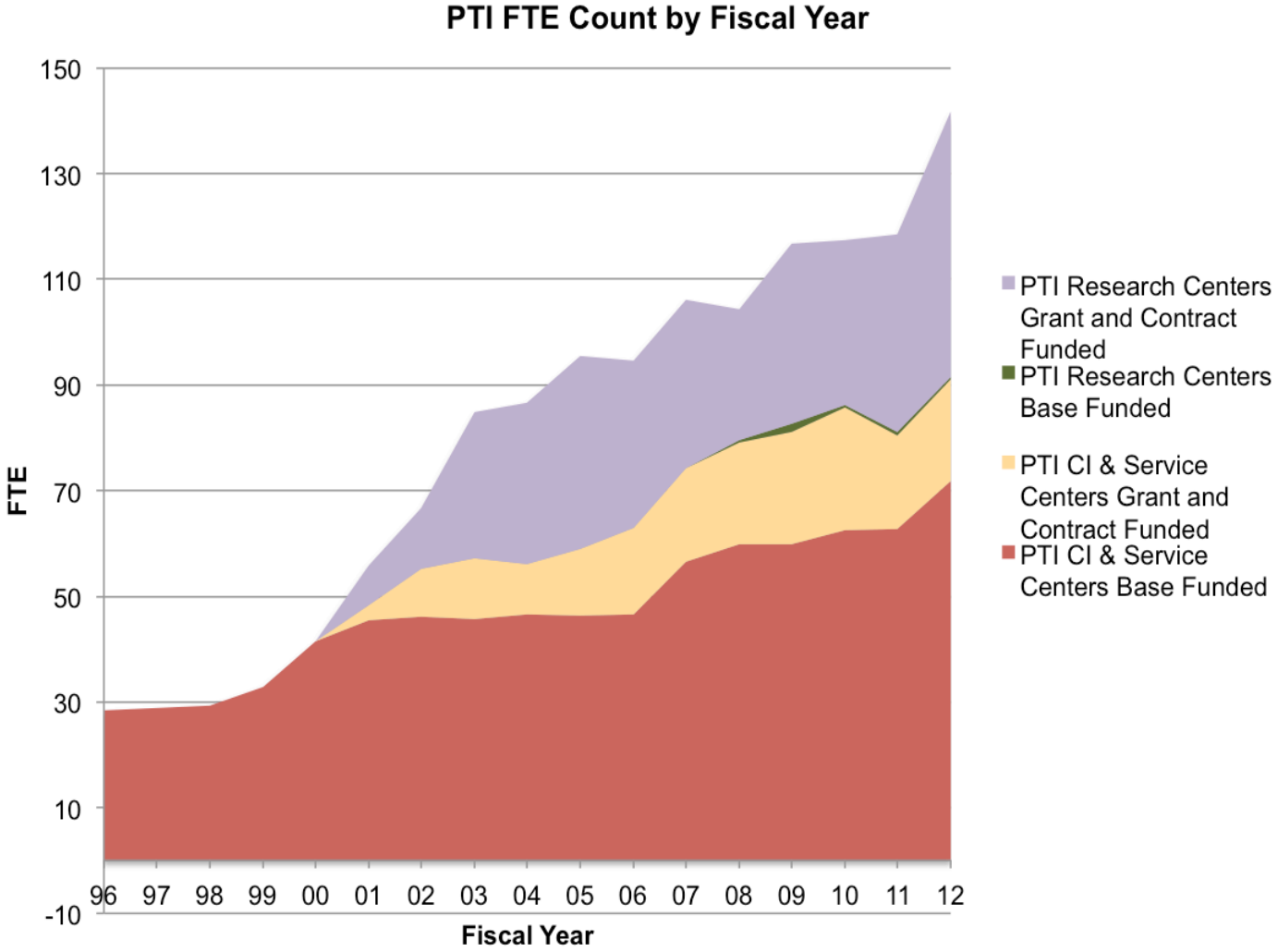
Cyberinfrastructure (primarily an US term): Cyberinfrastructure consists of computing systems, data storage systems, advanced instruments and data repositories, visualization environments, and people, all linked together by software and high performance networks to improve research productivity and enable breakthroughs not otherwise possible. (Stewart, 2007)

eScience (primarily an EU term): “In the future, e-Science will refer to the large scale science that will increasingly be carried out through distributed global collaborations enabled by the Internet. Typically, a feature of such collaborative scientific enterprises is that they will require access to very large data collections, very large scale computing resources and high performance visualization back to the individual user scientists.” (National e-Science Centre, 2010)

