“Using Supercomputing & Advanced Analytic Software to Discover Radical Changes in the Human Microbiome in Health and Disease”

Invited Remote Presentation To Weekly Team Meeting
Dermot McGovern, Director, Translational Medicine,
Inflammatory Bowel and Immunobiology Research Institute,
Gastroenterology, Cedars-Sinai

Los Angeles, CA

April 28, 2015

Dr. Larry Smarr

Director, California Institute for Telecommunications and Information Technology

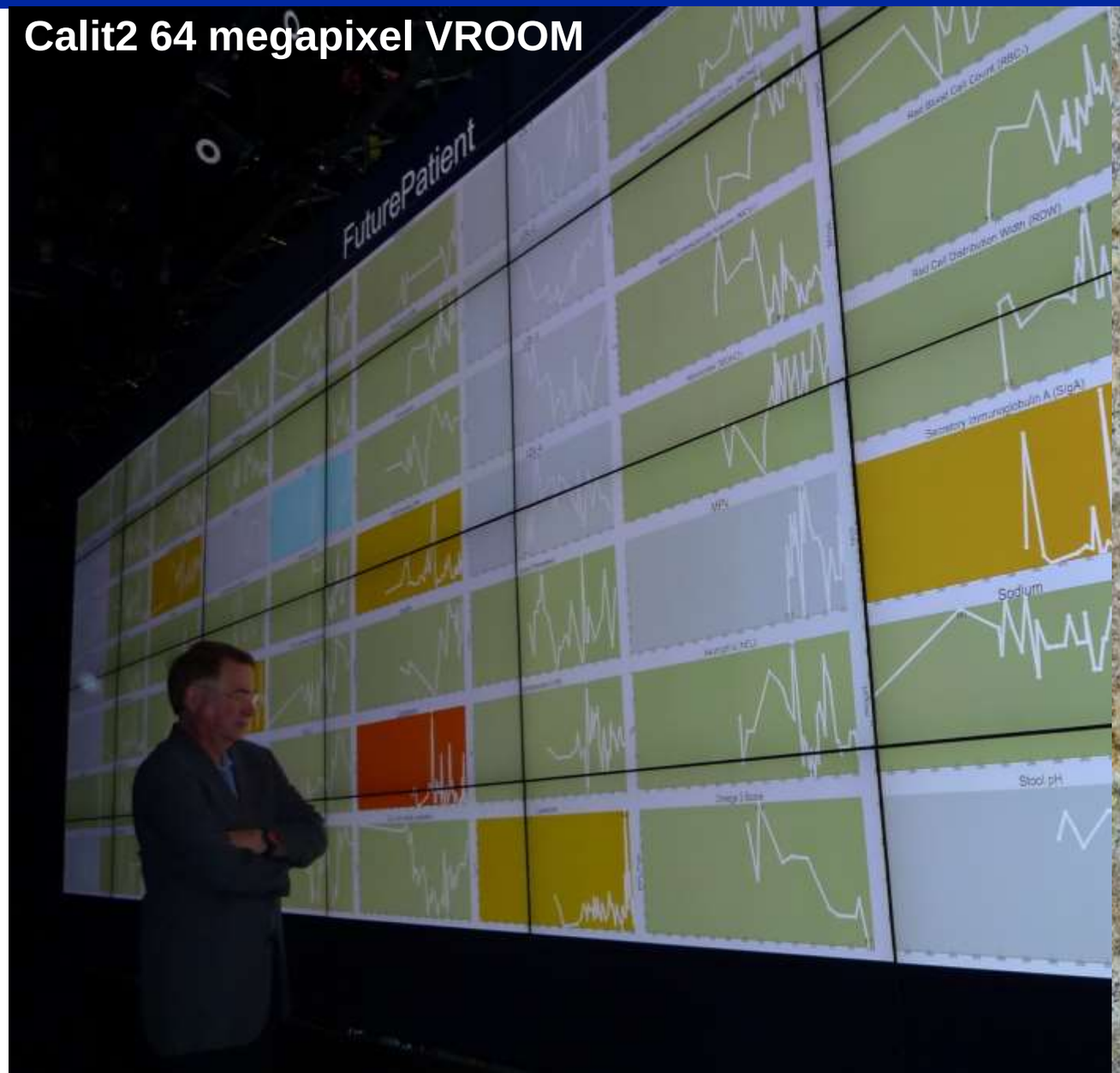
Harry E. Gruber Professor,

Dept. of Computer Science and Engineering

Jacobs School of Engineering, UCSD

http://lsmarr.calit2.net

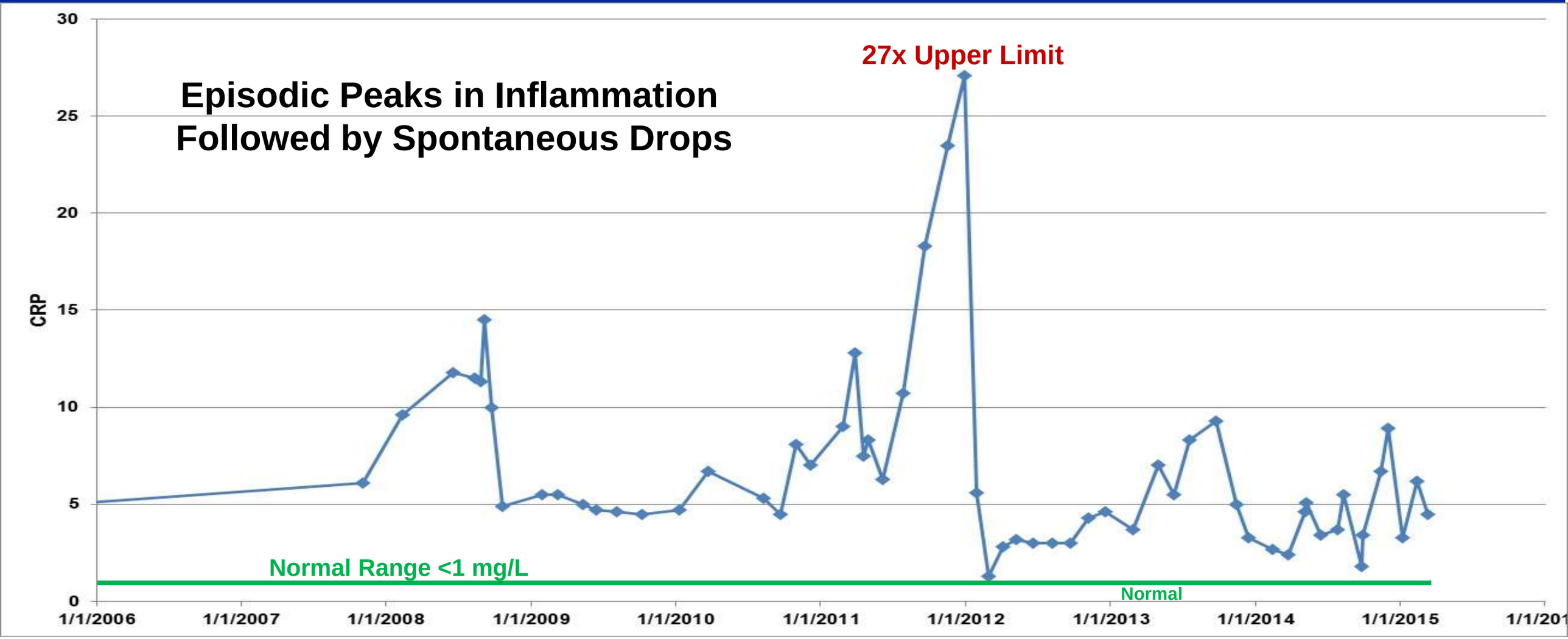
I Discovered I Had IBD By Analyzing 150 Blood and Stool Variables, Each Over 5-10 Years



Calit2 64 megapixel VROOM

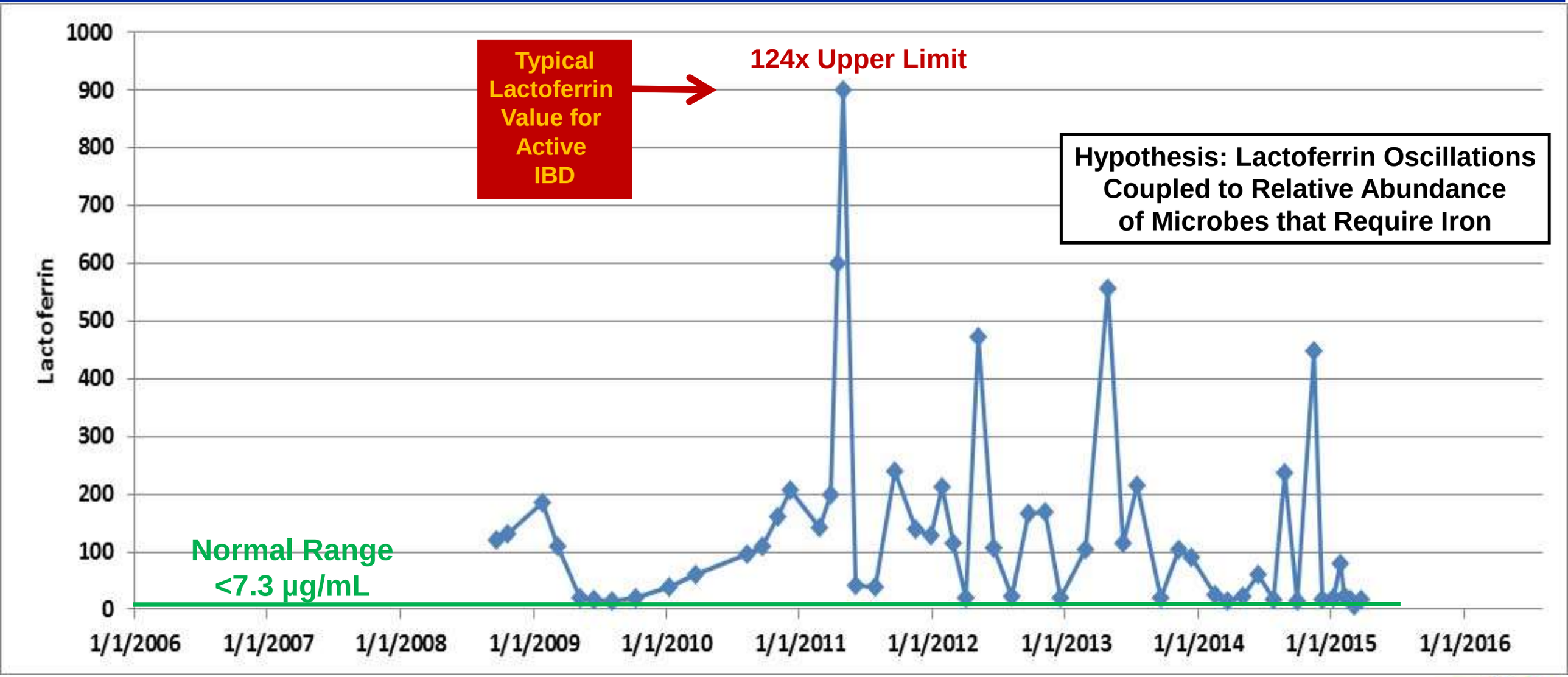

One Blood Draw For Me

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



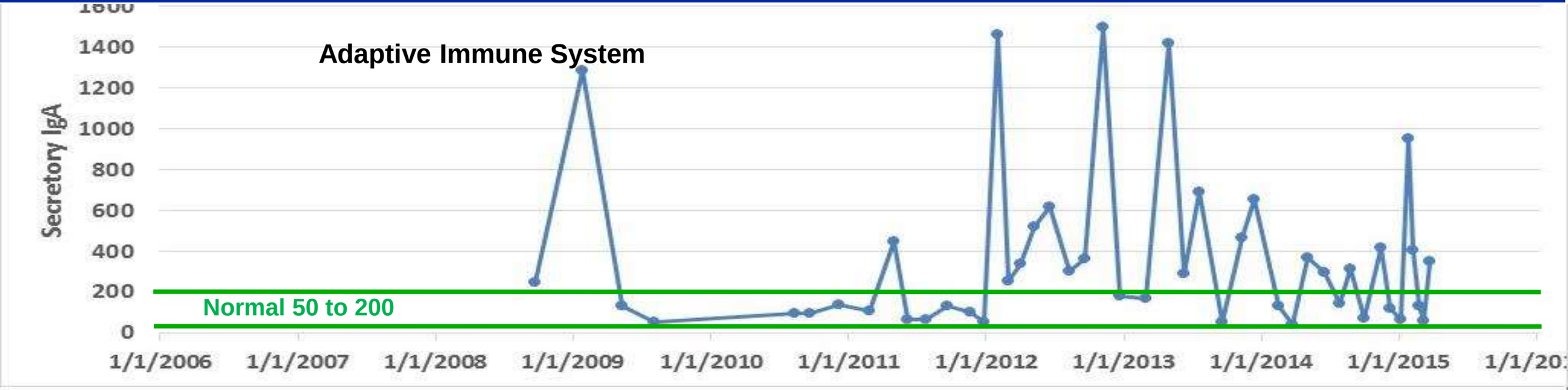
Complex Reactive Protein (CRP) is a Blood Biomarker for Detecting Presence of Inflammation

