“Quantifying Your Superorganism Body Using Big Data Supercomputing”

ACM International Workshop on Big Data in Life Sciences

BigLS 2014

Newport Beach, CA

September 20, 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 this "superorganism" is comprised of hundreds of species spread over many taxonomic phyla. The human immune system is tightly coupled with this microbial ecology and in cases of autoimmune disease, both the immune system and the microbial ecology can have dynamic excursions far from normal. Our research starts with trillions of DNA bases, produced by Illumina Next Generation sequencers, of the human gut microbial DNA taken from my own body, as well as from hundreds of people sequenced under the NIH Human Microbiome Project. To decode the details of the microbial ecology we feed this data into parallel supercomputers, running sophisticated bioinformatics software pipelines. We then use Calit2/SDSC designed Big Data PCs to manage the data and drive innovative scalable visualization systems to examine the complexities of the changing human gut microbial ecology in health and disease. Finally, I will show how advanced data analytics tools find patterns in the resulting microbial distribution data that suggest new hypotheses for clinical application.

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

Newsweek

A Doctor's Vision of the Future of Medicine

Leroy Hood
NEWSWEEK

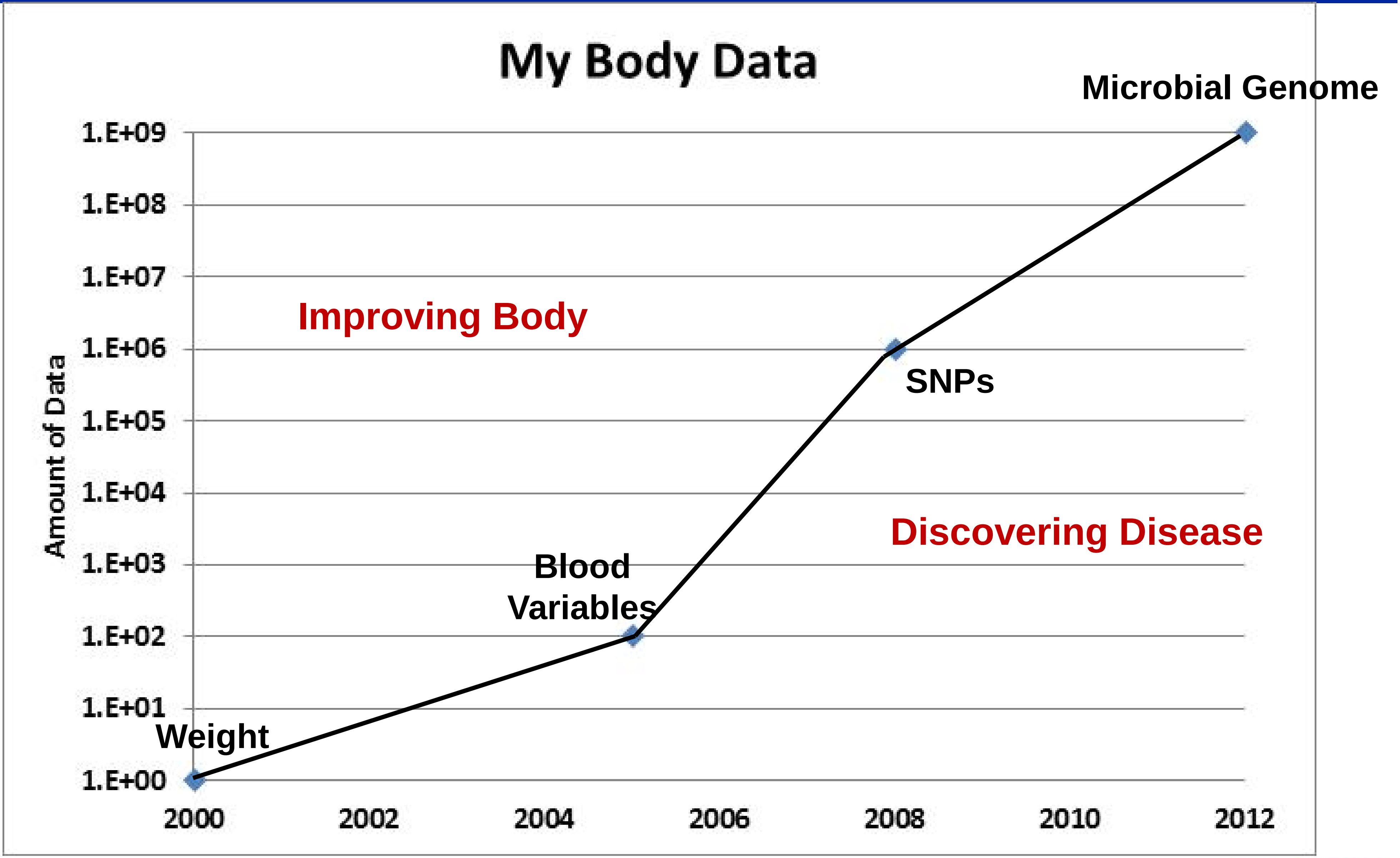
From the magazine issue dated Jul 13, 2009

PIONEER 100
HUNDRED PERSON WELLNESS PROJECT
Institute for Systems Biology

Will Grow to 1000, then 10,000

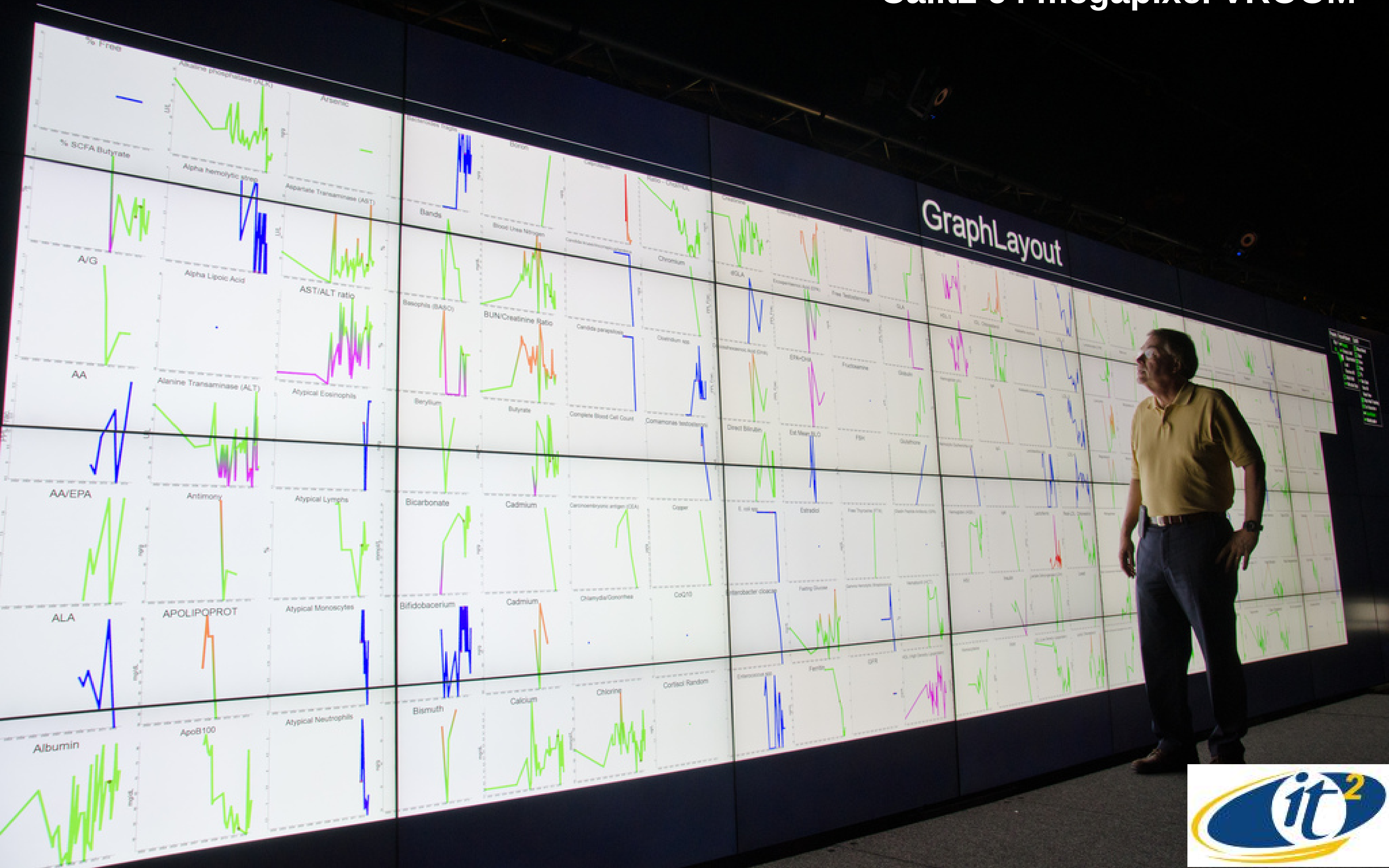
www.newsweek.com/2009/06/26/a-doctor-s-vision-of-the-future-of-medicine.html

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

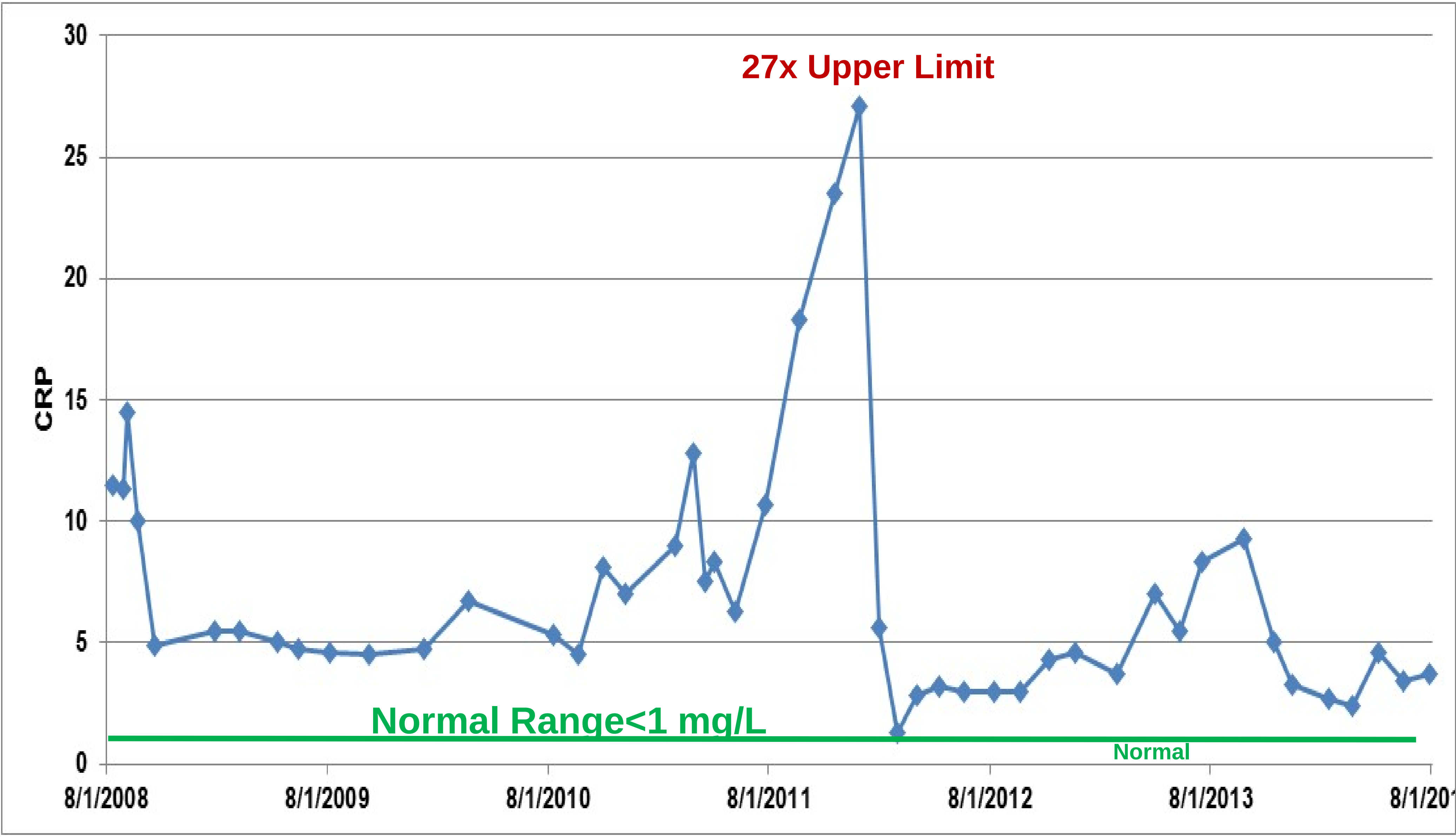


Visualizing Time Series of 150 LS Blood and Stool Variables, Each Over 5-10 Years

Calit2 64 megapixel VROOM

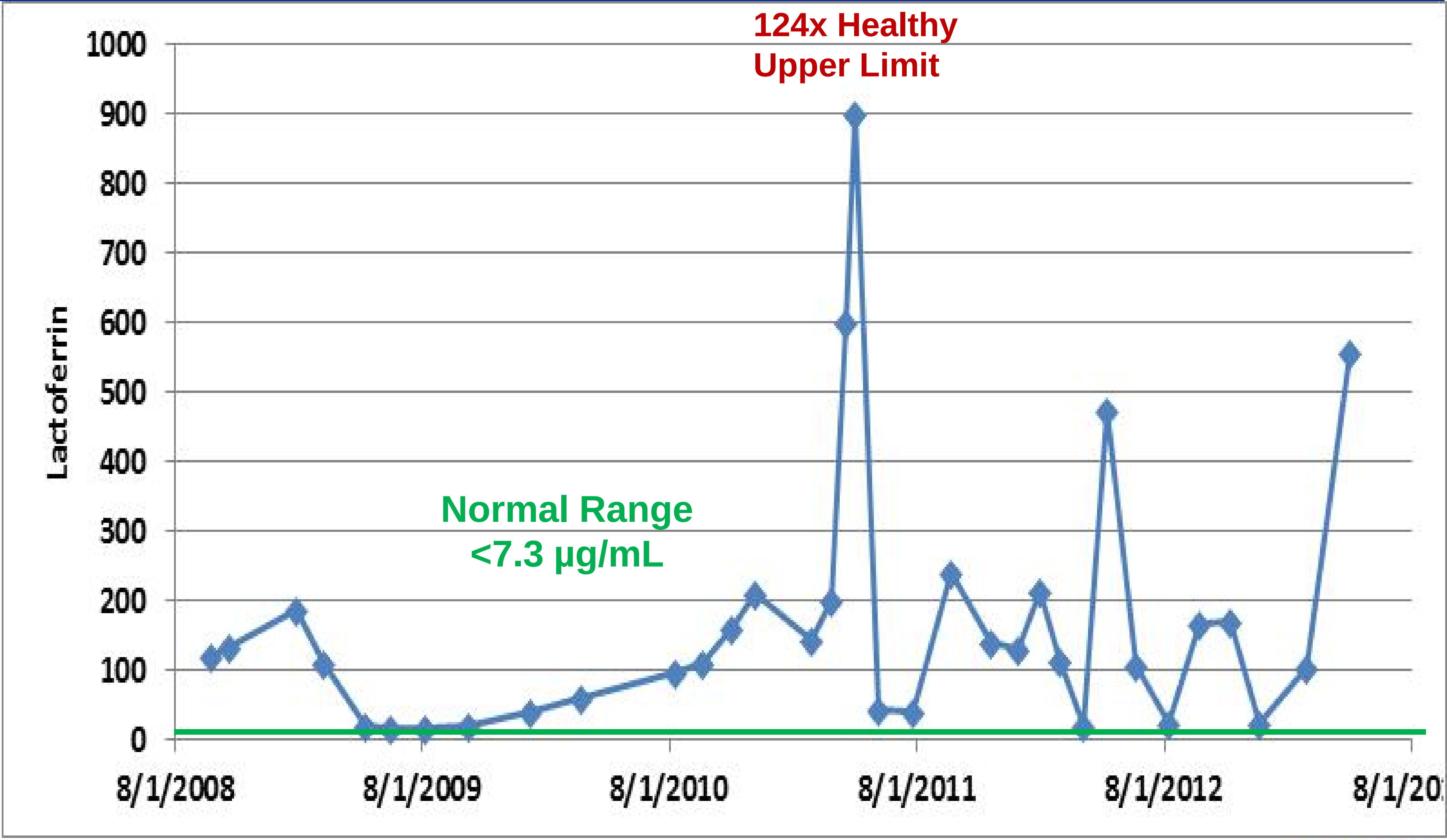


One of My Blood Measurements
Was Far Out of Range--Indicating Chronic Inflammation



