“Machine Learning Opportunities in the Explosion of Personalized Precision Medicine”

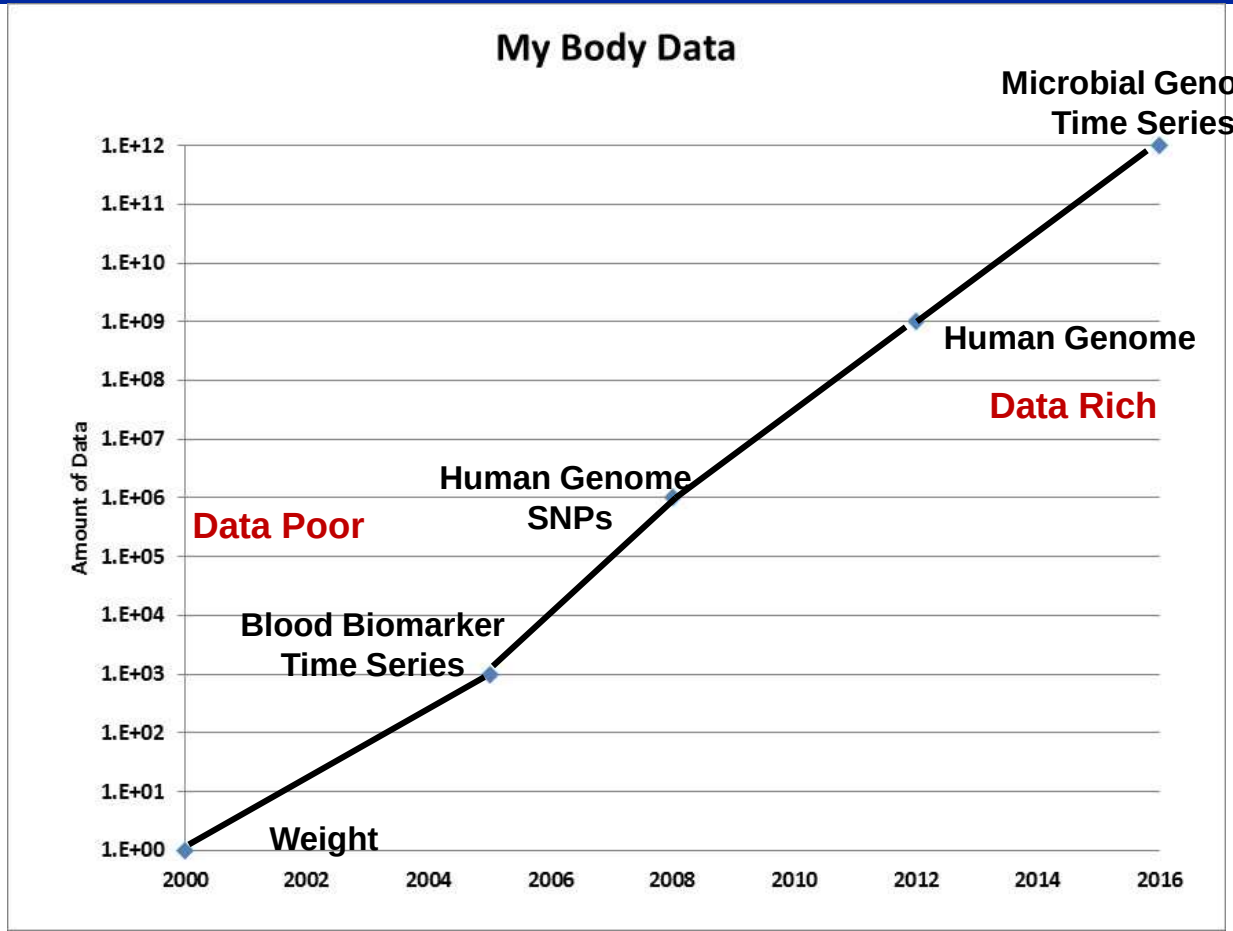
Invited Presentation
Machine Learning in Healthcare
Saban Research Institute
Los Angeles, CA
August 19, 2016

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

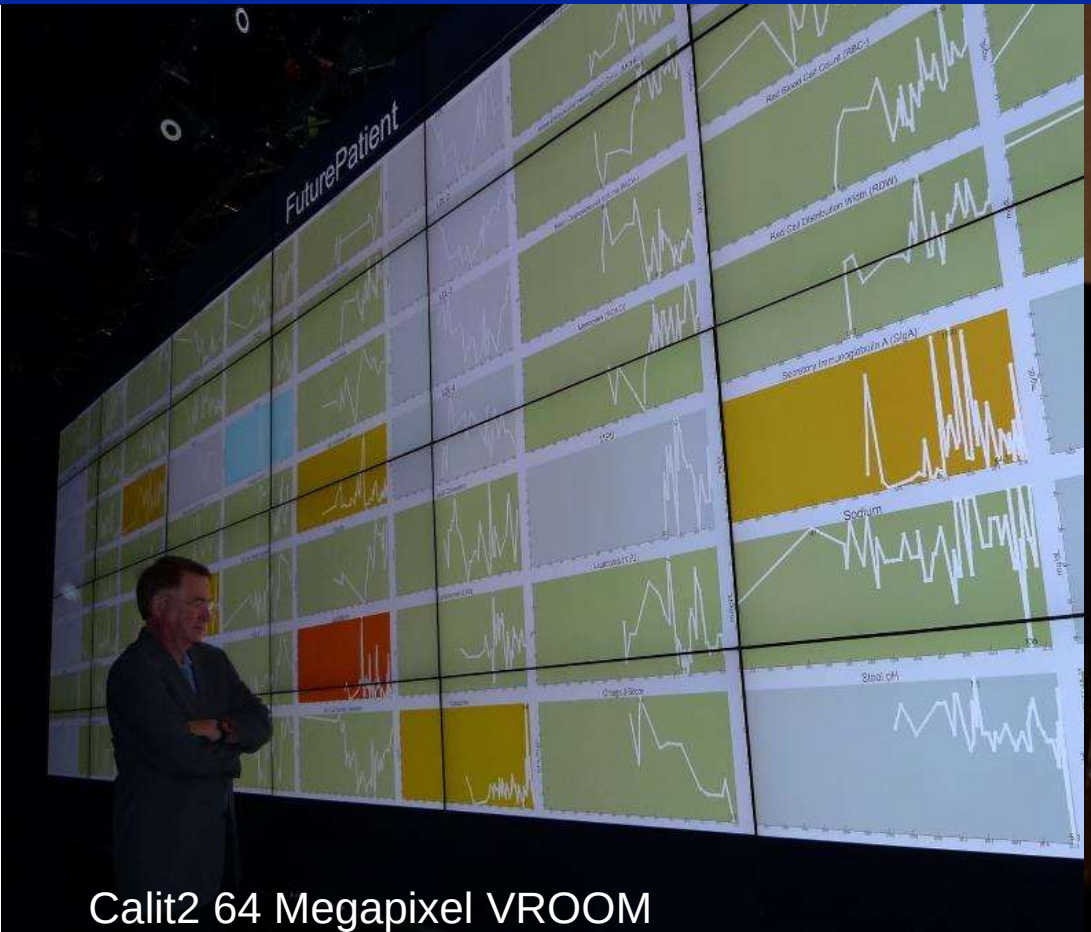
We have reached the take off point in the generation of massive datasets from individuals and across populations, both of which are necessary for personalized precision medicine. I will give an example of my N=1 self-study, in which I have my human genome as well as multi-year time series of my gut microbiome genomics and over one hundred blood biomarkers. This is now being augmented with time series of my metabolome and immunome. These are then compared with hundreds of healthy people's gut microbiomes, revealing major shifts between health and disease. Multiple companies and organizations will soon be carrying out similar levels of analysis on hundreds of thousands of individuals. Machine learning techniques will be essential to bring the patterns out of these exponentially growing datasets.

Calit2's Future Patient Project: How Does Medicine Transform in a Data-Rich World?



My Body Produces 1 Trillion Times as Much Data in Only 15 Years!

