Sequencing Genomics: The New Big Data Driver

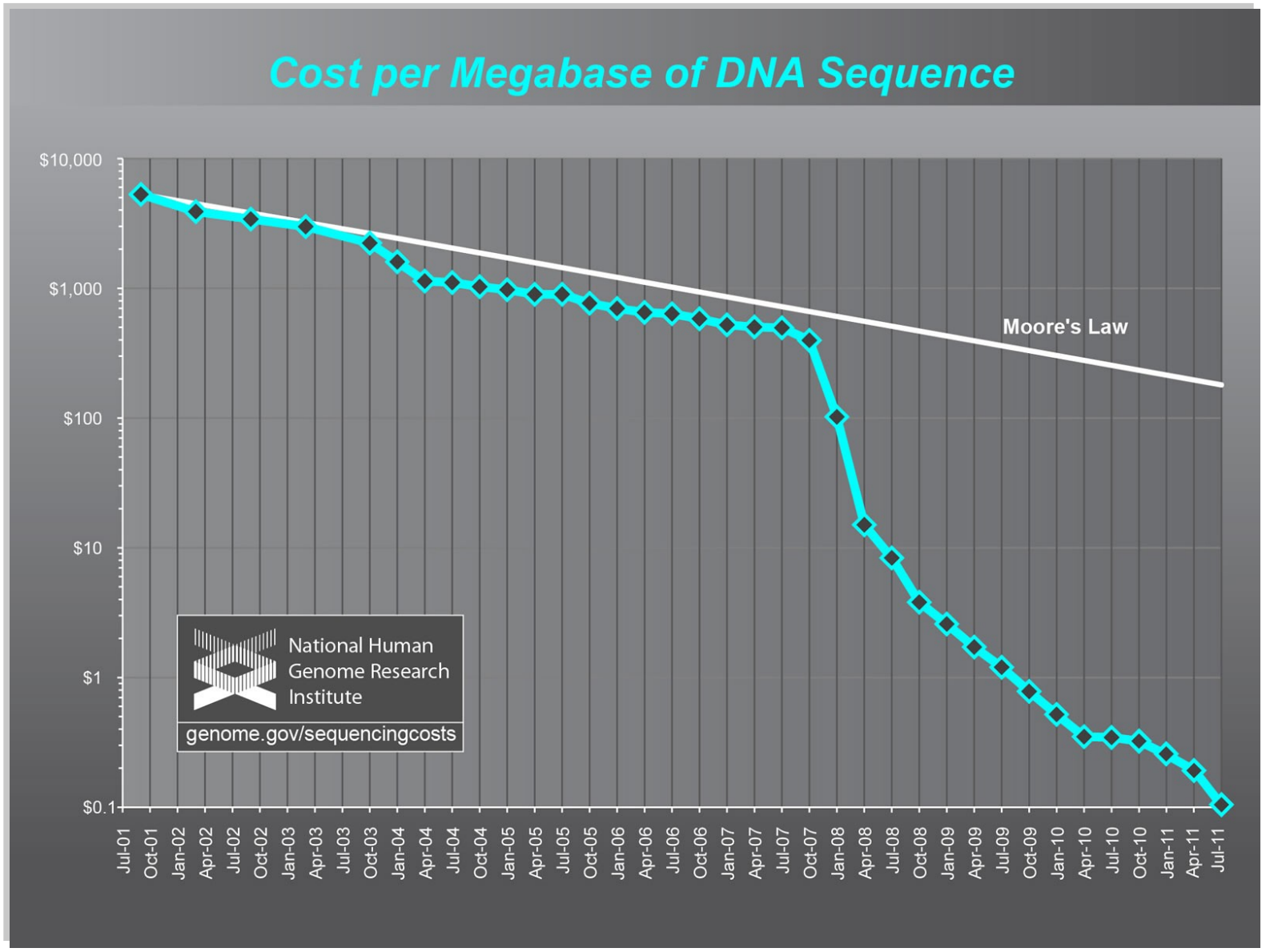
IntermezzoTalk
SURFnet7, Part of GigaPort3
Utrecht, Netherlands
December 7, 2011

