“Big Data and Superorganism Genomics – Microbial Metagenomics Meets Human Genomics”

NGS and the Future of Medicine

Illumina Headquarters

La Jolla, CA

February 27, 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

I Arrived in La Jolla in 2000 After 20 Years in the Midwest and Decided to Move Against the Obesity Trend

Age 41



Age 51



Age 61



I Reversed My Body's Decline By Quantifying and Altering Nutrition and Exercise

http://lsmarr.calit2.net/repository/LS_reading_recommendations_FiRe_2011.pdf

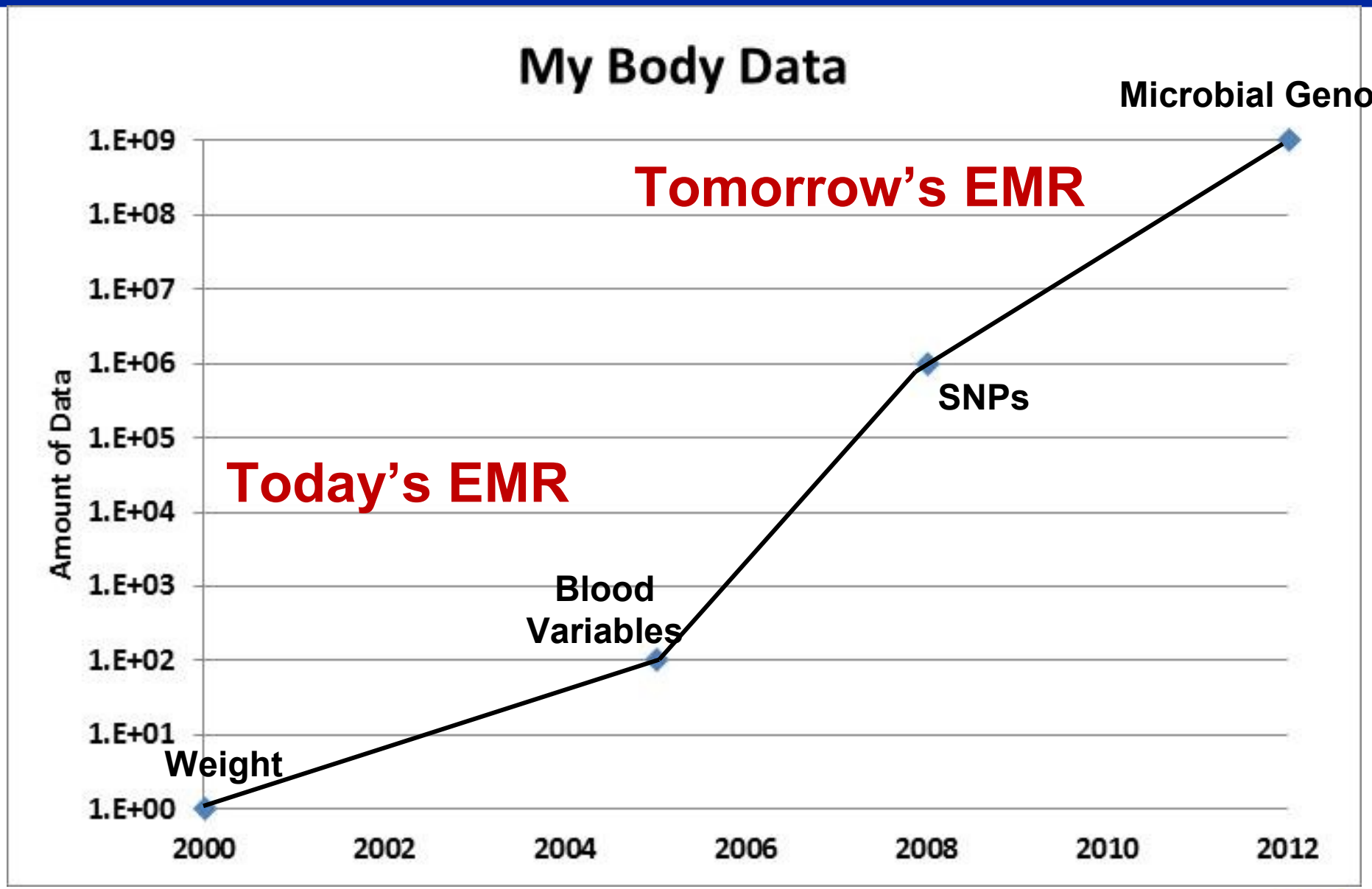
Consumer Self Measurement is Exploding Totally Outside of the Medical Complex

From the First San Francisco QS Meetup in 2008
To 116 Cities in 37 Countries in Four Years

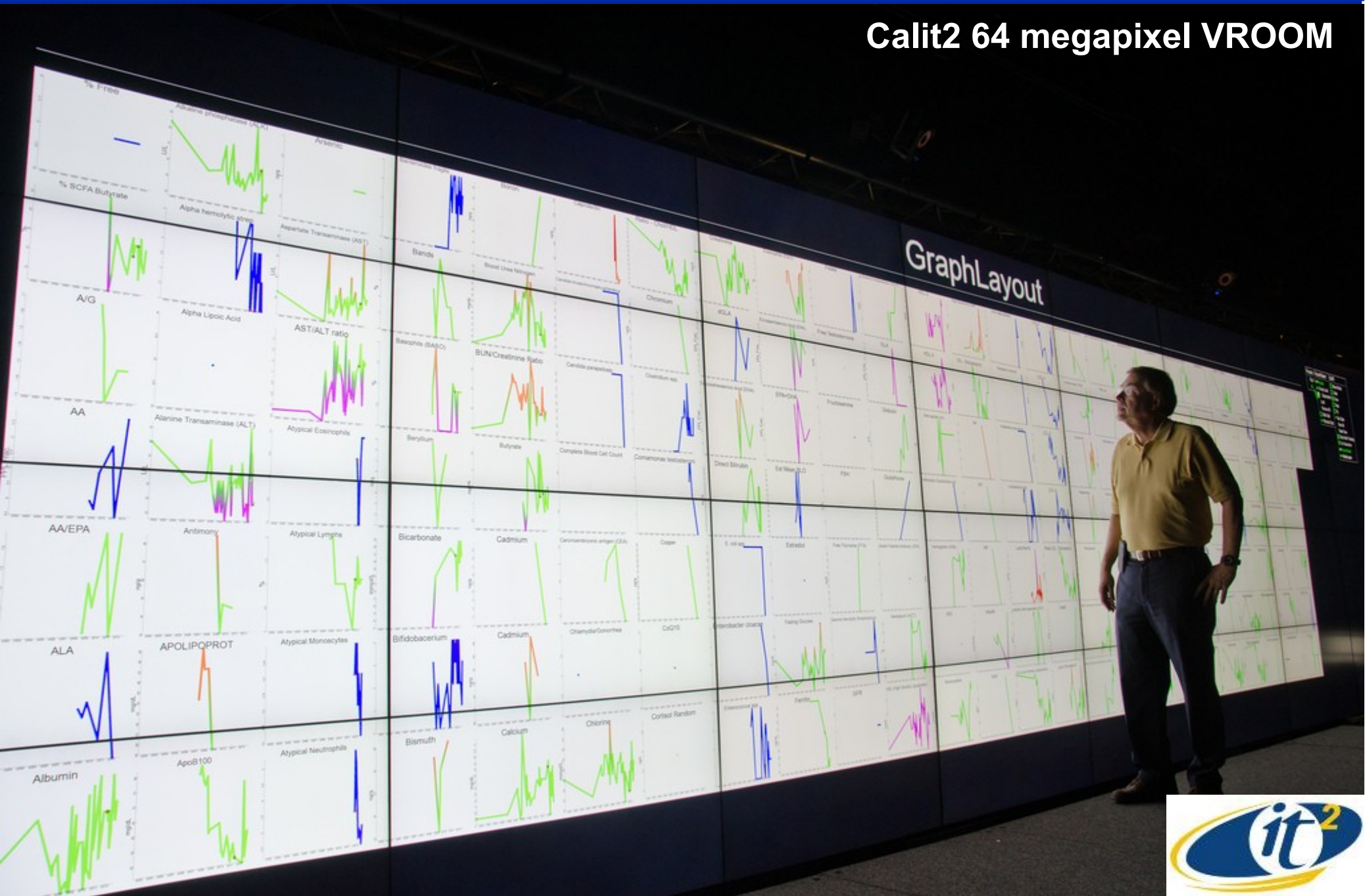


Groups	Members	Interested	Cities	Countries
157	27,814	7,373	116	37

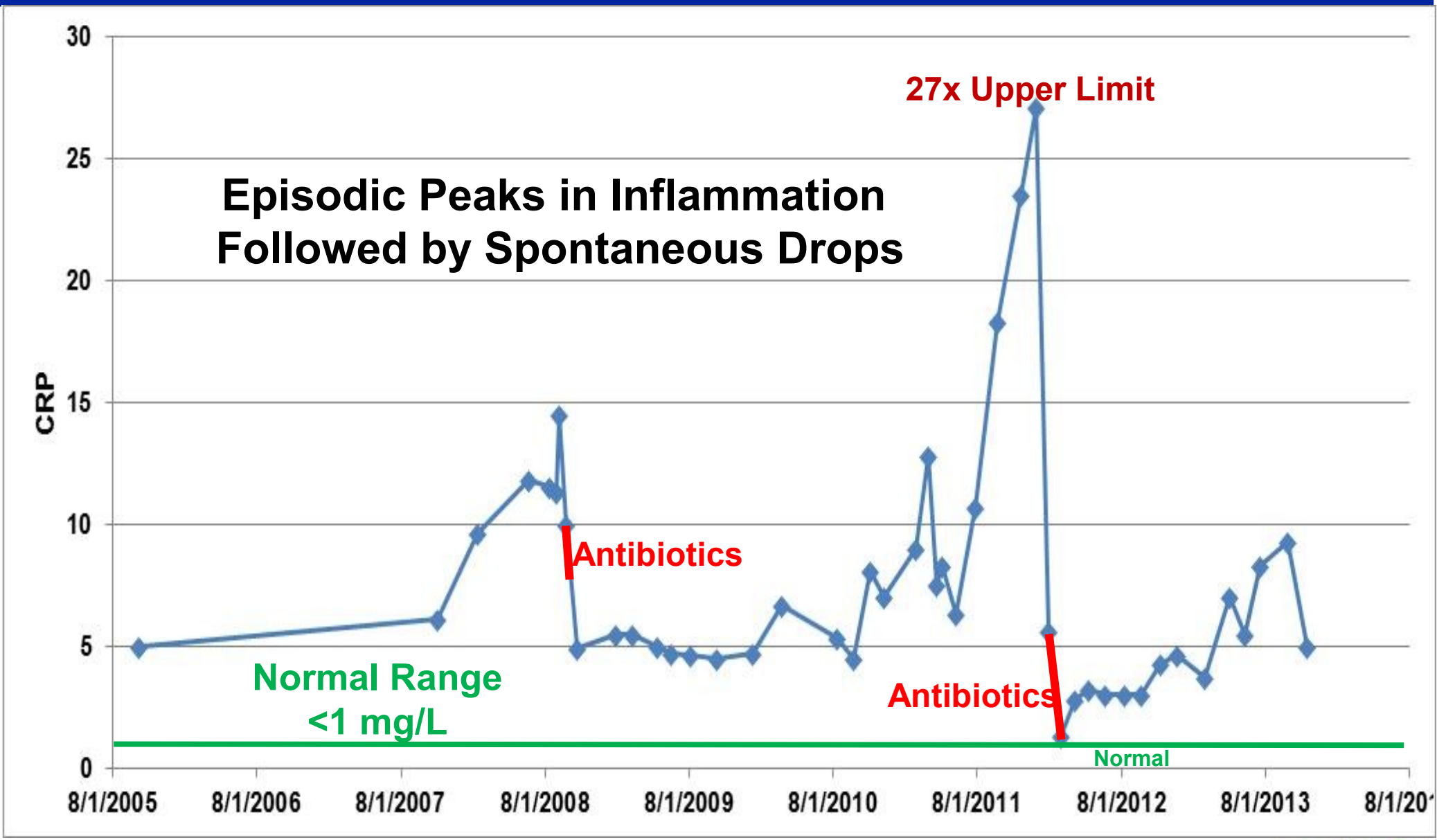
From One to a Billion Data Points Defining Me: Big Data Coming to the Electronic Medical Record (EMR)