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

Stool Samples Revealed
Episodic Autoimmune Response



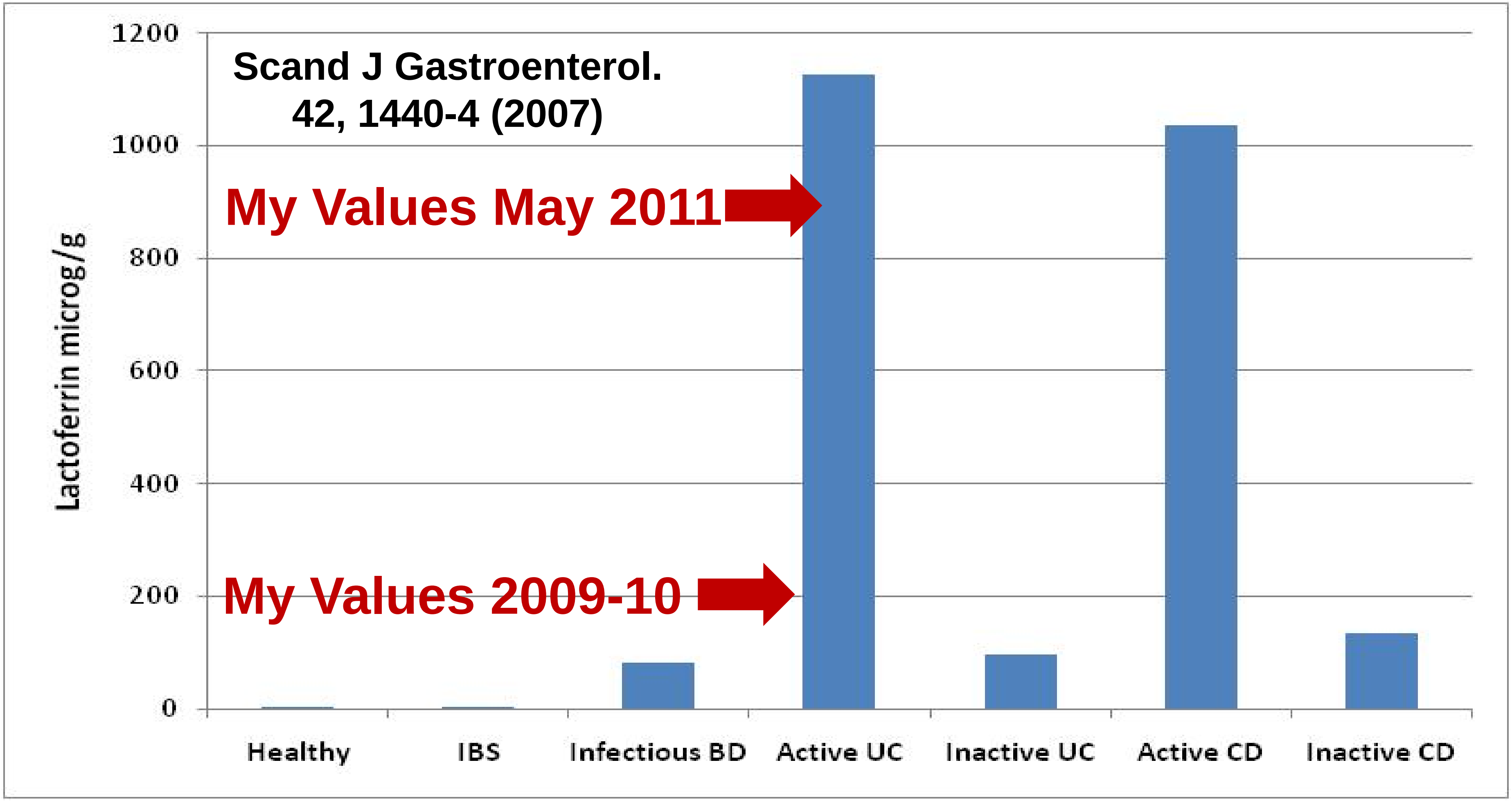
124x Healthy
Upper Limit

Normal Range
<7.3 µg/mL

Lactoferrin is an Antibacteria Glycoprotein
Shed from Attacking WBC Neutrophils

High Lactoferrin Biomarker Led Me to Hypothesis I Had Inflammatory Bowel Disease (IBD)

IBD is an Autoimmune Disease Which Comes in Two Subtypes: Crohn's and Ulcerative Colitis



High Level of Calprotectin Confirmed Hypothesis

Why Did I Have an Autoimmune Disease like IBD?

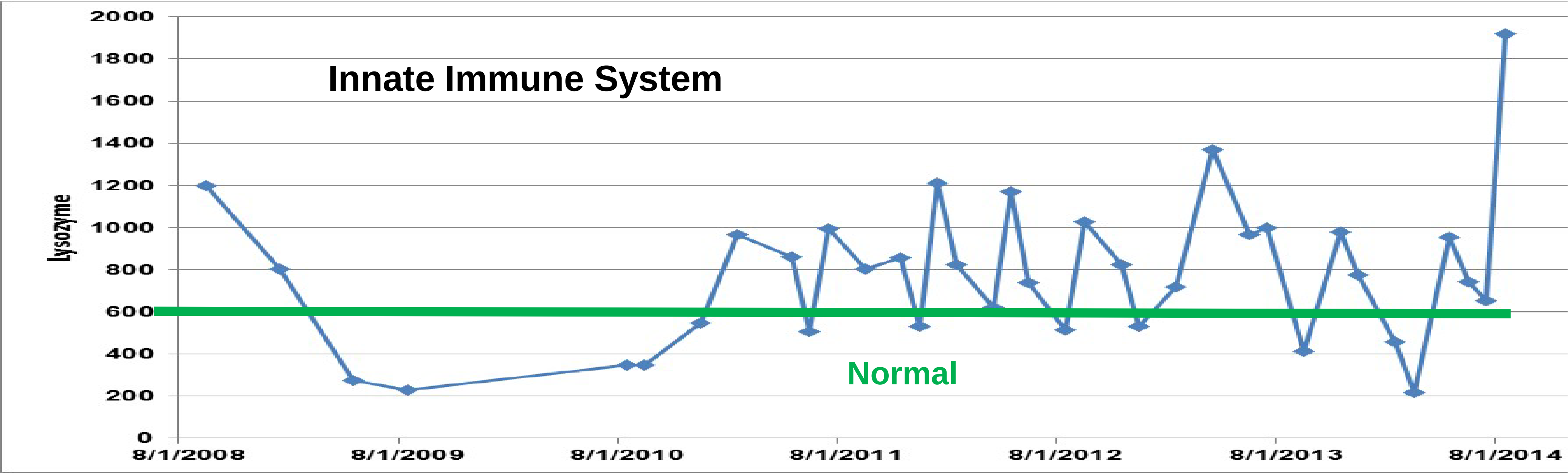
Despite decades of research,
the etiology of Crohn's disease
remains unknown.
Its pathogenesis may involve
a complex interplay between
host genetics,
immune dysfunction,
and microbial or environmental factors.

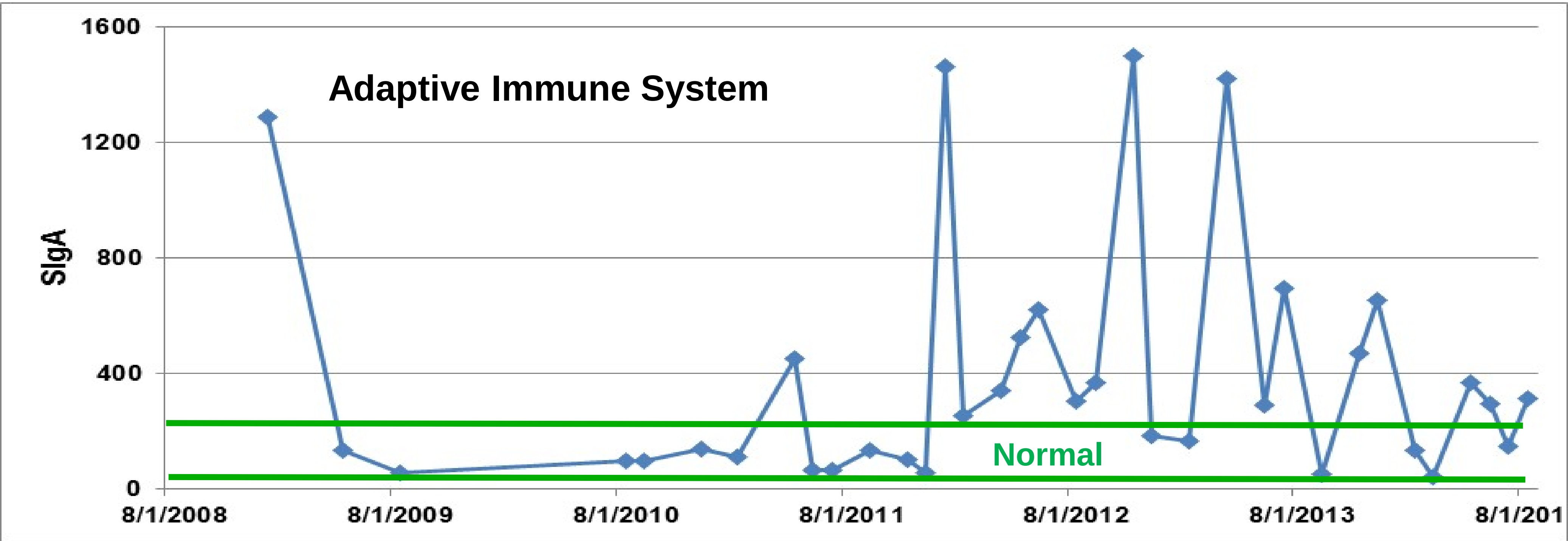
--*The Role of Microbes in Crohn's Disease*

So I Set Out to Quantify All Three!

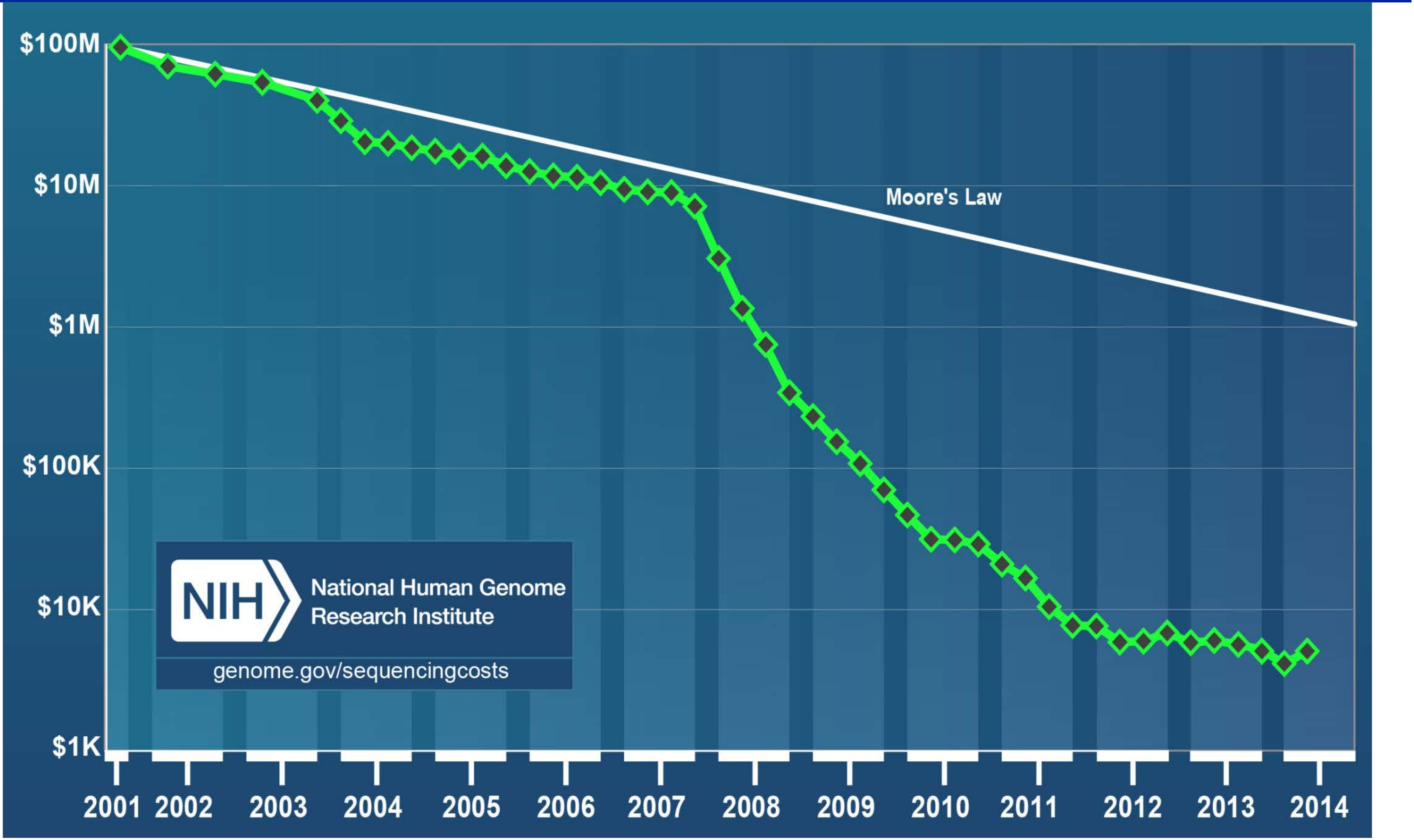
Paul B. Eckburg & David A. Relman
Clin Infect Dis. 44:256-262 (2007)