I Decided to Track My Internal Biomarkers To Understand My Body's Dynamics

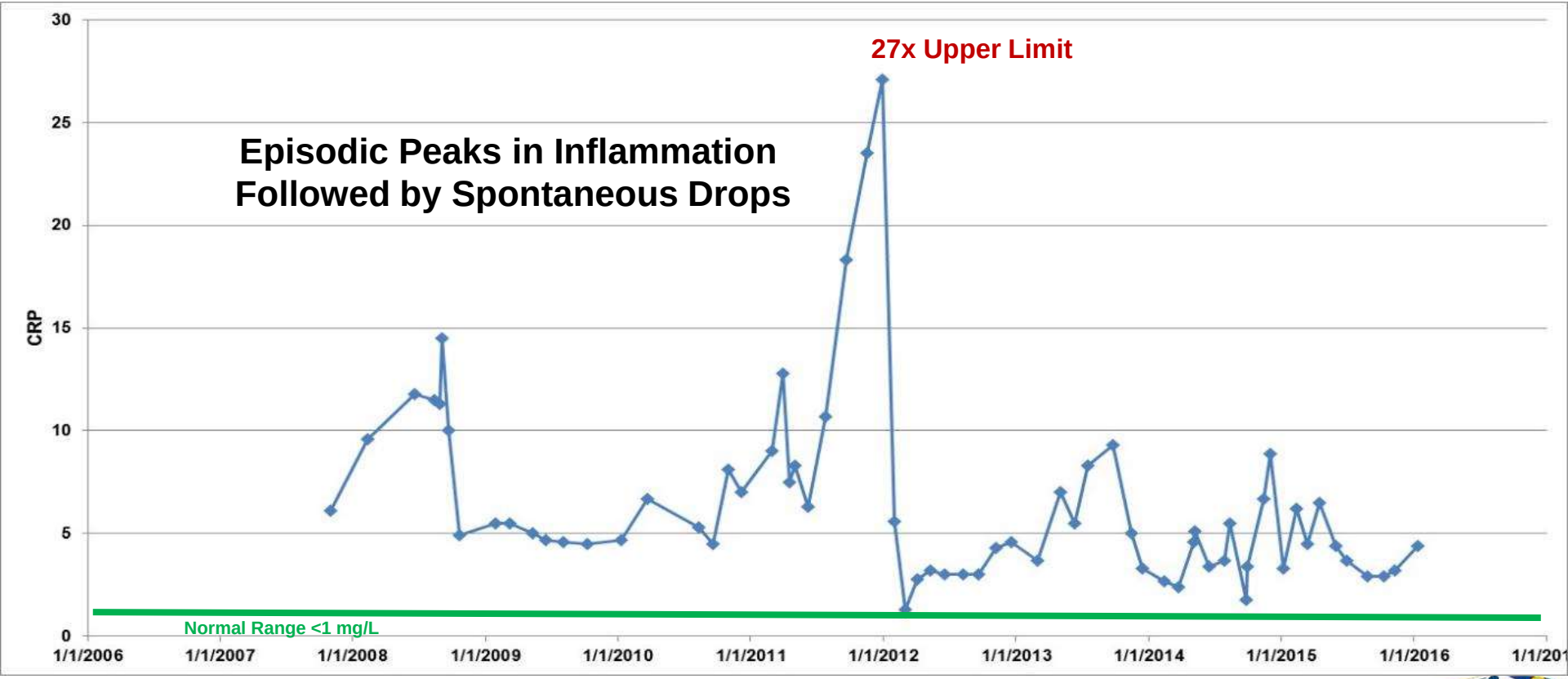


Calit2 64 Megapixel VROOM



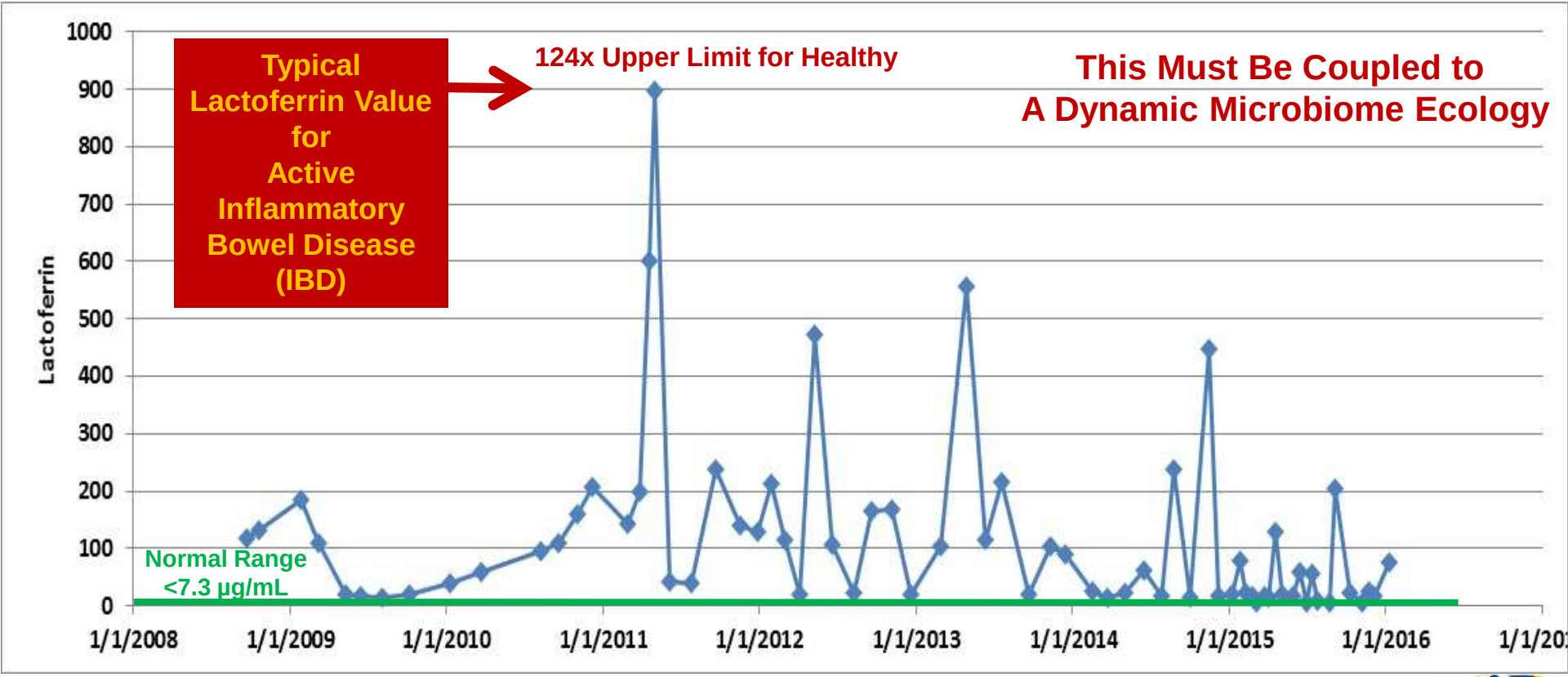
My Quarterly Blood Draw

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



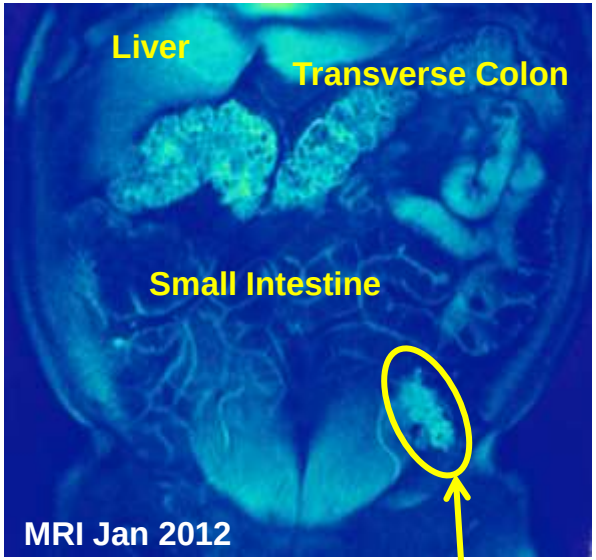
Complex Reactive Protein (CRP) is a Blood Biomarker for Detecting Presence of Inflammation

Adding Stool Tests Revealed
Oscillatory Behavior in an Immune Variable Which is Antibacterial



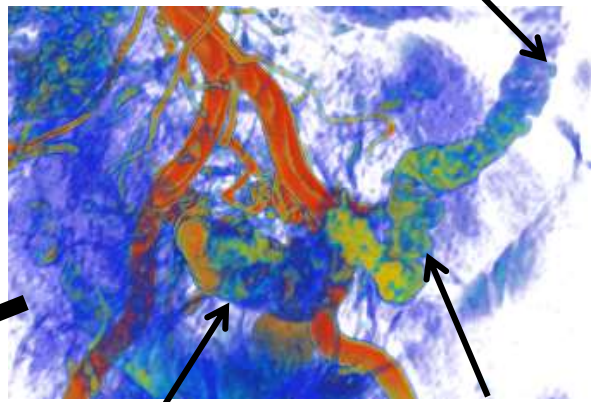
Lactoferrin is a Protein Shed from Neutrophils -
An Antibacterial that Sequesters Iron

Confirming the IBD (Colonic 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

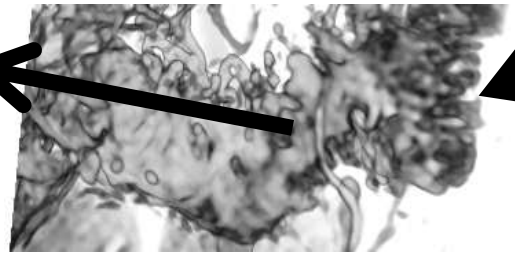
Descending Colon



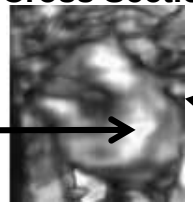
Major Kink

Sigmoid Colon Threading Iliac Arteries

Diseased Sigmoid Colon



Cross Section



Severe Colon Wall Swelling

To Understand the Autoimmune Dynamics of the Immune System We Must Consider the Human Microbiome

Your Microbiome is
Your "Near-Body" Environment
and its Cells
Contain 100x as Many DNA Genes
As Your Human DNA-Bearing Cells

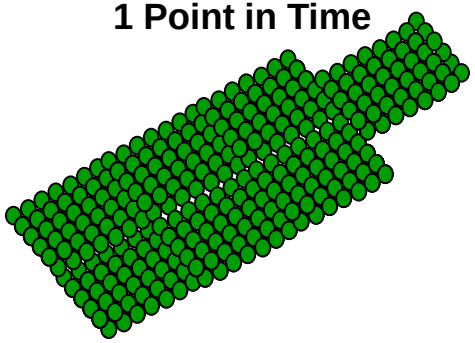
Inclusion of the "Dark Matter" of the Body
Will Radically Alter Medicine

We Downloaded Metagenomic Sequencing of the Gut Microbiome of Healthy and IBD Patients and Compared with My Time Series

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

"Healthy" Individuals

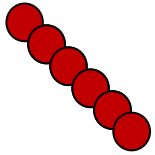
250 Subjects
1 Point in Time



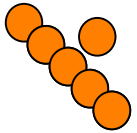
Inflammatory Bowel Disease (IBD) Patients

Larry Smarr
(Colonic Crohn's)

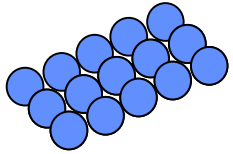
7 Points in Time
Over 1.5 Years



2 Ulcerative Colitis Patients,
6 Points in Time



5 Ileal Crohn's Patients,
3 Points in Time



NIH

HMP

J. Craig Venter INSTITUTE

Total of 27 Billion Reads
Or 2.7 Trillion Bases

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

it²

To Map Out the Dynamics of Autoimmune Microbiome Ecology Couples Next Generation Genome Sequencers to *Big Data* Supercomputers

Source: Weizhong Li, UCSD

Our Team Used 25 CPU-years to Compute Comparative Gut Microbiomes Starting From 2.7 Trillion DNA Bases from My Time Samples and 255 Healthy and 20 IBD Controls

Illumina HiSeq 2000 at JCVI



SDSC Gordon Data Supercomputer



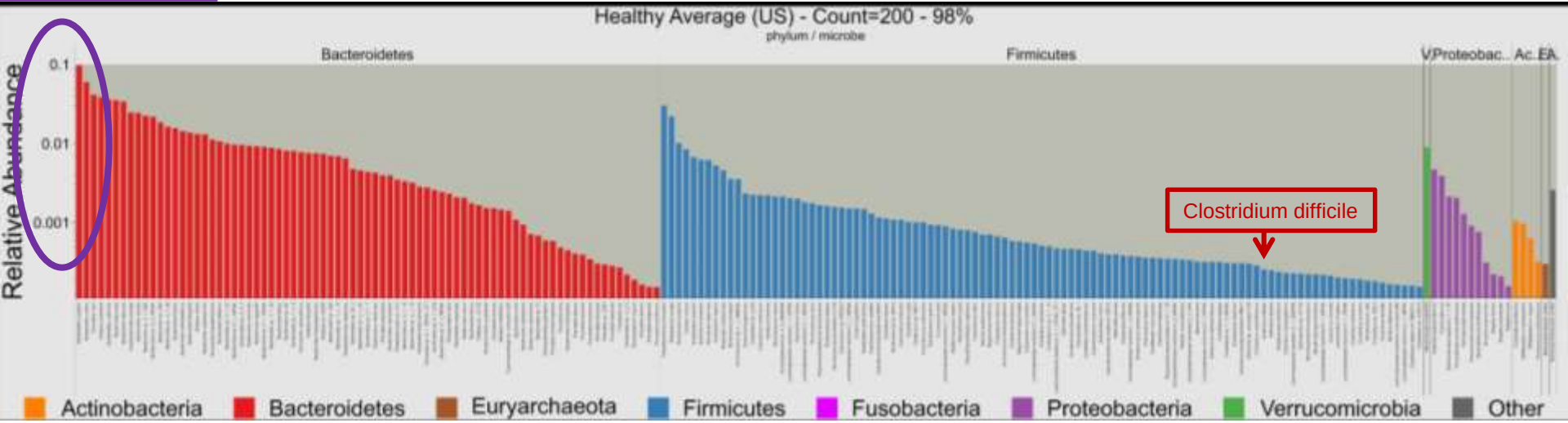
Dell Solutions Center
Industry Solutions Lab
SANGER DSU