Probably cyberinfrastructure = eScience + support staff



From not much at all to a mid-sized cyberinfrastructure center



Some foundational issues regarding life science research

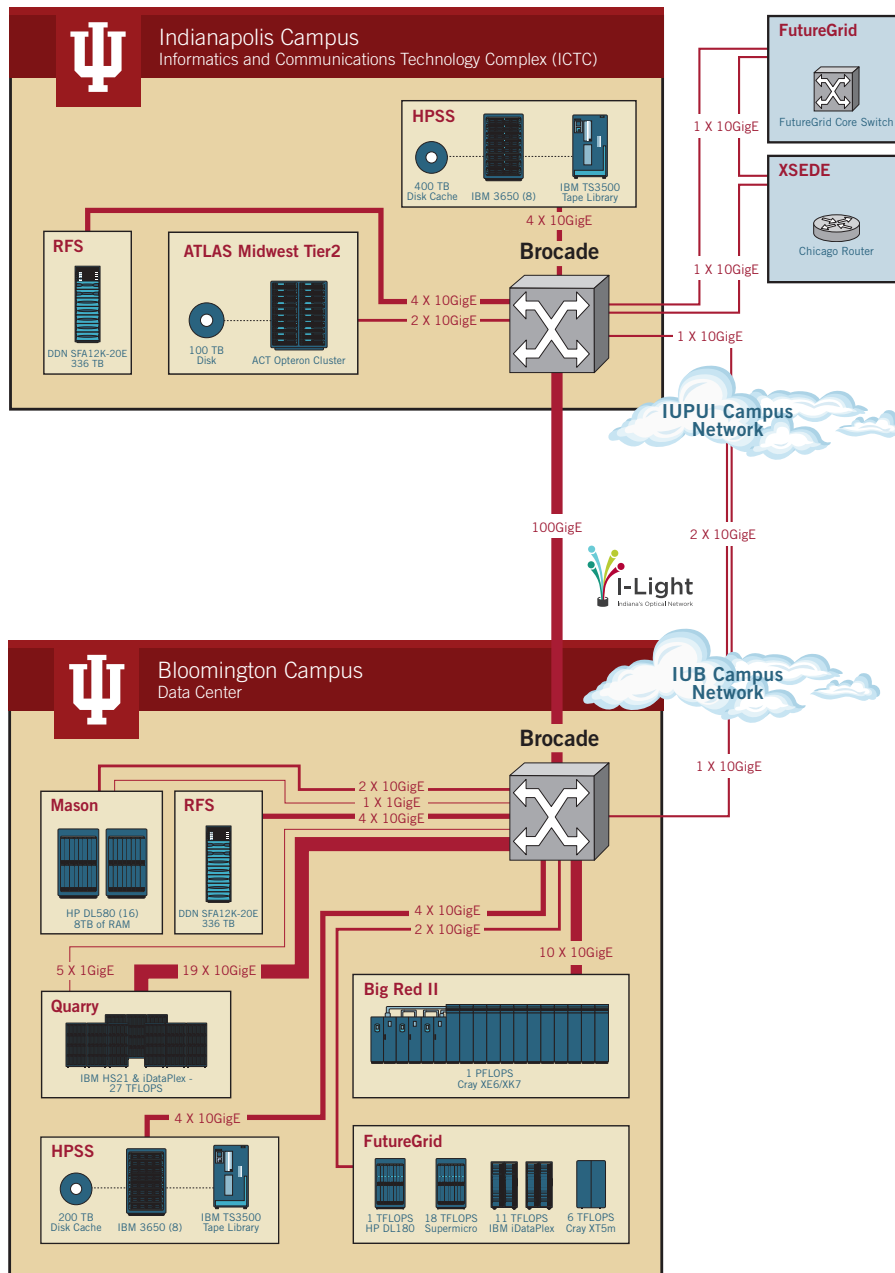
Life sciences data matter – for a long time

HIPAA (Health Information Privacy and Portability Act)

Life science data are

- Increasingly born digital
- Often collected in ways that are driven by the research, not by a desire to have tidy data structures
- Often researchers want to analyze their data very quickly as a part of their ongoing, daily research activities
- And life sciences are more complicated than physics





