"Discovering Yourself with Computational Bioinformatics"

Rutgers Discovery Informatics Institute (RDI²) Distinguished Seminar

Rutgers University

New Brunswick, NJ

May 9, 2013

Dr. Larry Smarr

Director, California Institute for Telecommunications and Information Technology

Harry E. Gruber Professor,

Dept. of Computer Science and Engineering

Jacobs School of Engineering, UCSD

UCSD

UCIrvine

Abstract

For over a decade, Calit2 has had a driving vision that healthcare is being transformed into “digitally enabled genomic medicine.” Combined with advances in nanotechnology and MEMS, a new generation of body sensors is rapidly developing. As these real-time data streams are stored in the cloud, cross population comparisons becomes increasingly possible and the availability of biofeedback leads to behavior change toward wellness. To put a more personal face on the "patient of the future," I have been increasingly quantifying my own body over the last ten years. In addition to external markers I also currently track over 100 blood biomarkers and dozens of molecular and microbial variables in my stool. Using my saliva 23andme.com obtained 1 million single nucleotide polymorphisms (SNPs) in my human DNA. My gut microbiome has been metagenomically sequenced by the J. Craig Venter Institute, yielding 25 billion DNA bases. I will show how one can discover emerging disease states before they develop serious symptoms using this Big Data approach. Hundreds of thousands of supercomputer CPU-hours were used in this voyage of self-discovery.

UCSD
UCIrvine

it²

Where I Believe We are Headed: Predictive, Personalized, Preventive, & Participatory Medicine

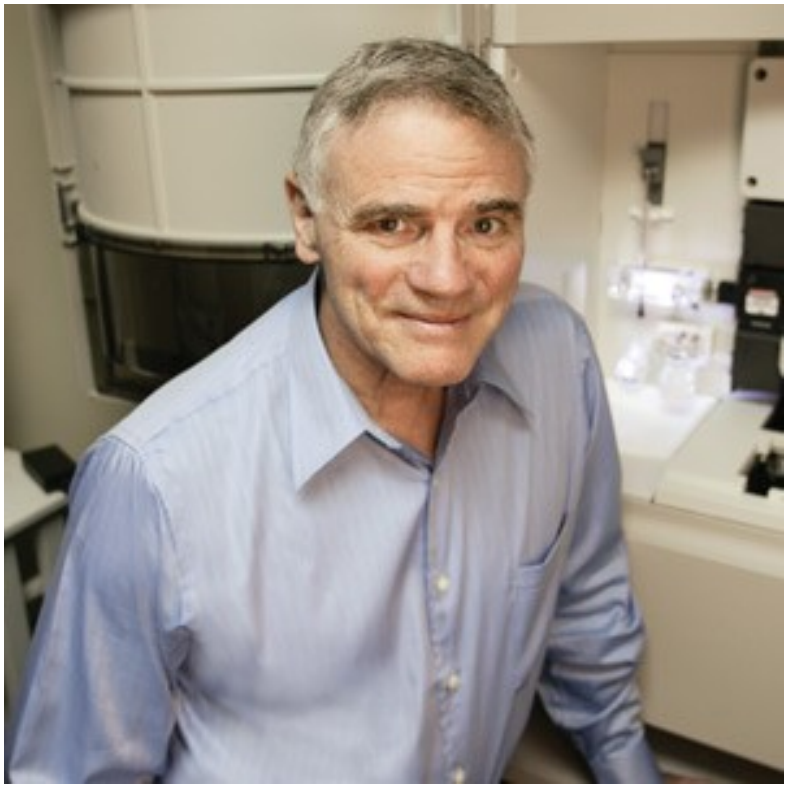
Newsweek

I am Lee Hood's Lab Rat!

A Doctor's Vision of the Future of Medicine

Leroy Hood
NEWSWEEK

From the magazine issue dated Jul 13, 2009



UCSD
UCIrvine

www.newsweek.com/2009/06/26/a-doctor-s-vision-of-the-future-of-medicine.html

it²

Calit2 Has Been Had a Vision of "the Digital Transformation of Health" for a Decade

www.bodymedia.com



- **Next Step—Putting You On-Line!**
  - Wireless Internet Transmission
  - Key Metabolic and Physical Variables
  - Model -- Dozens of Processors and 60 Sensors / Actuators Inside of our Cars
- **Post-Genomic Individualized Medicine**
  - Combine
    - Genetic Code
    - Body Data Flow
  - Use Powerful AI Data Mining Techniques

UCSD
UCIrvine

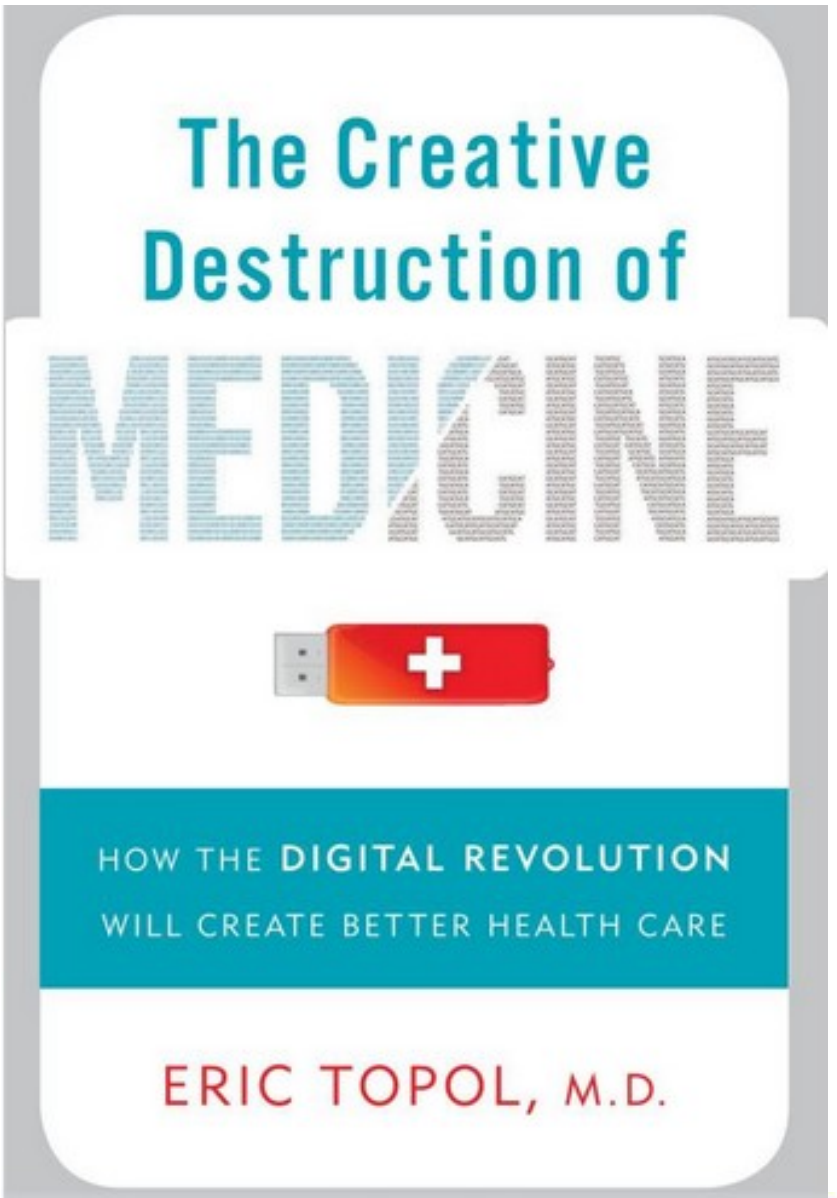
The Content of This Slide from 2001 Larry Smarr Calit2 Talk on Digitally Enabled Genomic Medicine

it²

The Calit2 Vision of Digitally Enabled Genomic Medicine is an Emerging Reality



July/August 2011



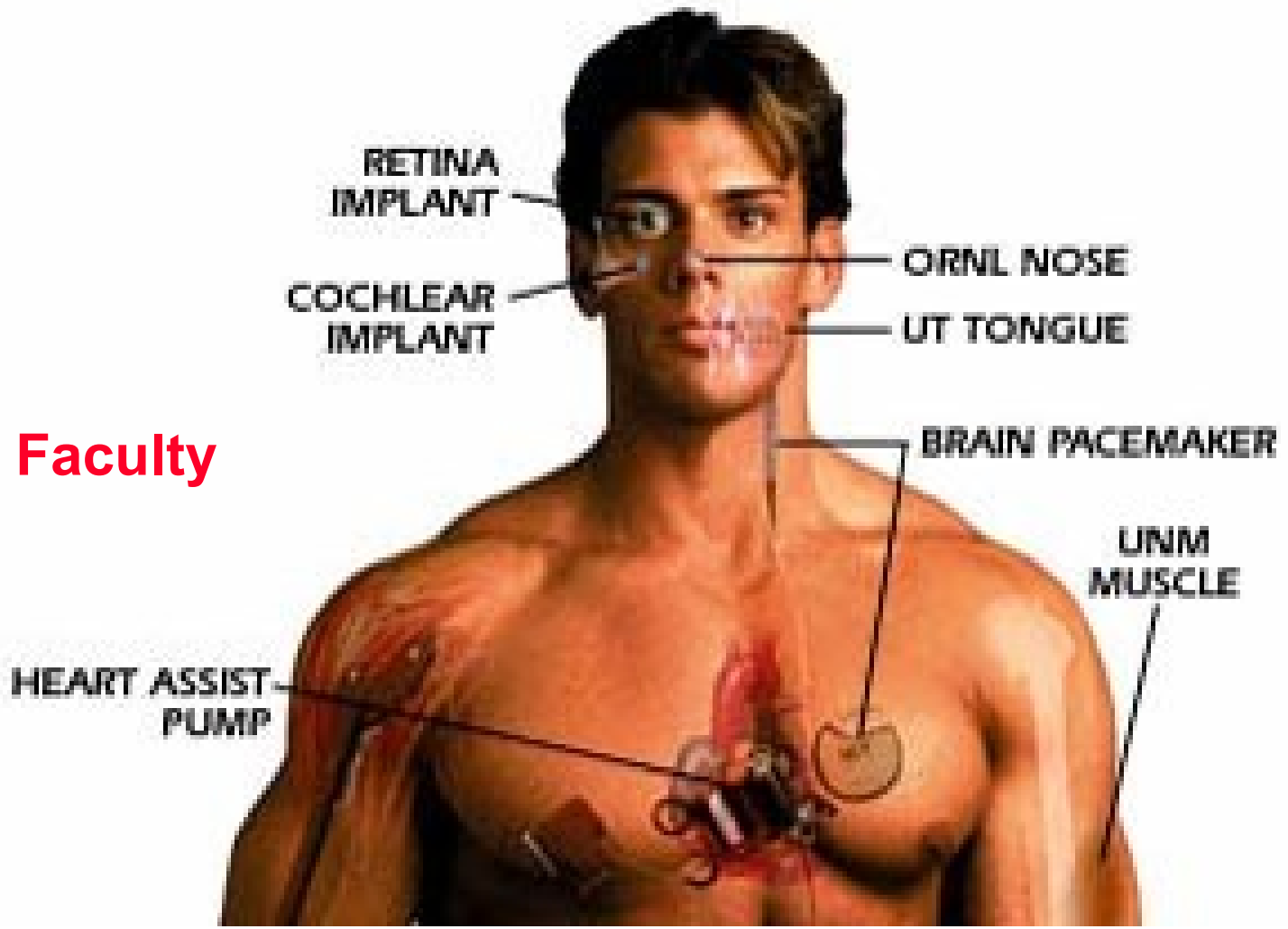
February 2012

UCSD
UCIrvine

it²

Lifechips--Merging Two Major Industries: Microelectronic Chips & Life Sciences

65 UCI Faculty



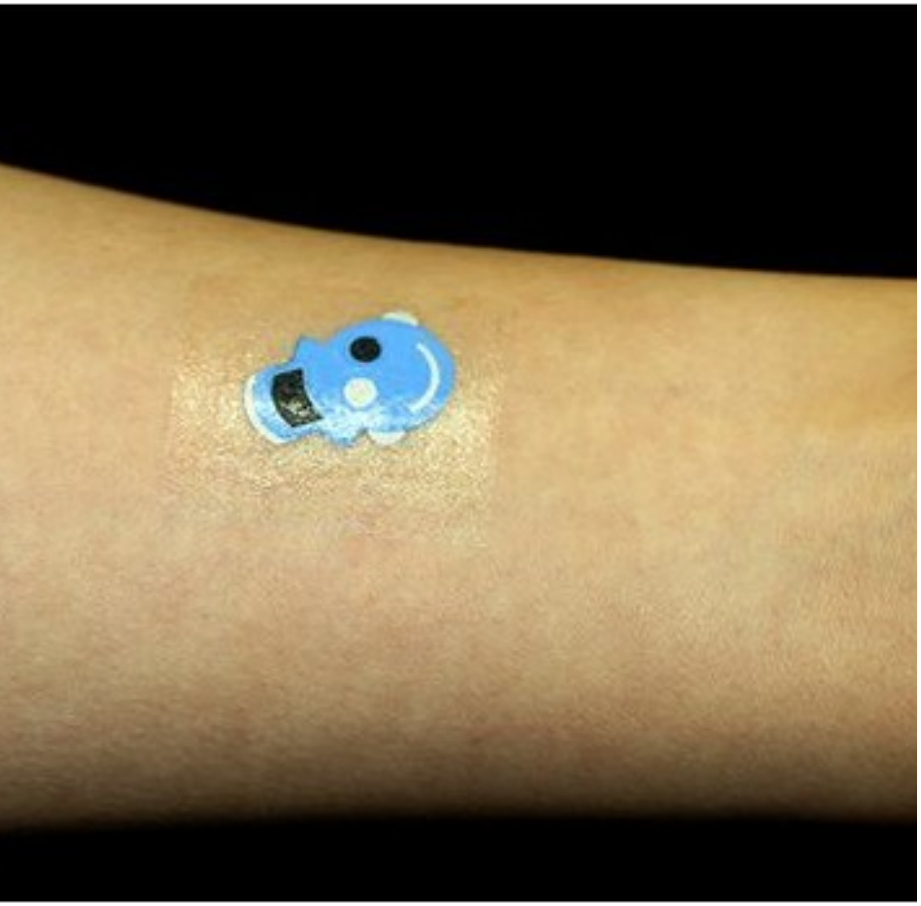
UCSD
UCIrvine

lifechips
university of california, irvine
www.lifechips.org

it²

Temporary Tattoo Biosensors Can Measure pH and Lactate in Sweat

From the UCSD Jacobs School of Engineering
Laboratory for Nanobioelectronics-Prof. Joe Wang





UCSD
UCIrvine

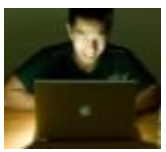
www.jacobsschool.ucsd.edu/news/news_releases/release.sfe?id=1353

it²

CitiSense –UCSD NSF Grant for Fine-Grained "Exposome" Sensing Using Cell Phones



4oz
30 compounds

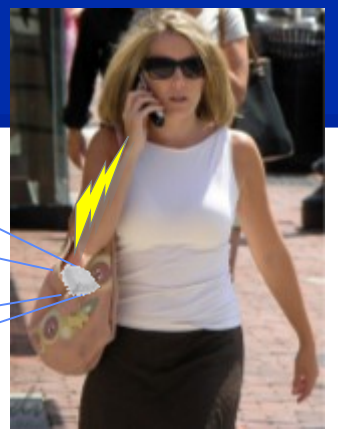


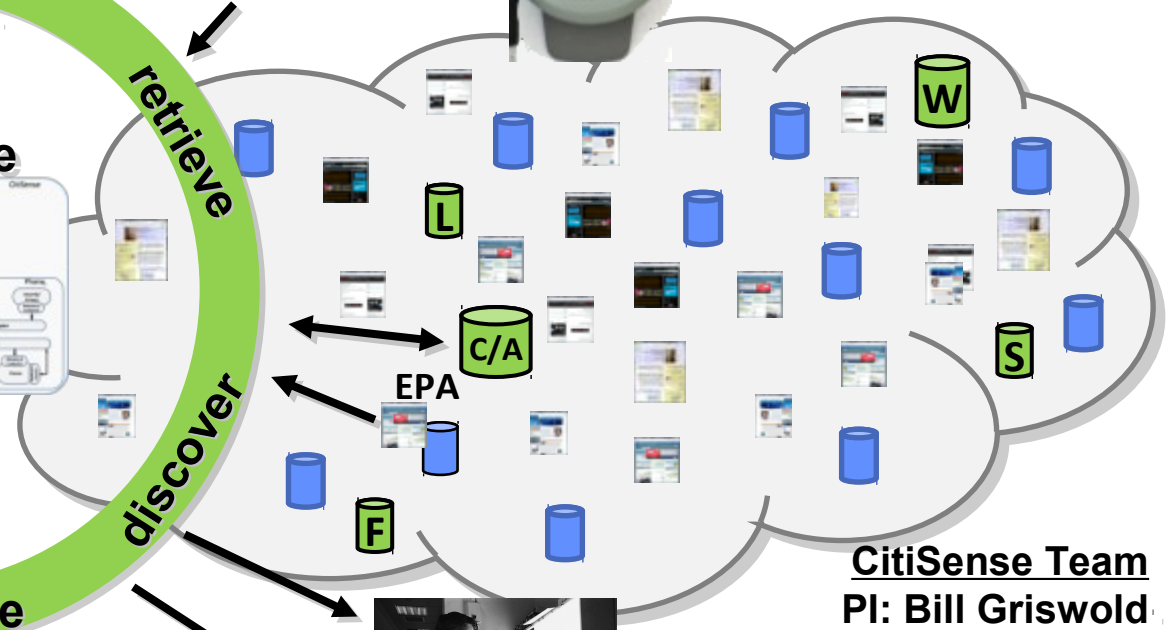
Seacoast Sci.