J. Craig Venter INSTITUTE · SDSC · DELL · NIH · HMP · CRBS · it²

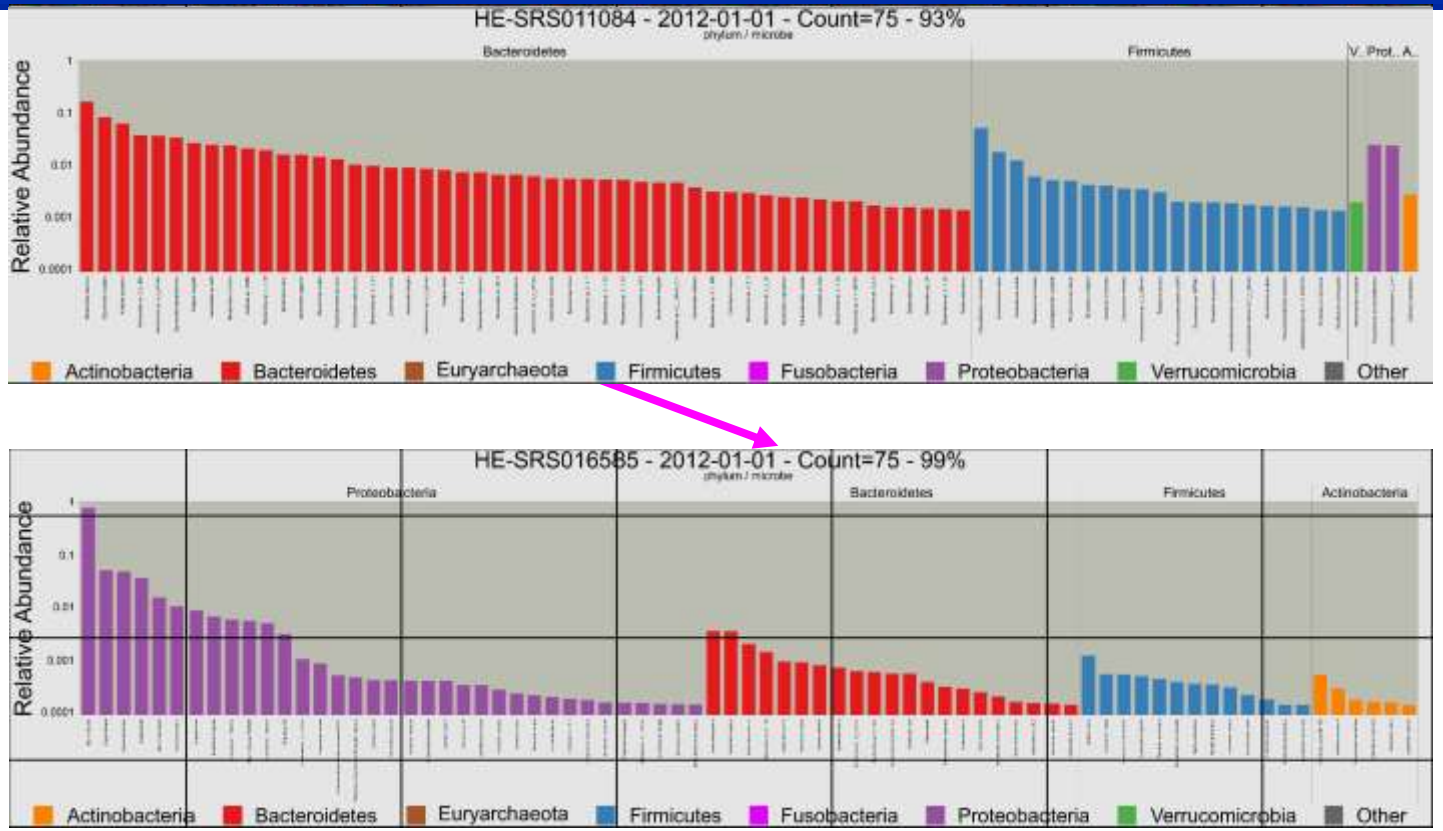
Results Include Relative Abundance of Hundreds of Microbial Species

Average Over 250 Healthy People
From NIH Human Microbiome Project

Note Log Scale

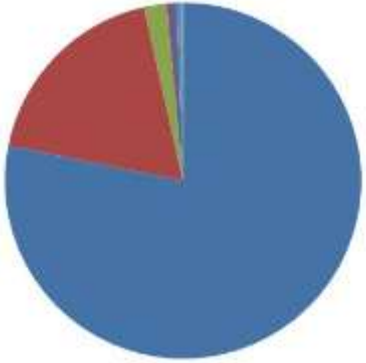


Using Microbiome Profiles to Survey 155 Subjects for Unhealthy Candidates



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

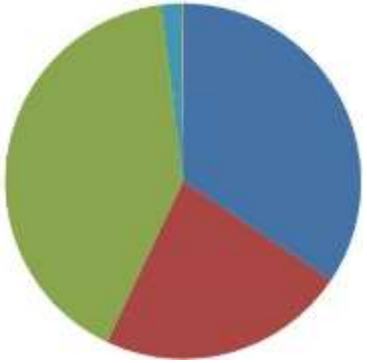
Average HE



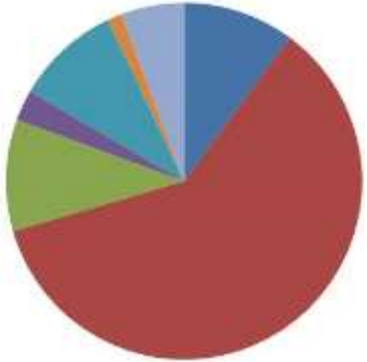
Most Common Microbial Phyla

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

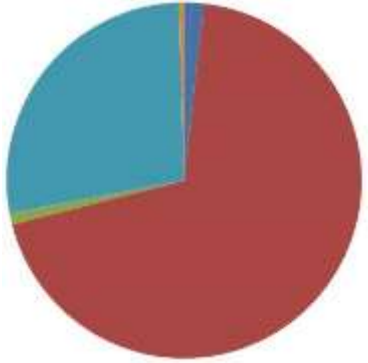
Average Ulcerative Colitis



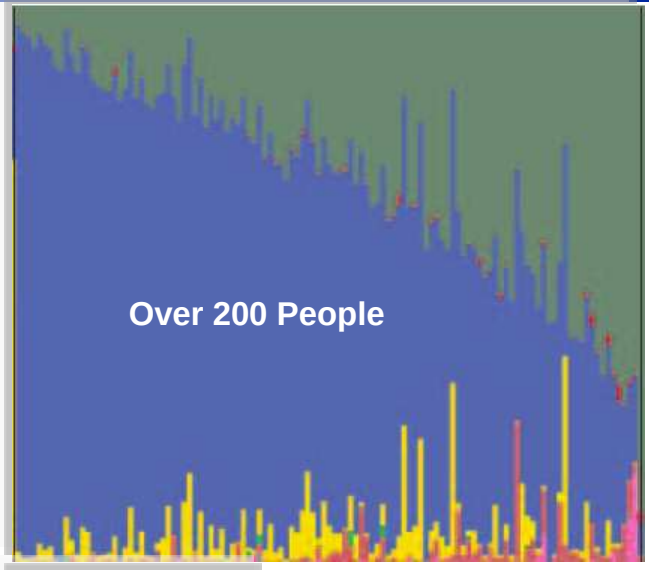
Average LS Colonic Crohn's Disease

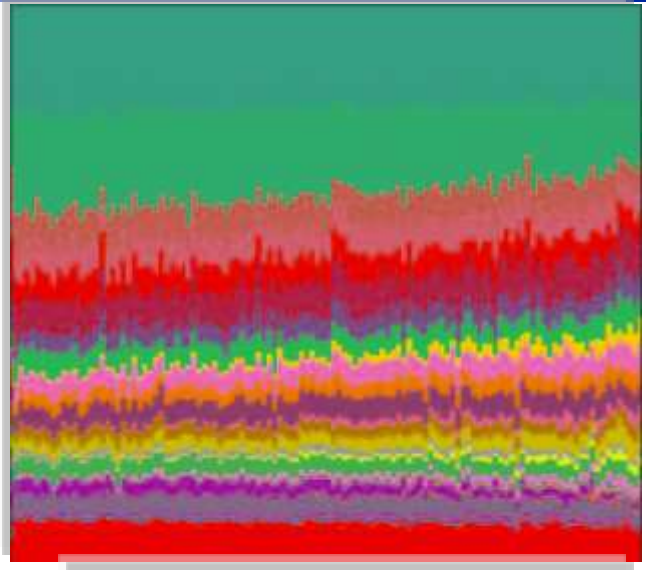


Average Ileal Crohn's Disease



In a “Healthy” Gut Microbiome:
Large Taxonomy Variation, Low Protein Family Variation