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

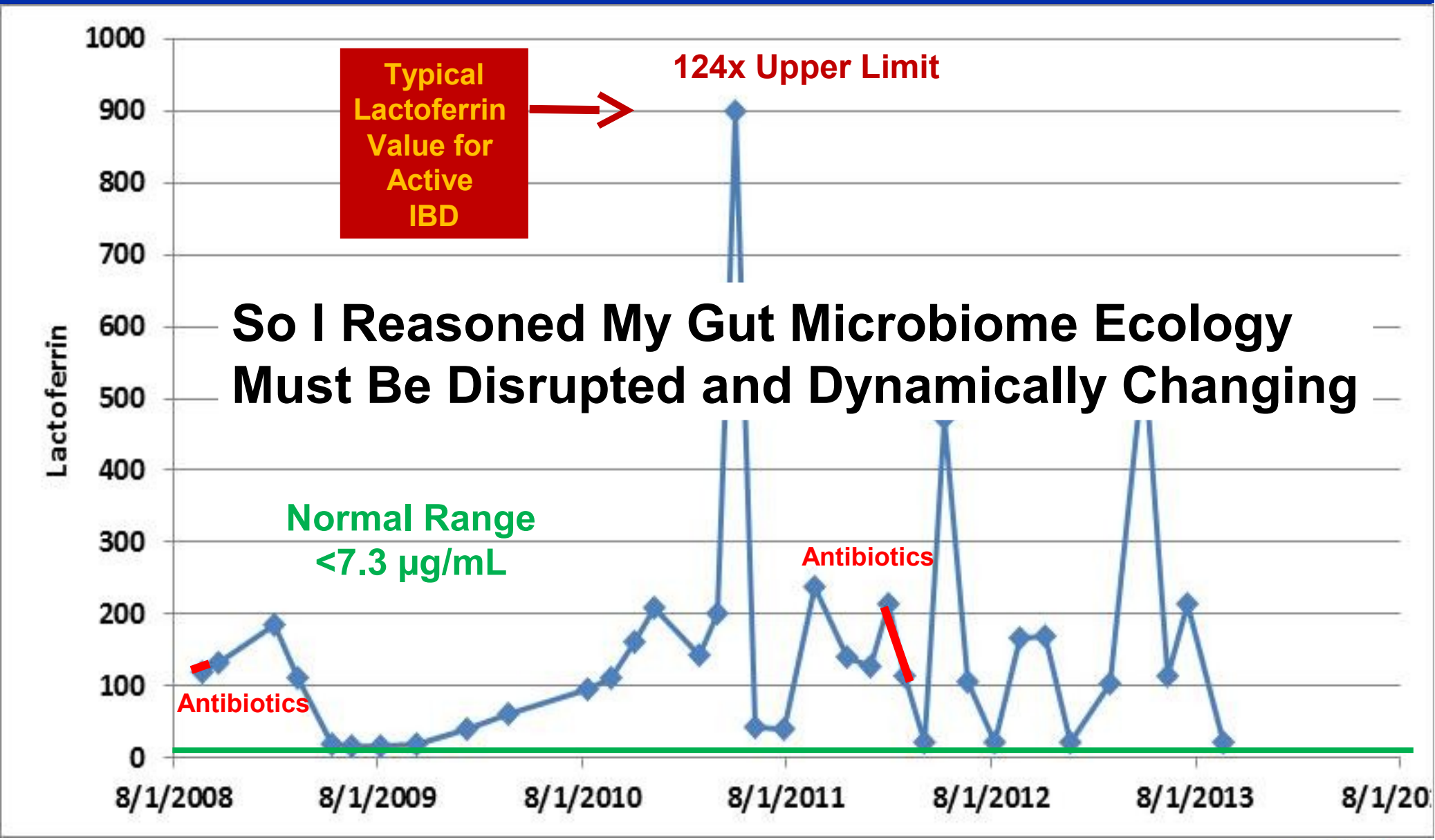


Only One of My Blood Measurements Was Far Out of Range--Indicating Chronic Inflammation



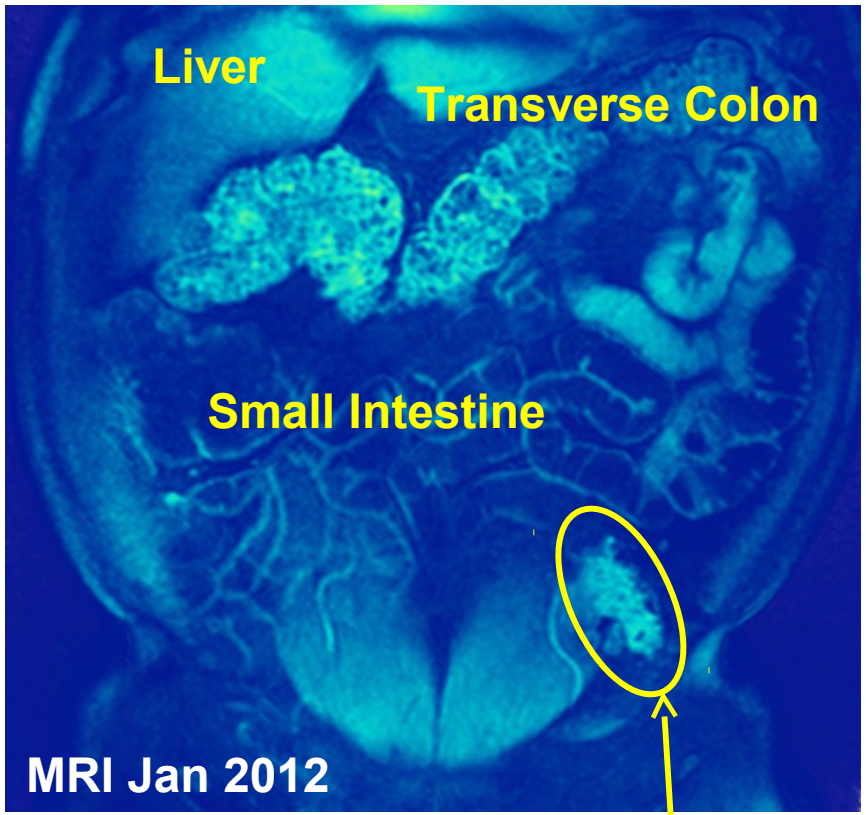
Complex Reactive Protein (CRP) is a Blood Biomarker for Detecting Presence of Inflammation

But by Using Stool Analysis Time Series, I Discovered I Had Episodically Excursions of My Immune System



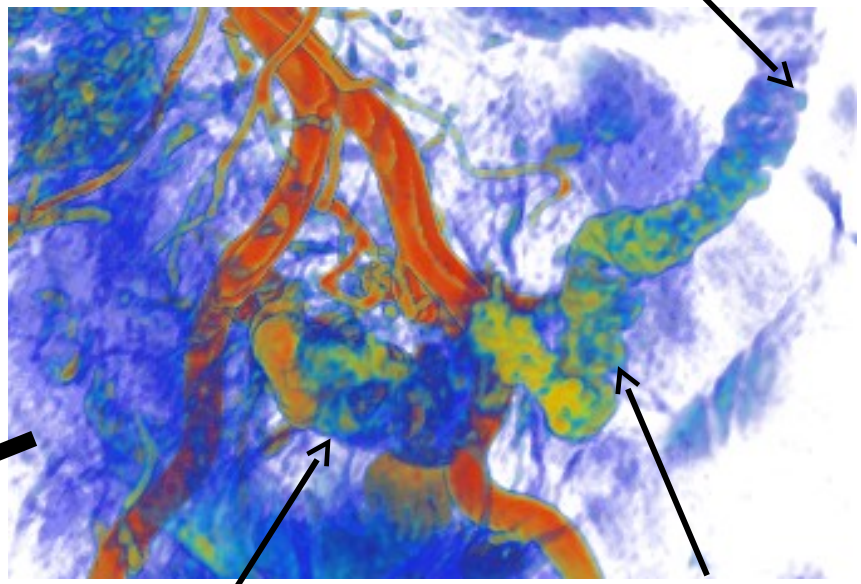
Lactoferrin is a Protein Shed from Neutrophils -
An Immune System Antibacterial that Sequesters Iron

Confirming the IBD Hypothesis: Finding the "Smoking Gun" with MRI Imaging



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

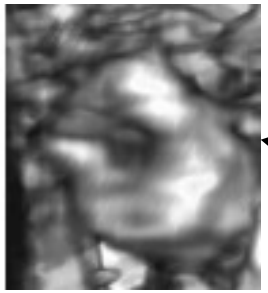
Descending Colon



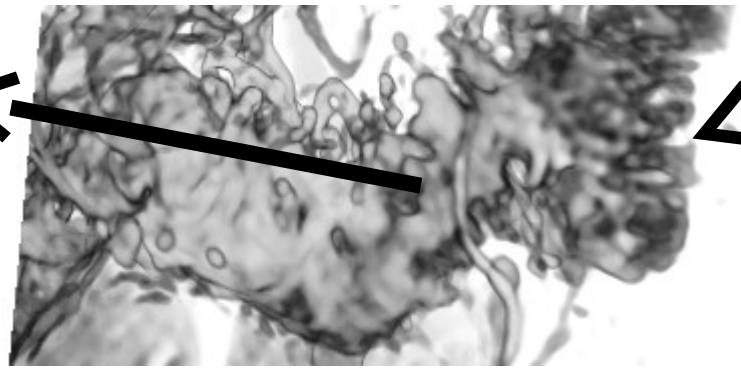
Major Kink

Sigmoid Colon Threading Iliac Arteries

Cross Section



Diseased Sigmoid Colon



Why Did I Have an Autoimmune Disease like IBD?

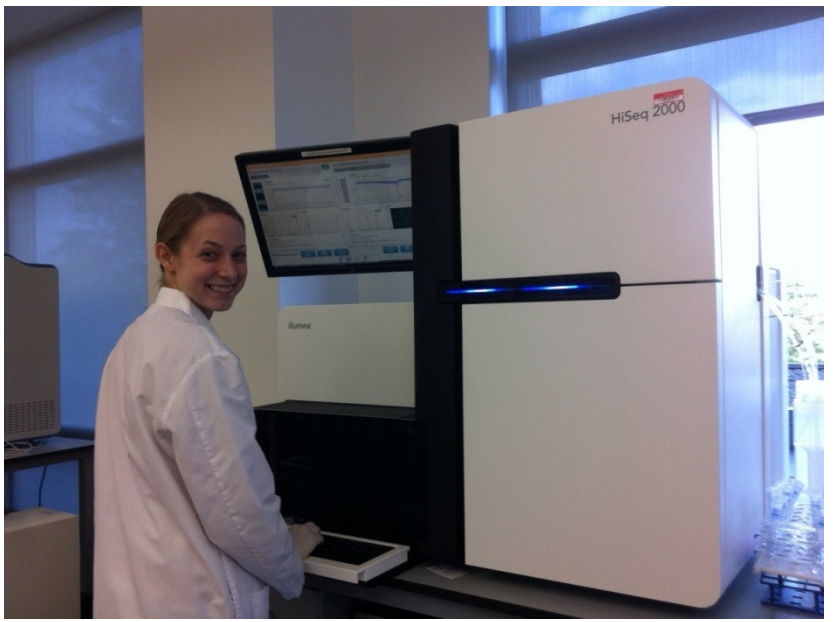
Despite decades of research,
the etiology of Crohn's disease
remains unknown.
Its pathogenesis may involve
a complex interplay between
host genetics,
immune dysfunction,
and microbial or environmental factors.
--*The Role of Microbes in Crohn's Disease*

So I Set Out to Quantify All Three!

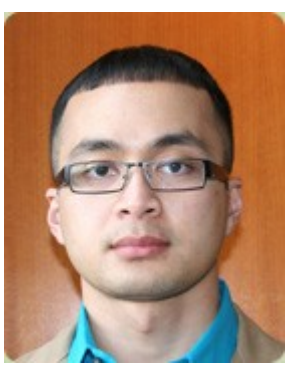
Paul B. Eckburg & David A. Relman
Clin Infect Dis. 44:256-262 (2007)

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

- JCVI Did Metagenomic Sequencing on Six of My Stool Samples Over 1.5 Years
- Sequencing on Illumina HiSeq 2000
  - Generates 100bp Reads
  - Run Takes ~14 Days
  - My 6 Samples Produced
    - 190.2 Gbp of Data
- 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