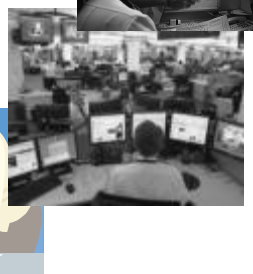
Intel MS











CitiSense Team
PI: Bill Griswold
Ingolf Krueger
Tajana Simunic Rosing
Sanjoy Dasgupta
Hovav Shacham
Kevin Patrick

CW PHS
CENTER FOR WIRELESS & POPULATION HEALTH SYSTEMS

it²

CitiSense Atmospheric Sensor Platform: Sensors Will Miniaturize and Diversify





UCSD
UCIrvine

www.jacobsschool.ucsd.edu/news/news_releases/release.sfe?id=1353

it²

I Arrived in La Jolla in 2000 After 20 Years in the Midwest and Decided to Move Against the Obesity Trend

Age 41



Age 51



Age 61



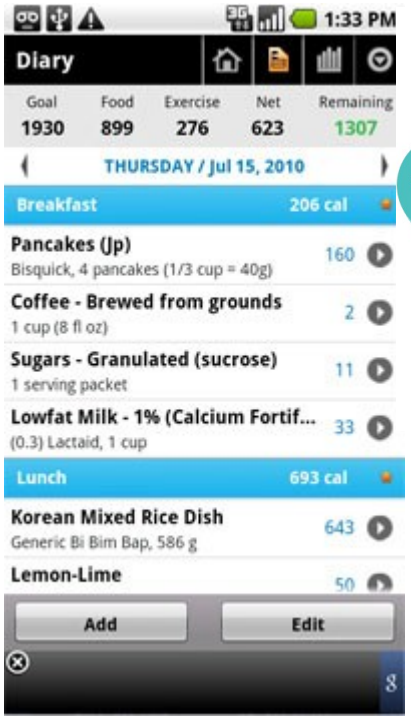
I Reversed My Body's Decline By Quantifying and Altering Nutrition and Exercise

UCSD
UCIrvine

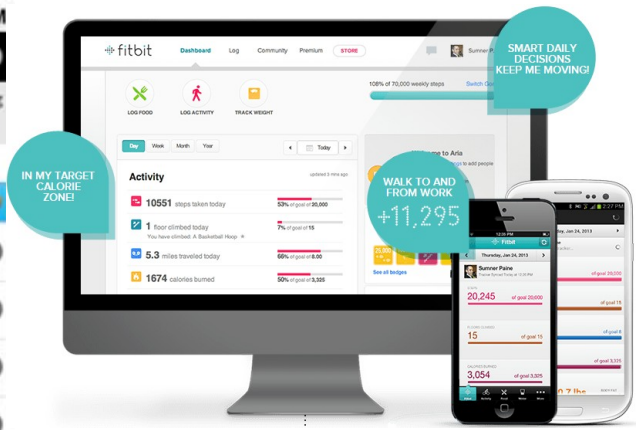
http://lsmarr.calit2.net/repository/LS_reading_recommendations_FiRe_2011.pdf

it²

Challenge-Develop Standards to Enable MashUps of Personal Sensor Data Across Private Clouds



MyFitnessPal-
Calories Ingested



FitBit -
Daily Steps &
Calories Burned



Withing/iPhone-
Blood Pressure





EM Wave PC-
Stress



Azumio-Heart Rate



Zeo-Sleep

UCSD
UCIrvine

it²

From Measuring Macro-Variables to Measuring Your Internal Variables

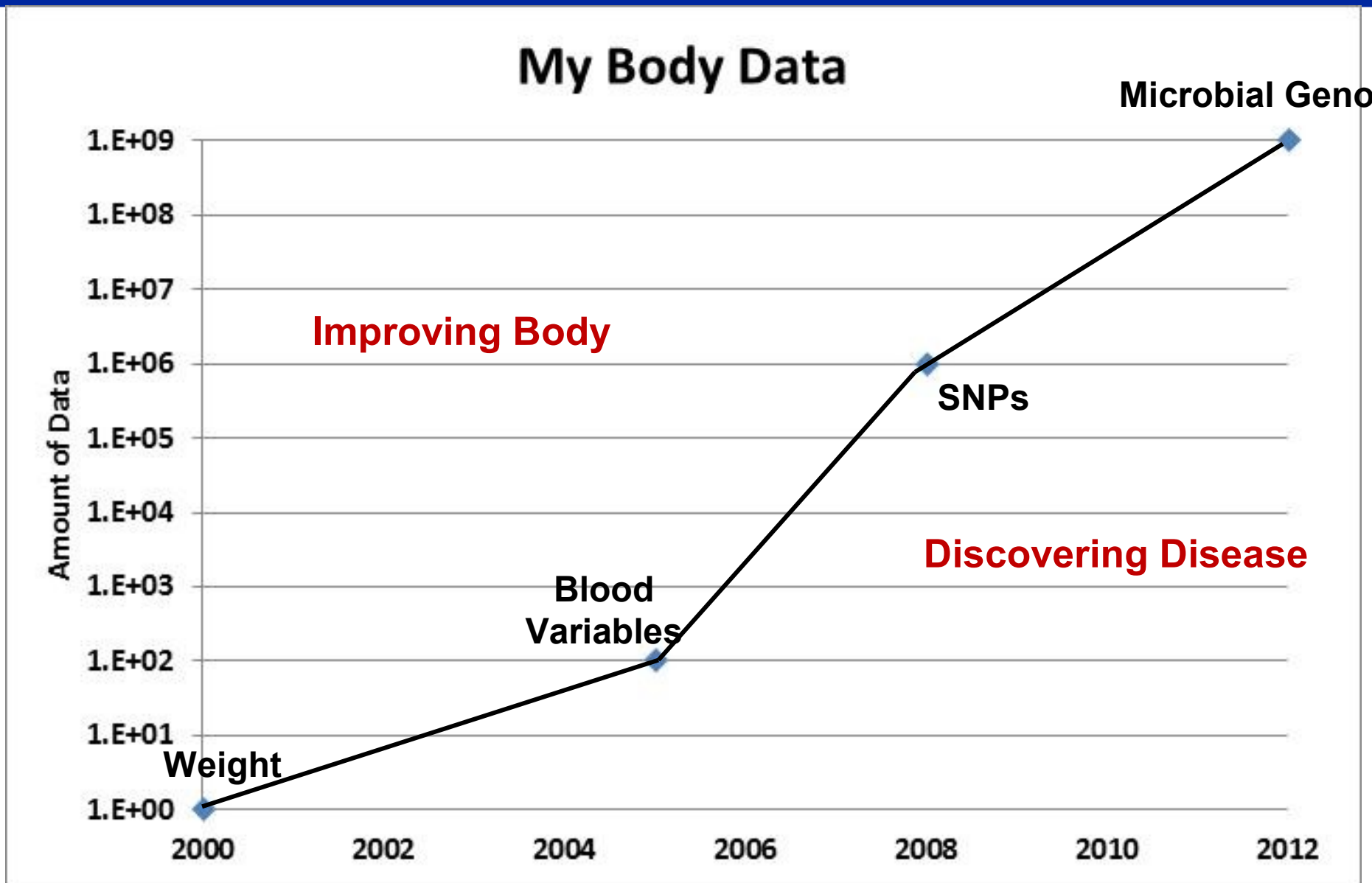


www.technologyreview.com/biomedicine/39636

UCSD
UCIrvine

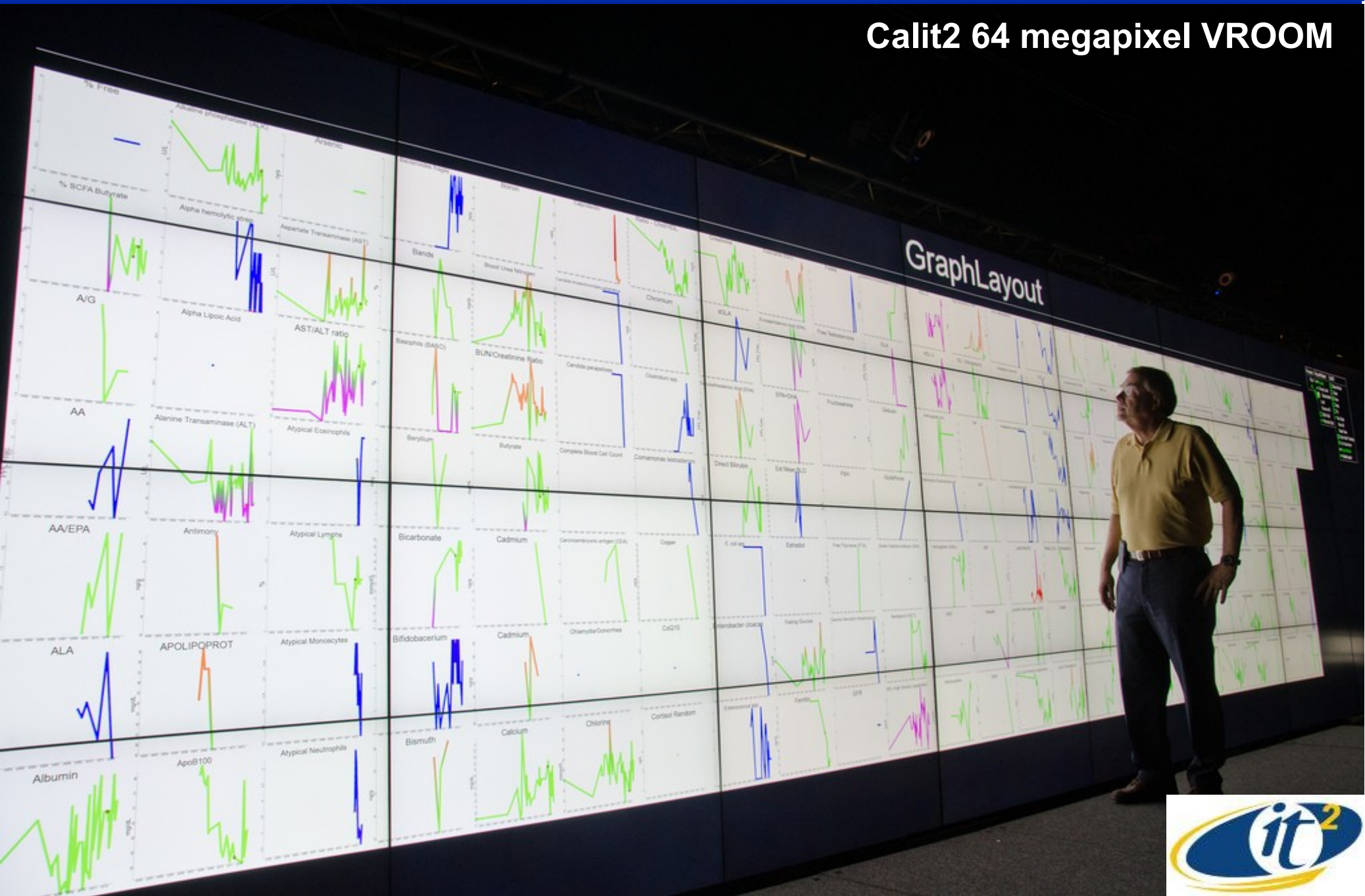
it²

From One to a Billion Data Points Defining Me: The Exponential Rise in Body Data in Just One Decade!

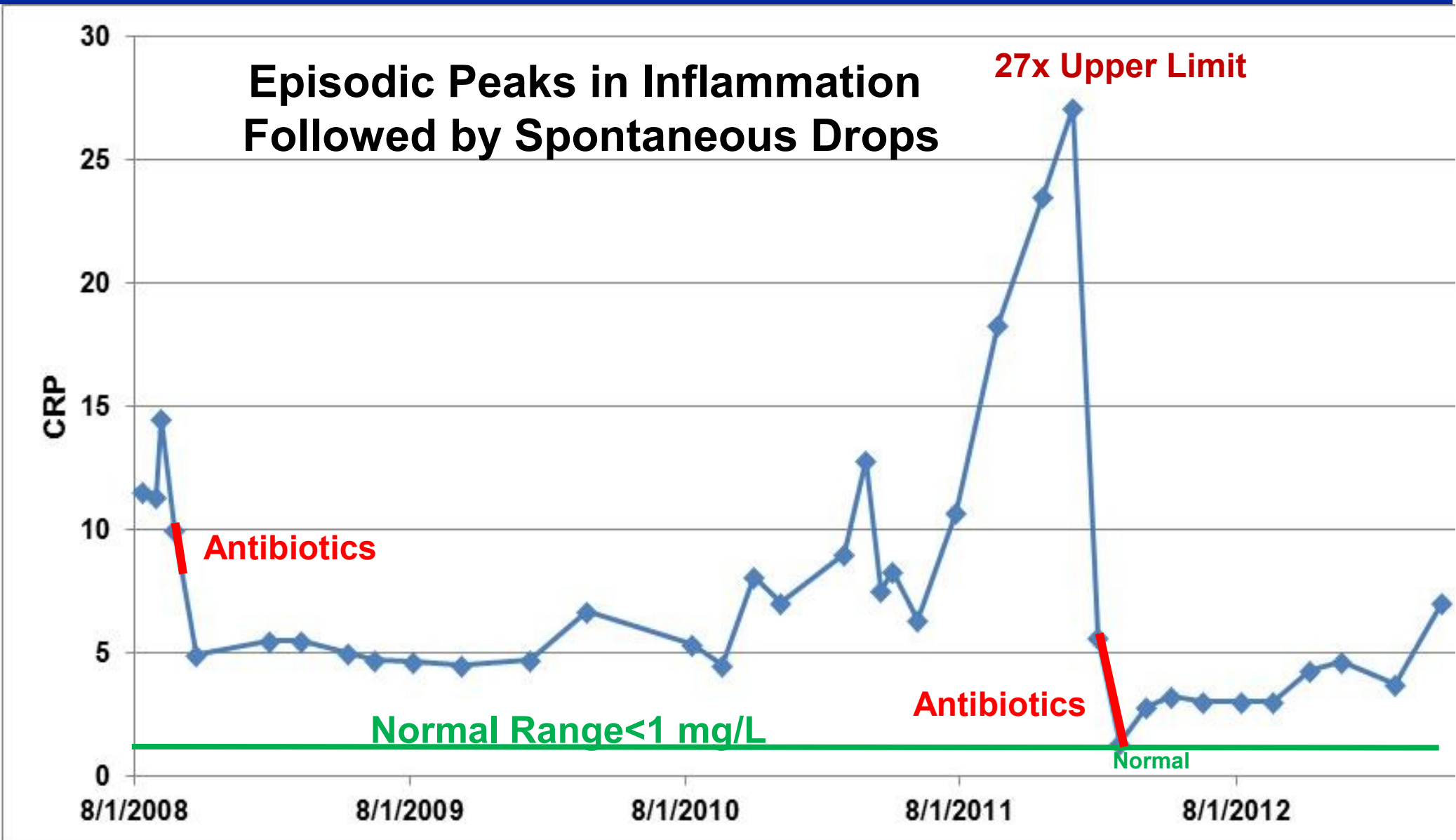


UCSD
UCIrvine

Visualizing Time Series of 150 LS Blood and Stool Variables, Each Over 5-10 Years



Only One of My Blood Measurements Was Far Out of Range--Indicating Chronic Inflammation