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

UCSD
UCIrvine

it²

Cost Per Megabase in Sequencing DNA is Falling Much Faster Than Moore's Law



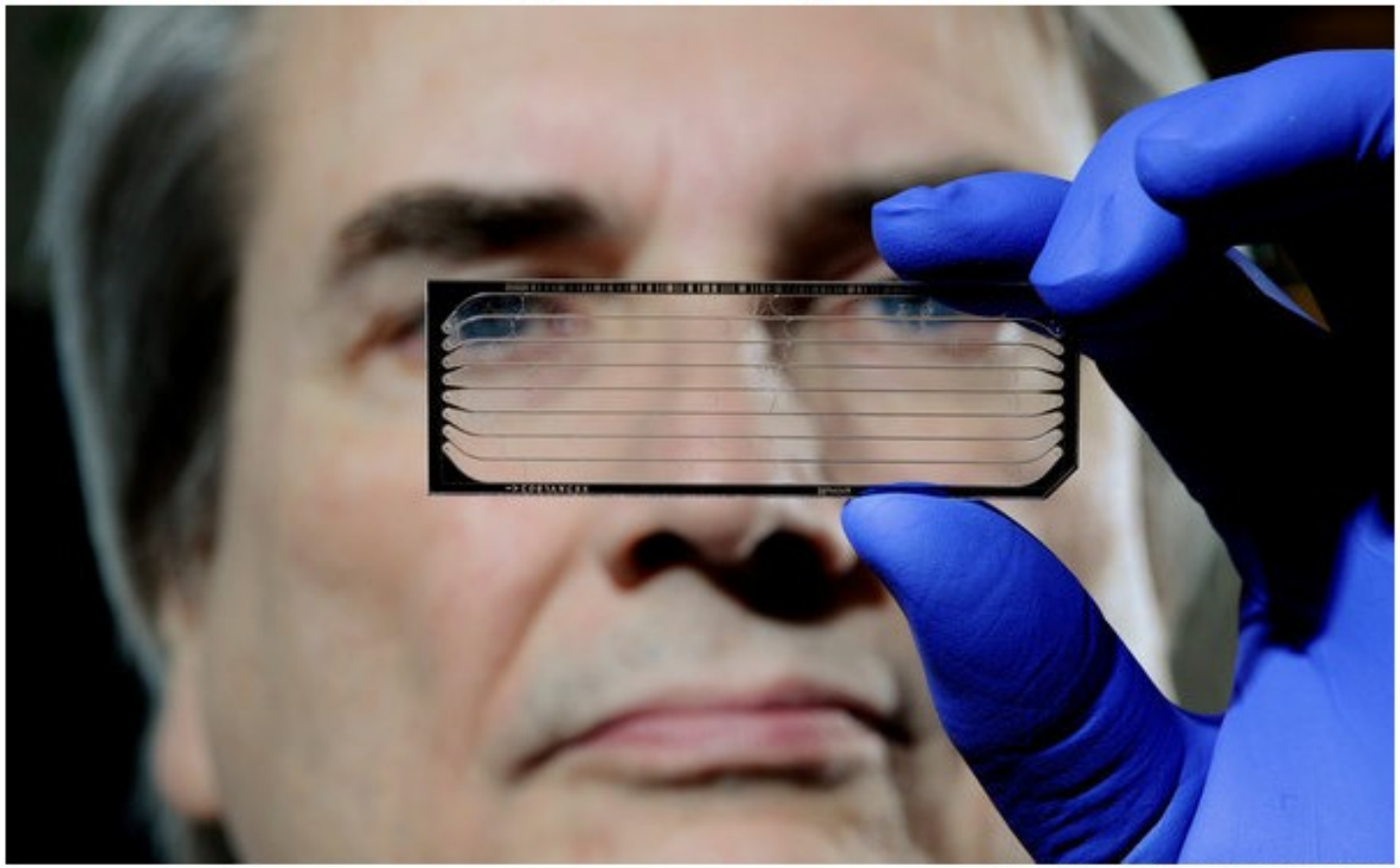
UCSD
UCIrvine

www.genome.gov/sequencingcosts/

it²

Genomic Sequencing is Driving Big Data

DNA Sequencing Caught in Deluge of Data



Kathy Kmonicek for The New York Times

The New York Times

Business Day

November 30, 2011

UCSD
UCIrvine

it²

BGI—The Beijing Genome Institute is the World's Largest Genomic Institute





- **Main Facilities in Shenzhen and Hong Kong, China**
  - **Branch Facilities in Copenhagen, Boston, UC Davis**
- **137 Illumina HiSeq 2000 Next Generation Sequencing Systems**
  - **Each Illumina Next Gen Sequencer Generates 25 Gigabases/Day**
- **Supported by Supercomputing ~160TF, 33TB Memory**
  - **Large-Scale (12PB) Storage**