Genome analysis

BLAST

Phylogenetics

Genome assembly



BLAST

Perhaps the most commonly used bioinformatics application ever

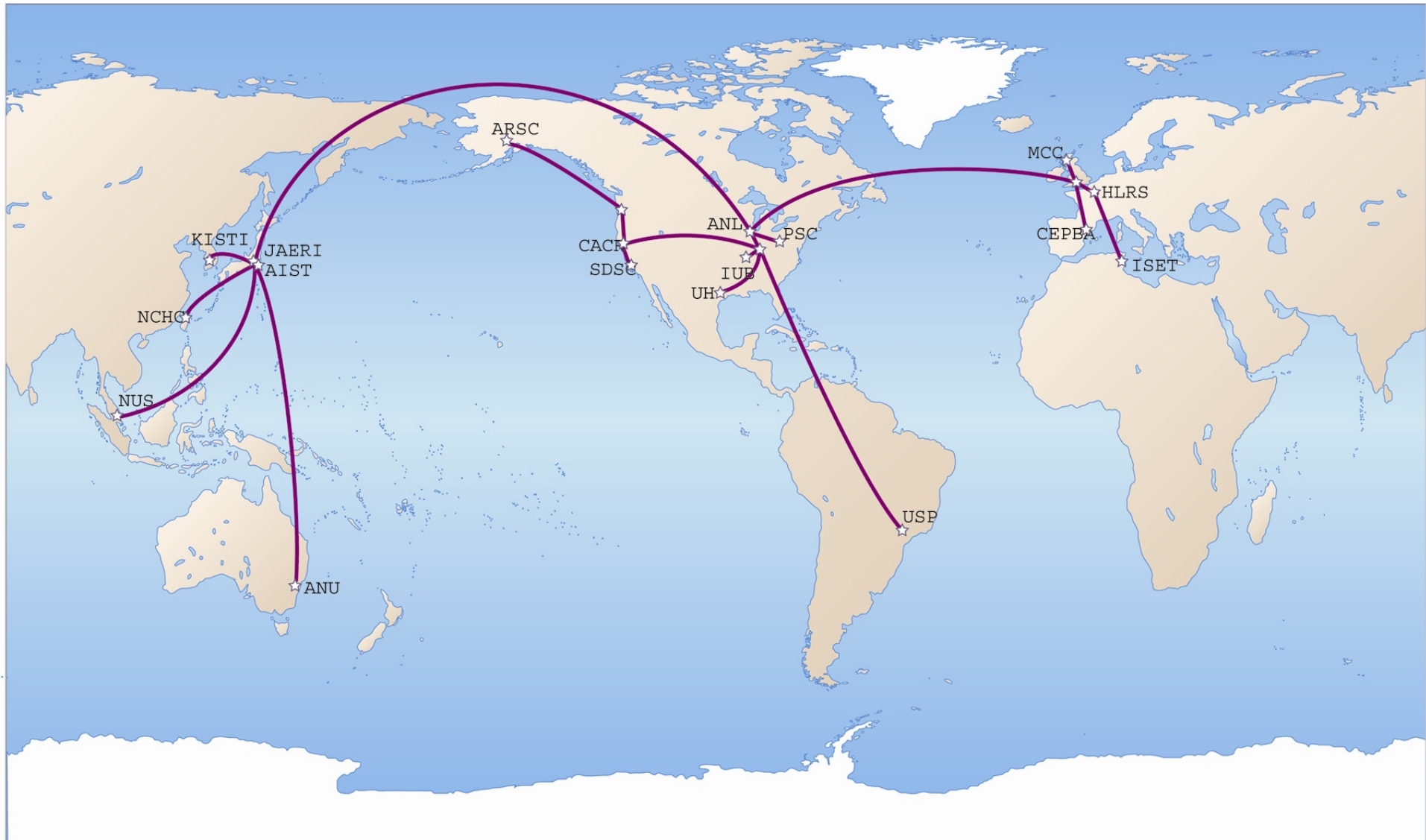
There are many variants of BLAST

From 2003 to today we have constantly experimented with new approaches to running BLAST

- MPI-BLAST
- BLAST on Cell processors
- BLAST as a high throughput application
- The need for BLAST is so constant that this approach of constantly experimenting has worked out well.



GleiderfüsslerGrid



Why this project on a grid?

- Important & time-sensitive biological question requiring massive computer resources
- A biologically-oriented code that scales well
- Grid middleware environment & collaboration tool well suited to the task at hand
- Opportunity to create a grid spanning every continent on earth (except Antarctica)
- Great way to get computer centers over the world to donate computing time for free
- And a few general biological results from fastDNAmI in general
 - Fungi are more closely related to animals than plants or microbes
 - Horizontal movement of genes in plants
 - Timing jump of AIDS from primates to humans




And in 2003 this is how tired we looked when we collected our award for most distributed HPC application at SC2003



National infrastructure serving genome science

Creators of
new software



NCGAS – small, serving large community largely reactively

- Trinity
- Galaxy
- ABySS
- Velvet

iPlant – large collaborative serving plant science

- DNA Subway
- iPlant Discovery Environment
- Many bioinformatics software applications planned as part of group strategy

XSEDE – designed to serve all research communities

- Stampede
- Gordon
- Blacklight
- Comet
- Mason
- Wrangler
- FutureGrid

Network – essentially independent of any particular research community

- Internet2
- Regional providers



XSEDE (eXtreme Science and Engineering Discovery Environment)





iPlant Cloud Services

PROJECT ATMOSPHERE

Customized cloud platform for computing on your terms !

