“Using Supercomputers to Discover the 100 Trillion Bacteria Living Within Each of Us”

Invited Talk

Supercomputing 2014

New Orleans, LA

November 19, 2014

Dr. Larry Smarr

Director, California Institute for Telecommunications and Information Technology

Harry E. Gruber Professor,

Dept. of Computer Science and Engineering

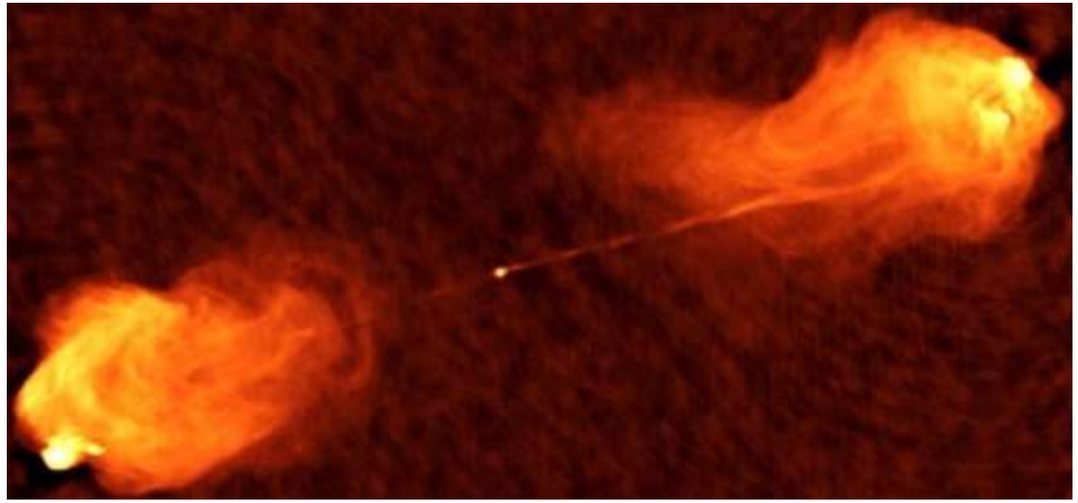
Jacobs School of Engineering, UCSD

http://lsmarr.calit2.net

Abstract

The human body is host to 100 trillion microorganisms, ten times the number of cells in the human body and these microbes contain 100 times the number of DNA genes that our human DNA does. The microbial component of our “superorganism” is comprised of hundreds of species with immense biodiversity. Thanks to the National Institutes of Health’s Human Microbiome Program researchers have been discovering the states of the human microbiome in health and disease. To put a more personal face on the “patient of the future,” I have been collecting massive amounts of data from my own body over the last five years, which reveals detailed examples of the episodic evolution of this coupled immune-microbial system. An elaborate software pipeline, running on high performance computers, reveals the details of the microbial ecology and its genetic components. We can look forward to revolutionary changes in medical practice over the next decade.

Past Four Decades-From Observational Taxonomy to Nonlinear Dynamics of Astrophysical Systems



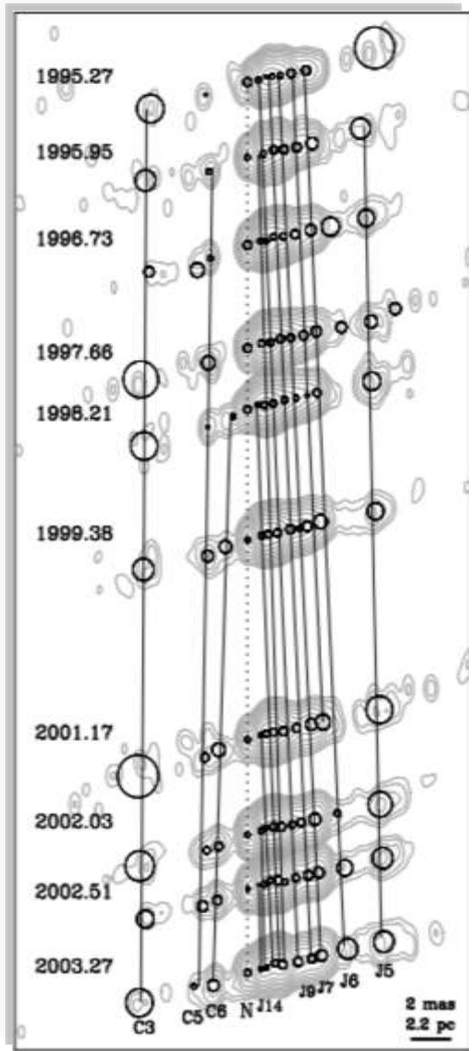
Radio Observation of Cygnus A (NRAO)



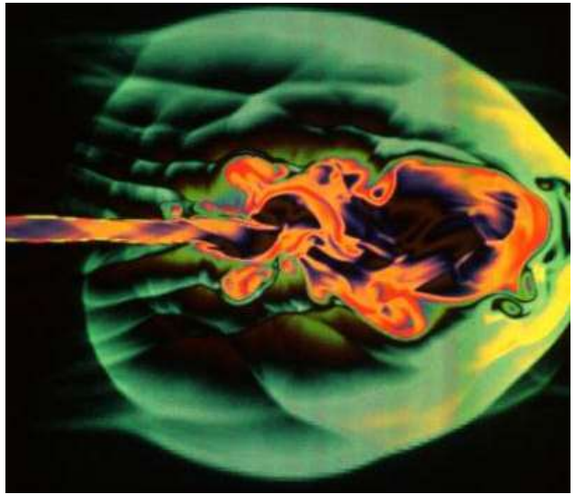
Gas Accreting
Onto a Black Hole

Hawley and Smarr 1985

Key Roles of Time Series and Imaging



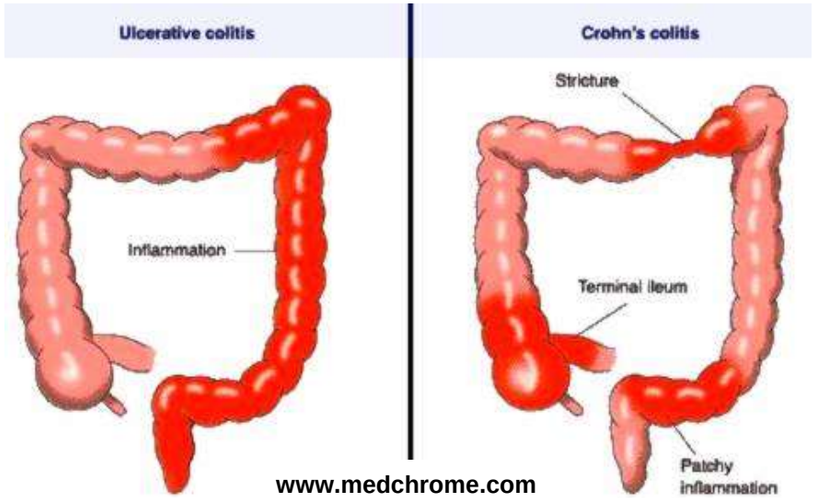
Time Series Observation
Of Inner Jet
Bach, et al. 2003



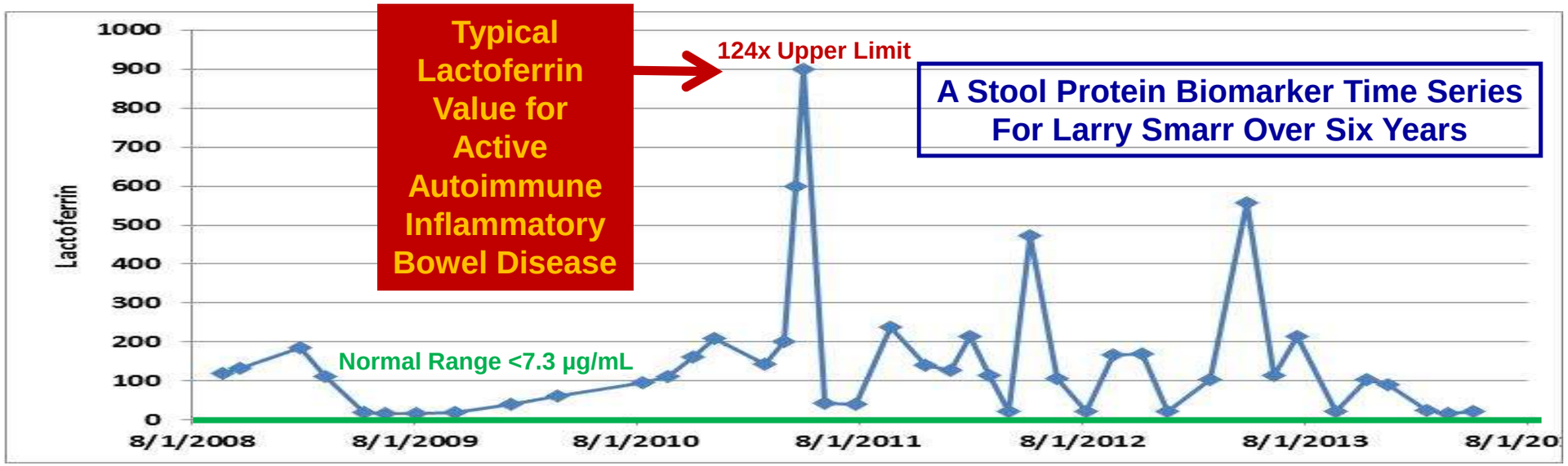
Hydrodynamics Simulation
of a Supersonic Gas Jet
Norman, Cox (1988)

Next Few Decades-From Observational Taxonomy To Nonlinear Dynamics of Biological Systems

Inflammatory Bowel Disease is one of 80 Autoimmune Diseases Largely Defined By Symptoms and Episodic Flares

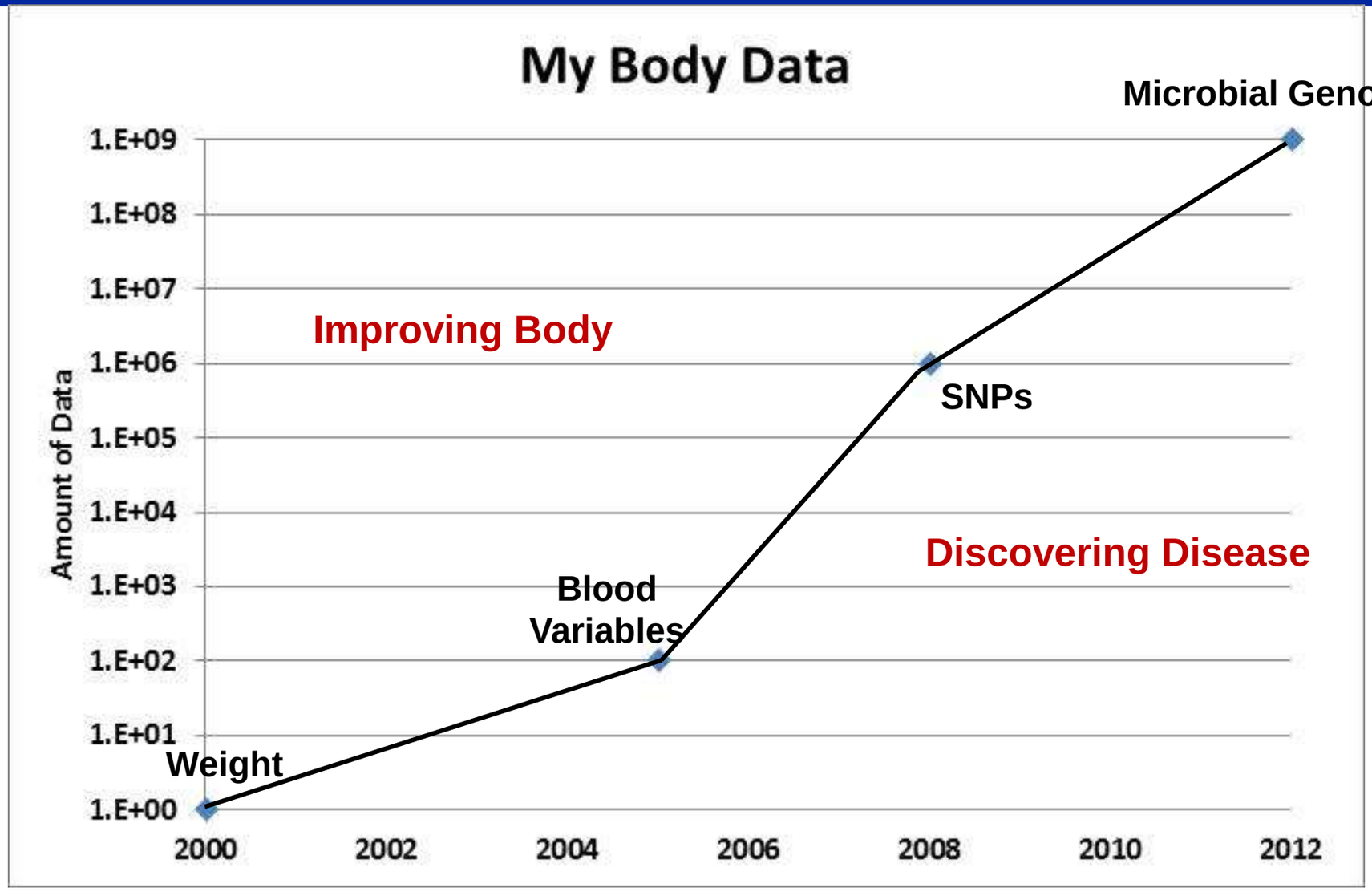


Can Time Series, Genetic Sequencing, and Dynamical Systems Lead to Root Cause?