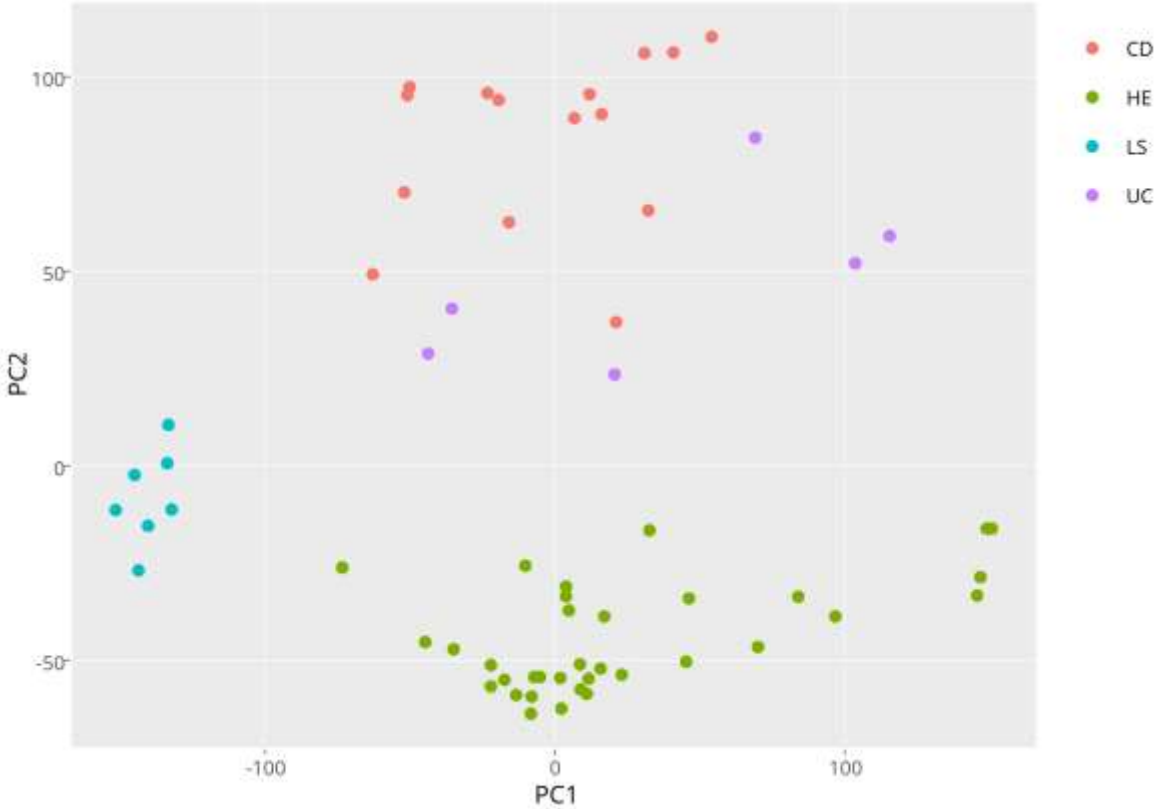
- Firmicutes
- Actinobacteria
- Bacteroidetes
- Proteobacteria
- Fusobacteria
- Tenericutes
- Spirochaetes
- Cyanobacteria
- Verrucomicrobia
- TM7

- Central carbohydrate metabolism
- Cofactor and vitamin biosynthesis
- Oligosaccharide and polyol transport system
- Purine metabolism
- ATP synthesis
- Phosphate and amino acid transport system
- Aminoacyl tRNA
- Pyrimidine metabolism
- Ribosome
- Aromatic amino acid metabolism

Source: Nature, 486, 207-212 (2012)

We Supercomputed ~10,000 Microbiome Protein Families (KEGGs) Which Clearly Separate Disease Subtypes Using PCA

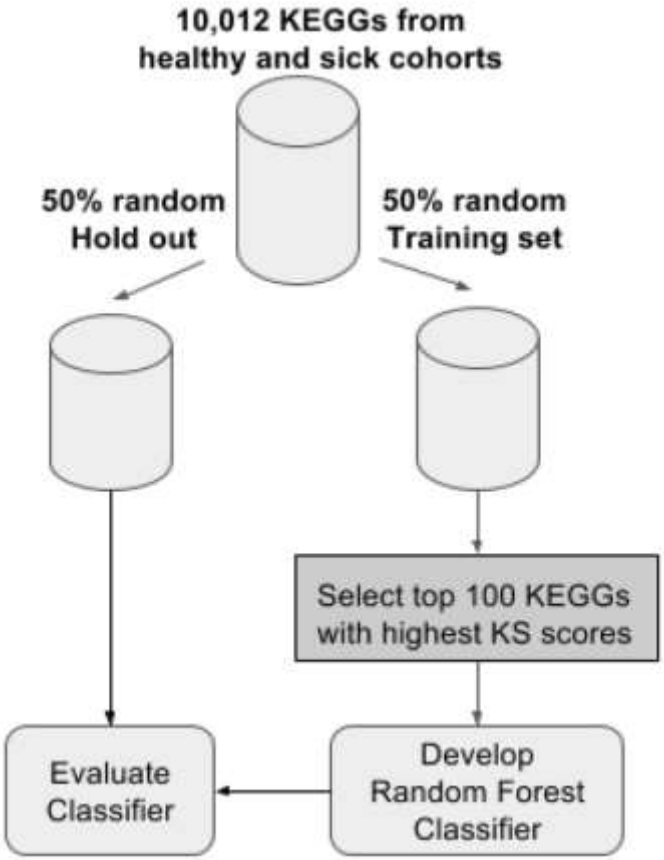
Implies That Disease Subtypes Have Distinct Protein Distributions



Computing KEGGs Required 10 CPU-Years On SDSC's Gordon Supercomputer

Source: Computing Weizhong Li, PCA Mehrdad Yazdani, Calit2

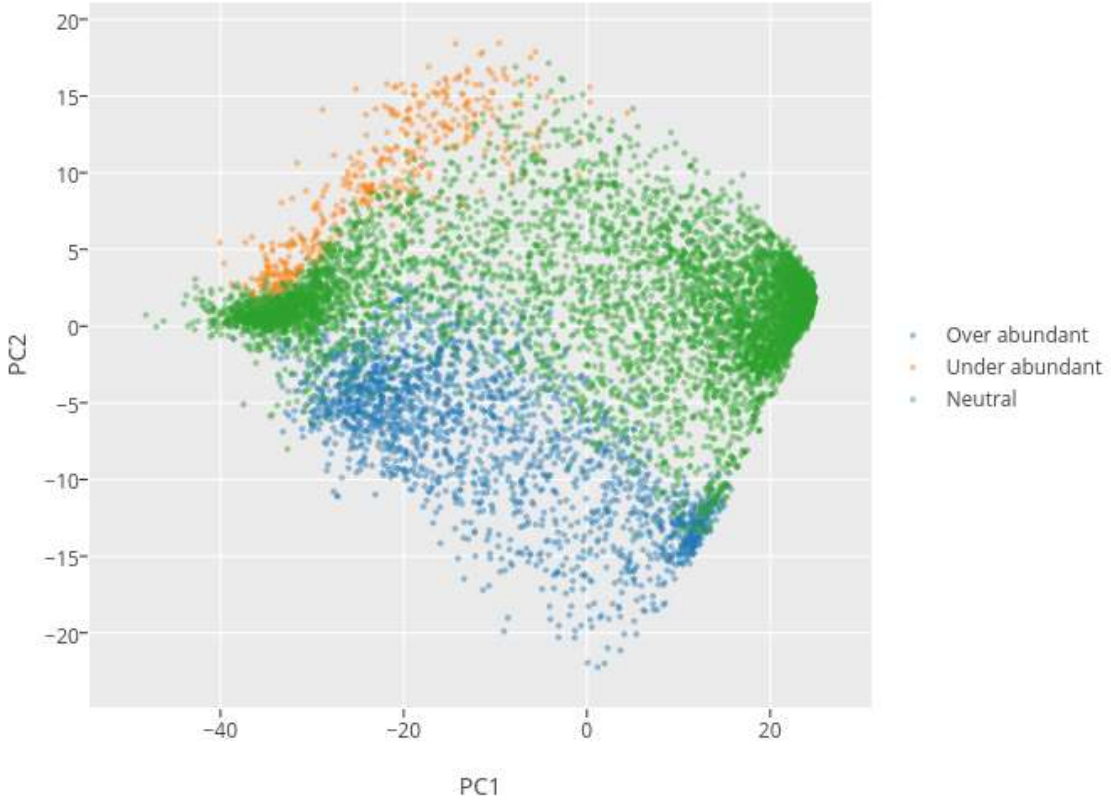
Using Machine Learning to Identify Protein Families That Are Over or Under Abundant in Disease State



- Split KEGGs into 50% Training and Holdout Sets
- In Training set, Compute Kolmogorov-Smirnov Test to Find Statistically Most Significant KEGGs That Differentiate Healthy and Disease States
- Train a Random Forest as a Probabilistic Binary Classifier on 100 KEGGs with Highest KS Scores
- Use Trained RF to Classify all KEGGs as Over or Under Abundant

PCA Plot of the Random Forest Classifier Probability Confidence Level Applied to All 10,012 KEGGs

