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

Invited Talk

Ayasdi

Menlo Park, CA

December 5, 2014

Dr. Larry Smarr

Director, California Institute for Telecommunications and Information Technology

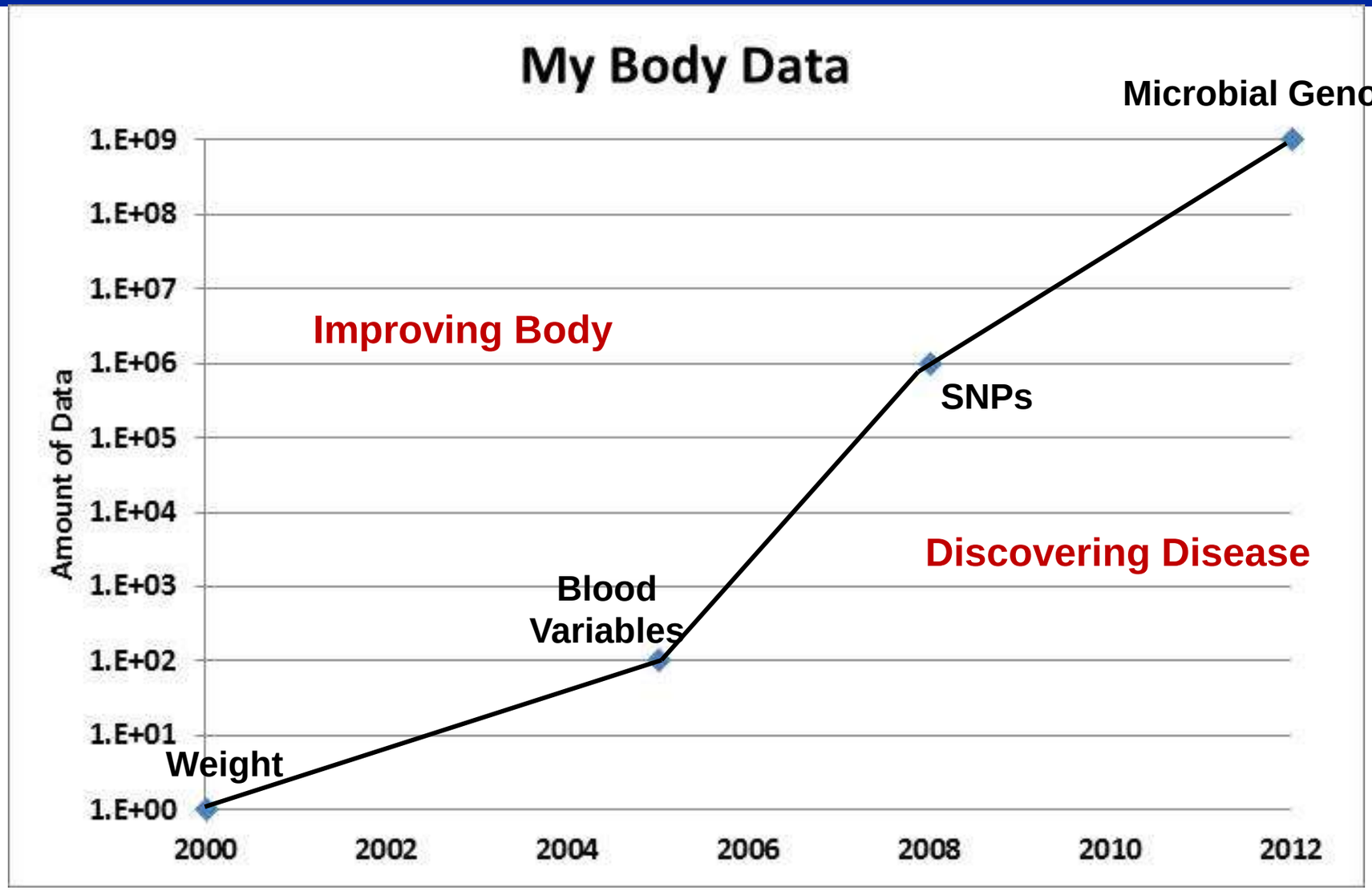
Harry E. Gruber Professor,

Dept. of Computer Science and Engineering

Jacobs School of Engineering, UCSD

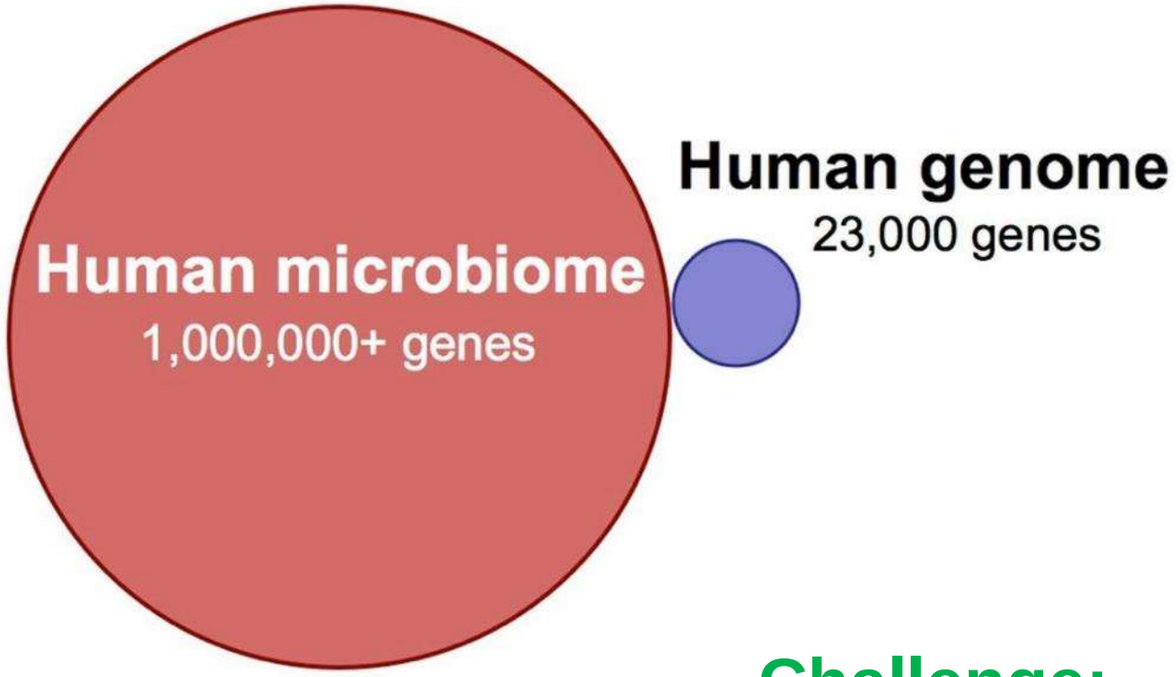
http://lsmarr.calit2.net

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



How Will Detailed Knowledge of Microbiome Ecology Radically Change Medicine and Wellness?

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



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

Challenge:
Map Out Microbial Ecology and Function in Health and Disease States

Intense Scientific Research is Underway on Understanding the Human Microbiome



June 8, 2012



June 14, 2012



August 18, 2012

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

Example: Inflammatory Bowel Disease (IBD)

- **Metagenomic Sequencing**
  - **JCVI Produced**
    - **~150 Billion DNA Bases From Seven of LS Stool Samples Over 1.5 Years**
  - **We Downloaded ~3 Trillion DNA Bases From NIH Human Microbiome Program Data Base**
    - **255 Healthy People, 21 with IBD**
- **Supercomputing** (Weizhong Li, JCVI/HLI/UCSD):
  - **~20 CPU-Years on SDSC's Gordon**
  - **~4 CPU-Years on Dell's HPC Cloud**
- **Produced Relative Abundance of**
  - **~10,000 Bacteria, Archaea, Viruses in ~300 People**
  - **~3Million Filled Spreadsheet Cells**

Illumina HiSeq 2000 at JCVI



SDSC Gordon Data Supercomputer



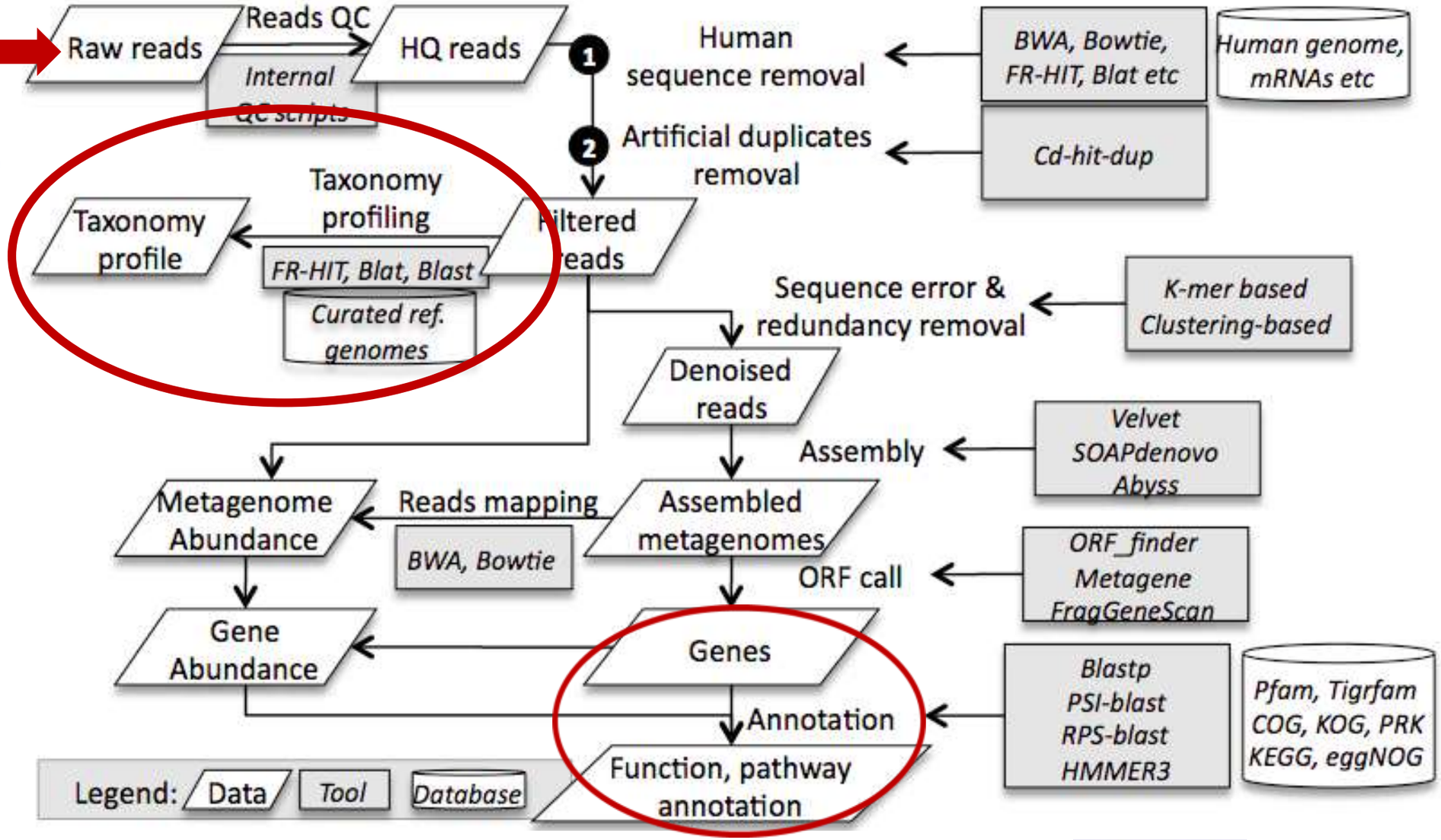
Dell Solutions Center
Industry Solutions Lab
SANGER DSU

J. Craig Venter INSTITUTE SDSC DELL NIH HMP CRBS it²

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)

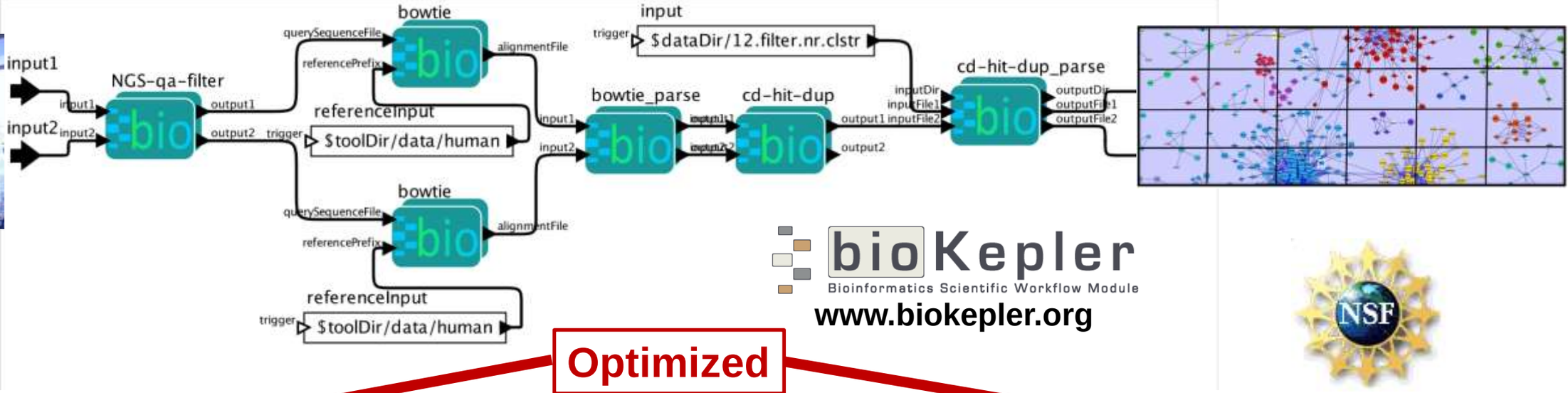
CRBS

it²

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

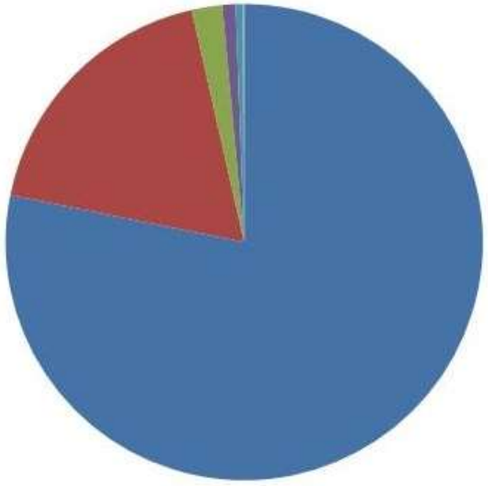
it²

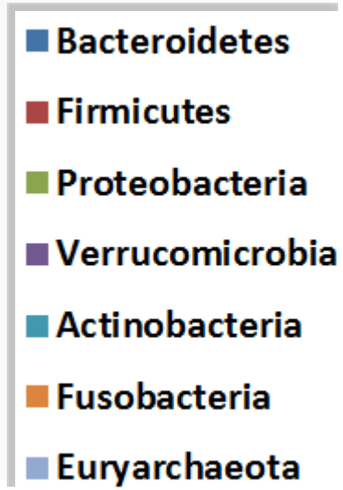
How Best to Analyze The Microbiome Datasets to Discover Patterns in Health and Disease?

Can We Find New Noninvasive Diagnostics In Microbiome Ecologies?