Adding Stool Tests Revealed A Likelihood of My Having IBD



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

Dynamical Innate and Adaptive Immune Oscillations From Stool Samples



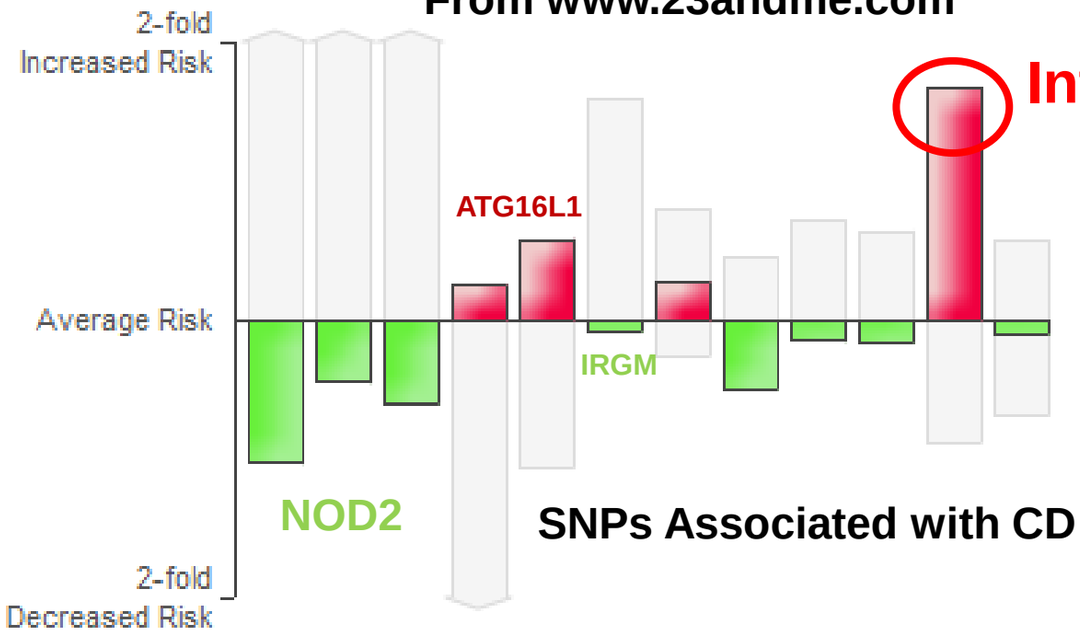


Correlating Immune/Inflammation Time Series With Symptom/Sign, Pharmaceuticals, and Stool Metagenomics Time Series



I Found I Had One of the Earliest Known SNPs Associated with Crohn's Disease

From www.23andme.com



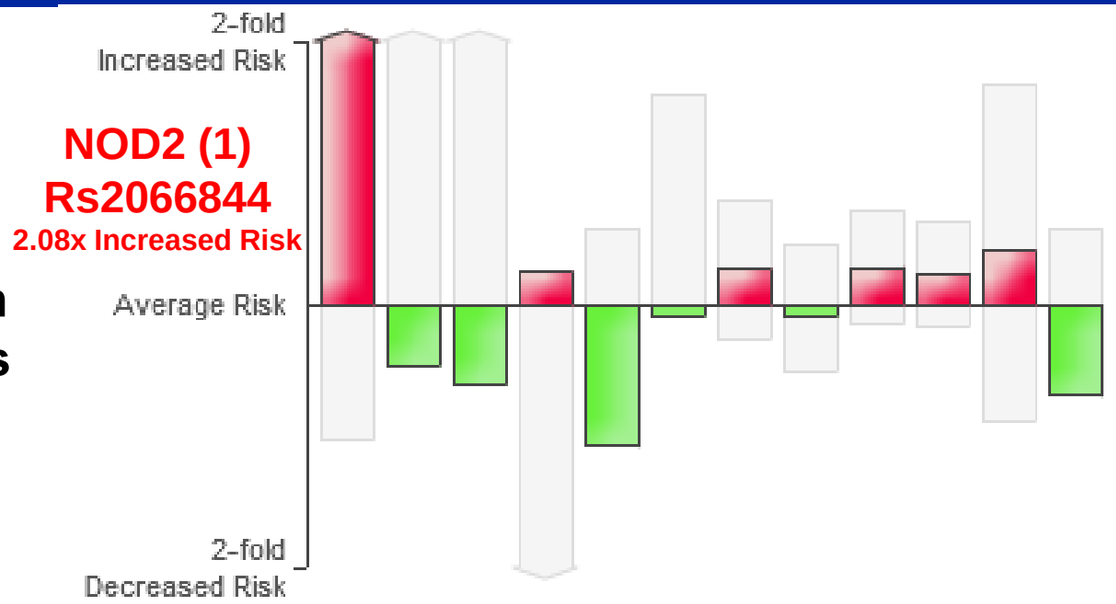
Polymorphism in Interleukin-23 Receptor Gene — 80% Higher Risk of Pro-inflammatory Immune Response rs1004819

A Genome-Wide Association Study Identifies *IL23R* as an Inflammatory Bowel Disease Gene

Richard H. Düerr, *et al.*
Science **314**, 1461 (2006)

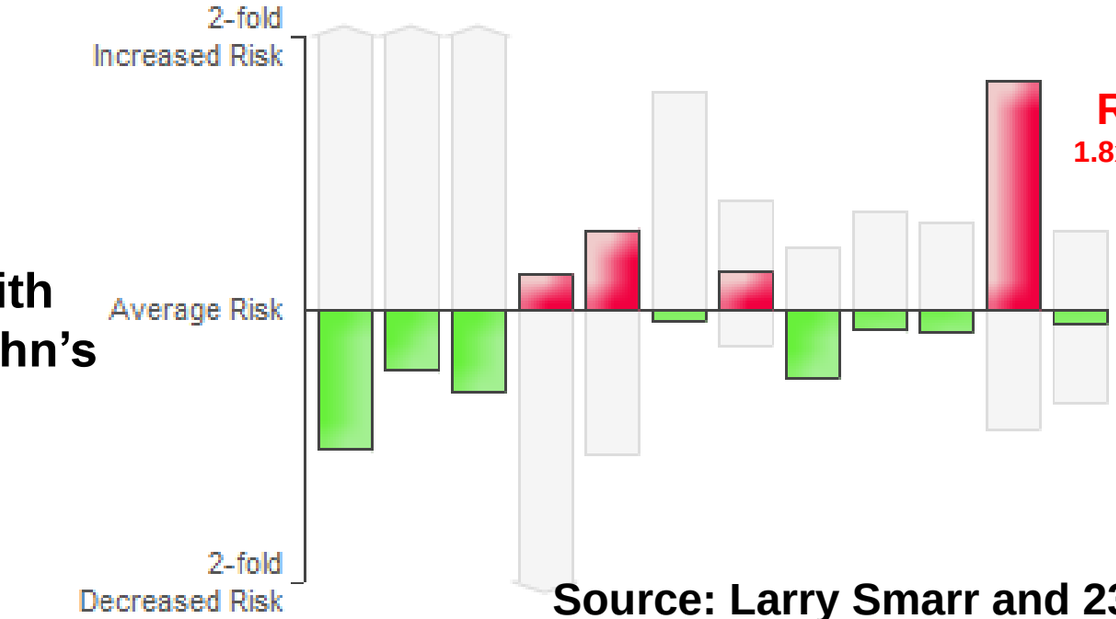
There Is Likely a Correlation Between CD SNPs and Where and When the Disease Manifests

Subject with Ileal Crohn's



Female
CD Onset
At 20-Years Old

Subject with Colonic Crohn's



Me-Male
CD Onset
At 60-Years Old

Source: Larry Smarr and 23andme

A Statistical Study is Needed to Determine If NOD2 and IL23R Are Associated with Different Disease Phenotypes

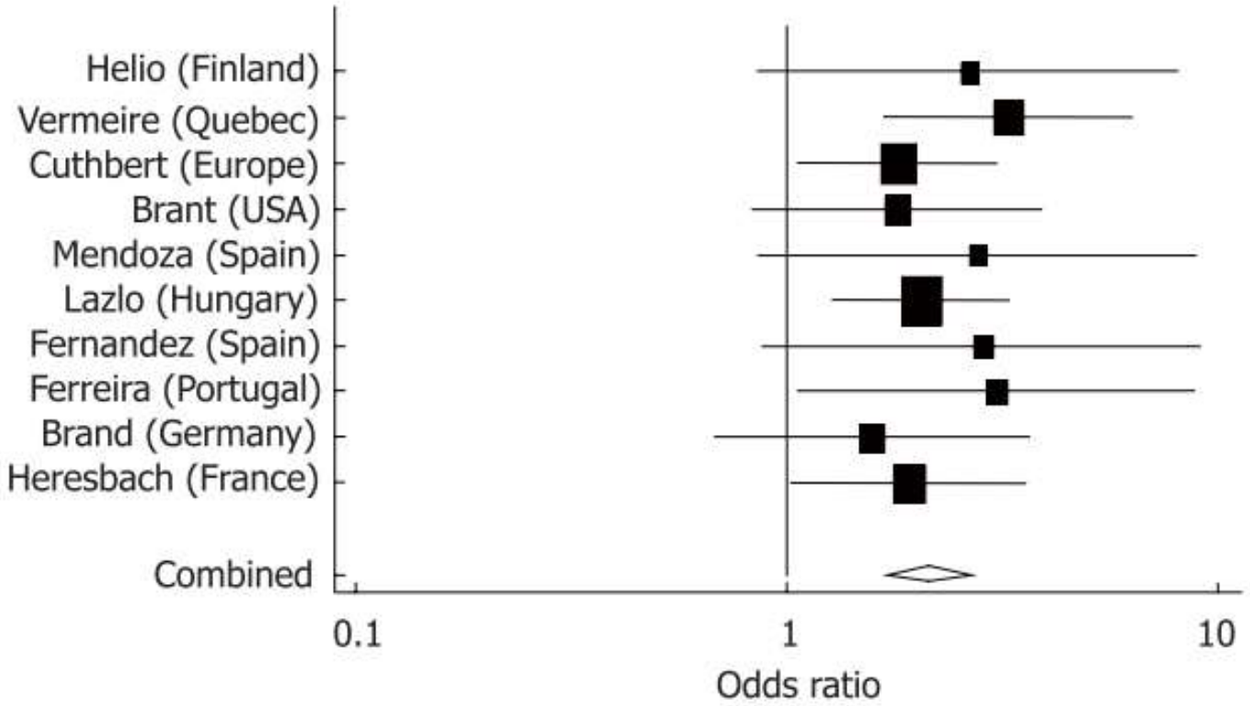


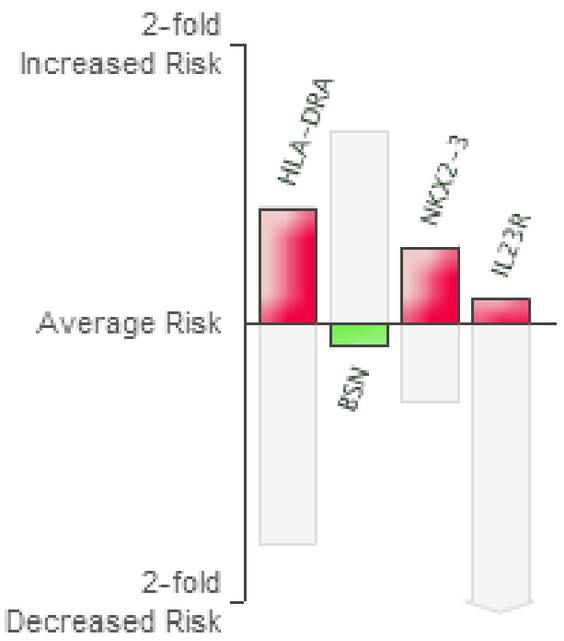
Figure 2 Meta-analysis comparing odds of having a NOD2 variant among Ileocolonic *versus* Colonic CD patients in 10 studies. Pooled estimate = 2.13 (95% CI 1.7-2.7), $P < 0.0001$, test of heterogeneity among studies, $P = 0.927$.