New biology priorities going forward:

- Expand Scope to Non-Plant Species
- Continue Support for NGS
- Deliver CI Platform for Modeling, Molecular Breeding
- Expand Support for Ecophysiology
- Continue Range Map Creation for Biodiversity
- Integrate Environmental Information
- Support Additional Molecular Profiling Tech



NCGAS – National Center for Genome Analysis Support

- Mason provided as “facilities” with IU funding, for use by national research community, through XSEDE, as part of this award
- IU also hosts the commercially owned Rockhopper system – owned and managed by Penguin Computing, a “pay to use” system, software installed and supported by NCGAS
- ~ \$0.7M / year budget (award of \$1.5M over 3 years + match)
- Focused on user-driven needs
- ~ 4 FTEs (Full Time Equivalents = 1 person) total
- Newest of the projects discussed – funded starting in 2011 (implies that situation prior to 2011 was not optimal)



National Center for Genome Analysis Support

“Mind the Gap”

Gap	How we fill it
System configurations offered by XSEDE and what people doing genome assembly need	Mason (IU contribution to facilities)
Software on XSEDE is not what people need	NCGAS installs and maintains
Software works slowly	NCGAS tunes / re-engineers
People just need help	NCGAS provides consulting NCGAS goes to conferences and informs people about our services
<i>People need storage</i>	<i>NCGAS provides tape storage (IU facilities)</i>
<i>People need to publish data sets</i>	<i>IU provides resources via IU Scholarworks</i>



Trinity

RNA-Seq De novo Assembly Using Trinity

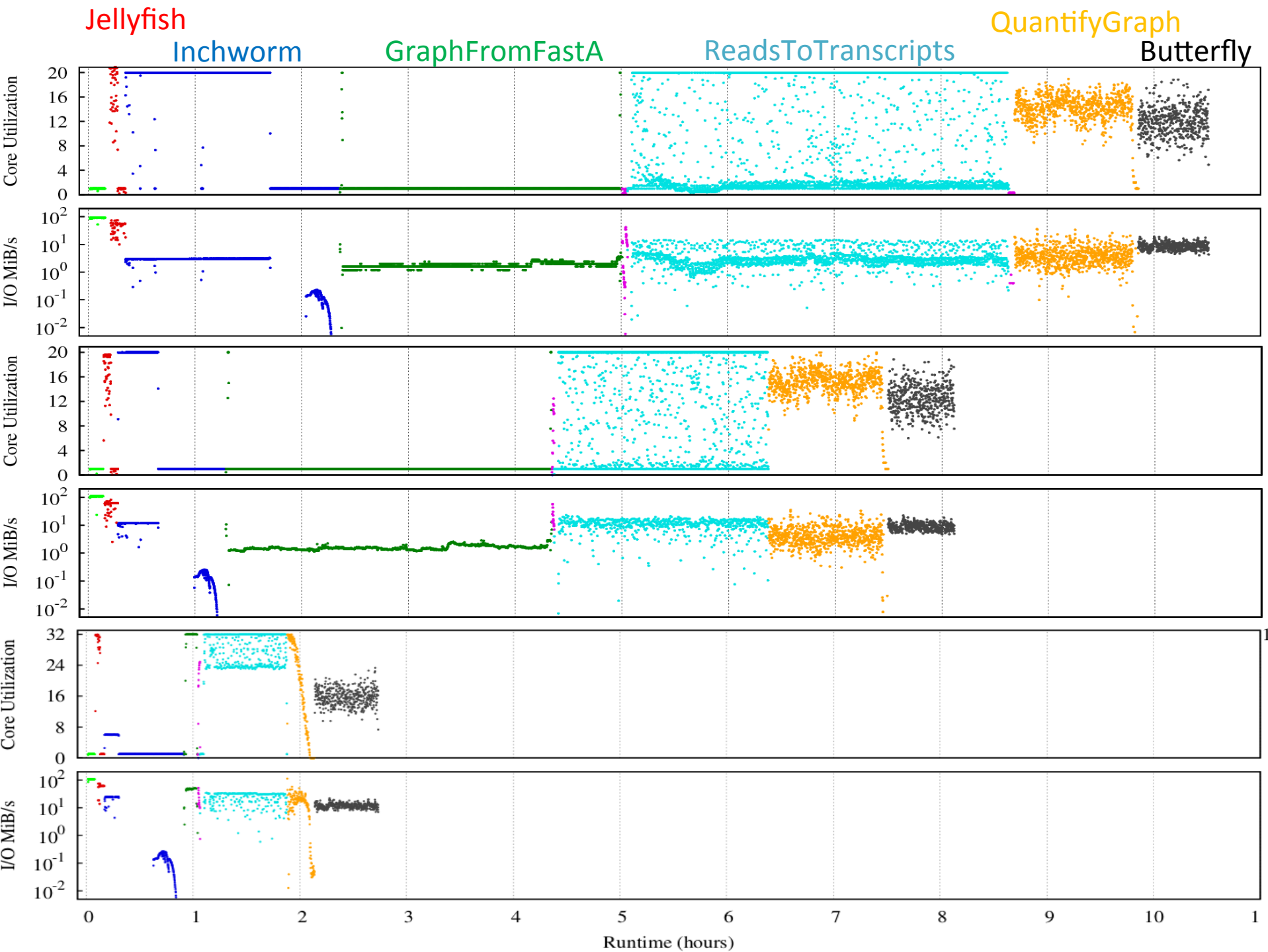


Trinity, developed at the [Broad Institute](#) and the [Hebrew University of Jerusalem](#), represents a novel method for the efficient and robust de novo reconstruction of transcriptomes from RNA-seq data. Trinity combines three independent software modules: Inchworm, Chrysalis, and Butterfly, applied sequentially to process large volumes of RNA-seq reads. Trinity partitions the sequence data into many individual de Bruijn graphs, each representing the transcriptional complexity at a given gene or locus, and then processes each graph independently to extract full-length splicing isoforms and to tease apart transcripts derived from paralogous genes. Briefly, the process works like so:

- **Inchworm** assembles the RNA-seq data into the unique sequences of transcripts, often generating full-length transcripts for a dominant isoform, but then reports just the unique portions of alternatively spliced transcripts.
- **Chrysalis** clusters the Inchworm contigs into clusters and constructs complete de Bruijn graphs for each cluster. Each cluster represents the full transcriptional complexity for a given gene (or sets of genes that share sequences in common). Chrysalis then partitions the full read set among these disjoint graphs.
- **Butterfly** then processes the individual graphs in parallel, tracing the paths that reads and pairs of reads take within the graph, ultimately reporting full-length transcripts for alternatively spliced isoforms, and teasing apart transcripts that corresponds to paralogous genes.

From: <http://trinityrnaseq.sourceforge.net> - no copyright terms stated





Science results supported

Helped assemble
genomes of

- Pine tree
- Cocoa
- Zooplankton
- Full RNA transcriptome of fruit fly
- Evolution of beetles

en.wikipedia.org/wiki/File:Scarabaeus_viettei_01.jpg
licensed under the [Creative Commons Attribution-Share Alike 3.0 Unported, 2.5 Generic, 2.0 Generic and 1.0 Generic license.](#)



Indiana Clinical and Translational Studies Institute

Funded by the Clinical and Translational Science Award program within National Center for Advancing Translational Sciences (NCATS)

Established in 2008

Partnership between Indiana University, Purdue University and Notre Dame University

Mission - To improve the health and economy of Indiana by:

- Creating a home for translational research
- Building resources to accelerate research
- Training a new cadre of research workforce
- Possession of one of these awards is the difference between having the possibility of being a top ranked medical school and not having that possibility



Indiana CTSI Partnerships

- Eli Lilly & Company
- Cook Medical Group
- Roche Diagnostics
- Takeda Pharmaceuticals
- Biocrossroads
- State of Indiana Government
- Indiana University Health
- Eskenazi Health
- Roudabush Veterans Affairs Medical Center
- Regenstrief Institute
- Fairbanks Foundation





Are You an Investigator Needing Help?

- About
- News & Events
- Research Resources
- Training & Education
- Grants & Funding
- Community Engagement
- Volunteer for Research
- Tools
- Contribute

- Support & Feedback
- Citing CTSI
- Newsletter
- Grants Login

This website is powered by HUBzero. Supported by Indiana and Purdue universities. Design by IUPUI; concept by Tufts CTSI. Copyright © 2012 Indiana CTSI

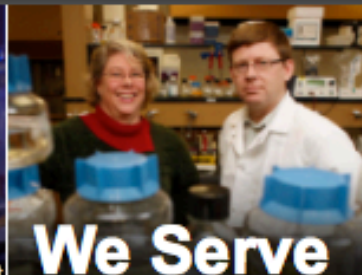
Contact: info@indianactsi.org



We Provide

- Clinical Research Center Access
- Collaboration Opportunities
- Community Access
- Biospecimen Access
- Education & Training
- Grants & Funding

- A Letter of Support
- Project Development Help
- Research Studies Information
- Statistics & Data Management Help
- Recruitment, Feasibility & Data Request
- Core Technology & Lab Resources



We Serve

- Community Members
- Health Providers
- Institutional Partners
- New Faculty Members
- Research Coordinators
- Research Volunteers
- Scientists & Researchers
- Students & Postdocs

News Alerts

Indiana Drug Discovery Alliance seeks applications for new RFA - due July 1

Request for Applications – Strategic Pharma-Academic Research Consortium – Letters of Intent due July 29

Mark your calendars! Indiana CTSI Sixth Annual meeting scheduled Sept. 26

The Indiana CTSI is funded by the National Institutes of Health through its National Center for Advancing Translational Sciences

CTSA Clinical & Translational Science Awards
Translating Discoveries to Medical Practice