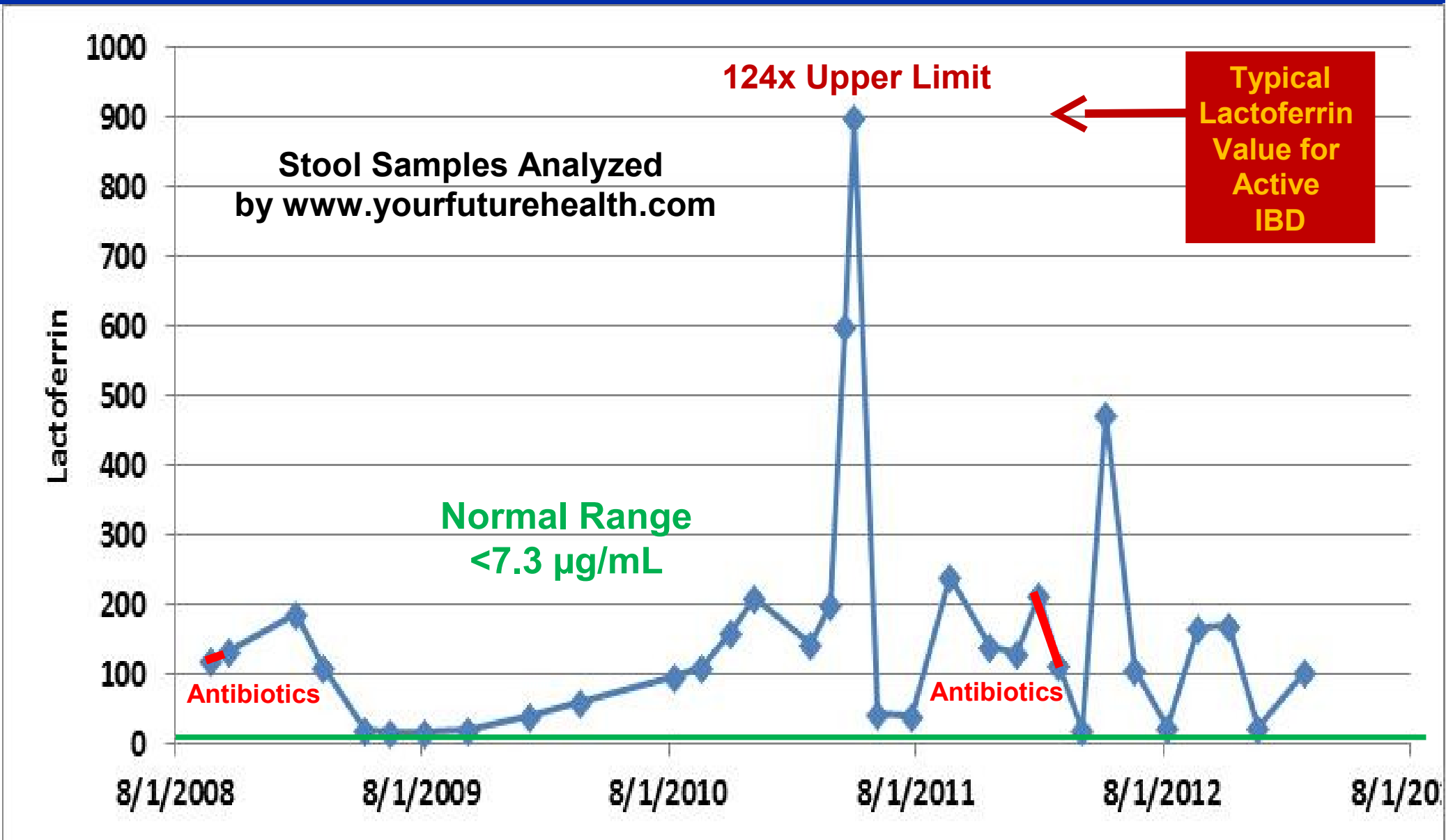


UCSD
UCIrvine

Complex Reactive Protein (CRP) is a Blood Biomarker for Detecting Presence of Inflammation

it²

High Values of Lactoferrin (Shed from Neutrophils) From Stool Sample Suggested Inflammation in Colon

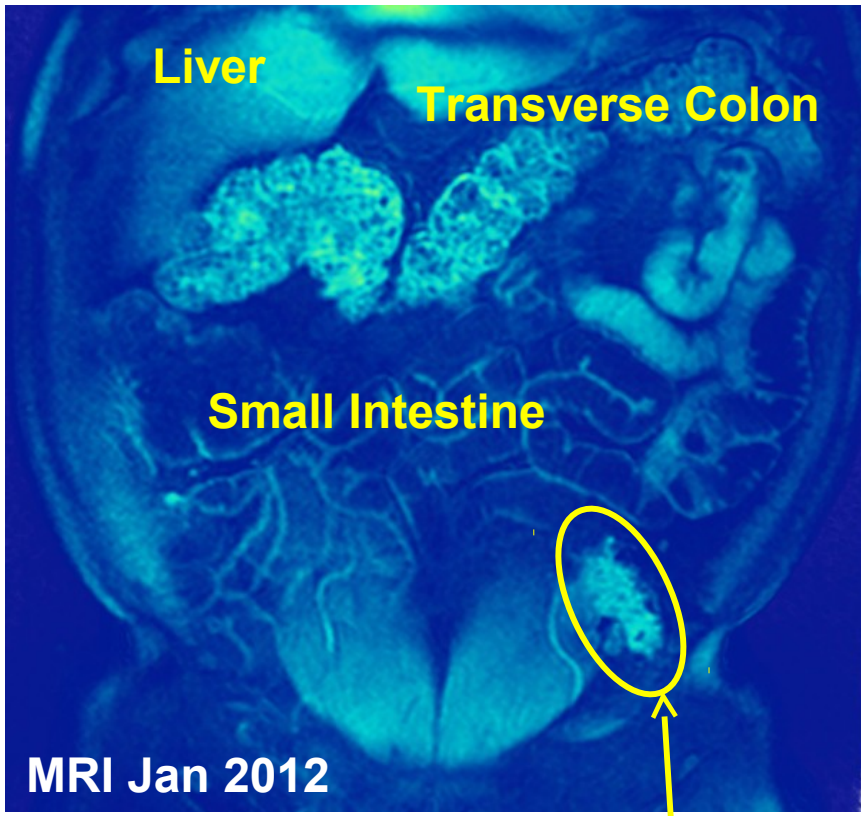


UCSD UCIrvine

Lactoferrin is a Sensitive and Specific Biomarker for Detecting Presence of Inflammatory Bowel Disease (IBD)

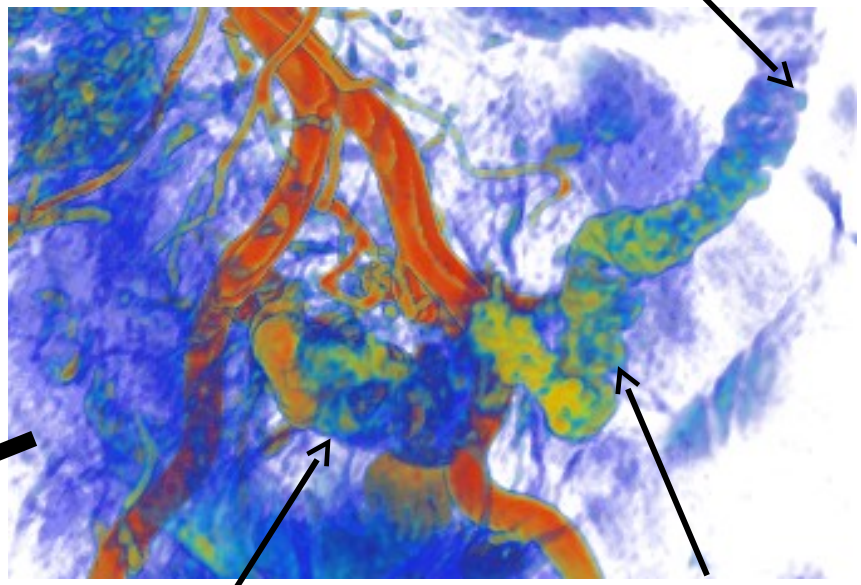
it²

Confirming the IBD (Crohn's) Hypothesis: Finding the "Smoking Gun" with MRI Imaging



I Obtained the MRI Slices From UCSD Medical Services and Converted to Interactive 3D Working With Calit2er Jurgen Schulze's DeskVOX Software

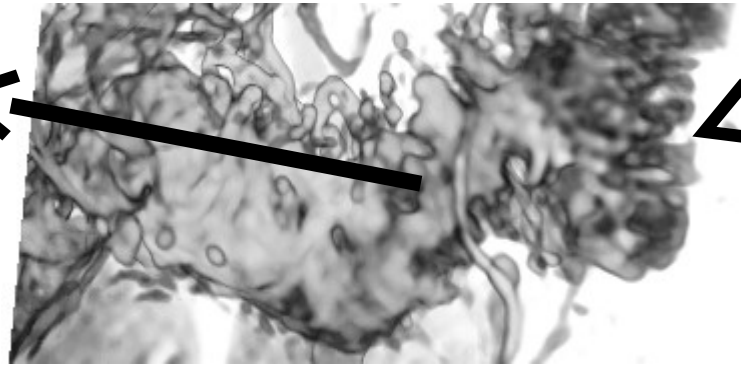
Descending Colon



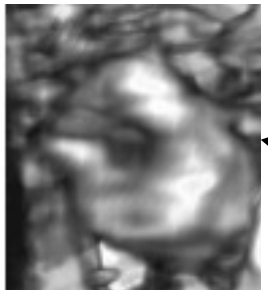
Major Kink

Sigmoid Colon
Threading Iliac Arteries

Diseased Sigmoid Colon



Cross Section



UCSD
UCIrvine

it²

An MRI Shows Sigmoid Colon Wall Thickened Indicating Probable Diagnosis of Crohn's Disease



Why Did I Have an Autoimmune Disease like IBD?

Despite decades of research,
the etiology of Crohn's disease
remains unknown.
Its pathogenesis may involve
a complex interplay between
host genetics,
immune dysfunction,
and microbial or environmental factors.
--The Role of Microbes in Crohn's Disease

So I Set Out to Quantify All Three!

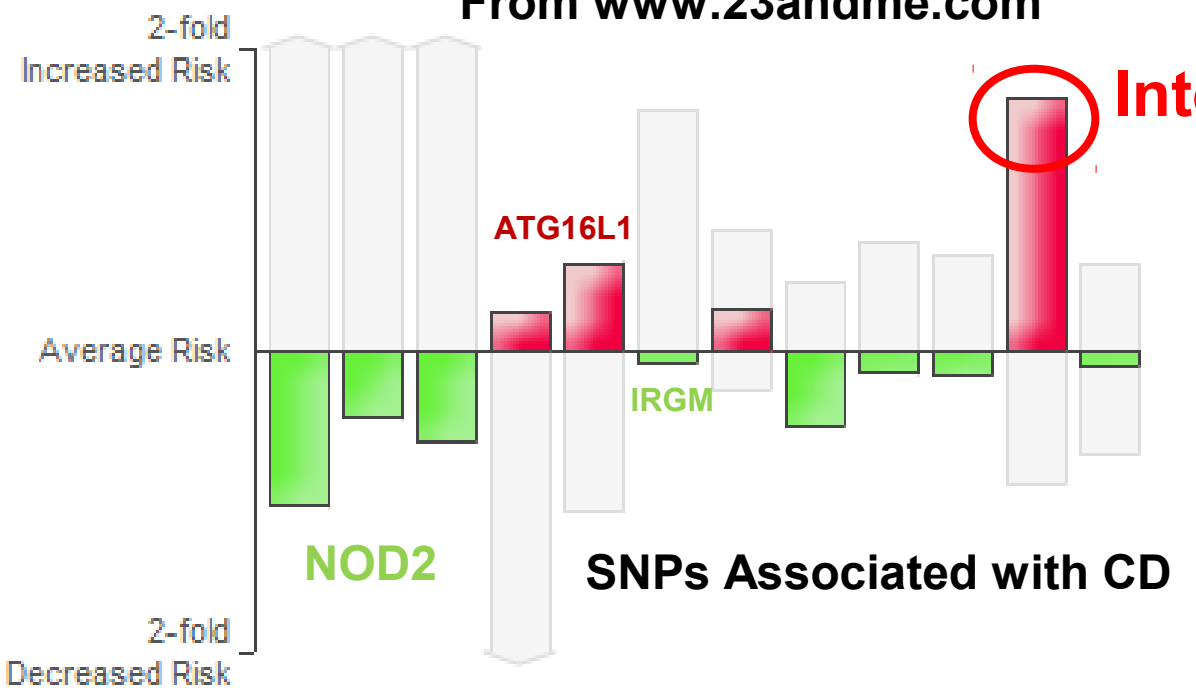
UCSD
UCIrvine

Paul B. Eckburg & David A. Relman
Clin Infect Dis. 44:256-262 (2007)

it²

I Wondered if Crohn's is an Autoimmune Disease, Did I Have a Personal Genomic Polymorphism?

From www.23andme.com



Polymorphism in Interleukin-23 Receptor Gene — 80% Higher Risk of Pro-inflammatory Immune Response

Now Comparing 163 Known IBD SNPs with 23andme SNP Chip

News and Views

Nature Medicine **13**, 26 - 28 (2007)
doi:10.1038/nm0107-26

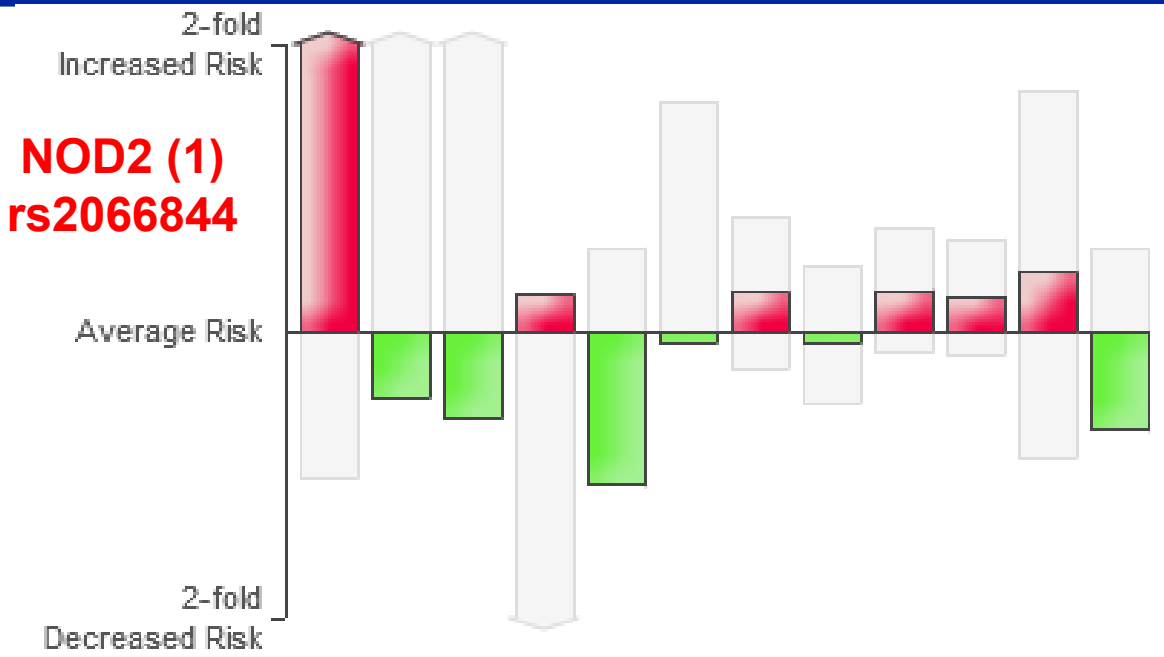
IL-23: a master regulator in Crohn disease

Markus F Neurath[1]