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

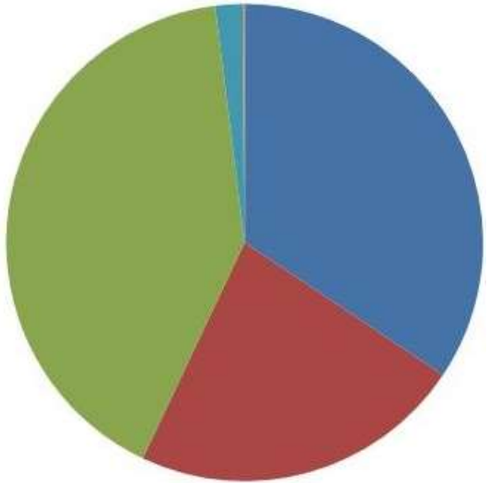
Average HE





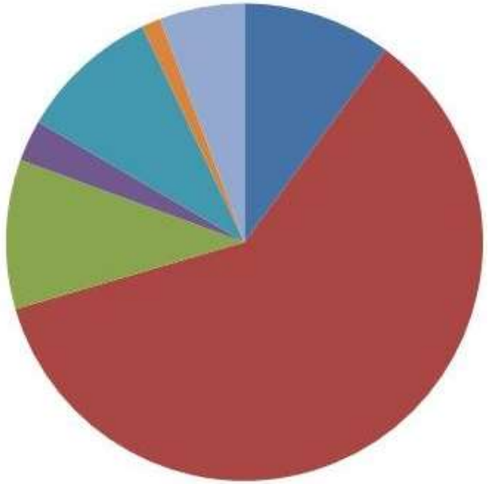
Most Common Microbial Phyla

Average Ulcerative Colitis



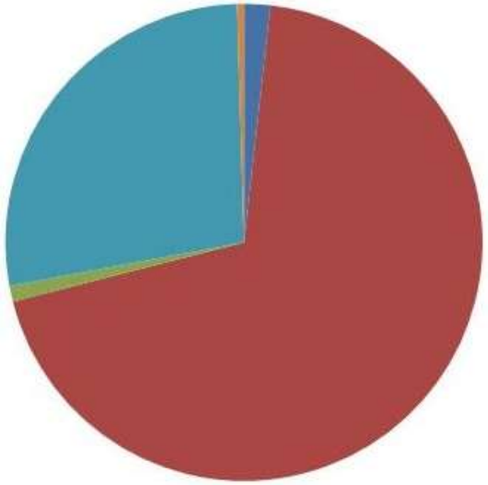
Explosion of Proteobacteria

Average LS



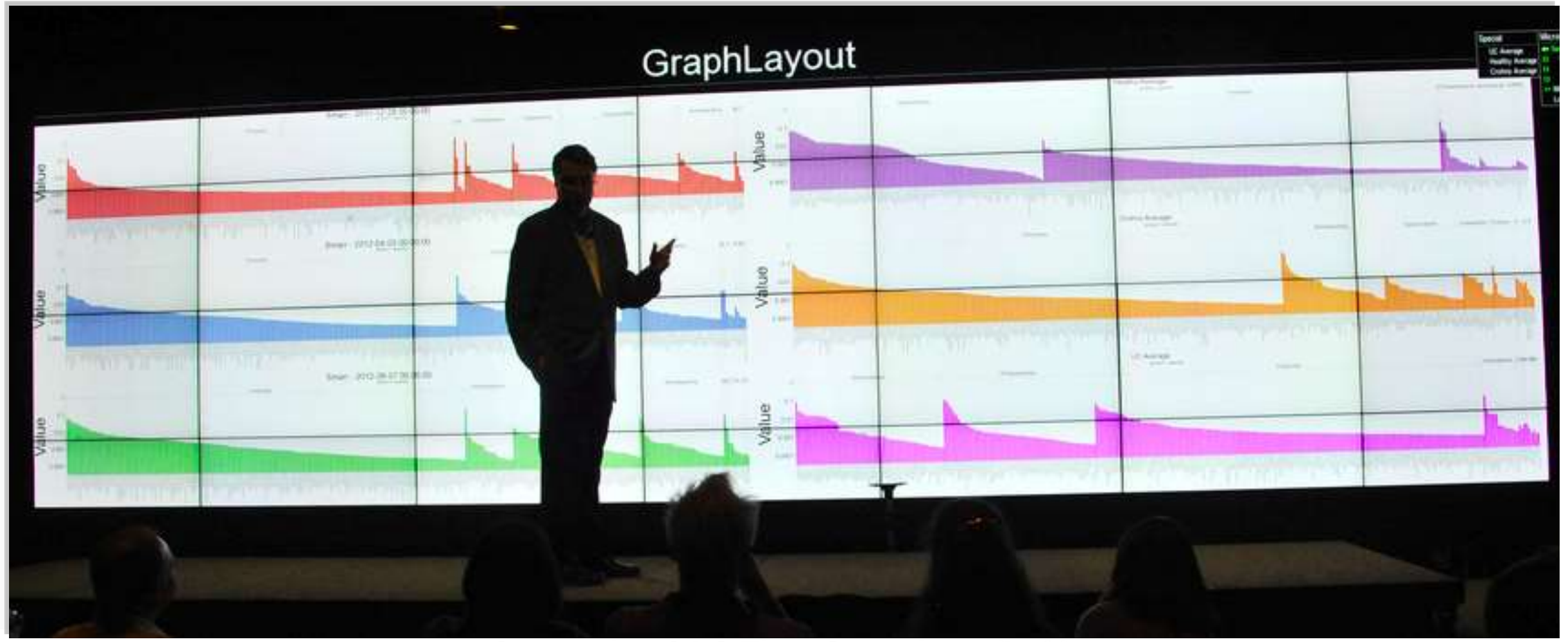
Hybrid of UC and CD
High Level of Archaea

Average Crohn's Disease



Collapse of Bacteroidetes
Explosion of Actinobacteria

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



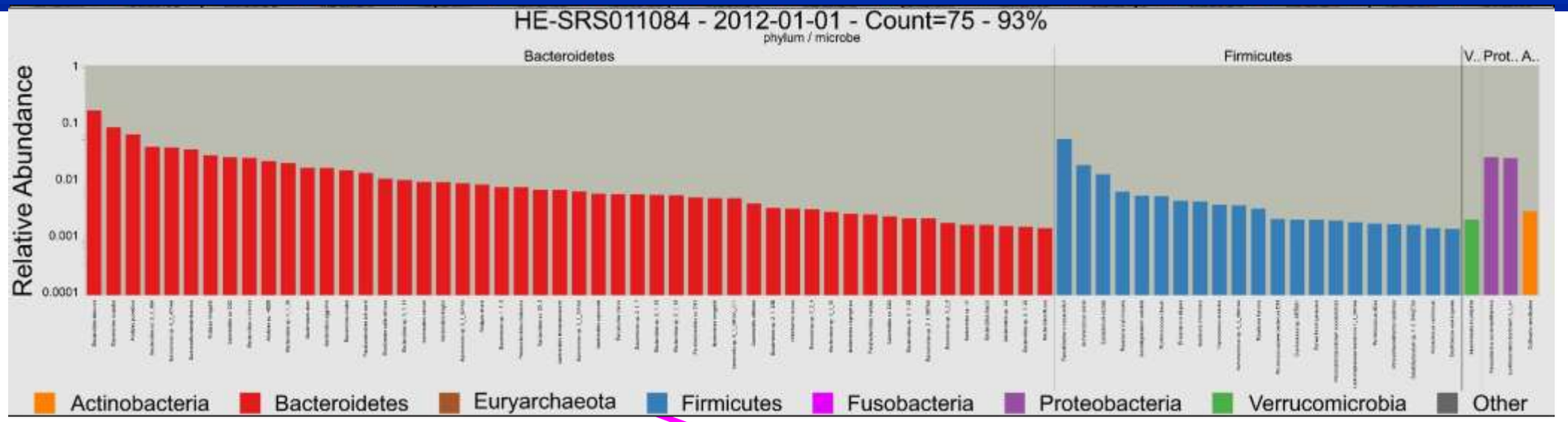
Comparing 3 LS Time Snapshots (Left)
with Healthy, Crohn's, Ulcerative Colitis (Right Top to Bottom)

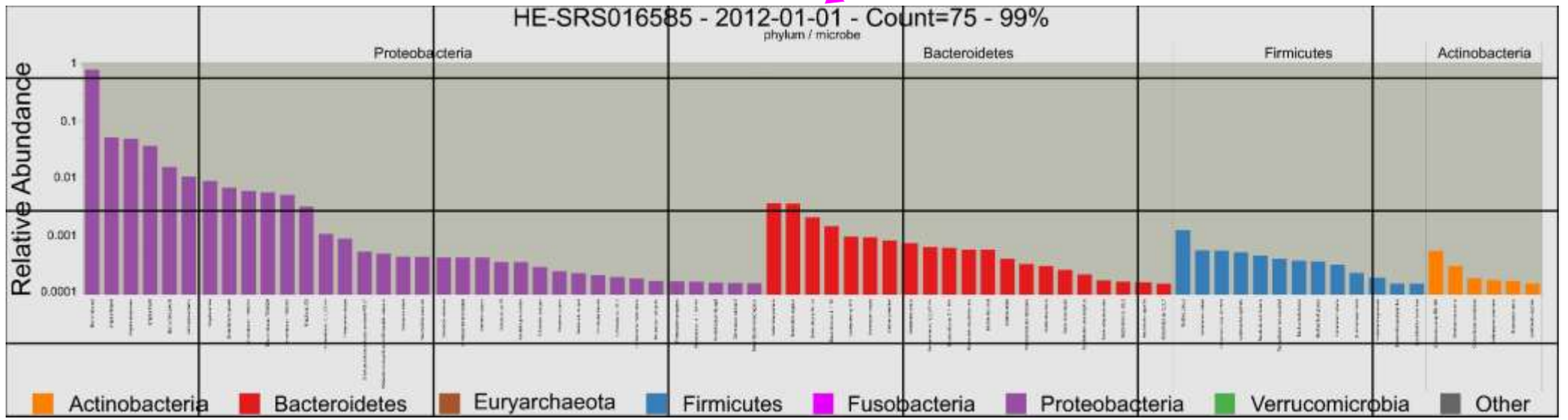
Calit2 VROOM-FuturePatient Expedition

Using Dell HPC Cloud and Dell Analytics to Discover Microbial Diagnostics for Disease Dynamics

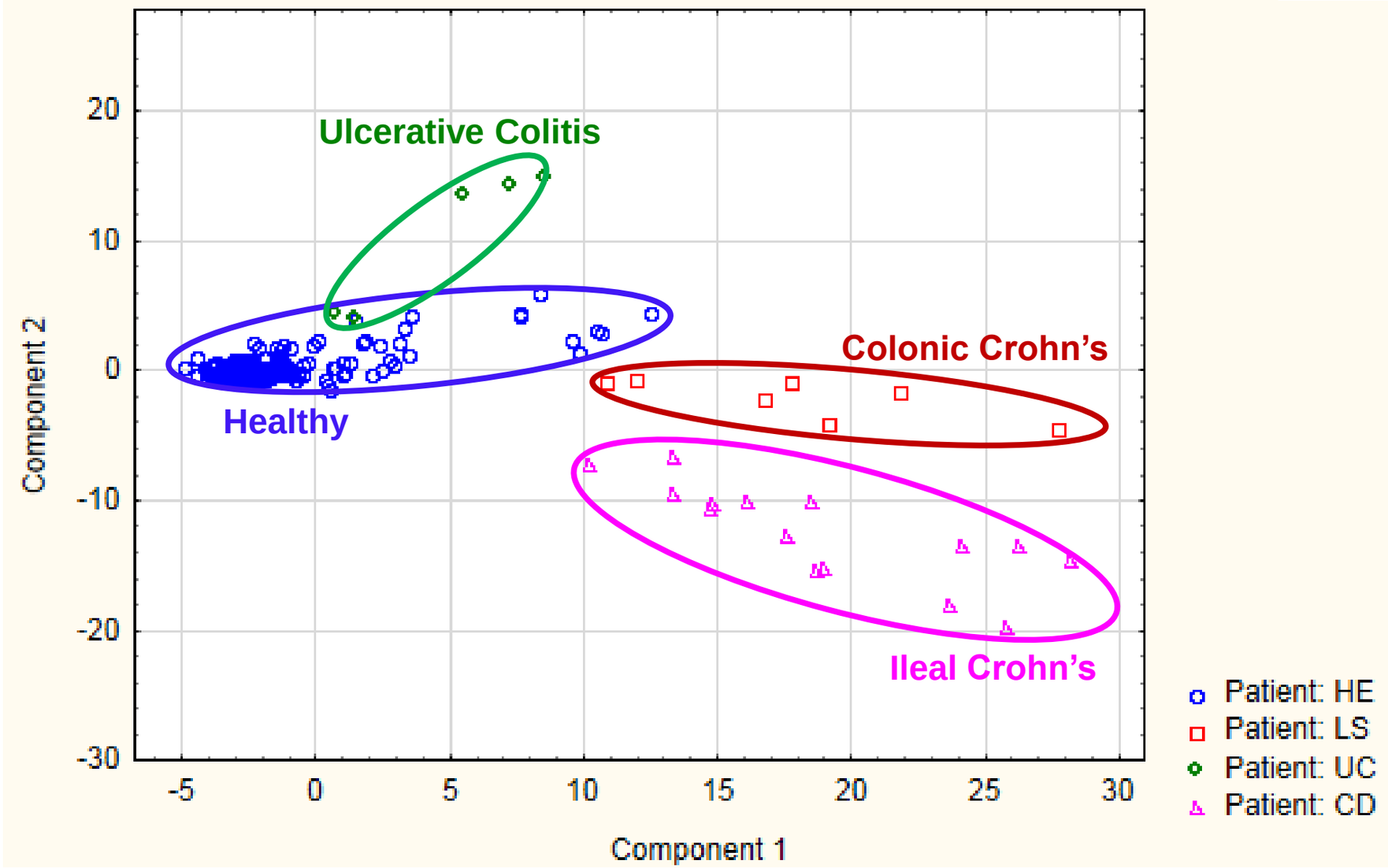
- **Can We Distinguish Noninvasively Between Health and Disease States?**
- **Are There Subsets of Health or Disease States?**
- **Can We Track Time Development of the Disease State?**
- **Can Novel Microbial Diagnostics Differentiate Health and Disease States?**

Using Microbiome Profiles to Survey 155 Subjects for Unhealthy Candidates



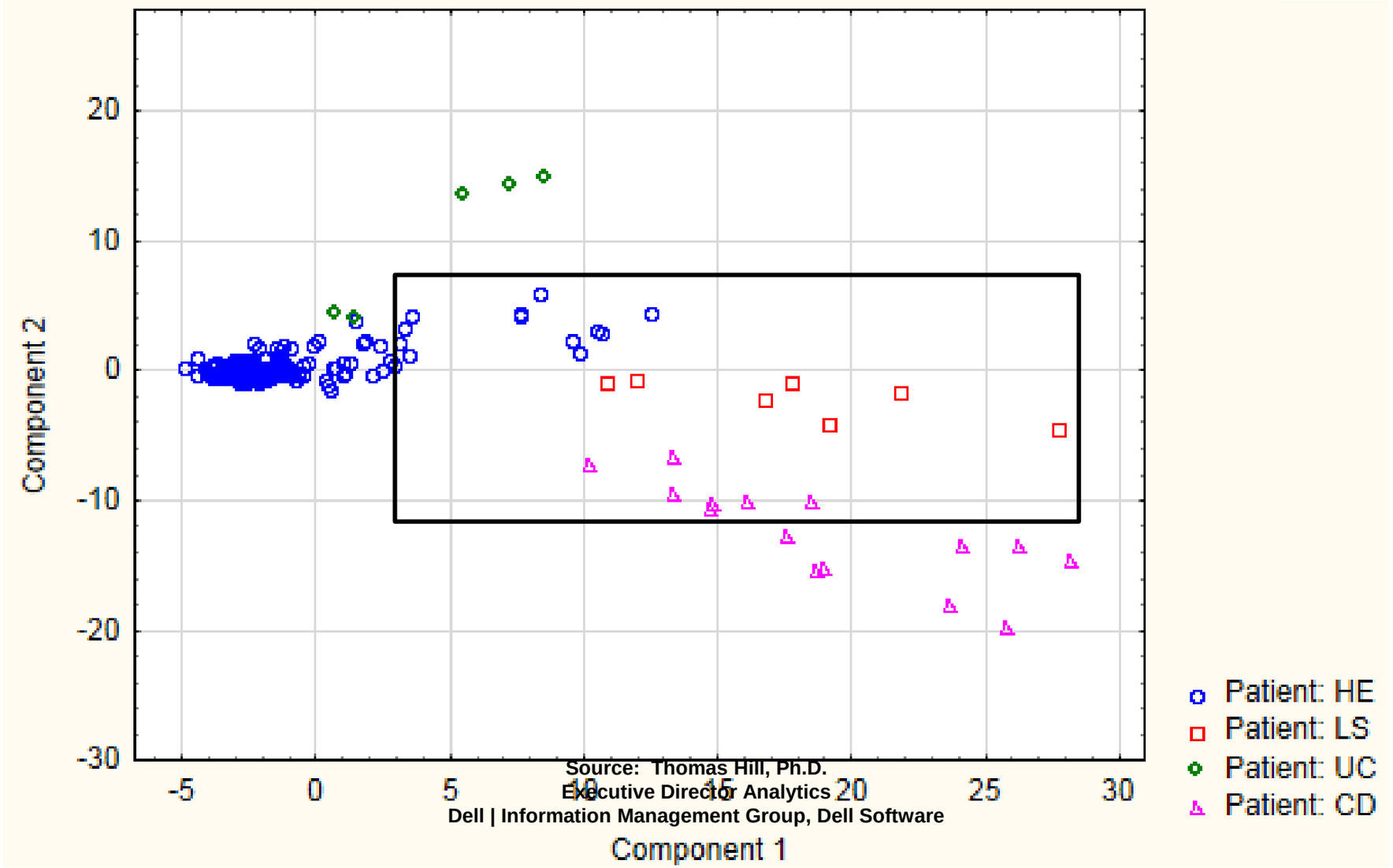


Dell Analytics Separates The 4 Patient Types in Our Using Our Microbiome Species Data