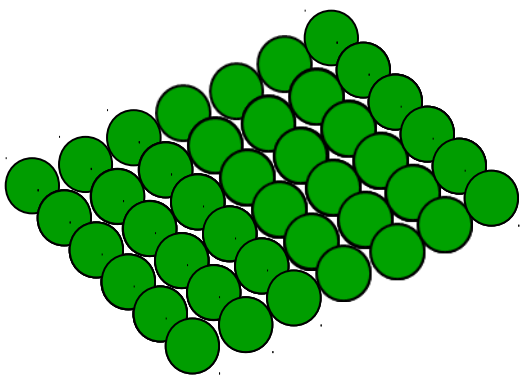
it2

We Downloaded Additional Phenotypes from NIH HMP For Comparative Analysis

Download Raw Reads
~100M Per Person

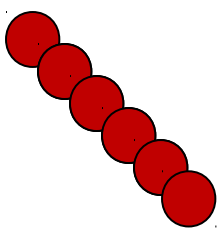
"Healthy" Individuals

35 Subjects
1 Point in Time



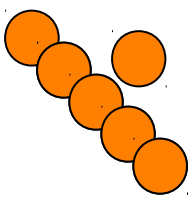
Larry Smarr

6 Points in Time

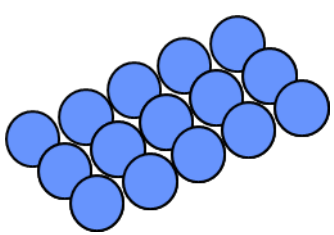


IBD Patients

2 Ulcerative Colitis Patients,
6 Points in Time



5 Ileal Crohn's Patients,
3 Points in Time



NIH

HMP

Total of 5 Billion Reads

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

it²

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

Now to Align Our 5 Billion Reads Against the Reference Database

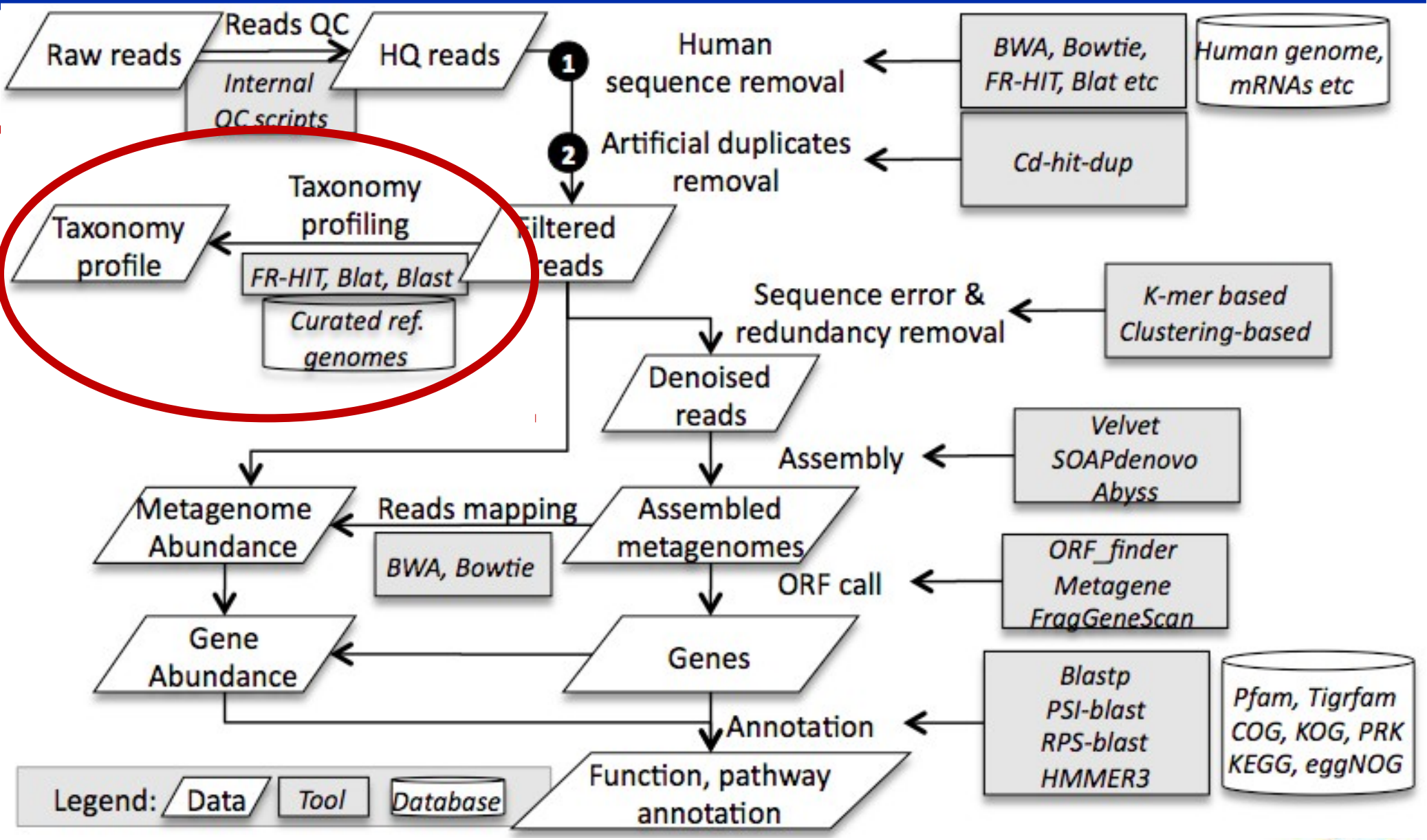
NIH

J. Craig Venter INSTITUTE

Source: Weizhong Li, Sitao Wu, CRBS, UCSD

it²

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

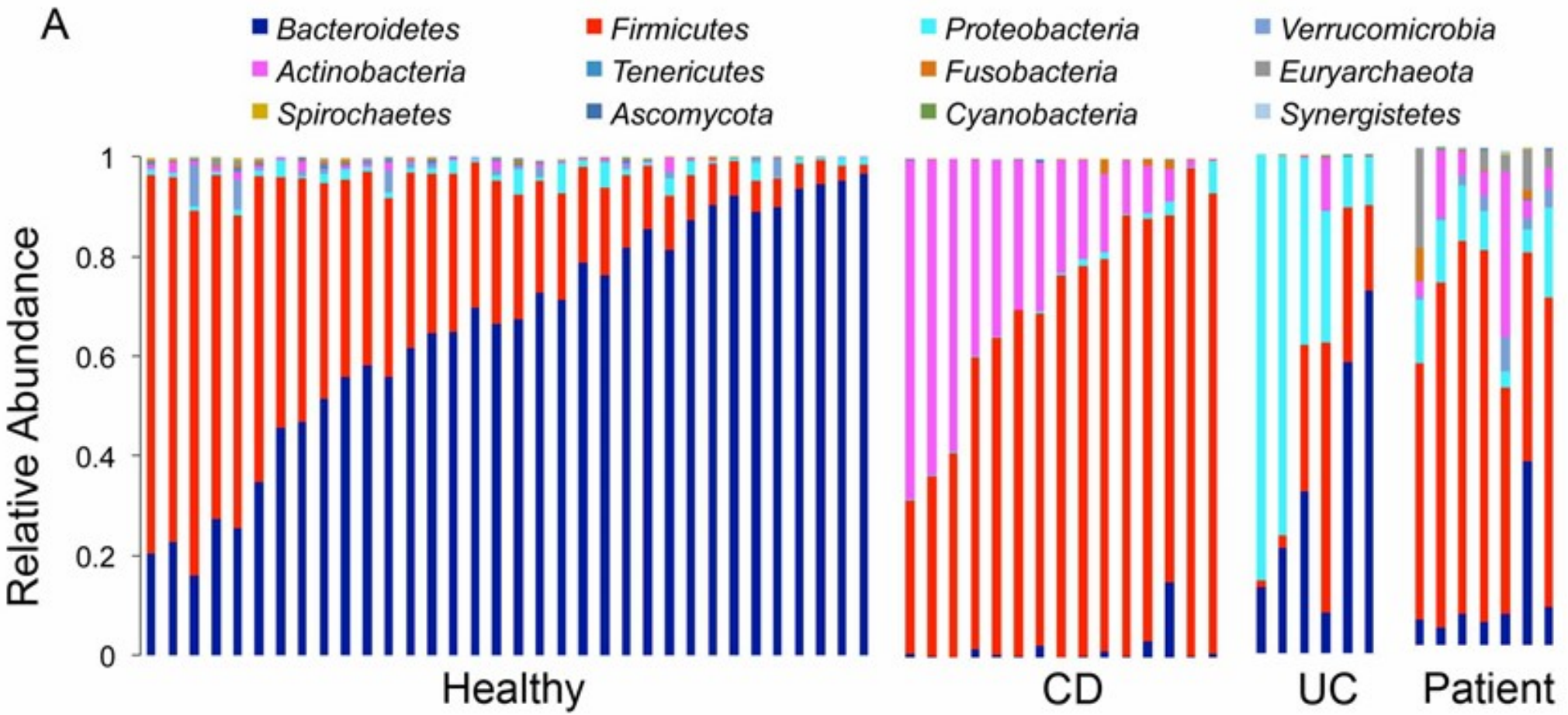
Enabled by
a Grant of Time
on Gordon from SDSC
Director Mike Norman

NSF NIH HMP

J. Craig Venter INSTITUTE

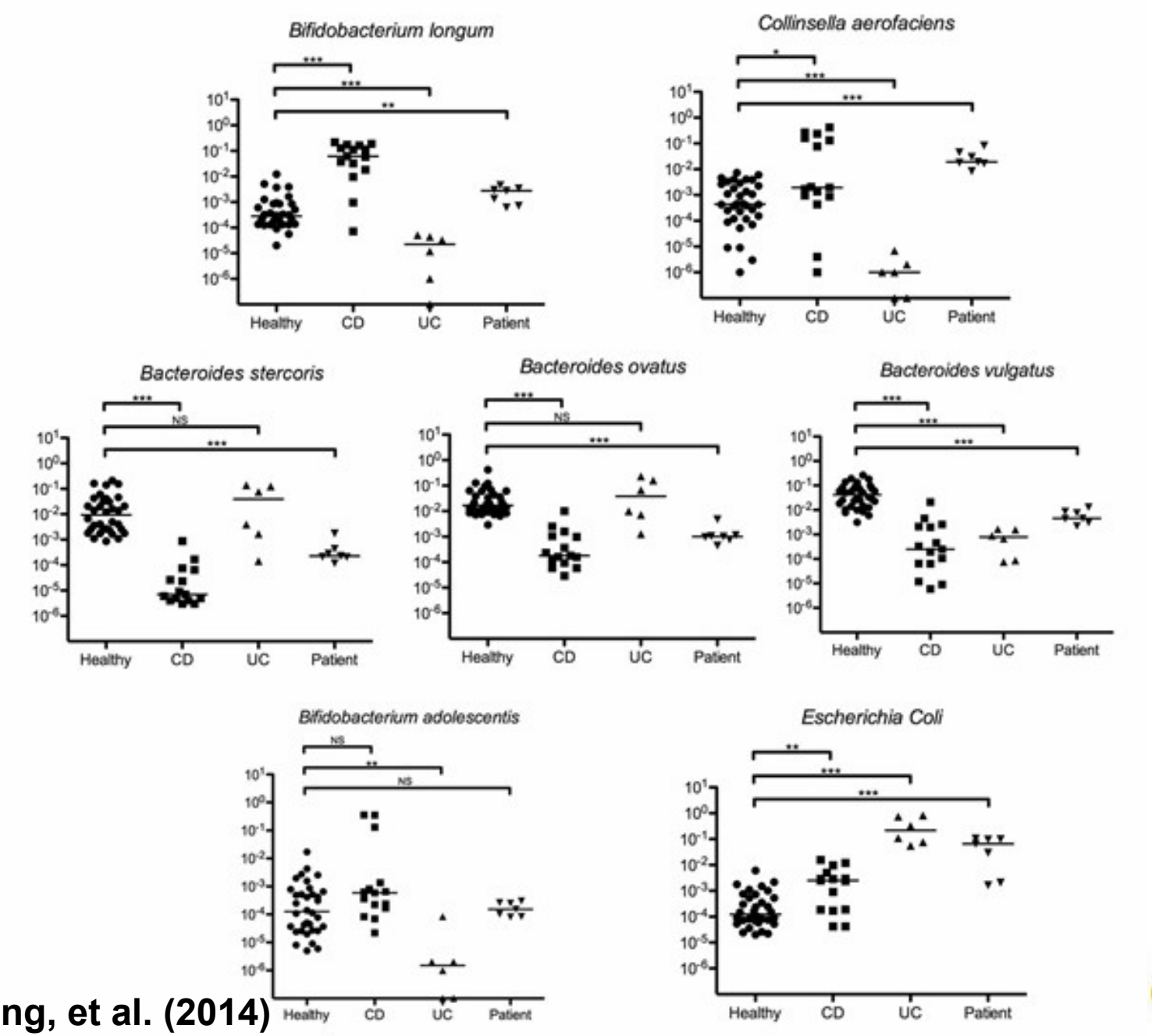
CRBS it²

The Emergence of Microbial Genomics Diagnostics



Source: Chang, et al. (2014)

Bacterial Species Which PCA Indicates Best Separate the Four States



Source: Chang, et al. (2014)

We Used Dell's Supercomputer (Sanger) to Analyze additional 219 HMP and 110 MetaHIT samples

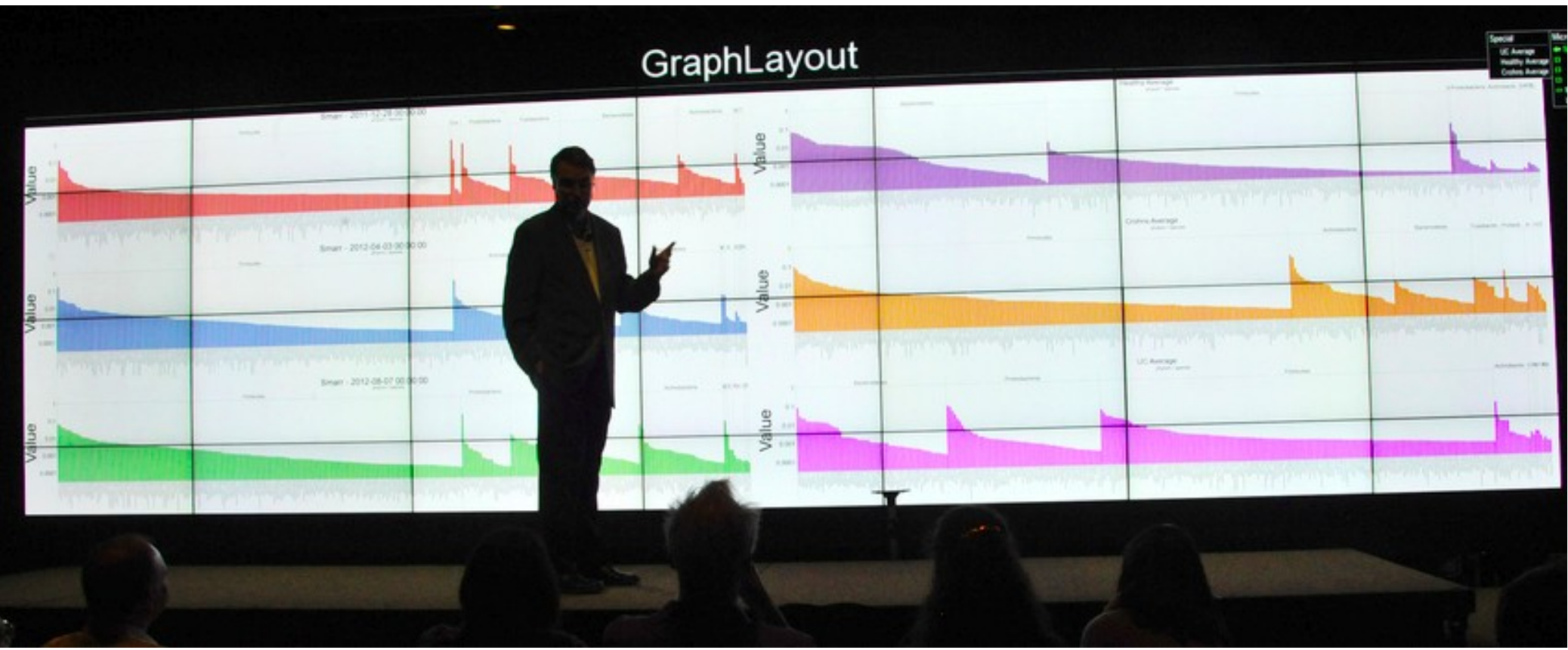
- **Dell's Sanger cluster**
  - **32 nodes, 512 cores,**
  - **48GB RAM per node**
  - **50GB SSD local drive, 390TB Lustre file system**
- **We used faster but less sensitive method with a smaller reference DB (duo to available 48GB RAM)**
- **Only processed to taxonomy mapping**
  - **~35,000 Core-Hrs on Dell's Sanger**
  - **30 TB data**