Fine Time-Resolution Sampling Reveals Dynamical Innate and Adaptive Immune Dysfunction





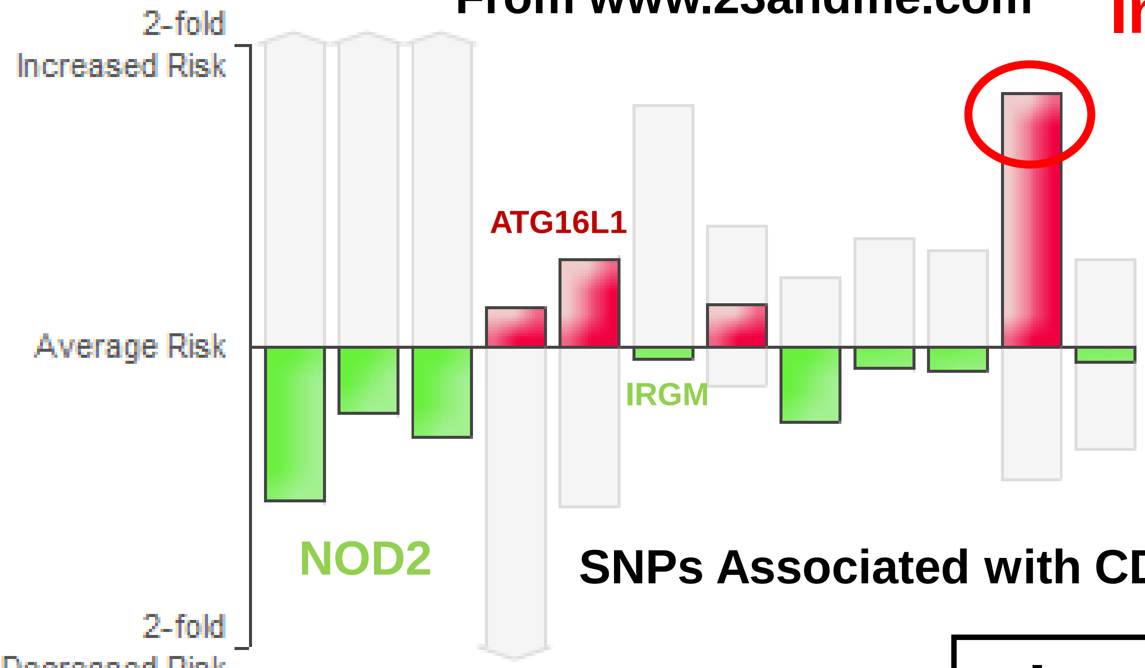
The Cost of Sequencing a Human Genome Has Fallen Over 10,000x in the Last Ten Years



This Has Enabled Sequencing of Both Human and Microbial Genomes

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

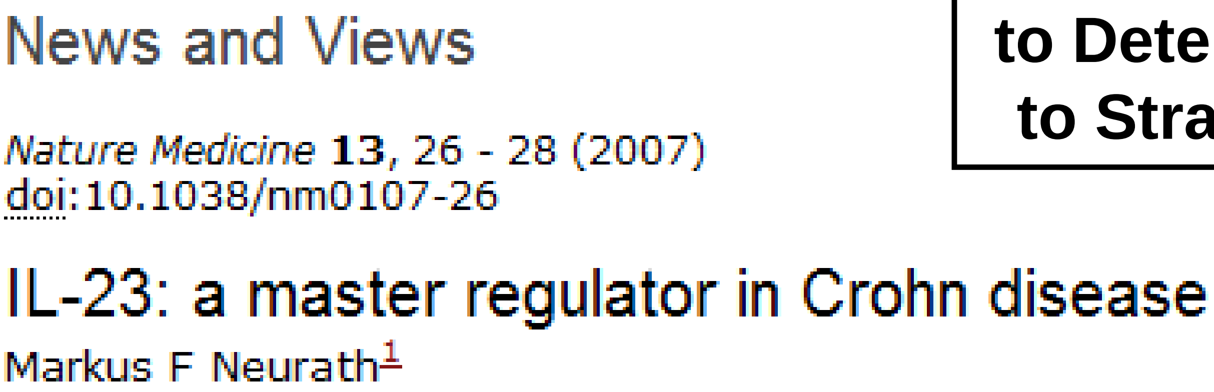
From www.23andme.com



SNPs Associated with CD

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

I am an Advisor to 23andme Who Are Seeking 10,000 Volunteers with IBD to Determine SNP Distribution to Stratify Disease Spectrum



News and Views

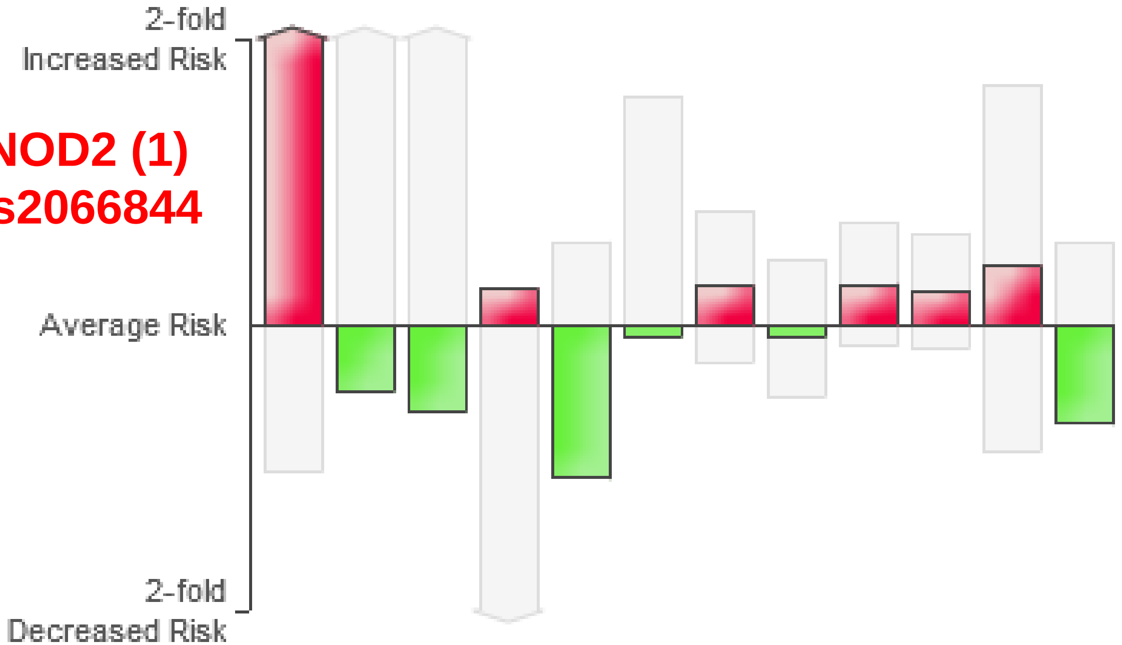
Nature Medicine **13**, 26 - 28 (2007)
doi:10.1038/nm0107-26

IL-23: a master regulator in Crohn disease

Markus F Neurath[1]

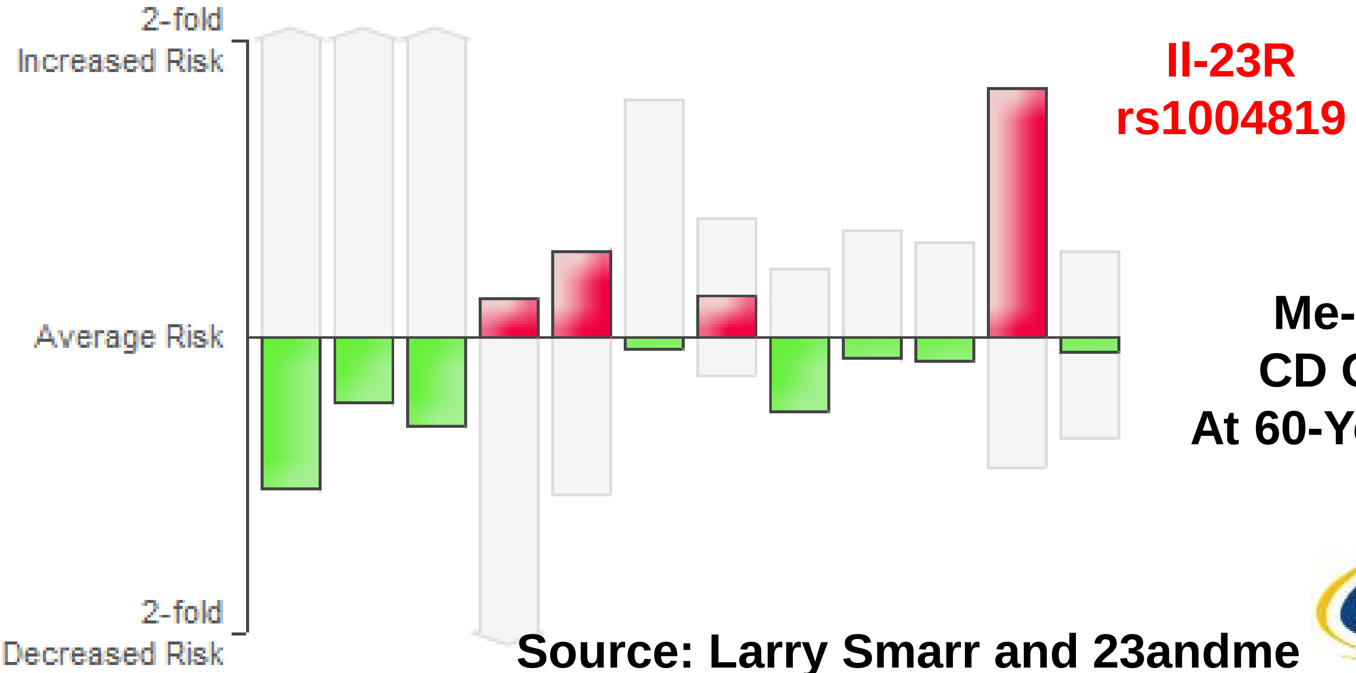
There Is Likely a Correlation Between CD SNPs and Where and When the Disease Manifests

Subject with Ileal Crohn's (ICD)



Female CD Onset At 20-Years Old

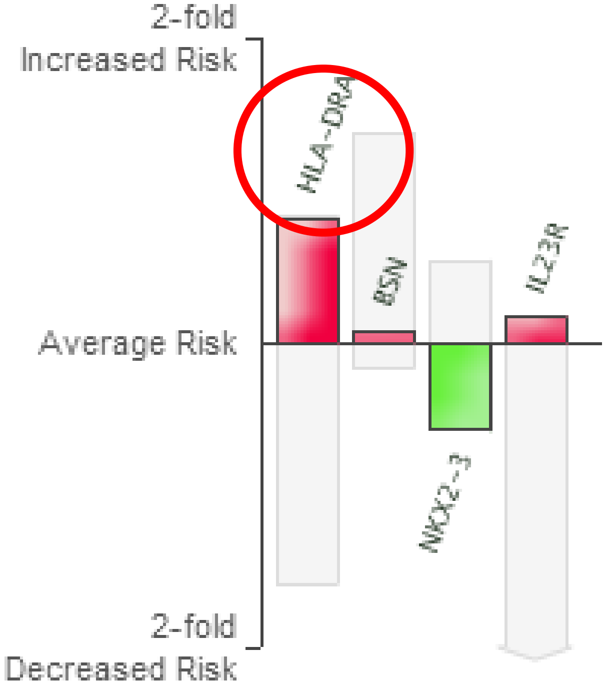
Subject with Colon Crohn's (CCD)

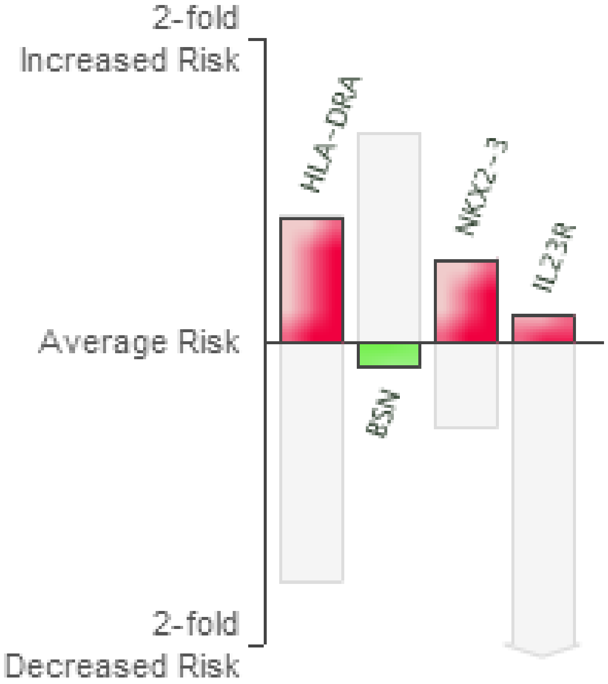


Me-Male CD Onset At 60-Years Old

Source: Larry Smarr and 23andme

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





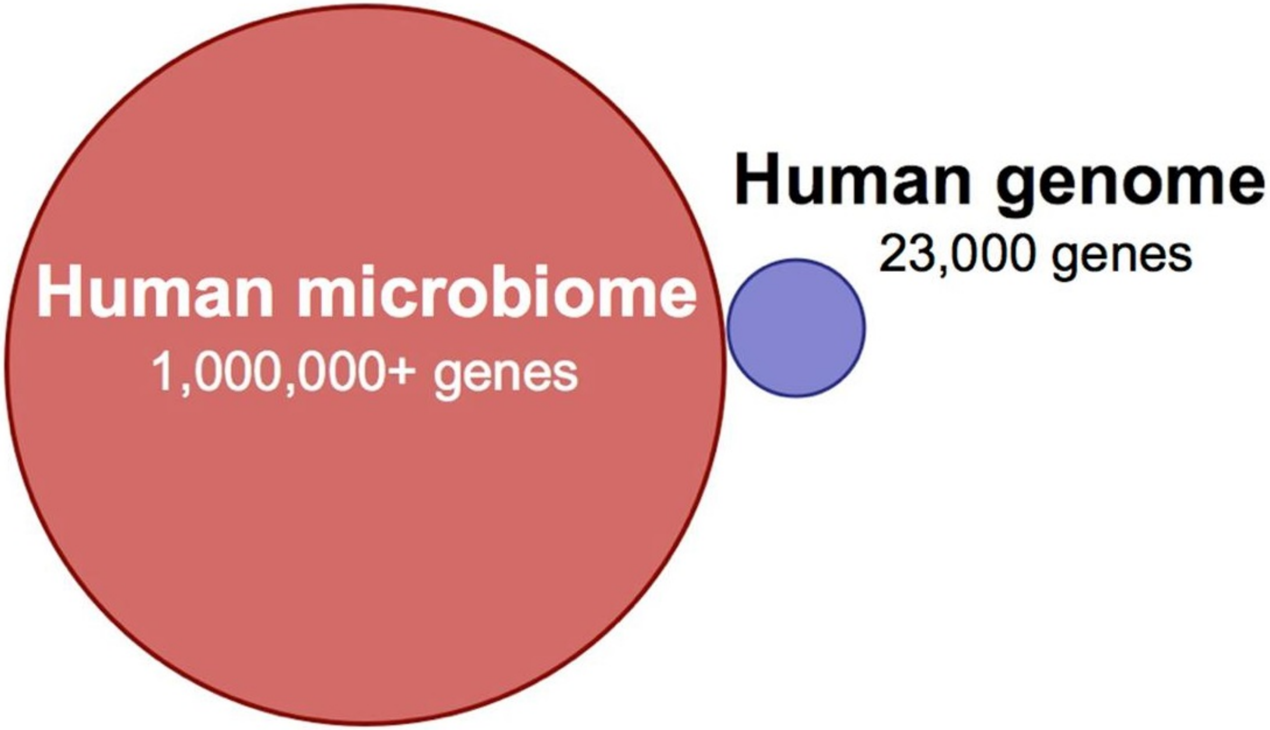
**I Have a
33% Increased Risk
for Ulcerative Colitis
HLA-DRA (rs2395185)**