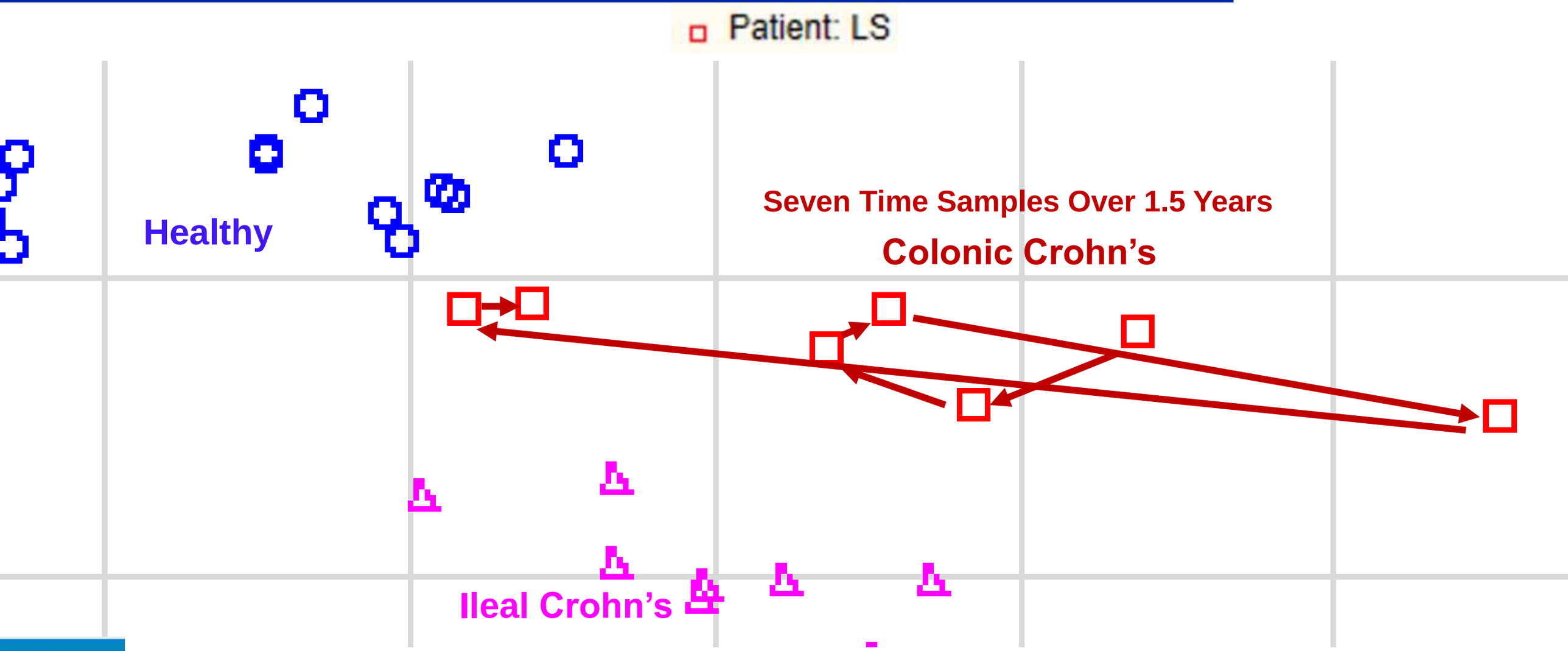
Source: Thomas Hill, Ph.D.
Executive Director Analytics
Dell | Information Management Group, Dell Software

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

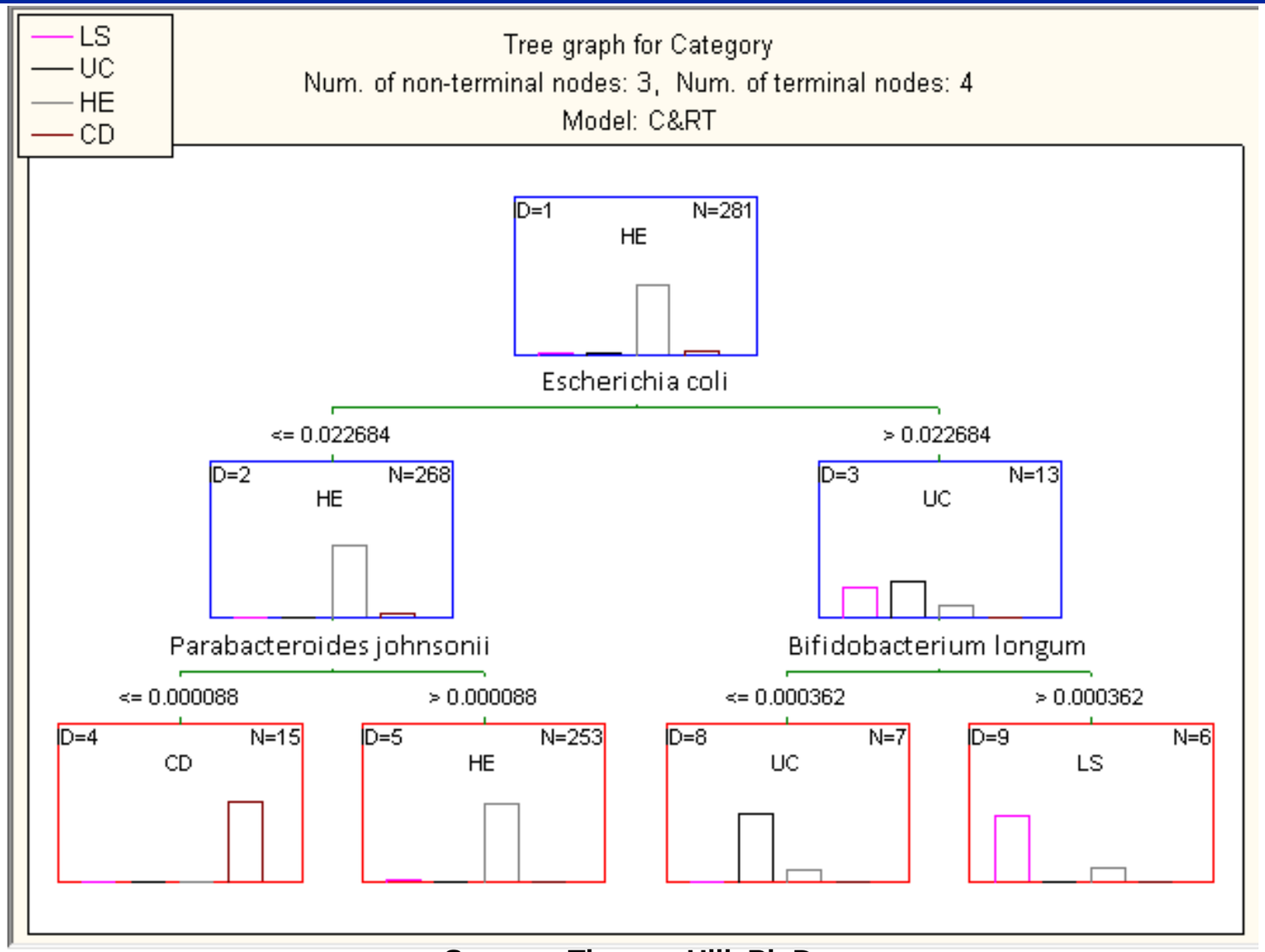


Source: Thomas Hill, Ph.D.
Executive Director Analytics
Dell | Information Management Group, Dell Software

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

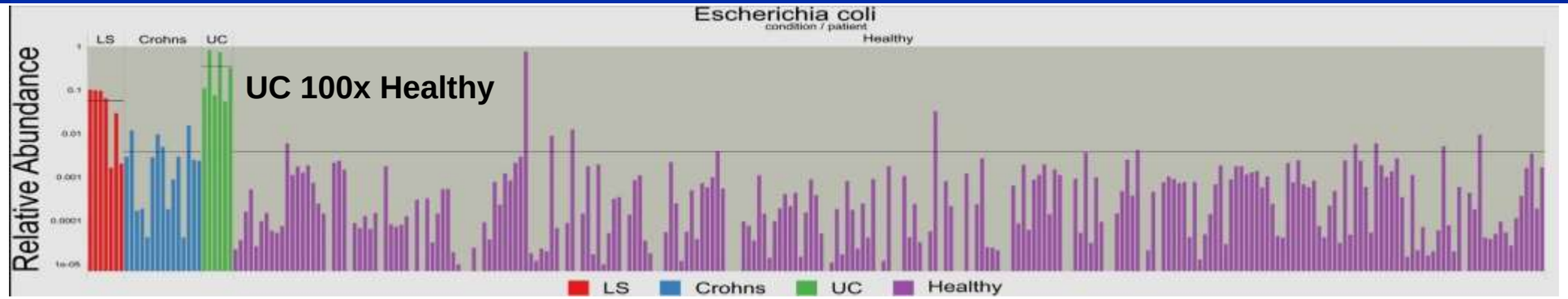


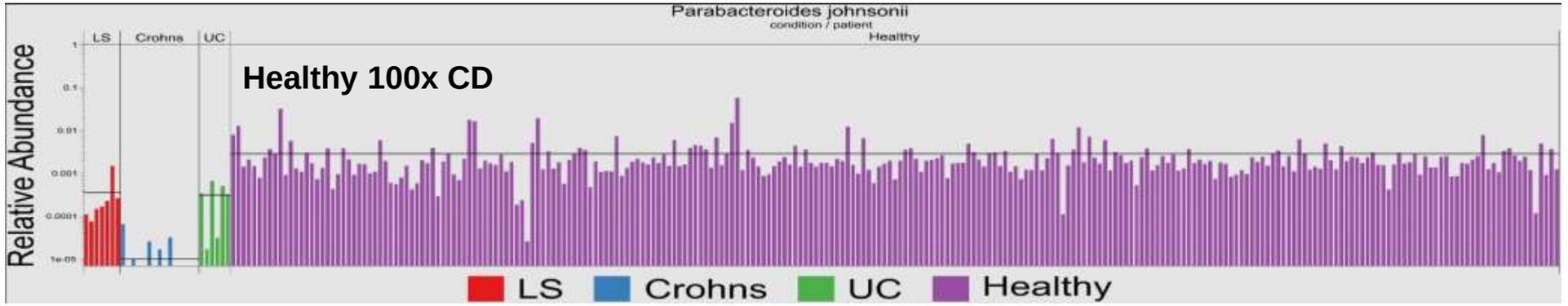
Dell Analytics Tree Graphs Classifies the 4 Health/Disease States With Just 3 Microbe Species

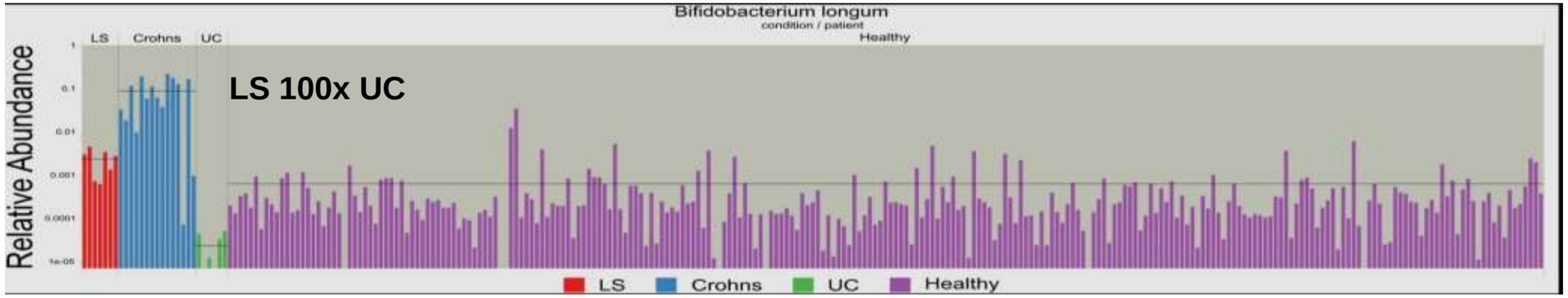


Source: Thomas Hill, Ph.D.
Executive Director Analytics
Dell | Information Management Group, Dell Software

Our Relative Abundance Results Across ~300 People Show Why Dell Analytics Tree Classifier Works







We Produced Similar Results for ~2500 Microbial Species

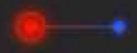
Using Ayasdi's Advanced Topological Data Analysis to Separate Healthy from Disease States

Using Ayasdi Categorical Data Lens

All Ileal Crohn's



All Healthy



All Healthy





All Healthy



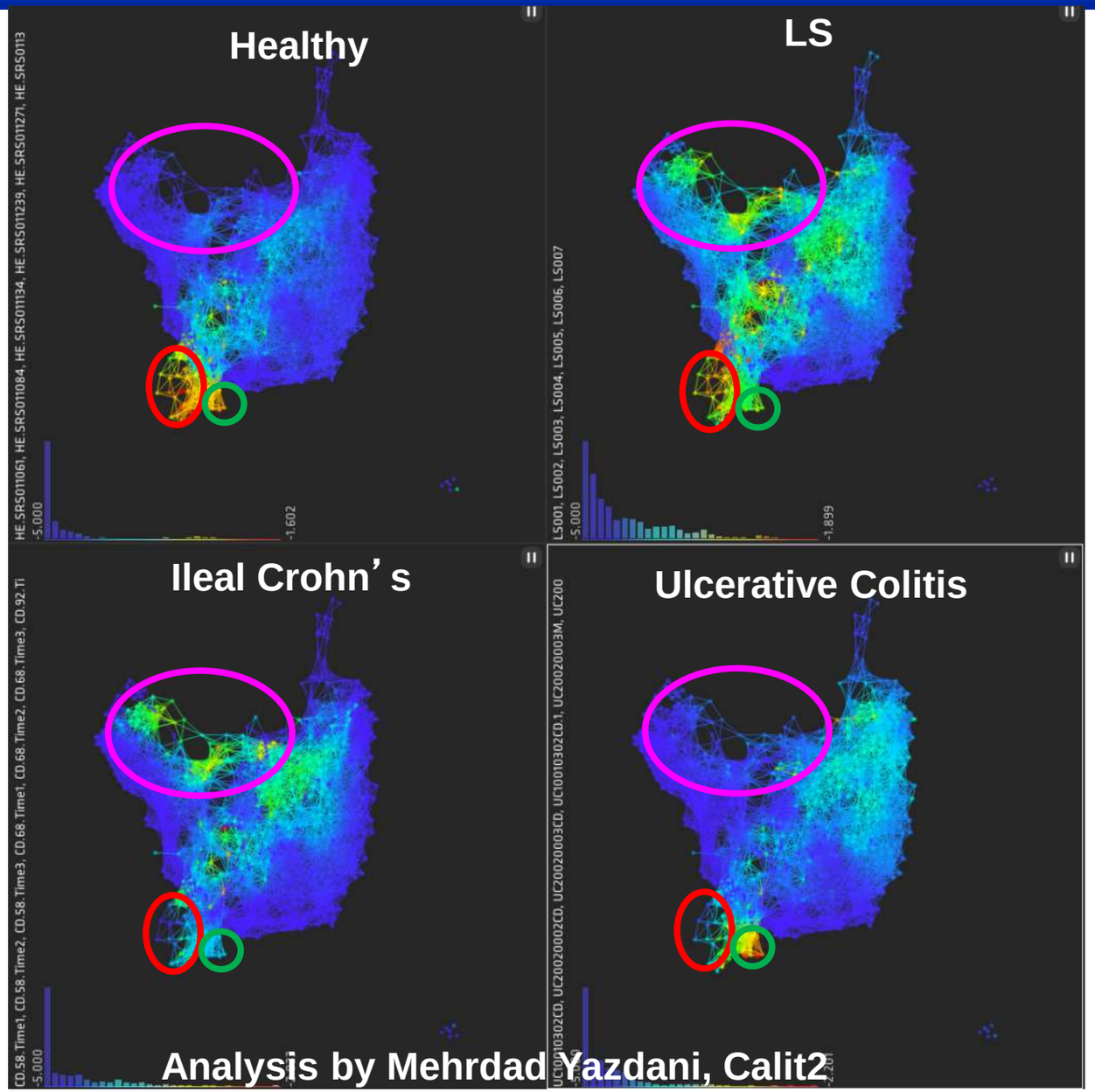
Healthy, Ulcerative Colitis, and LS

Analysis by Mehrdad Yazdani, Calit2

Ayasdi Cure™
Turn Data into Therapies

Talk to Ayasdi in the Intel Booth at SC14

Ayasdi Enables Discovery of Differences Between Healthy and Disease States Using Microbiome Species