UCSD
UCIrvine

it²

Next Generation Genome Sequencers Produce Large Data Sets

Data Volume	Total	Final	Comment
HiSeq 2000 200G run			
Image Data	32 TB	0	
Intensity Data	2 TB	0	Optionally transferred
Base Call / Quality Score Data	250 GB	250 GB	1 byte/base (raw) assuming qseq generation offline
Alignment Output	6 TB (3 TB)	1.2 TB	Remove intermediate files
GA_{IIx} 50G run			
Image Data	6.9	0	Optionally transferred
Intensity Data	0.93	0.93	
Base Call / Quality Score Data	0.17	0.17	
Alignment Output	1.2TB	1.2 TB	

UCSD
UCIrvine

Source: Chris Misleh, SOM/Calit2 UCSD

it²

Needed: Interdisciplinary Teams Made From Computer Science, Data Analytics, and Genomics

We believe
the field of bioinformatics
for genetic analysis
will be one of the biggest areas of
disruptive innovation
in life science tools
over the next few years,"
--Isaac Ro, an analyst at Goldman Sachs

UCSD
UCIrvine

Calit2 Brings Together Computer Science and Bioinformatics

Center for Algorithmic and Systems Biology

National Biomedical Computation Resource
an NIH supported resource center

NCMIR NATIONAL CENTER for MICROSCOPY and IMAGING RESEARCH at San Diego, an NIH supported resource center

camera Marine Microbial Ecology

JCSG Joint Center for Structural Genomics

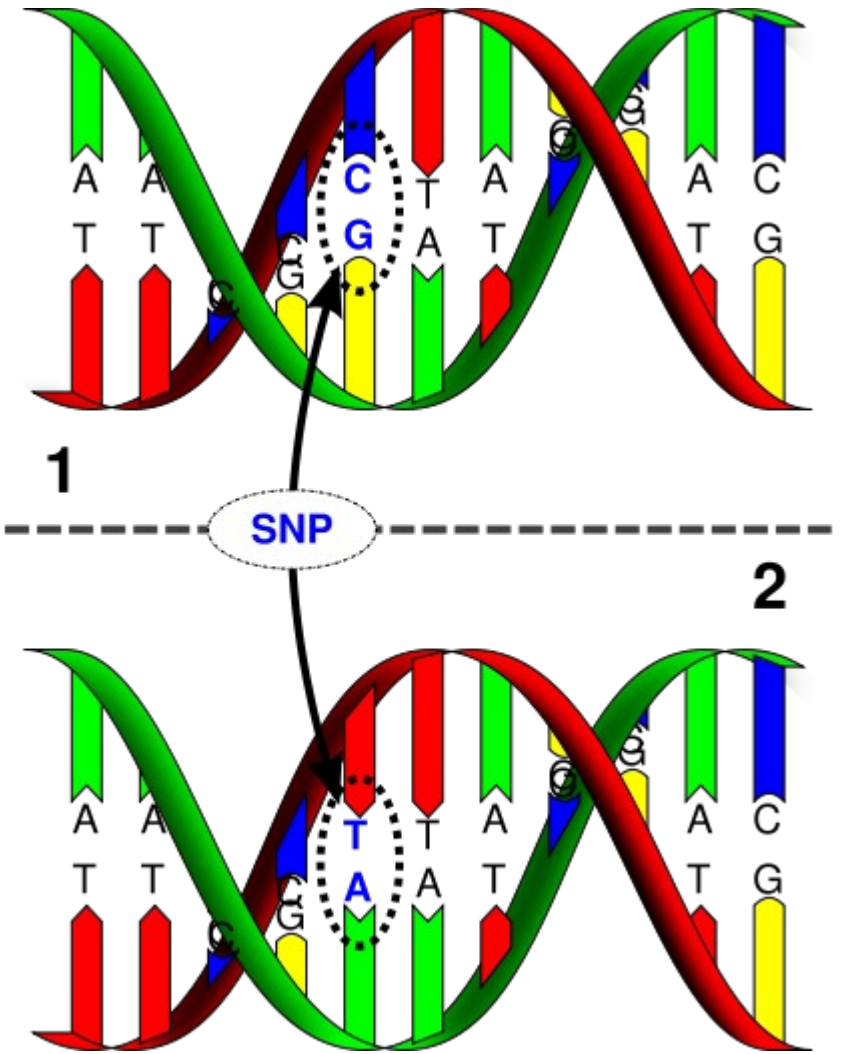
Institute for Genomics and Bioinformatics
University of California, Irvine