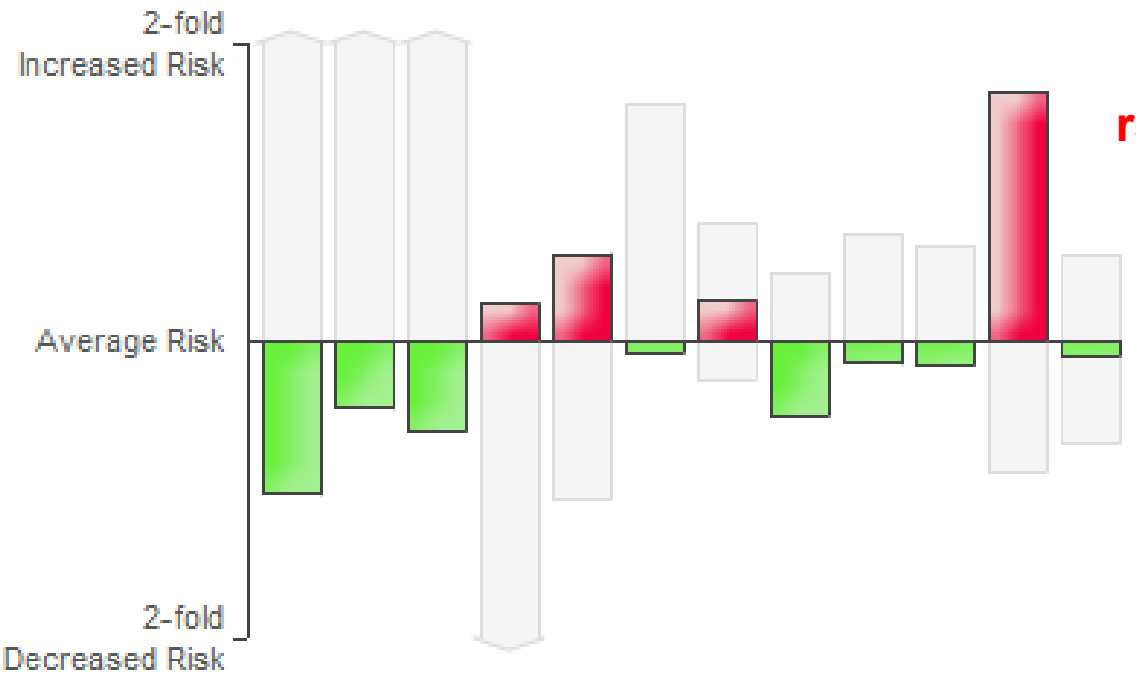
UCSD
UCIrvine

it²

Crohn's May be a Related Set of Diseases Driven by Different SNPs



Female
CD Onset
At 20-Years Old



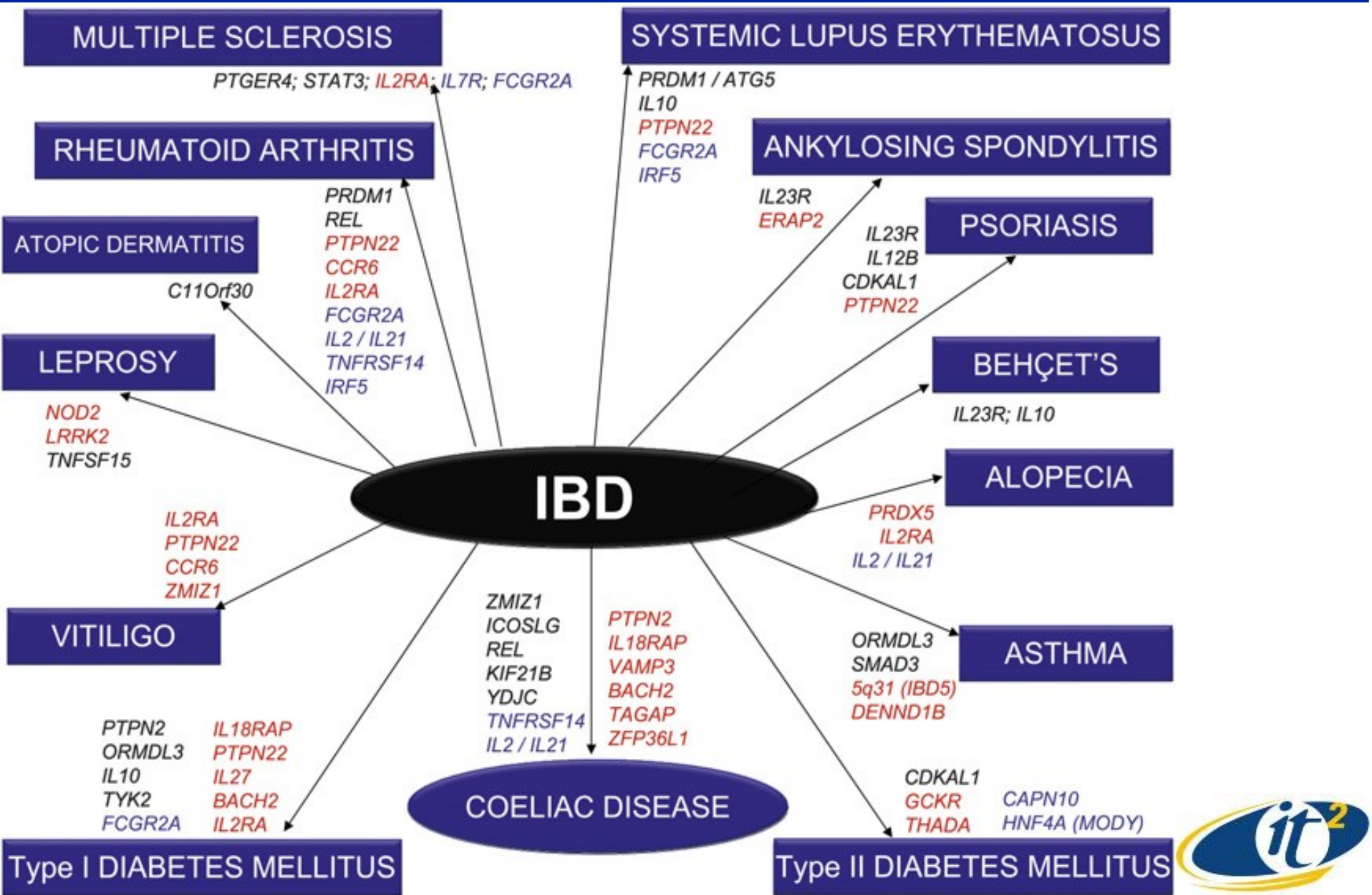
Il-23R
rs1004819

Me-Male
CD Onset
At 60-Years Old

UCSD
UCIrvine

Autoimmune Disease Overlap from SNP GWAS

Gut Lees, et al. 60:1739-1753 (2011)



Imagine Crowdsourcing 23andme SNPs For Even a Small Portion of Crohnology!



UCSD
UCIrvine

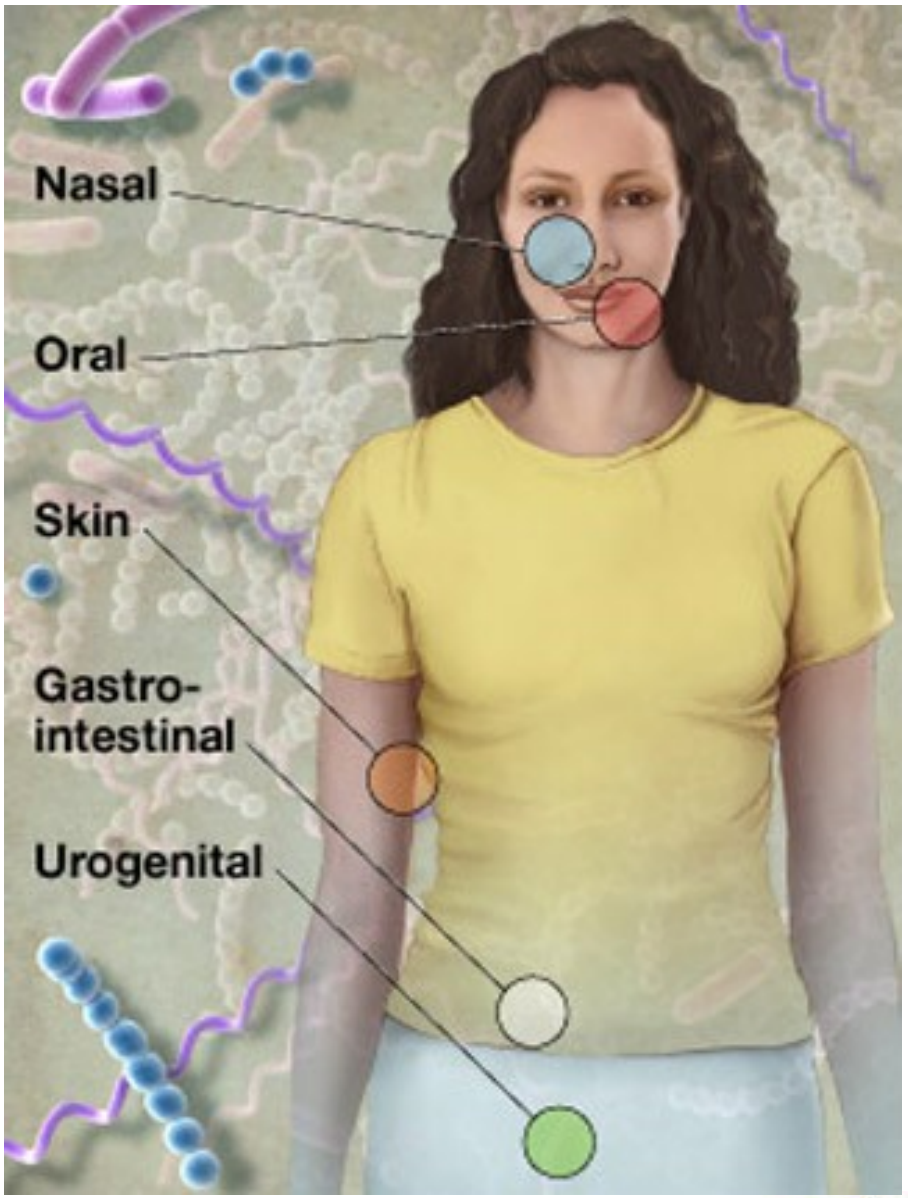
www.crohnology.com

it²

But the Human Genome Contains Less Than 1% of the Bodies Genes

Human Microbiome Project

The Total Number of These Bacterial Cells is 10 Times the Number of Human Cells in Your Body



The NIH Common Fund
WE ACCELERATE DISCOVERY.

UCSD
UCIrvine

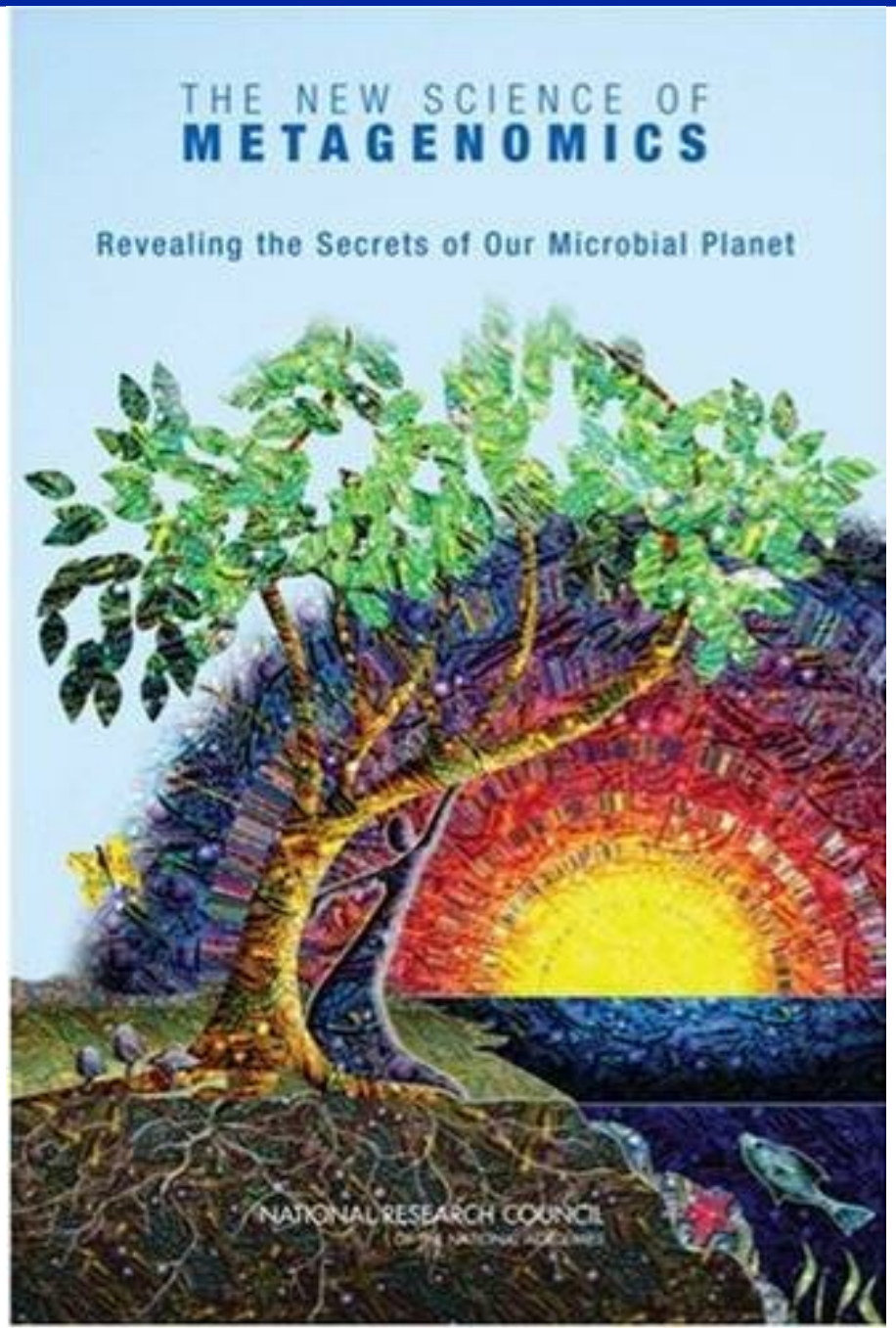
http://commonfund.nih.gov/hmp/

it²

But How Can You Determine Which Microbes Are Within You?

NRC Report:

Metagenomic data should be made publicly available in international archives as rapidly as possible.



"The emerging field of metagenomics, where the DNA of entire communities of microbes is studied simultaneously, presents the greatest opportunity -- perhaps since the invention of the microscope – to revolutionize understanding of the microbial world." –

National Research Council
March 27, 2007

THE NATIONAL ACADEMIES
Advisers to the Nation on Science, Engineering, and Medicine

UCSD
UCIrvine

it²

Calit2 Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA)







UCSD Triton



NSF/SDSC
Gordon



NSF/SDSC
Trestles



NSF/TACC
Lonestar



NSF/TACC
Ranger



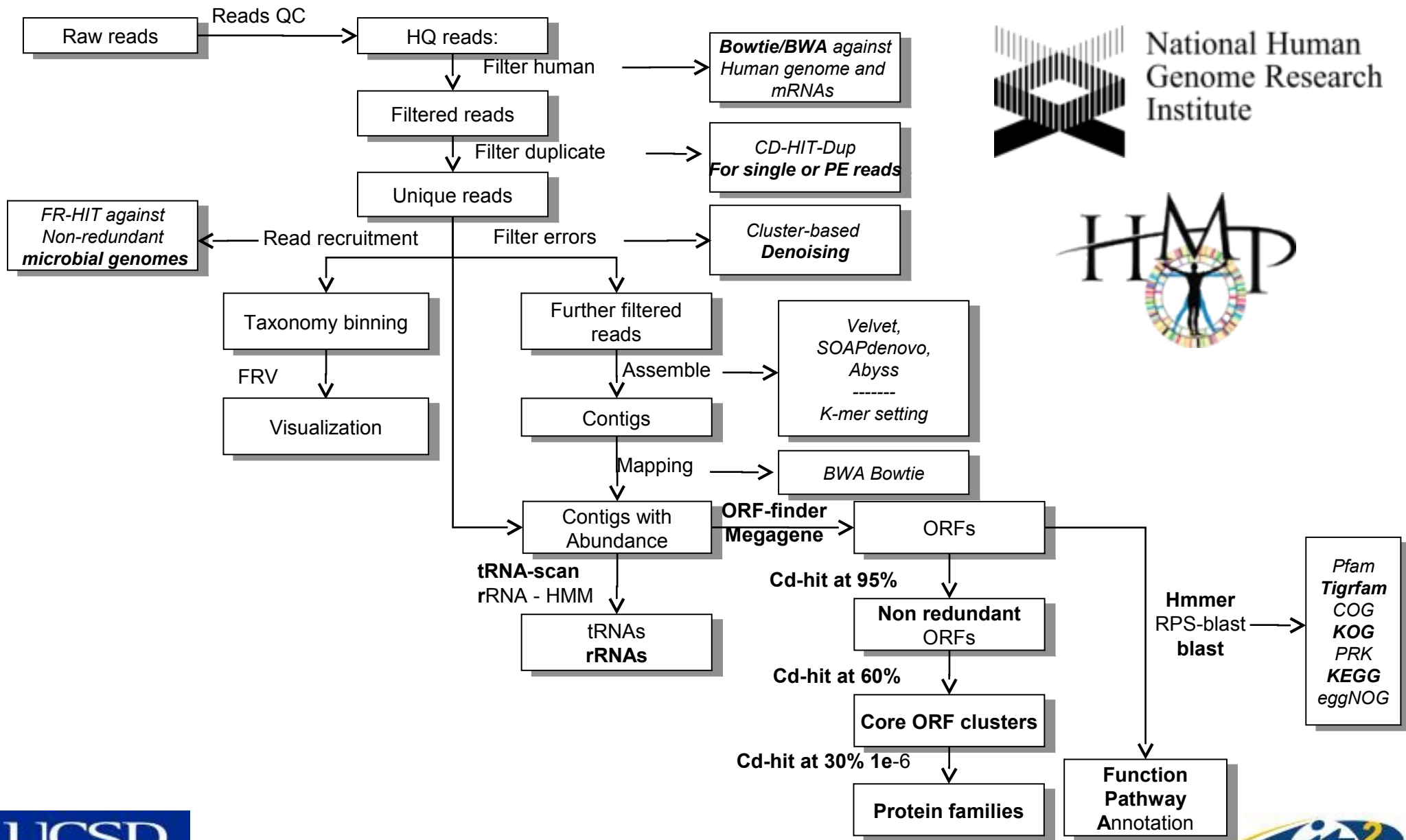
NSF/RCAC
Steele

Infrastructure Services Extend CAMERA Computations to 3rd Party Compute Resources

Source:
Jeff Grethe,
CRBS, UCSD

>5000 Users
>90 Countries

CAMERA and NIH Funded Weizhong Li Group's Metagenomic Computational NextGen Sequencing Pipeline



National Human Genome Research Institute

HMP

UCSD UCIrvine

PI: (Weizhong Li, UCSD):
NIH R01HG005978 (2010-2013, $1.1M)

it²

We Used SDSC's Gordon Data-Intensive Supercomputer to Analyze a Wide Range of Gut Microbiomes

- **Analyzed Healthy and IBD Patients:**
  - LS, 13 Crohn's Disease & 11 Ulcerative Colitis Patients, + 150 HMP Healthy Subjects
- **Gordon Compute Time**
  - ~1/2 CPU-Year Per Sample
  - > 200,000 CPU-Hours so far
- **Gordon RAM Required**
  - 64GB RAM for Most Steps
  - 192GB RAM for Assembly
- **Gordon Disk Required**
  - 8TB for All Subjects
  - Input, Intermediate and Final Results

Venter Sequencing of LS Gut Microbiome:
230 M Reads
101 Bases Per Read
23 Billion DNA Bases

Enabled by a Grant of Time on Gordon from SDSC Director Mike Norman

SDSC

UCSD UCIrvine

HMP NIH NSF

it²

2012 Was
the Year of Human Microbiome





UCSD
UCIrvine

it²

When We Think About Biological Diversity
We Typically Think of the Wide Range of Animals