Ayasdi Cure™
Turn Data into Therapies

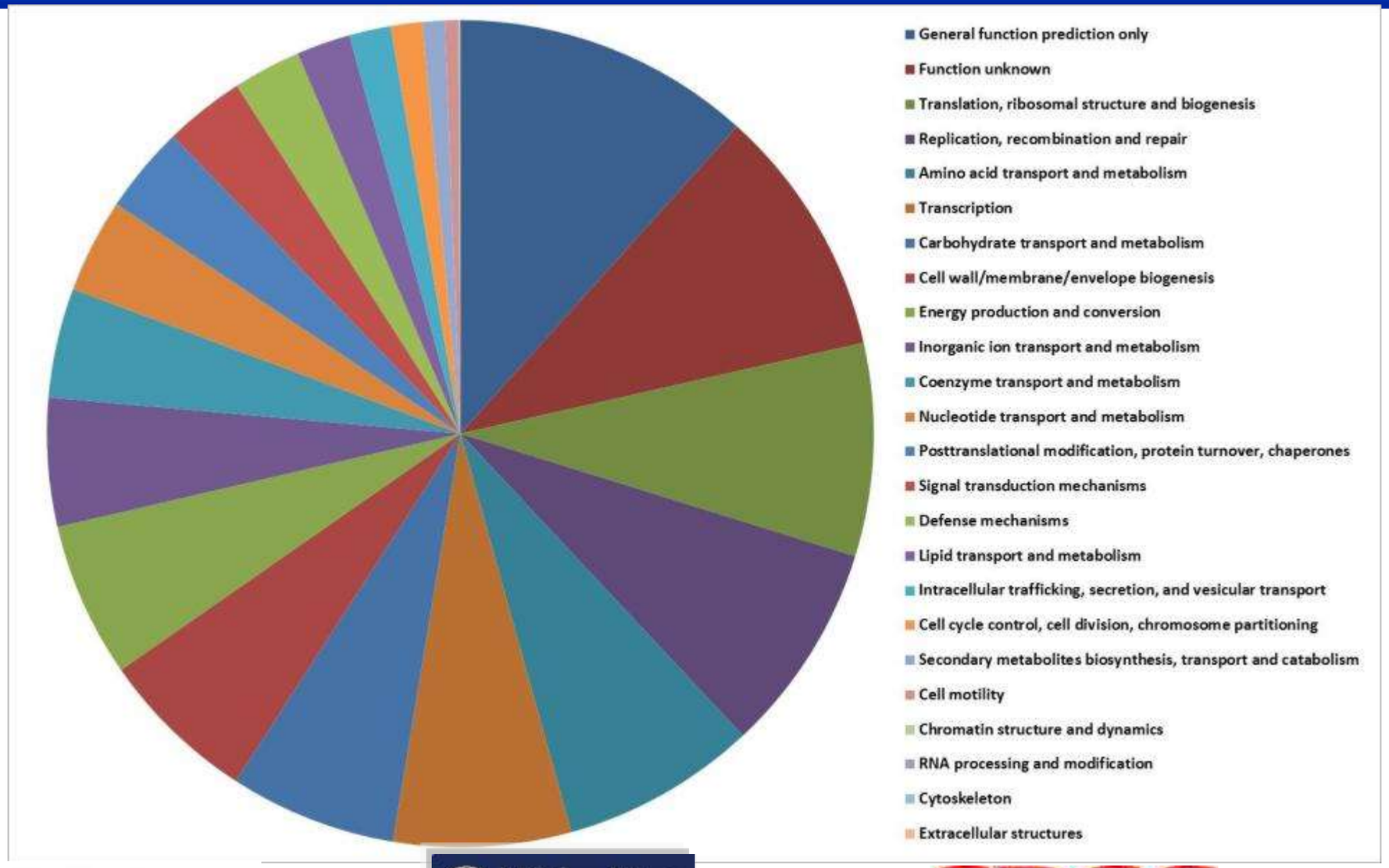
High in Healthy and LS

High in Healthy and Ulcerative Colitis

High in Both LS and Ileal Crohn's Disease

Using Multidimensional Scaling Lens with Correlation Metric

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



J. Craig Venter INSTITUTE

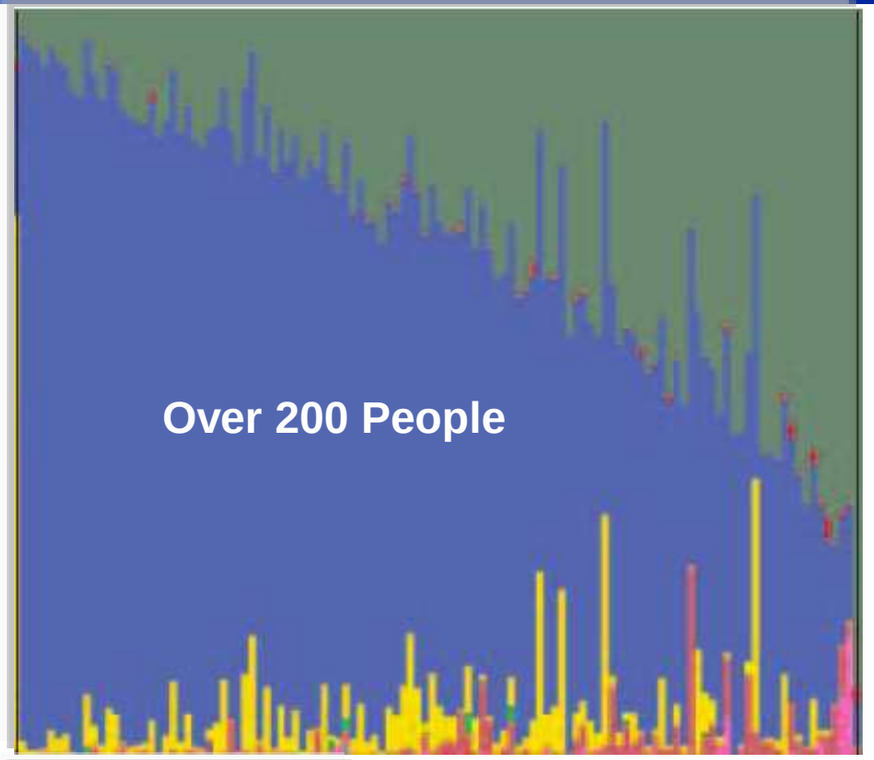
UC San Diego HEALTH SCIENCES

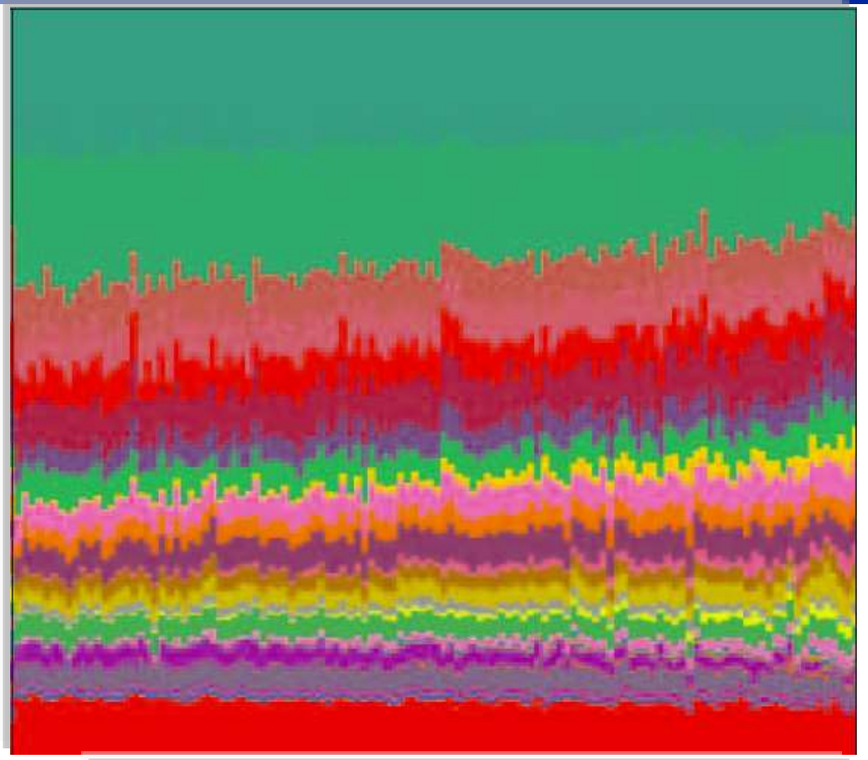
SDSC

it²

Analysis: Weizhong Li & Sitao Wu, UCSD

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



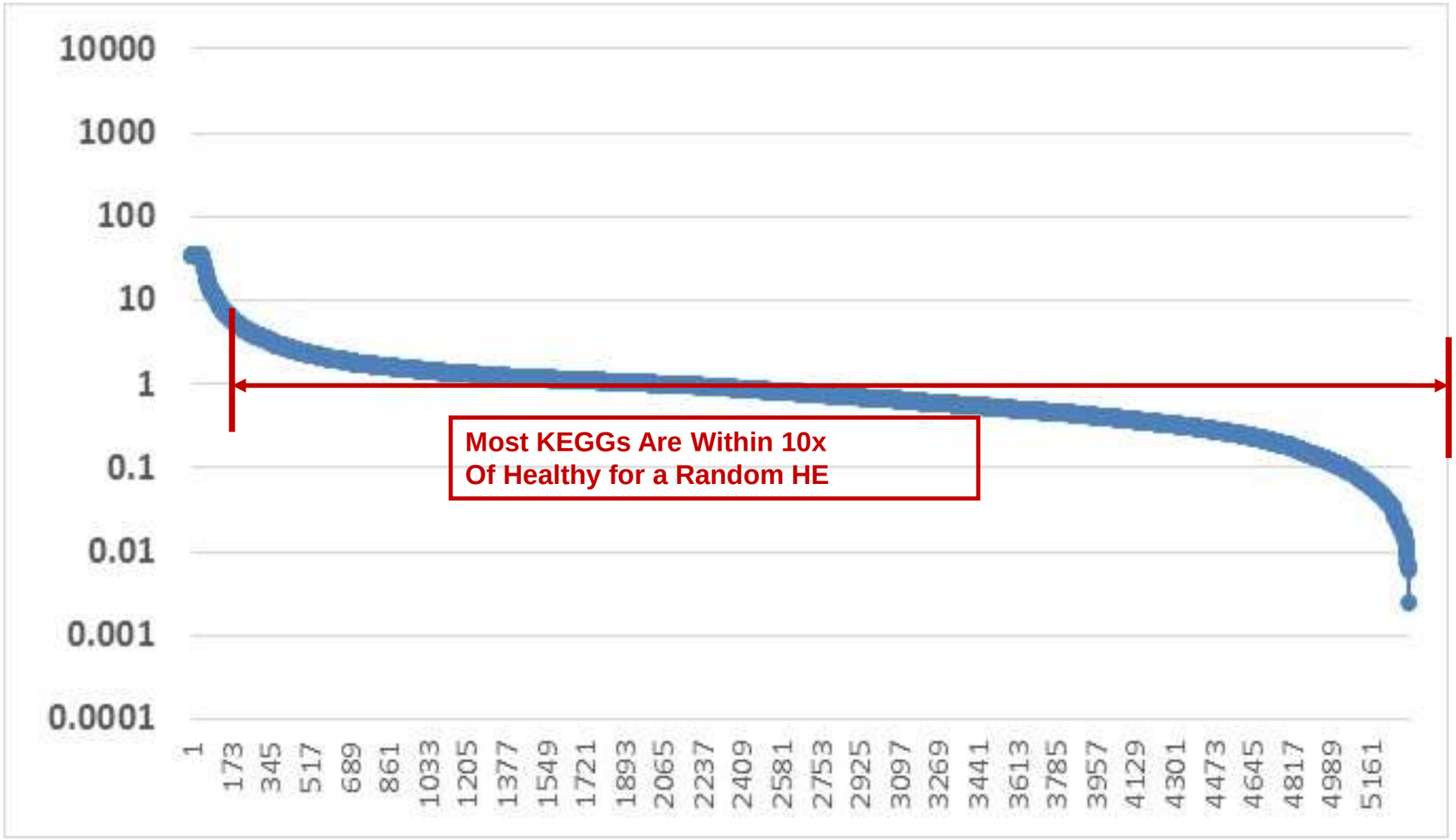


- Firmicutes
- Actinobacteria
- Bacteroidetes
- Proteobacteria
- Fusobacteria
- Tenericutes
- Spirochaetes
- Cyanobacteria
- Verrucomicrobia
- TM7

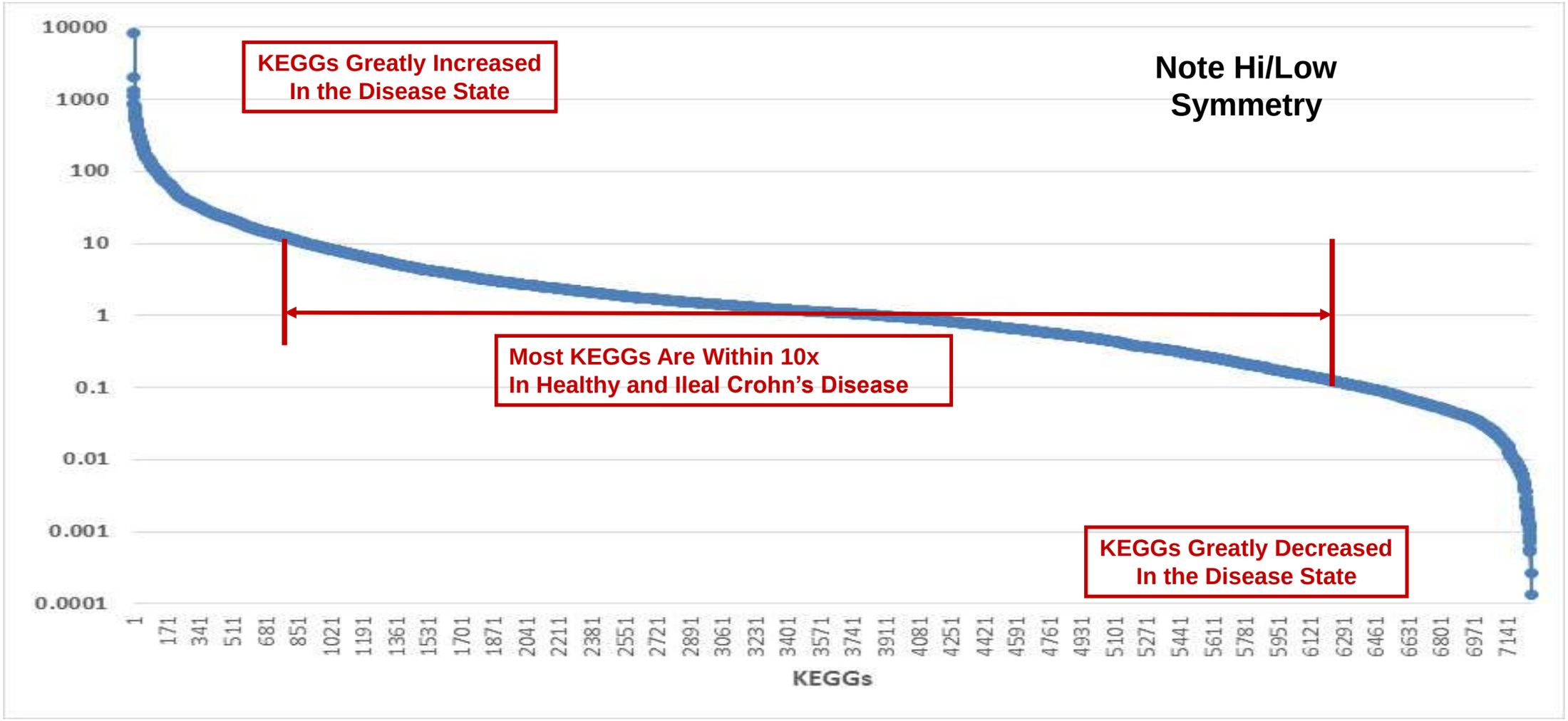
- Central carbohydrate metabolism
- Cofactor and vitamin biosynthesis
- Oligosaccharide and polyol transport system
- Purine metabolism
- ATP synthesis
- Phosphate and amino acid transport system
- Aminoacyl tRNA
- Pyrimidine metabolism
- Ribosome
- Aromatic amino acid metabolism

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

Ratio of HE11529 to Ave HE
Test to see How Much Variation There is Within Healthy

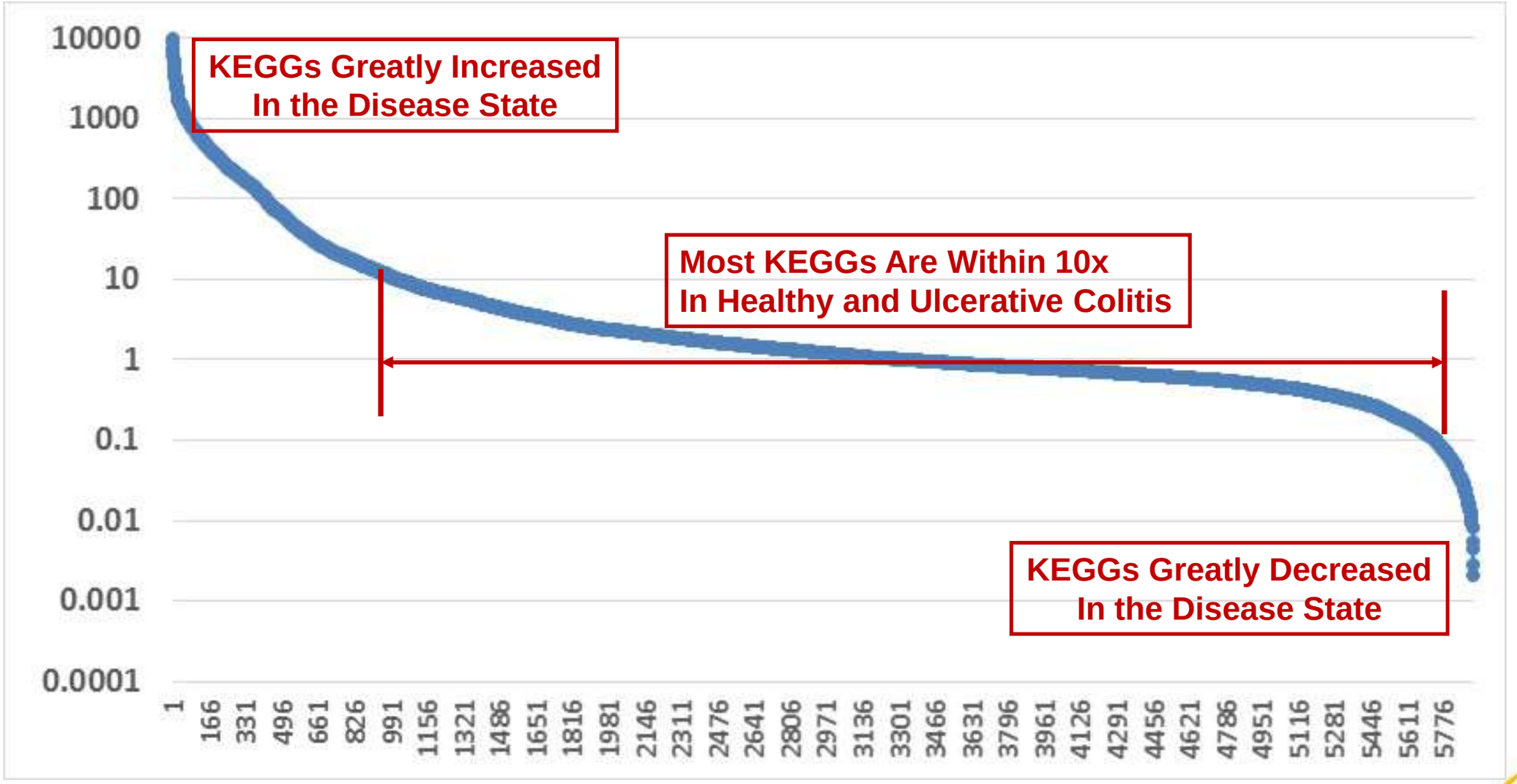


However, Our Research Shows Large Changes in Protein Families Between Health and Disease