lifechips
university of california, irvine
www.lifechips.org

UCSD
UCIrvine

Single Nucleotide Polymophisms (SNPs): Human DNA Base Pairs May Differ At Some Points

Person A



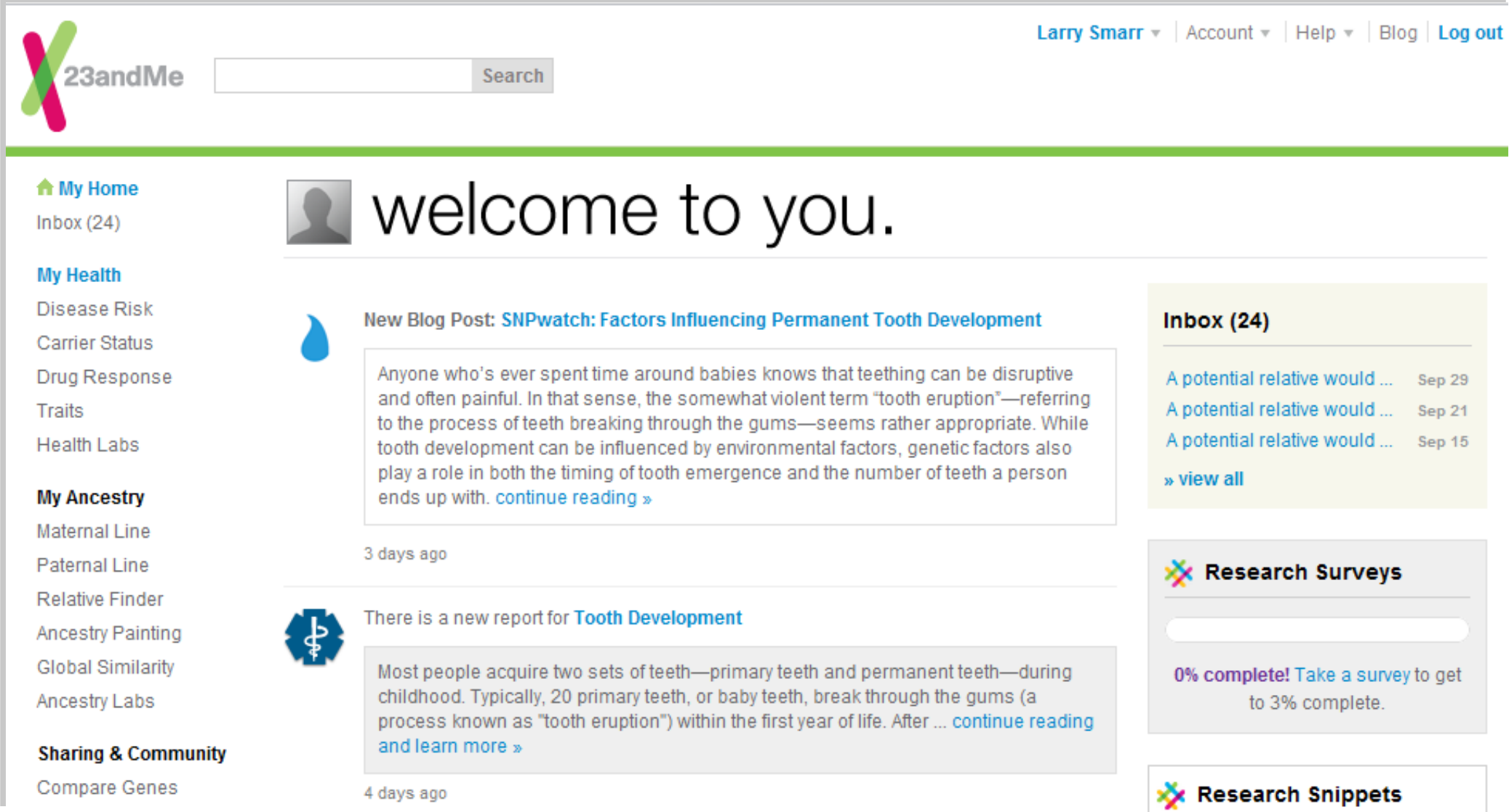
Person B

http://en.wikipedia.org/wiki/File:Dna-SNP.svg

Why We Study SNPs

99.9% of One's Individual DNA
Sequence will be Identical
to that of Another Person.
Of the 0.1% Difference,
Over 80% will be
Single Nucleotide Polymorphisms
(SNPs).