"Associations Between NOD2/CARD15 Genotype and Phenotype in Crohn's Disease-Are We there Yet?," Radford-Smith and Pandeya, World J. of Gastroentrology, 28, 7097-7103 (2006)

I Also Had an Increased Risk for Ulcerative Colitis, But a SNP that is Also Associated with Colonic CD



I Have a 33% Increased Risk for Ulcerative Colitis HLA-DRA (rs2395185)

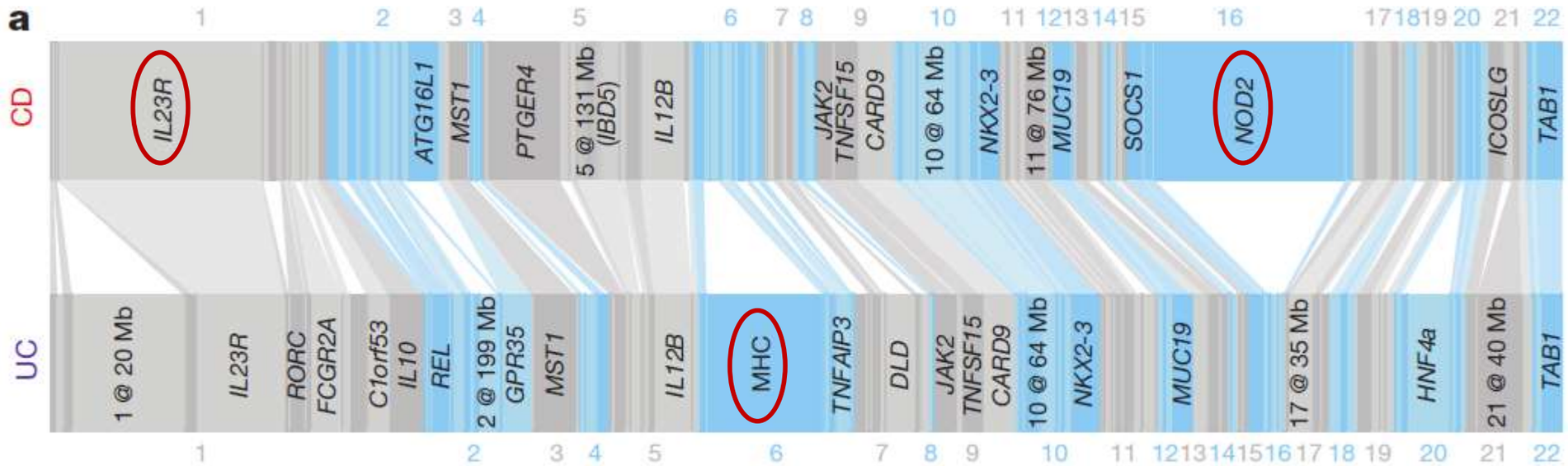


I Have the Same Level of HLA-DRA Increased Risk as Another Male Who Has Had Ulcerative Colitis for 20 Years

"Our results suggest that at least for the SNPs investigated [including HLA-DRA], colonic CD and UC have common genetic basis."
-Waterman, et al., IBD 17, 1936-42 (2011)

So IBD May be Stratified by a Personalized Combination of the 163 Known SNPs Associated with IBD

The Current Division of IBD Into Crohn's Disease and Ulcerative Colitis May Turn Out to be Superseded by a More Accurate Human Genetic Stratification



- The width of the bar is proportional to the variance explained by that locus
- Bars are connected together if they are identified as being associated with both phenotypes
- Loci are labelled if they explain more than 1% of the total variance explained by all loci

"Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease," Jostins, et al. Nature 491, 119-124 (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

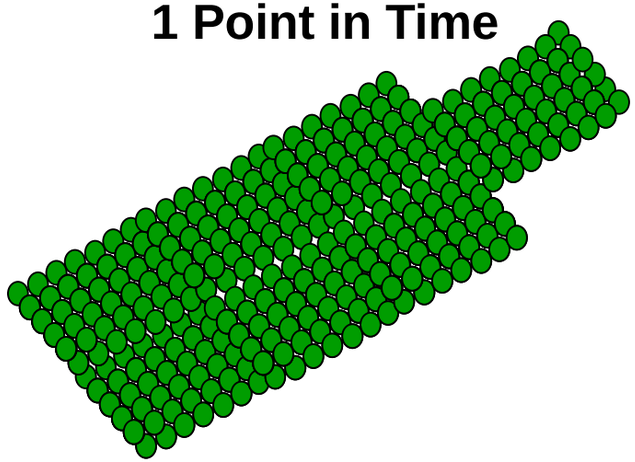
J. Craig Venter INSTITUTE SDSC DELL NIH HMP CRBS it²

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

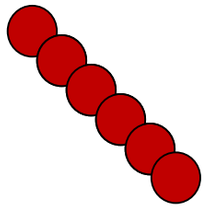
250 Subjects
1 Point in Time



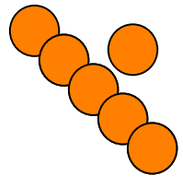
Inflammatory Bowel Disease (IBD) Patients

Larry Smarr
(Colonic Crohn's)

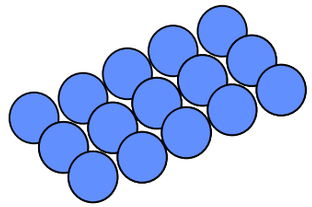
7 Points in Time



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**



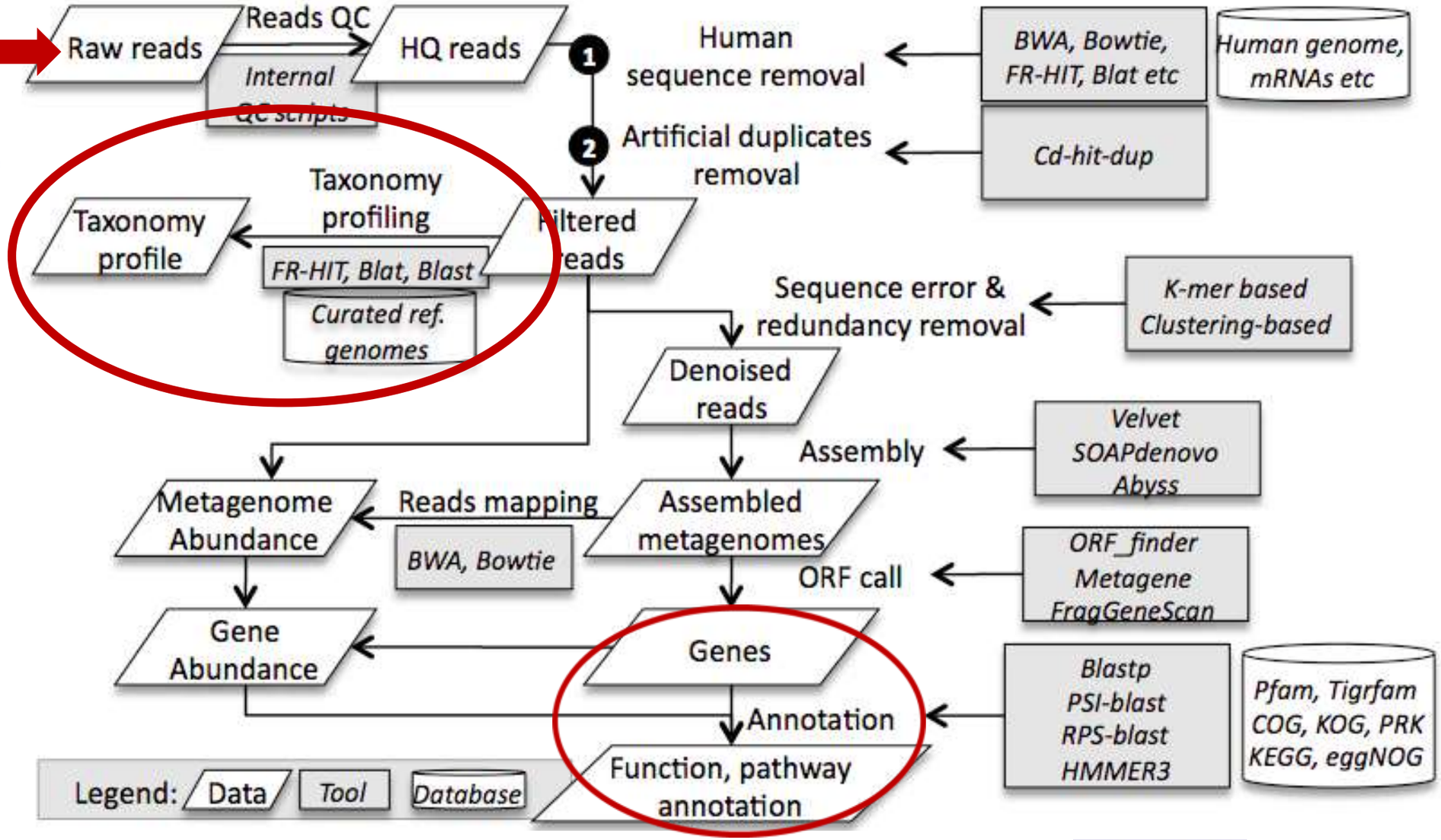
Now to Align Our 27 Billion Reads Against the Reference Database

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)