Dell Solutions Center
Industry Solutions Lab
SANGER DSU

Source: Weizhong Li, UCSD

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

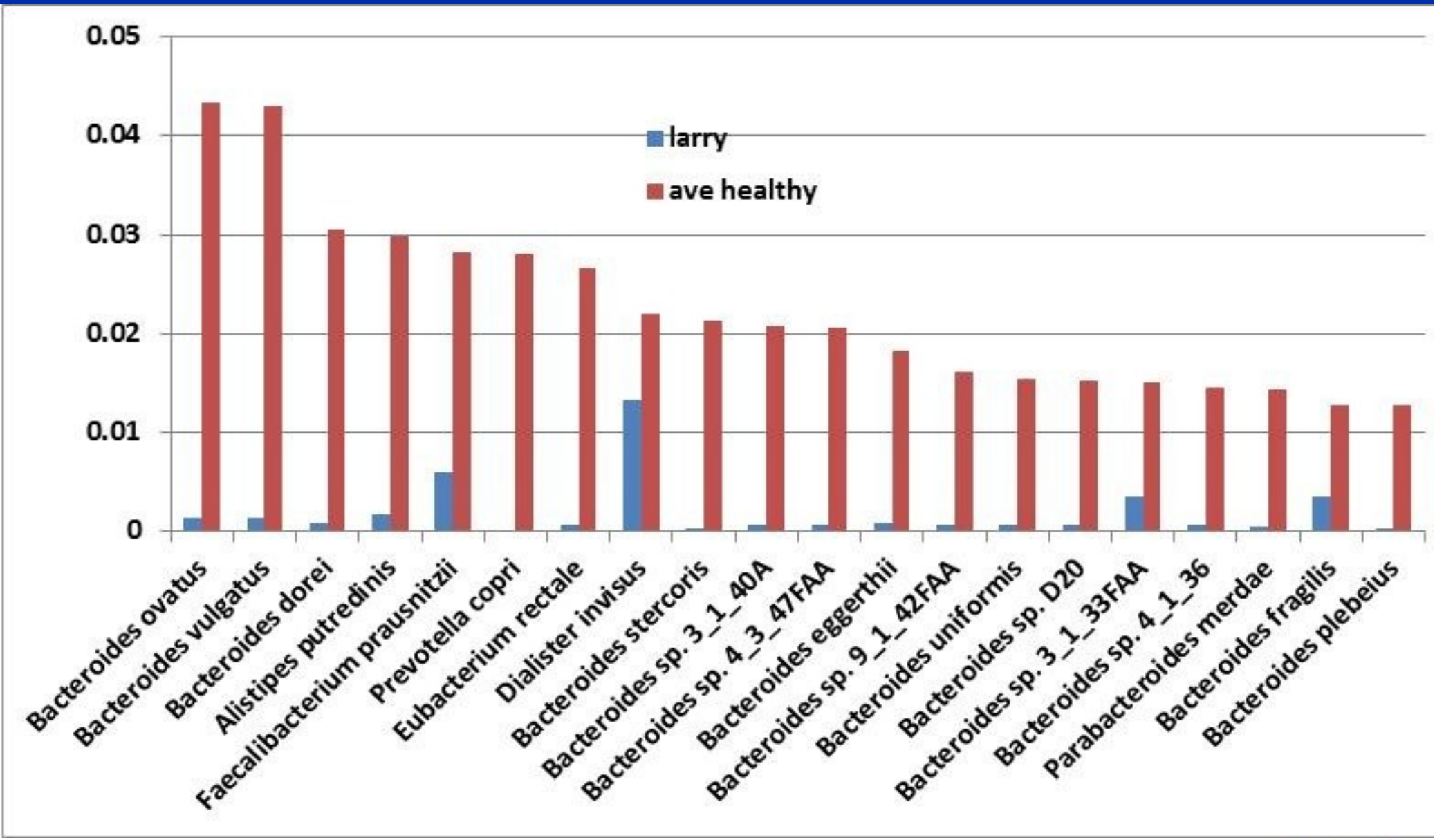
Lessons From Ecological Dynamics
Invasive Species Dominate After Major Species Destroyed



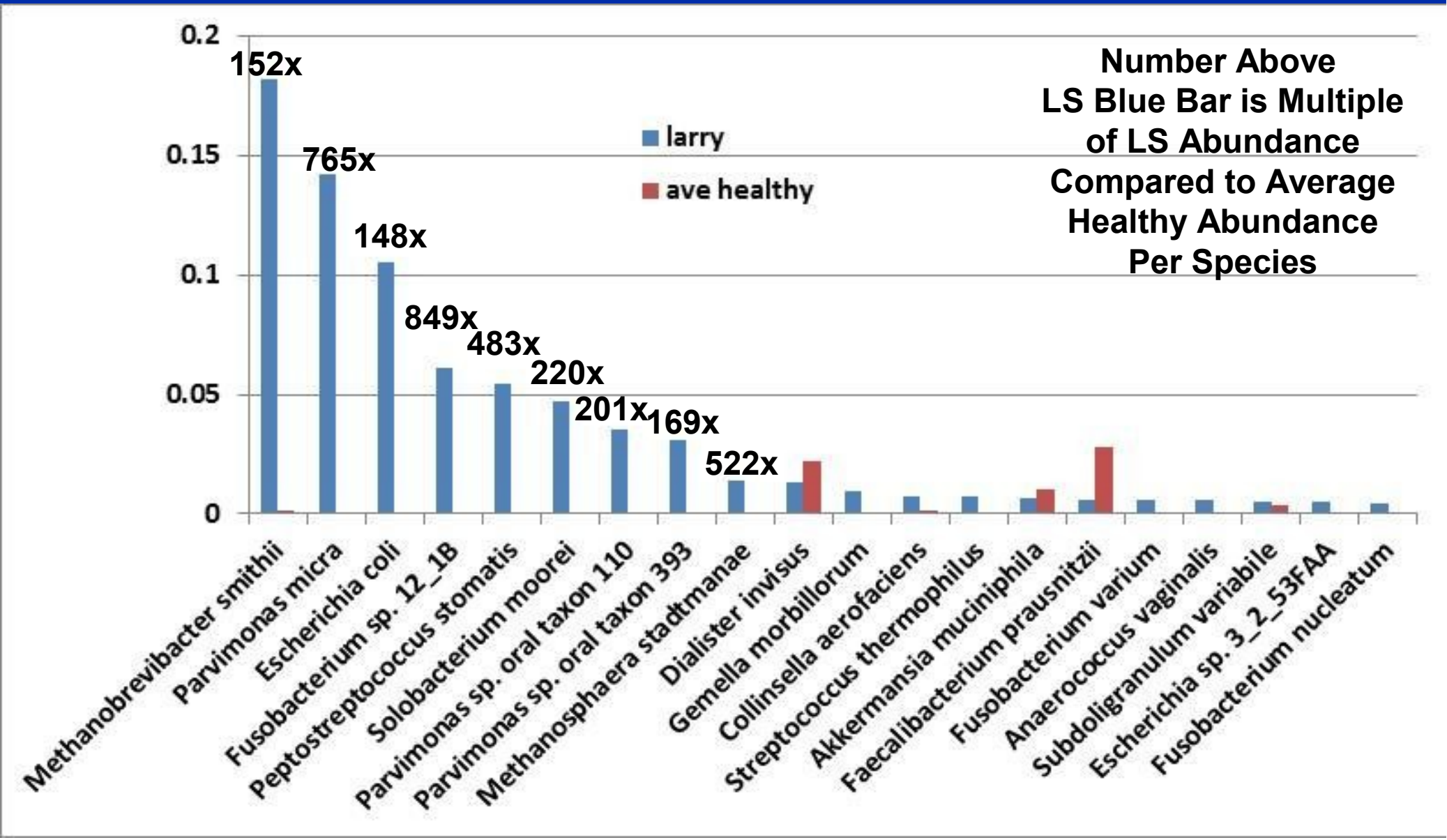
”In many areas following these burns invasive species are able to establish themselves, crowding out native species.”

Source: Ponderosa Pine Fire Ecology
http://cpluhna.nau.edu/Biota/ponderosafire.htm

Almost All Abundant Species (≥1%) in Healthy Subjects Are *Severely Depleted* in Larry's Gut Microbiome

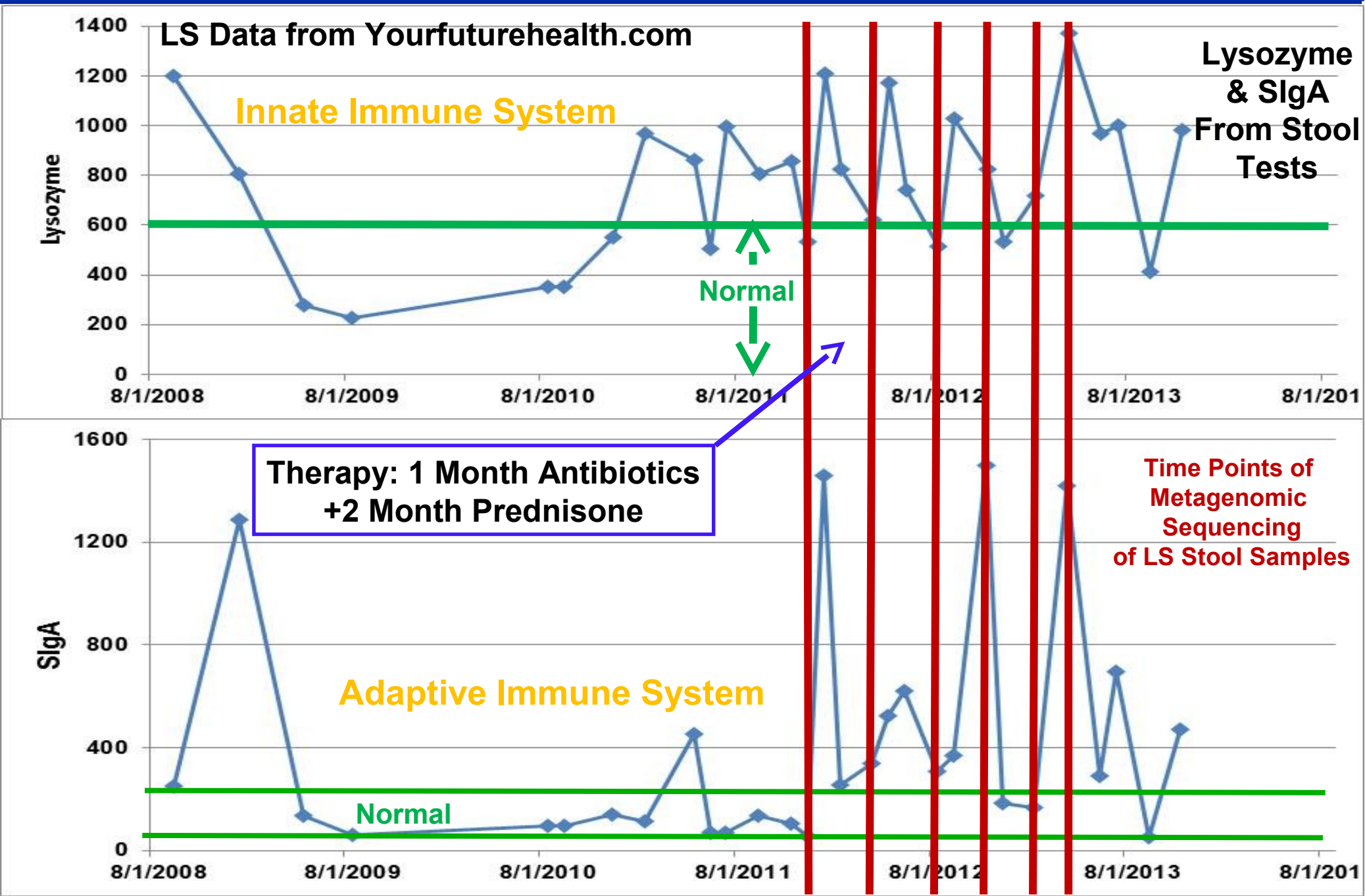


Top 20 Most Abundant Microbial Species In LS vs. Average Healthy Subject

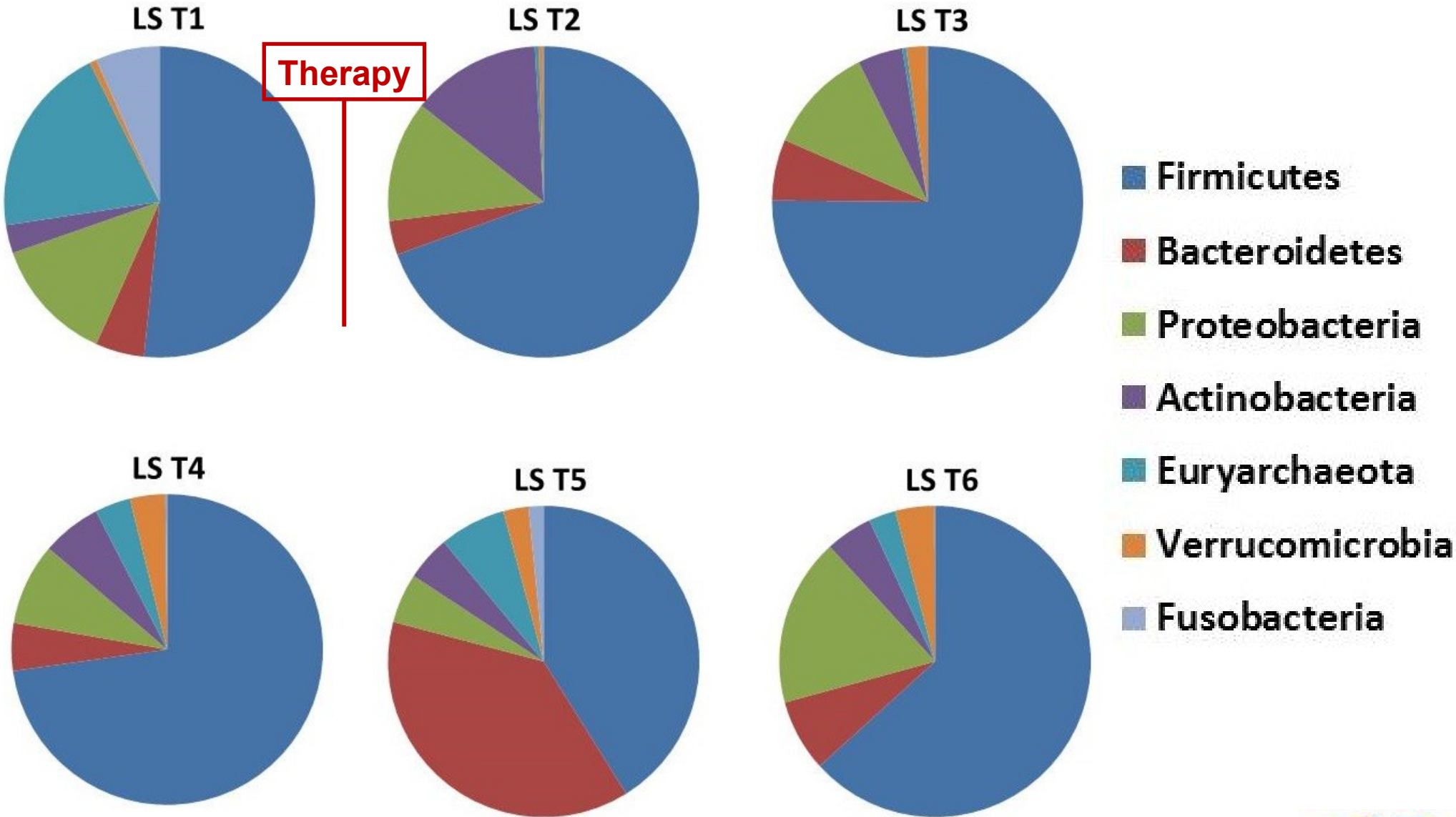


Source: Sequencing JCVI; Analysis Weizhong Li, UCSD
LS December 28, 2011 Stool Sample