Note Tight Clustering of Over and Under Abundant Protein Families

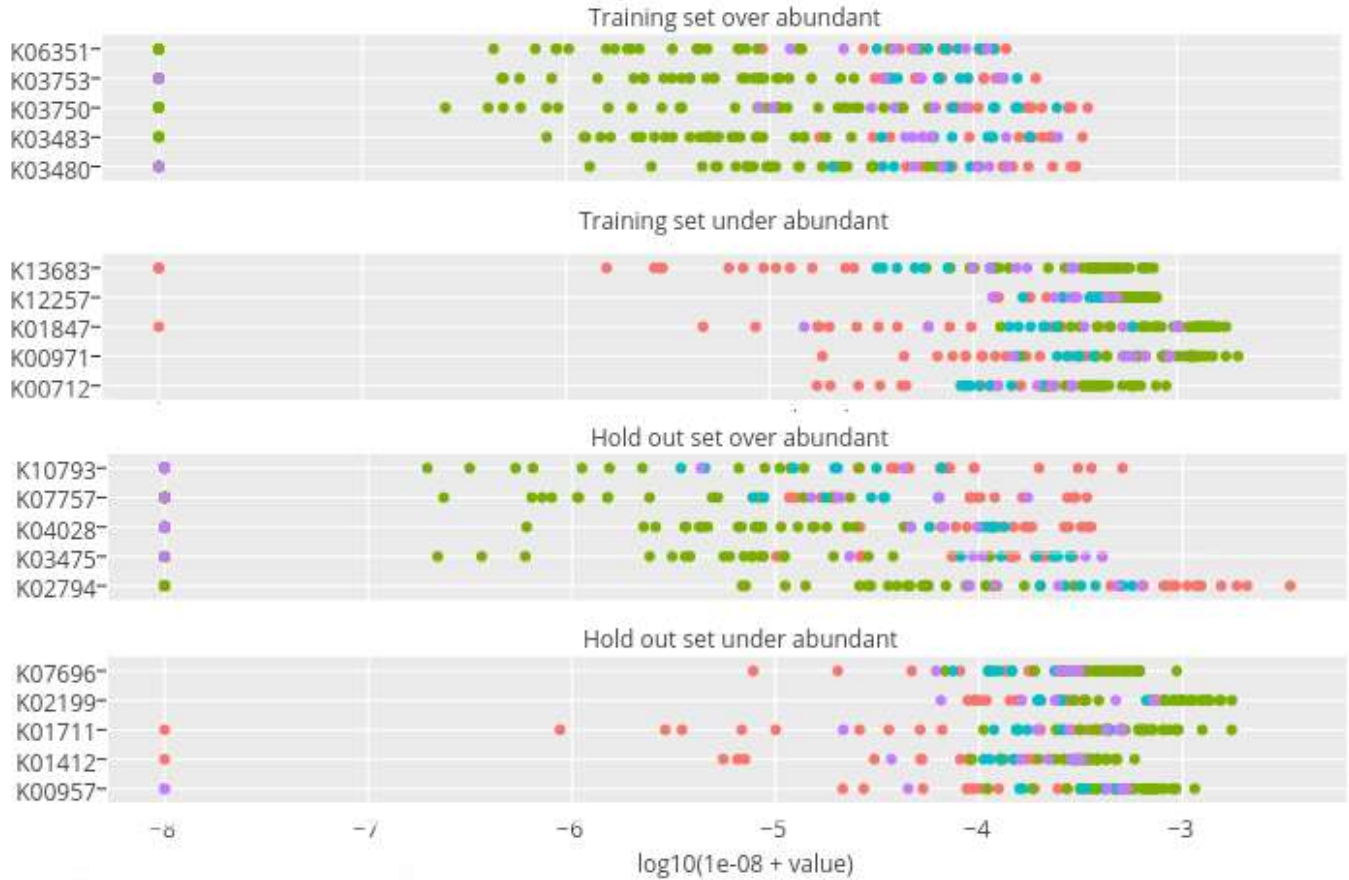


Source: Computing Weizhong Li, PCA Mehrdad Yazdani, Calit2

Examples of the Most Statistically Significant KEGGs That Differentiate Between the Disease and Healthy Cohorts

Selected from Top 100 KS Scores

Selected by Random Forest Classifier From Holdout Set

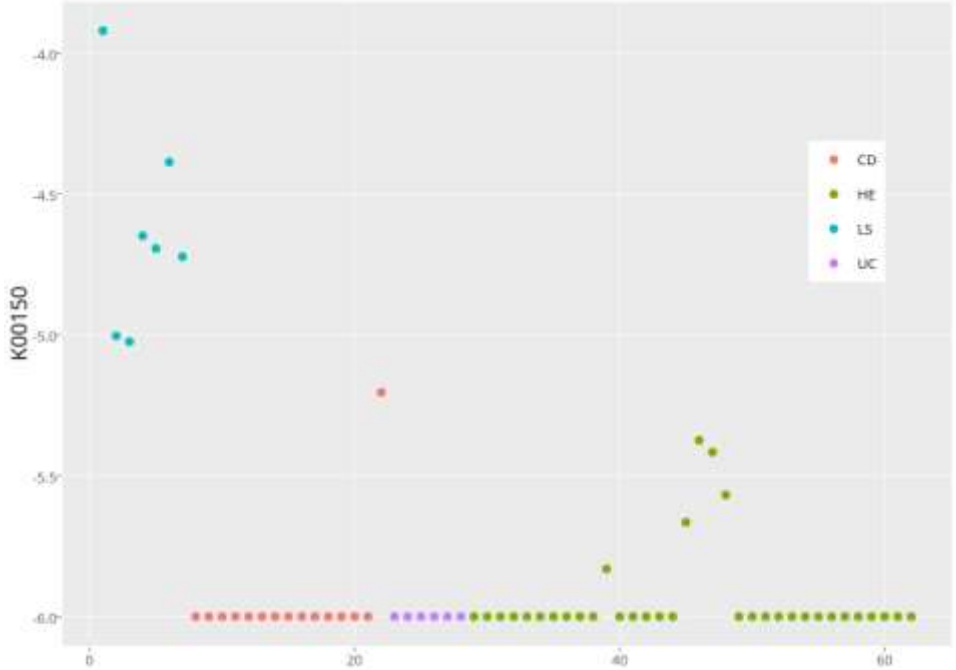


Note: Orders of Magnitude Increase or Decrease in Protein Families Between Health and Disease

Source: Computing Weizhong Li, PCA Mehrdad Yazdani, Calit2

So Which Protein Families Define My Disease State?

We Ran a Linear Classifier for Each of the 10,012 KEGGs
And Chose the Ones with the Lowest Error





Next Step: Investigate Biochemical Pathways of Key KEGGs

Source: Computing Weizhong Li, PCA Mehrdad Yazdani, Calit2

To Expand IBD Project the Knight/Smarr Labs Were Awarded ~ 1 CPU-Century Supercomputing Time



SDSC COMET

- **Smarr Gut Microbiome Time Series**
  - From 7 Samples Over 1.5 Years
  - To 75 Samples Over 5 Years
- **IBD Patients: From 5 Crohn's Disease and 2 Ulcerative Colitis Patients to ~100 Patients**
- **New Software Suite from Knight Lab**
  - Re-annotation of Reference Genomes, Functional / Taxonomic Variations
  - From 10,000 KEGGs to ~1 Million Genes
  - Novel Compute-Intensive Assembly Algorithms from Pavel Pevzner

8x Compute Resources Over Prior Study

KNIGHTLAB

XSEDE Extreme Science and Engineering Discovery Environment

it2

We are Genomically Analyzing My Stool Time Series in a Collaboration with the UCSD Knight Lab

LOCAL TOPICS | SCIENCE & ENVIRONMENT | SCIENCE

UCSD hires star biologist Rob Knight

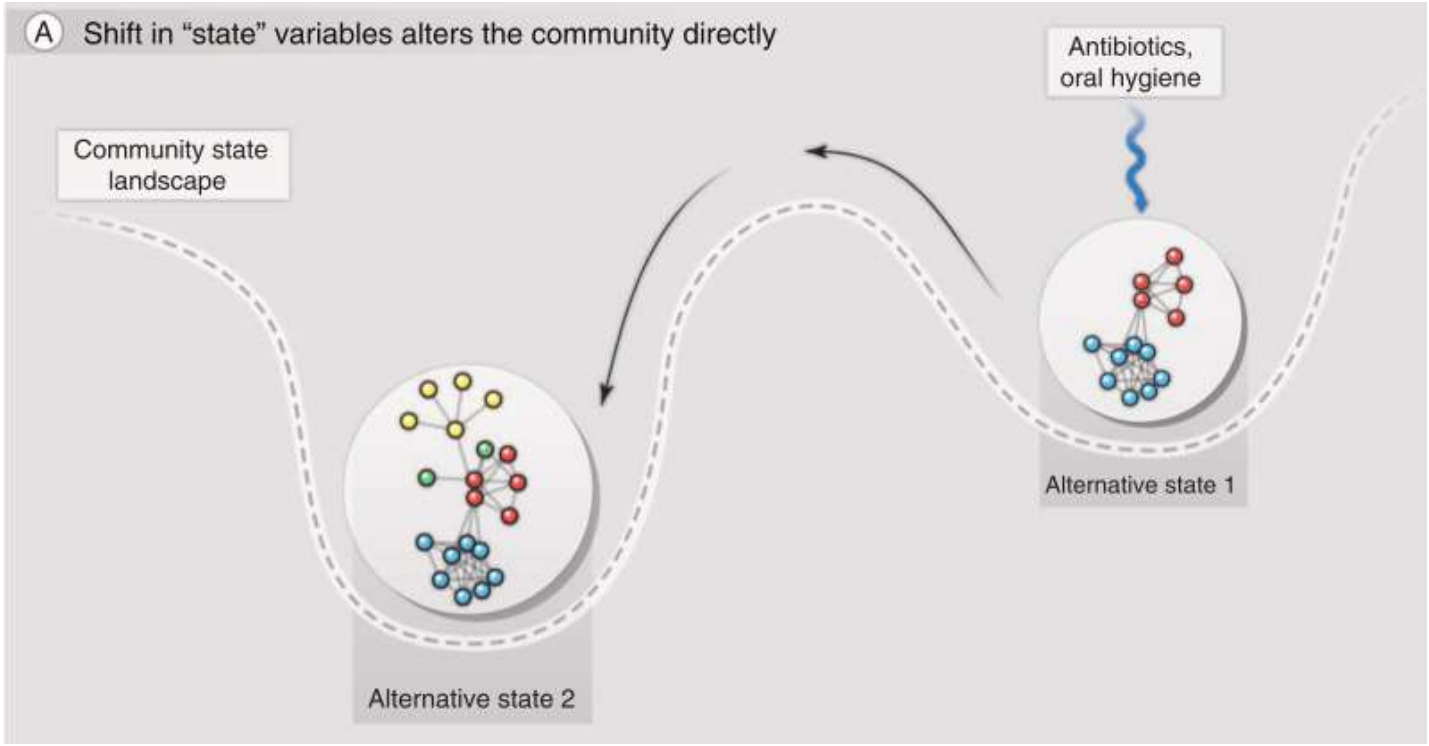
By Gary Robbins 1:30 P.M. JAN. 11, 2015 Updated 12:17 P.M. JAN. 12, 2015 PRINT COMMENTS 0