But All These Animals Are in One SubPhylum Vertebrata of the Chordata Phylum

UCSD
UCIrvine

All images from Wikimedia Commons.
Photos are public domain or by Trisha Shears & Richard Bartz

Think of These Phyla of Animals When You Consider the Biodiversity of Microbes Inside You



Phylum Chordata



Phylum Cnidaria



Phylum Echinodermata



Phylum Annelida



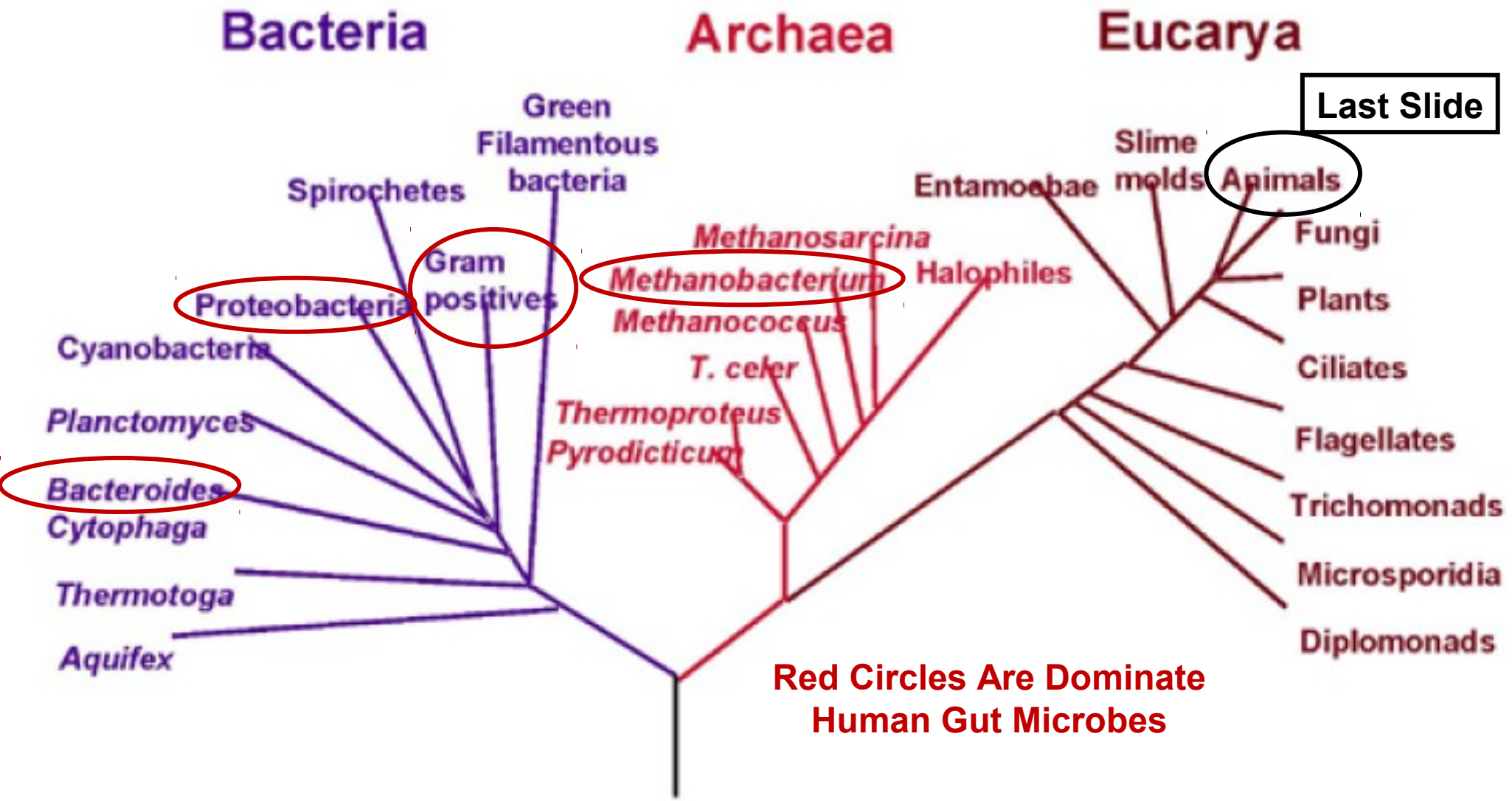
Phylum Mollusca



Phylum Arthropoda

All images from WikiMedia Commons.
Photos are public domain or by Dan Hershman, Michael Linnenbach, Manuae, B_cool

Most Biological Diversity on Earth is in the Microbial World



Evolutionary Distance Derived from Comparative Sequencing of 16S or 18S Ribosomal RNA

Source: Carl Woese, et al

UCSD
UCIrvine

Intense Scientific Research is Underway on Understanding the Human Microbiome



June 8, 2012



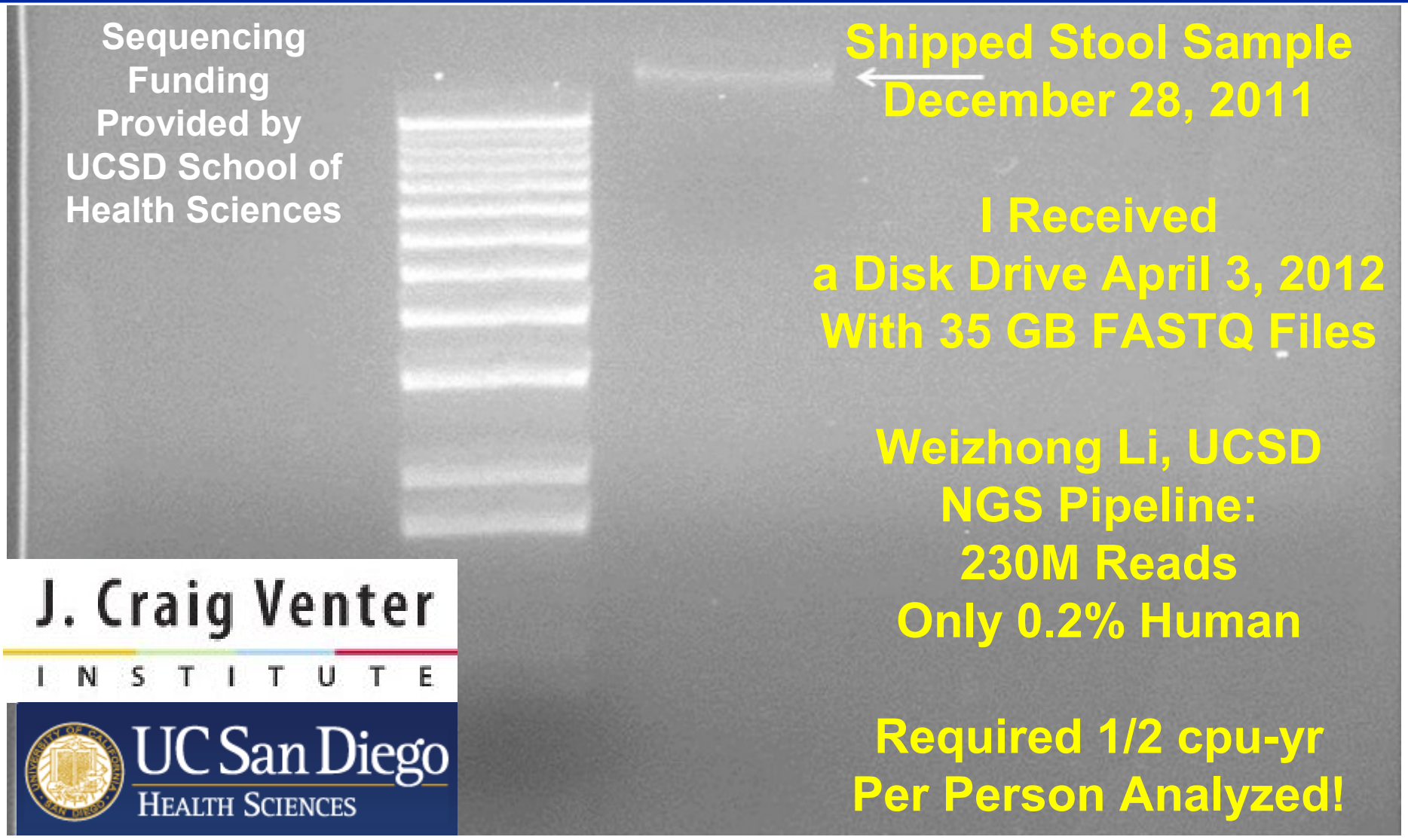
June 14, 2012

UCSD
UCIrvine

From Culturing Bacteria to Sequencing Them

it²

To Map My Gut Microbes, I Sent a Stool Sample to the Venter Institute for Metagenomic Sequencing



Sequencing Funding Provided by UCSD School of Health Sciences

Shipped Stool Sample ← December 28, 2011

I Received a Disk Drive April 3, 2012 With 35 GB FASTQ Files

Weizhong Li, UCSD NGS Pipeline: 230M Reads Only 0.2% Human

Required 1/2 cpu-yr Per Person Analyzed!

J. Craig Venter INSTITUTE

UC San Diego HEALTH SCIENCES

Gel Image of Extract from Smarr Sample-Next is Library Construction

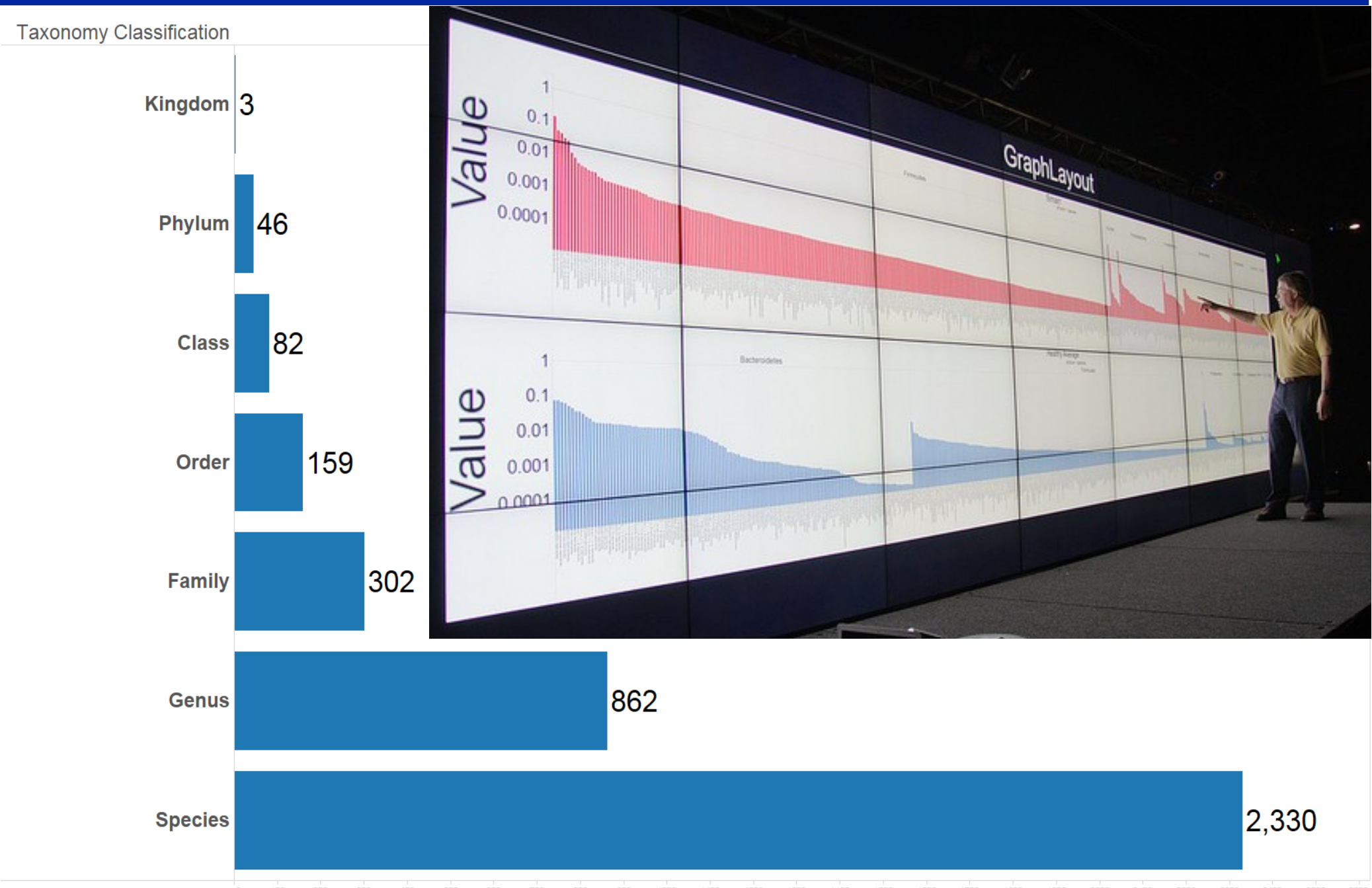
Manny Torralba, Project Lead - Human Genomic Medicine
J Craig Venter Institute

January 25, 2012

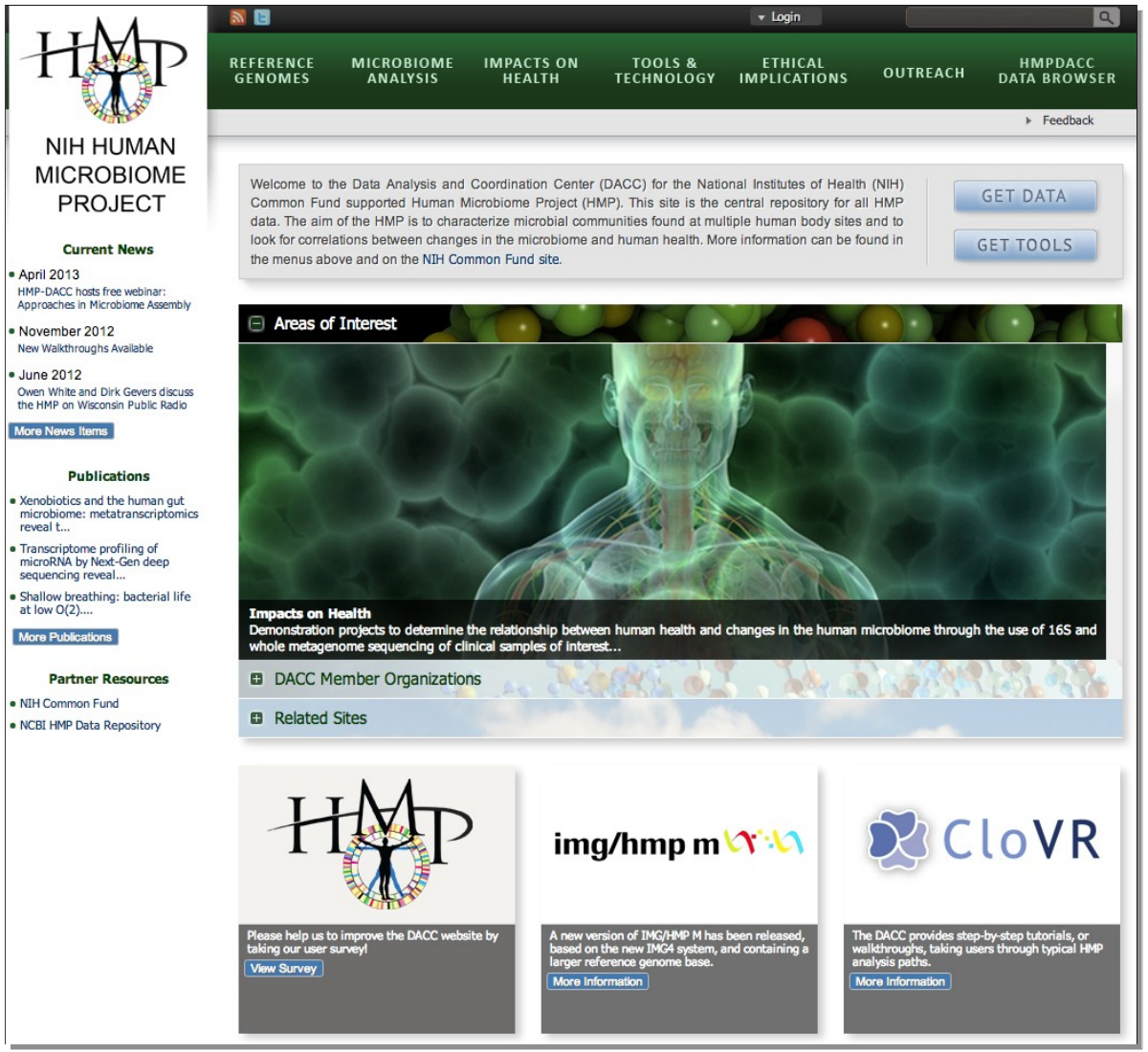
UCSD UCIrvine

it²

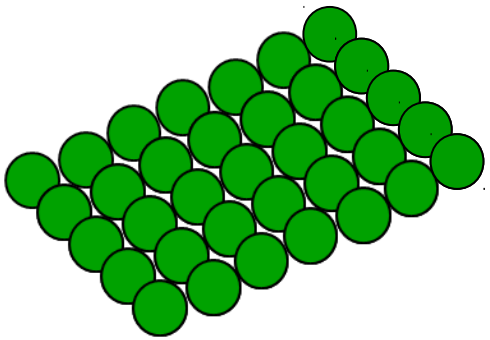
We Computationally Align 230M Illumina Short Reads With a Reference Genome Set & Then Visually Analyze