Featured Item

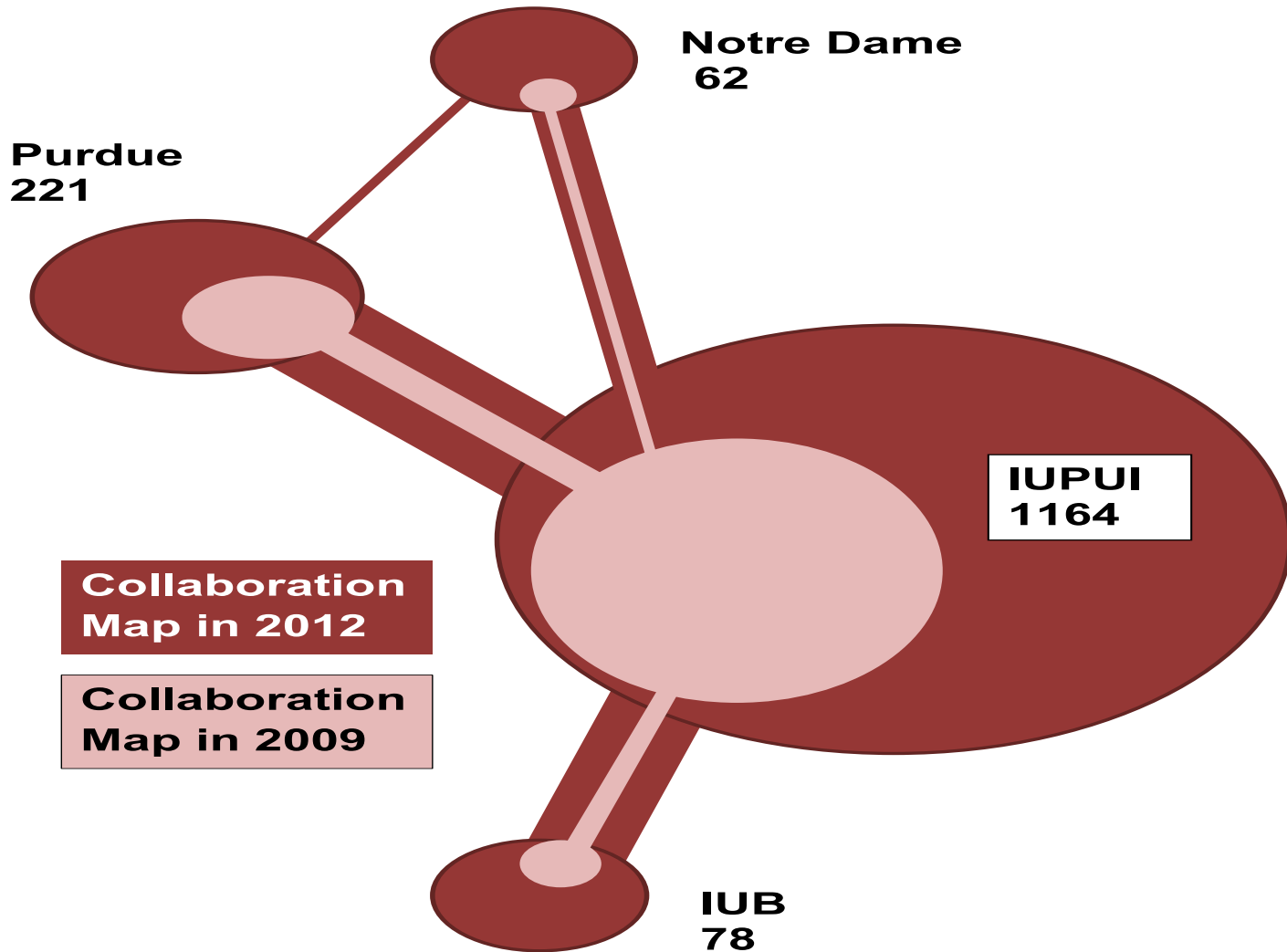


Indiana CTSI launches multi-state consortium to spark translational medicine collaborations. [\[Read More\]](#)

- News & Announcements
- Events Calendar

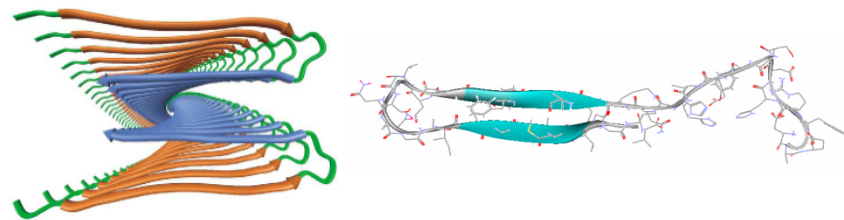
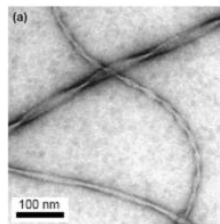
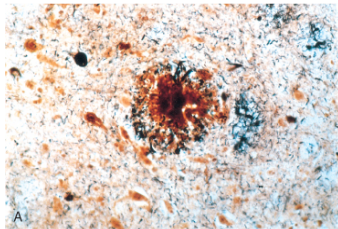
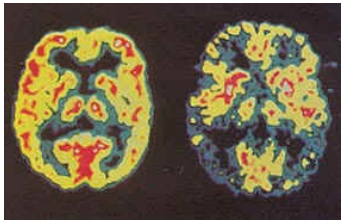