Comparing Changes in Gut Microbiome Ecology with Oscillations of the Innate and Adaptive Immune System



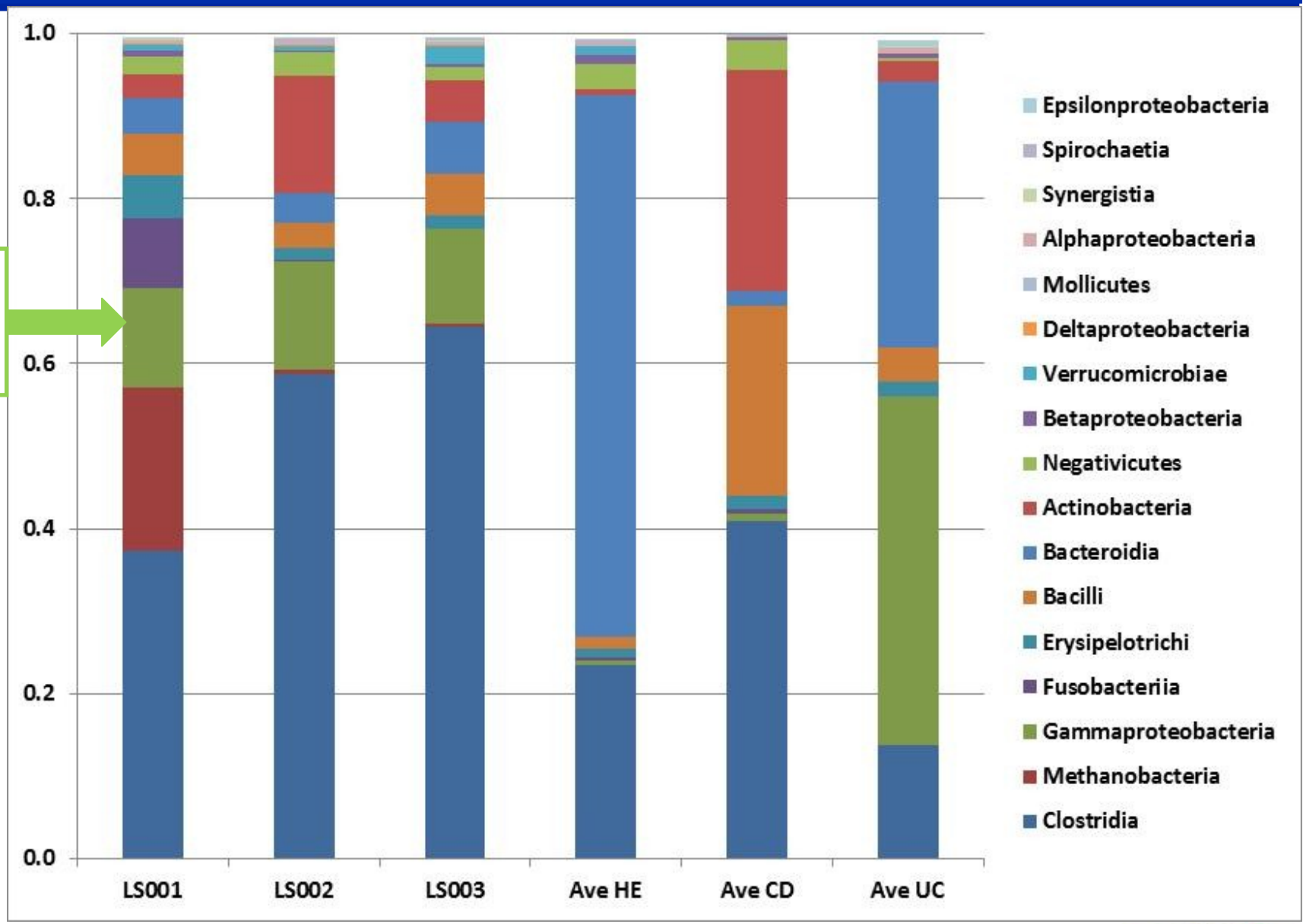
Time Series Reveals Autoimmune Dynamics of Gut Microbiome by Phyla



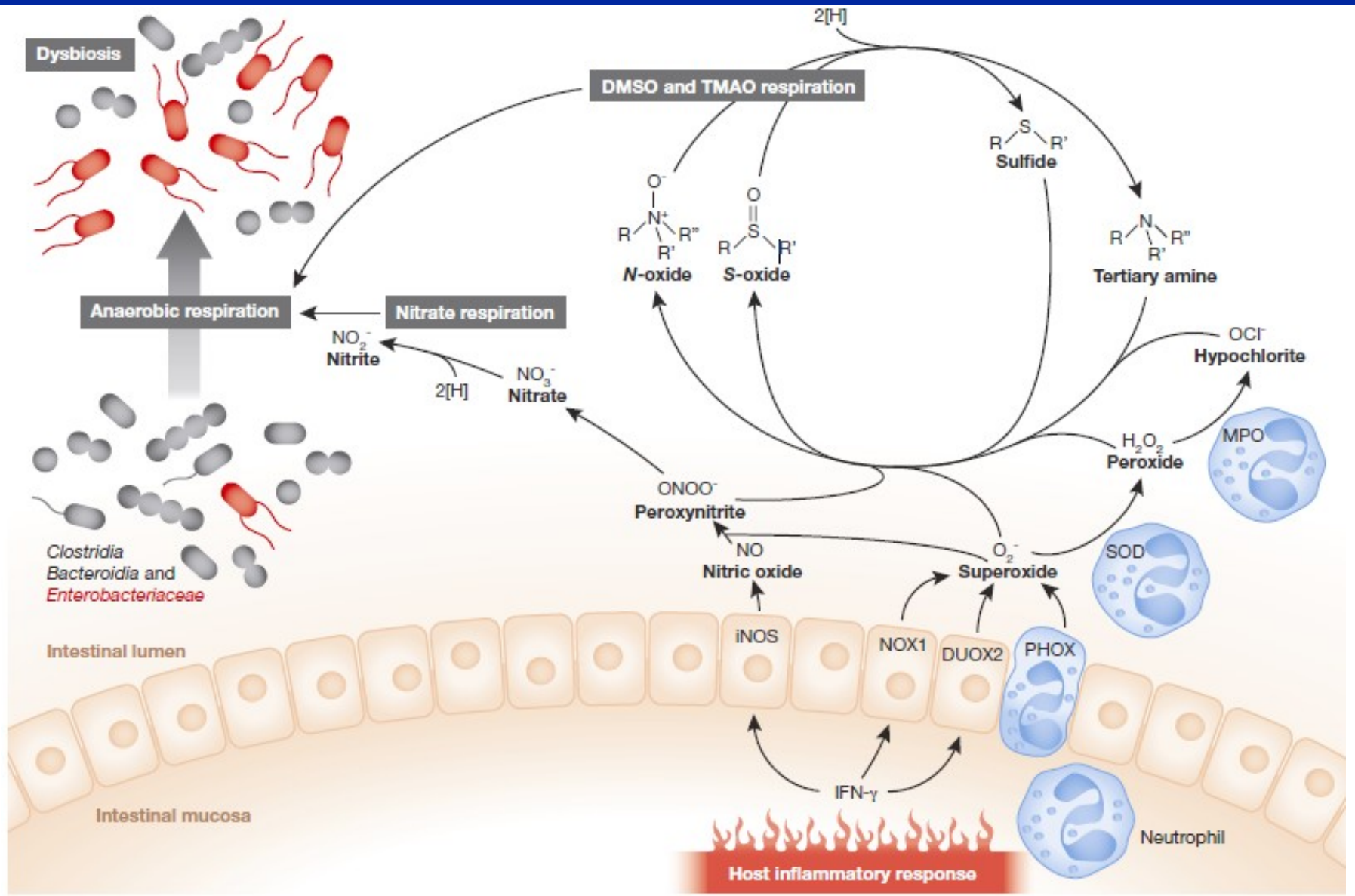
Six Metagenomic Time Samples Over 16 Months

LS Time Series Gut Microbiome Classes vs. Healthy, Crohn's, Ulcerative Colitis

Class Gamma-proteobacteria



Inflammation Enables Anaerobic Respiration Which Leads to Phylum-Level Shifts in the Gut Microbiome



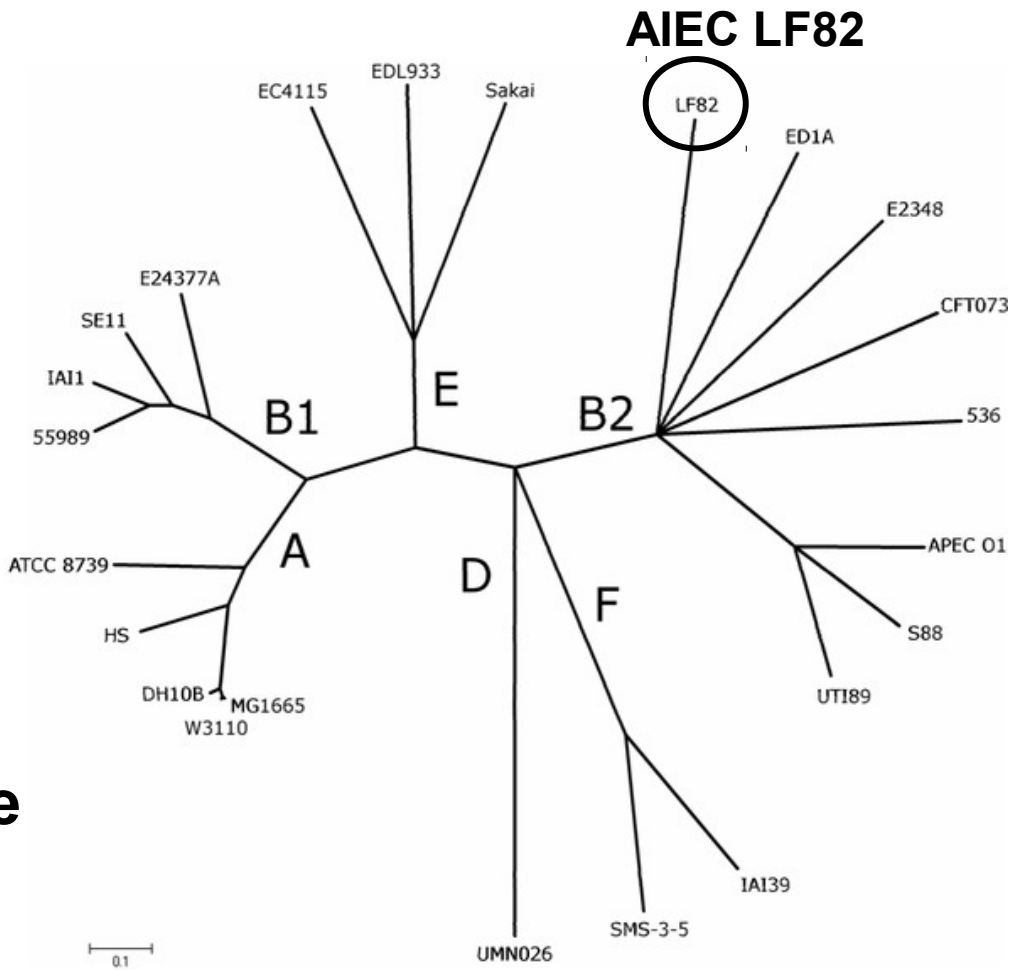
Sebastian E. Winter, Christopher A. Lopez & Andreas J. Bäumler, EMBO reports VOL 14, p. 319-327 (2013)

Does Intestinal Inflammation Select for Pathogenic Strains That Can Induce Further Damage?

"Adherent-invasive E. coli (AIEC) are isolated more commonly from the intestinal mucosa of individuals with Crohn's disease than from healthy controls."

"Thus, the mechanisms leading to dysbiosis might also select for intestinal colonization with more harmful members of the Enterobacteriaceae* —such as AIEC— thereby exacerbating inflammation and interfering with its resolution."