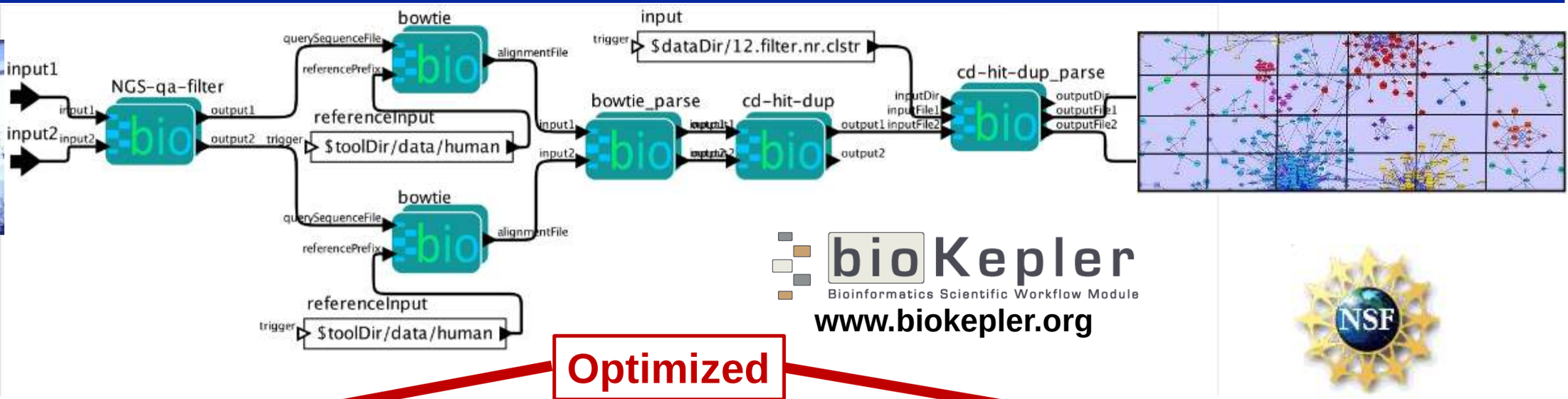
CRBS

it²

Next Step

Programmability, Scalability and Reproducibility using bioKepler





bioKepler
Bioinformatics Scientific Workflow Module
www.biokepler.org

NSF

Source: Ilkay Altintas, SDSC

Optimized

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

WorDS

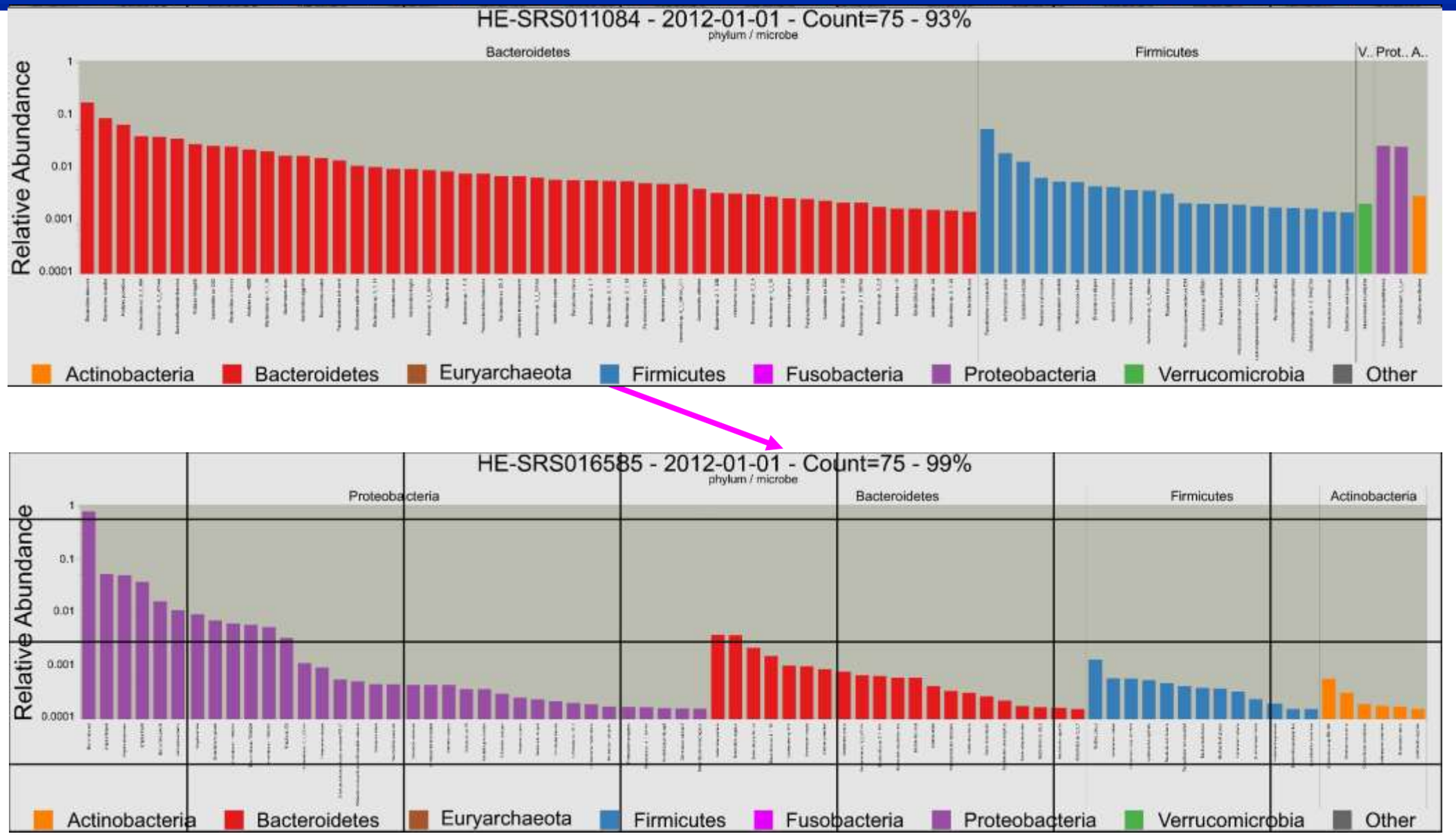
Kepler

www.kepler-project.org

SDSC

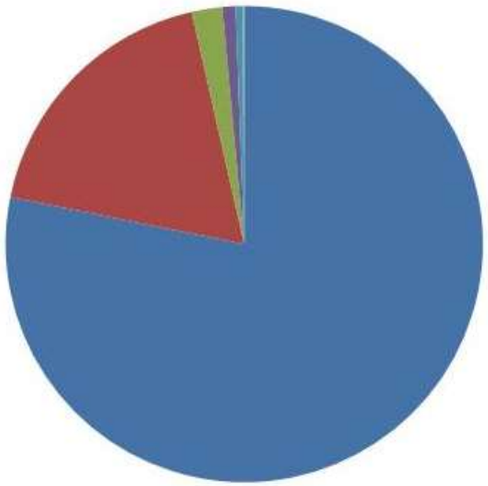
it2

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

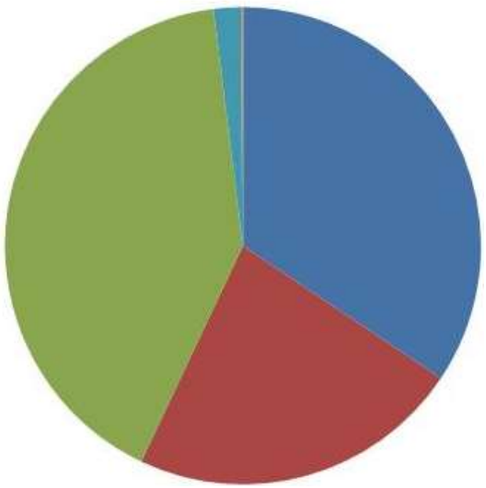
Average HE



Most Common Microbial Phyla

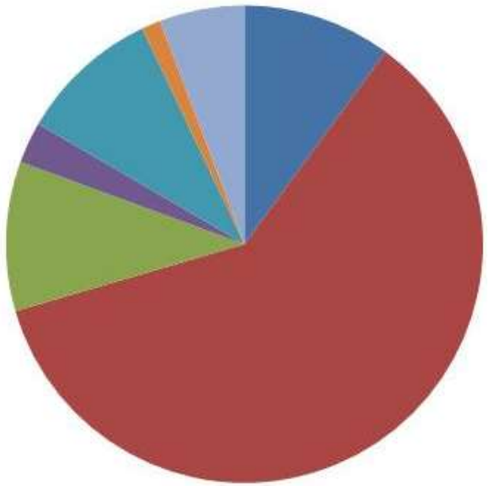
- Bacteroidetes
- Firmicutes
- Proteobacteria
- Verrucomicrobia
- Actinobacteria
- Fusobacteria
- Euryarchaeota

Average Ulcerative Colitis



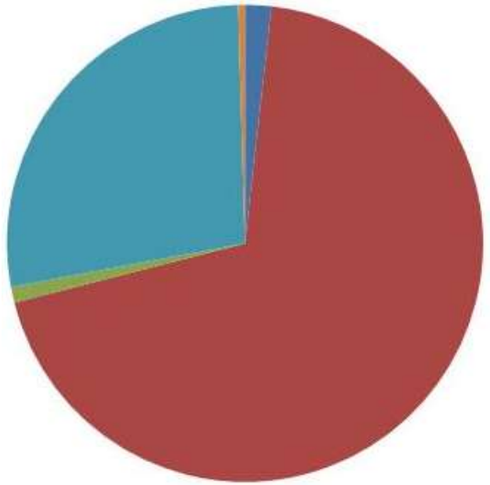
Explosion of Proteobacteria

Average LS Colonic Crohn's Disease



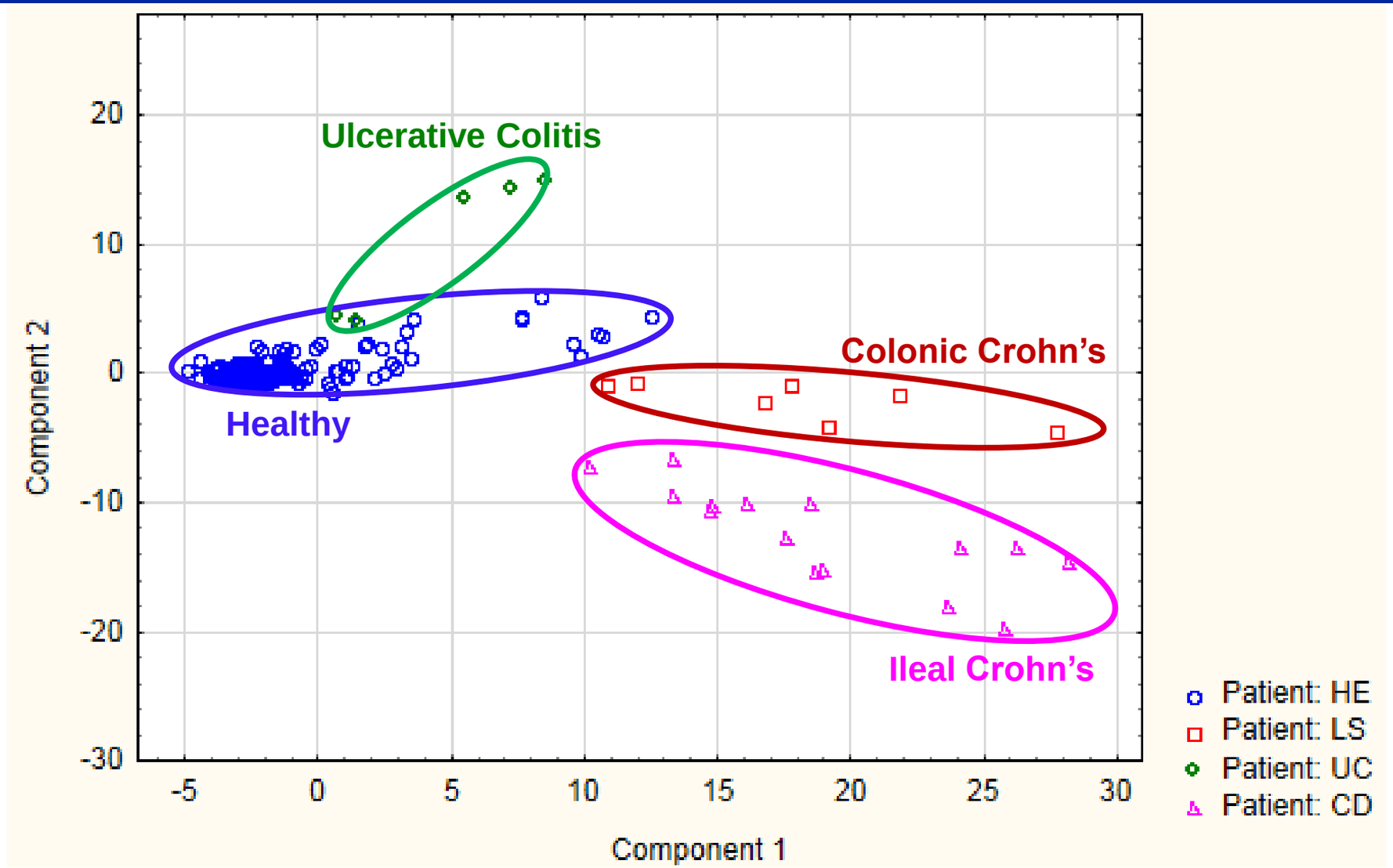
Hybrid of UC and CD
High Level of Archaea