**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)**

Now I am Observing the 100 Trillion Non-Human Cells in My Body

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

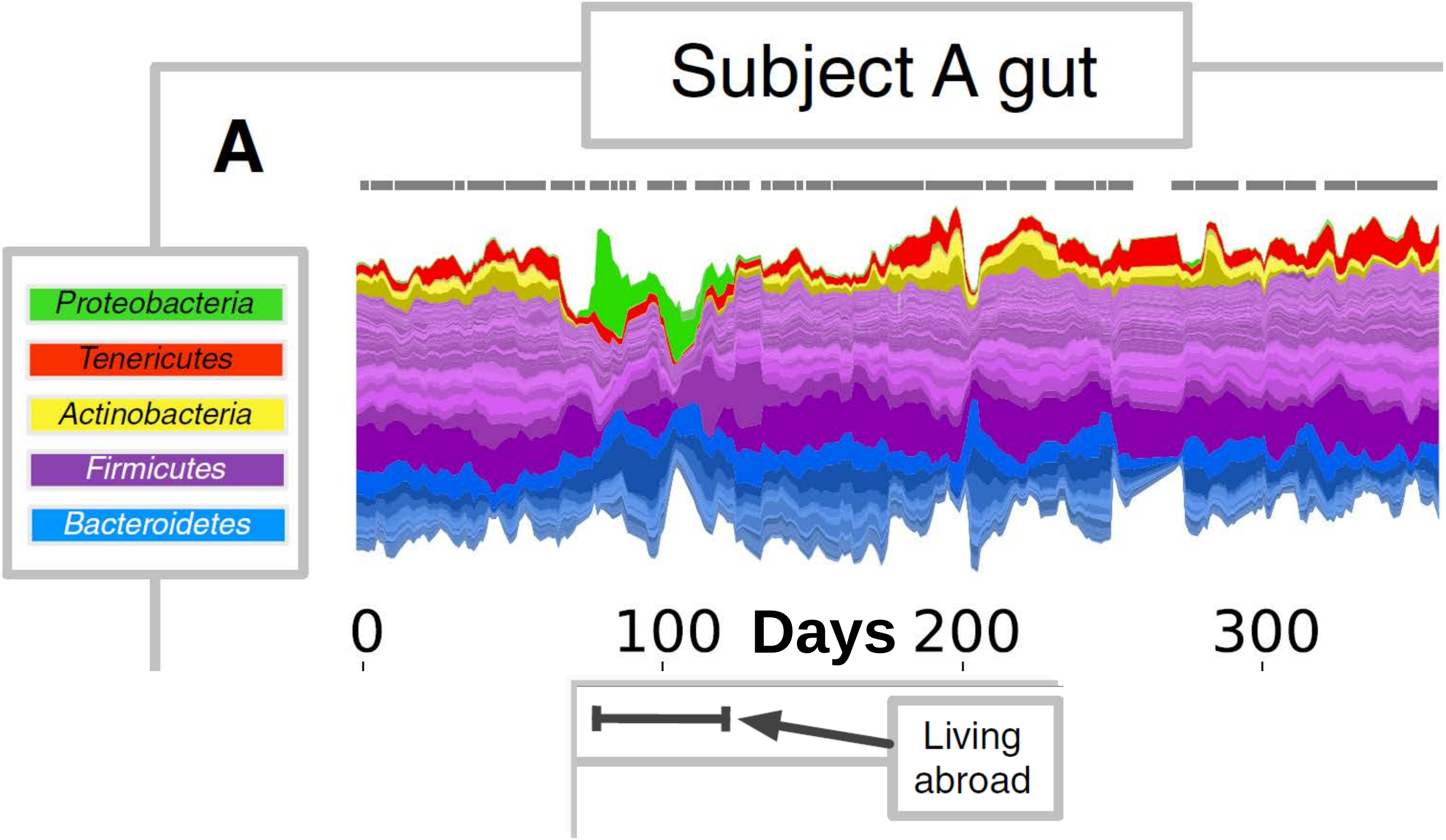


99% of Your
DNA Genes
Are in Microbe Cells
Not Human Cells

Inclusion of the Microbiome
Will Radically Change Medicine

J. Craig Venter INSTITUTE

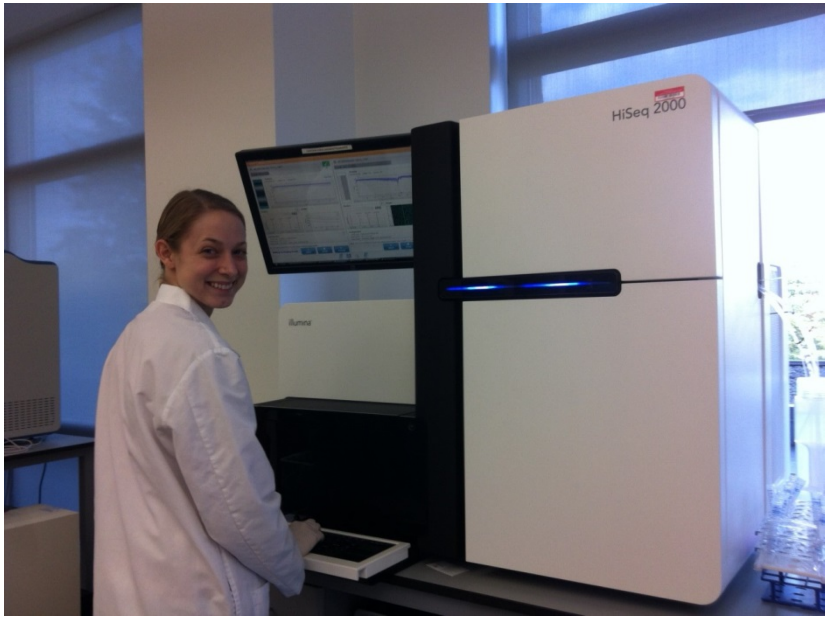
A Year of Sequencing a Healthy Gut Microbiome Daily - Remarkable Stability with Abrupt Changes



Genome Biology (2014)
David, et al.

To Map Out the Dynamics of My Microbiome Ecology I Partnered with the J. Craig Venter Institute

- JCVI Did Metagenomic Sequencing on Seven of My Stool Samples Over 1.5 Years
- Sequencing on Illumina HiSeq 2000
  - Generates 100bp Reads
- JCVI Lab Manager, Genomic Medicine
  - Manolito Torralba
- IRB PI Karen Nelson
  - President JCVI



Illumina HiSeq 2000 at JCVI

Manolito Torralba, JCVI

Karen Nelson, JCVI

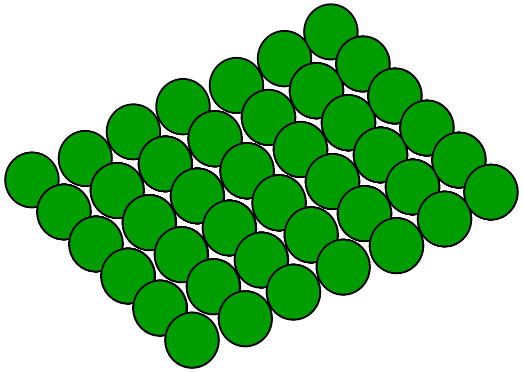
J. Craig Venter INSTITUTE

We Downloaded Additional Phenotypes from NIH HMP For Comparative Analysis

Download Raw Reads
~100M Per Person

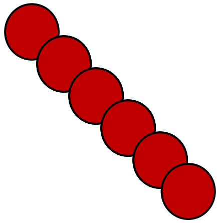
“Healthy” Individuals

35 Subjects
1 Point in Time



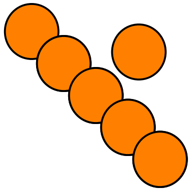
Larry Smarr

6 Points in Time

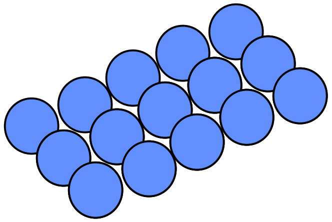


IBD Patients

2 Ulcerative Colitis Patients,
6 Points in Time



5 Ileal Crohn’s Patients,
3 Points in Time



Total of 5 Billion Reads

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

NCBI National Center for Biotechnology Information

HMP

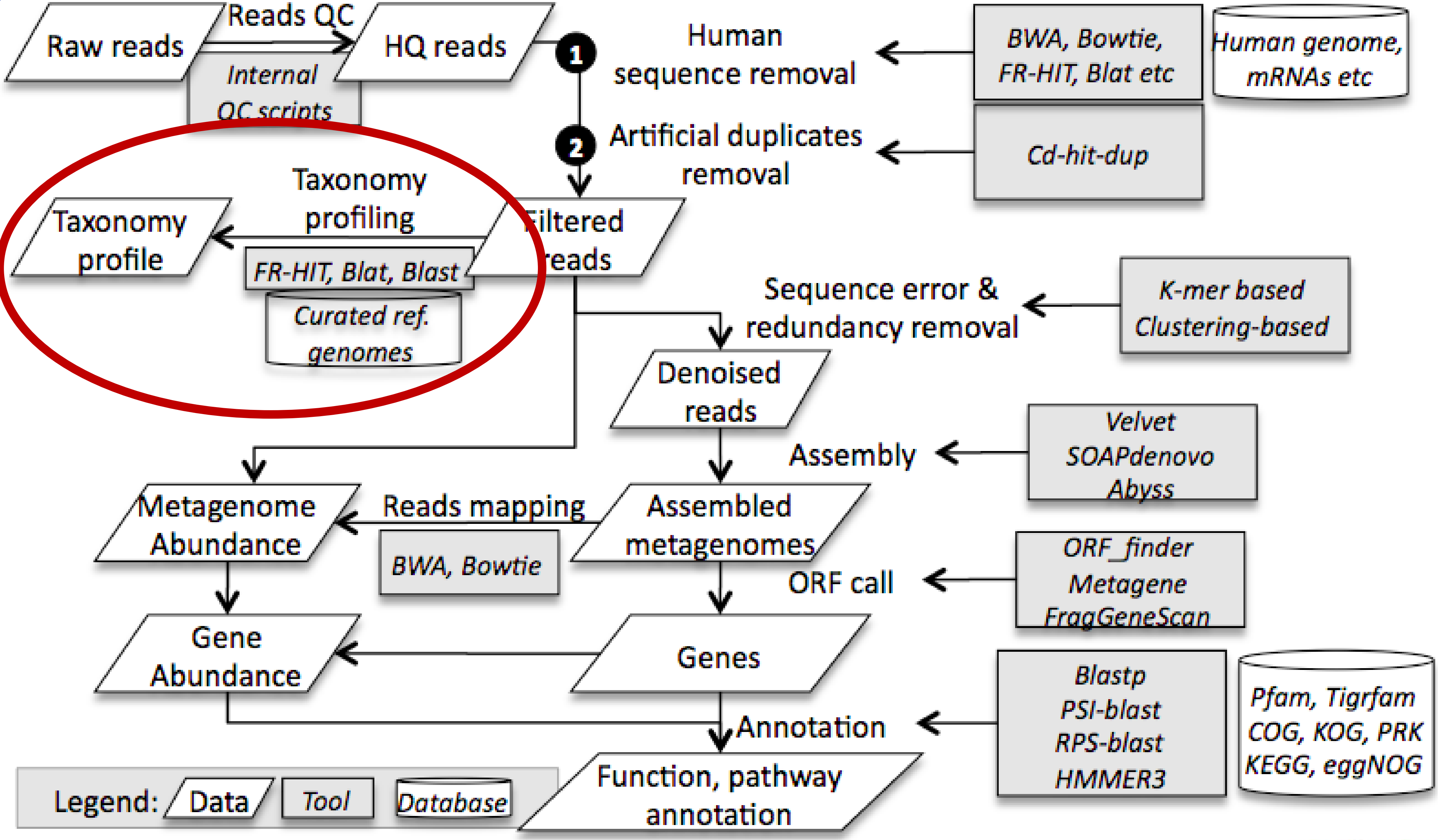
NIH

Now to Align Our 5 Billion Reads Against the Reference Database

J. Craig Venter INSTITUTE

Source: Weizhong Li, Sitao Wu, CRBS, UCSD

Computational NextGen Sequencing Pipeline: From “Big Equations” to “Big Data” Computing



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 Analyze a Wide Range of 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