Sebastian E. Winter , et al.,
EMBO reports VOL 14, p. 319-327 (2013)

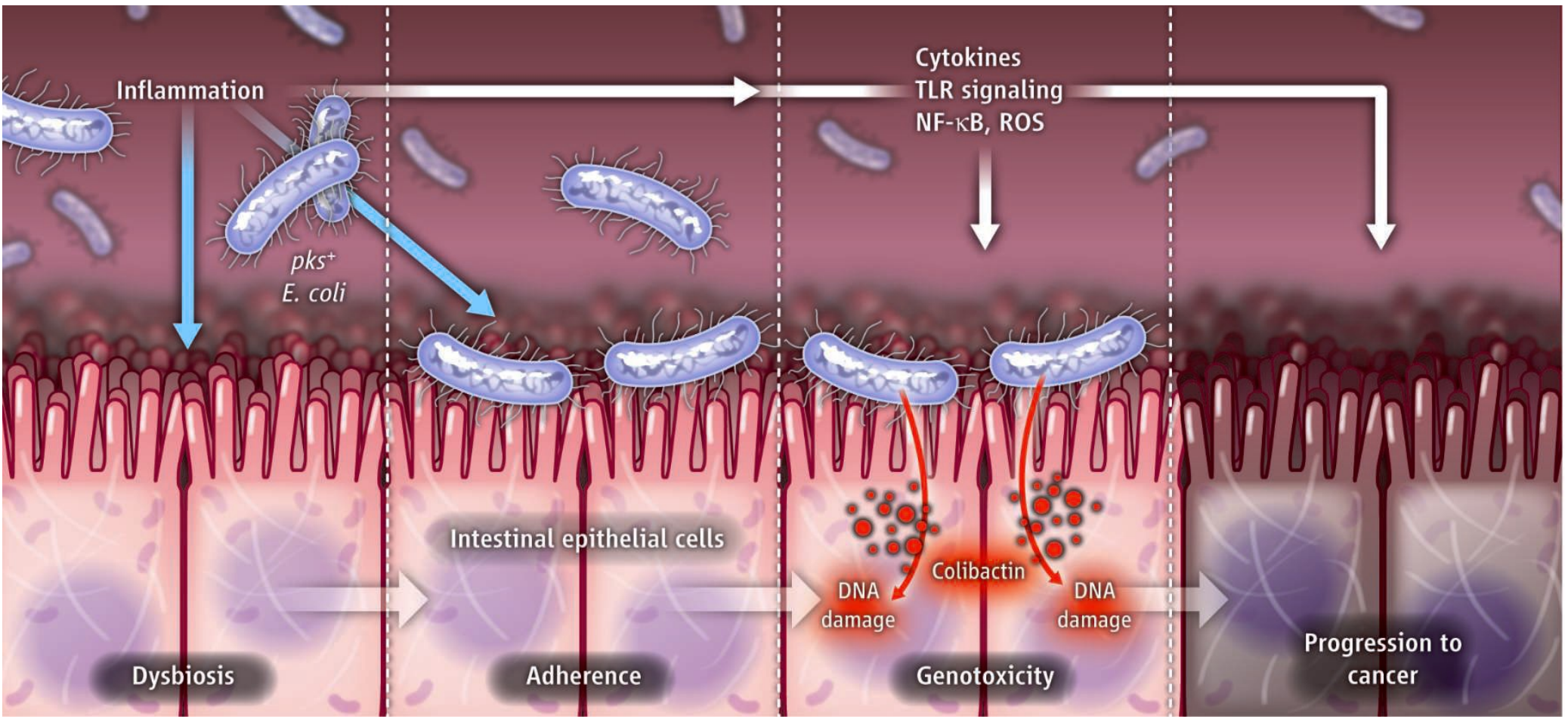


E. coli/Shigella Phylogenetic Tree
Miquel, et al.
PLOS ONE, v. 5, p. 1-16 (2010)

*Family Containing E. coli

Chronic Inflammation Can Accumulate Cancer-Causing Bacteria in the Human Gut

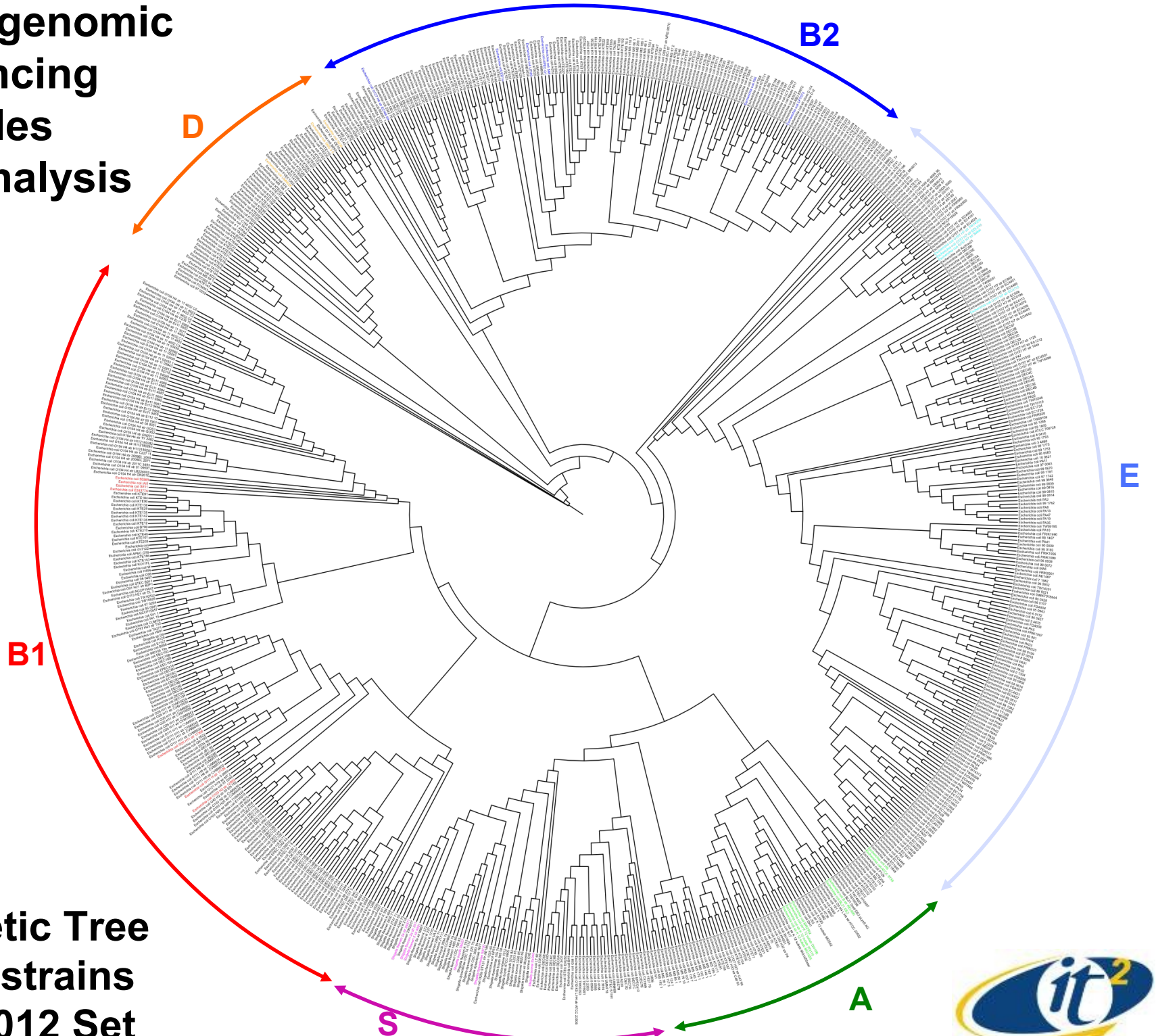
Escherichia coli Strain NC101



Science
AAAS

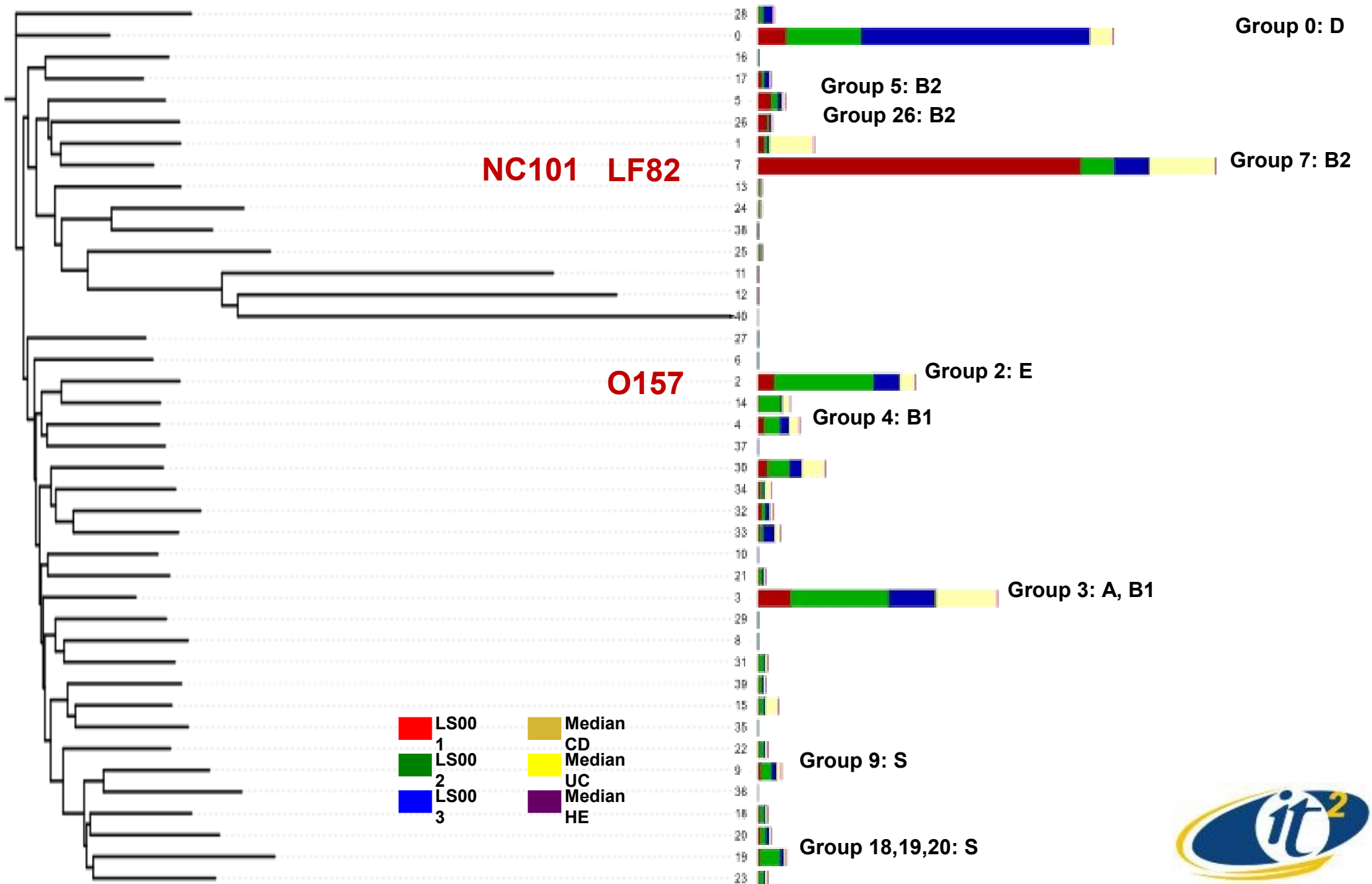
Bacteria Deliver a Genotoxic Hit
Robert F. Schwabe and Timothy C. Wang
Science **338**, 52 (2012);
DOI: 10.1126/science.1229905

Deep Metagenomic Sequencing Enables Strain Analysis

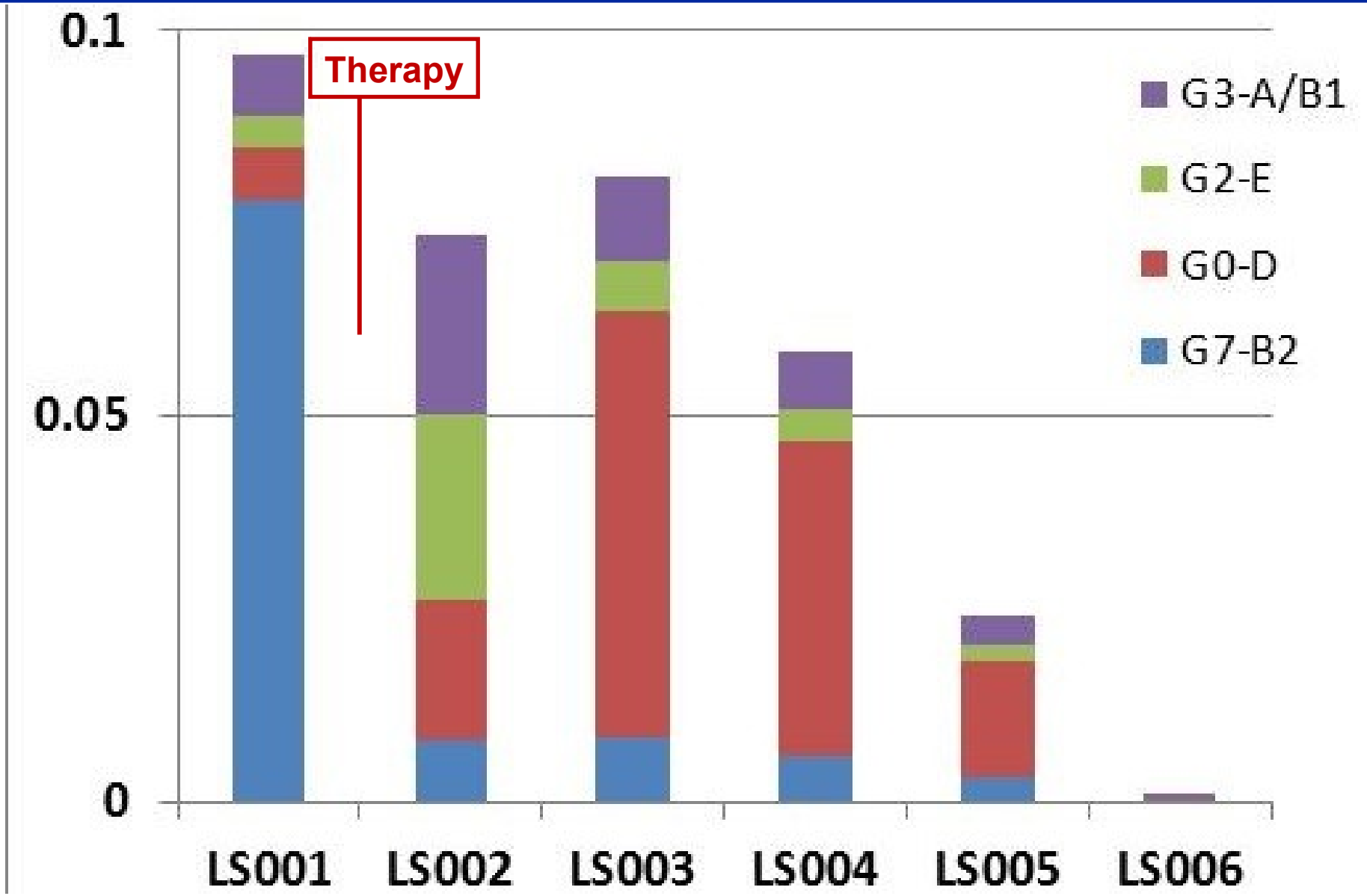


Phylogenetic Tree
778 Ecoli strains
=6x our 2012 Set

We Divided the 778 E. coli Strains into 40 Groups, Each of Which Had 80% Identical Genes