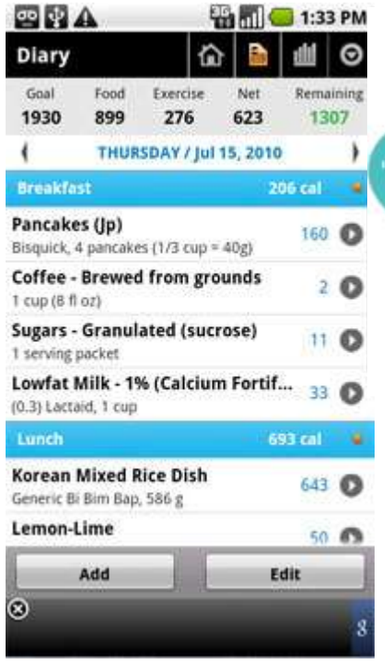


Lactoferrin is a Protein Shed from Neutrophils - An Immune System Antibacterial that Sequesters Iron

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



I Used a Variety of Emerging Personal Sensors To Quantify My Body & Drive Behavioral Change



MyFitnessPal-
Calories Ingested



FitBit -
Daily Steps &
Calories Burned



Withings/iPhone-
Blood Pressure





Withings WiFi Scale -
Daily Weight



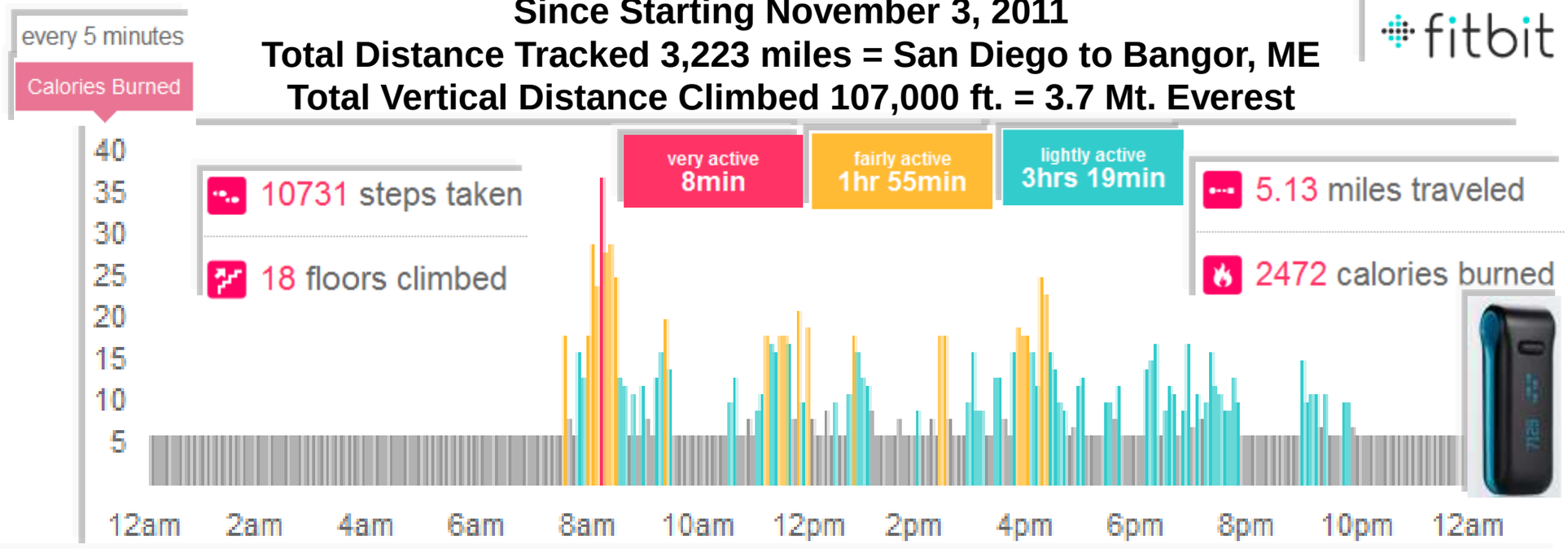
Zeo-Sleep

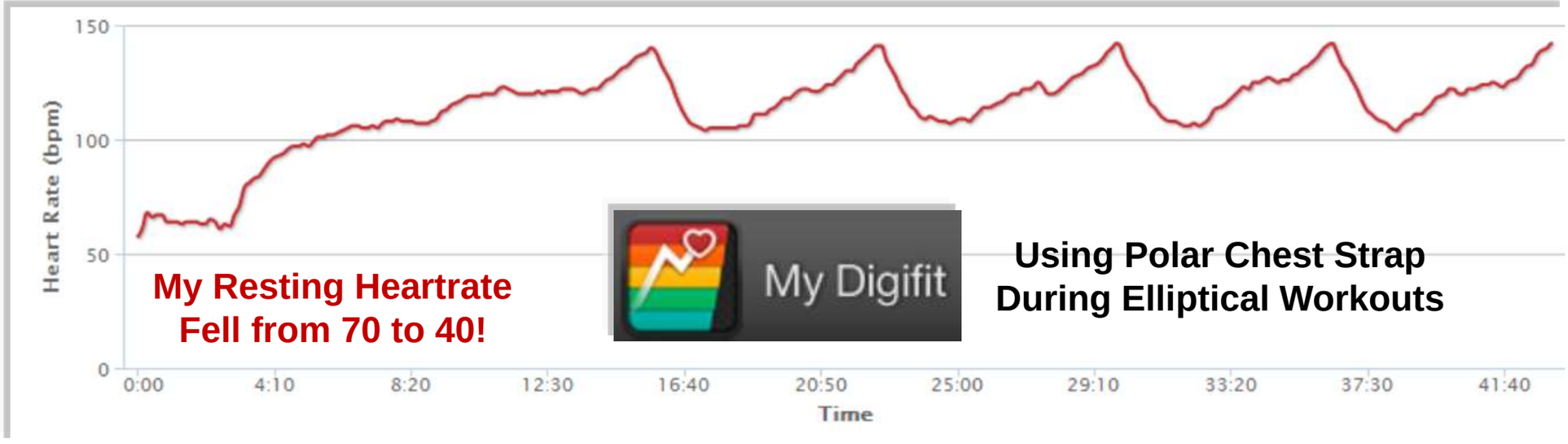


Azumio-Heart Rate

Wireless Monitoring
Produced Time Series That Helped Me Improve My Health

Since Starting November 3, 2011
Total Distance Tracked 3,223 miles = San Diego to Bangor, ME
Total Vertical Distance Climbed 107,000 ft. = 3.7 Mt. Everest





My Resting Heartrate Fell from 70 to 40!