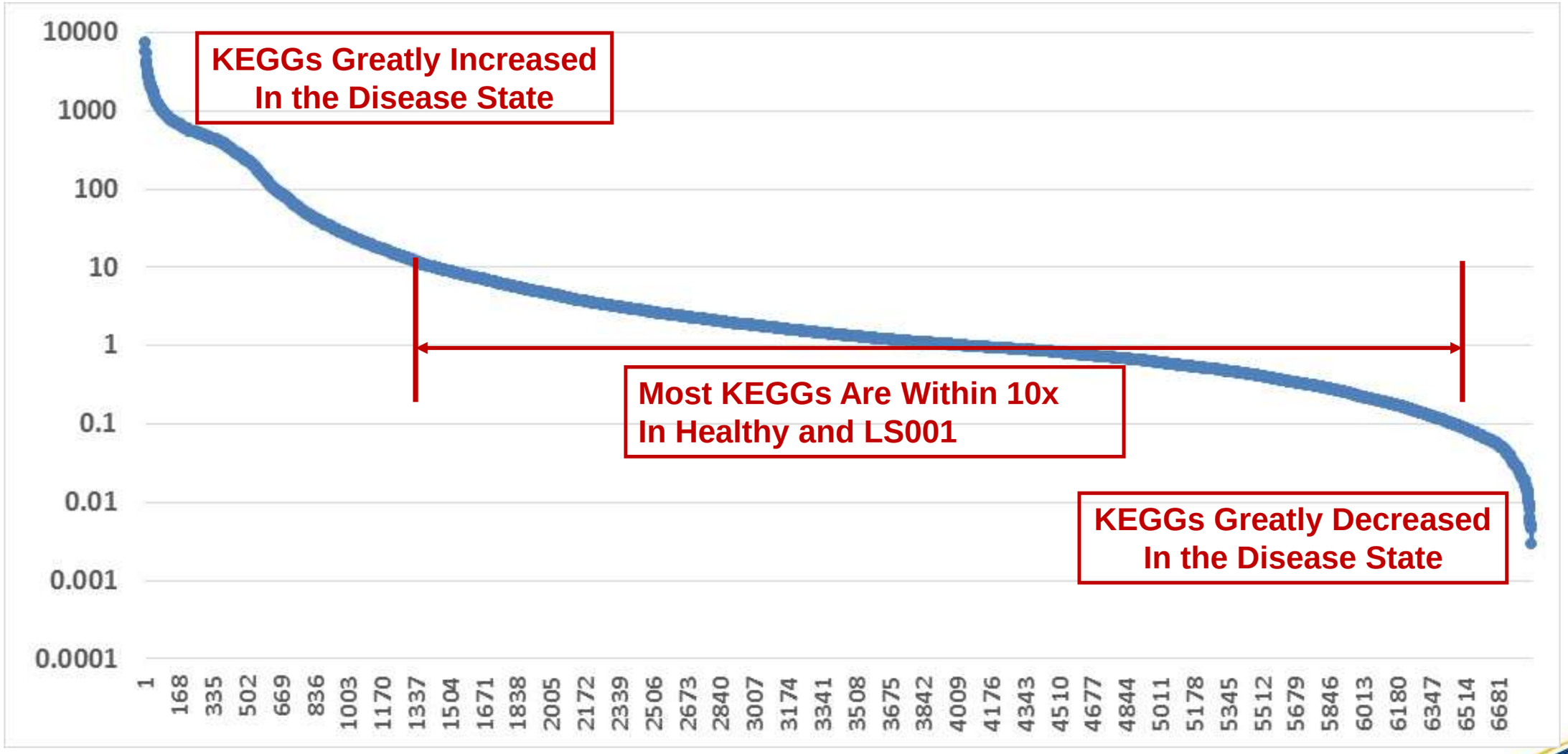


Over 7000 KEGGs Which Are Nonzero in Health and Disease States

Note UC Has Many Few KEGGs that are Much Smaller than HE; Also Fewer KEGGs That are Nonzero; Note Asymmetry Between High & Low



Note LS001 Has Many Few KEGGs that are Much Smaller than HE; ~Same # KEGGs That are Nonzero; Note Asymmetry Between High & Low



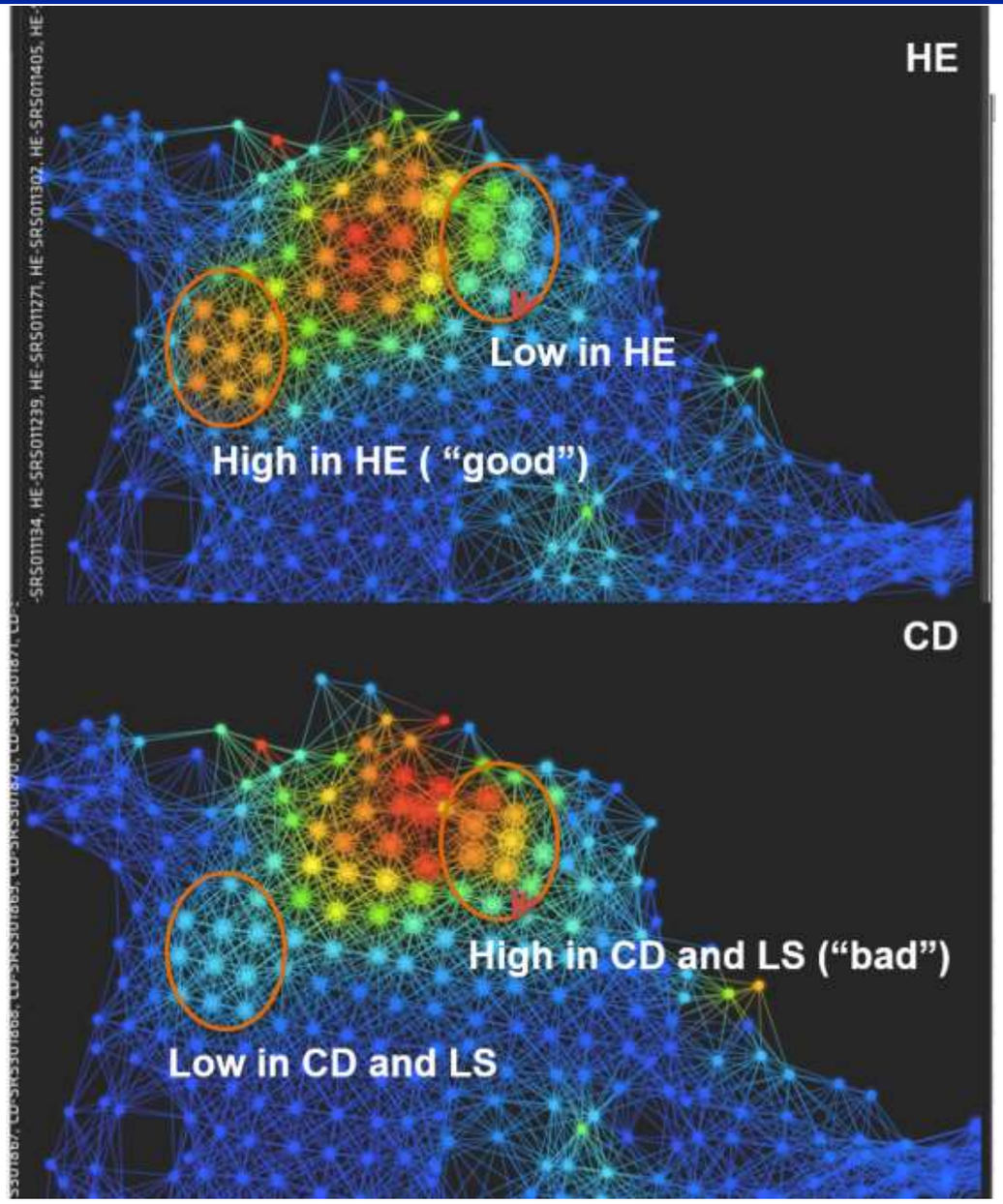
We Can Define a Subgroup of the 10,000 KEGGs Which Are Extreme in the Disease State

- **Look for KEGGs That Have the Properties:**
  - **Are 100x in All Four Disease States**
    - **LS001/Ave HE**
    - **Ave CD/ Ave HE**
    - **Ave UC/Ave HE**
    - **Sick HE Person/Ave HE**
- **There are 48 of These Extreme KEGGs**

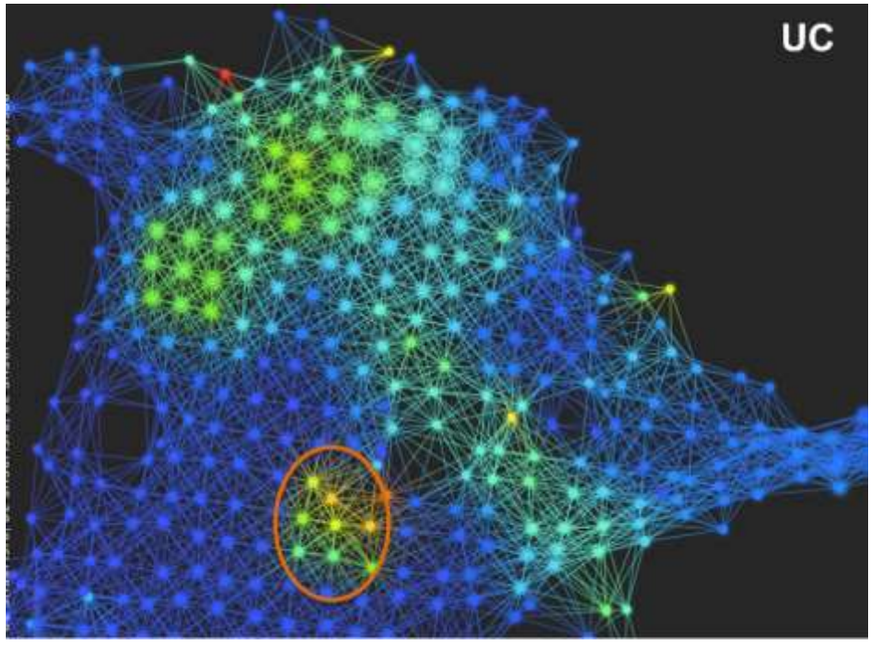
- **A New Way to Define What is Wrong with the Microbiome in Disease?**

- **Can We Devise an Ayasdi Lens That Can Separates These Extreme KEGGs?**

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

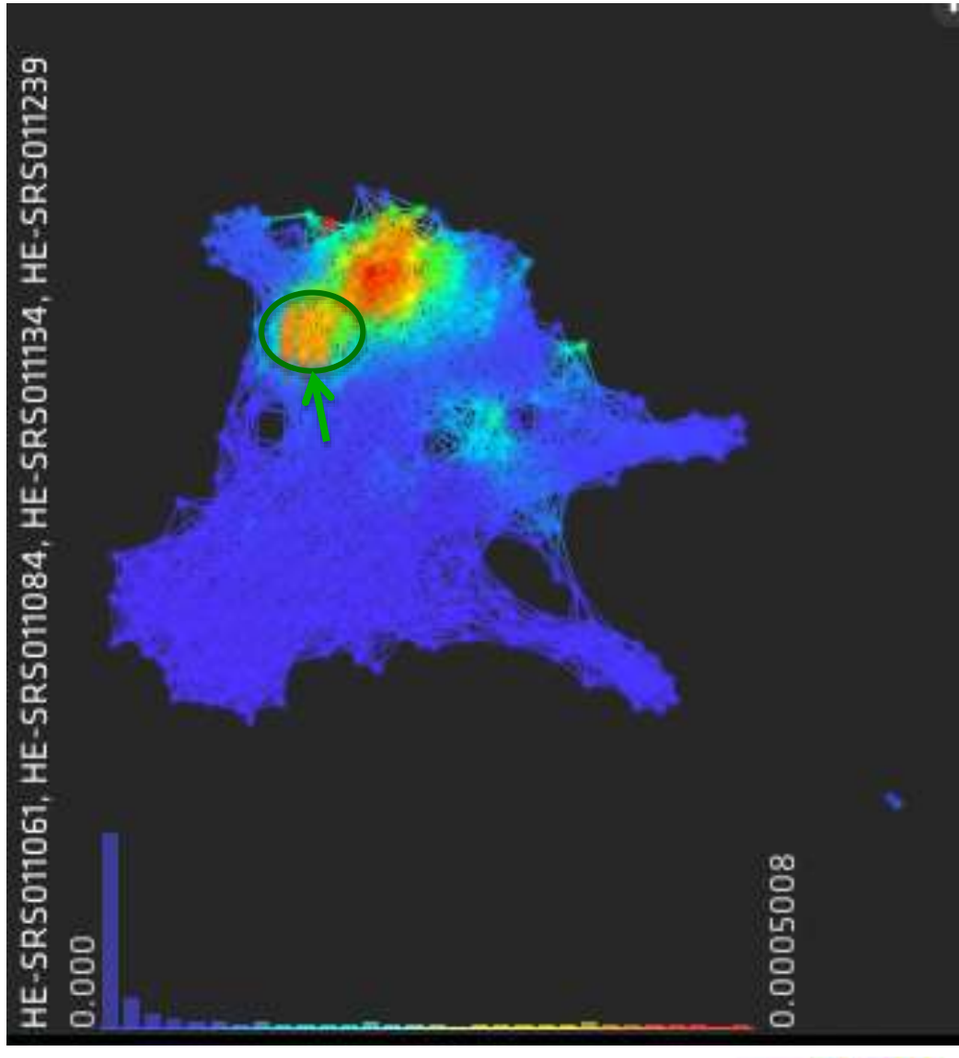


CD is Missing a Population of Bacteria That Exists in High Quantities in HE (Circled with Arrow)

- **Problem is That These KEGGs Have Moderate Values of Ave CD/ Ave HE**
- **How Can We Change the Ayasdi Lenses So That We Pick Out The Very High Values of Ratios to Ave HE?**