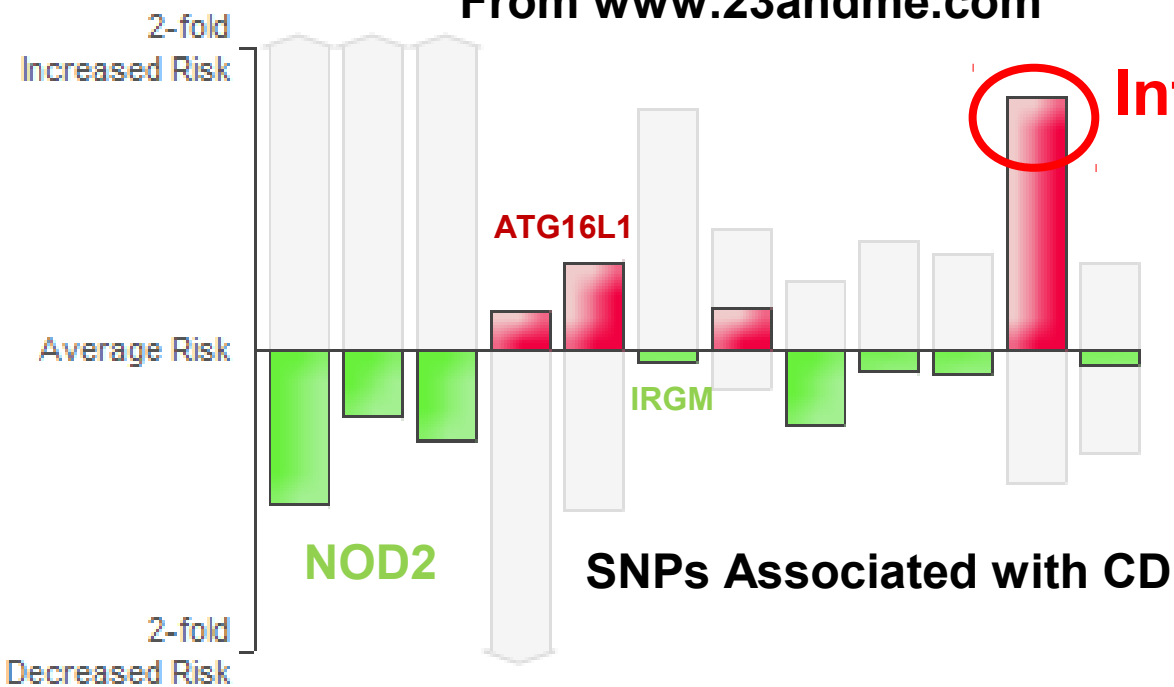
Reduction in E. coli Over Time
With Major Shifts in Strain Abundance



Strains >0.5% Included

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)

News and Views

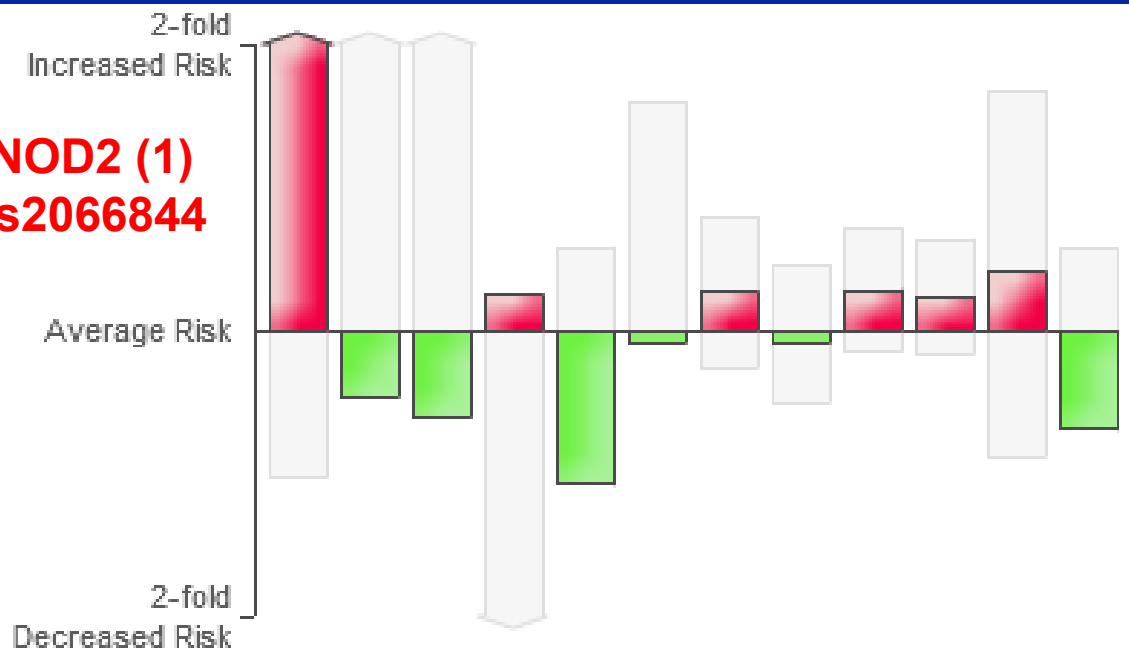
Nature Medicine **13**, 26 - 28 (2007)
doi:10.1038/nm0107-26

IL-23: a master regulator in Crohn disease

Markus F Neurath[1]

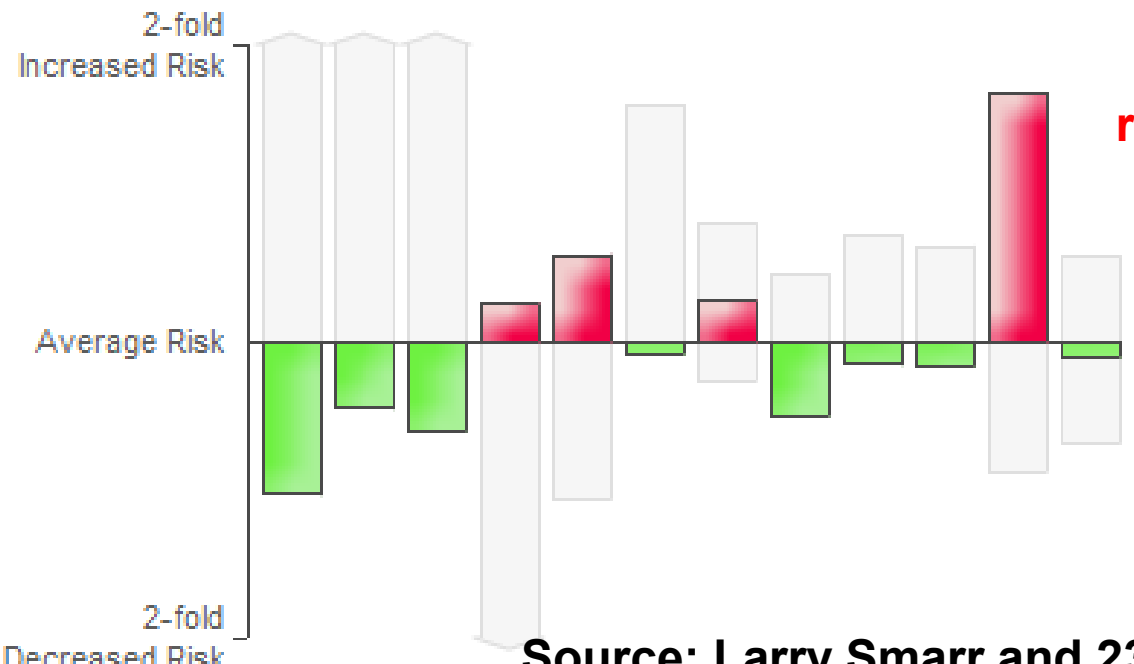
There Is Likely a Correlation Between CD SNPs and Where and When the Disease Manifests

Subject with Ileal Crohn's



Female
CD Onset
At 20-Years Old

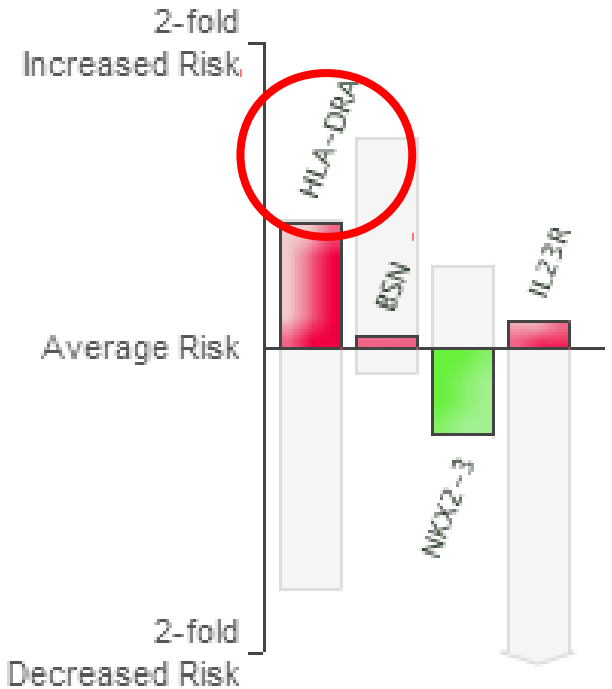
Subject with Colon Crohn's



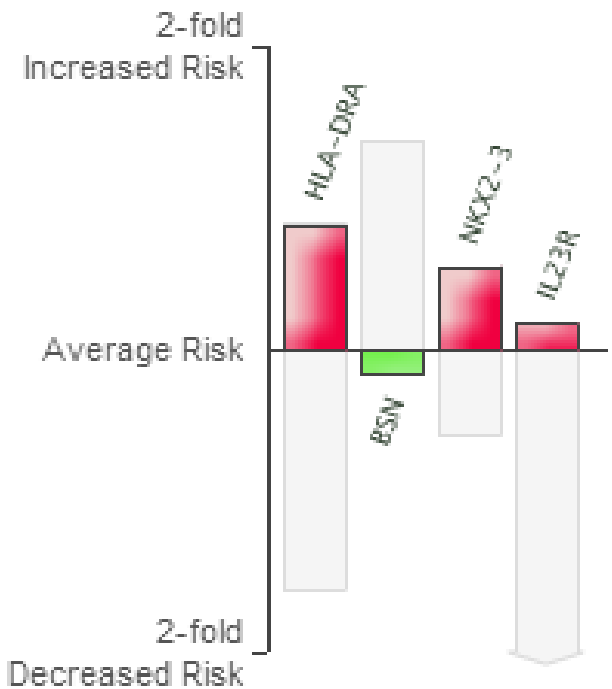
Me-Male
CD Onset
At 60-Years Old

Source: Larry Smarr and 23andme

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



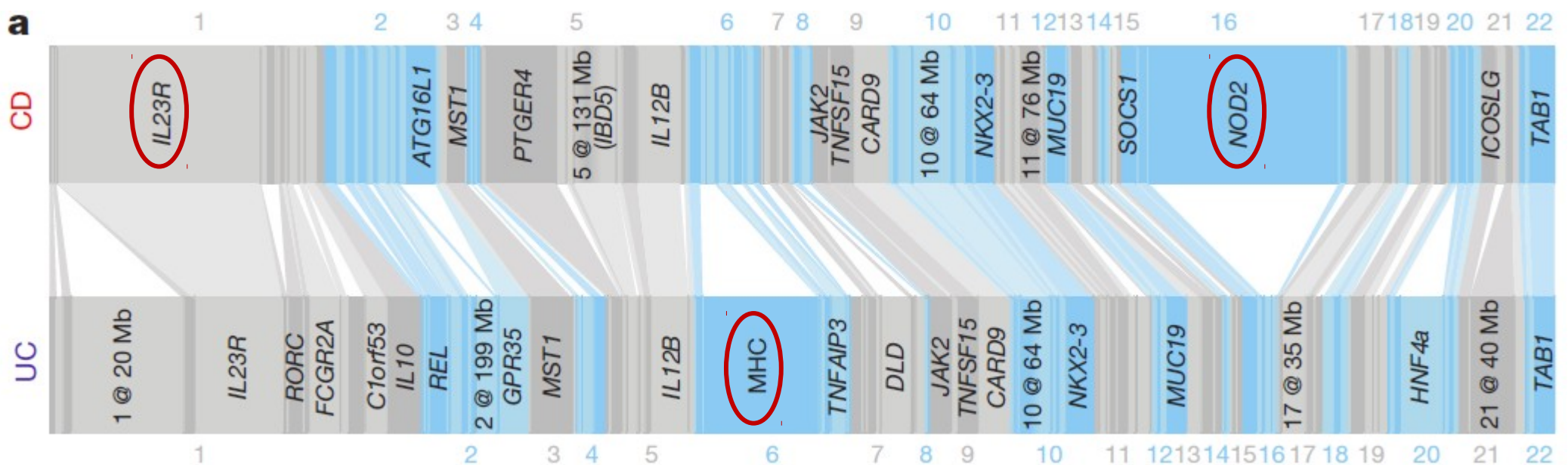
I Have a 33% Increased Risk for Ulcerative Colitis HLA-DRA (rs2395185)



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)

I Compared my 23andme SNPs With the 163 Known SNPs Associated with IBD



- 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)