Consumer Companies Provide Your SNPs

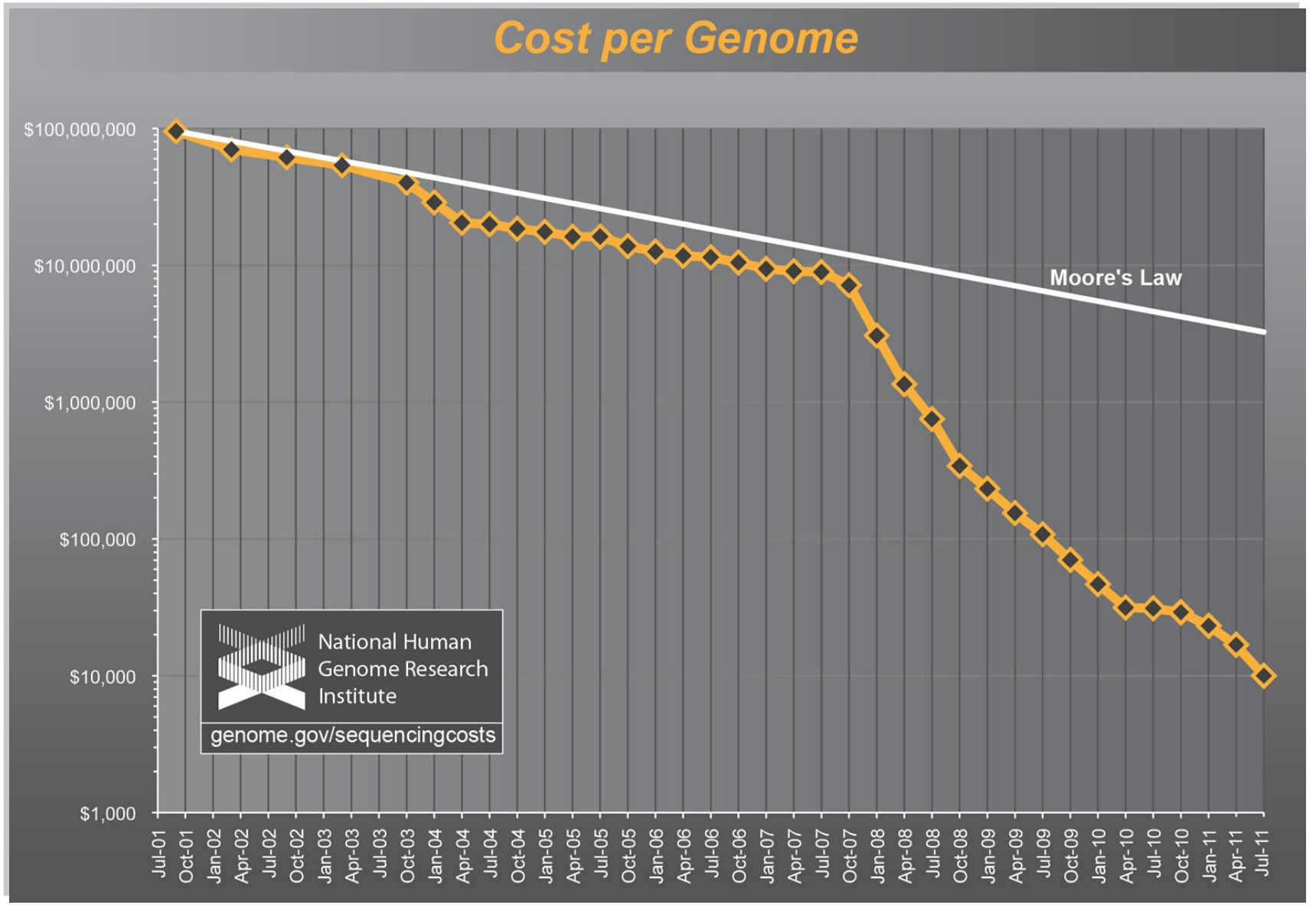


UCSD
UCIrvine

www.23andme.com

it²

Cost of Sequencing Human Genome is Rapidly Becoming Affordable



UCSD
UCIrvine

it²

The Rise of Individual and Societal Genomic Testing- Promise and Concerns

Technology Review
PUBLISHED BY MIT

How Personal Genomics Could Change Health Care