Average Ileal Crohn's Disease



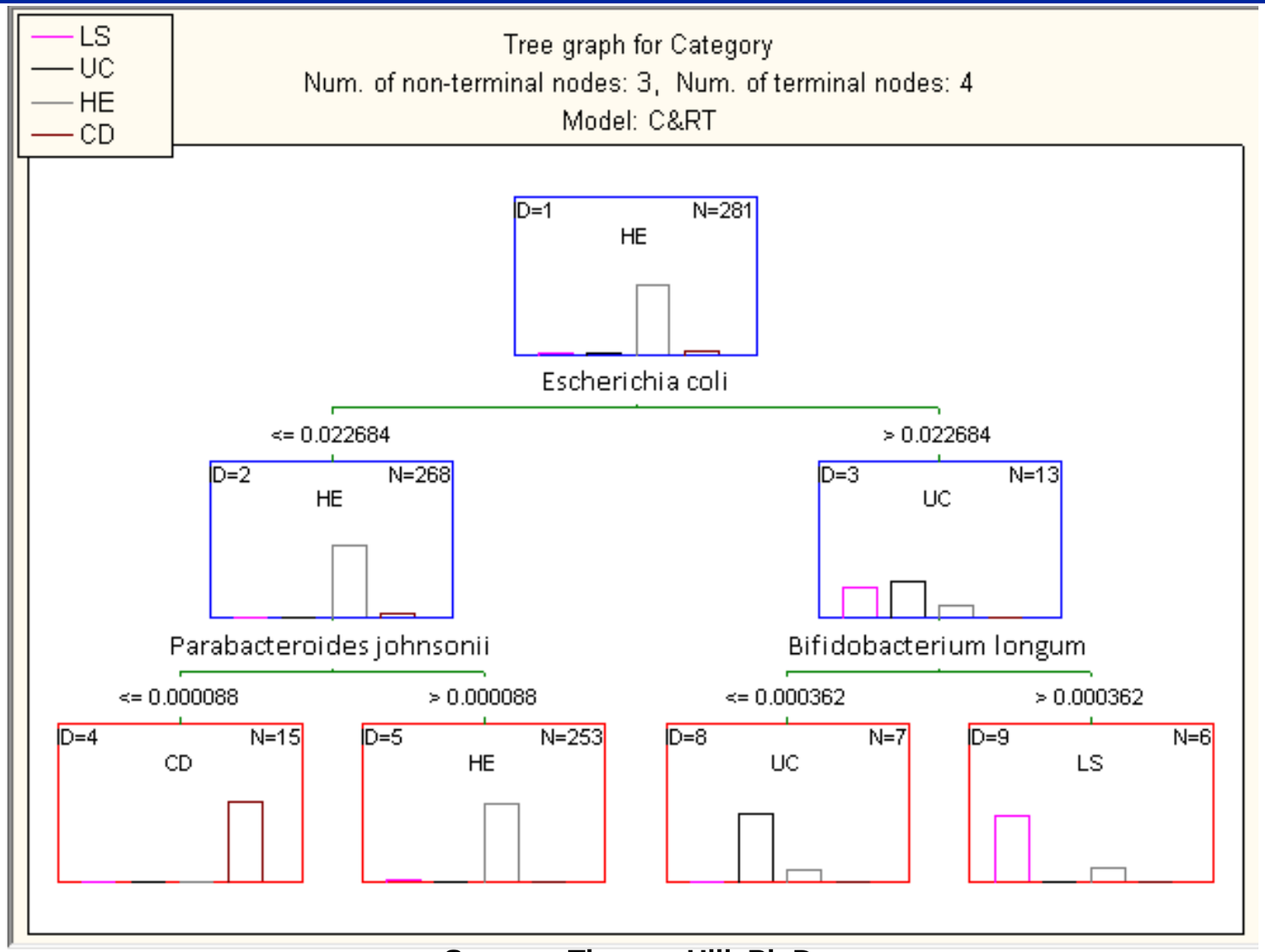
Collapse of Bacteroidetes
Explosion of Actinobacteria

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



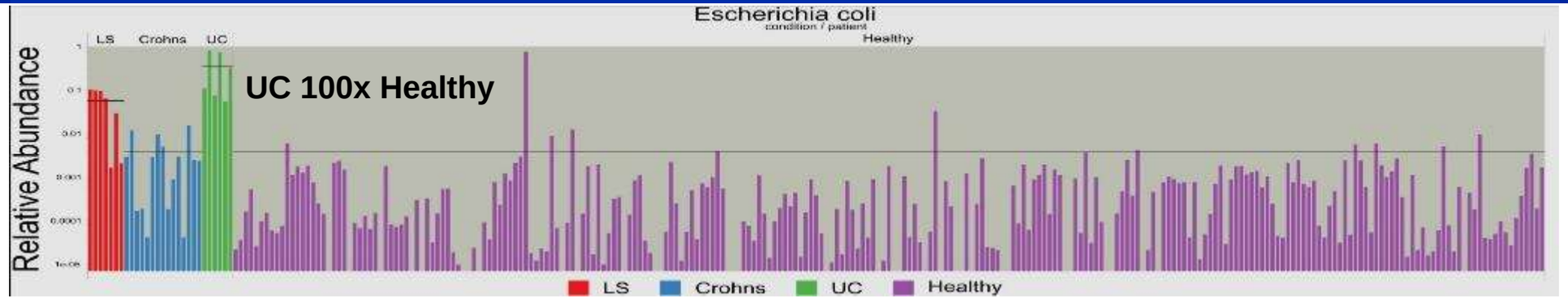
Source: Thomas Hill, Ph.D.
Executive Director Analytics
Dell | Information Management Group, Dell Software

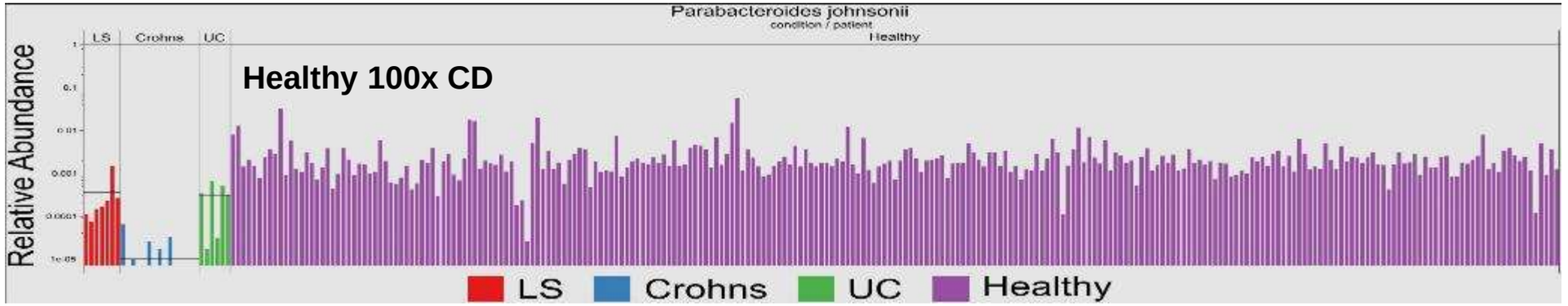
Dell Analytics Tree Graphs Classifies the 4 Health/Disease States With Just 3 Microbe Species

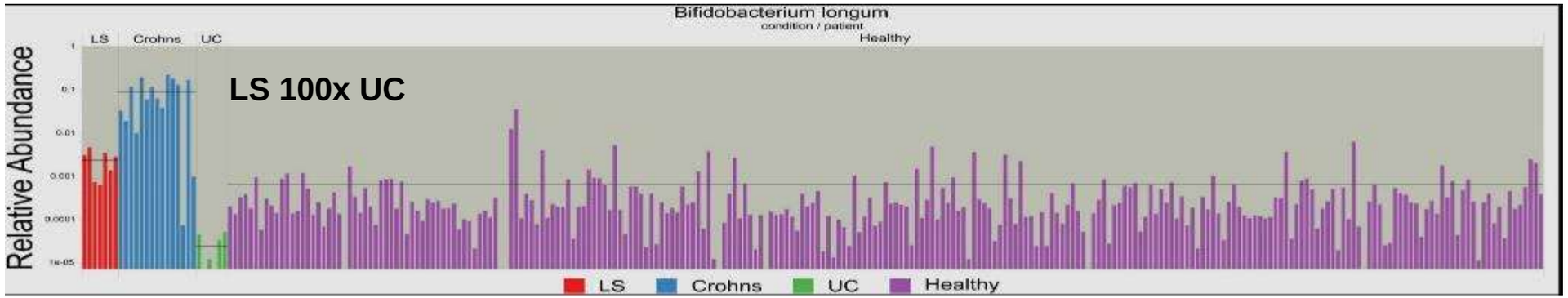


Source: Thomas Hill, Ph.D.
Executive Director Analytics
Dell | Information Management Group, Dell Software

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

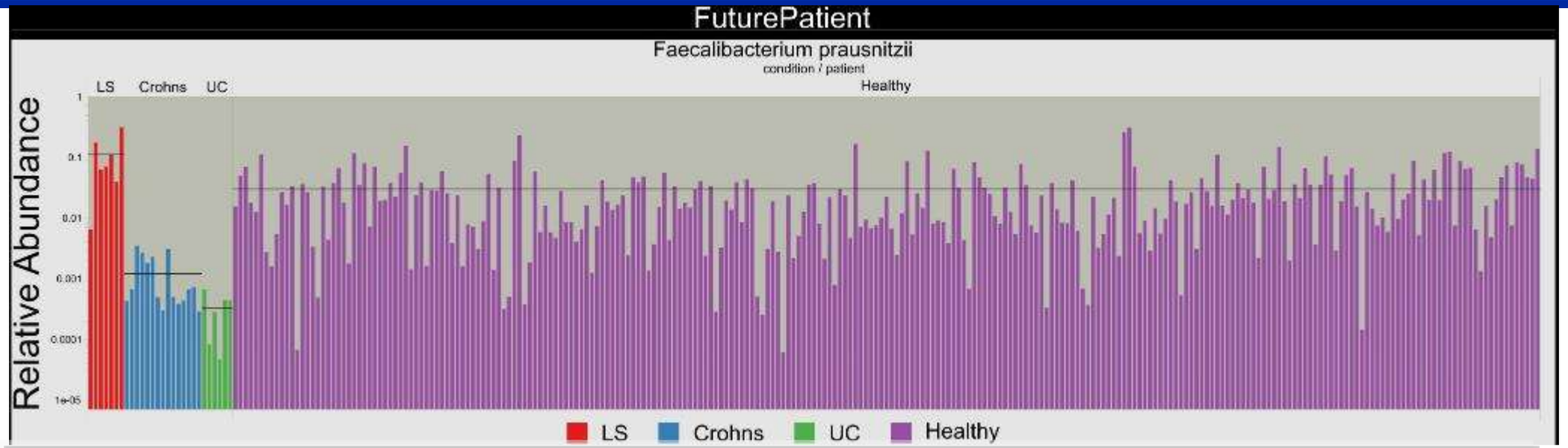






We Produced Similar Results for ~2500 Microbial Species

Ileal Crohn's and UC Patients Have Reduced Abundance of Anti-Inflammatory *Faecalibacterium prausnitzii*



Decreased abundance of *Faecalibacterium prausnitzii* in the gut microbiota of Crohn's disease

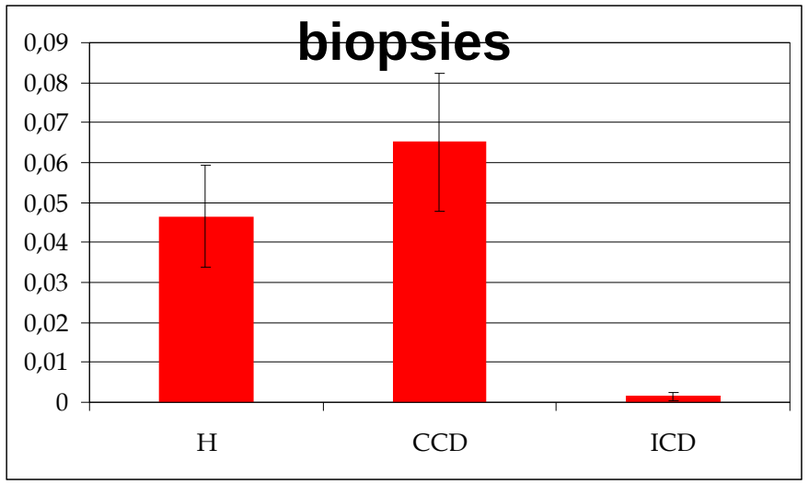
Journal of Gastroenterology and Hepatology **28** (2013) 613–619

Takehide Fujimoto,* Hirotsugu Imaeda,† Kenichiro Takahashi,* Eiji Kasumi,* Shigeki Bamba,* Yoshihide Fujiyama† and Akira Andoh*

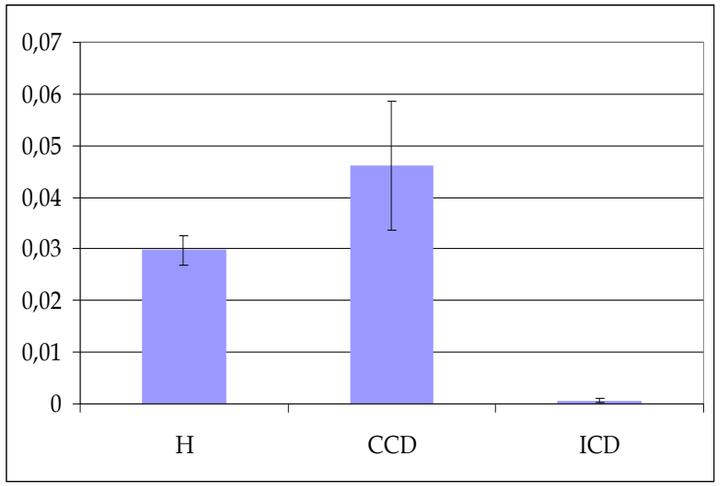
However, Colonic Crohn's (LS) Have *Increased* Abundance