Low in CD and LS

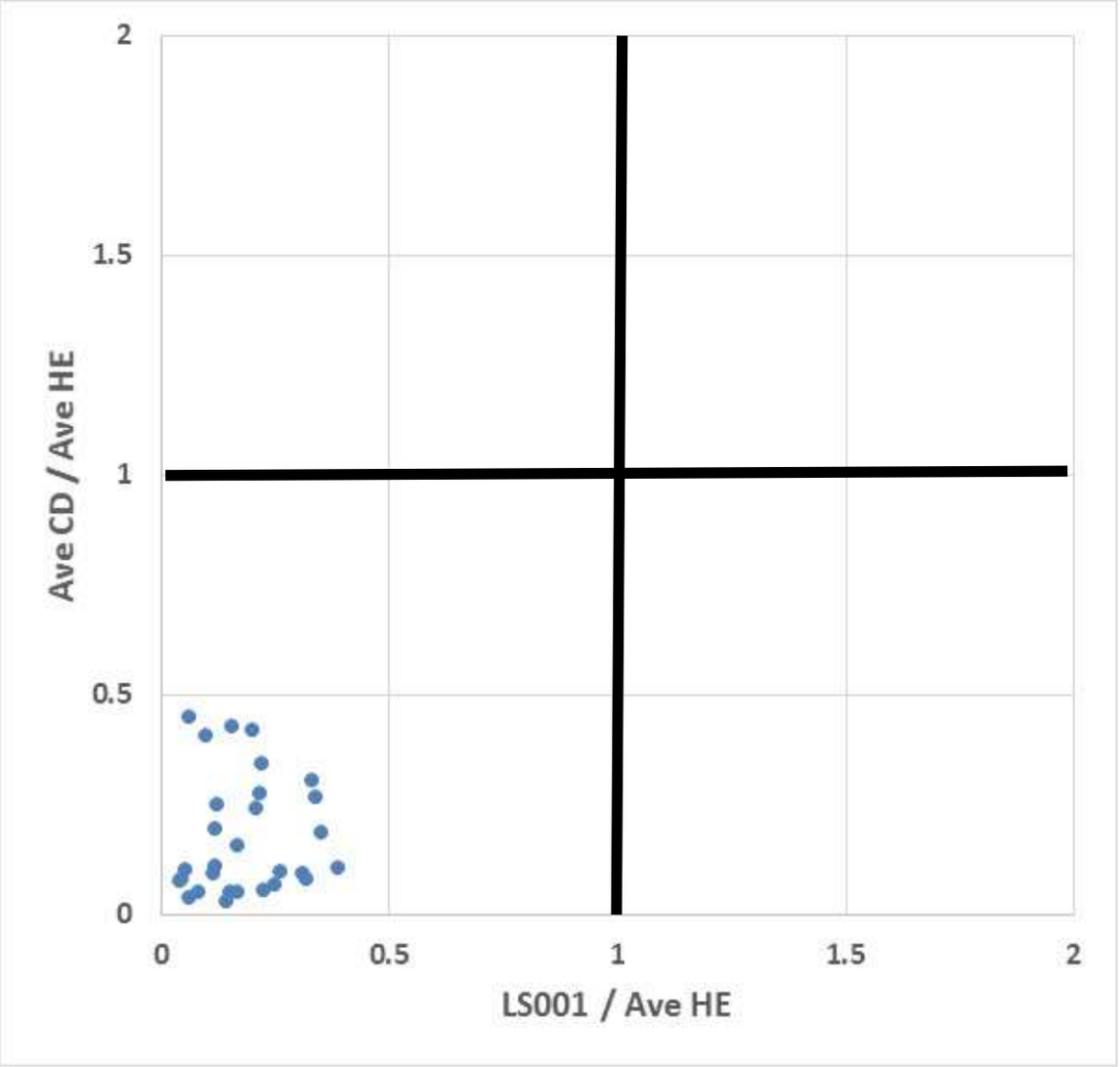
	LS0
K01190(beta-galactosidase)	0.0
K12373(hexosaminidase)	0.0
K02014(iron_complex_outermembrane_recepter_...	0.0
K03088(RNA_polymerase_sigma-70_factor,_ECF_...	0.0
K01195(beta-glucuronidase)	0.0
K07114(N-acetylmuramic_acid_6-phosphate_eth...	0.0
K01238(myrosinase)	0.0
K01198(xylan_1,4-beta-xylosidase)	0.0
K03407(two-component_system,_chemotaxis_fa...	0.0
K01130(arylsulfatase)	0.0
K01768(adenylate_cyclase)	0.0
K01209(alpha-N-arabinofuranosidase)	0.0
K07165(transmembrane_sensor)	0.0
K01138(N-acetylglucosamine-6-sulfatase)	0.0
K03287(outer_membrane_factor,_OMF_family)	0.0
K01213(galacturan_1,4-alpha-galacturonidase)	7.7
K03296(hydrophobic/amphiphilic_exporter-1_(m...	0.0
K01134(arylsulfatase_A)	0.0
K01077(alkaline_phosphatase)	0.0
K01051(pectinesterase)	0.0
K12340(outer_membrane_channel_protein_TolC)	0.0
K01184(polygalacturonase)	5.5
K03585(membrane_fusion_protein)	0.0
K02005(HlyD_family_secretion_protein)	0.0
K15532()	6.5
K03797(carboxyl-terminal_processing_protease)	0.0
K07713(two-component_system,_NtrC_family,_re...	0.0
K01137(N-acetylglucosamine-6-sulfatase)	0.0
K03350(anaphase-promoting_complex_subunit_3)	0.0
K00971(mannose-1-phosphate_guanylyltransfer...	0.0
K01847(methylmalonyl-CoA_mutase)	0.0



Source: Pek Lum,
Formerly Chief Data Scientist, Ayasdi

This Ayasdi Lens Does Identify KEGGs In Which Ave CD and LS001 Are Less Than Ave HE

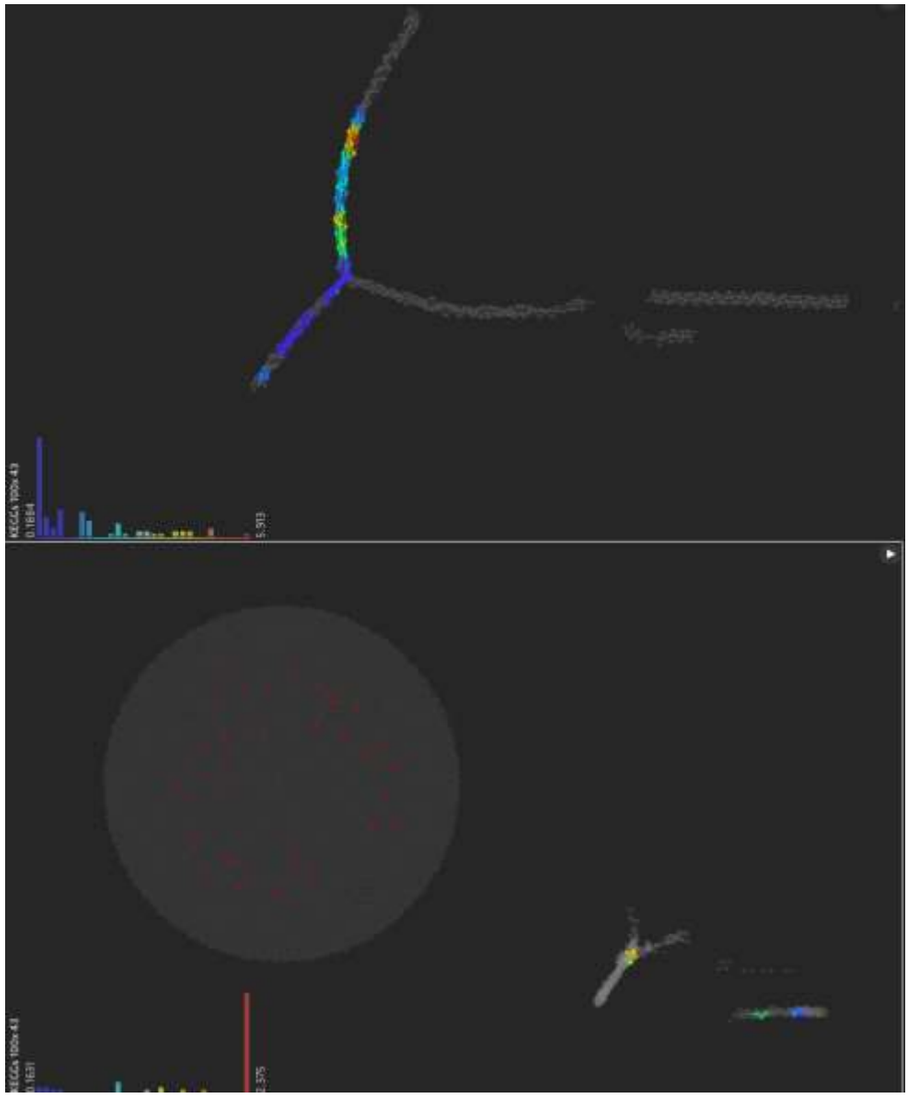
- **Problem is That These KEGGs Have Moderate Low Values of Ave CD/ Ave HE**
- **How Can We Change the Ayasdi Lenses So That We Pick Out The Very High Values of Ratios to Ave HE?**



We Found a Set of Lenes That Clearer Find the 43 Extreme KEGGs

L-Infinity Centrality Lens Using Norm Correlation as Metric (Resolution: 242, Gain: 5.7)

Entropy & Variance Lens Using Angle as Metric (Resolution: 30, Gain 3.00)

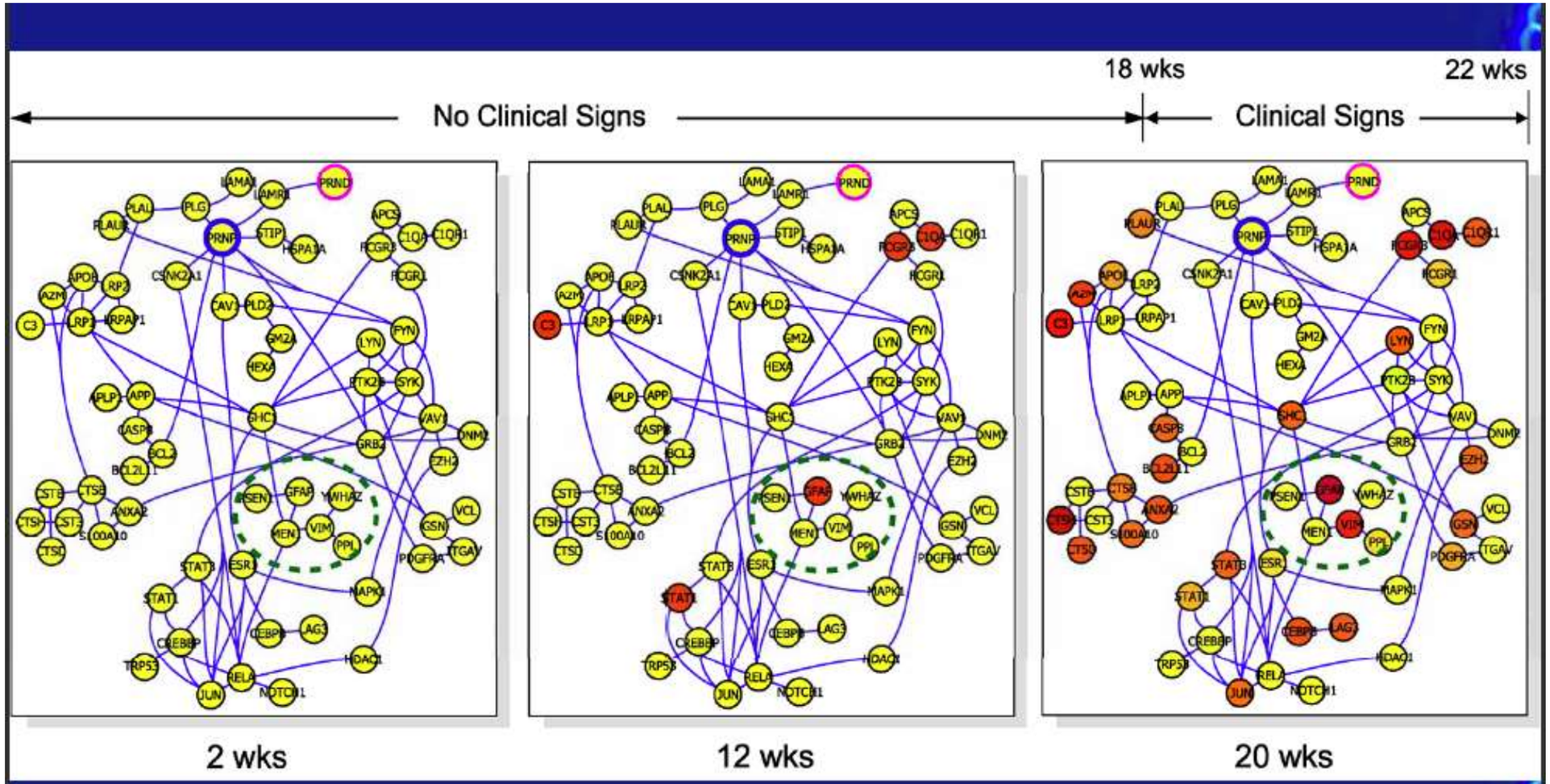


K00108(choline_dehydrogenase)
K00673(arginine_N-succinyltransferase)
K00867(type_I_pantothenate_kinase)
K01169(ribonuclease_I_(enterobacter_ribonuclease))
K01484(succinylarginine_dihydrolase)
K01682(aconitate_hydratase_2)
K01690(phosphogluconate_dehydratase)
K01825(3-hydroxyacyl-CoA_dehydrogenase_/_enoyl-CoA_hydratase_/3-hydroxybutyryl-CoA_epimerase_/_e
K02173(hypothetical_protein)
K02317(DNA_replication_protein_DnaT)
K02466(glucitol_operon_activator_protein)
K02846(N-methyl-L-tryptophan_oxidase)
K03081(3-dehydro-L-gulonate-6-phosphate_decarboxylase)
K03119(taurine_dioxygenase)
K03181(chorismate--pyruvate_lyase)
K03807(AmpE_protein)
K05522(endonuclease_VIII)
K05775(maltose_operon_periplasmic_protein)
K05812(conserved_hypothetical_protein)
K05997(Fe-S_cluster_assembly_protein_SufA)
K06073(vitamin_B12_transport_system_permease_protein)
K06205(MioC_protein)
K06445(acyl-CoA_dehydrogenase)
K06447(succinylglutamic_semialdehyde_dehydrogenase)
K07229(TrkA_domain_protein)
K07232(cation_transport_protein_ChaC)
K07312(putative_dimethyl_sulfoxide_reductase_subunit_YnfH_(DMSO_reductaseanchor_subunit))
K07336(PKHD-type_hydroxylase)
K08989(putative_membrane_protein)
K09018(putative_monooxygenase_RutA)
K09456(putative_acyl-CoA_dehydrogenase)
K09998(arginine_transport_system_permease_protein)
K10748(DNA_replication_terminus_site-binding_protein)
K11209(GST-like_protein)
K11391(ribosomal_RNA_large_subunit_methyltransferase_G)
K11734(aromatic_amino_acid_transport_protein_AroP)
K11735(GABA_permease)
K11925(SgrR_family_transcriptional_regulator)
K12288(pilus_assembly_protein_HofM)
K13255(ferric_iron_reductase_protein_FhuF)
K14588()
K15733()
K15834()

Analysis by Mehrdad Yazdani, Calit2

Disease Arises from Perturbed Protein Family Networks: Dynamics of a Prion Perturbed Network in Mice

Source: Lee Hood, ISB



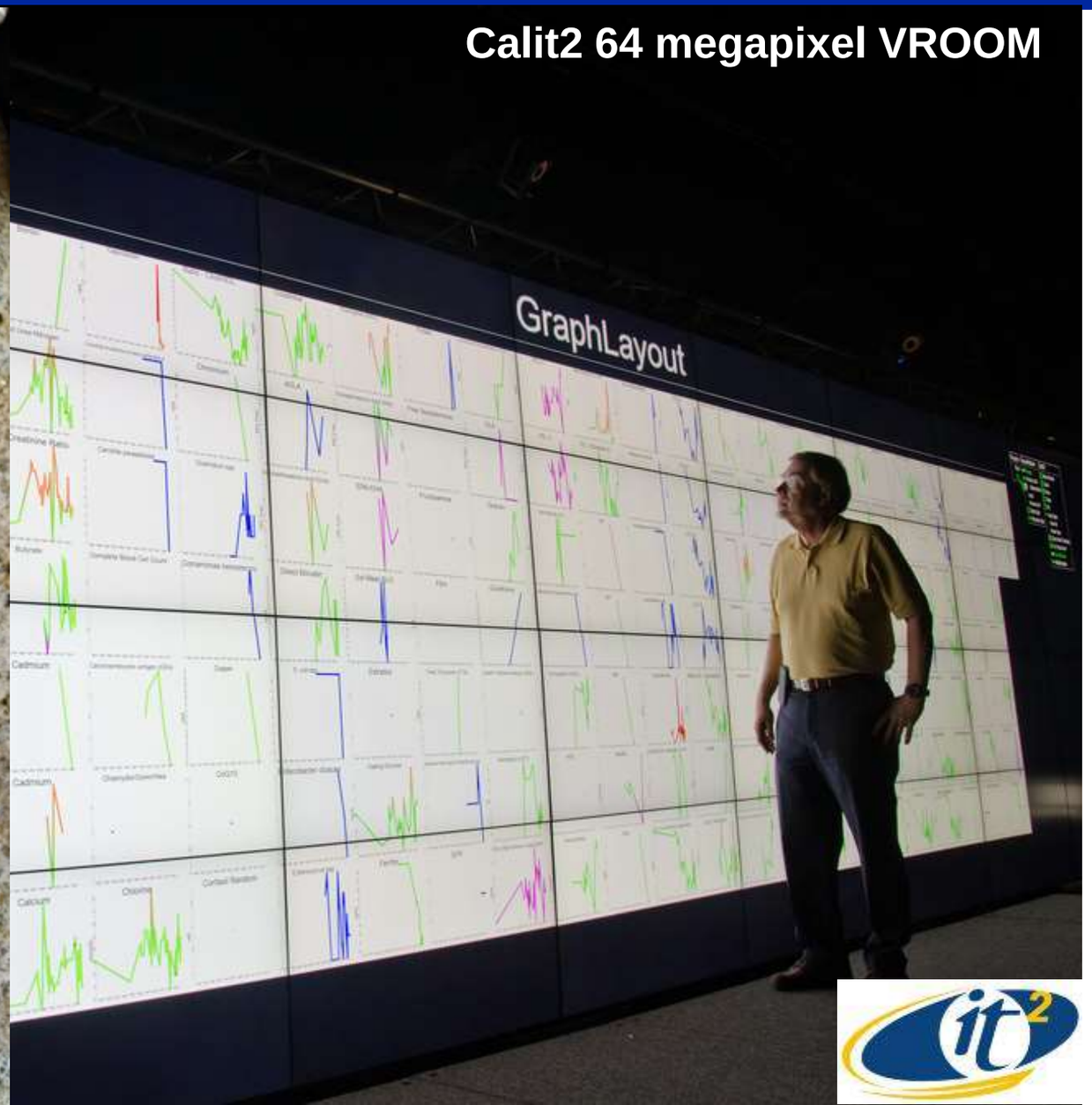
Our Next Goal is to Create Such Perturbed Networks in Humans