SDSC

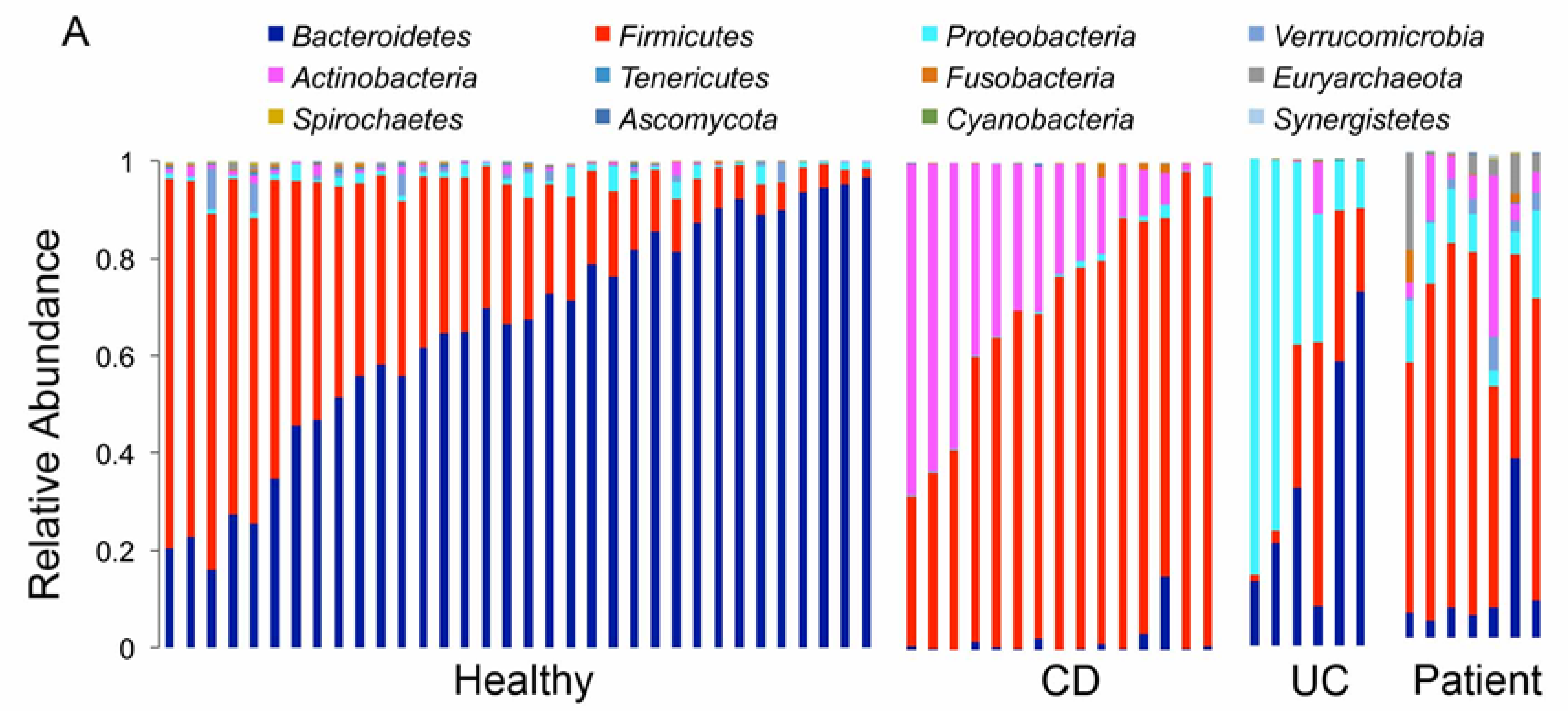
Enabled by a Grant of Time on Gordon from SDSC Director Mike Norman

NSF NIH HMP

J. Craig Venter INSTITUTE

CRBS it2

The Emergence of
Microbial Genomics Diagnostics



**Microbial Ecology Is Radically Altered in Disease States,
But Differently in the Two Forms of IBD**

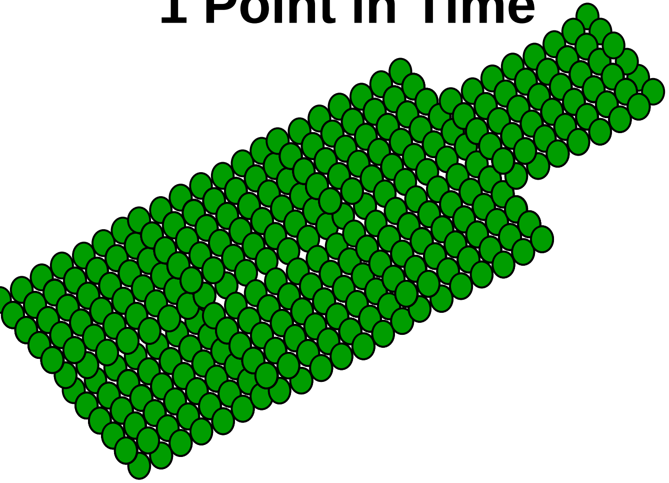
Source: Chang, et al. (2014)

We Expaned Our Healthy Cohort to All Gut Microbiomes from NIH HMP For Comparative Analysis

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

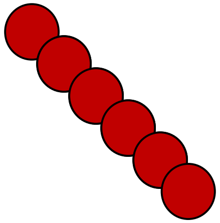
“Healthy” Individuals

250 Subjects
1 Point in Time



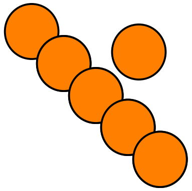
Larry Smarr

7 Points in Time

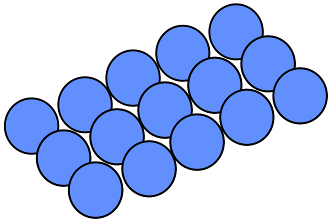


IBD Patients

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

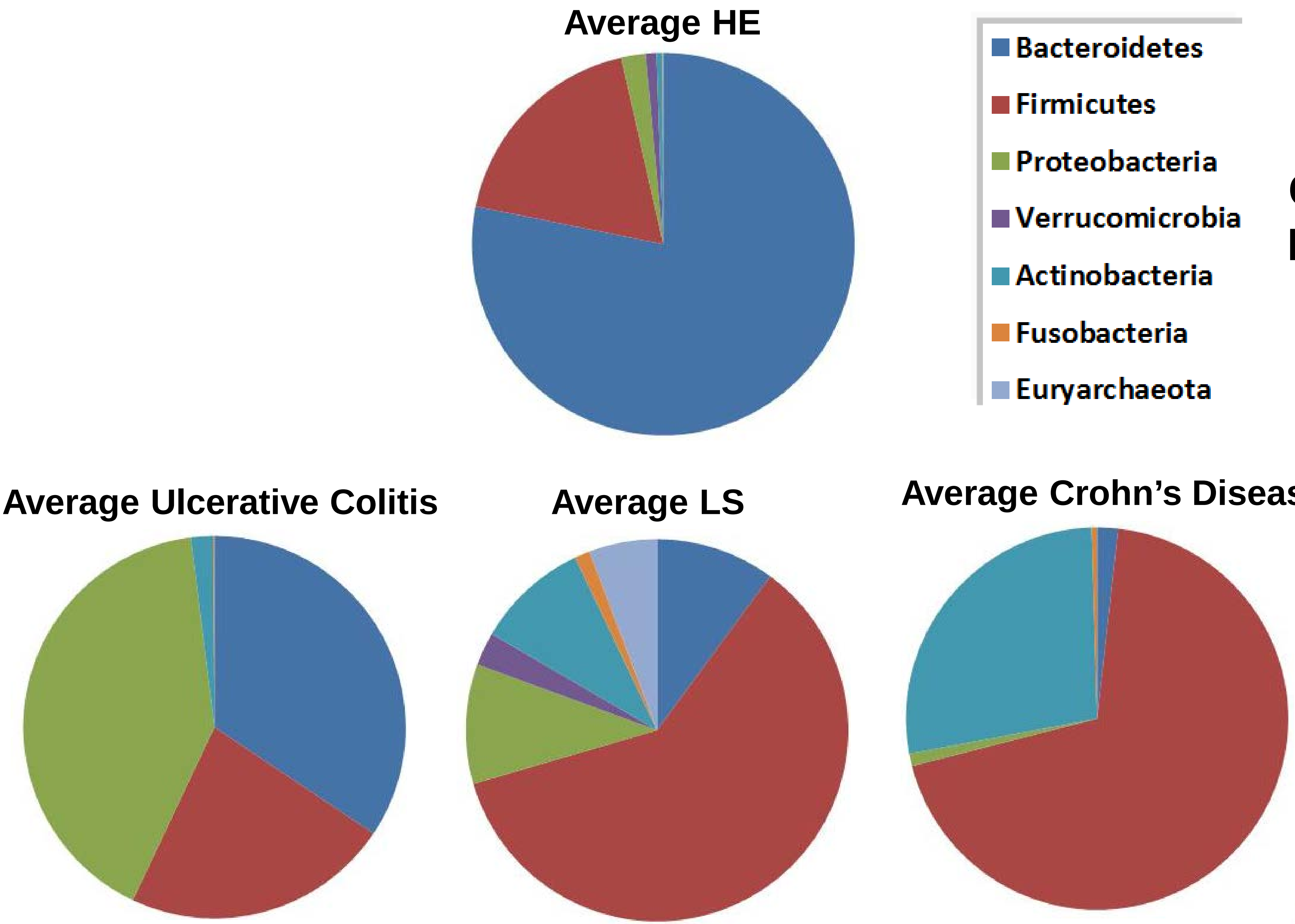
We Used Dell's HPC Cloud to Analyze All of Our Human Gut Microbiomes

Dell Solutions Center
Industry Solutions Lab
SANGER DSU

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

Source: Weizhong Li, UCSD

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



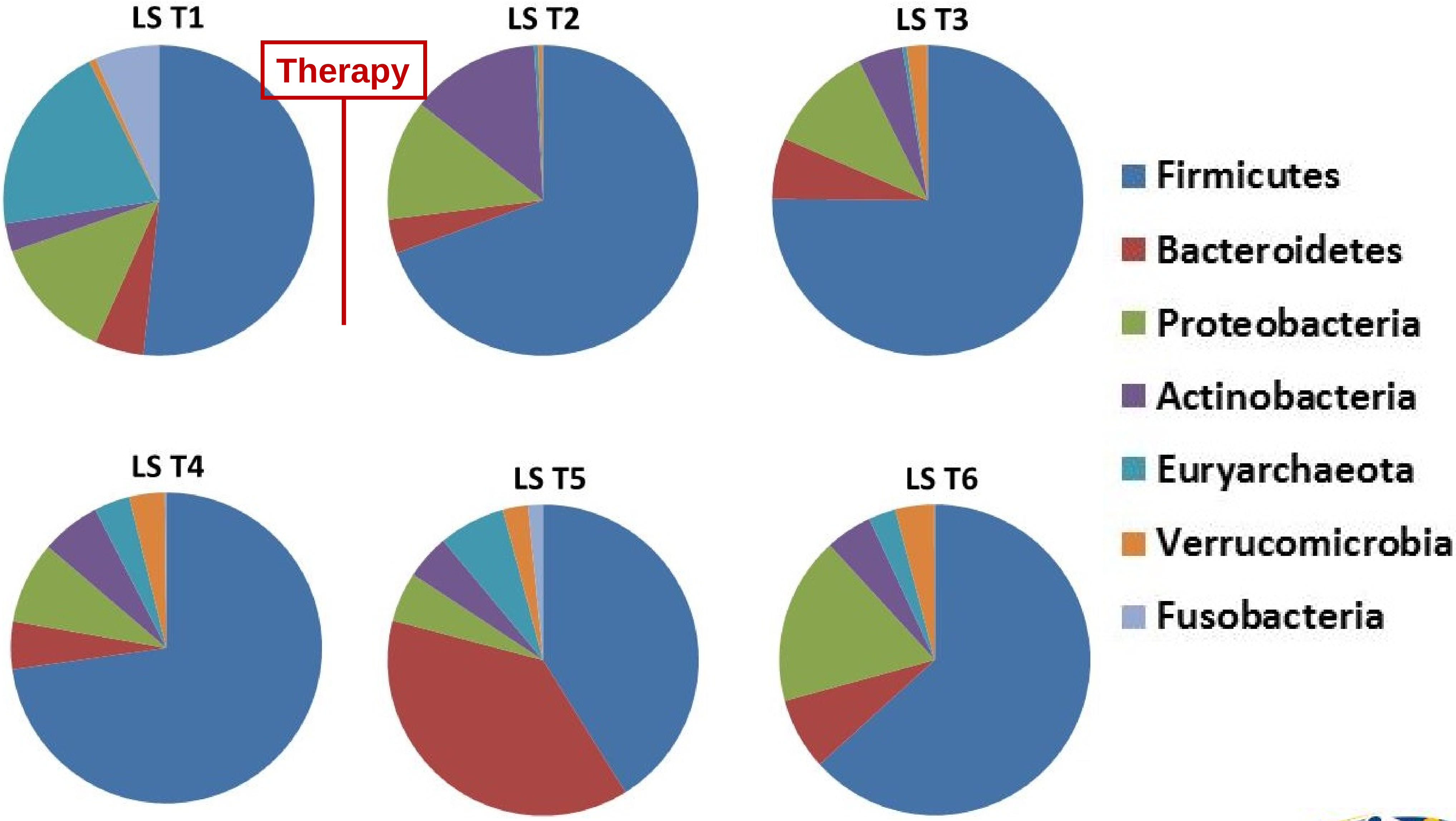
Most Common Microbial Phyla

Explosion of Proteobacteria

Hybrid of UC and CD
High Level of Archaea

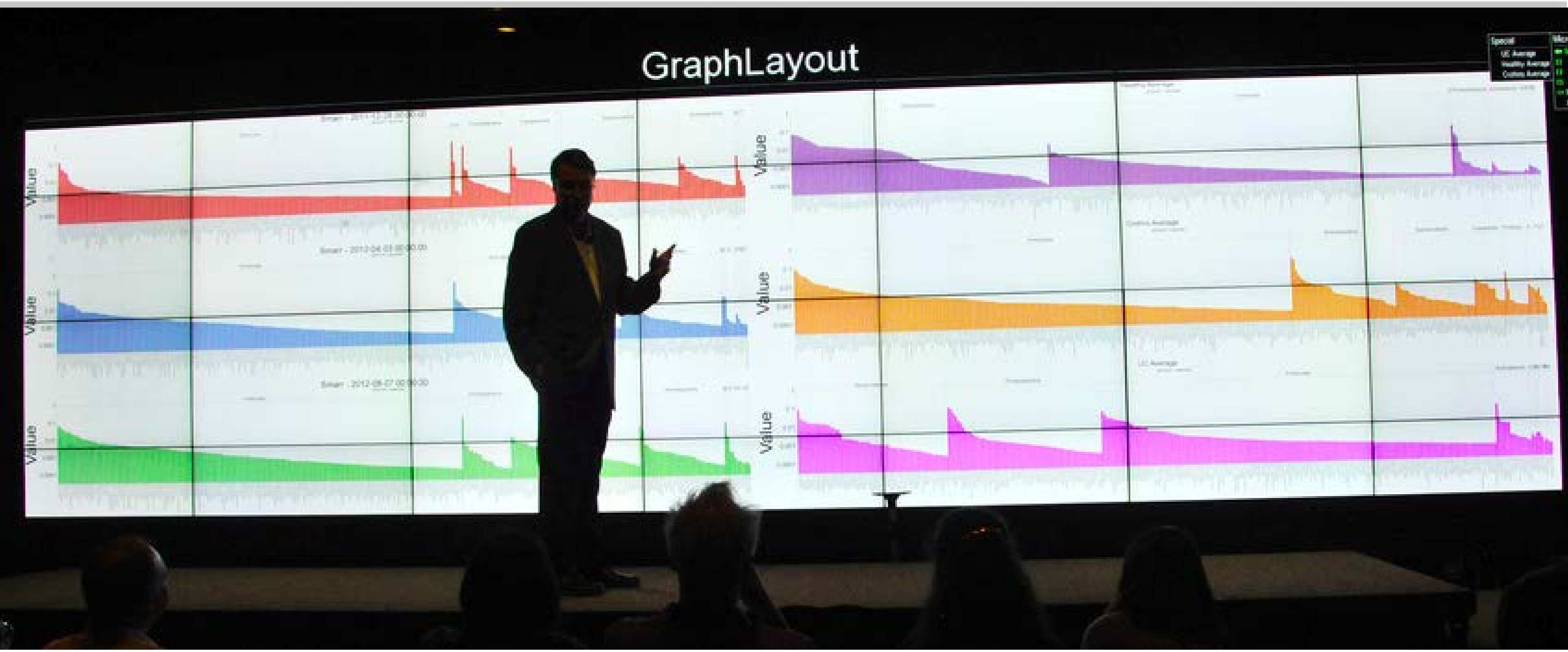
Collapse of Bacteroidetes
Explosion of Actinobacteria