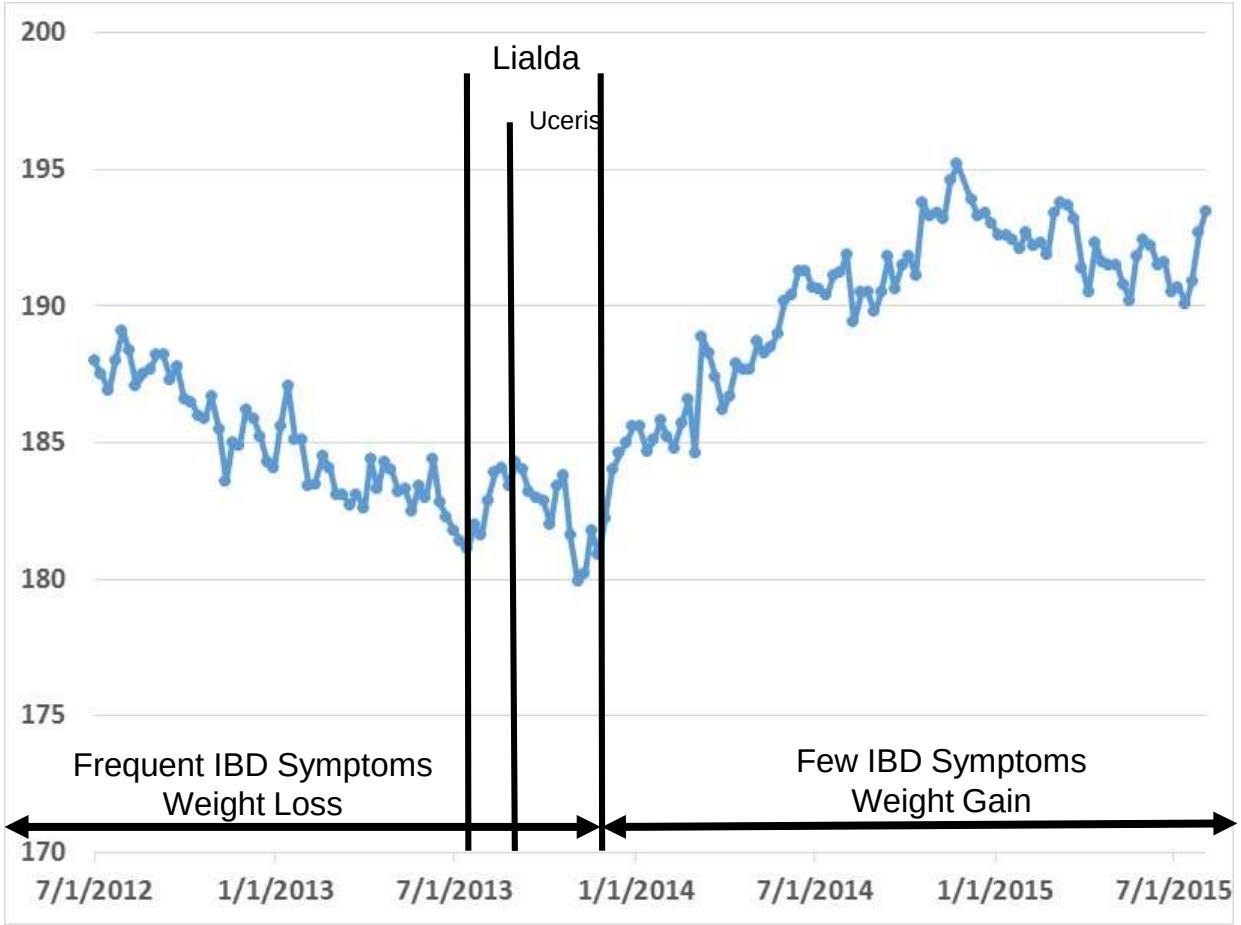
Larry's 40 Stool Samples Over 3.5 Years
to Rob's lab on April 30, 2015

Lessons from Ecological Dynamics: Gut Microbiome Has Multiple Relatively Stable Equilibria



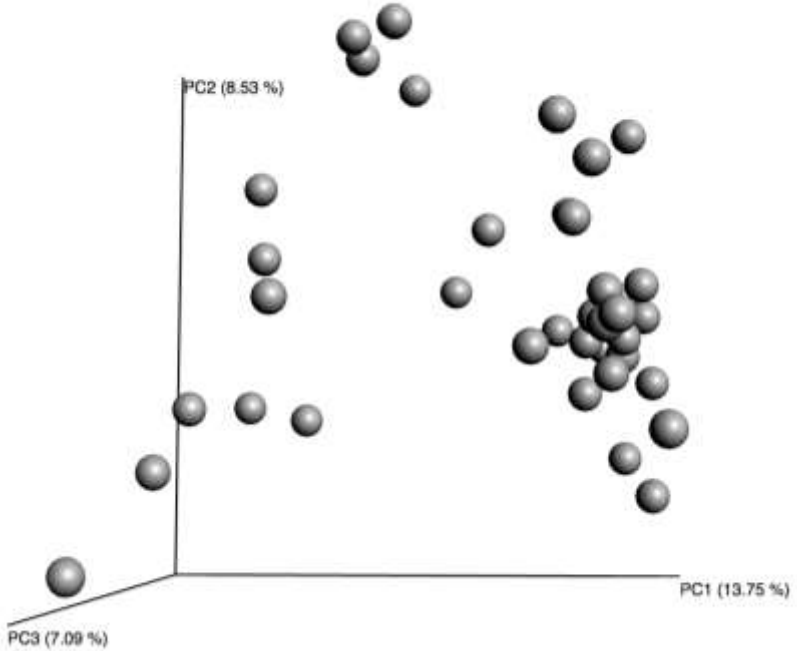
"The Application of Ecological Theory Toward an Understanding of the Human Microbiome," Elizabeth Costello, Keaton Stagaman, Les Dethlefsen, Brendan Bohannan, David Relman Science 336, 1255-62 (2012)

LS Weekly Weight During Period of 16S Microbiome Analysis
Abrupt Change in Weight and in Symptoms at January 1, 2014



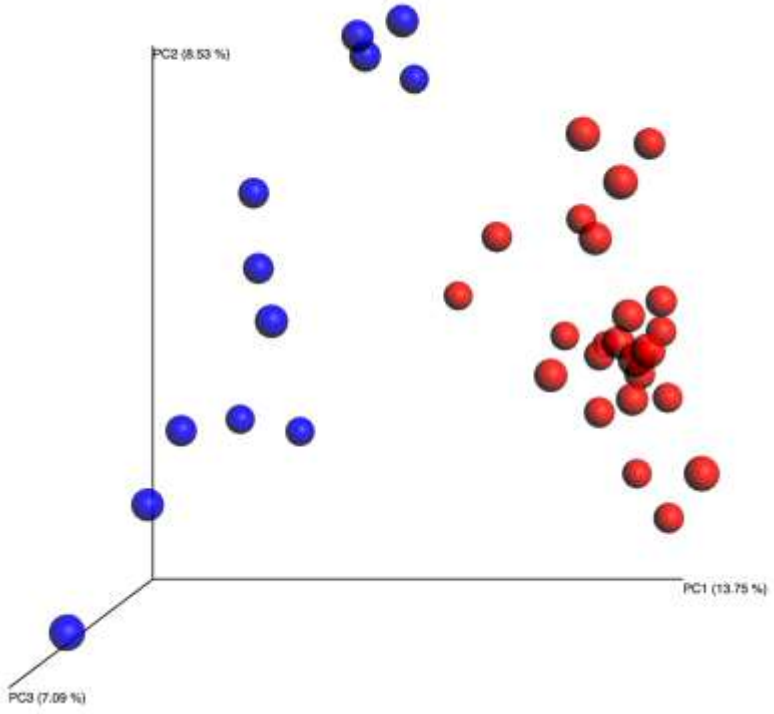
Source: Larry Smarr, UCSD

My Microbiome Ecology
Time Series Over 3 Years



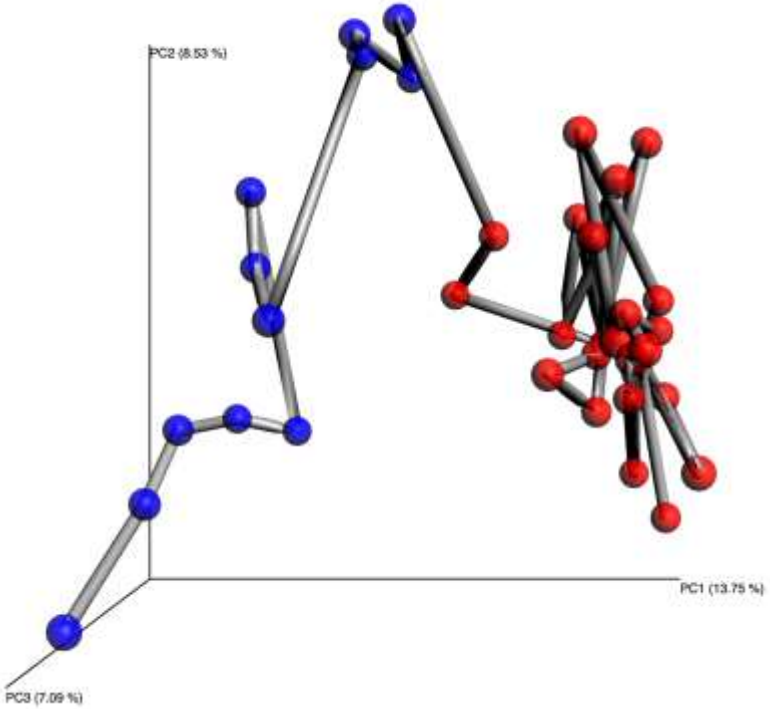
Source Justine Debelius, Knight Lab, UC San Diego

Coloring Samples Before (Blue) and After (Red) January 2014 Reveals Clustering



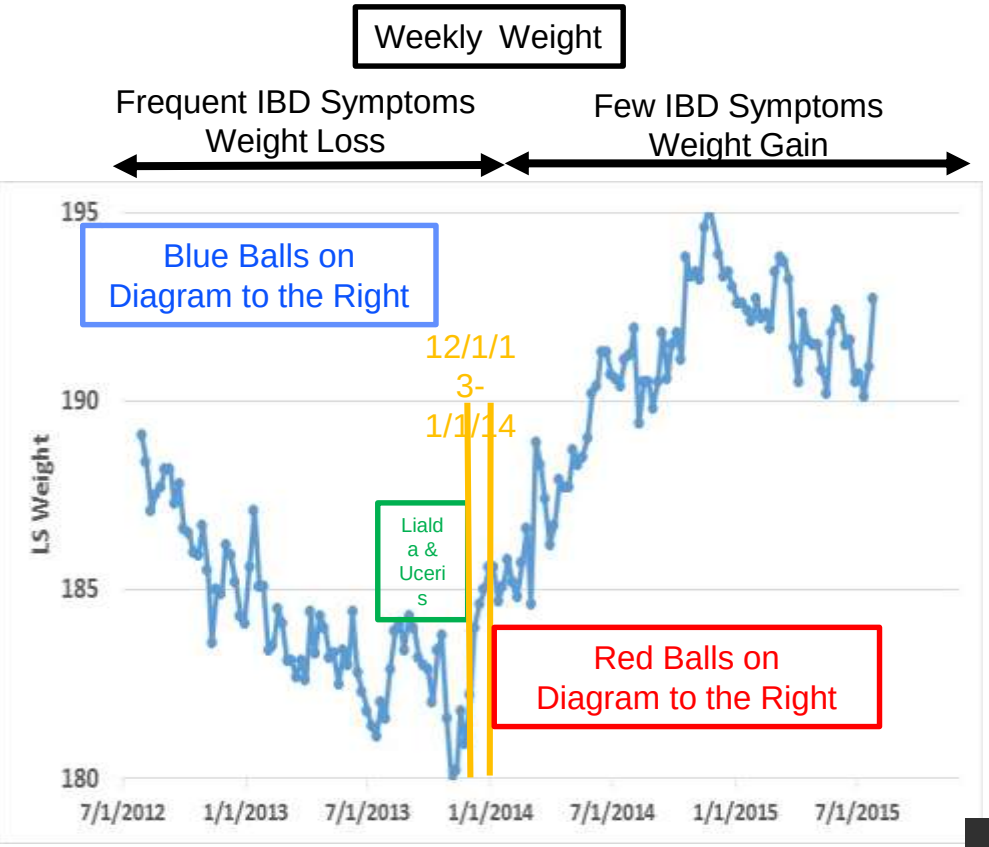
Source Justine Debelius, Knight Lab, UC San Diego