Using Polar Chest Strap During Elliptical Workouts

From Measuring Macro-Variables to Measuring Your Internal Variables

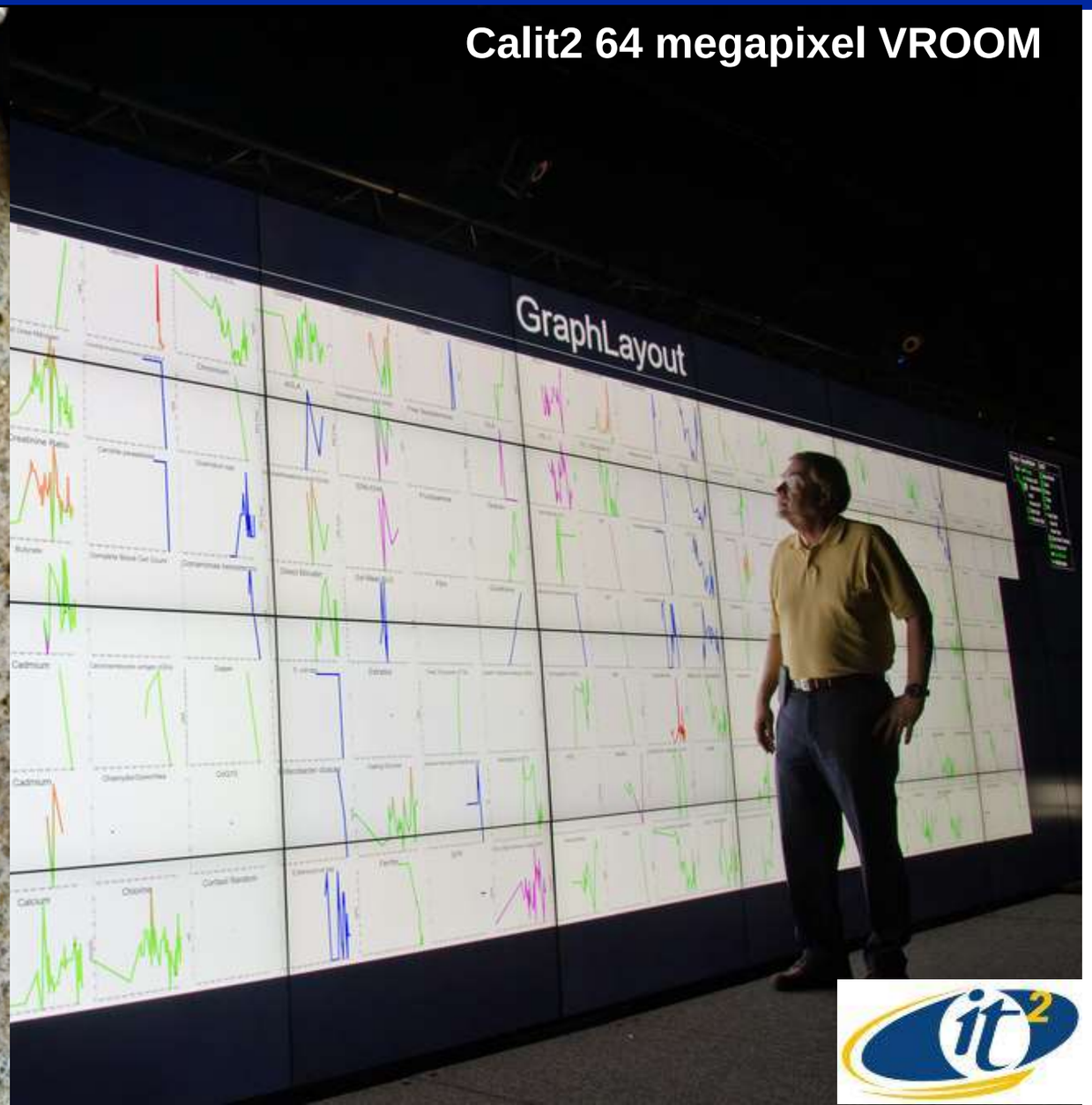


www.technologyreview.com/biomedicine/39636

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

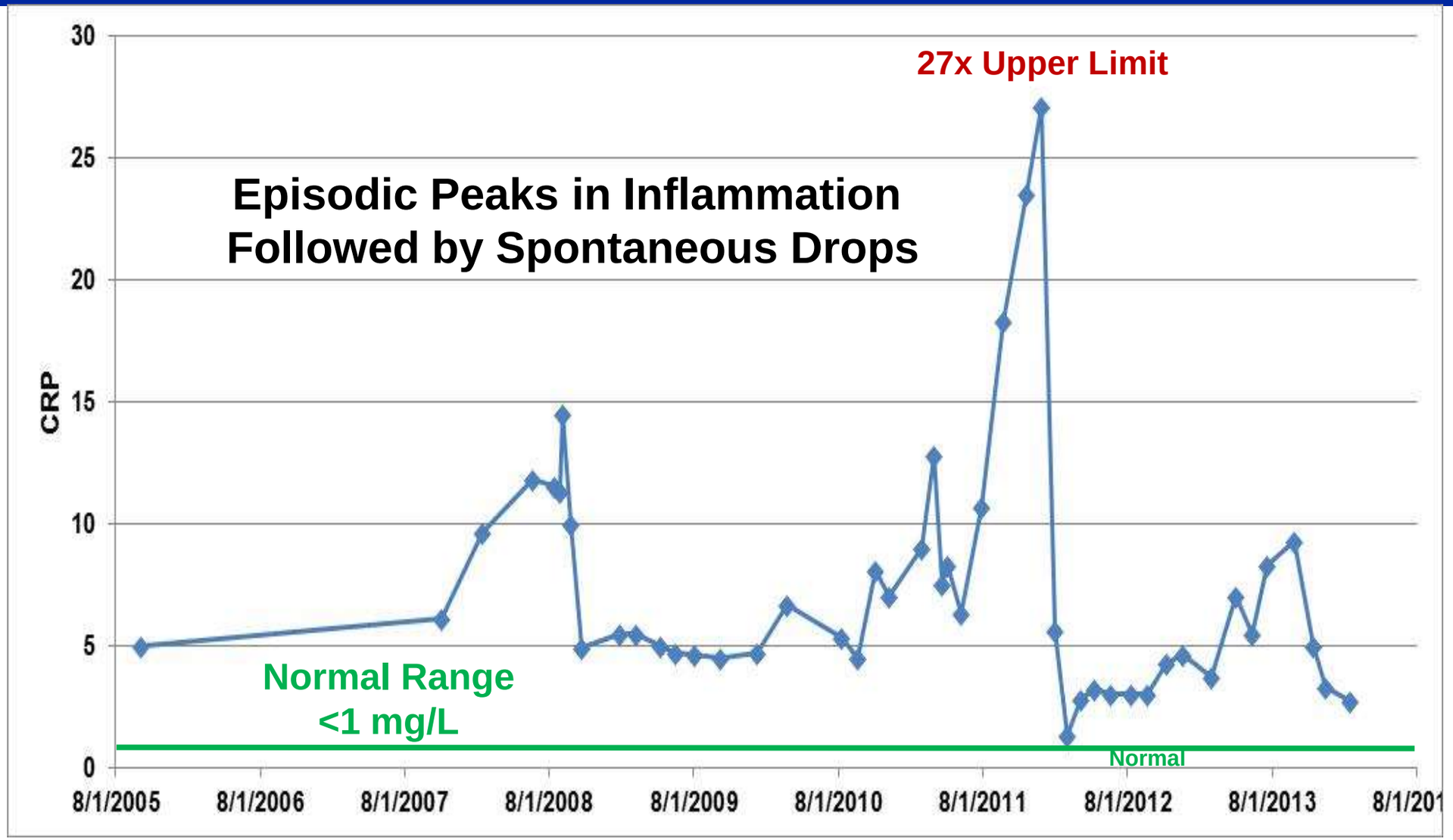


One Blood Draw For Me



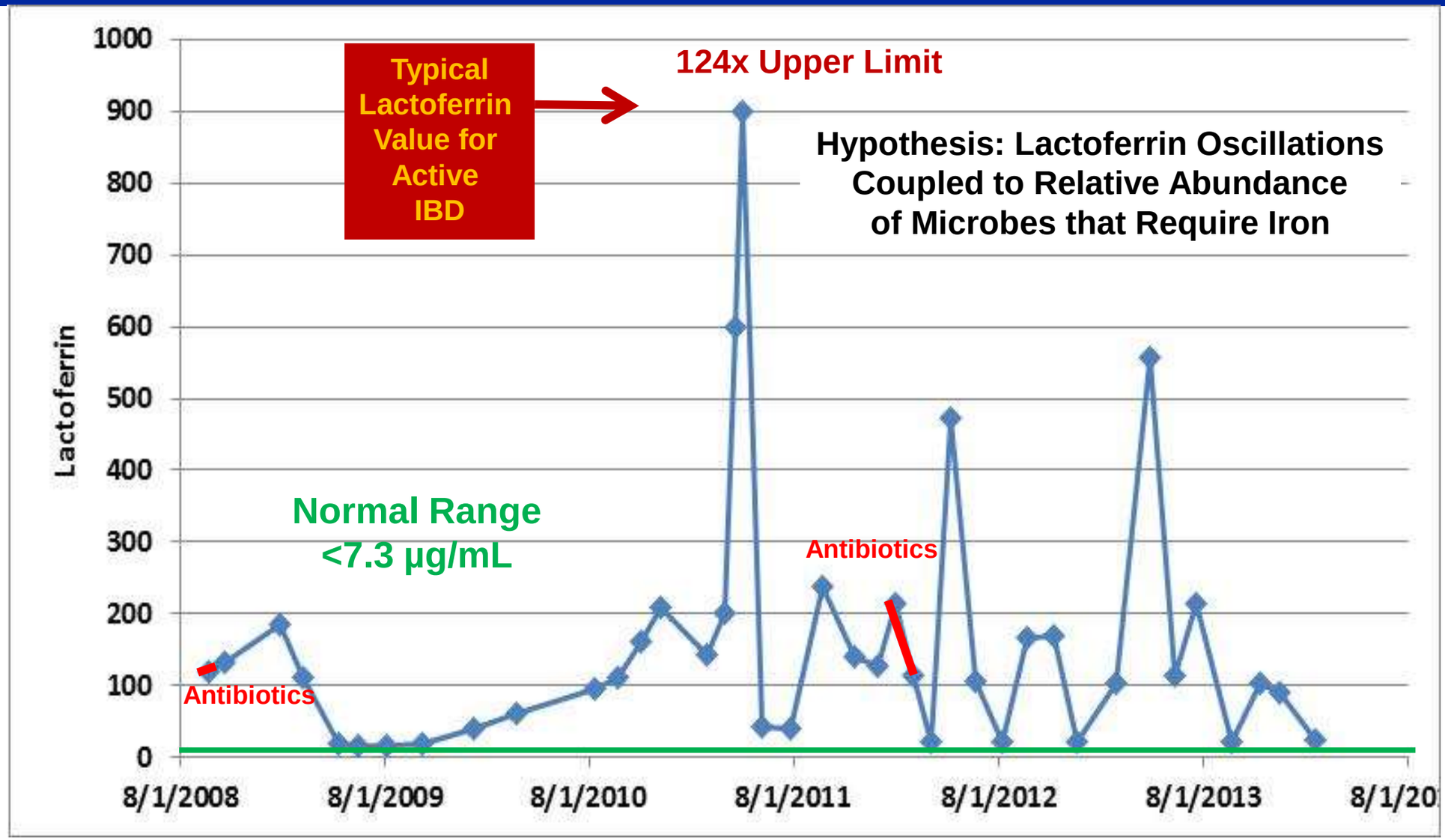
Calit2 64 megapixel VROOM

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



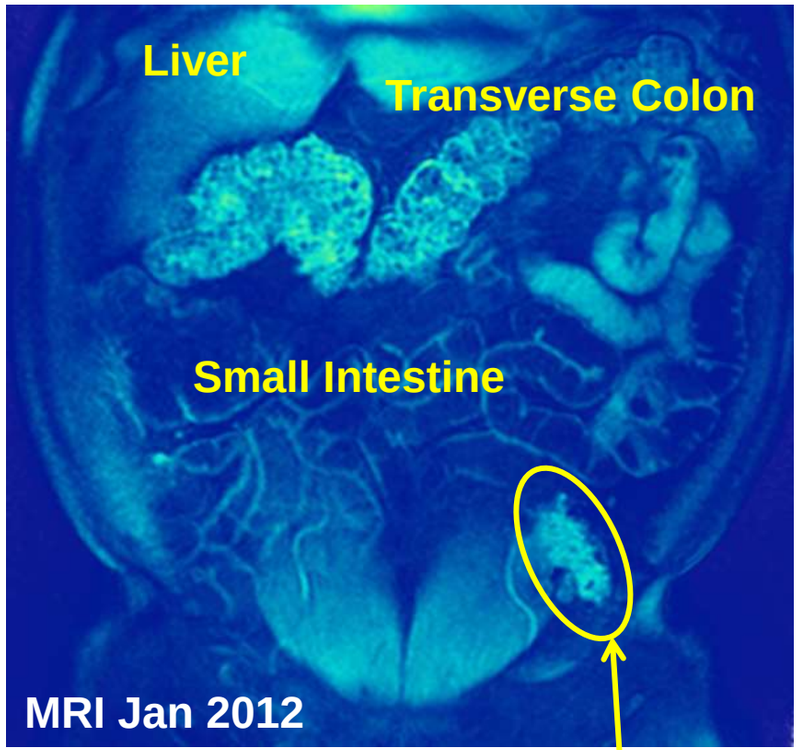
Complex Reactive Protein (CRP) is a Blood Biomarker for Detecting Presence of Inflammation

Adding Stool Tests Revealed Oscillatory Behavior in an Immune Variable



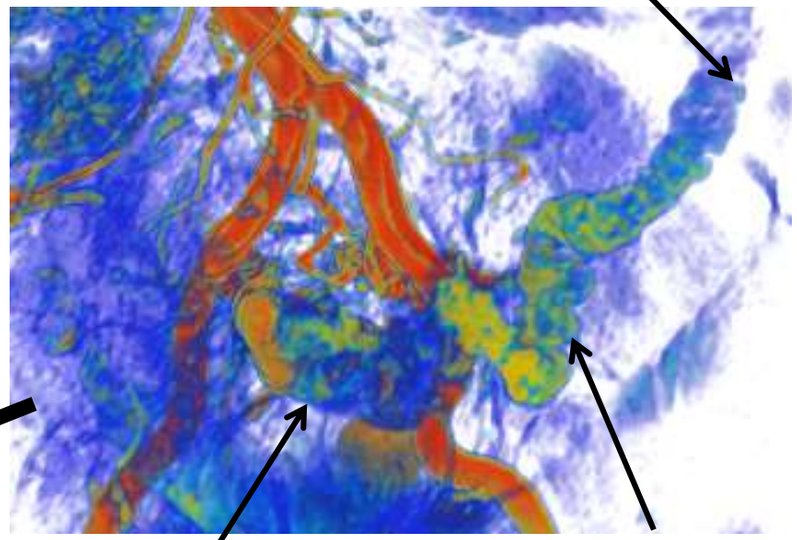
Lactoferrin is a Protein Shed from Neutrophils - An Antibacterial that Sequesters Iron

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



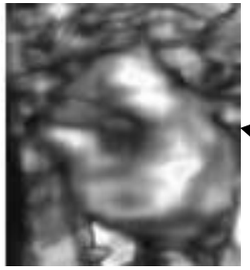
I Obtained the MRI Slices From UCSD Medical Services and Converted to Interactive 3D Working With Calit2 Staff & DeskVOX Software

Descending Colon

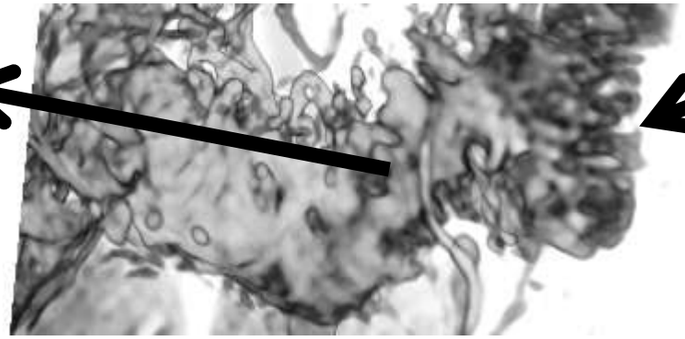


Major Kink

Cross Section



Diseased Sigmoid Colon



Sigmoid Colon
Threading Iliac Arteries

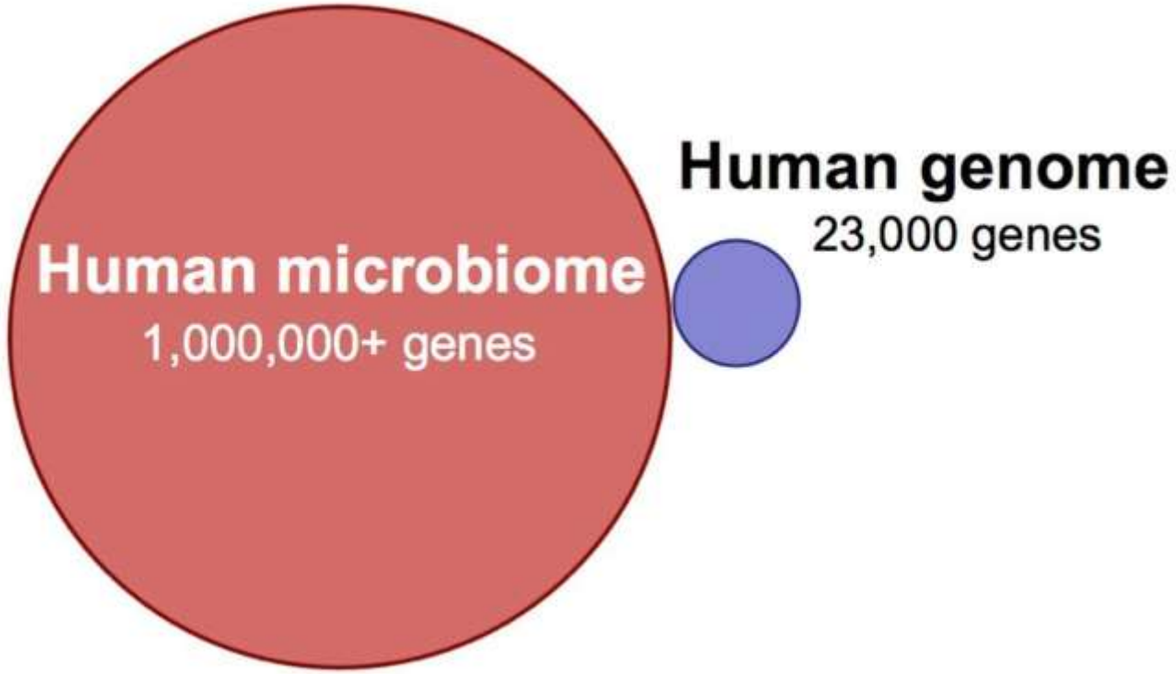
Source: Jurgen Schulze, Calit2

From Observing the 100 Billion Stars in the Andromeda Galaxy in the 1980s



To Observing the 100 Trillion Non-Human Cells in The Human Body in the 2000s

Your Body Has 10 Times As Many Microbe Cells As Human Cells



99% of Your DNA Genes Are in Microbe Cells Not Human Cells

Inclusion of the "Dark Matter" of the Body Will Radically Alter Medicine

J. Craig Venter™ INSTITUTE

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

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



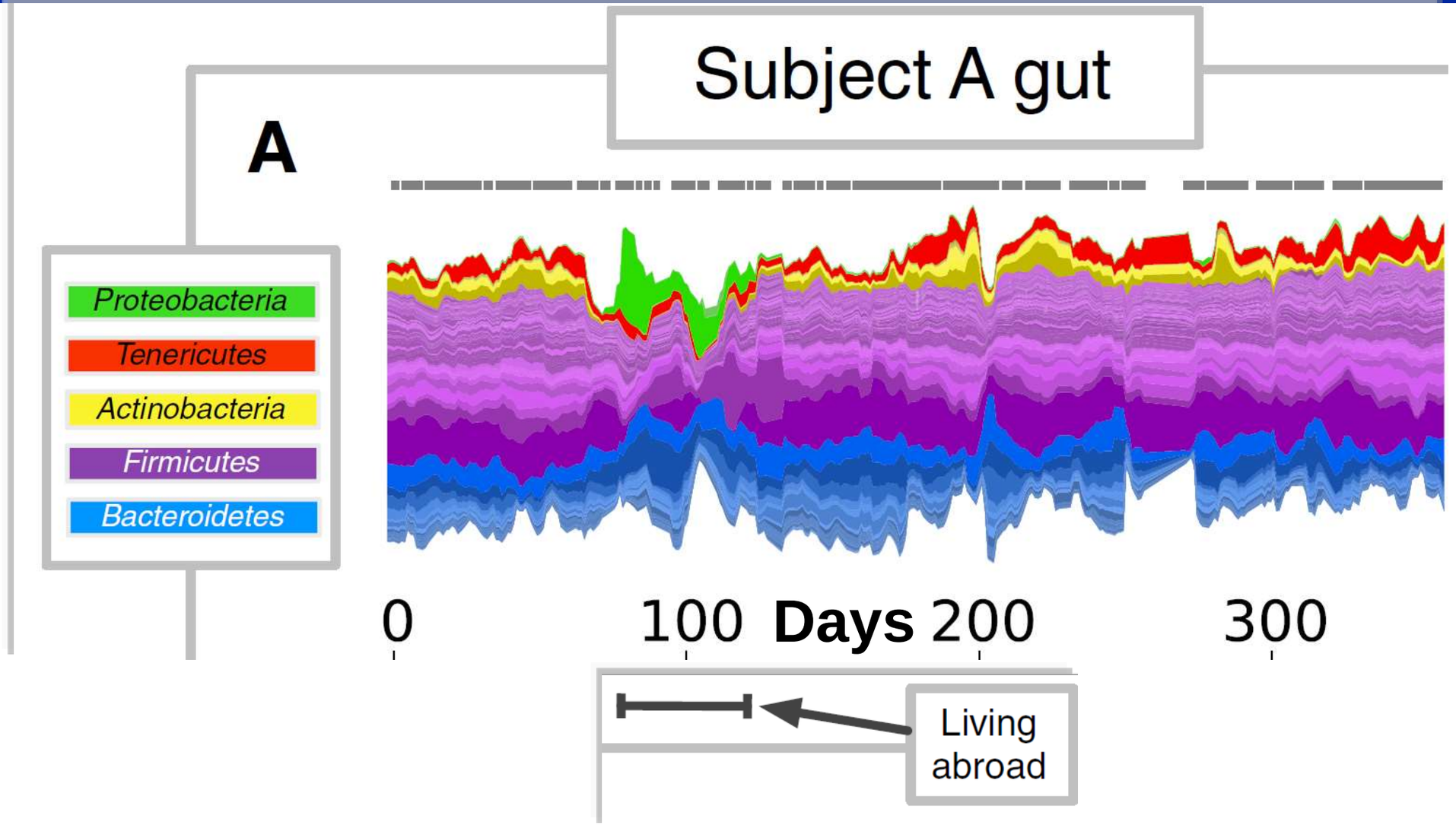
Phylum Mollusca



Phylum Arthropoda

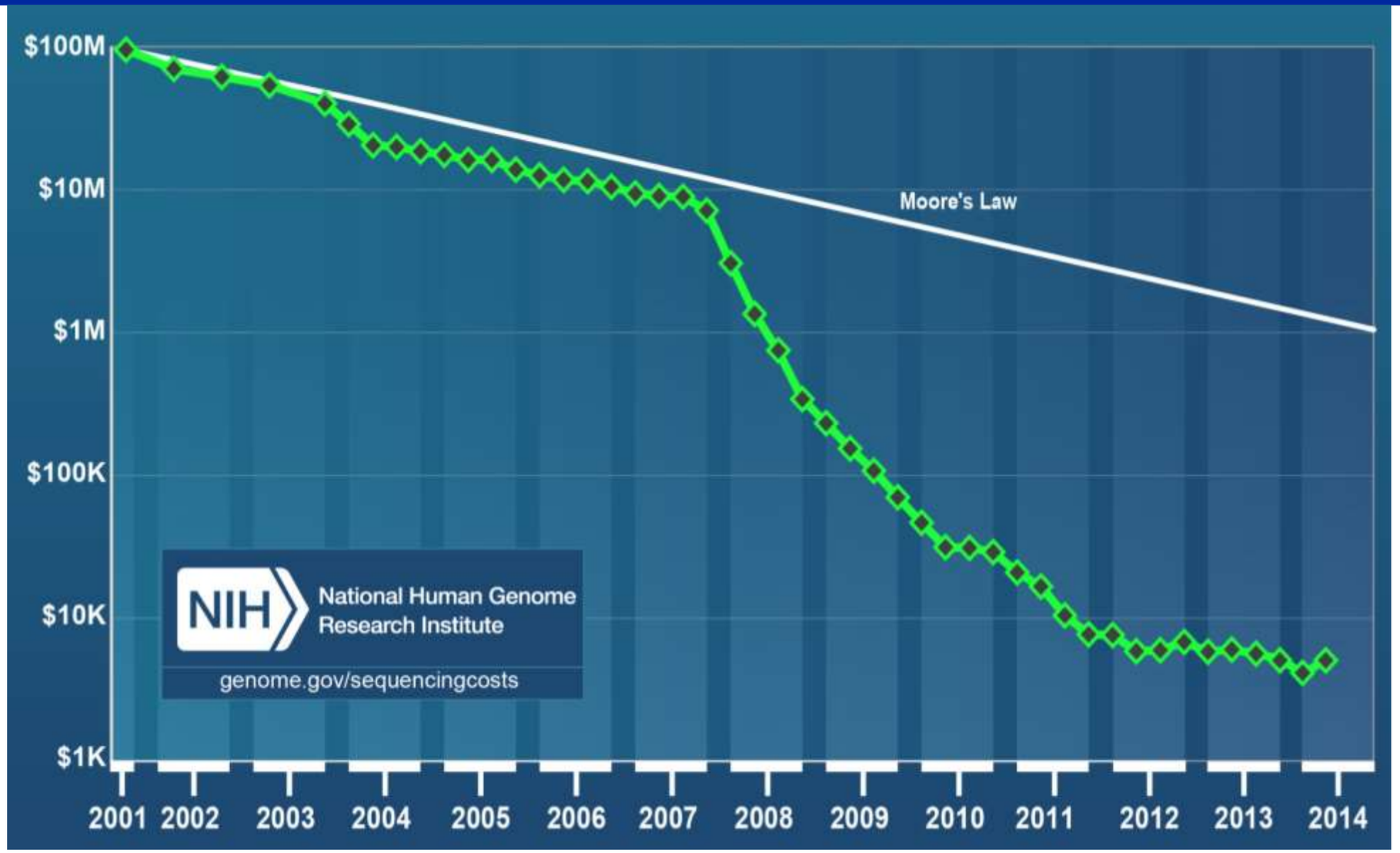
All images from WikiMedia Commons.
Photos are public domain or by Dan Hershman, Michael Linnenbach, Manuae, B_cool

A Year of Sequencing a Healthy Gut Microbiome Daily - Remarkable Stability with Abrupt Changes



Genome Biology (2014)
David, et al.