Time Series Reveals Autoimmune Dynamics of Gut Microbiome by Phyla



Six Metagenomic Time Samples Over 16 Months

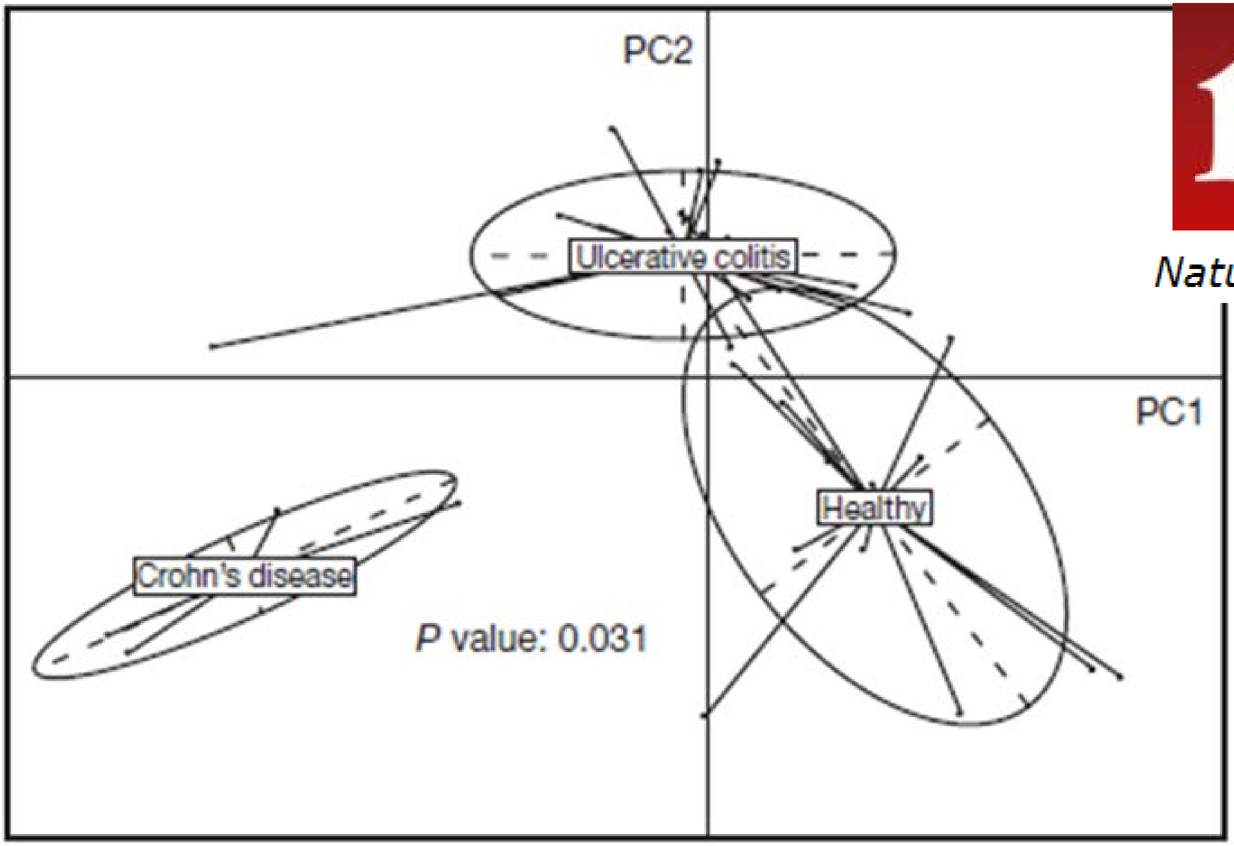
Using Scalable Visualization Allows Comparison of the Relative Abundance of 200 Microbe Species



Comparing 3 LS Time Snapshots (Left)
with Healthy, Crohn’s, UC (Right Top to Bottom)

Calit2 VROOM-FuturePatient Expedition

Can Microbial Metagenomics Diagnose Disease States?



nature

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

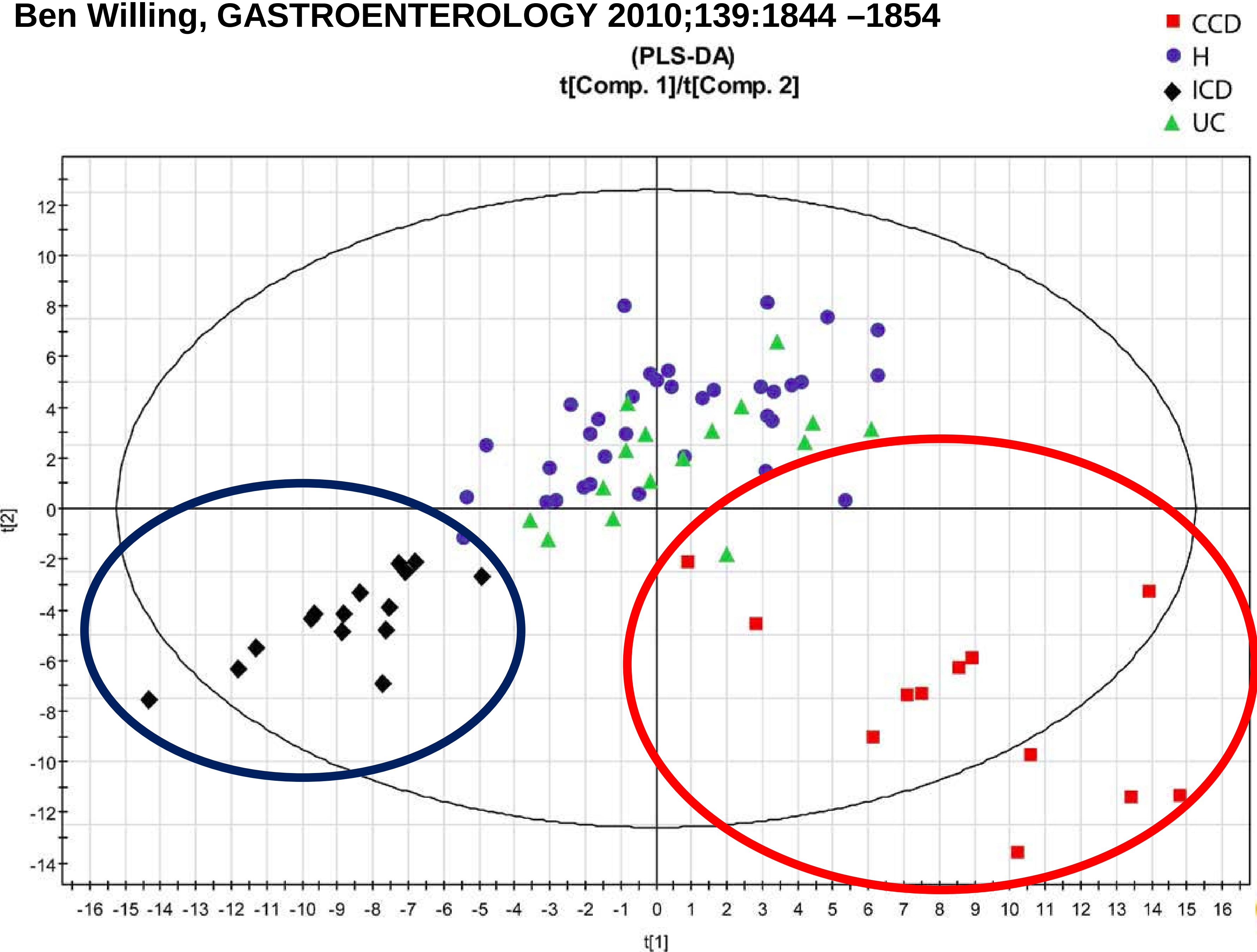
Figure 4 | Bacterial species abundance differentiates IBD patients and healthy individuals. Principal component analysis with health status as

A human gut microbial gene catalogue established by metagenomic sequencing

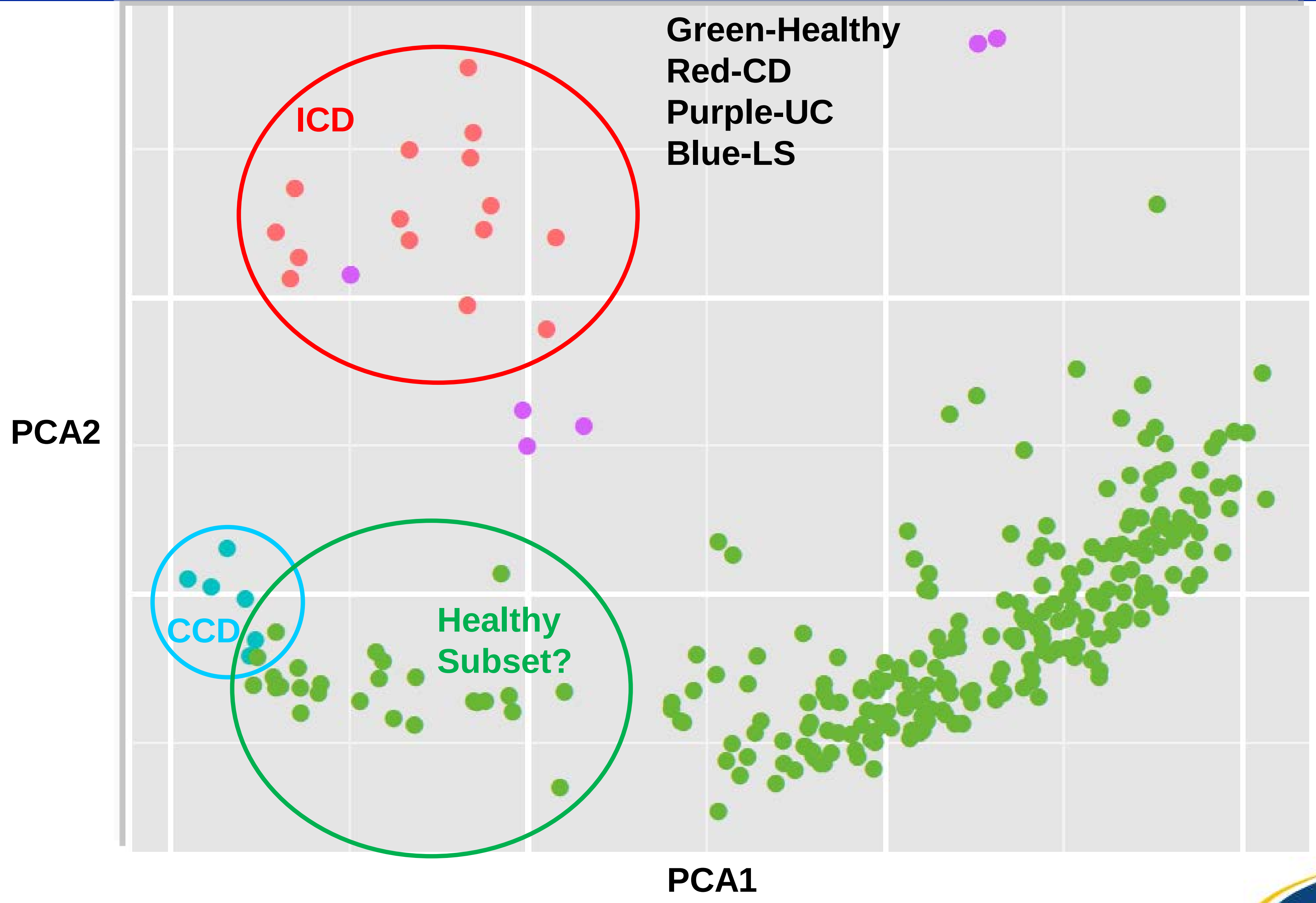
Junjie Qin[1]*, Ruiqiang Li[1]*, Jeroen Raes[2,3], Manimozhiyan Arumugam[2], Kristoffer Solvsten Burgdorf[4],