A Noninvasive Diagnostic?? - Faecalibacterium is Depleted in Ileal CD and Increased in Colonic CD

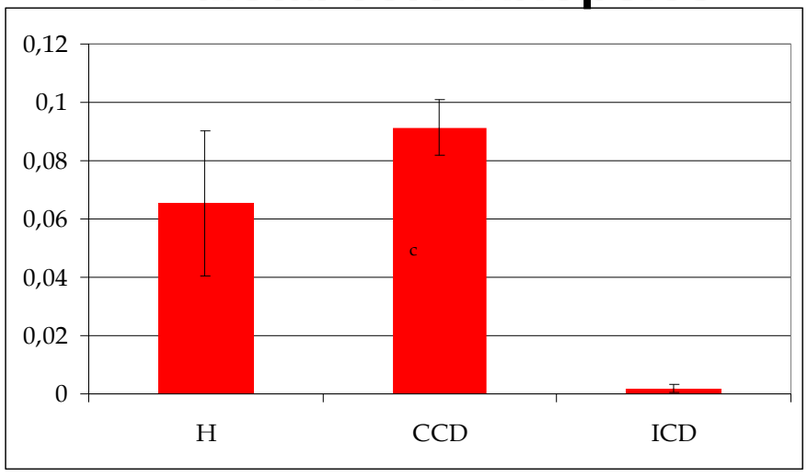
ileum biopsies



feces



distal colon biopsies



Willing et al., 2009.Inflammatory Bowel Diseases

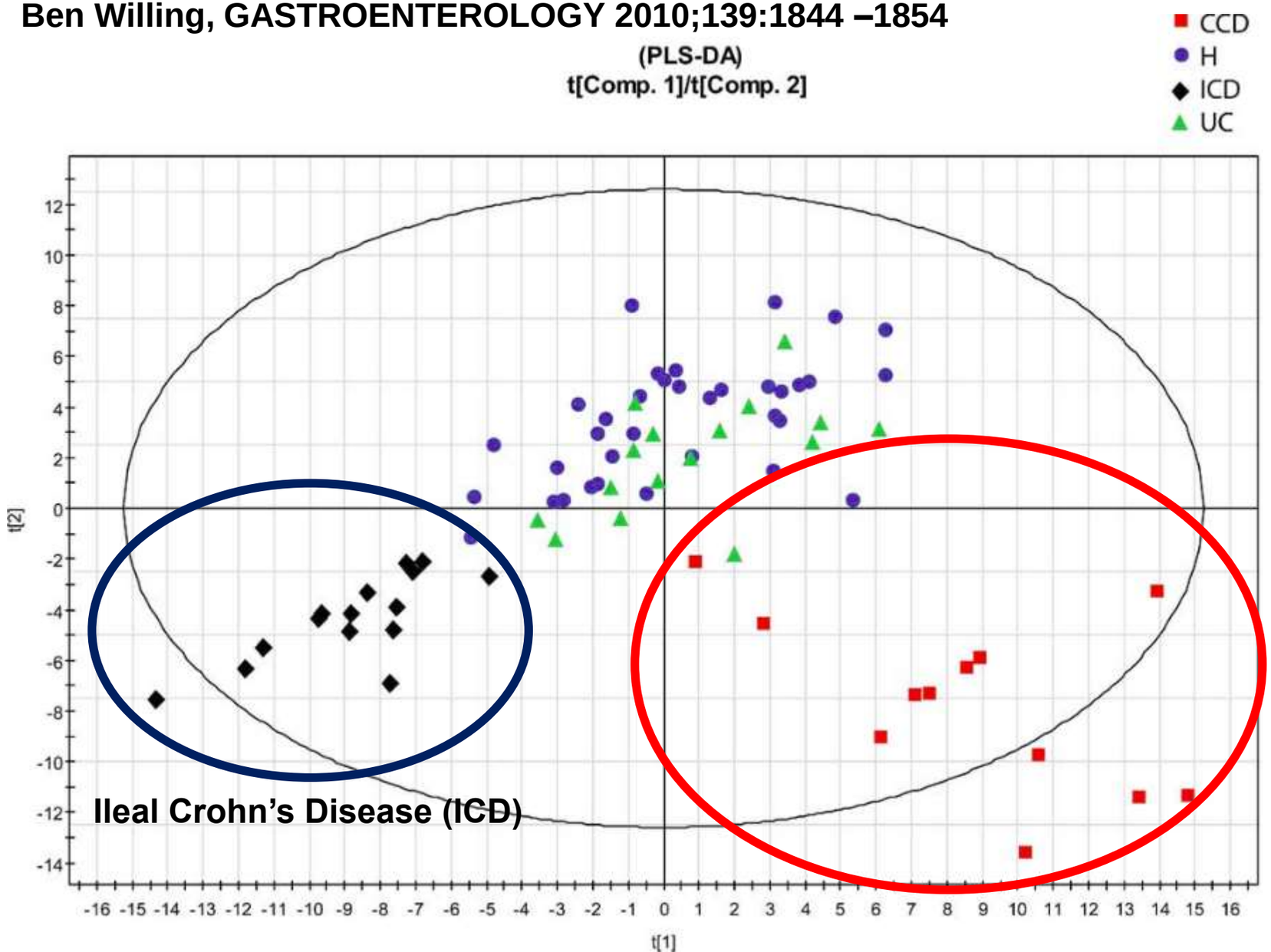
Faecalibacterium prausnitzii

One of the main producers of butyrate Important for colonic health.

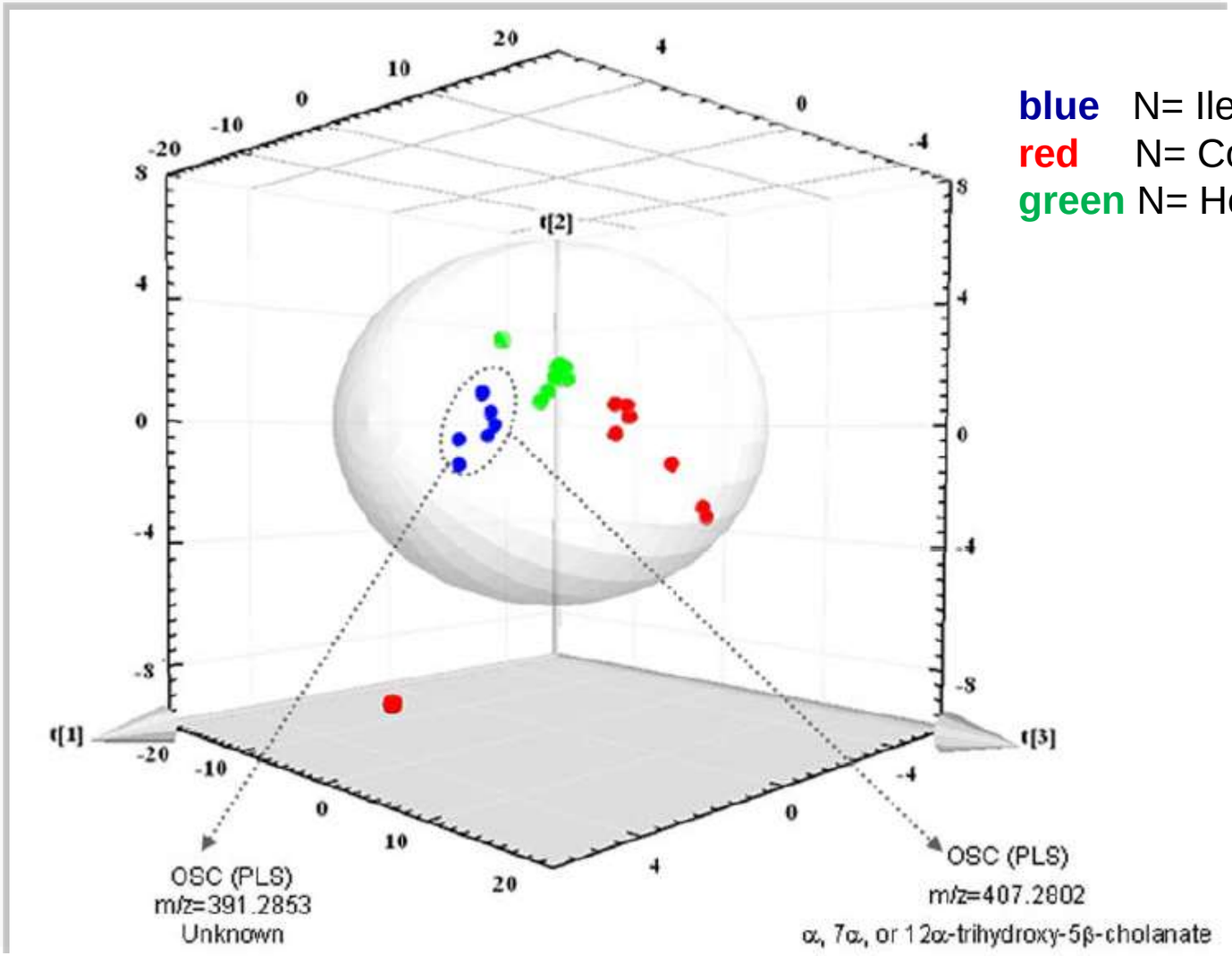
Slide from Janet Jansson, PNNL

Is the Gut Microbial Ecology Different in Crohn's Disease Subtypes?

Ben Willing, GASTROENTEROLOGY 2010;139:1844 –1854



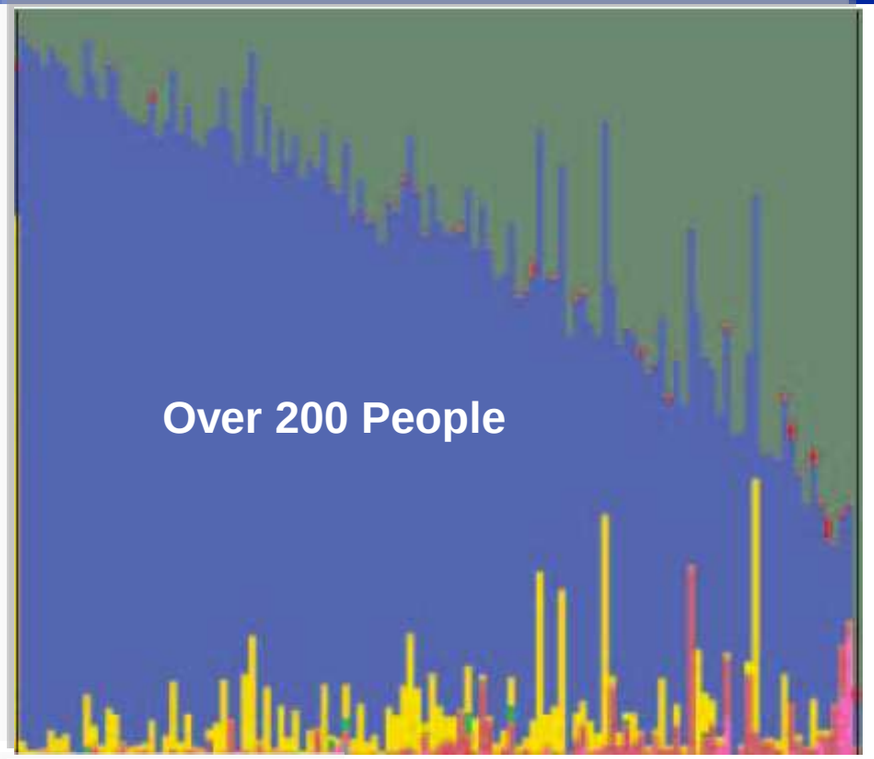
It Appears That Metabolomics Can Differentiate Ileum vs. Colon Inflammation in Crohn's Disease