Institute for Systems Biology

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

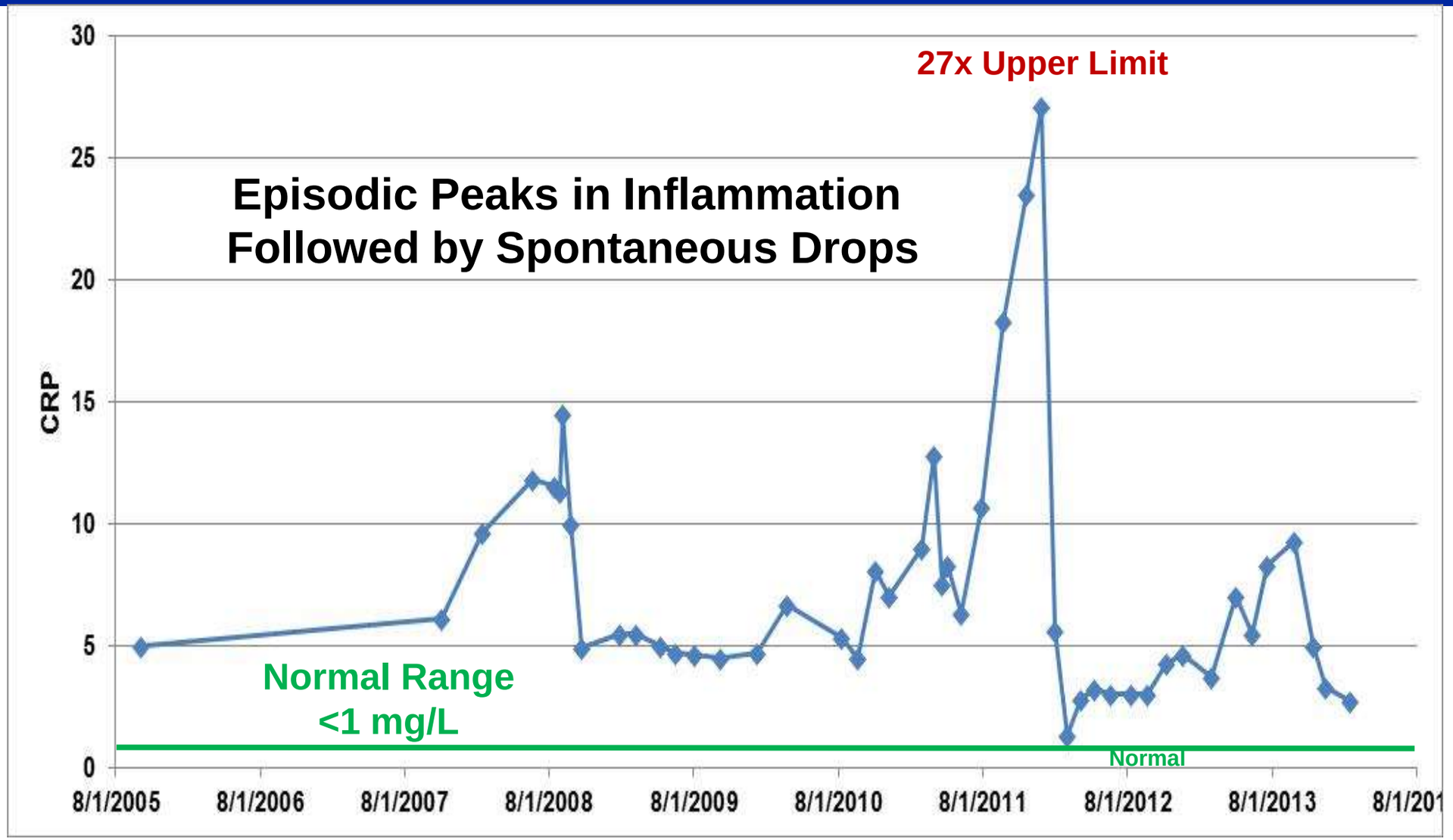


One Blood Draw For Me



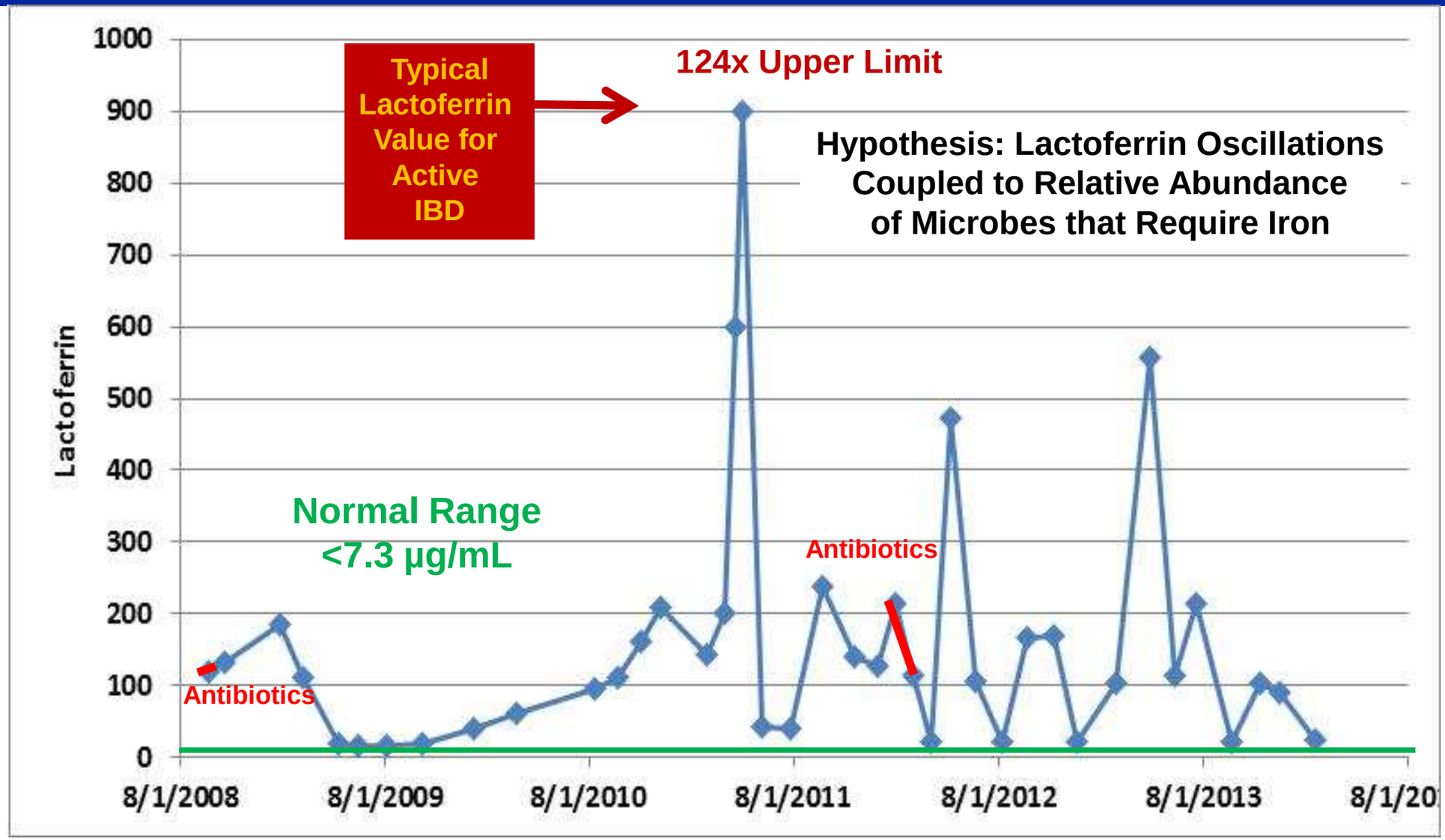
Calit2 64 megapixel VROOM

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



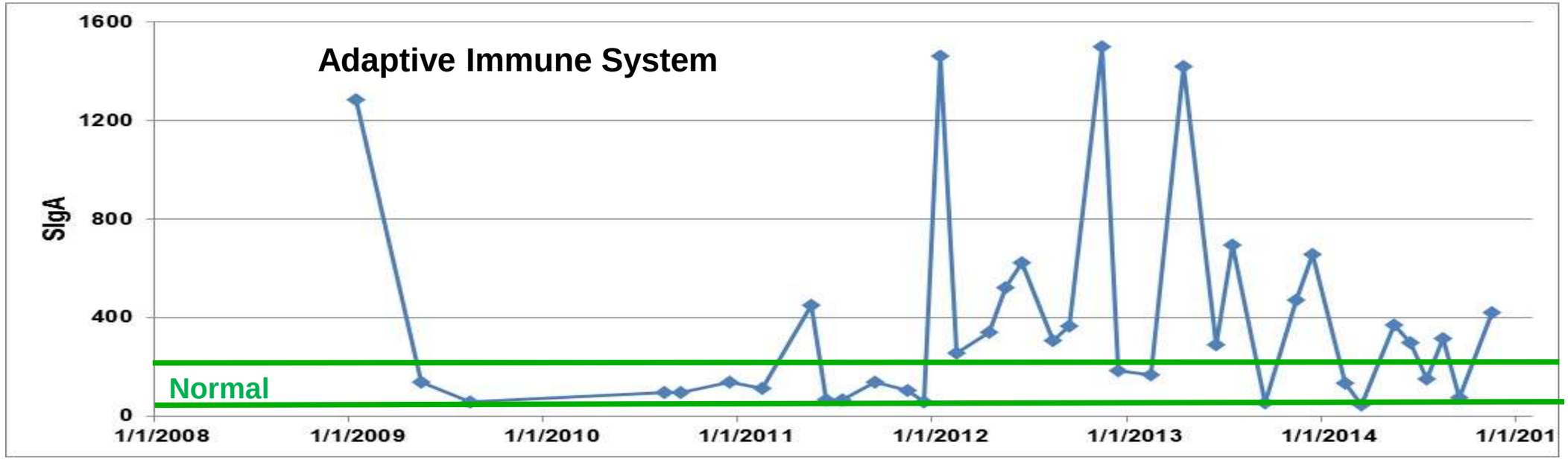
Complex Reactive Protein (CRP) is a Blood Biomarker for Detecting Presence of Inflammation

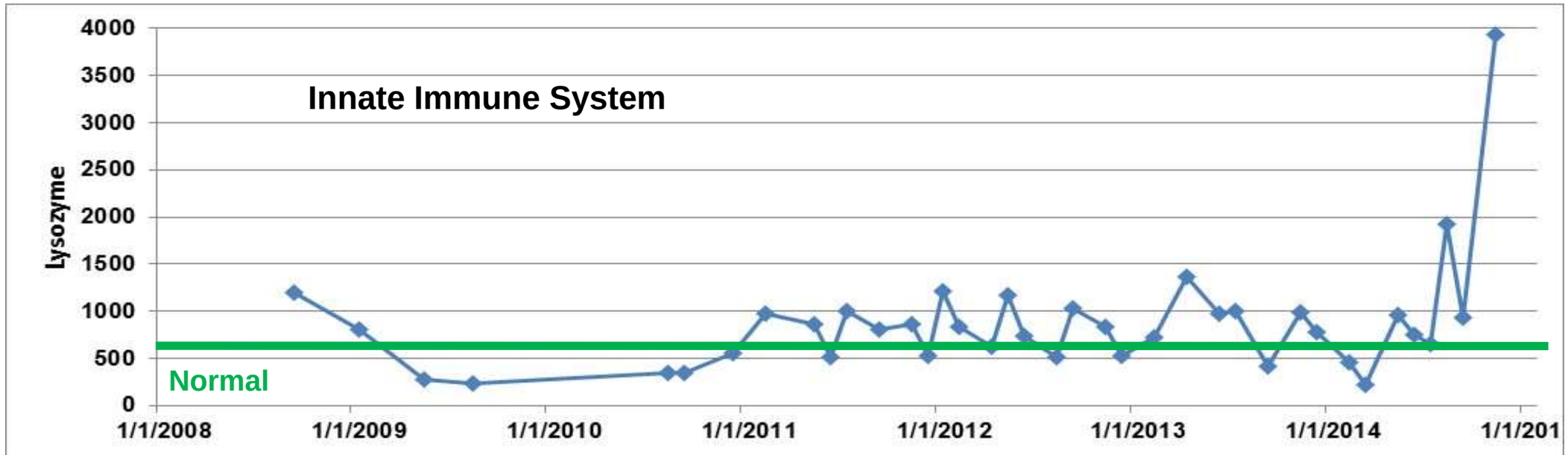
Adding Stool Tests Revealed Oscillatory Behavior in an Immune Variable



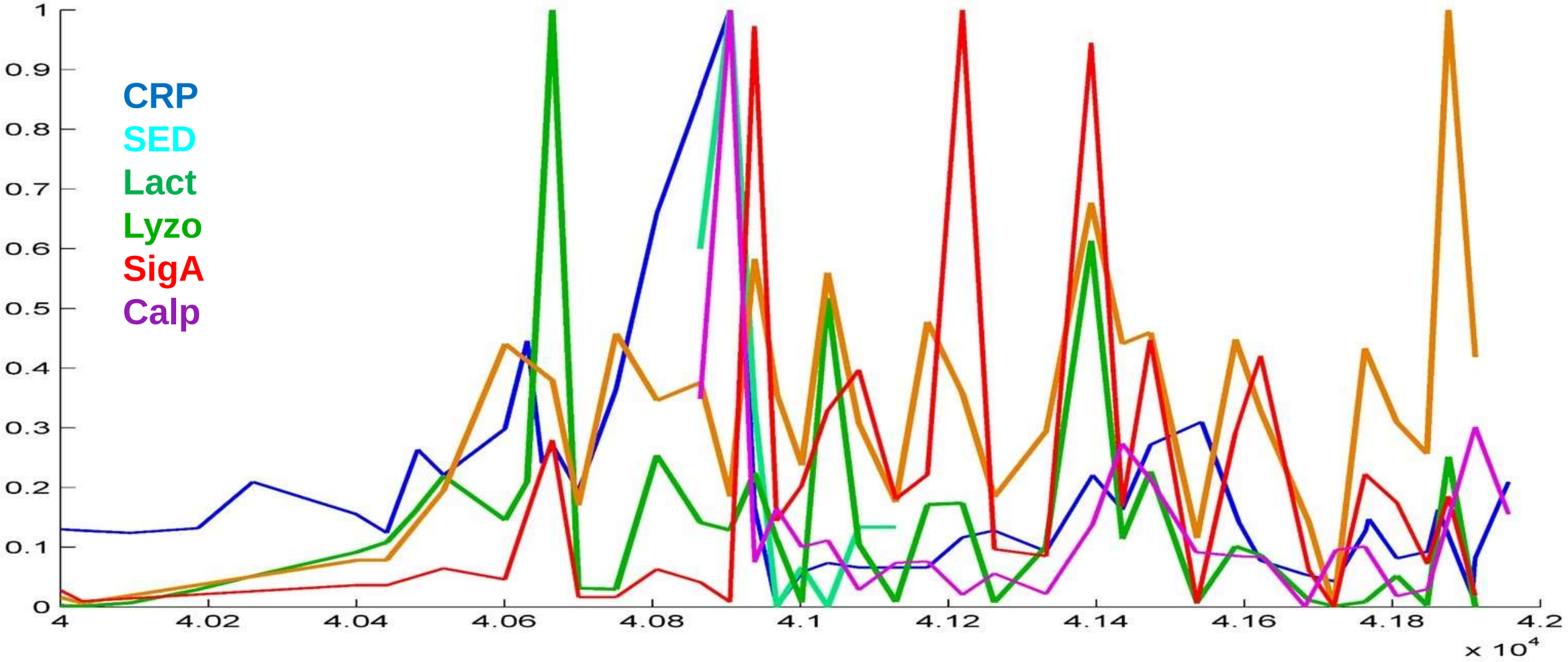
Lactoferrin is a Protein Shed from Neutrophils - An Antibacterial that Sequesters Iron

Fine Time-Resolution Sampling Enables Analysis of Dynamical Innate and Adaptive Immune Dysfunction