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

Ben Willing, GASTROENTEROLOGY 2010;139:1844 –1854

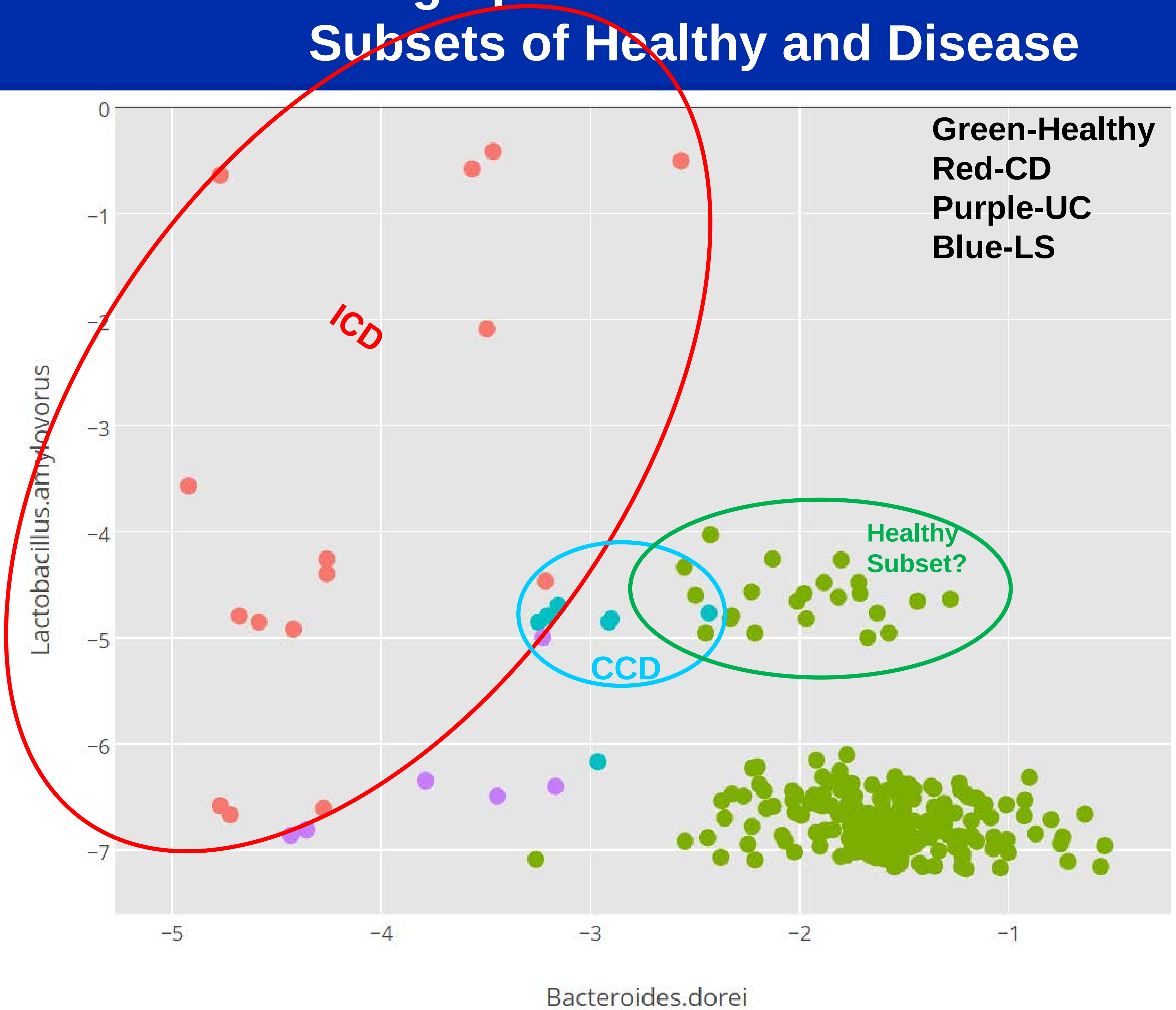


PCA Analysis
on Species Abundance Across People



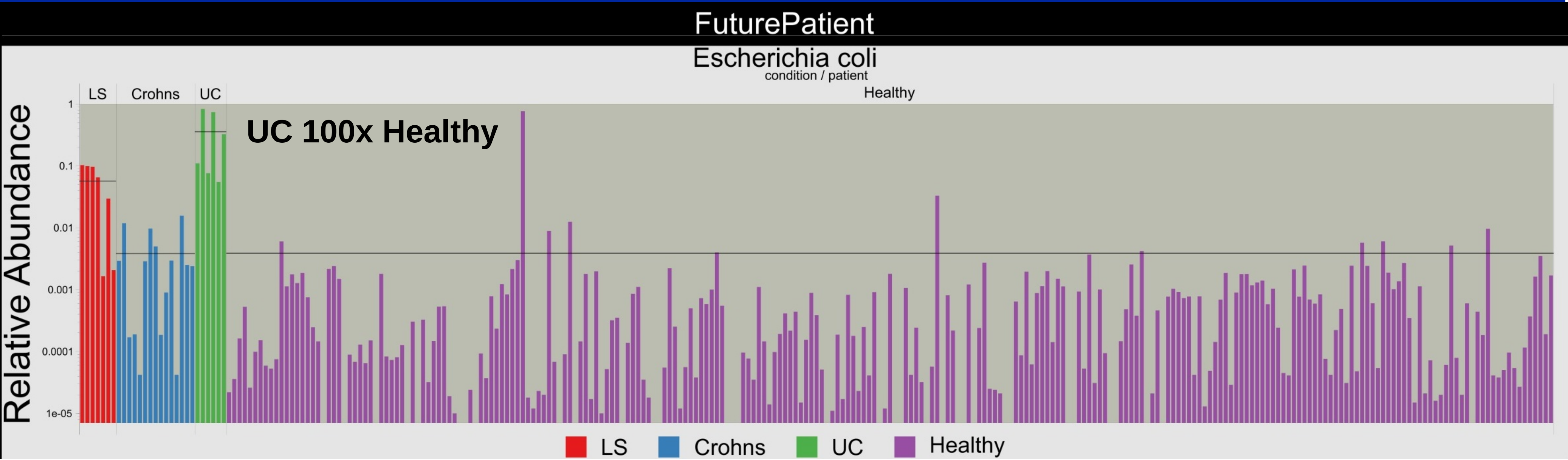
Analysis by Mehrdad Yazdani, Calit2

Finding Species Which Differentiate Subsets of Healthy and Disease

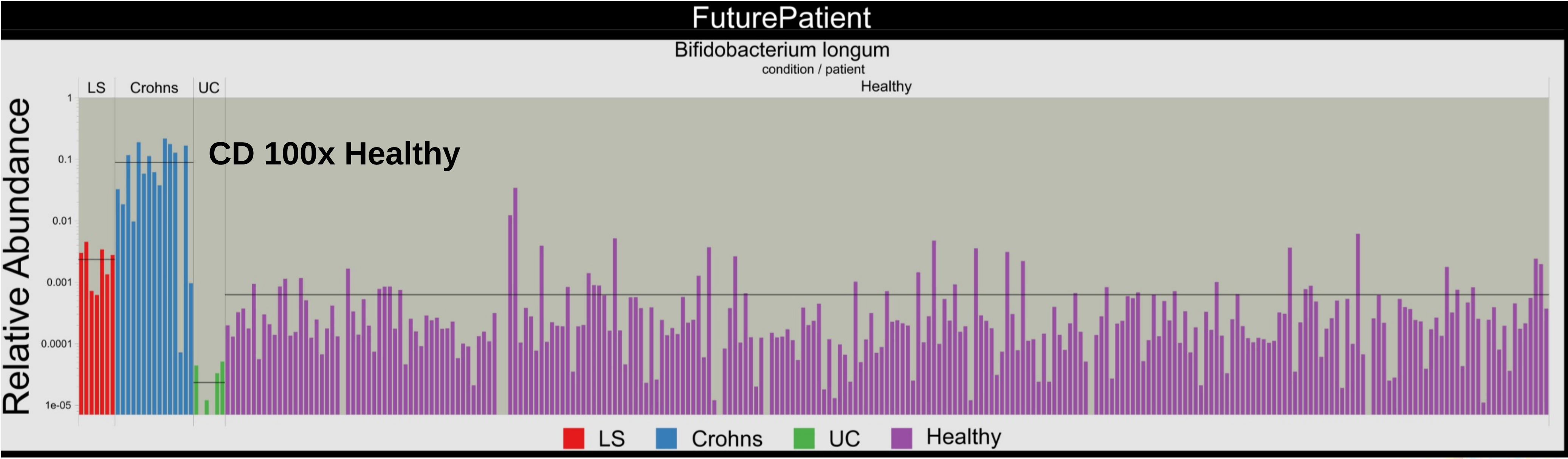


Green-Healthy
Red-CD
Purple-UC
Blue-LS

Dell Cloud Results Are Leading Toward Microbiome Disease Diagnosis



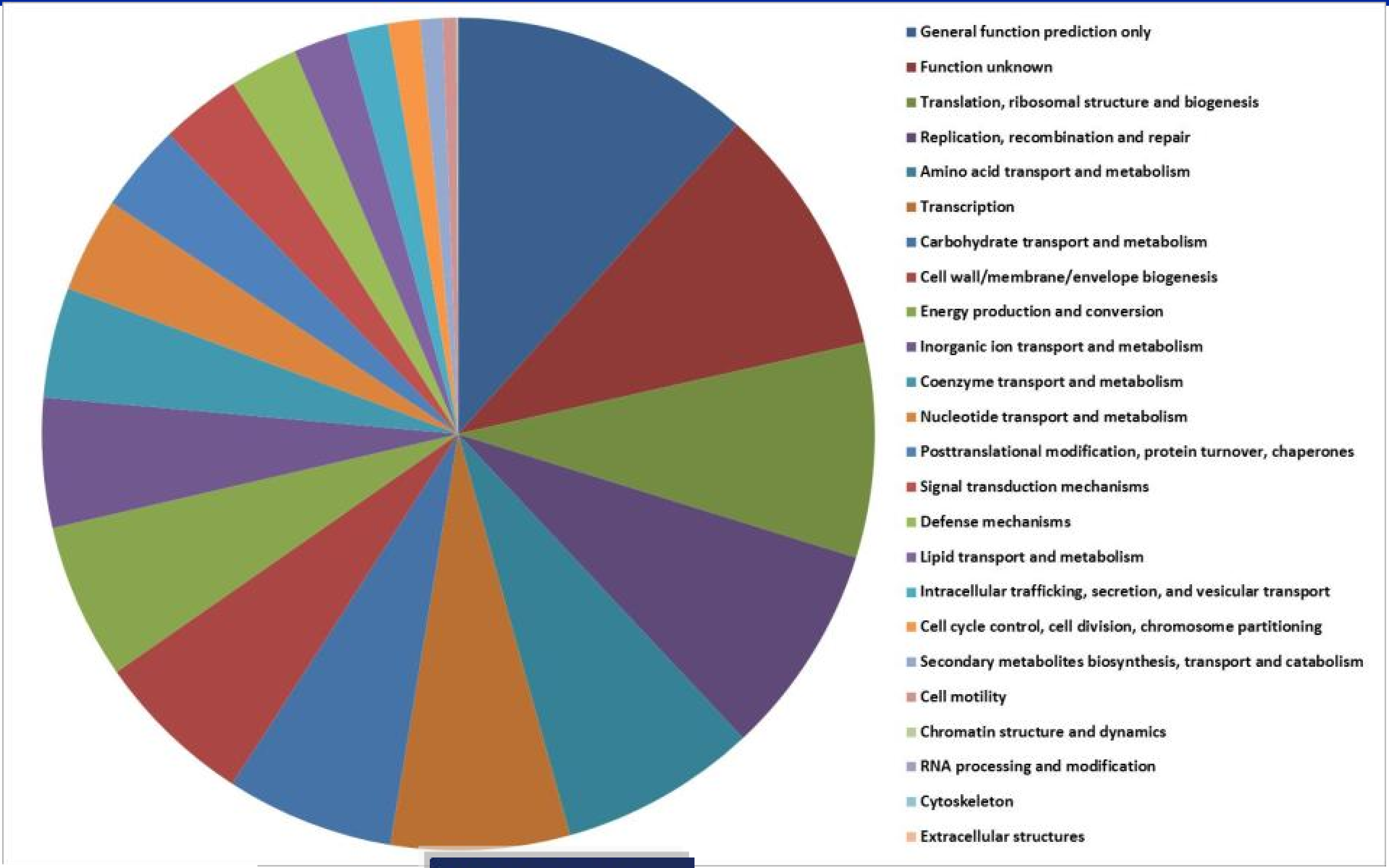
UC 100x Healthy



CD 100x Healthy

We Produced Similar Results for ~2500 Microbial Species

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



J. Craig Venter INSTITUTE

UC San Diego HEALTH SCIENCES

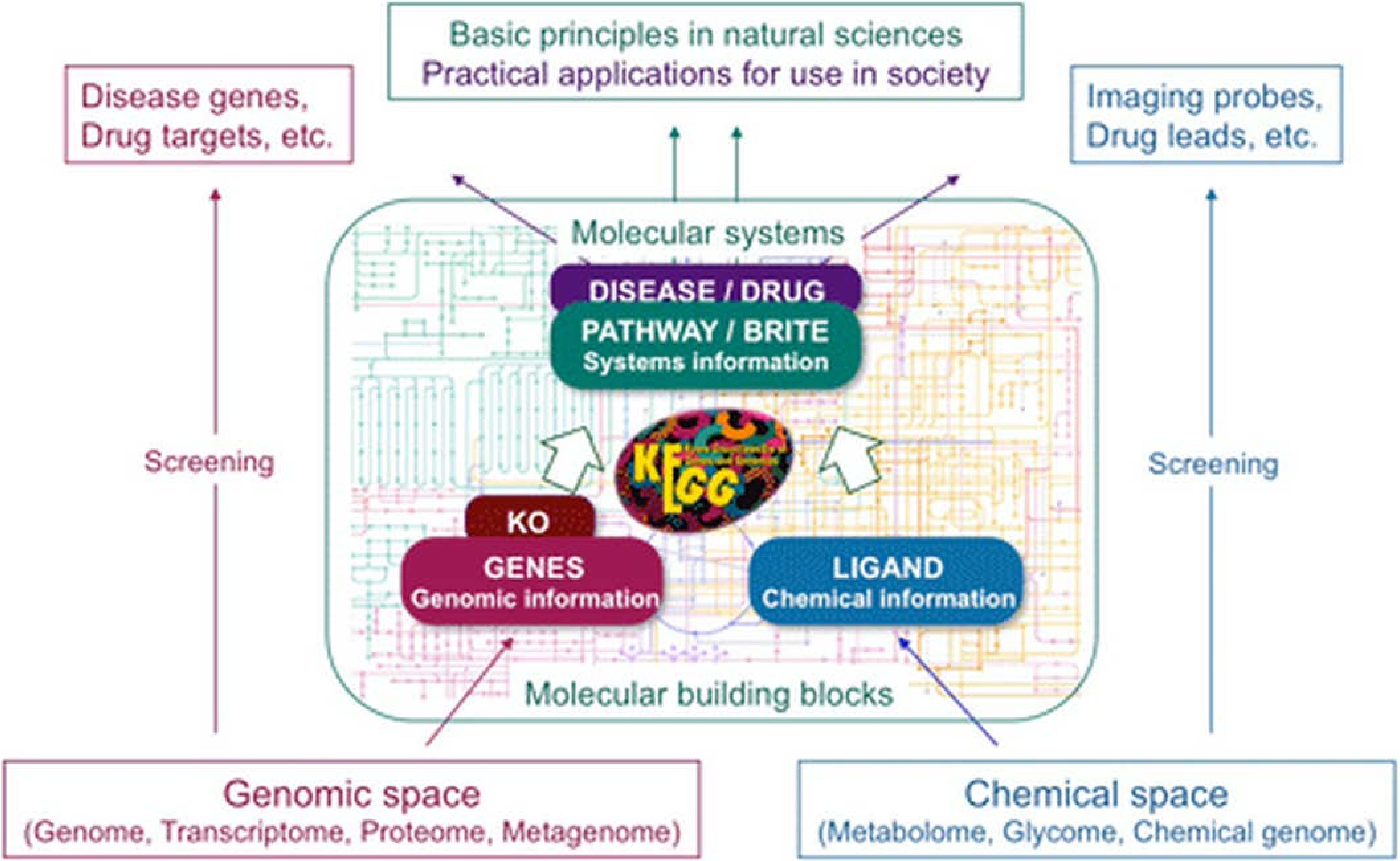
SDSC

it2

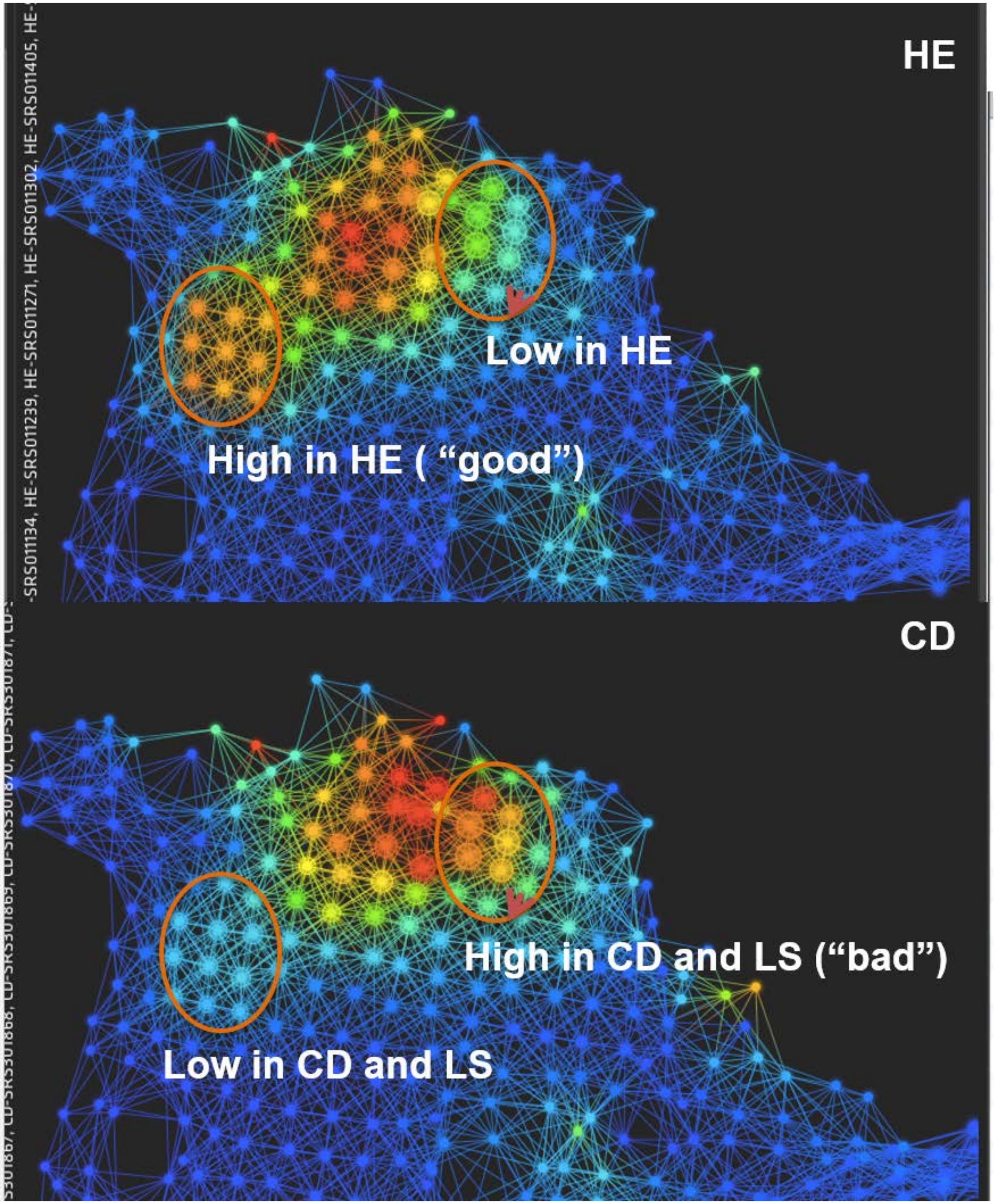
Analysis: Weizhong Li & Sitao Wu, UCSD

KEGG: a Database Resource for Understanding High-Level Functions and Utilities of the Biological System

http://www.genome.jp/kegg/



Using Ayasdi To Discover Patterns in KEGG Dataset



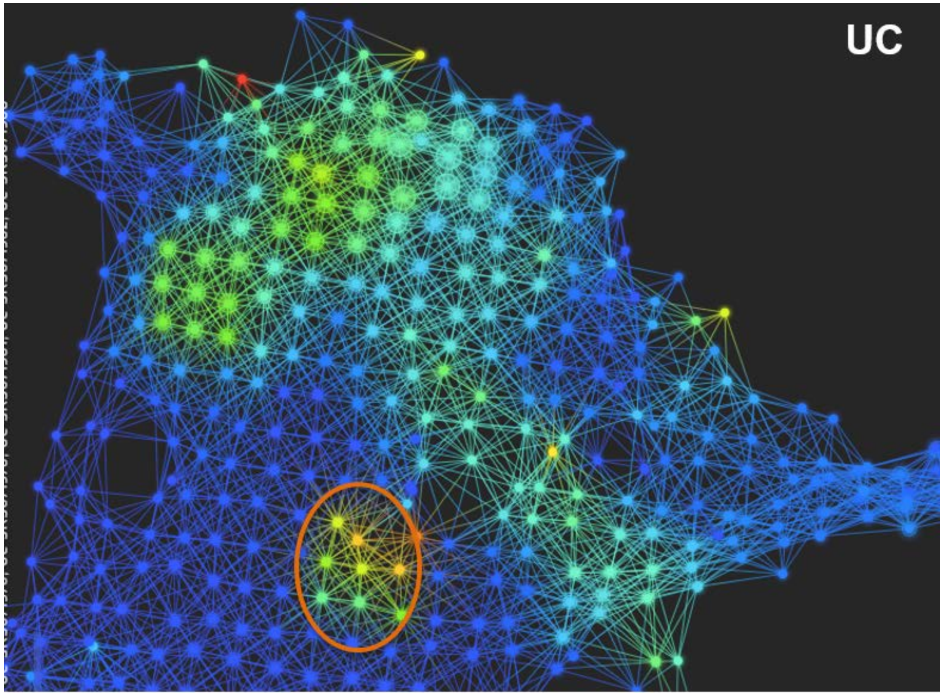
Ayasdi Advanced Analytics

topological data analysis

Ayasdi Cure™

Turn Data into Therapies

Source: Pek Lum, Chief Data Scientist, Ayasdi



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

Next Step:
Compute Genes and Function

Full Processing to Function
(COGs, KEGGs)

Would Require
~1-2 Million
Core-Hours

Plus Dedicated Network to Move Data
From R Systems / Dell to Calit2@UC San Diego

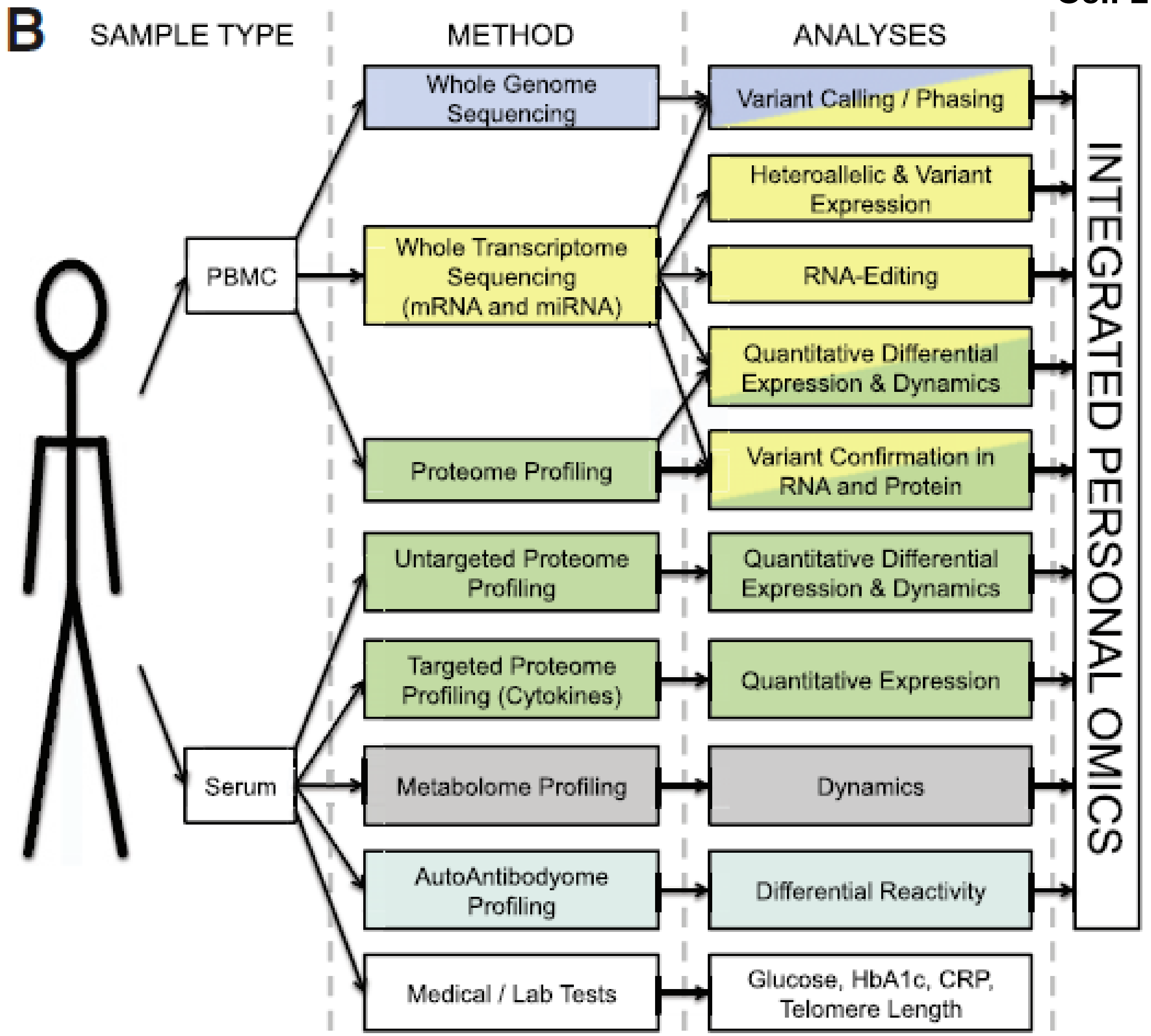
Next Step: Time Series of Metagenomic Gut Microbiomes and Immune Variables in an N=100 Clinic Trial

Goal: Understand
The Coupled Human Immune-Microbiome Dynamics
In the Presence of Human Genetic Predispositions