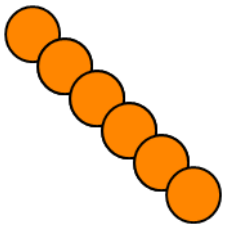
Additional Phenotypes Added from NIH HMP For Comparative Analysis



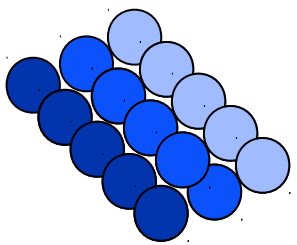
35 "Healthy" Individuals
1 Point in Time



6 Ulcerative Colitis, 1 Point in Time



5 Ileal Crohn's, 3 Points in Time



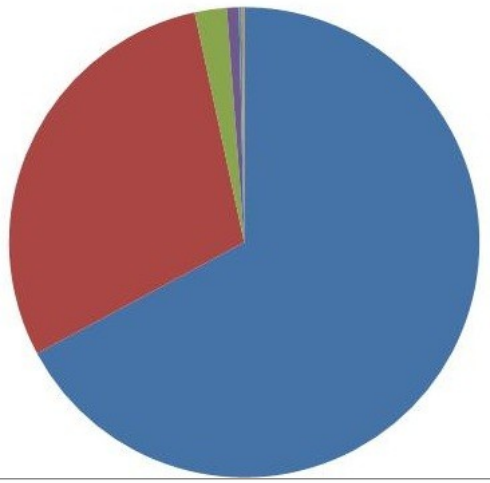
UCSD
UCIrvine

it²

We Find Major Shifts in Microbial Ecology Between Healthy and Two Forms of IBD

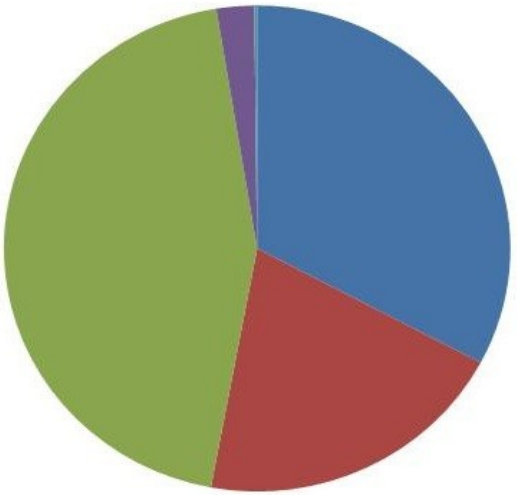
Microbiome "Dysbiosis" or "Mass Extinction"?

Ave. Healthy

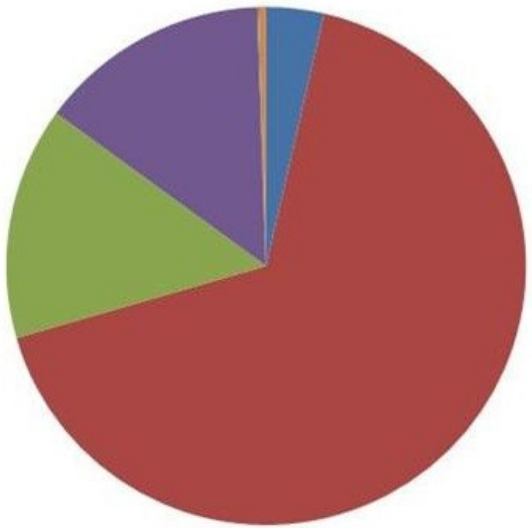


- Bacteroidetes
- Firmicutes
- Proteobacteria
- Actinobacteria
- Fusobacteria
- Euryarchaeota

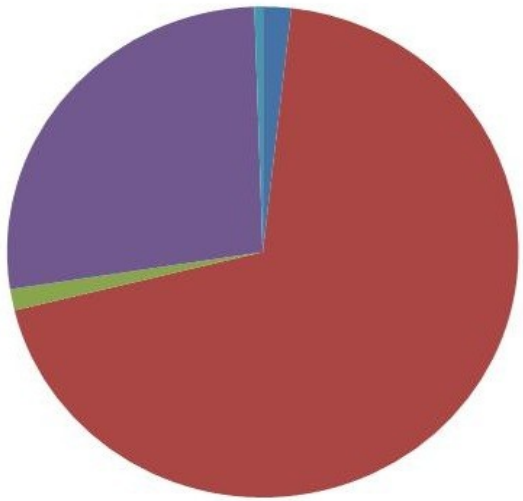
Ave. Ulcerative Colitis



LS Apr 2012



Ave. Crohn's



UCSD UCIrvine

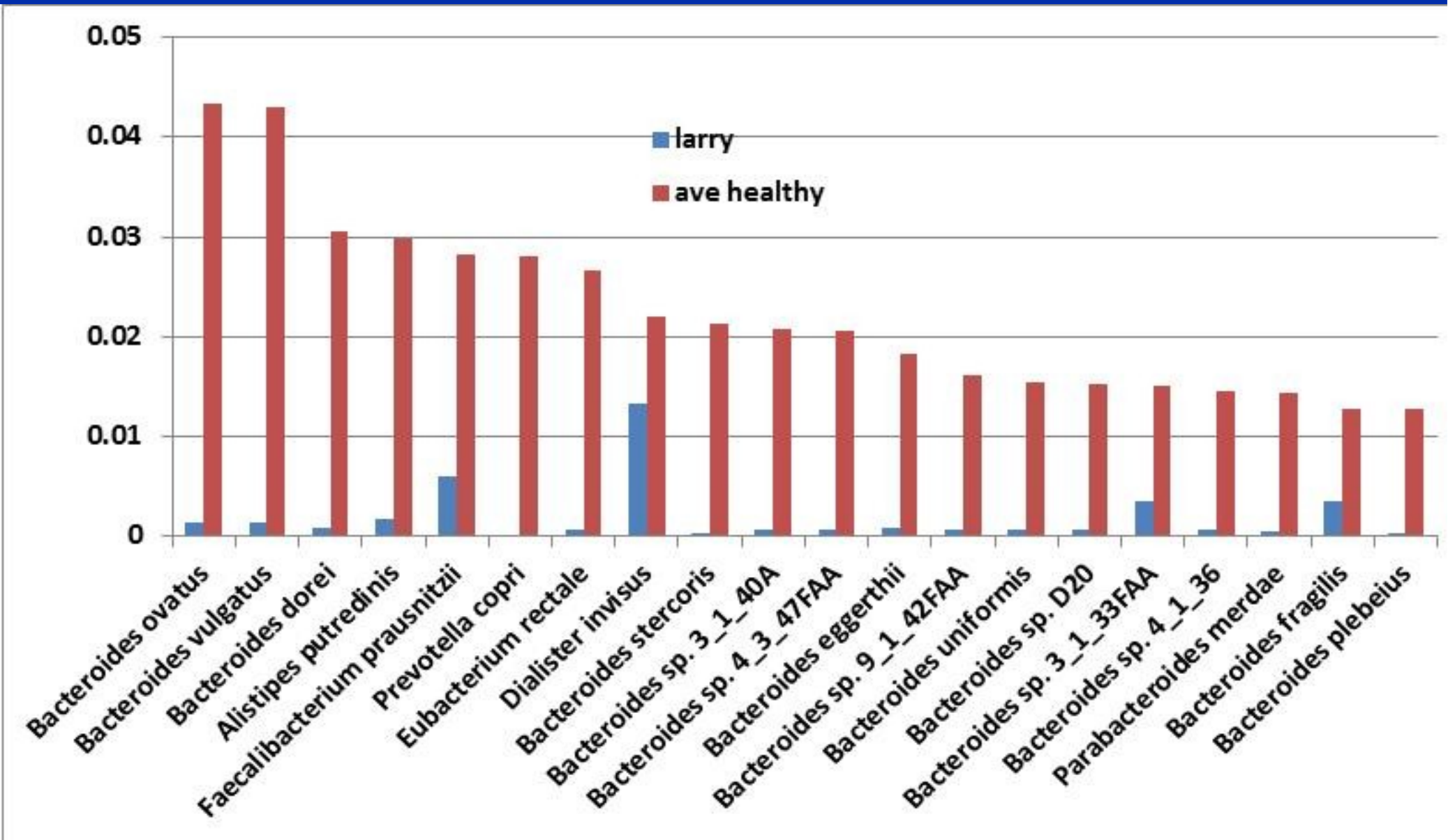
Explosion of Proteobacteria

On the IBD Spectrum

Collapse of Bacteroidetes

it²

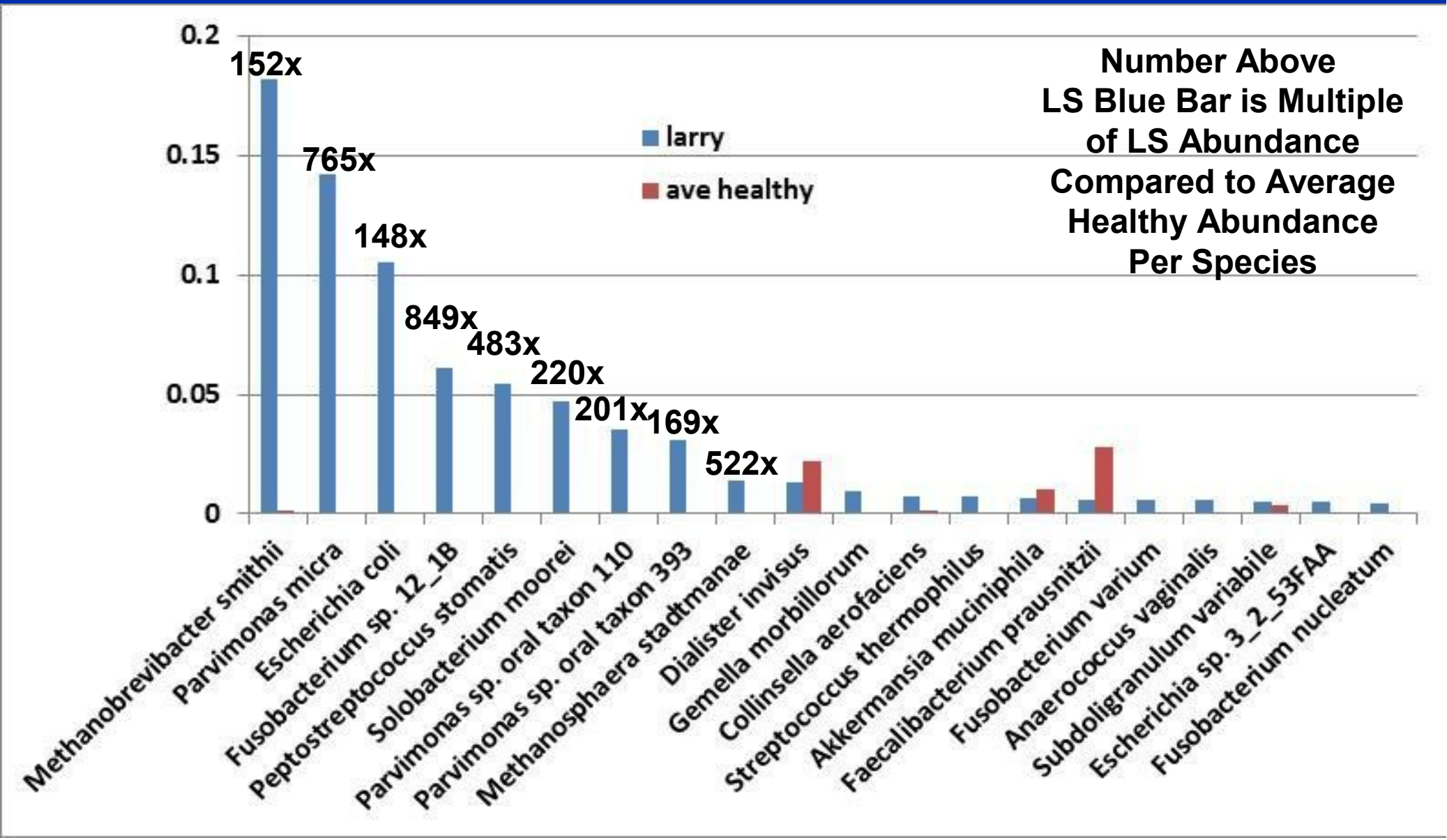
Almost All Abundant Species (≥1%) in Healthy Subjects Are *Severely Depleted* in LS Gut



UCSD
UCIrvine

it²

Top 20 Most Abundant Microbial Species In LS vs. Average Healthy Subject

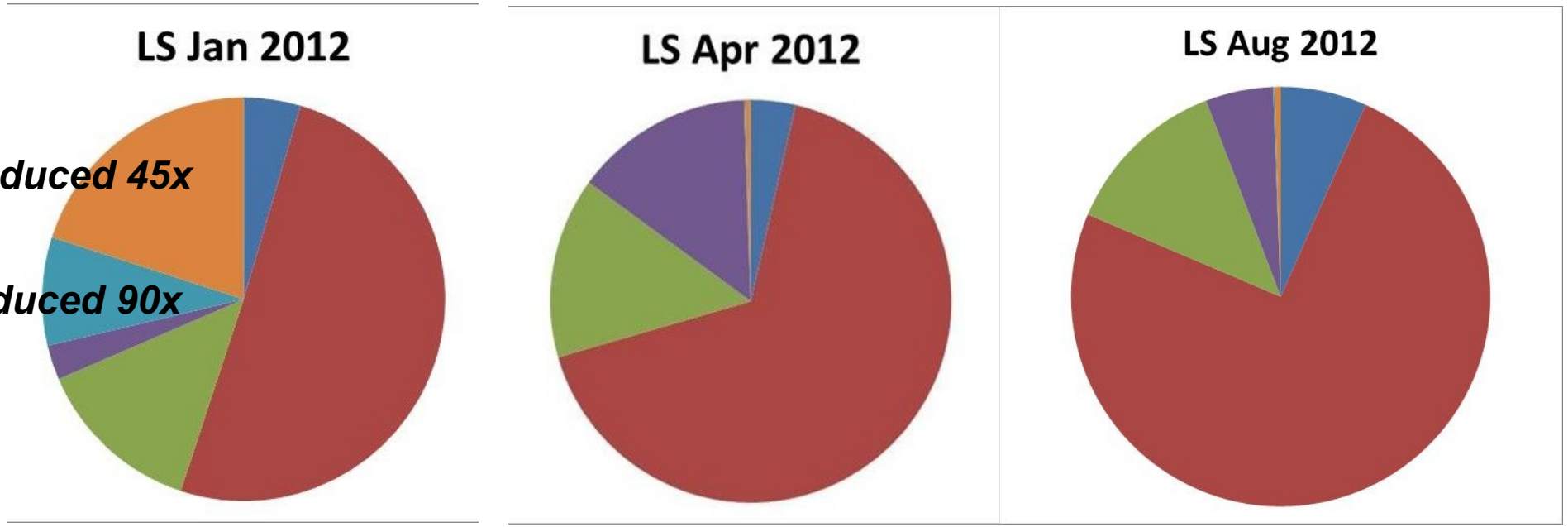


UCSD
UCIrvine

Source: Sequencing JCVI; Analysis Weizhong Li, UCSD
LS December 28, 2011 Stool Sample

it²

Major Changes in LS Microbiome Before and After 1 Month Antibiotic & 2 Month Prednisone Therapy



Therapy Greatly Reduced Two Phyla, But Massive Reduction in Bacteroidetes And Large % Proteobacteria Remain

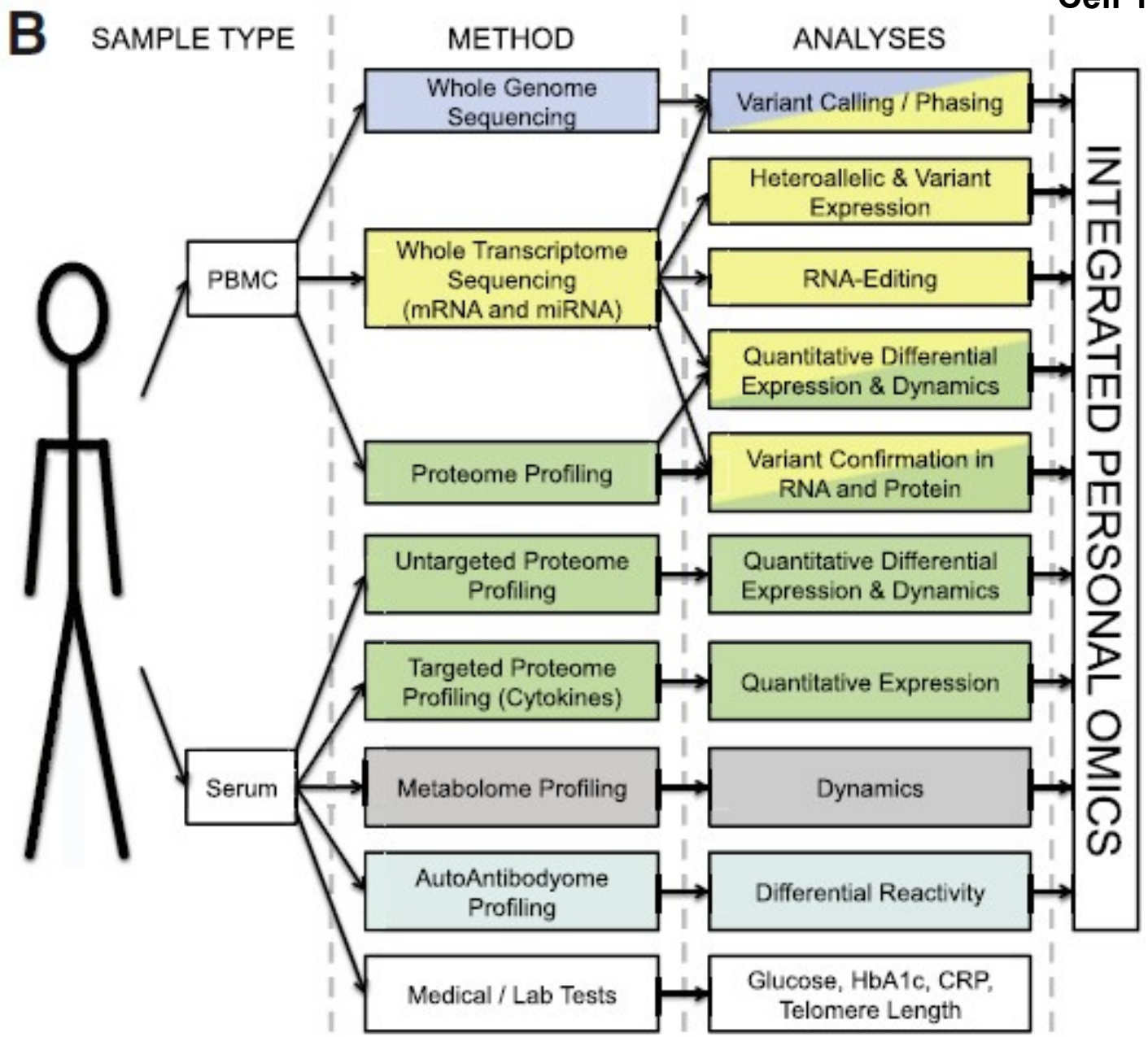
Small Changes With No Therapy

- Bacteroidetes
- Firmicutes
- Proteobacteria
- Actinobacteria
- Fusobacteria
- Euryarchaeota

How Does One Get Back to a "Healthy" Gut Microbiome?