The Cost of Sequencing a Human Genome Has Fallen Over 10,000x in the Last Ten Years



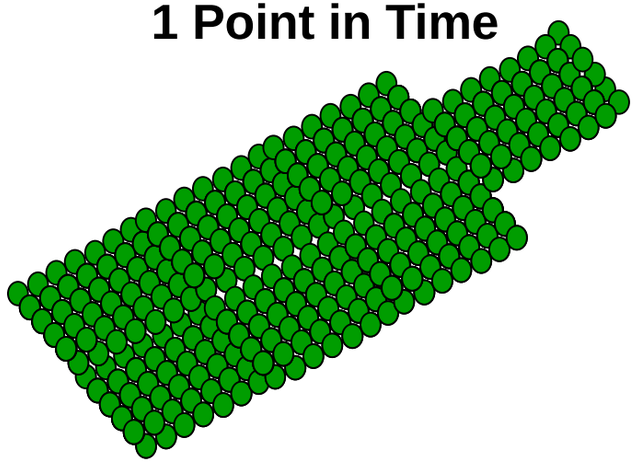
This Has Enabled Sequencing of Both Human and Microbial Genomes

JCVI Sequenced My Gut Microbiome and We Downloaded ~270 More from the NIH Human Microbiome Project For Comparative Analysis

Each Sample Has 100-200 Million Illumina Short Reads (100 bases)

"Healthy" Individuals

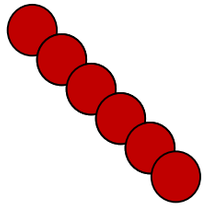
250 Subjects
1 Point in Time



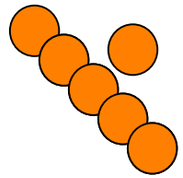
Inflammatory Bowel Disease (IBD) Patients

Larry Smarr
(Colonic Crohn's)

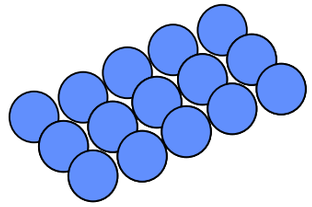
7 Points in Time



2 Ulcerative Colitis Patients,
6 Points in Time



5 Ileal Crohn's Patients,
3 Points in Time



Total of 27 Billion Reads
Or 2.7 Trillion Bases

Source: Jerry Sheehan, Calit2
Weizhong Li, Sitao Wu, CRBS, UCSD

We Created a Reference Database Of Known Gut Genomes

- **NCBI April 2013**
  - **2471 Complete + 5543 Draft Bacteria & Archaea Genomes**
  - **2399 Complete Virus Genomes**
  - **26 Complete Fungi Genomes**
  - **309 HMP Eukaryote Reference Genomes**
- **Total 10,741 genomes, ~30 GB of sequences**



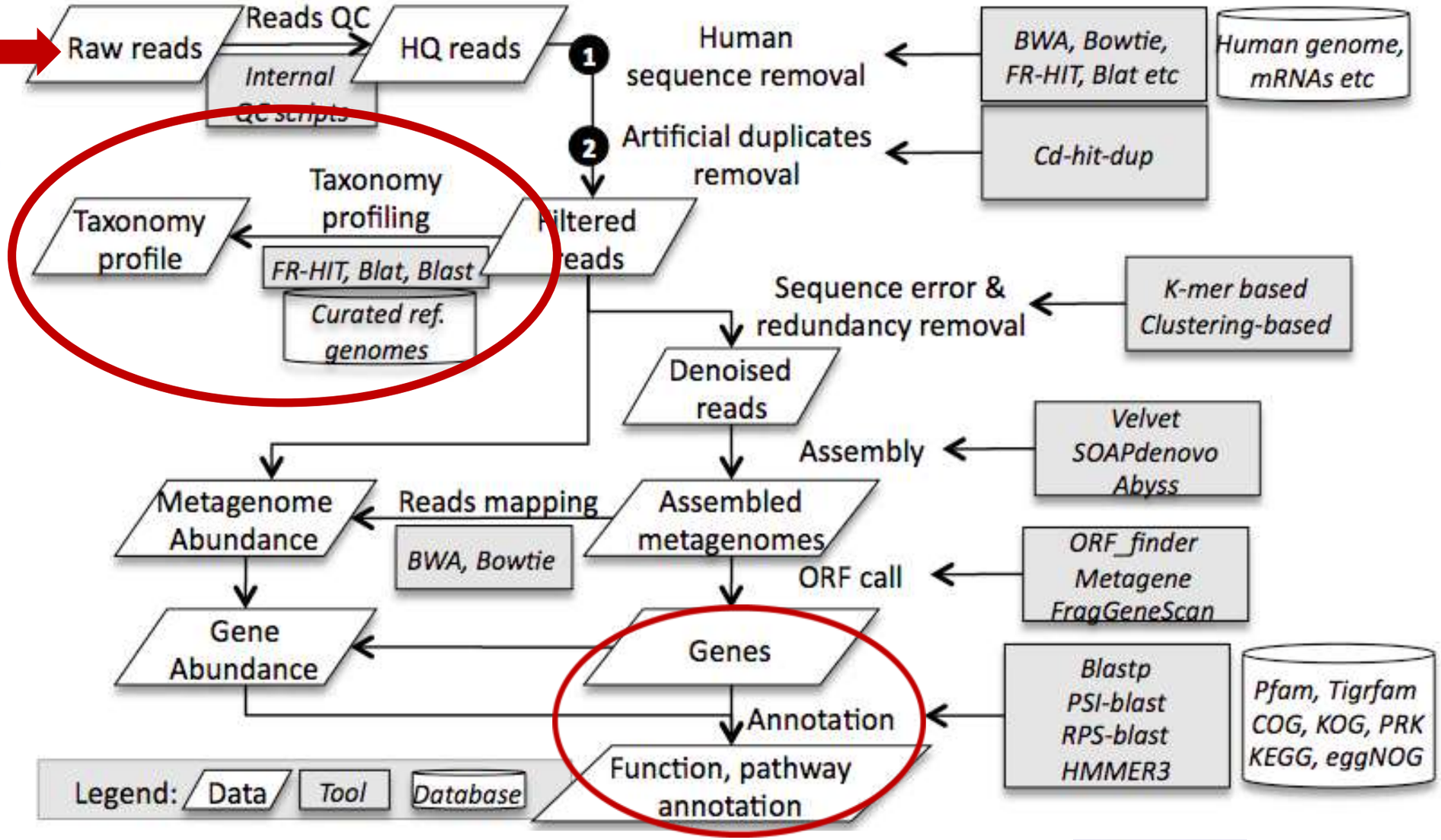
Now to Align Our 27 Billion Reads Against the Reference Database

Source: Weizhong Li, Sitao Wu, CRBS, UCSD

Computational NextGen Sequencing Pipeline: From Sequence to Taxonomy and Function



J. Craig Venter INSTITUTE



National Human Genome Research Institute

HMP

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

CRBS

it²

We Used SDSC's Gordon Data-Intensive Supercomputer to Completely Analyze a Subset of Our Gut Microbiomes

- **~180,000 Core-Hrs on Gordon**
  - KEGG function annotation: 90,000 hrs
  - Mapping: 36,000 hrs
    - Used 16 Cores/Node and up to 50 nodes
  - Duplicates removal: 18,000 hrs
  - Assembly: 18,000 hrs
  - Other: 18,000 hrs
- **Gordon RAM Required**
  - 64GB RAM for Reference DB
  - 192GB RAM for Assembly
- **Gordon Disk Required**
  - Ultra-Fast Disk Holds Ref DB for All Nodes
  - 8TB for All Subjects



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

Source: Weizhong Li, UCSD

We Used Dell's HPC Cloud to Extend Our Taxonomic Analysis to All of Our Human Gut Microbiomes

- **Dell's Sanger Cluster**
  - 32 Nodes, 512 Cores
  - 48GB RAM per Node

Dell Solutions Center
Industry Solutions Lab
SANGER DSU

- **We Processed the Taxonomic Relative Abundance**
  - Used ~35,000 Core-Hours on Dell's Sanger
- **Produced Relative Abundance of ~10,000 Bacteria, Archaea, Viruses in ~300 People**
  - ~3Million Spreadsheet Cells



- **New System: R Bio-Gen System**
  - 48 Nodes, 768 Cores
  - 128 GB RAM per Node

Dell Cloud for HPC with R-Systems

Enabled by a Grant of Time From Dell/R Systems

R SYSTEMS powering innovation

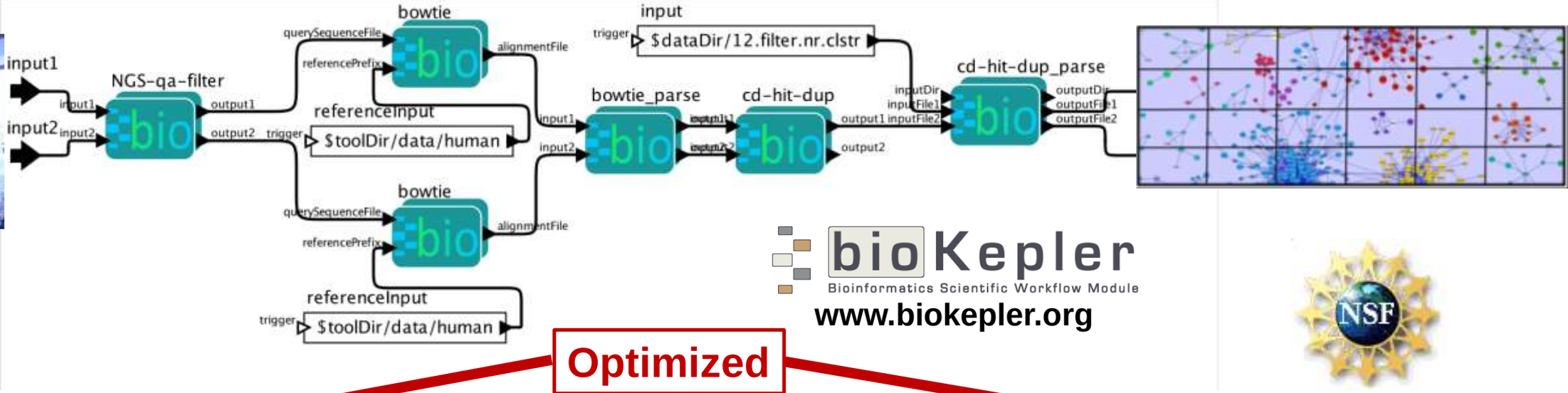
it²

Source: Weizhong Li, UCSD

Next Step

Programmability, Scalability and Reproducibility using bioKepler





bioKepler
Bioinformatics Scientific Workflow Module
www.biokepler.org

NSF

Optimized

Source: Ilkay Altintas, SDSC

Local Cluster Resources



NVIDIA TESLA

hadoop

Cloud Resources

amazon web services

Google Cloud Platform

DELL RSYSTEMS powering innovation

National Resources



(Gordon)



(Comet)

SDSC



(Lonestar)



(Stampede)

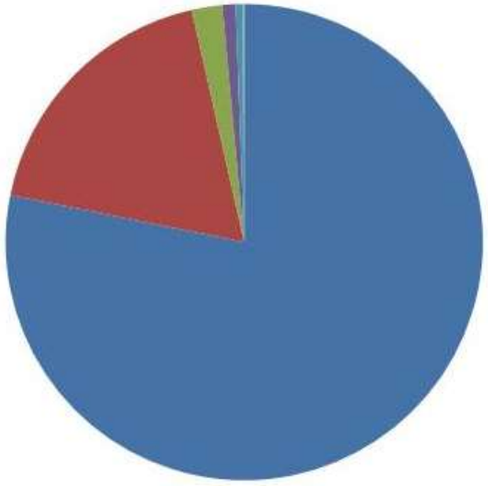
TACC

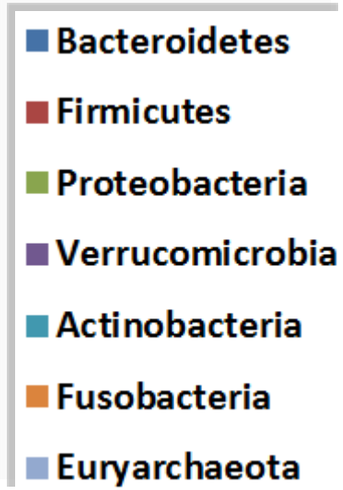
Kepler
www.kepler-project.org

it²

We Found Major State Shifts in Microbial Ecology Phyla Between Healthy and Two Forms of IBD