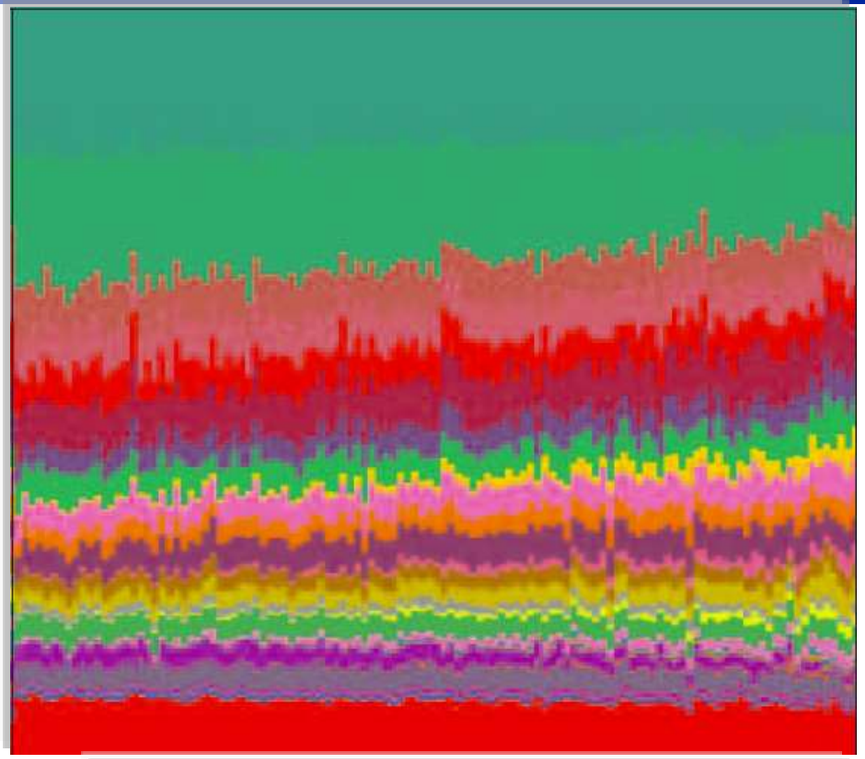


blue N= Ileum (ICD)
red N= Colon (CCD)
green N= Healthy

Jansson, et al. PLOS ONE, July 2009 | Volume 4 | Issue 7 | e6386

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





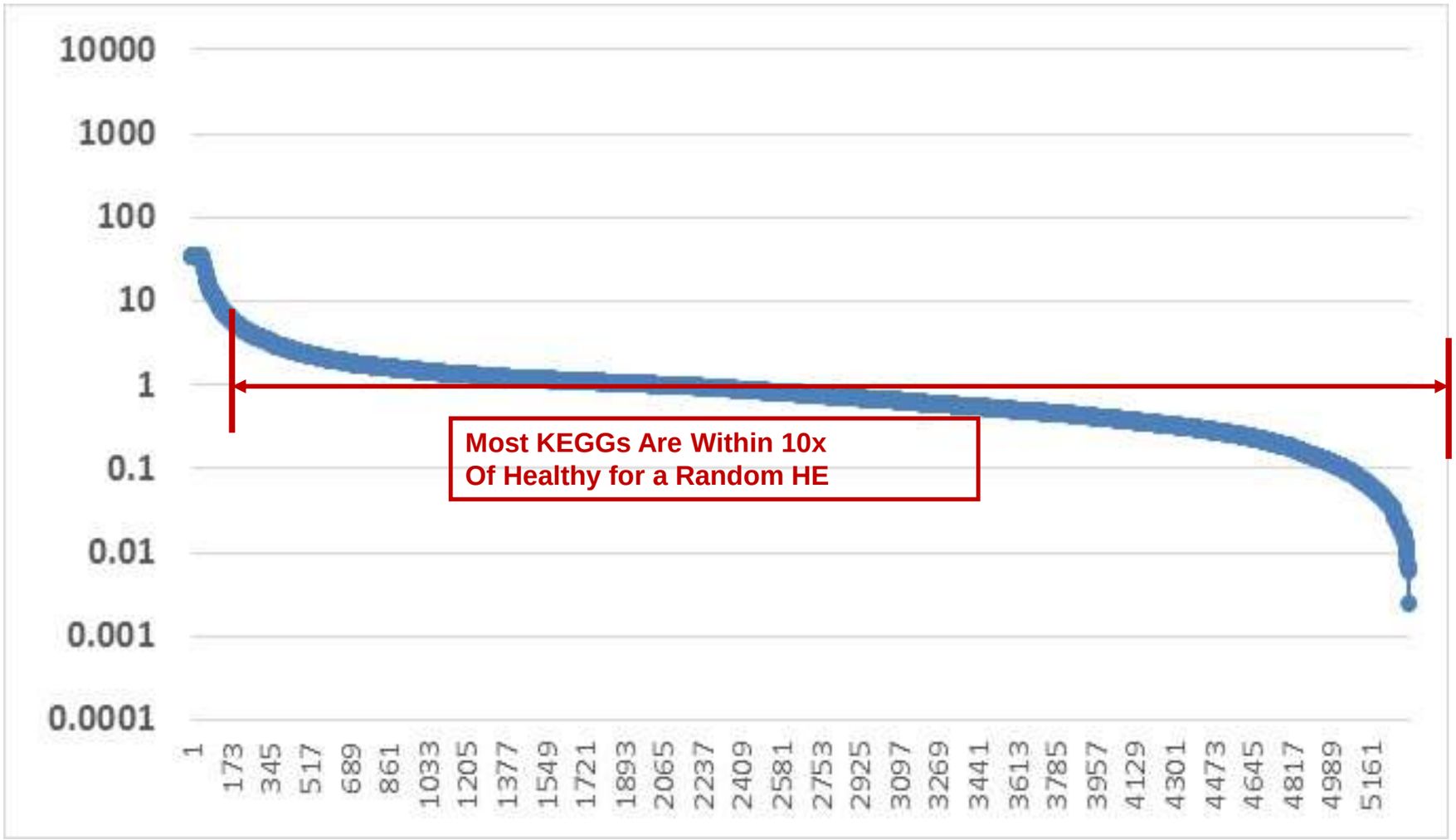
- Firmicutes
- Actinobacteria
- Bacteroidetes
- Proteobacteria
- Fusobacteria
- Tenericutes
- Spirochaetes
- Cyanobacteria
- Verrucomicrobia
- TM7

- 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

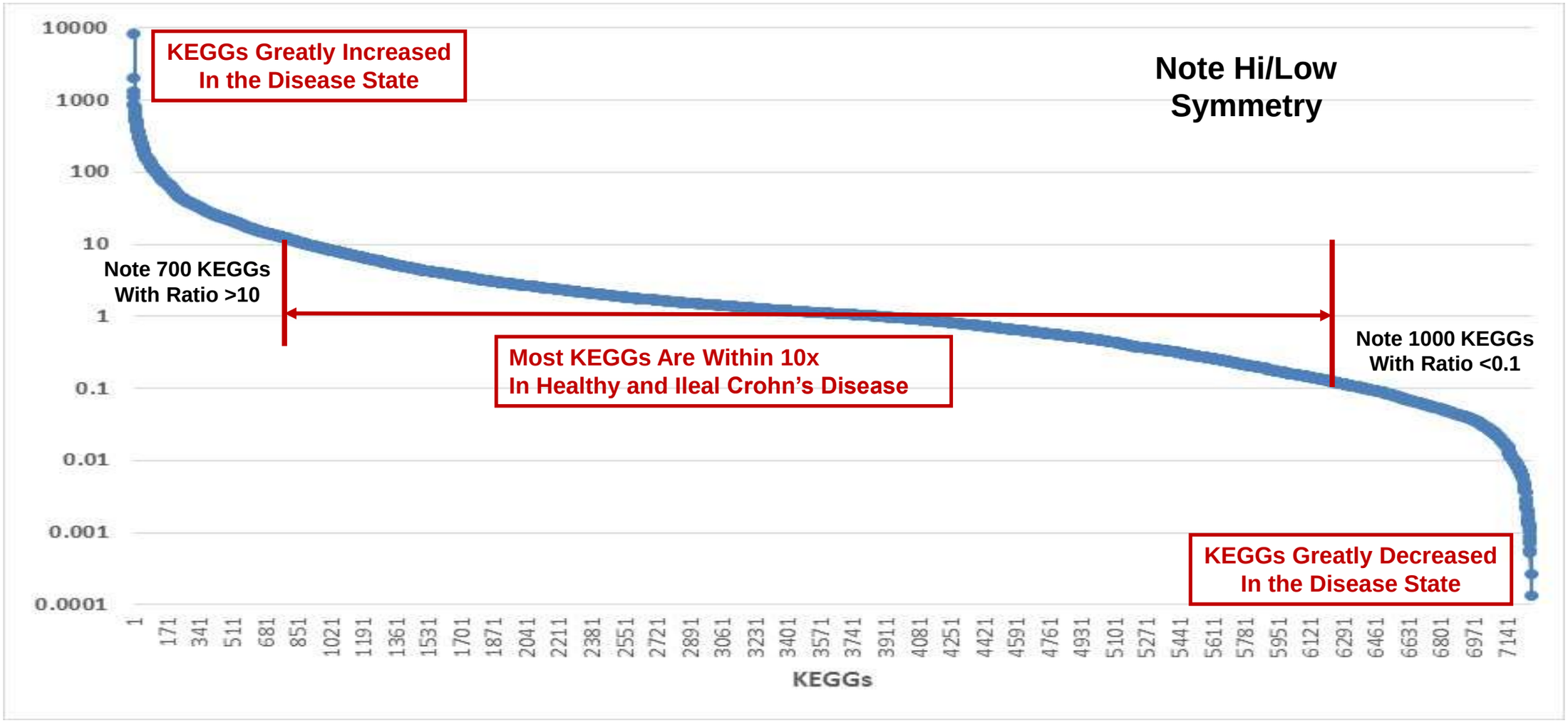
Source: Nature, 486, 207-212 (2012)

Ratio of One of the Healthy Subjects to the Average KEGG for 35 Healthy: Test to see How Much Inter-Personal Variation There is Within Healthy

We Computed the Relative Abundance of 10,000 KEGGs in 35 Healthy And 25 IBD Patients



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



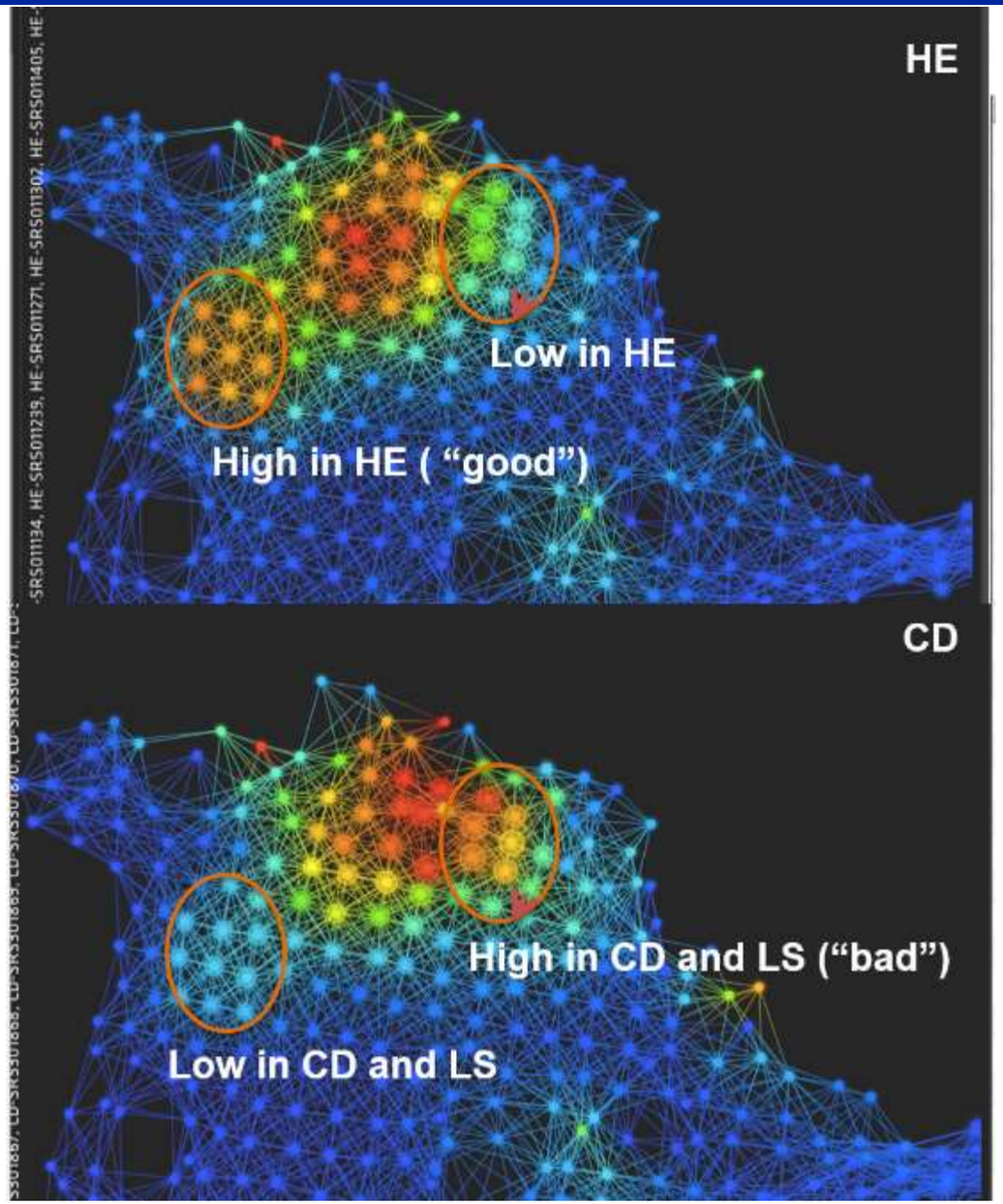
Over 7000 KEGGs Which Are Nonzero in Health and Disease States