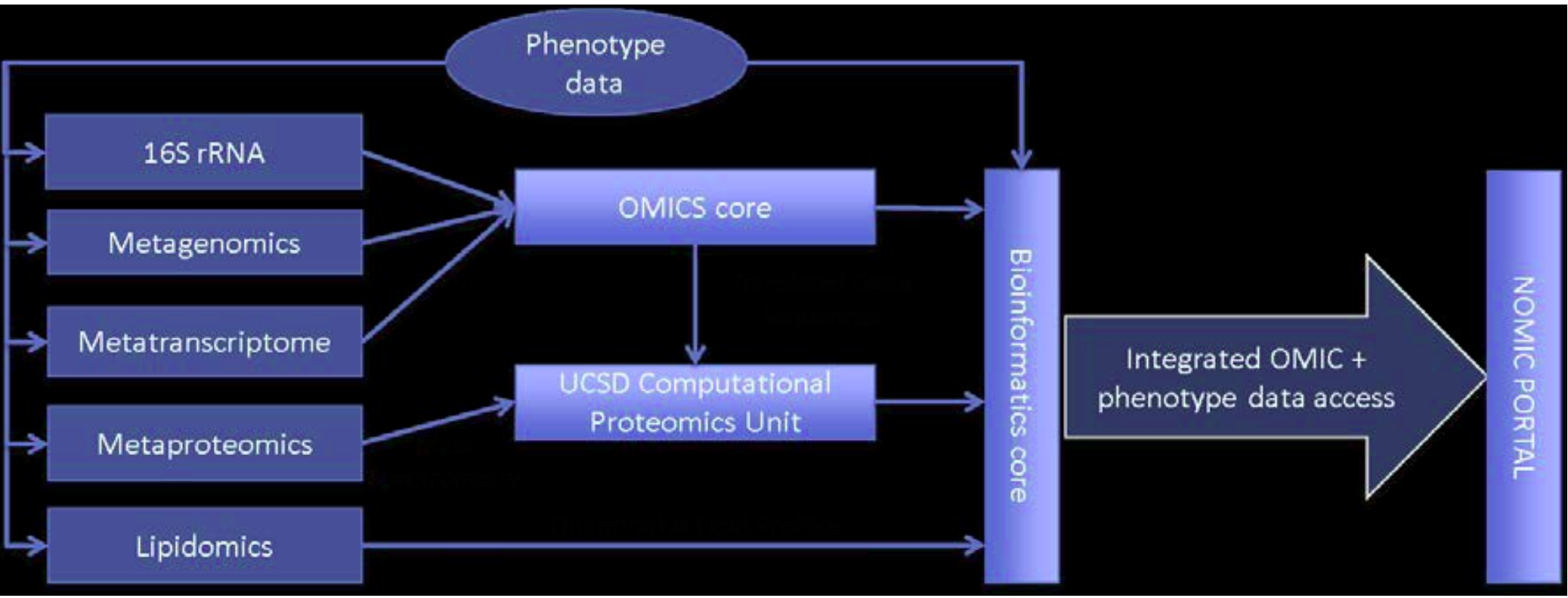
Integrative Personal Omics Profiling Using 100x My Quantifying Biomarkers

Cell 148, 1293–1307, March 16, 2012



- **Michael Snyder, Chair of Genomics Stanford Univ.**
- **Genome 140x Coverage**
- **Blood Tests 20 Times in 14 Months**
  - **tracked nearly 20,000 distinct transcripts coding for 12,000 genes**
  - **measured the relative levels of more than 6,000 proteins and 1,000 metabolites in Snyder's blood**

Proposed UCSD/JCVI Integrated Omics Pipeline



J. Craig Venter INSTITUTE

UCSD UCIrvine

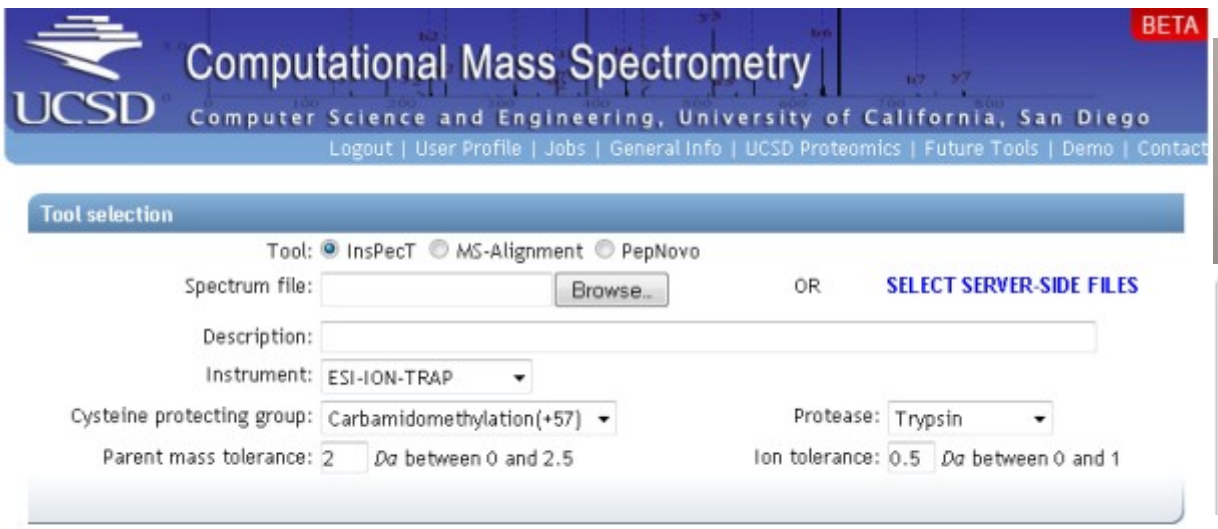
UC San Diego HEALTH SCIENCES

it²

Source: Nuno Bandiera, UCSD

UCSD Center for Computational Mass Spectrometry Becoming Global MS Repository

***ProteoSAFe*: Compute-intensive discovery MS at the click of a button**



***MassIVE*: repository and identification platform for all MS data in the world**

Molecular & Cellular Proteomics

Journal of proteome research

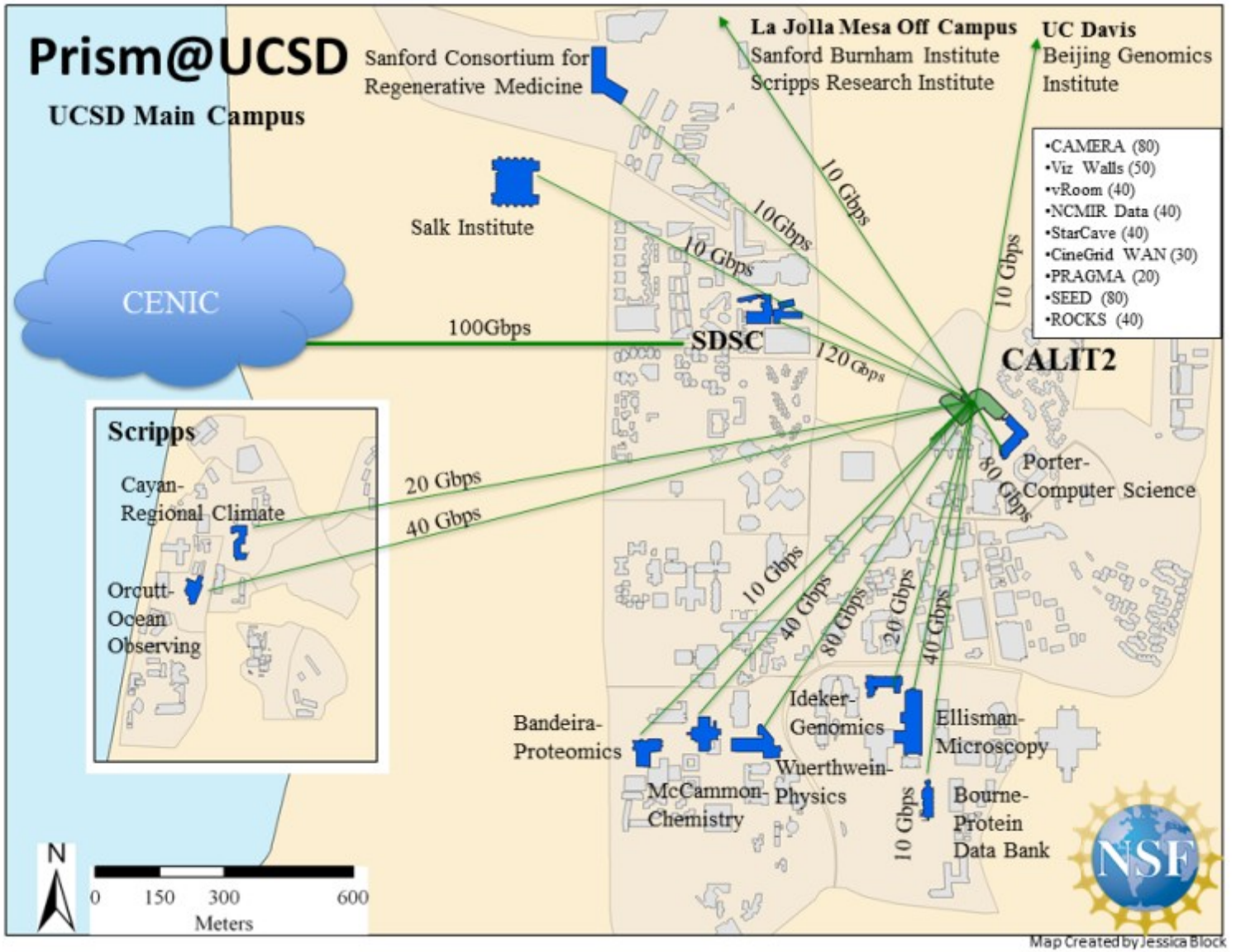
Statistics	
Members	1544
Groups & Projects	355
Data Size (uncompressed)	23.3 TB
Data Files	13,264,659



Source:
Nuno Bandeira,
Vineet Bafna,
Pavel Pevzner,
Ingolf Krueger,
UCSD
proteomics.ucsd.edu

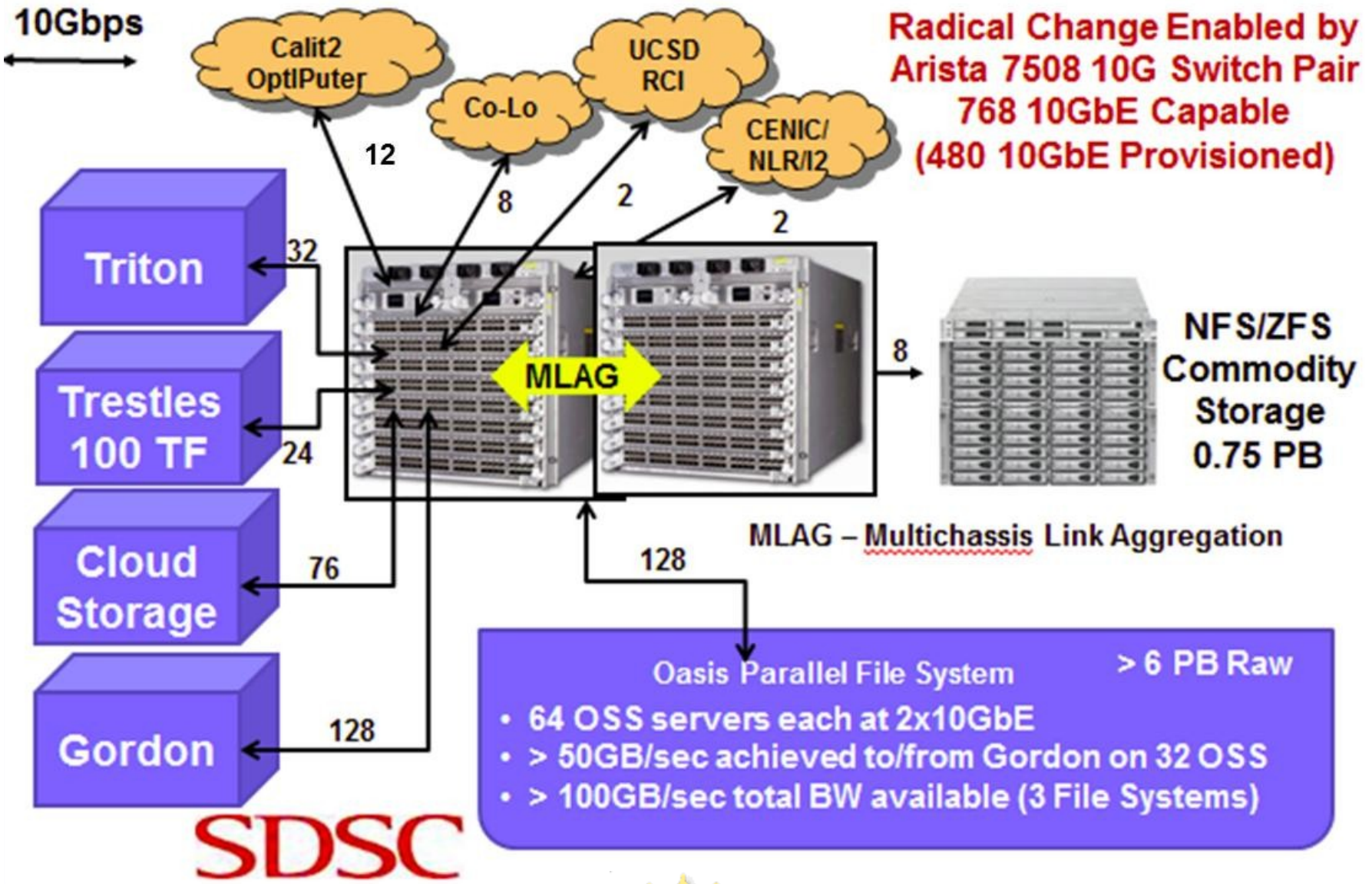
National Institutes of Health
National Institute of General Medical Sciences
BASIC DISCOVERIES FOR BETTER HEALTH

A "Big Data Freeway System" Connecting Users to Remote Campus Clusters & Scientific Instruments