Indiana CTSI – Increased Collaborations



Alzheimer's disease

- Alzheimer's disease is associated with **amyloid plaques** in brain tissue.
- These plaques are formed by the aggregation of **short peptides**, the Amyloid- β or A β peptides into insoluble fibrils



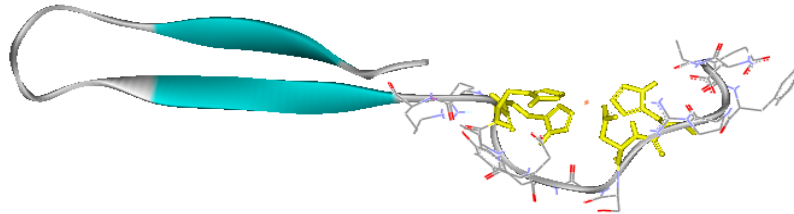
Alzheimer's brain showing buildup of amyloid plaque

Proposed models for structure of single fibril and monomer peptide



Where is the Weak Spot?

Mookie Baik of IU has proposed for the **a high-resolution structure** based on massive molecular modeling efforts.



Building on this work, we can now ask:

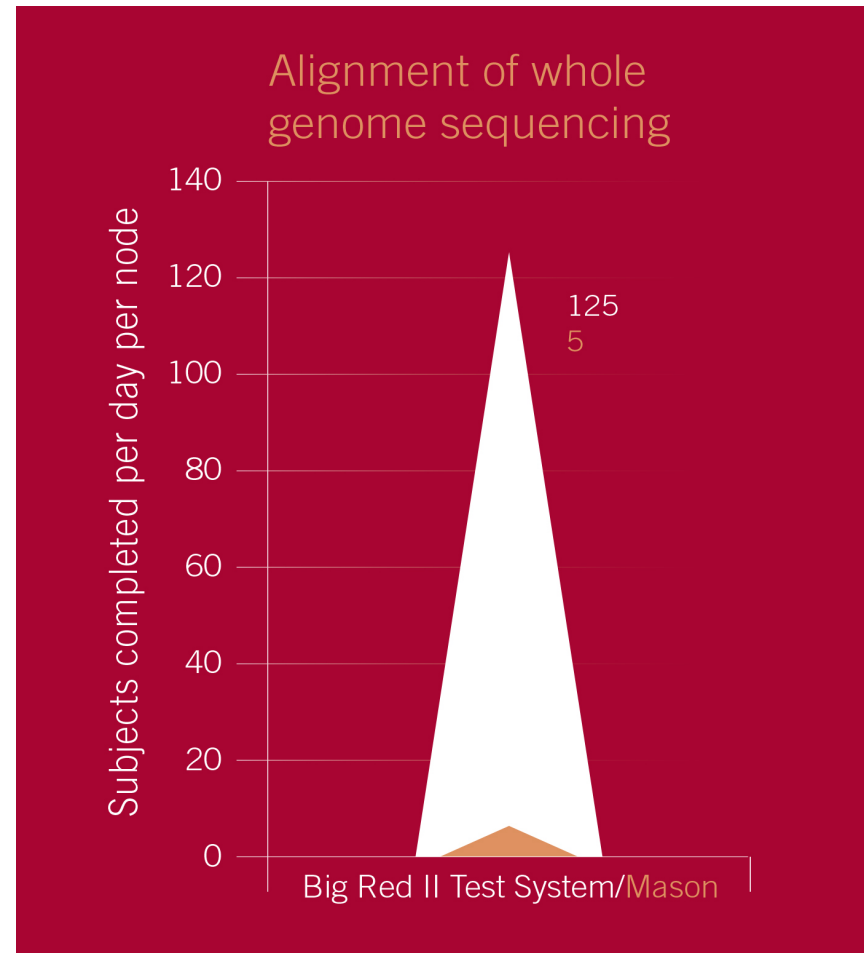
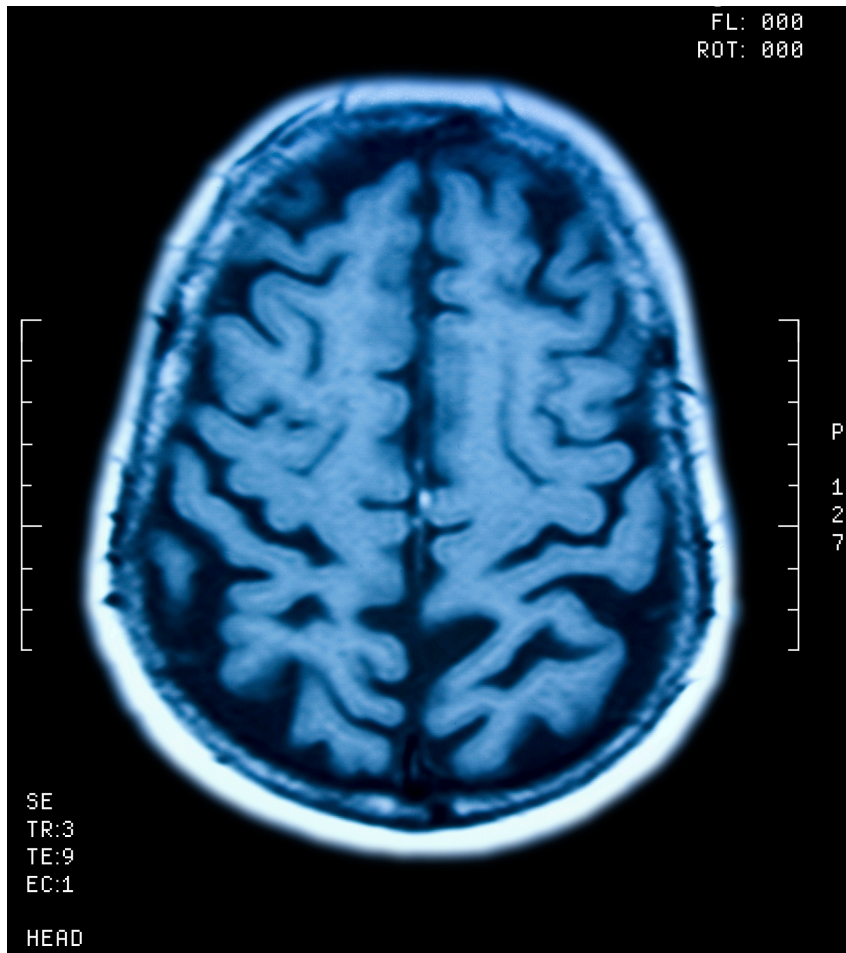
- **Which part of Ab is most critical for the structural integrity?**
- That is: Which part should we attack to cause **maximal damage** and potentially **destroy** the plaques?