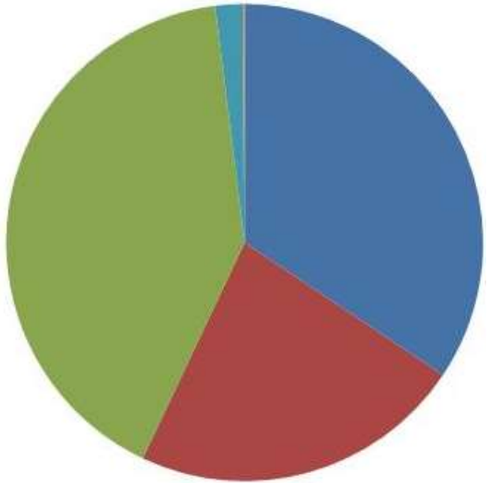
Average HE





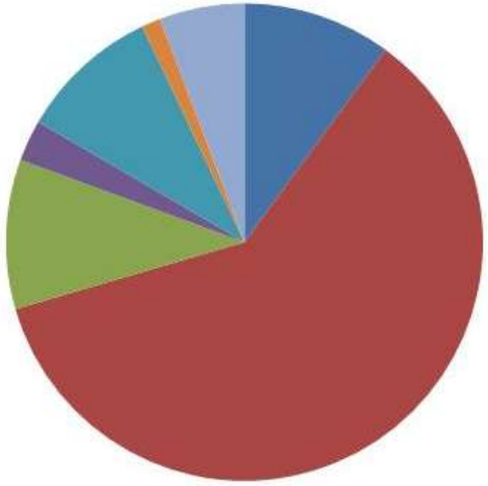
Most Common Microbial Phyla

Average Ulcerative Colitis



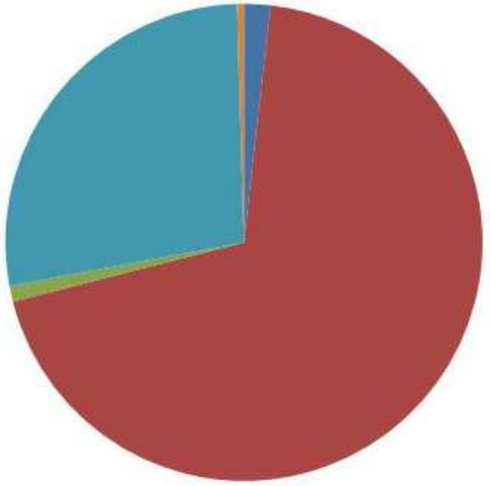
Explosion of Proteobacteria

Average LS



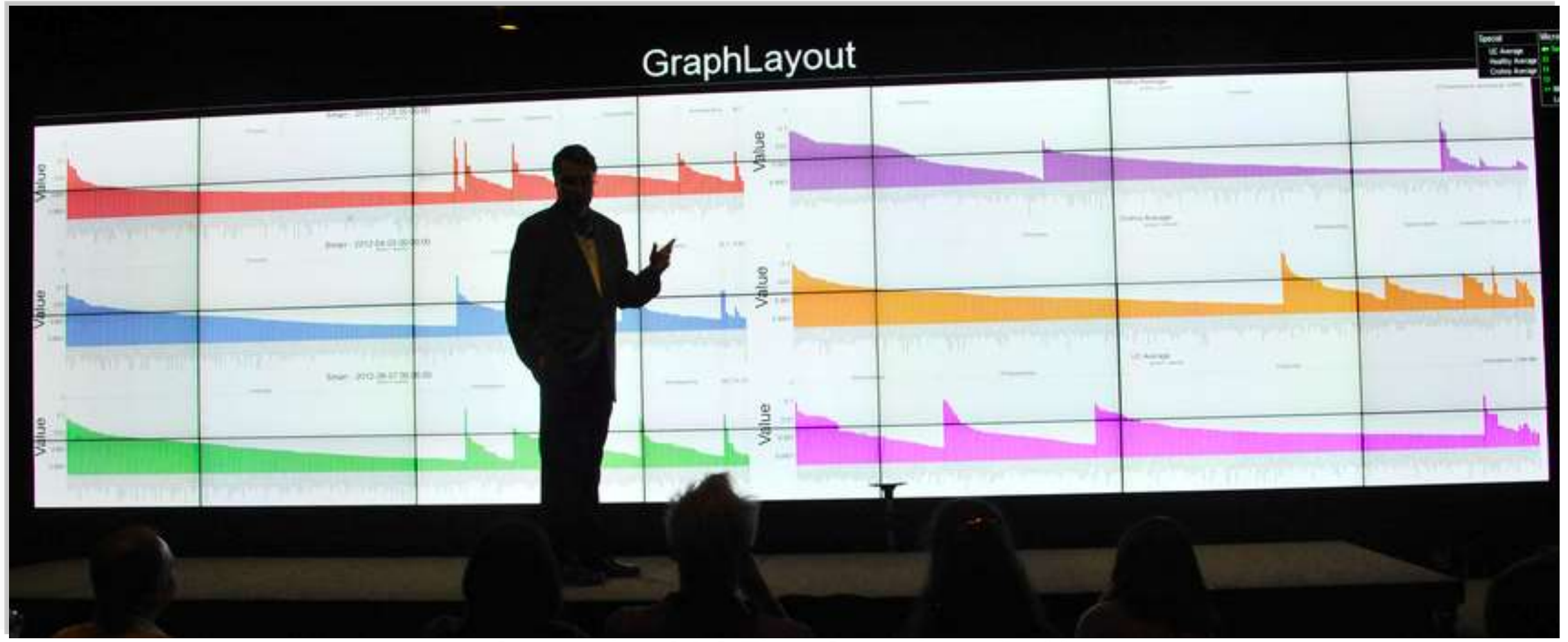
Hybrid of UC and CD
High Level of Archaea

Average Crohn's Disease



Collapse of Bacteroidetes
Explosion of Actinobacteria

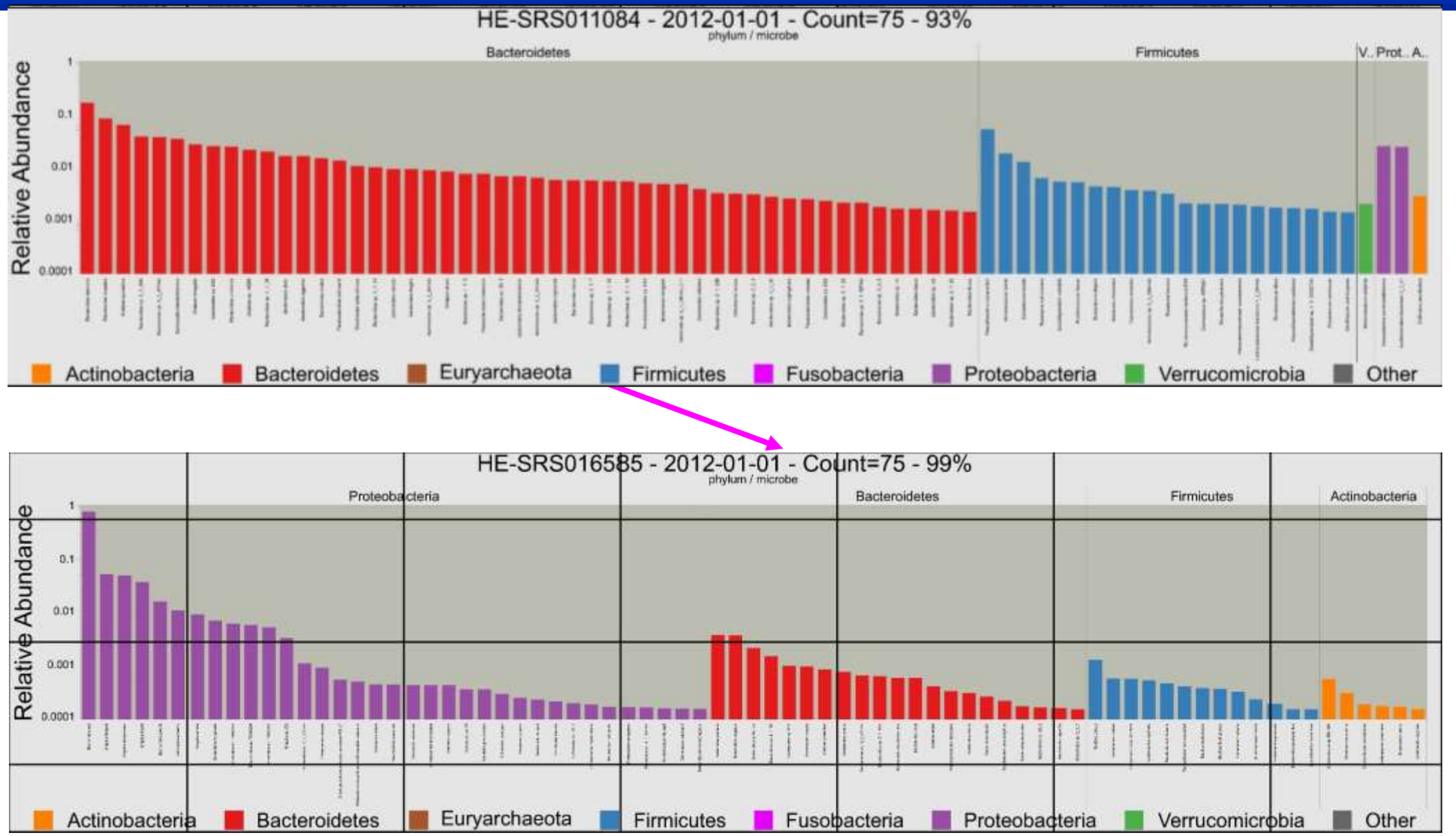
Using Scalable Visualization Allows Comparison of the Relative Abundance of 200 Microbe Species



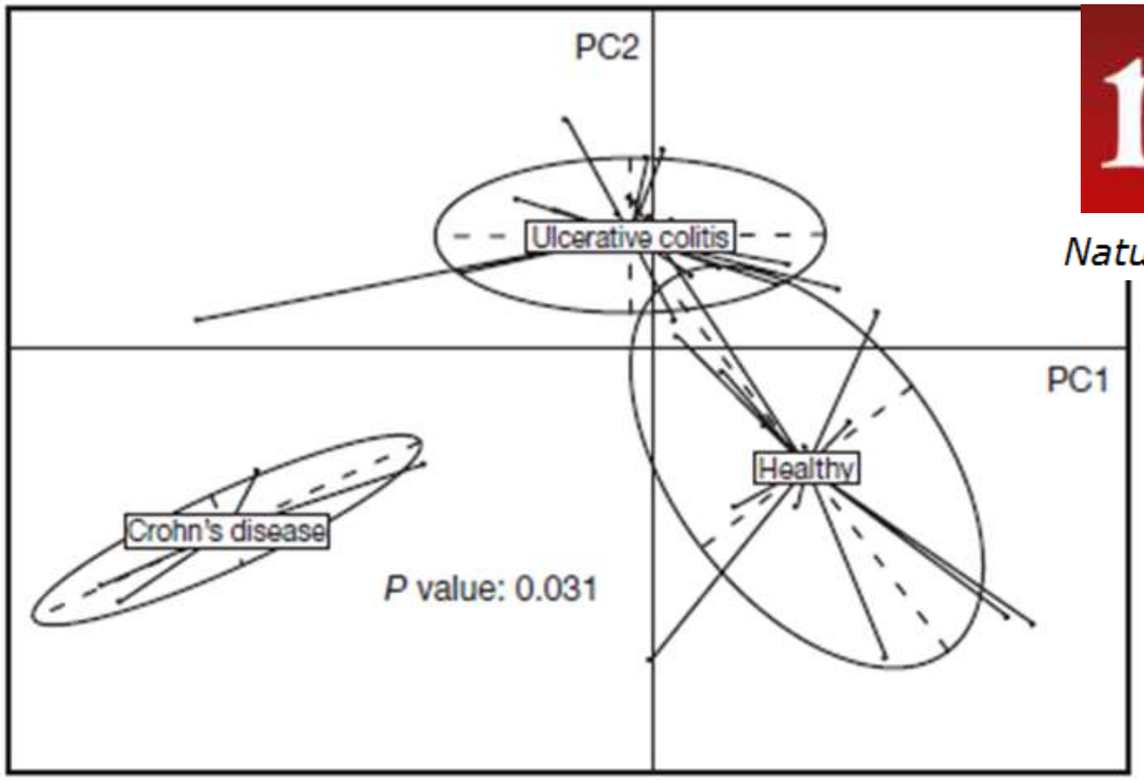
Comparing 3 LS Time Snapshots (Left)
with Healthy, Crohn's, Ulcerative Colitis (Right Top to Bottom)

Calit2 VROOM-FuturePatient Expedition

Using Microbiome Profiles to Survey 155 Subjects for Unhealthy Candidates



Using Principal Component Analysis To Stratify Disease States From Healthy States



nature

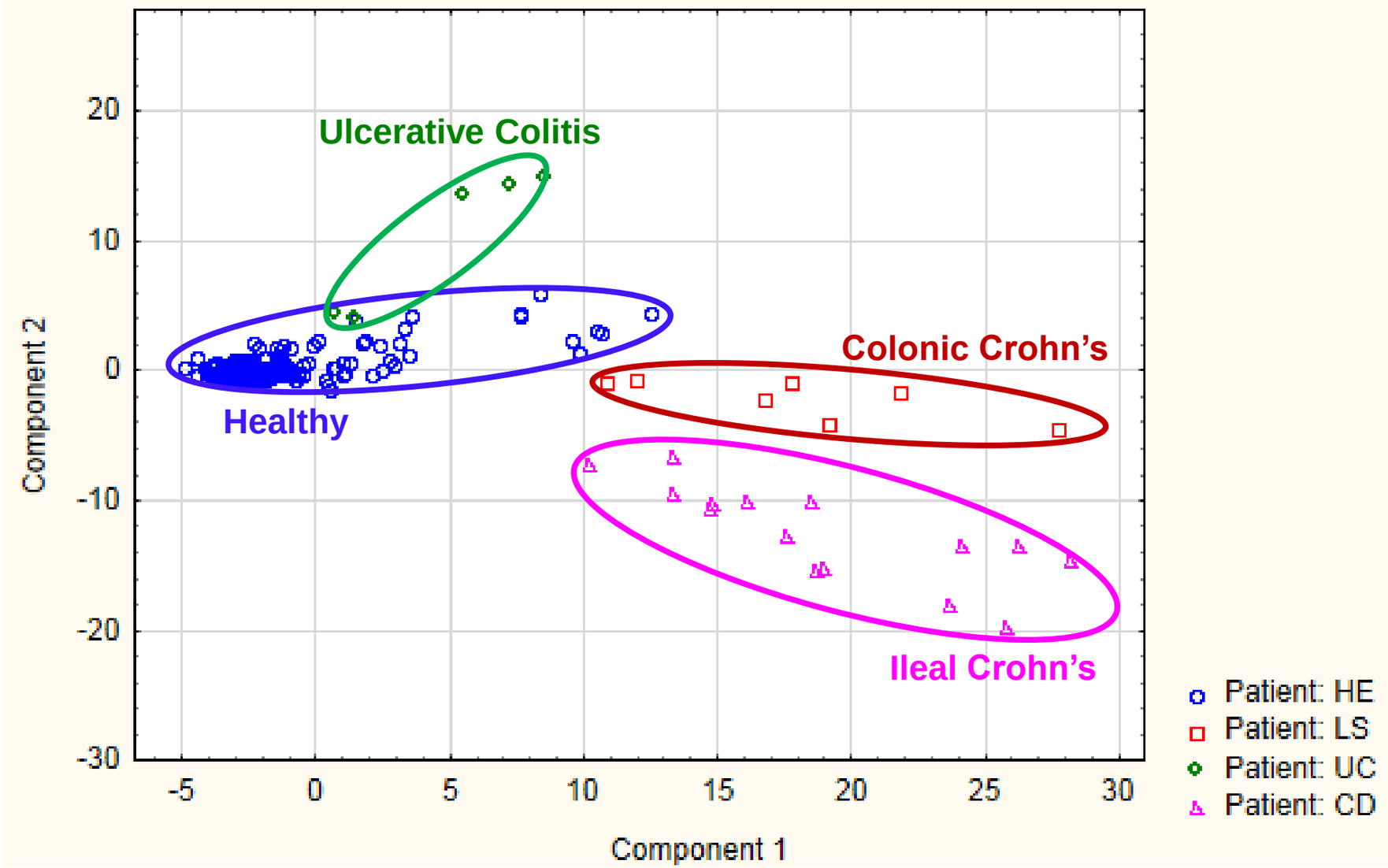
Nature **464**, 59-65 (4 March 2010)

Figure 4 | Bacterial species abundance differentiates IBD patients and healthy individuals. Principal component analysis with health status as