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

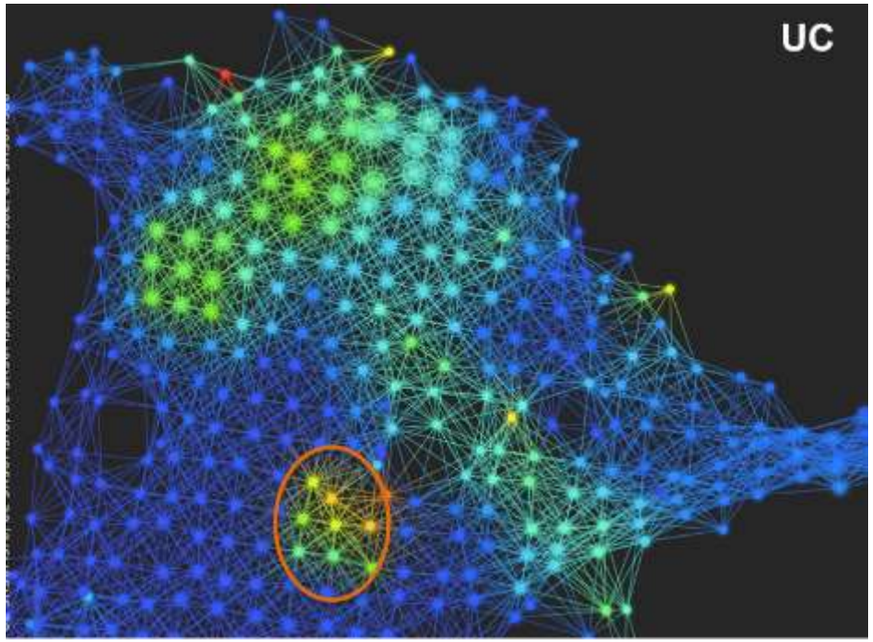
- **Look for KEGGs That Have the Properties:**
  - **Are 100x in All Four Disease States**
    - **LS001/Ave HE**
    - **Ave CD/ Ave HE**
    - **Ave UC/Ave HE**
    - **Sick HE Person/Ave HE**
- **There are 48 of These Extreme KEGGs (see spreadsheet)**

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

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

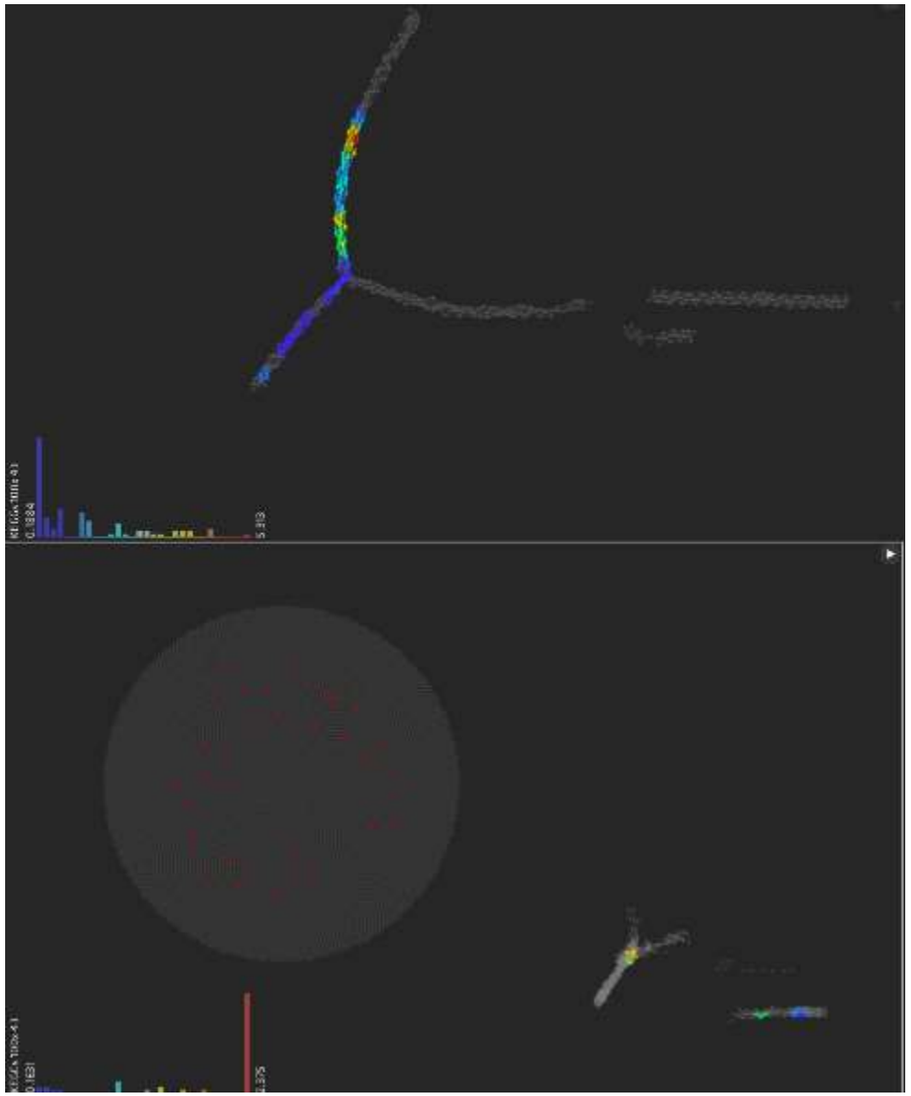


Source: Pek Lum,
Formerly Chief Data Scientist, Ayasdi



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)

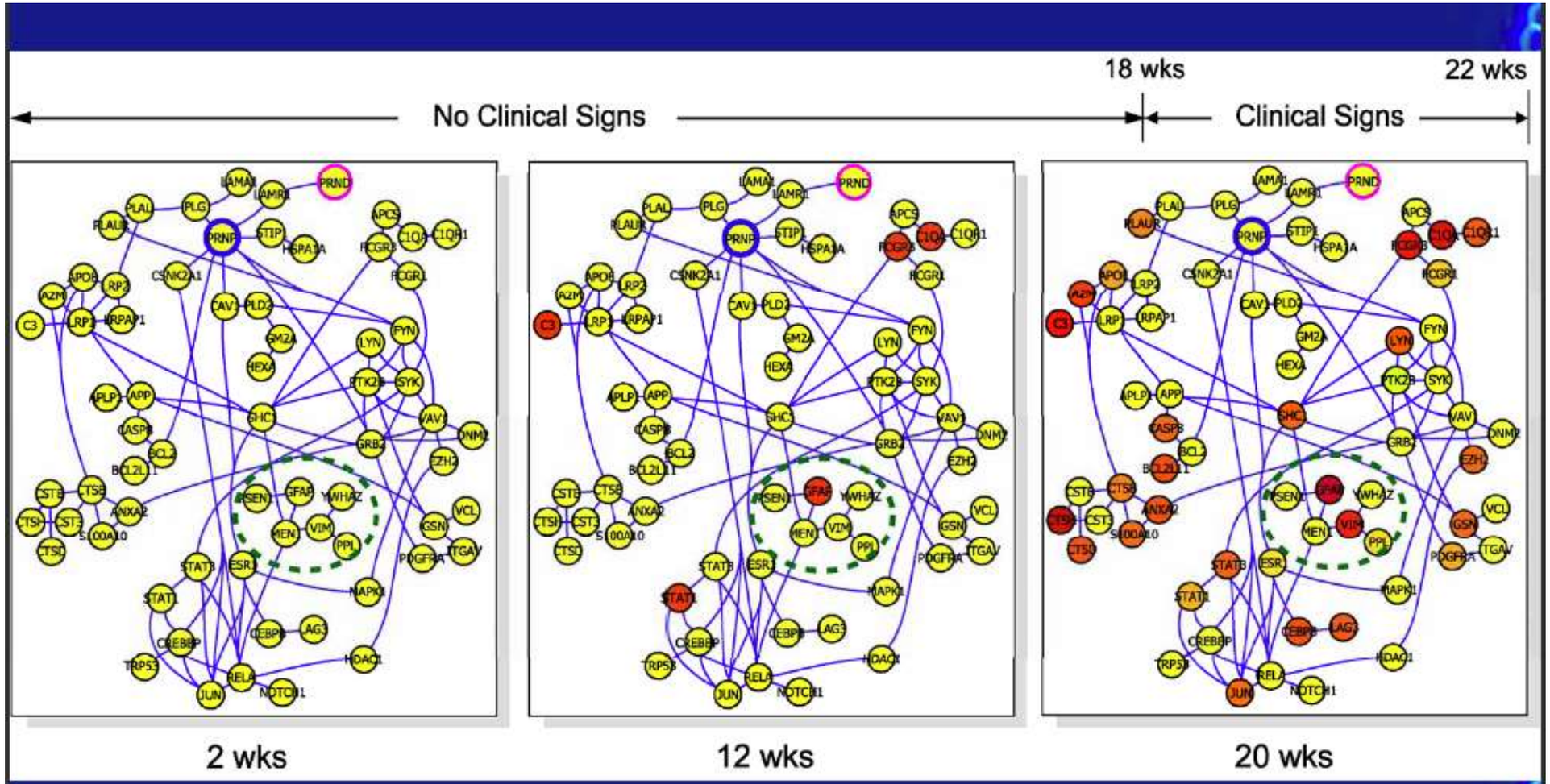
K00108(choline_dehydrogenase)
K00673(arginine_N-succinyltransferase)
K00867(type_I_pantothenate_kinase)
K01169(ribonuclease_I_(enterobacter_ribonuclease))
K01484(succinylarginine_dihydrolase)
K01682(aconitate_hydratase_2)
K01690(phosphogluconate_dehydratase)
K01825(3-hydroxyacyl-CoA_dehydrogenase_/_enoyl-CoA_hydratase_/3-hydroxybutyryl-CoA_epimerase_/_e
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()

Ayasdi Cure™
Turn Data into Therapies

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

Next Step: Compute Genes and Function For All ~300 People's Gut Microbiome

Full Processing to Function:
Genes & Protein Families
(COGs, KEGGs)

Would Require
~1-2 Million
Core-Hours

Source: Weizhong Li, UCSD; Brian Kucic, R Systems

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

Announced November 7, 2014!

Inflammatory Bowel Disease Biobank For Healthy and Disease Patients

Already 185 Enrolled,
Goal is 1500

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

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



JCVI Team
Karen Nelson
Shibu Yooseph
Manolito Torralba

SDSC Team
Michael Norman
Ilkay Altintas
Shweta Purawat
Mahidhar Tatineni
Robert Sinkovits

Dell/R Systems and Dell Analytics
Brian Kucic
John Thompson
Tom Hill

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