Now Working with 23andme Comparing 163 Known IBD SNPs with 23andme SNP Chip

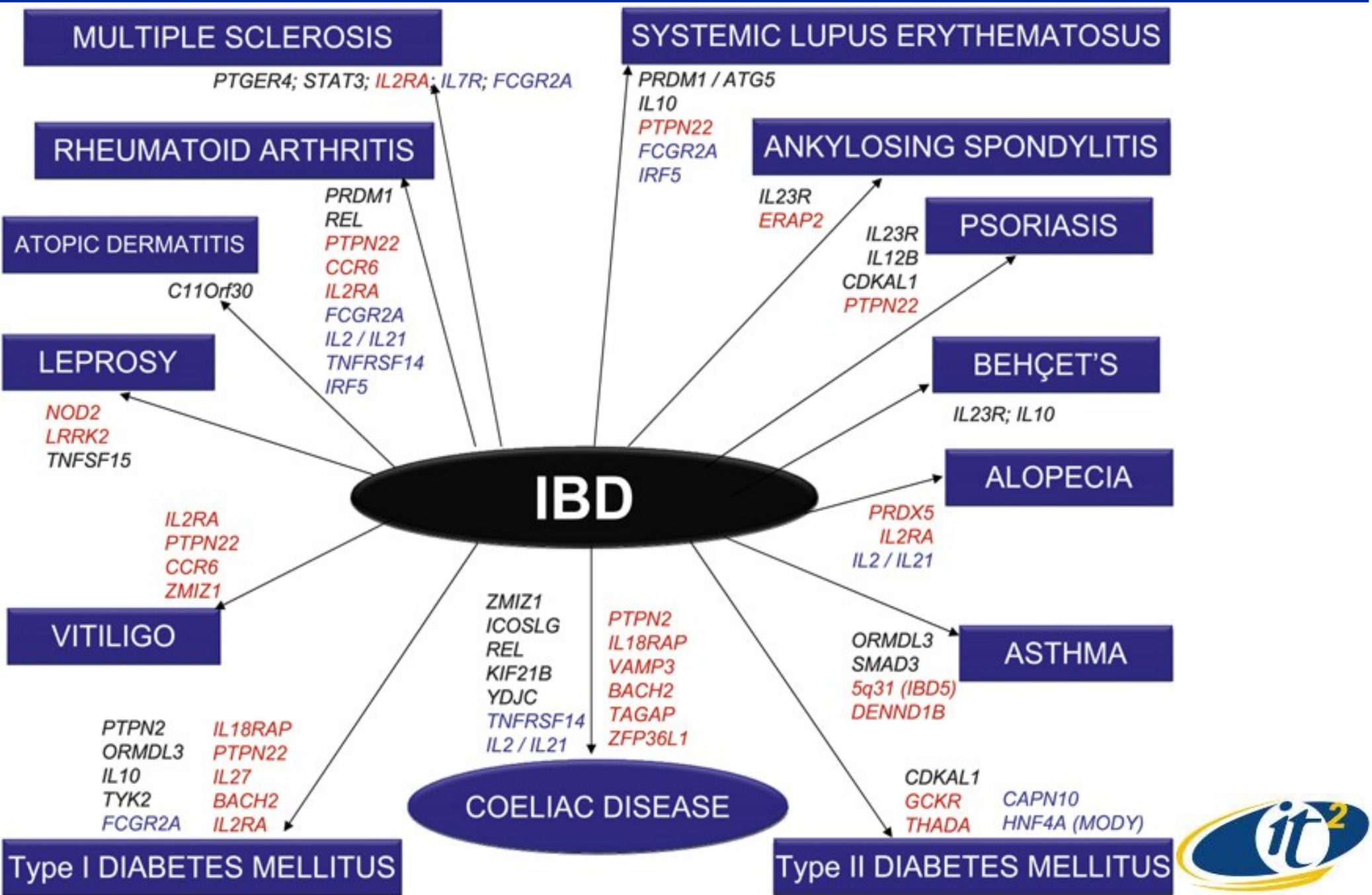
- **Currently 300,000 23andme Members**
  - Growing Rapidly to One Million
- **IBD Affects ~1/300 Americans**
  - Implies ~3000 IBD Subjects
  - Detailed IBD Survey to Members for Phenotyping
- **Enables Internal GWAS**

- **Also Working with Crohnology (Sean Ahrens)**
  - Encouraging His >5000 Crohn's Members to Use 23andme
  - Combine SNPs with Detailed Phenotyping and Drug Impacts



Autoimmune Disease Overlap from SNP GWAS

Gut Lees, et al. 60:1739-1753 (2011)



Large-Scale Genomic Analysis Enabled by SDSC's Gordon

Kristopher Standish[*^], Tristan M. Carland[*], Glenn K. Lockwood[+^], Mahidhar Tatineni[+^], Wayne Pfeiffer[+^], Nicholas J. Schork[*^]

[*] Scripps Translational Science Institute

[+] San Diego Supercomputer Center

[^] University of California San Diego

Project funding provided by Janssen R&D

Scripps Translational Science Institute

SDSC SAN DIEGO SUPERCOMPUTER CENTER

janssen PHARMACEUTICAL COMPANIES OF Johnson & Johnson

A Large-Scale Human Genome Trial

- **Janssen R&D Performed Whole-Genome Sequencing on 438 Patients Undergoing Treatment for Rheumatoid Arthritis**
- **Problem: Correlate Response or Non-Response to Drug Therapy with Genetic Variants**
- **Solution Combines Multi-Disciplinary Expertise**
  - **Genomic Analytics from Janssen R&D and Scripps Translational Science Institute (STSI)**
  - **Data-Intensive Computing from San Diego Supercomputer Center (SDSC)**

Source: Wayne Pfeiffer, SDSC

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

Big Data Technical Challenges

- **Data Volume: Raw Reads from 438 Full Human Genomes**
  - 50 TB of Compressed Data from Janssen R&D
  - Encrypted on 8x 6 TB SATA RAID Enclosures
- **Compute: Perform Read Mapping and Variant Calling on All Genomes**
  - 9-Step Pipeline to Achieve High-Quality Read Mapping
  - 5-Step Pipeline to do Group Variant Calling for Analysis
- **Project requirements:**
  - FAST Turnaround (Assembly in < 2 Months)
  - EFFICIENT (Minimum Core-Hours Used)

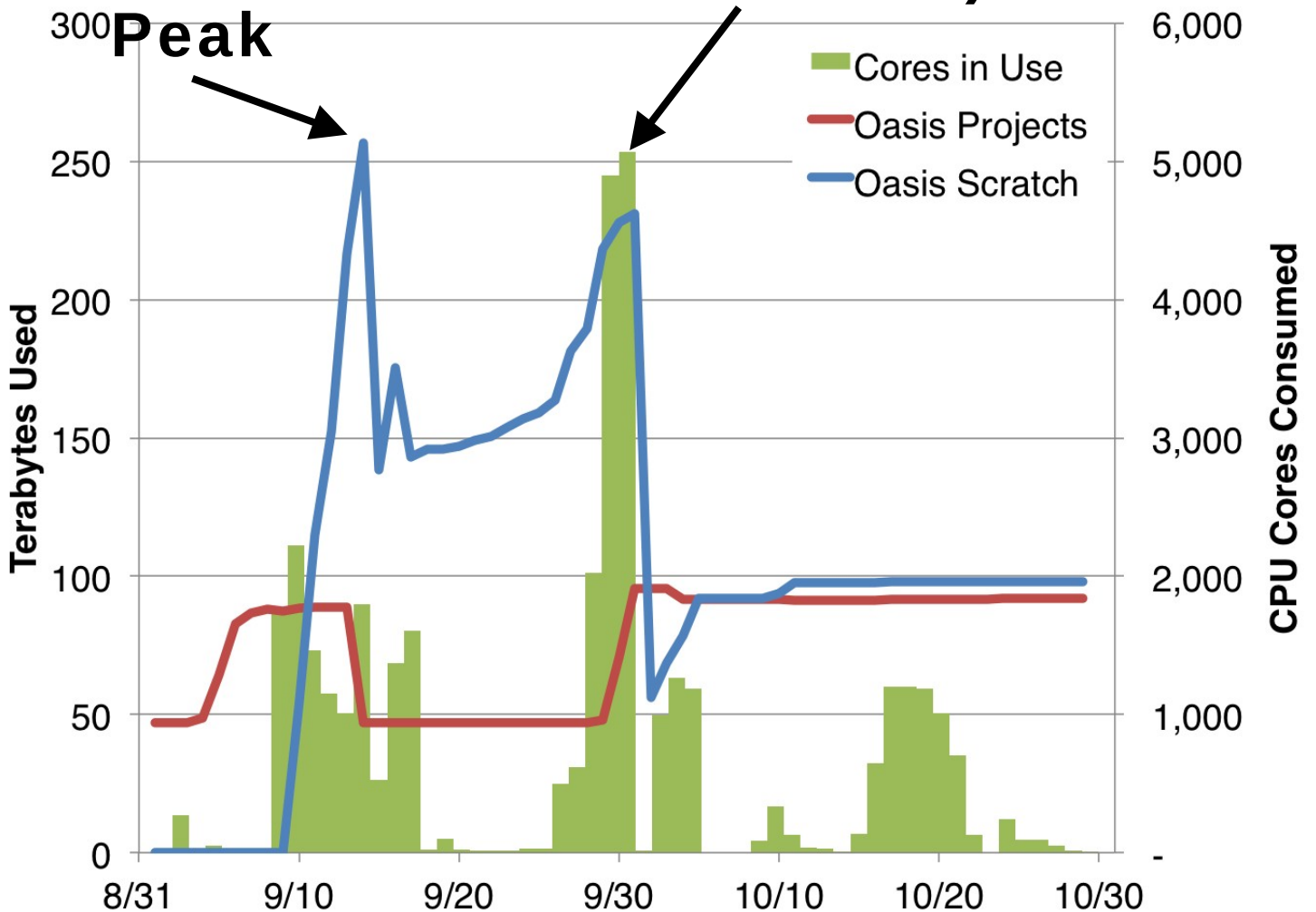
Source: Wayne Pfeiffer, SDSC

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

Footprint on Gordon: CPUs and Storage Used

257 TB Lustre Scratch Used at Peak

5,000 cores (30% of Gordon) in Use at Once

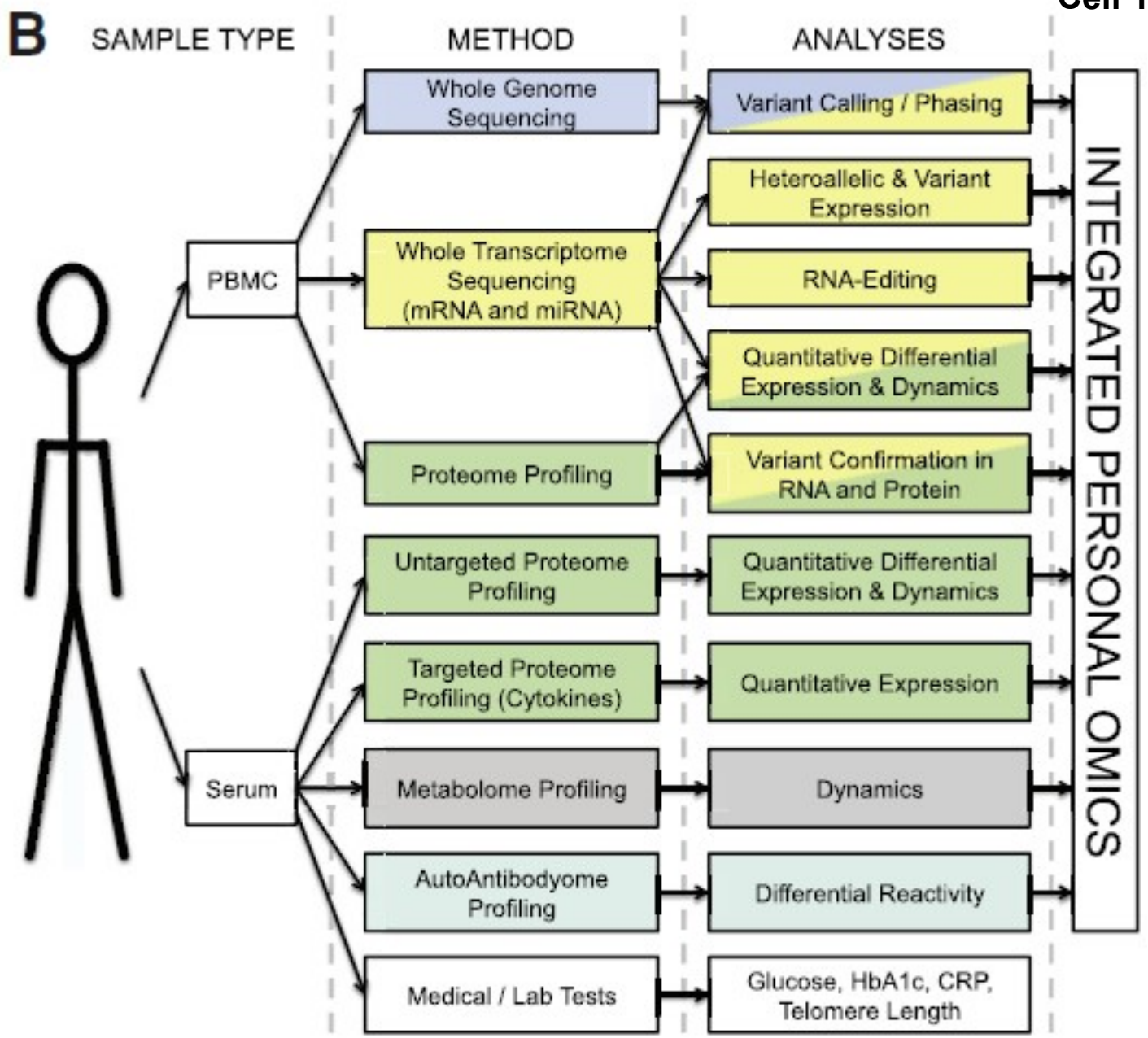


Source: Wayne Pfeiffer, SDSC

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

Integrative Personal Omics Profiling Reveals Details of Clinical Onset of Viruses and Diabetes

Cell 148, 1293–1307, March 16, 2012



- **Michael Snyder, Chair of Genomics Stanford Univ.**
- **Genome 140x Coverage**
- **Blood Tests 20 Times in 14 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**

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



www.delsaglobal.org

www.personalgenomes.org

it²

From N=1 Hypothesis Generation to N=100 Prospective Time Series Clinical Studies

- **Mike Snyder, Dept. of Genetics, Stanford Univ.**
  - **250 Pre-Diabetic Patients**
- **Lee Hood, Institute for Systems Biology**
  - **100 Person Wellness Project**
- **William Sandborn, School of Medicine, UC San Diego**
  - **150 Subjects, 50 Healthy, 50 UC, 50 CD**

I am a Subject in Each of These Studies

Thanks to Our Great Team!

UCSD Metagenomics Team

Weizhong Li
Sitao Wu

JCVI Team

Karen Nelson
Shibu Yooseph
Manolito Torralba

Calit2@UCSD Future Patient Team

Jerry Sheehan
Tom DeFanti
Kevin Patrick
Jurgen Schulze
Andrew Prudhomme
Philip Weber
Fred Raab
Joe Keefe
Ernesto Ramirez



SDSC Team

Michael Norman
Mahidhar Tatineni
Robert Sinkovits

UCSD Health Sciences Team

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