The select few who have had their genomes sequenced learn to use that information for health care.

By Emily Singer

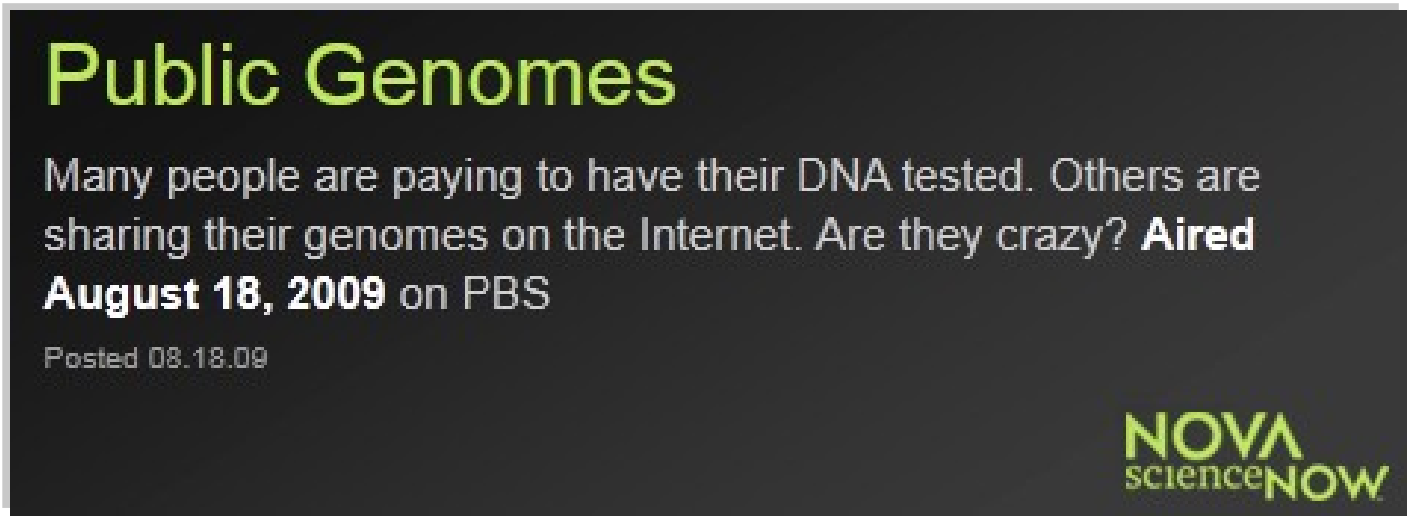
FRIDAY, APRIL 30, 2010

UCSD
UCIrvine

www.technologyreview.com/biomedicine/25218/

it²

Publically Sharing Your Genome and Medical Records: Is it Crazy or the Future?