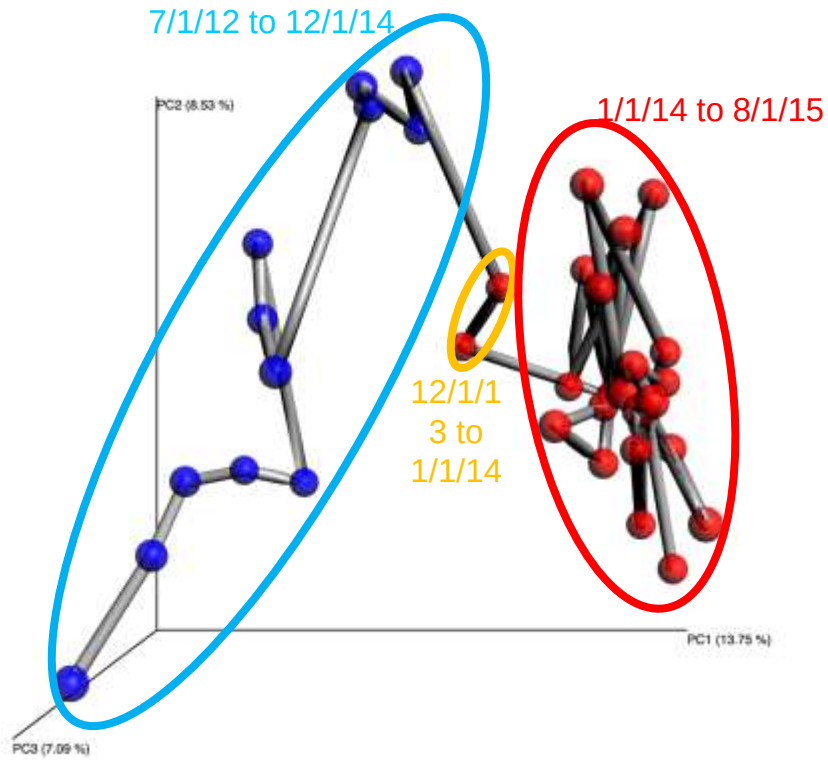
An Apparent Sudden Phase Change In the Microbiome Ecology Occurs



Source Justine Debelius, Knight Lab, UC San Diego

My Gut Microbiome Ecology Shifted After Drug Therapy Between Two Time-Stable Equilibriums Correlated to Physical Symptoms





Weight Data from Larry Smarr, Calit2, UCSD

PCoA by Justine Debelius and Jose Navas, Knight Lab, UCSD

What I Have Measured Is Rapidly Being Superseded to Include Deep Characterization of the Human Body

Precision Medicine Put to the Test: UnDx Consortium Gathers to Examine New Hypotheses for Undiagnosed Patients

GlobeNewswire • August 15, 2016

The five technology providers that have donated their services to the UnDx Consortium include:

- **Genome Profiling, LLC** of Wilmington, DE, providing epigenetic analysis;
- **The Lab of Rob Knight and the American Gut Project** at the University of California San Diego (UCSD), providing microbiome analysis;
- **KromaTiD, Inc.** of Fort Collins, CO, providing a chromosomal imaging platform for the detection of chromosomal rearrangements;
- **Metabolon, Inc.** of Durham, NC, providing metabolomic analysis; and
- **ORIG3N, Inc.** of Boston, MA, providing both the sample collection and stem cell analysis.

The Future Foundation of Medicine
is an Exponential Scaling-Up of the Number of Deeply Quantified Humans



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

Building a UC San Diego High Performance Cyberinfrastructure to Support Big Data Distributed Integrative Omics

NIH

NCBI
National Center for Biotechnology Information



Gordon

Knight 1024 Cluster
In SDSC Co-Lo





1.3Tbps

SDSC

PRP/
CENIC

CHERuB
100Gbps



Data Oasis
7.5PB,
200GB/s

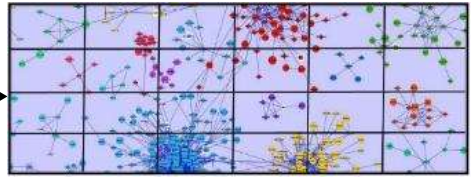
Knight Lab



ARISTA

120Gbps

Emperor & Other Vis Tools



64Mpixel Data Analysis Wall

10Gbps

40Gbps



Prism@UCSD

FIONA
12 Cores/GPU
128 GB RAM
3.5 TB SSD
48TB Disk
10Gbps NIC

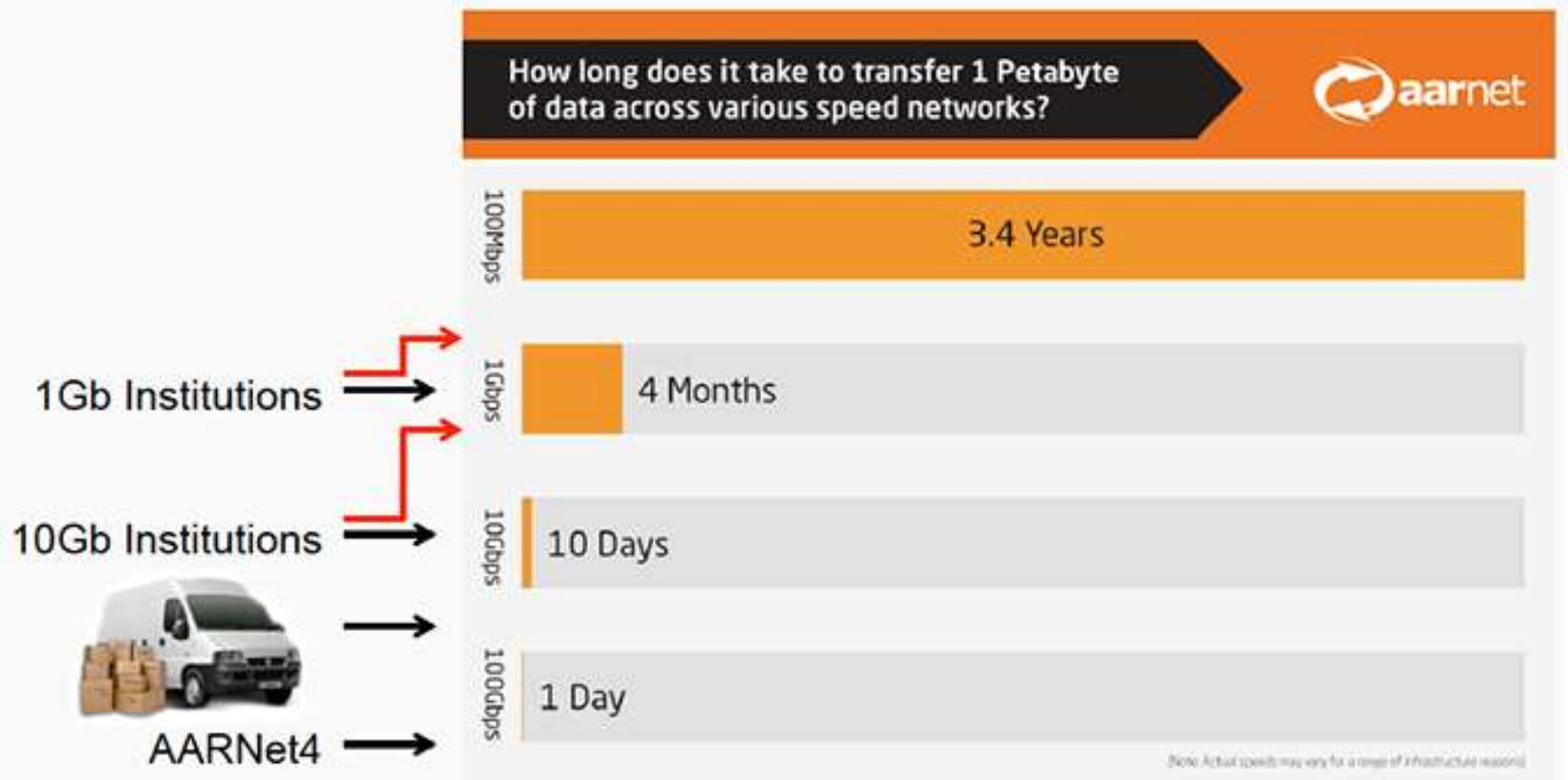


QUALCOMM INSTITUTE

IGM
Institute for Genomic Medicine

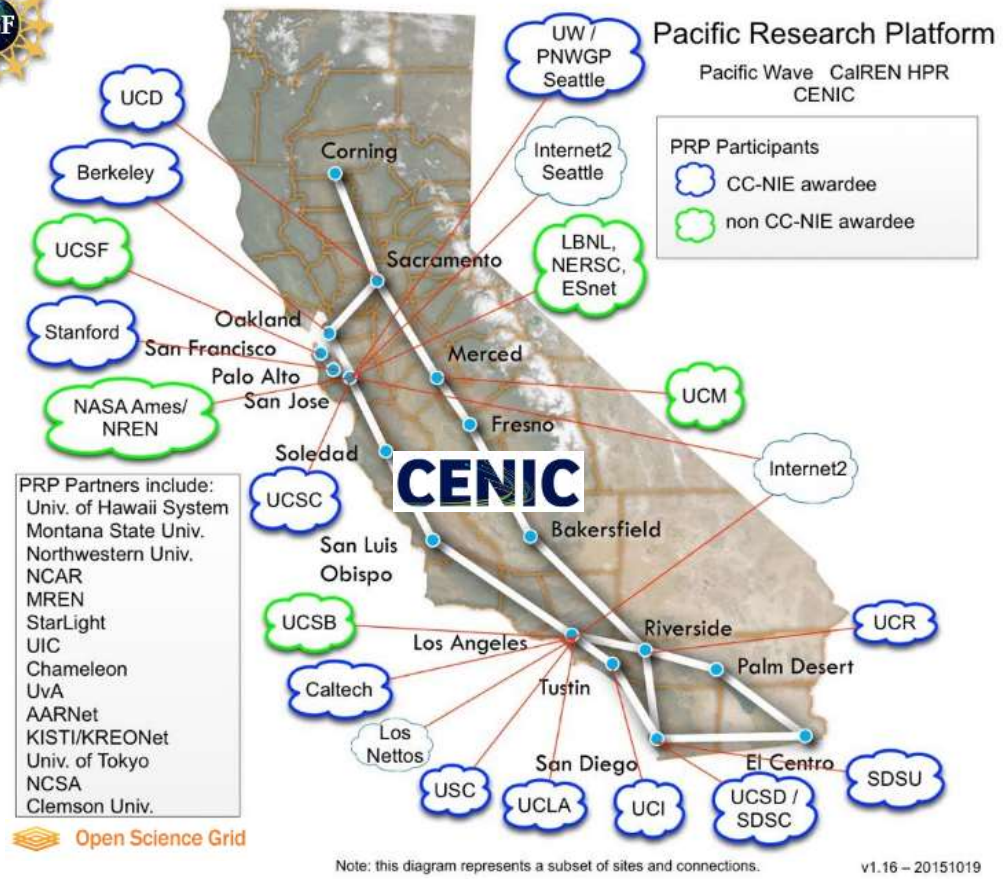
it²

Big Data Requires Big Bandwidth



http://news.aarnet.edu.au/data-movement-do-you-know-what-your-campus-network-is-actually-capable-of/