Computational approach:

- Computational mutagenesis – swap out Amino Acids and see what disrupts the formation of the amyloids the most?



Alzheimer's – genomes to brain scans to behavioral analyses – looking for the genetic basis of Alzheimer's



Mistakes we made, things we learned (1)

Mistakes we should try not to repeat

- Some times: too much tactic, not enough strategy (especially at times we were ahead of our faculty)
- Sometimes promising too much first, figuring out how to deliver later (=> too much stress). You have to promise somewhat more than you know how to deliver or you simply won't be at the front edge of technology. The key is 'how much depends upon miracles'?

Things that went wrong that we will repeat as necessary

- Pursuing a strategy and having that strategy collapse for external reasons
- But we try to get good data from the industry and community to improve our guesses



Things the literature tells us

- Technology adoption choices are based on perceived value and perceived ease of use

Things we learned

- ***First and second derivatives matter much more than current location***
- ***Collaborations are important especially early on.***
- Build on your unique capabilities to differentiate your organization
- Your opportunity to distinguish your organization depends upon supporting current & future distinguished researchers Work and responsibility flow to demonstrated competence
- Cloud computing is just a technology trend, and all we need to do is figure out how to deliver and support cloud services effectively



- Information technology can be an important strategic asset for many universities.
- In the coming several years, universities are likely to sort themselves into categories of those that treat IT as a commodity and those that treat IT as a strategic asset.
- ***Life sciences in particular:***
 - *Data management and storage can be particularly strategic in the life sciences, especially curation and archiving of data and analyses*
 - *Cloud computing may be strategic in terms of promoting scientific replicability in life sciences*
 - *Energy efficiency will be increasingly important in the future*
 - ***It is possible right now to be particularly effective in aiding life scientists if you pay close attention to the current needs of life scientists, and expect that those needs will change rapidly over time.***



Thanks!

- This talk represents the results of decades of work by thousands of staff of OVPIT, the groups that report to OVPIT and the predecessors of those groups, and the investment of hundreds of millions of dollars of taxpayer money from residents of Indiana and the US overall. All of these people deserve thaks.
- Thanks to the staff of OVPIT and especially PTI and the Research Technologies Division of University Information Technology Services.
- Thanks especially RT Directors / Senior Leaders (Eric Wernert, Matt Link, Therese Miller, Bill Barnett) and Managers (John Samuel, Stephen Simms, Mike Boyles, David Hancock, Richard Knepper, Matt Allen, Robert Quick, Robert Henschel, Marlon Pierce, Richard LeDuc, Robert Ping, Kristy Kallback-Rose, Ganesh Shankar, and Kurt Seiffert and George Turner, managers / tech leaders emeriti).
- Thanks to colleagues who contributed slides and data: Sue Workman, Rob Lowden. Jill Piedmont, Toni Usrey.
- Thanks to PTI colleagues: Beth Plale, Andrew Lumsdaine, Thomas Sterling, Martin Swany, Geoffrey Fox, Fred Cate, Von Welch.
- **Thanks To Prof.-Dr. Lang for the kind invitation to speak, and to you for your attention**

I never mistake the leader for the team



Questions?