A human gut microbial gene catalogue established by metagenomic sequencing

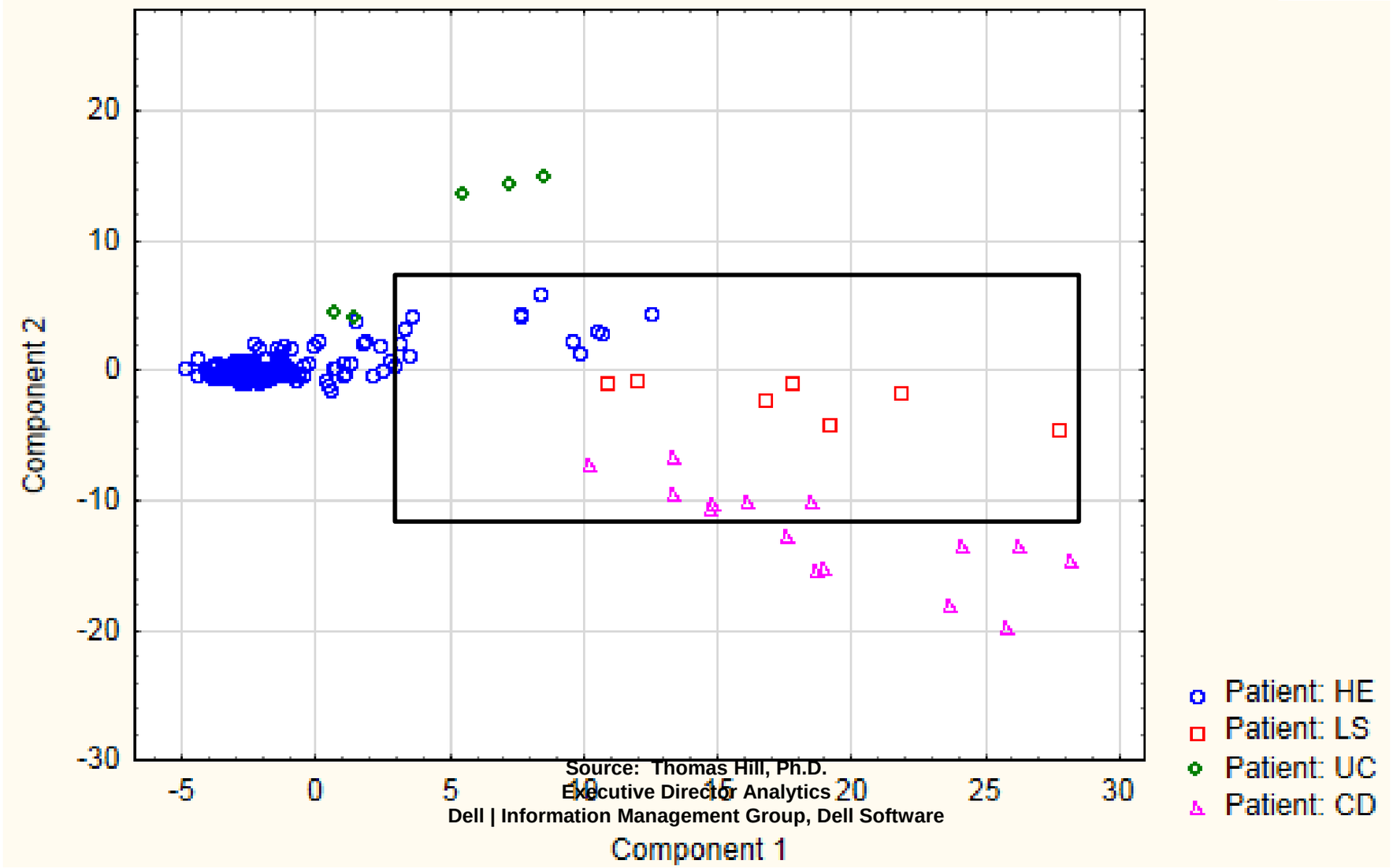
Junjie Qin[1]*, Ruiqiang Li[1]*, Jeroen Raes[2,3], Manimozhiyan Arumugam[2], Kristoffer Solvsten Burgdorf[4],

Dell Analytics Separates The 4 Patient Types in Our Using Our Microbiome Species Data



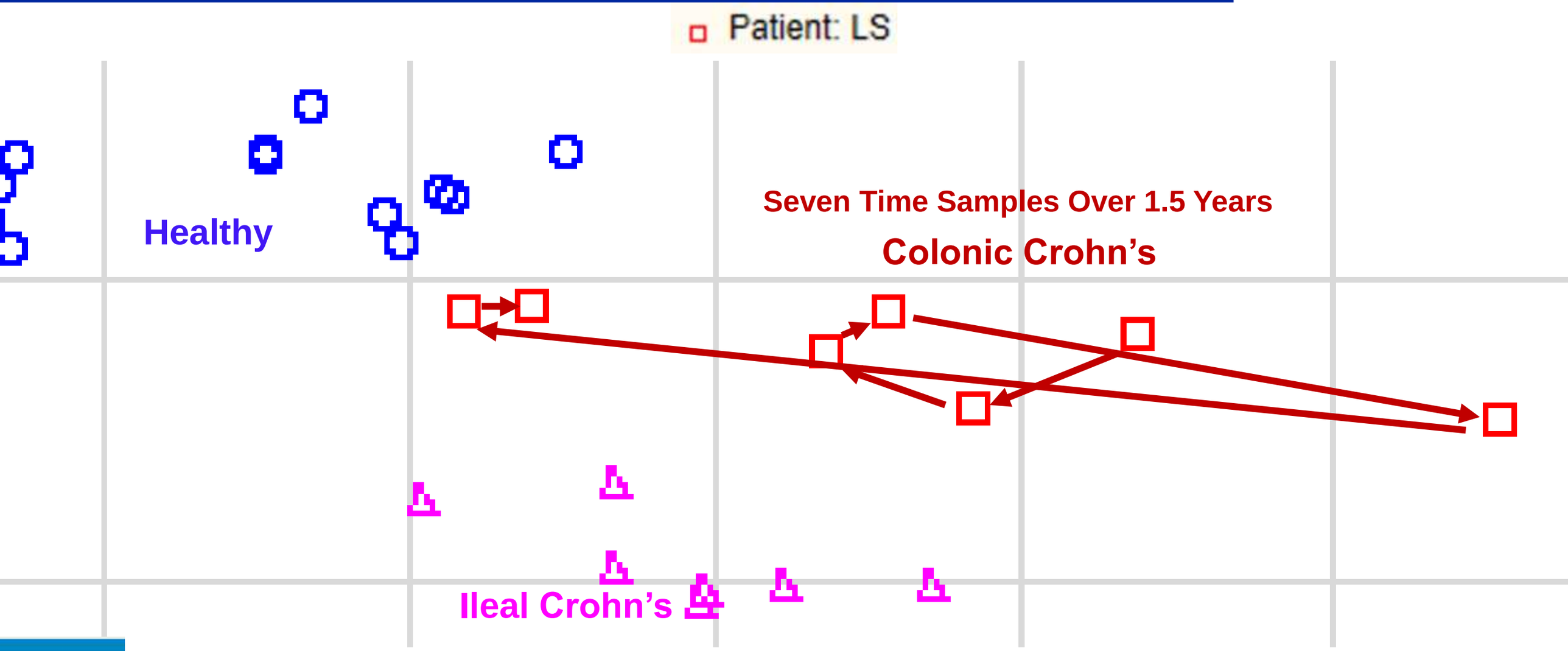
Source: Thomas Hill, Ph.D.
Executive Director Analytics
Dell | Information Management Group, Dell Software

I Built on Dell Analytics to Show Dynamic Evolution of My
Toward and Away from Healthy State – Colonic Cr



Source: Thomas Hill, Ph.D.
Executive Director Analytics
Dell | Information Management Group, Dell Software

I Built on Dell Analytics to Show Dynamic Evolution of My
Toward and Away from Healthy State – Colonic Cr



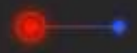
Using Ayasdi's Advanced Topological Data Analysis to Separate Healthy from Disease States

Using Ayasdi Categorical Data Lens

All Ileal Crohn's



All Healthy



All Healthy





All Healthy



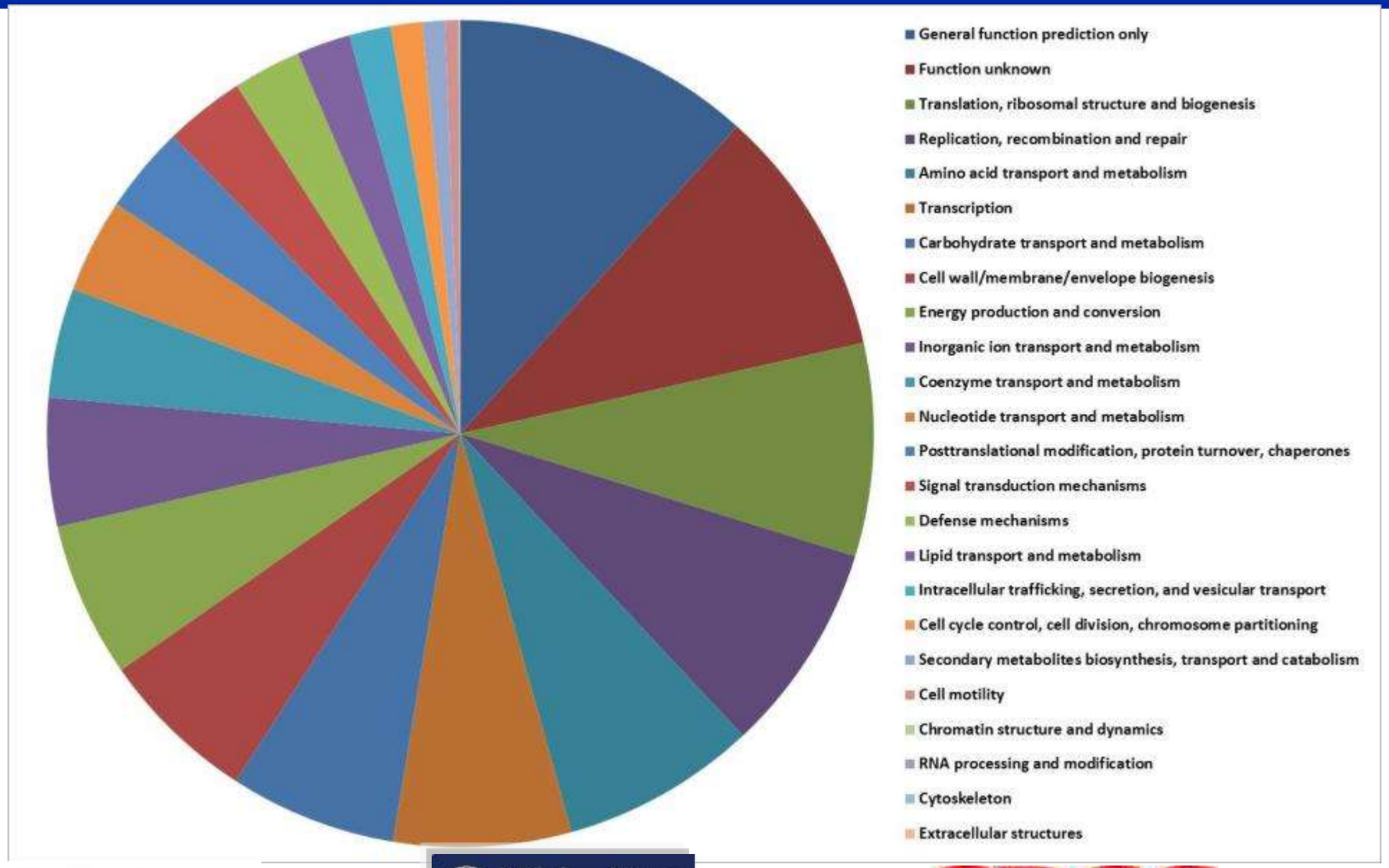
Healthy, Ulcerative Colitis, and LS

Analysis by Mehrdad Yazdani, Calit2

Ayasdi Cure™
Turn Data into Therapies

Talk to Ayasdi in the Intel Booth at SC14

From Taxonomy to Function: Analysis of LS Clusters of Orthologous Groups (COGs)