Personal Genome Project

Posted 06.01.08 | NOVA scienceNOW

The Personal Genome Project, spearheaded by George Church of Harvard's Center for Computational Genetics, aims to recruit 100,000 people willing to offer up their DNA and personal life histories. It's all in an effort to further knowledge of human genetics and why we get—or don't get—diseases.

UCSD
UCIrvine

it²

From 10,000 Human Genomes Sequenced in 2011 to 1 Million by 2015 Out of Less Than 5,000 sq. ft.!

Proprietary Sequencing Technology

Designed and Optimized for Whole Human Genome Sequencing **Quality**, **Cost** and **Scale**



Outsourced Business Model

A Turnkey Service Enabling Customers to **Outsource** WGS – Samples In, **Research Ready Data** Out



Complete genomics

UCSD
UCIrvine

4 Million Newborns / Year in U.S.

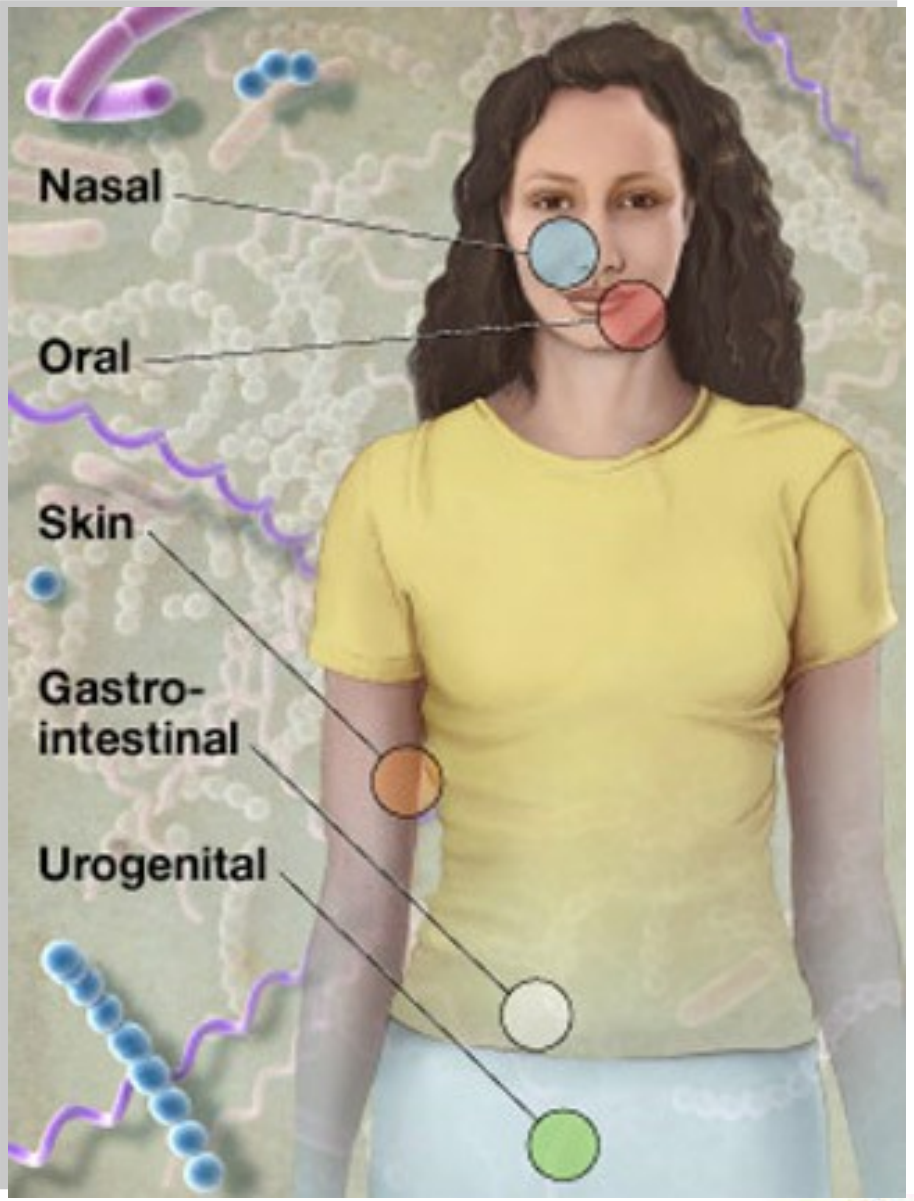
it²

But the Human Genome Contains Less Than 1% of the Bodies Genes

Human Microbiome Project

HMP

The Total Number of These Bacterial Cells
is 10 Times the Number
of Human Cells in Your Body



The NIH Common Fund
WE ACCELERATE DISCOVERY.