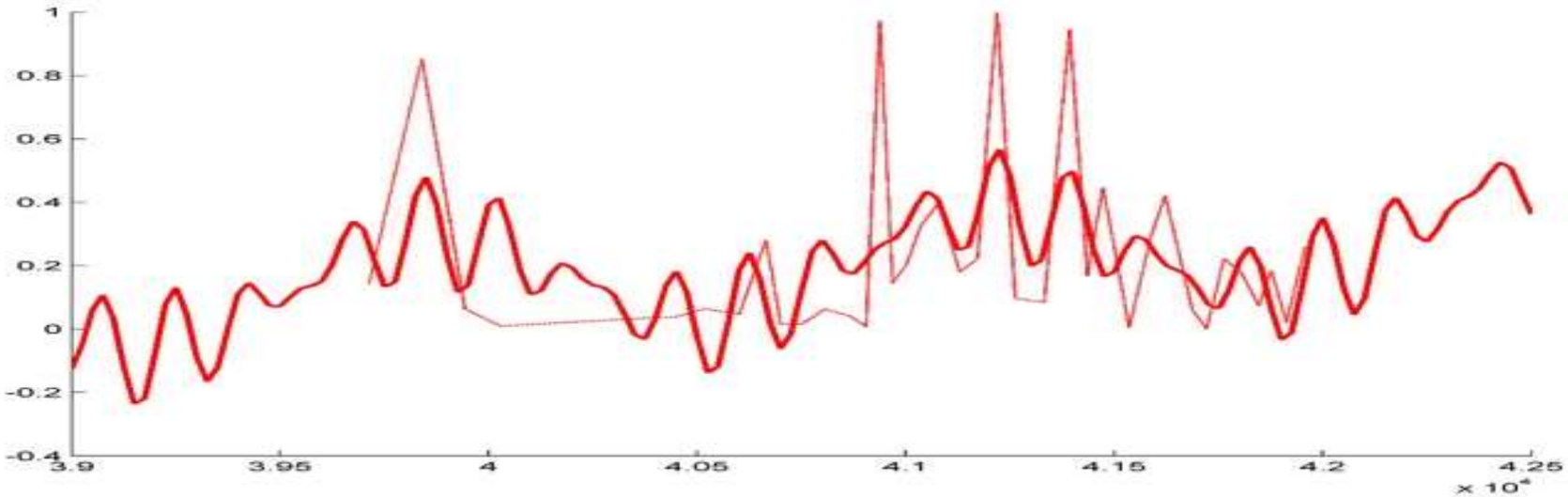


By Overlaying a Number of Immune/Inflammation Variables, It Appears There May be Phase Correlations



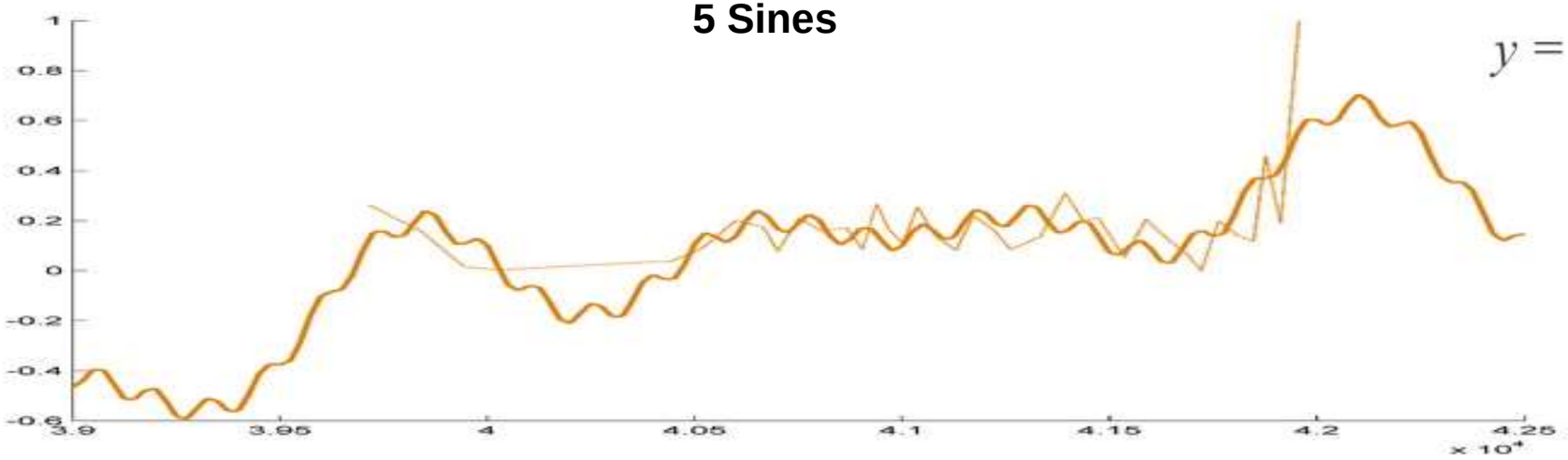
Data Analytics by Benjamin Smarr, UC Berkeley

One Can Use Sine Fitting with Least Squares To Try and Approximate the Time Series Dynamics



5 Sines

$$y = \sum_{i=1}^{n} a_i \sin(b_i x + c_i)$$

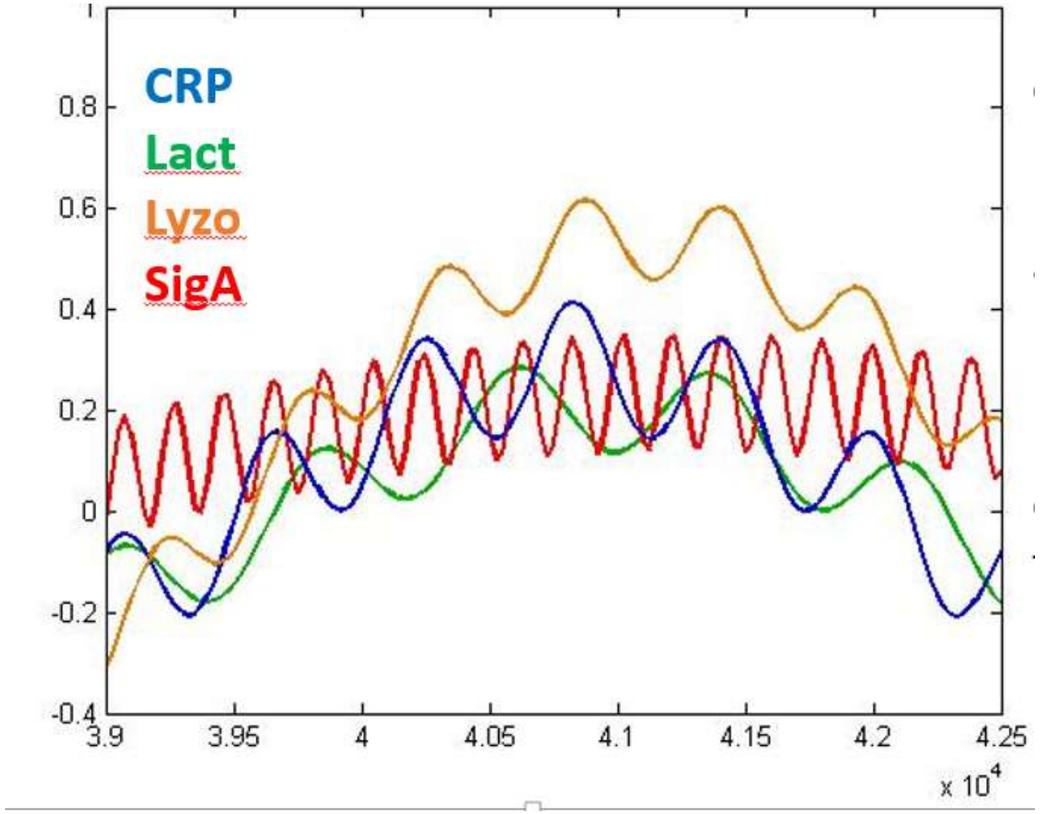


Data Analytics by Benjamin Smarr, UC Berkeley

With Low Resolution Sine Fitting, There Is Indication of Phase Correlation

r	CRP_sl	Lact_s2	Lyzo_sl	SigA_s2
CRP_sl	1.00	0.58	0.80	0.32
Lact_s2	0.58	1.00	0.66	0.43
Lyzo_sl	0.80	0.66	1.00	0.46
SigA_s2	0.32	0.43	0.46	1.00

p	CRP_sl	Lact_s2	Lyzo_sl	SigA_s2
CRP_sl	1.00E+00	5.02E-14	3.06E-33	1.37E-04
Lact_s2	5.02E-14	1.00E+00	5.88E-19	1.10E-07
Lyzo_sl	3.06E-33	5.88E-19	1.00E+00	6.93E-09
SigA_s2	1.37E-04	1.10E-07	6.93E-09	1.00E+00



2 Sines

Data Analytics by Benjamin Smarr, UC Berkeley

Are There Ayasdi Tools to More Deeply Analyze Such Time Series?

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

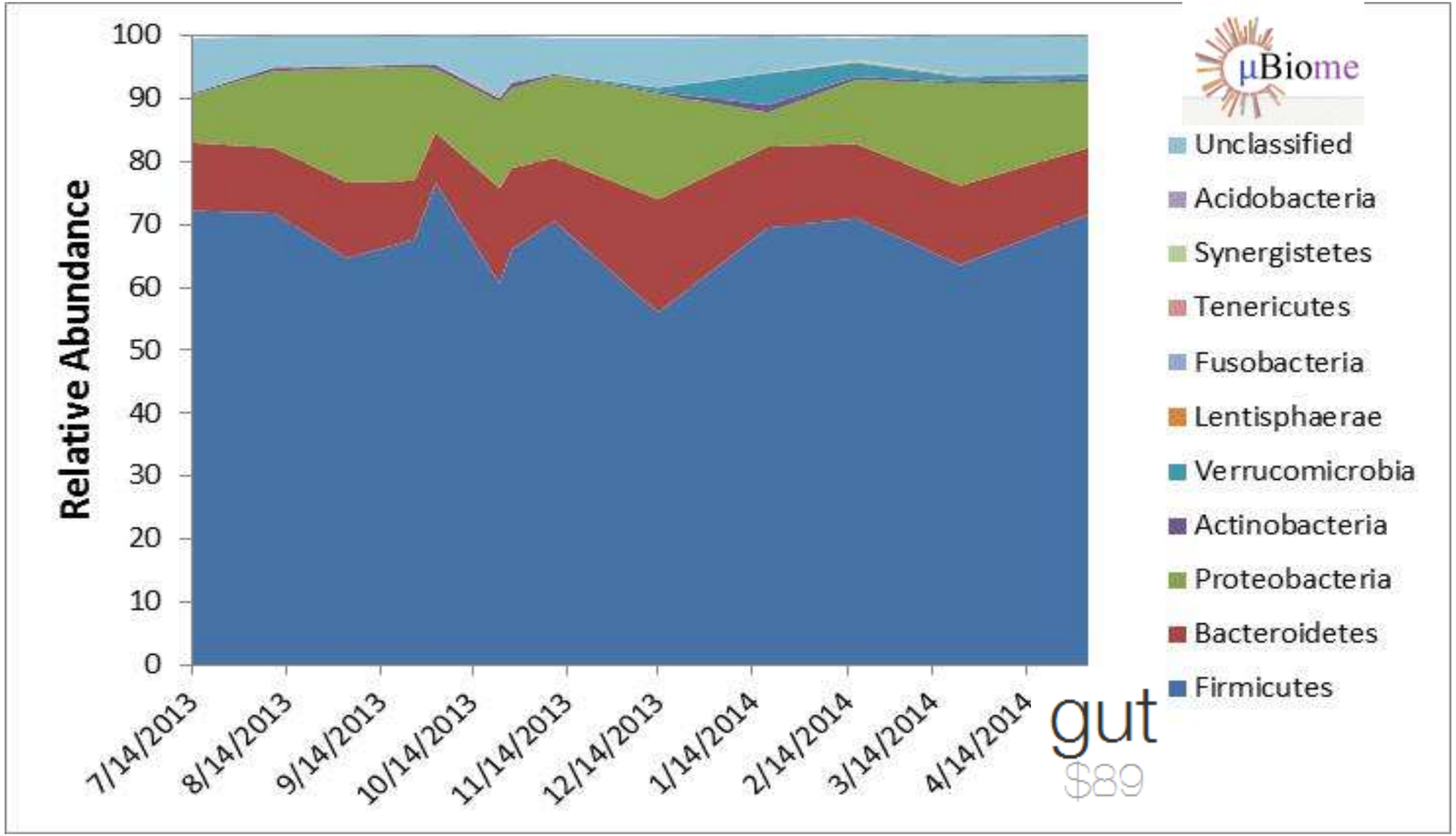
Announced Last Friday!

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

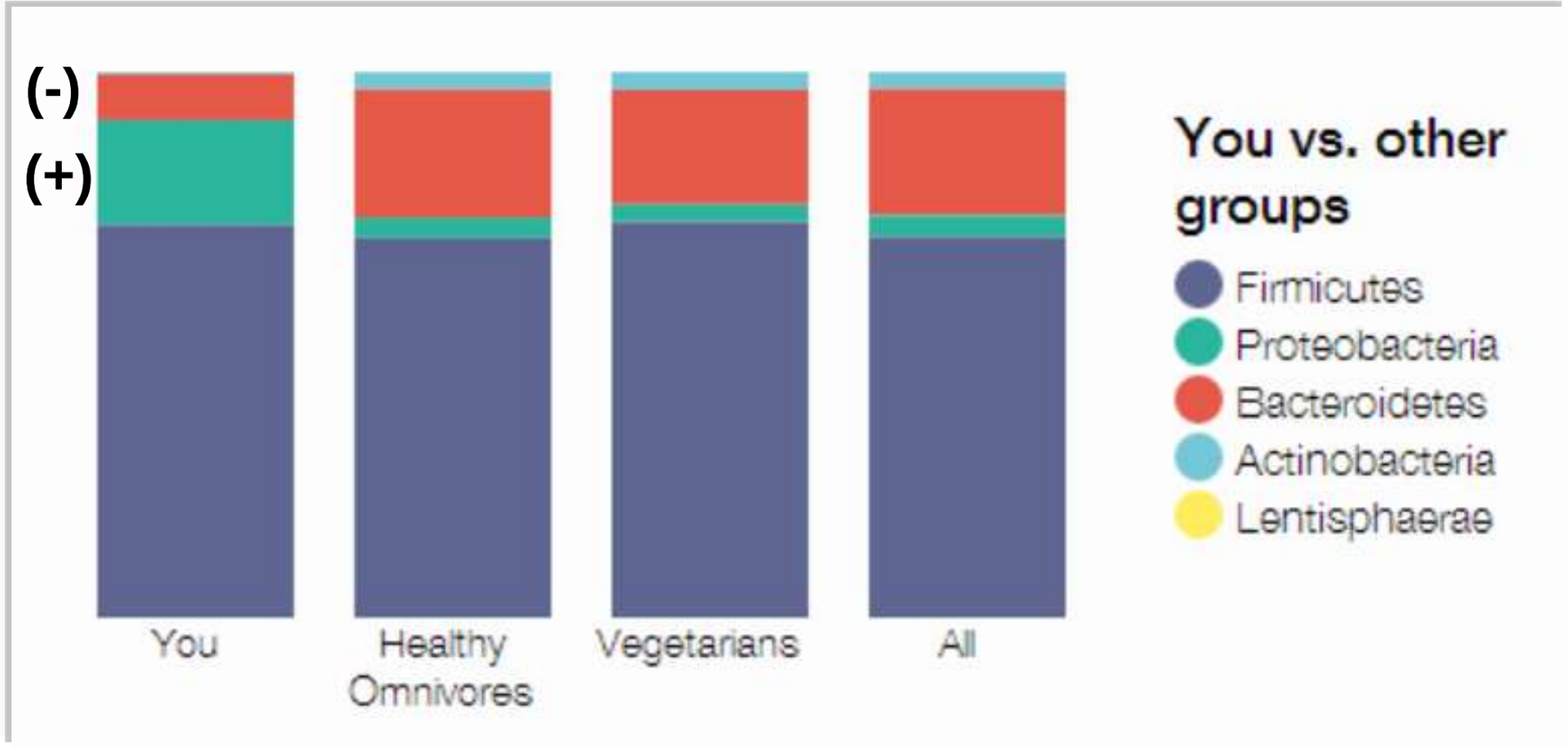
Inexpensive Consumer Time Series of Microbiome Now Possible Through Ubiome



Data source: LS (Stool Samples); Sequencing and Analysis Ubiome

By Crowdsourcing, Ubiome Can Show I Have a Major Disruption of My Gut Microbiome

LS Sample on September 24, 2014



µBiome

Visit Ubiome in the Exponential Medicine Healthcare Innovation Lab

it²

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

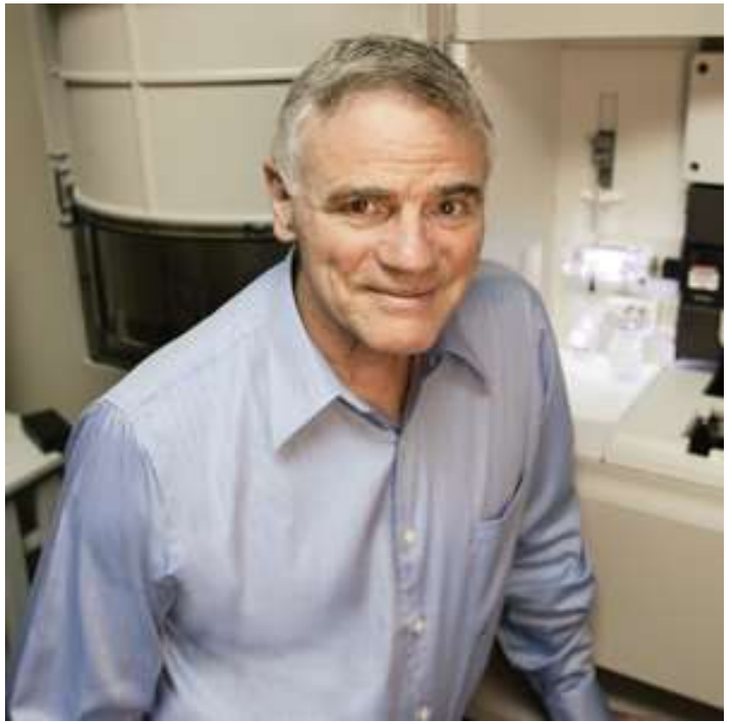
Newsweek

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