J. Craig Venter INSTITUTE

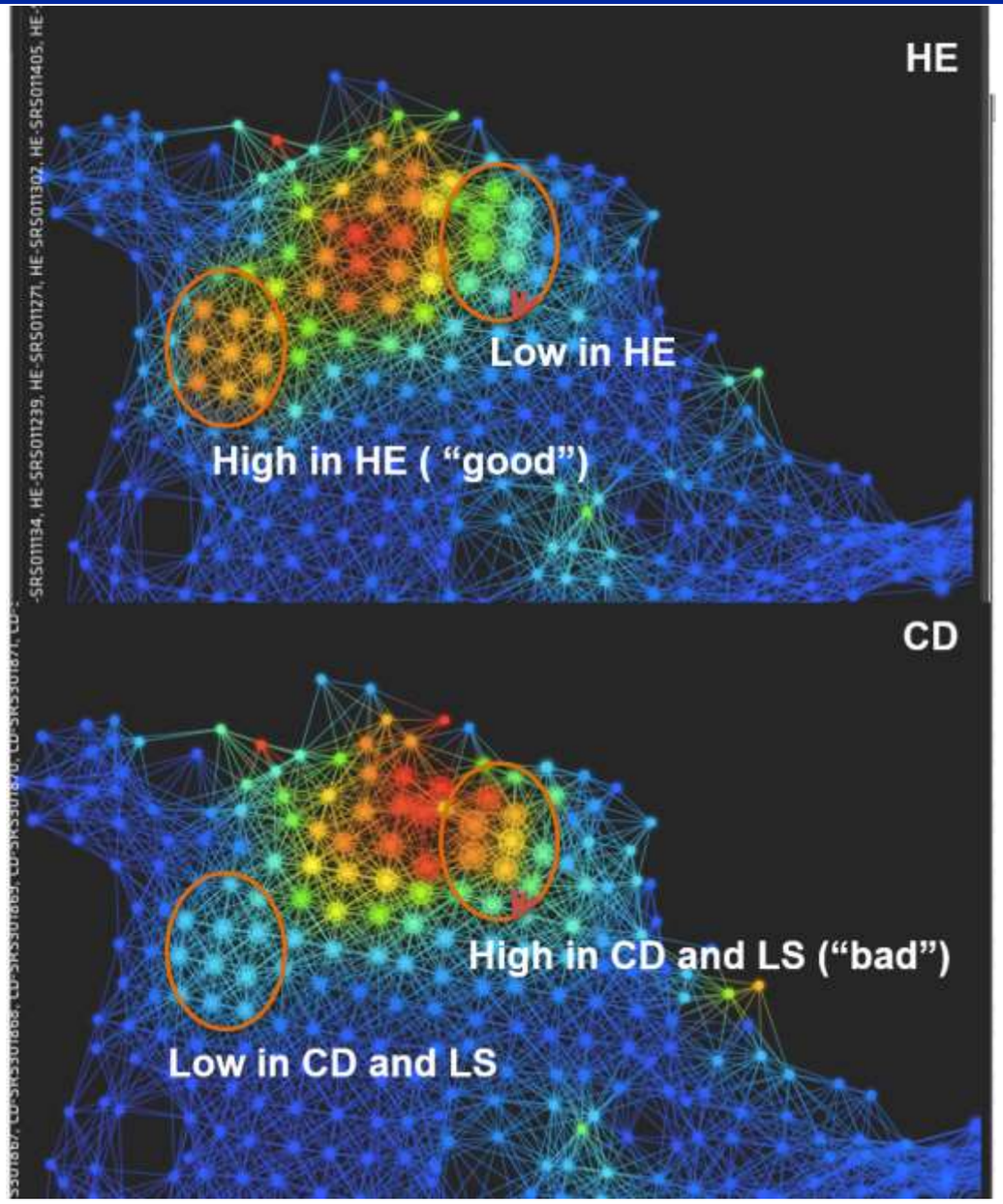
UC San Diego HEALTH SCIENCES

SDSC

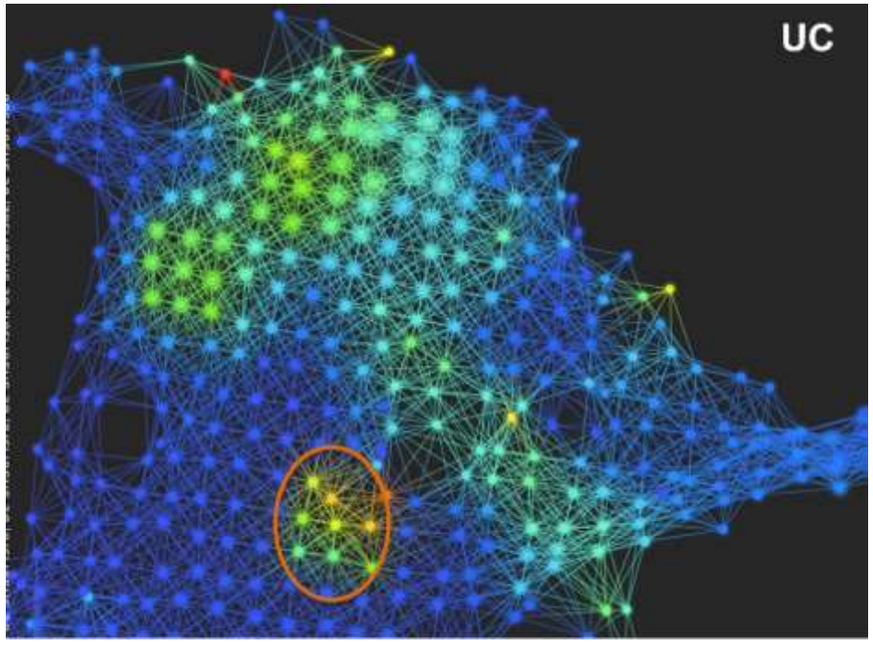
it²

Analysis: Weizhong Li & Sitao Wu, UCSD

Using Ayasdi Interactively to Explore Protein Families in Healthy and Disease States



Dataset from Larry Smarr Team
With 60 Subjects (HE, CD, UC, LS)
Each with 10,000 KEGGs -
600,000 Cells



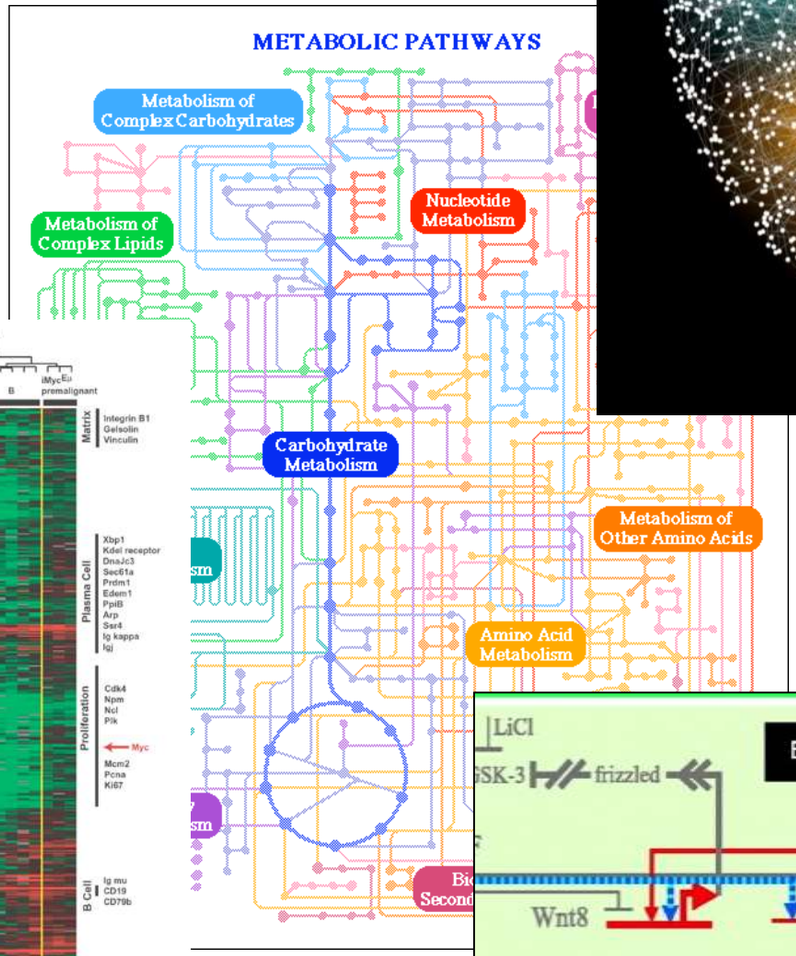
Source: Pek Lum,
Formerly Chief Data Scientist, Ayasdi



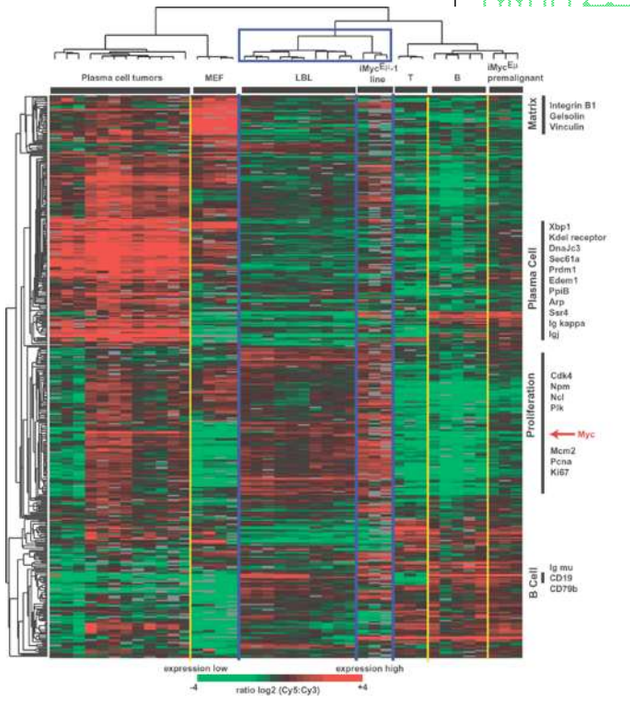
UCSD's Cytoscape Integrates and Visualizes Molecular Networks and Molecular Profiles

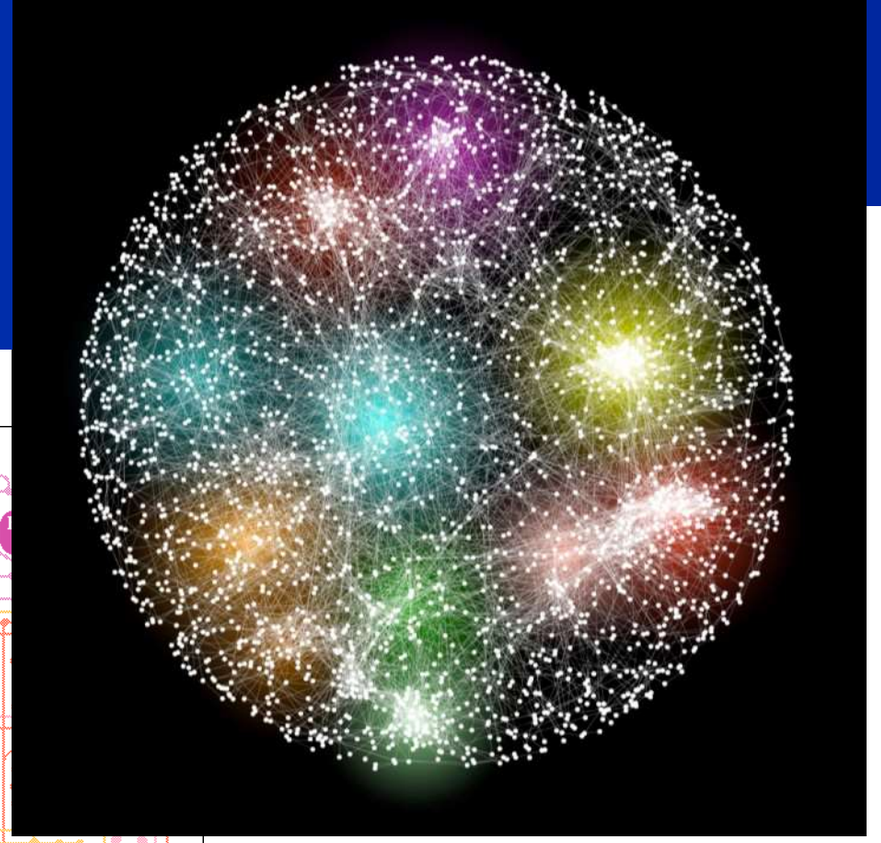
Source:
Trey Ideker, UCSD

Metabolic networks



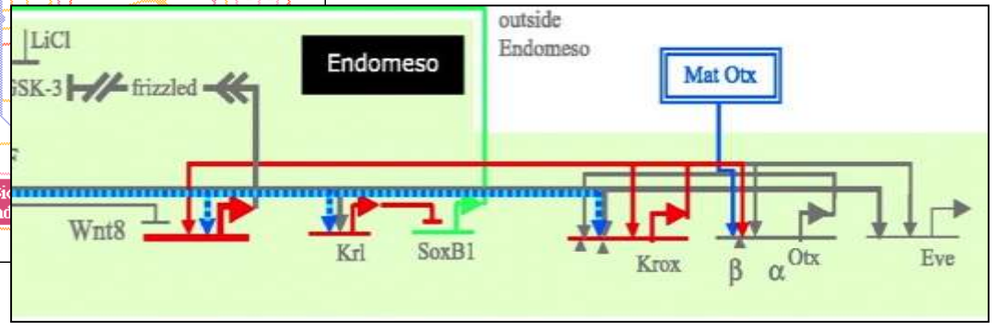
mRNA & protein expression



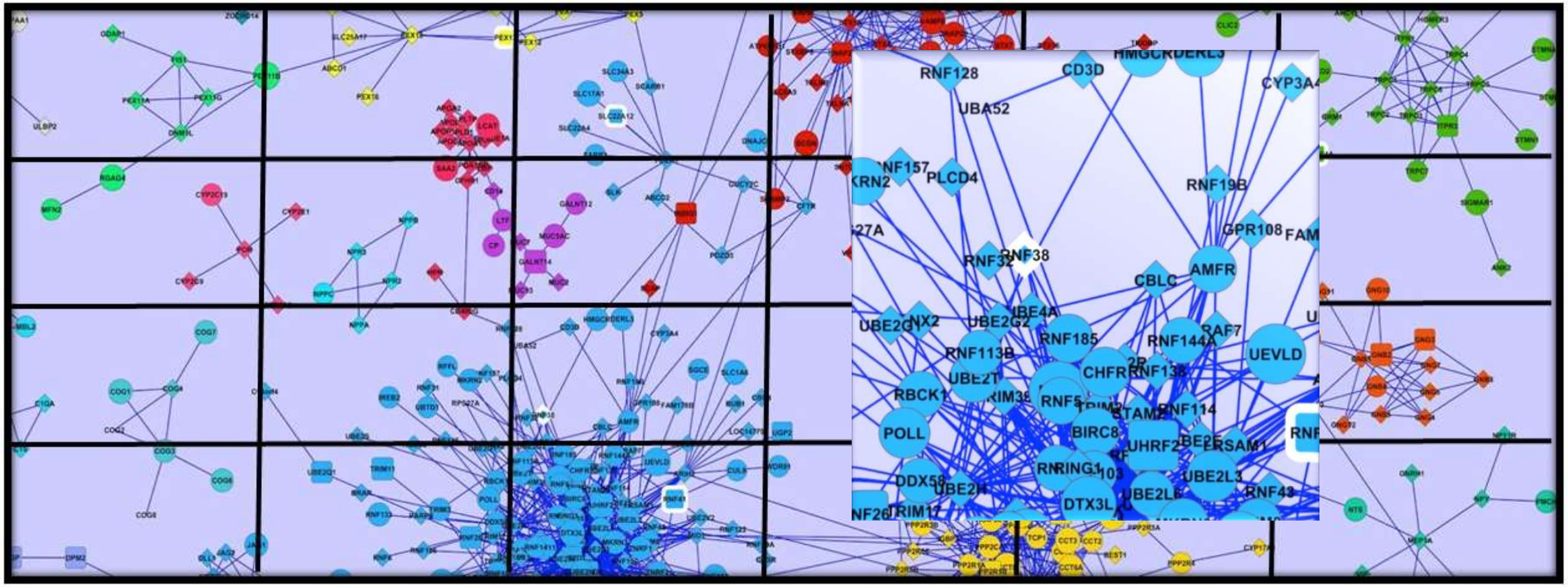


Genetic and protein interaction networks

Transcriptional networks



We Are Enabling Cytoscape to Run Natively on 64M Pixel Visualization Walls and in 3D in VR