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

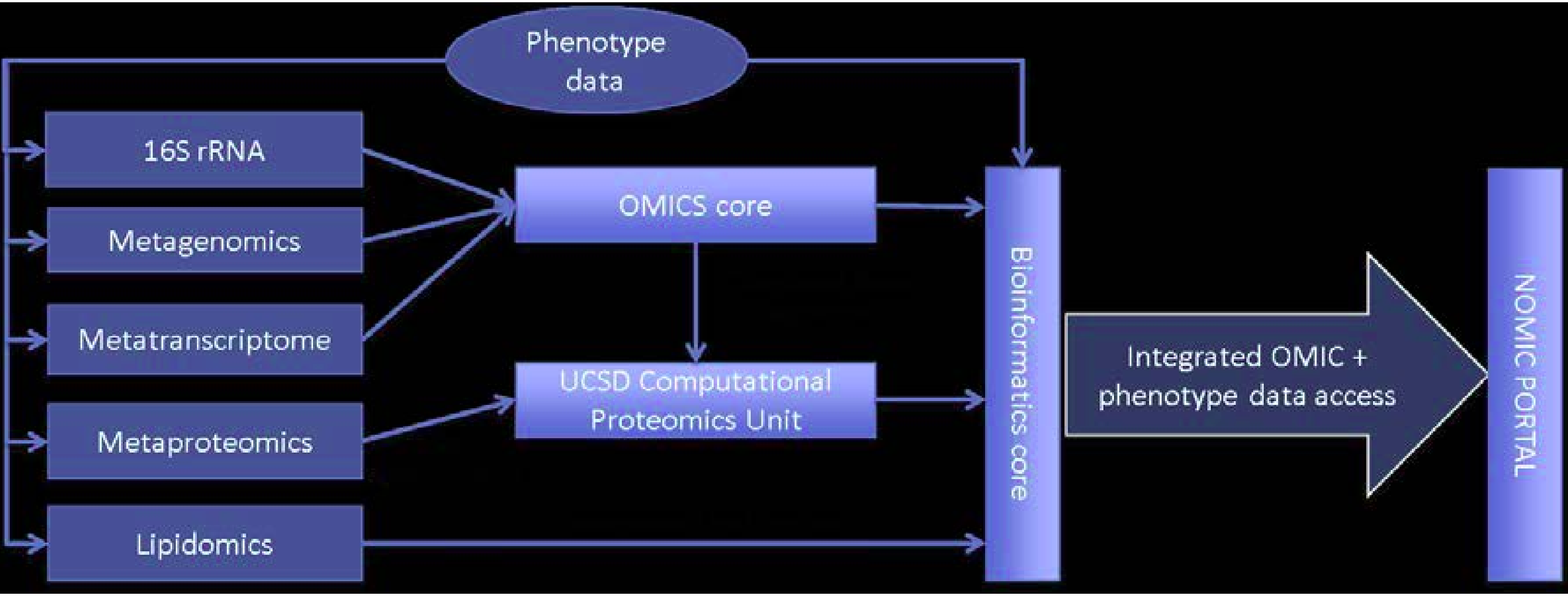
100x Beyond Current Medical Tests: Integrated Personal Time Series of Multiple 'Omics

Cell 148, 1293–1307, March 16, 2012


B
SAMPLE TYPE
METHOD
ANALYSES
PBMC
Serum
Whole Genome Sequencing
Whole Transcriptome Sequencing (mRNA and miRNA)
Proteome Profiling
Untargeted Proteome Profiling
Targeted Proteome Profiling (Cytokines)
Metabolome Profiling
AutoAntibodyome Profiling
Medical / Lab Tests
Variant Calling / Phasing
Heteroallelic & Variant Expression
RNA-Editing
Quantitative Differential Expression & Dynamics
Variant Confirmation in RNA and Protein
Quantitative Differential Expression & Dynamics
Quantitative Expression
Dynamics
Differential Reactivity
Glucose, HbA1c, CRP, Telomere Length
INTEGRATED PERSONAL OMICS

- 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

Proposed UCSD
Integrated Omics Pipeline



Source: Nuno Bandiera, UCSD

From Quantified Self to National-Scale Biomedical Research Projects

DELSA

DATA-ENABLED LIFE SCIENCES ALLIANCE

Accelerate the impact of

Data-Enabled Life Sciences Research

on the pressing needs of the global society.

The Quantified Human Initiative is an effort to combine our natural curiosity about self with new research paradigms. **Rich datasets of two individuals, Drs. Smarr and Snyder,** serve as 21st century personal data prototypes.

My Anonymized Human Genome is Available for Download

Personal Genome Project

GENOME: The Future is Now WEBISODE 1
George Church, Ph.D.
Founder, Personal Genome Project
2:05 / 7:22

www.delsaglobal.org

www.personalgenomes.org

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
Mahidhar Tatineni
Robert Sinkovits

UCSD Health Sciences Team

William J. Sandborn
Elisabeth Evans
John Chang
Brigid Boland
David Brenner