Phil Papadopoulos, SDSC, Calit2, PI

Arista Enables SDSC's Massively Parallel 10G Switched Data Analysis Resource



UCSD
UCIrvine

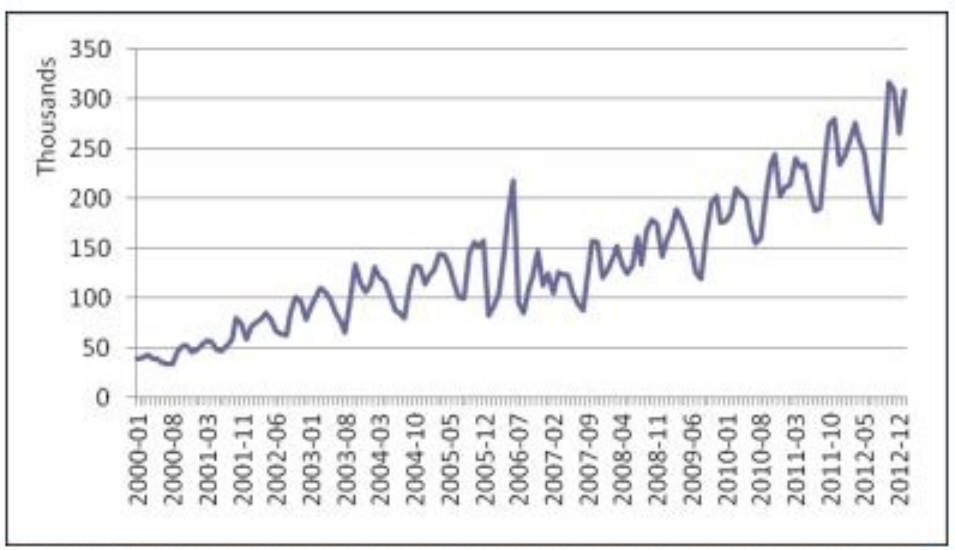
NSF

it²

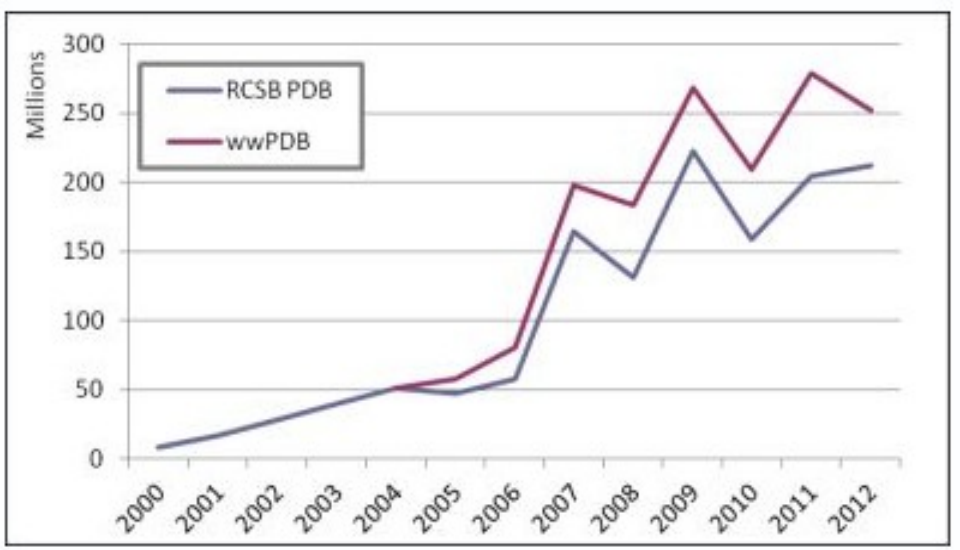
The Protein Data Bank (PDB) Usage Is Growing Over Time

- More than 300,000 Unique Visitors per Month
- Up to 300 Concurrent Users
- ~10 Structures are Downloaded per Second 7/24/365
- Increasingly Popular Web Services Traffic

RCSB PDB Website Unique Users by Month



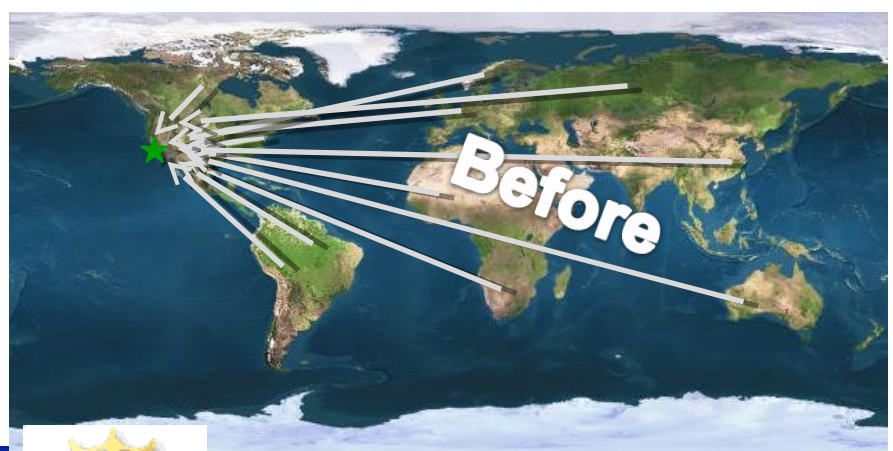
FTP Downloads by Year



Source: Phil Bourne and Andreas Prlić, PDB

PDB Plans to Establish Global Load Balancing

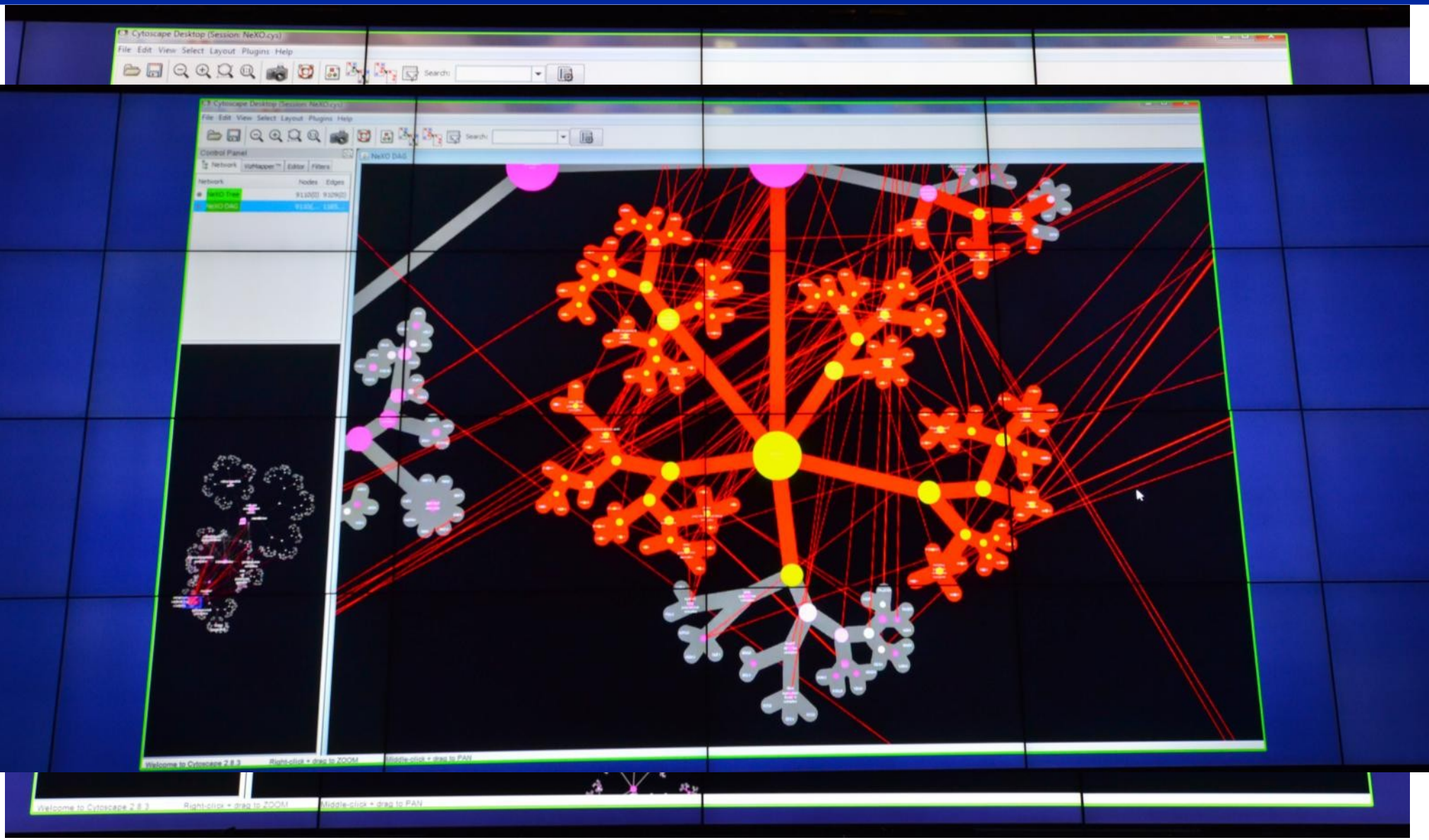
- **Why is it Important?**
  - Enables PDB to Better Serve Its Users by Providing Increased Reliability and Quicker Results
- **How Will it be Done?**
  - By More Evenly Allocating PDB Resources at Rutgers and UCSD
  - By Directing Users to the Closest Site
- **Need High Bandwidth Between Rutgers & UCSD Facilities**





Source: Phil Bourne and Andreas Prlić, PDB

Integrating Systems Biology Data: Cytoscape On Vroom-64MPixels Connected at 50Gbps



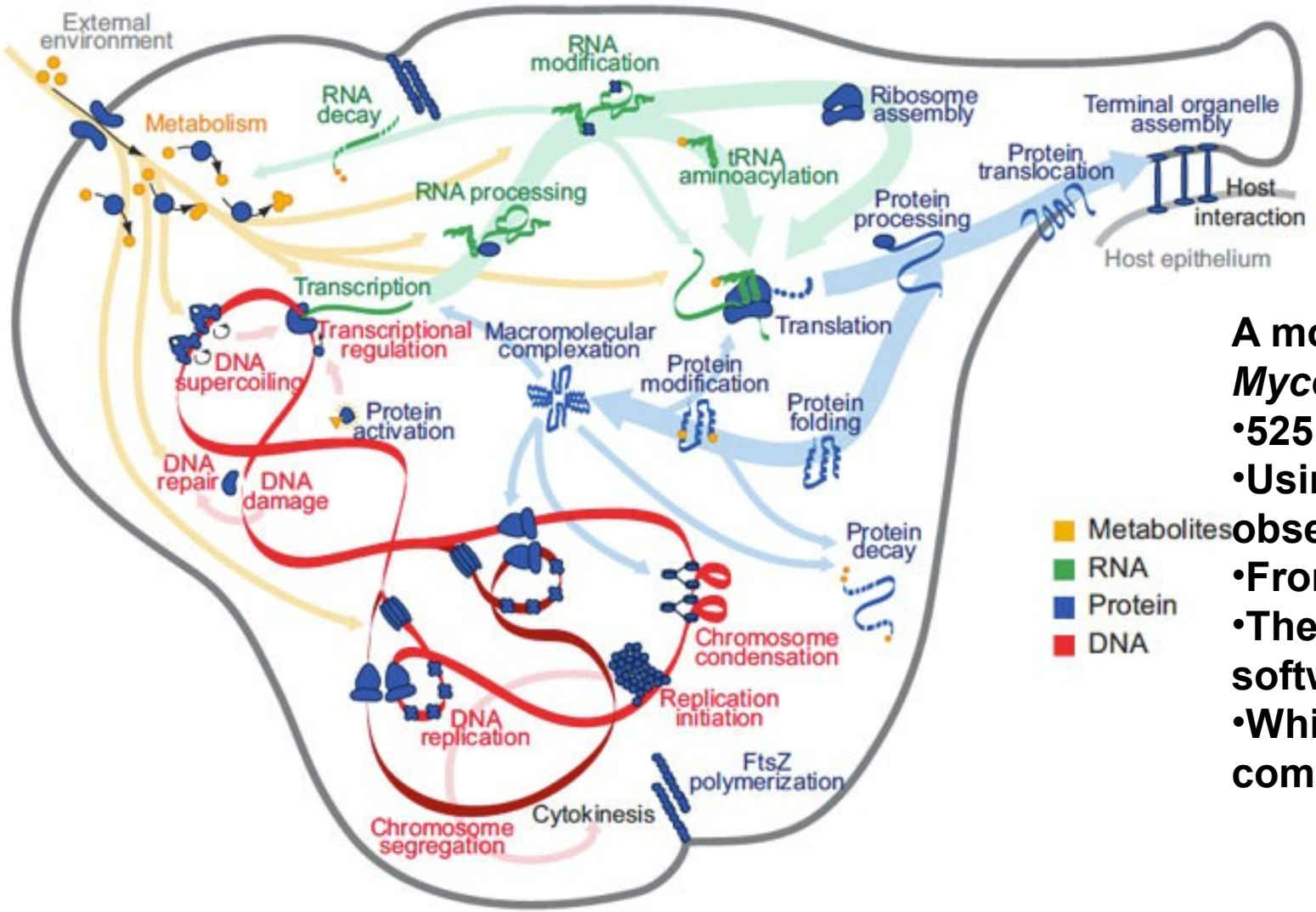
UCSD
UCIrvine

Calit2 Collaboration with Trey Idekar Group

www.cytoscape.org

it2

"A Whole-Cell Computational Model Predicts Phenotype from Genotype"



Cell

Cell *150*, 389–401, July 20, 2012

A model of *Mycoplasma genitalium,*
- **525 genes**
- **Using 1,900 experimental observations**
- **From 900 studies,**
- **They created the software model,**
- **Which requires 128 computers to run**

Jonathan R. Karr,[1,4] Jayodita C. Sanghvi,[2,4] Derek N. Macklin,[2] Miriam V. Gutschow,[2] Jared M. Jacobs,[2] Benjamin Bolival, Jr.,[2] Nacyra Assad-Garcia,[3] John I. Glass,[3] and Markus W. Covert[2,*]

[1]Graduate Program in Biophysics
[2]Department of Bioengineering
Stanford University, Stanford, CA 94305, USA
[3]J. Craig Venter Institute, Rockville, MD 20850, USA

The Dawn of Virtual Cell Biology

Early Attempts at Modeling the Systems Biology of the Gut Microbiome and the Human Immune System

Trends in Biotechnology June 2011, Vol. 29, No. 6

Prospects for systems biology and modeling of the gut microbiome

Fredrik H. Karlsson, Intawat Nookaew, Dina Petranovic and Jens Nielsen

Department of Chemical and Biological Engineering, Chalmers University of Technology, Kemivägen 10, SE-412 96 Göteborg, Sweden

2012 IEEE 26th International Parallel and Distributed Processing Symposium

High-Performance Interaction-Based Simulation of Gut Immunopathologies with ENteric Immunity SImulator (ENISI)

Keith Bisset*, Md. Maksudul Alam*, Josep Bassaganya-Riera*, Adria Carbo*, Stephen Eubank*, Raquel Hontecillas*, Stefan Hoops*, Yongguo Mei*, Katherine Wendelsdorf*, Dawen Xie*, Jae-Seung Yeom*†, and Madhav Marathe*†

**Virginia Bioinformatics Institute*

†Department of Computer Science

Virginia Tech, Blacksburg VA 24061

UCSD

UCIrvine

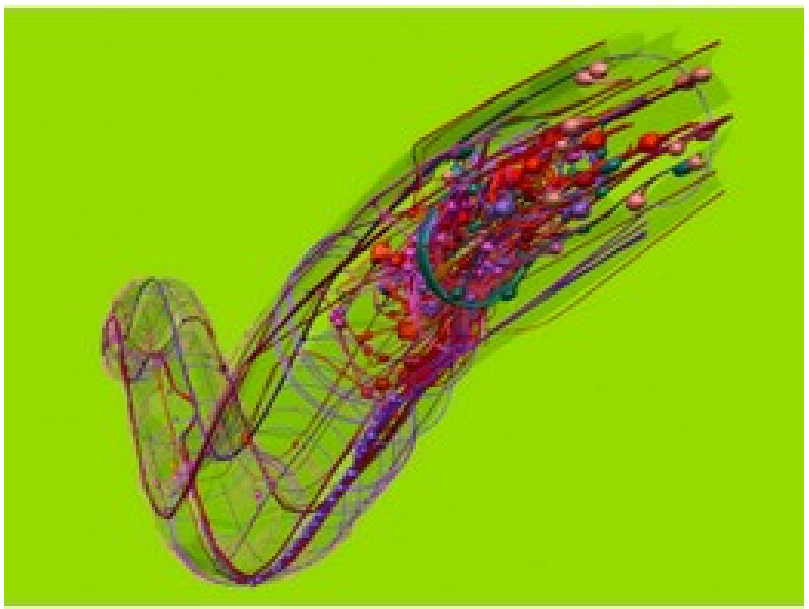
it²

Next Challenge:
Building a Multi-Cellular Organism Simulation

OpenWorm is an attempt to build a complete cellular-level simulation of the nematode worm Caenorhabditis elegans. Of the 959 cells in the hermaphrodite, 302 are neurons and 95 are muscle cells.

The simulation will model electrical activity in all the muscles and neurons. An integrated soft-body physics simulation will also model body movement and physical forces within the worm and from its environment.

- Latest news / current status
- Biology of C. elegans
- The nervous system
- Simulation approach
- Software stack
- Computer hardware
- 3D WebGL model



A 3D rendering of the C. elegans nervous system.

UCSD
UCIrvine

www.artificialbrains.com/openworm

it²

A Vision for Healthcare in the Coming Decades

Using this data, the planetary computer will be able to build a computational model of your body and compare your sensor stream with millions of others. Besides providing early detection of internal changes that could lead to disease, cloud-powered voice-recognition wellness coaches could provide continual personalized support on lifestyle choices, potentially staving off disease and making health care affordable for everyone.

ESSAY
An Evolution Toward a Programmable Universe
By LARRY SMARR
Published: December 5, 2011

UCSD
UCIrvine

The New York Times

Science

it²