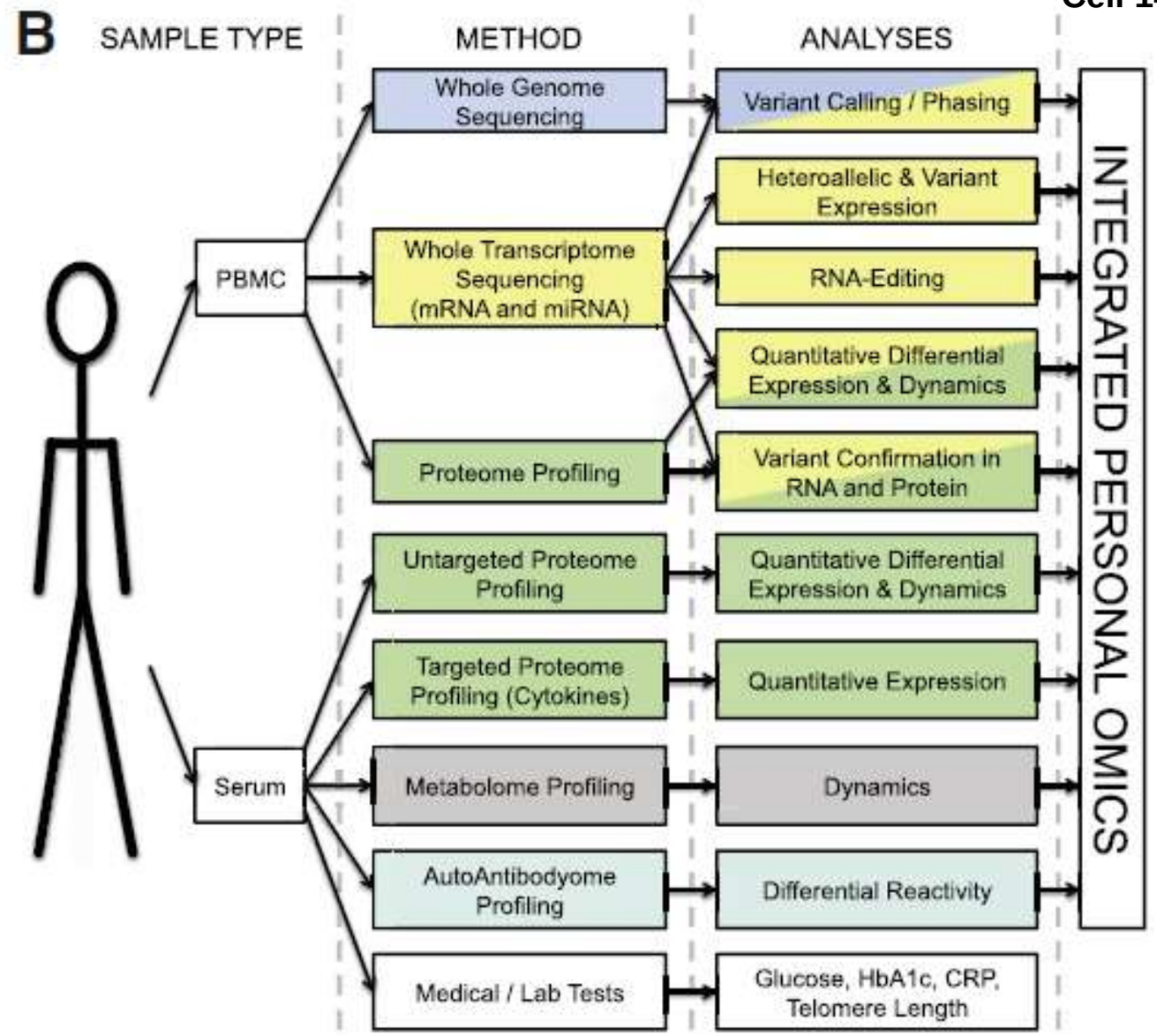


Simulation of Cytoscape Running on VROOM

Calit2 VROOM-FuturePatient Expedition

Cytoscape Example from Douglas S. Greer, J. Craig Venter Institute and Jurgen P. Schulze, Calit2's Qualcomm Institute

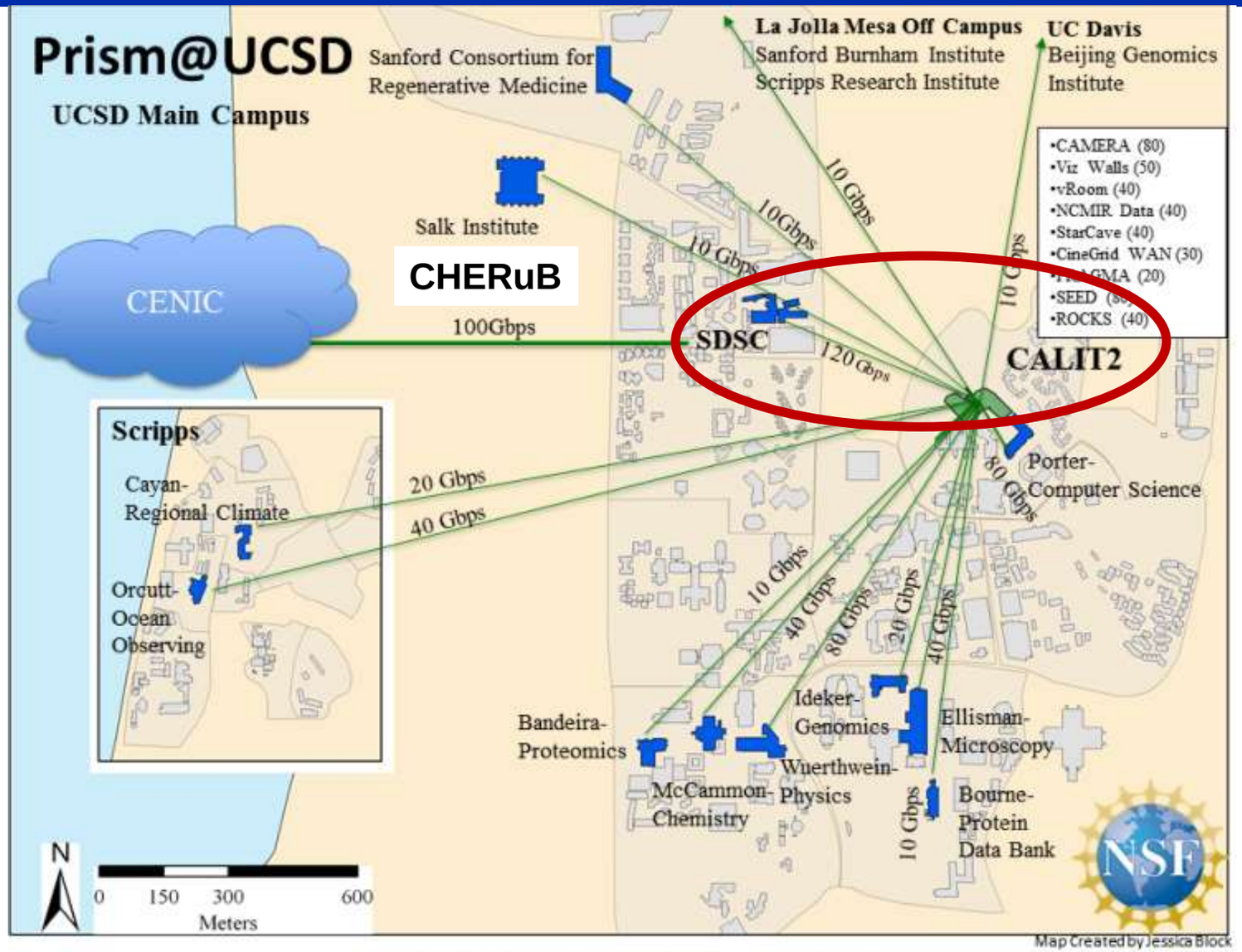
100x Beyond Current Medical Tests: Integrated Personal Time Series of Multiple 'Omics



Cell 148, 1293–1307, March 16, 2012

- **Michael Snyder, Chair of Genomics Stanford Univ.**
- **Blood Tests Time Series Over 40 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**

Connecting Data Generators and Storage/Compute Across a Campus at the Speed of a Cluster Backplane

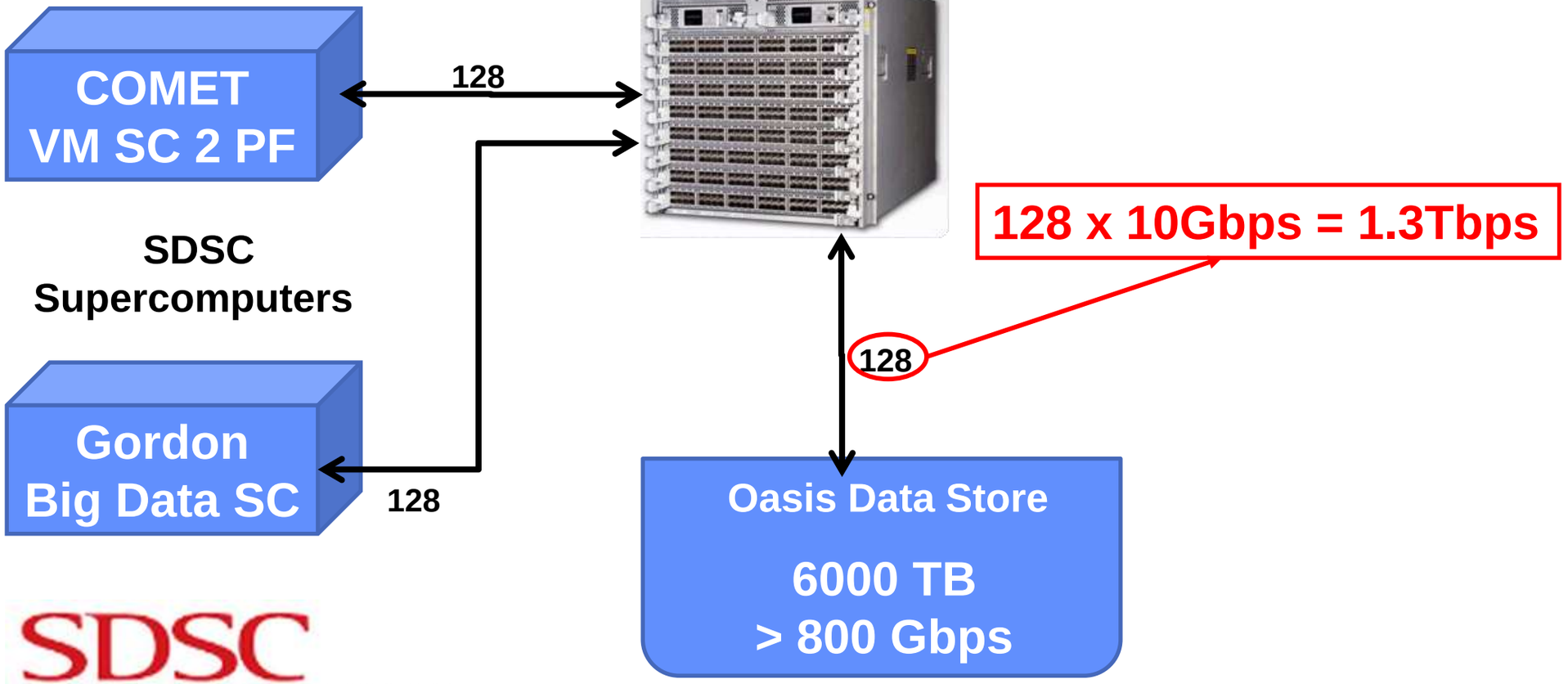


NSF CC-NIE Has Awarded 130 Such Grants To Over 100 U.S. Campuses

Big Data Compute/Storage Facility - Interconnected at Over 1 Tbps

# of Parallel 10Gbps Optical Light Paths

Arista Router
Can Switch
576 10Gps Light Paths



SDSC
SAN DIEGO SUPERCOMPUTER CENTER

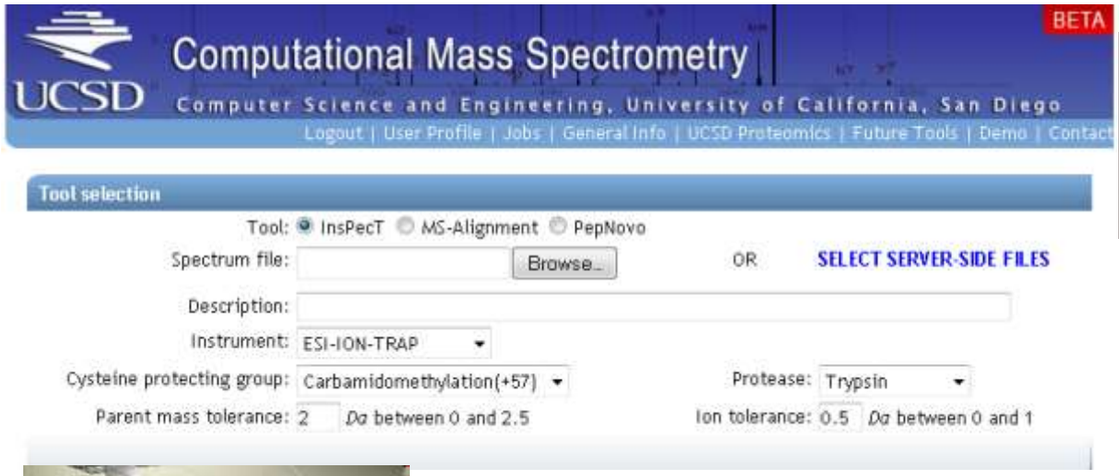
Source: Philip Papadopoulos, SDSC/Calit2

PRISM Will Link Computational Mass Spectrometry and Genome Sequencing Cores to the Big Data Freeway

Source: proteomics.ucsd.edu

***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





IGM Institute for Genomic Medicine

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

it²

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

UC San Diego Will Be Carrying Out a Major Clinical Study of IBD Using These Techniques

Announced Last Friday!

Inflammatory Bowel Disease Biobank For Healthy and Disease Patients

Already 120 Enrolled,
Goal is 1500

Drs. William J. Sandborn, John Chang, & Brigid Boland
UCSD School of Medicine, Division of Gastroenterology

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

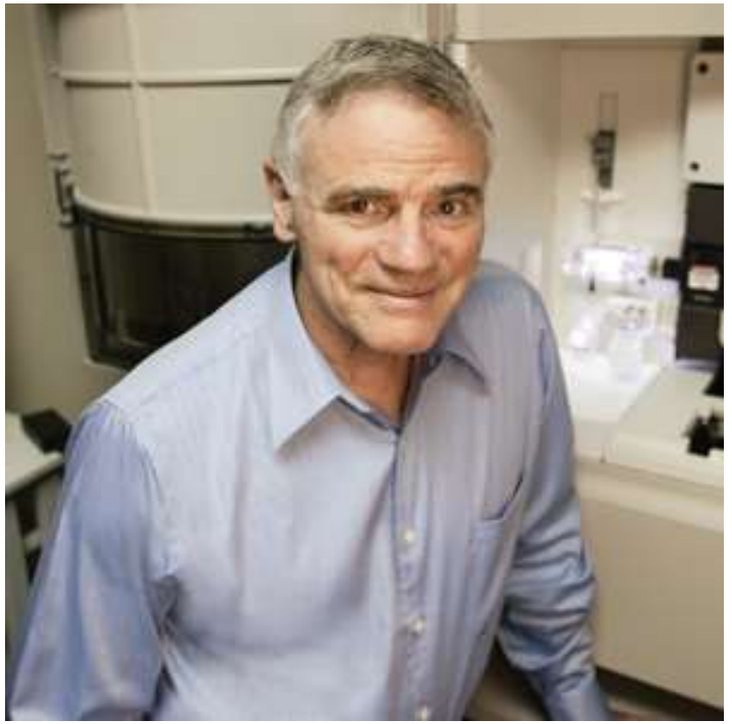
Newsweek

A Doctor's Vision of the Future of Medicine

Leroy Hood
NEWSWEEK

From the magazine issue dated Jul 13, 2009

PIONEER 100
HUNDRED PERSON WELLNESS PROJECT
Institute for Systems Biology



Will Grow to 1000, Then 10,000, Then 100,000

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

Genetic Sequencing of Humans and Their Microbes Is a Huge Growth Area and the Future Foundation of Medicine



Genomics england

The 100K Project
Institute for Systems Biology

REGENERON
GEISINGER
HEALTH SYSTEM

MVP

HUMAN LONGEVITY, INC.

Google BASELINE

Source: @EricTopol Twitter 9/27/2014

Thanks to Our Great Team!

UCSD Metagenomics Team
Weizhong Li
Sitao Wu

Calit2@UCSD Future Patient Team
Jerry Sheehan
Tom DeFanti
Kevin Patrick
Jurgen Schulze
Andrew Prudhomme
Philip Weber
Fred Raab
Joe Keefe
Ernesto Ramirez



Ayasdi
Devi Ramanan
Pek Lum



JCVI Team
Karen Nelson
Shibu Yooseph
Manolito Torralba

SDSC Team
Michael Norman
Mahidhar Tatineni
Robert Sinkovits

Dell/R Systems
Brian Kucic
John Thompson

UCSD Health Sciences Team
William J. Sandborn
Elisabeth Evans
John Chang
Brigid Boland
David Brenner