Next Step: The Pacific Research Platform Creates a Regional End-to-End Science-Driven "Big Data Freeway System"



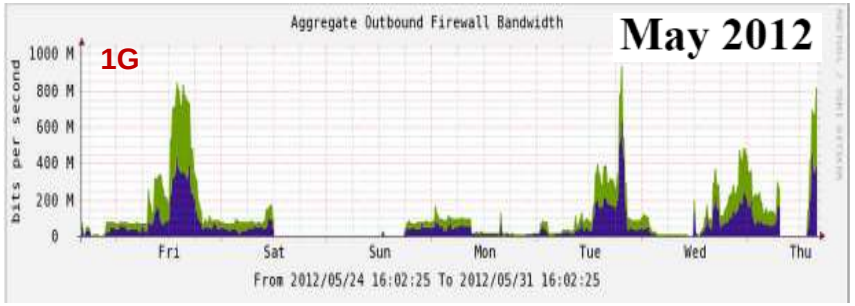
NSF CC*DNI Grant
$5M 10/2015-10/2020

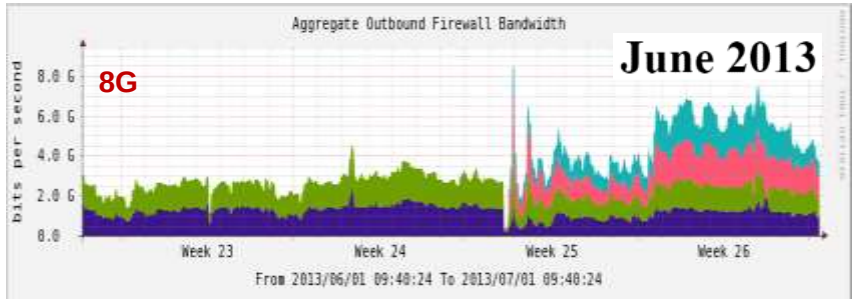
PI: Larry Smarr, UC San Diego Calit2
Co-Pis:

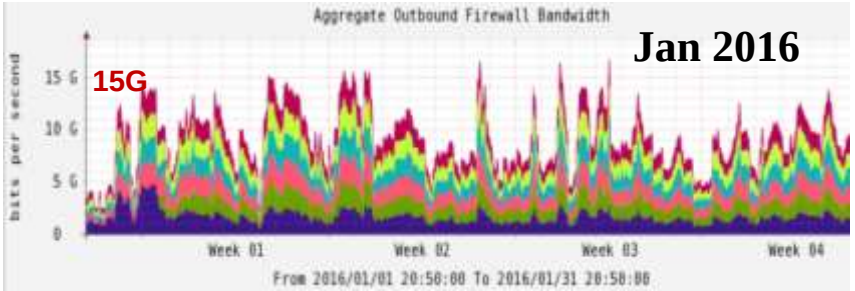
- **Camille Crittenden, UC Berkeley CITRIS,**
- **Tom DeFanti, UC San Diego Calit2,**
- **Philip Papadopoulos, UC San Diego SDSC,**
- **Frank Wuerthwein, UC San Diego Physics and SDSC**

Source:
John Hess, CENIC

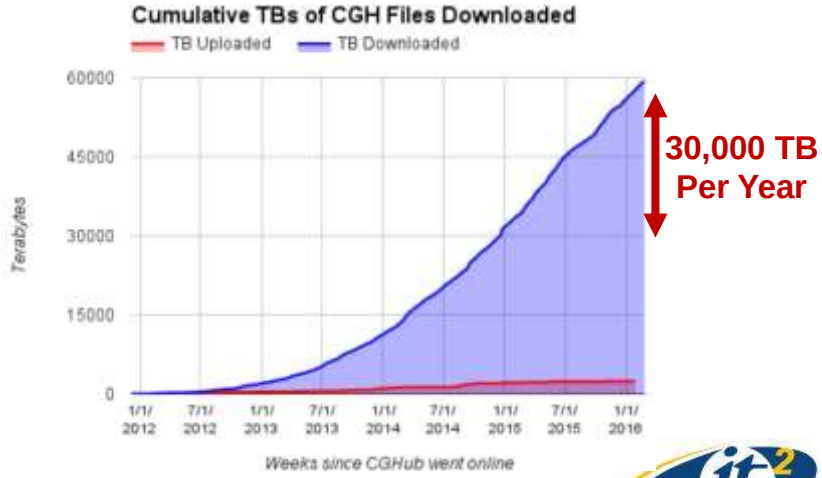
Cancer Genomics Hub (UCSC) is Housed in SDSC: Large Data Flows to End Users at UCSC, UCB, UCSF, …











Data Source: David Haussler, Brad Smith, UCSC

The Future of Supercomputing
Will Need More Than von Neumann Processors

“High Performance Computing Will Evolve Towards a Hybrid Model, Integrating Emerging Non-von Neumann Architectures, with Huge Potential in Pattern Recognition, Streaming Data Analysis, and Unpredictable New Applications.”



Horst Simon, Deputy Director,
U.S. Department of Energy’s
Lawrence Berkeley National Laboratory

U.S. DEPARTMENT OF ENERGY

BERKELEY LAB
Bringing Science Solutions to the World

Qualcomm Institute

it²

Calit2's Qualcomm Institute Has Established a Pattern Recognition Lab On the PRP, For Machine Learning on non-von Neumann Processors



UCSD ECE Professor Ken Kreutz-Delgado Brings the IBM TrueNorth Chip to Start Calit2's Qualcomm Institute Pattern Recognition Laboratory September 16, 2015

'It's only limited by money, not imagination,' Modha says."

Source: Dr. Dharmendra Modha
Founding Director, IBM Cognitive Computing Group

Dan Goldin Announced His Company KnuEdge June 6, 2016 - He Will Provide Chip to PRL This Year



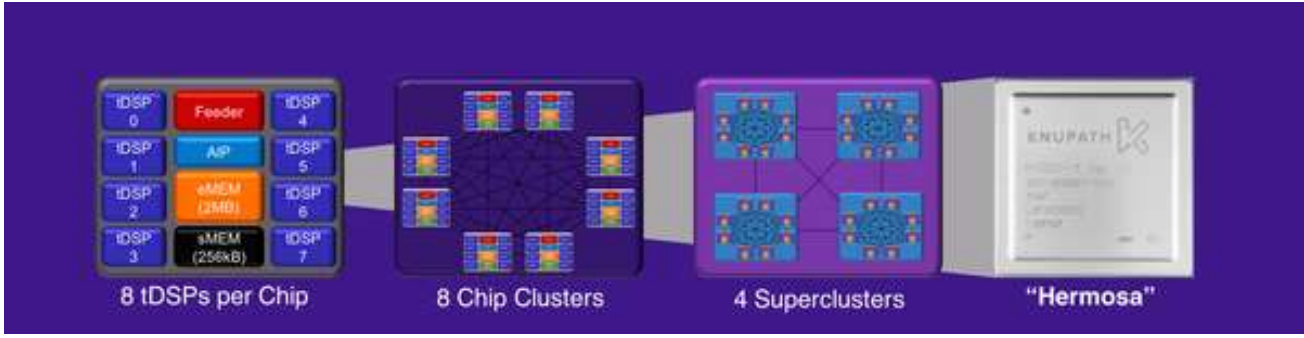
The KNUPATH Hermosa's super-scalable architecture is based on neurological design. It contains 256 digital signal processing cores (tDSPs) and 64 programmable DMA engines. Hermosa's performance-advantaged architecture "can deliver acceleration of targeted workloads far beyond legacy architectures," according to the company. [Photos courtesy KnuEdge, Inc.]

www.calit2.net/newsroom/release.php?id=2704

tom's HARDWARE
THE AUTHORITY ON TECH

KnuEdge Announces 'KnuVerse' Voice Authentication Technology And 'KnuPath' Neural Chip

By Lucian Armasu JUNE 6, 2016



www.tomshardware.com/news/knuedge-announces-knuverse-and-knupath,31981.html

KNUEDGE

Our Pattern Recognition Lab is Exploring Mapping Machine Learning Algorithm Families Onto Novel Architectures

- Deep & Recurrent Neural Networks (DNN, RNN)
- Graph Theoretic
- Reinforcement Learning (RL)
- Clustering and other neighborhood-based
- Support Vector Machine (SVM)
- Sparse Signal Processing and Source Localization
- Dimensionality Reduction & Manifold Learning
- Latent Variable Analysis (PCA, ICA)
- Stochastic Sampling, Variational Approximation
- Decision Tree Learning

Qualcomm Institute

Large Corporations
Are Already Using Non Specialized Accelerators

- **Microsoft Installs FPGAs into Bing Servers**



www.microsoft.com/en-us/research/project/project-catapult/

Google supercharges machine learning tasks with TPU custom chip

Wednesday, May 18, 2016

https://cloudplatform.googleblog.com/2016/05/Google-supercharges-machine-learning-tasks-with-custom-chip.html

Thanks to Our Great Team!

Calit2@UCSD
Future Patient Team
Jerry Sheehan
Tom DeFanti
- Joe Keefe
- John Graham

Kevin Patrick
Mehrdad Yazdani
Jurgen Schulze
- Andrew Prudhomme
- Philip Weber
- Fred Raab

Ernesto Ramirez



JCVI Team
Karen Nelson
- Shibu Yooseph
- Manolito Torralba

J. Craig Venter INSTITUTE

Ayasdi
Devi Ramanan
Pek Lum

Ayasdi Cure™
Turn Data into Therapies

UCSD Metagenomics Team
Weizhong Li
- Sitao Wu

UC San Diego

CRBS

SDSC Team
Michael Norman
- Mahidhar Tatineni
- Robert Sinkovits
- Ilkay Altintas

SDSC

Dell/R Systems
Brian Kucic
John Thompson
- Thomas Hill

RSYSTEMS powering innovation

DELL

UCSD Health Sciences Team
David Brenner
Rob Knight Lab
- Justine Debelius
- Jose Navas
- Bryn Taylor
- Gail Ackermann
- Greg Humphrey

William J. Sandborn Lab
- Elisabeth Evans
- John Chang
- Brigid Boland

UC San Diego HEALTH SCIENCES

it²