UCSD
UCIrvine

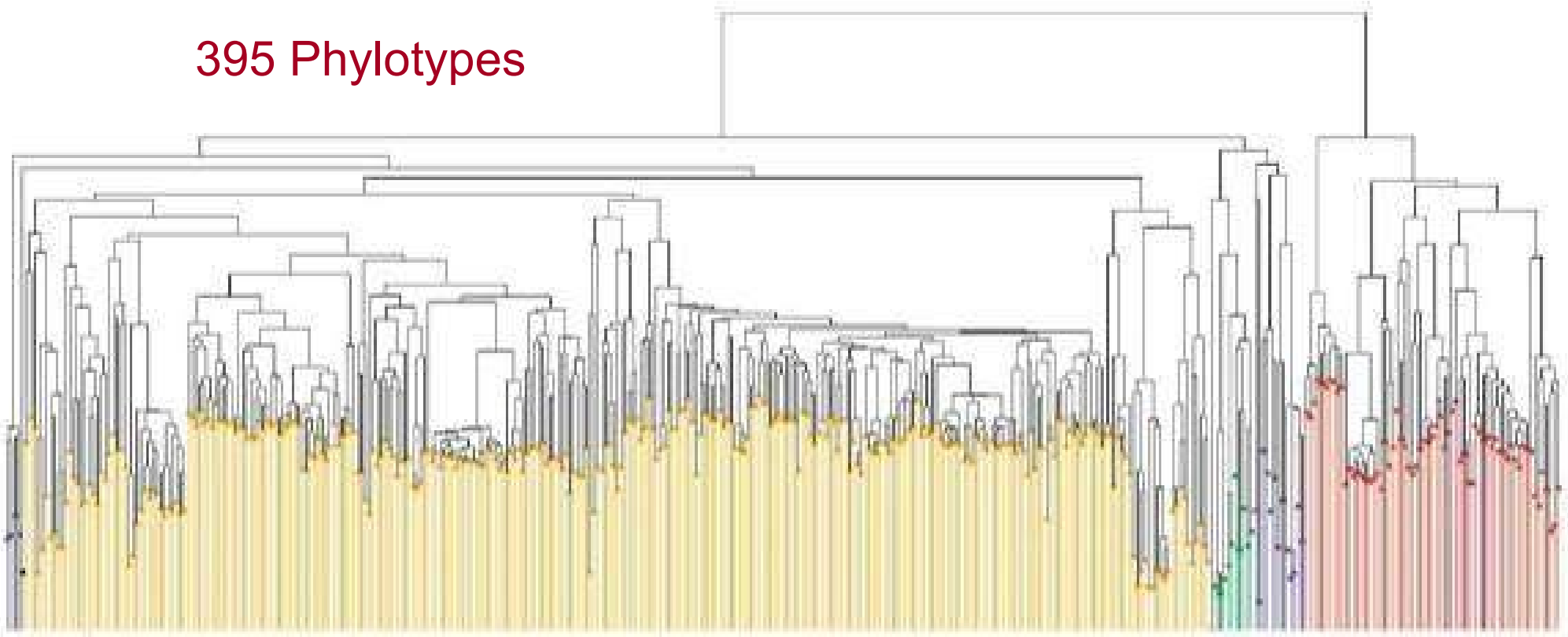
http://commonfund.nih.gov/hmp/

it²

The Human Microbiome is the Next Large NIH Drive to Understand Human Health and Disease

- **"A majority of the bacterial sequences corresponded to uncultivated species and novel microorganisms."**
- **"We discovered significant inter-subject variability."**
- **"Characterization of this immensely diverse ecosystem is the first step in elucidating its role in health and disease."**

395 Phylotypes



UCSD
UCIrvine

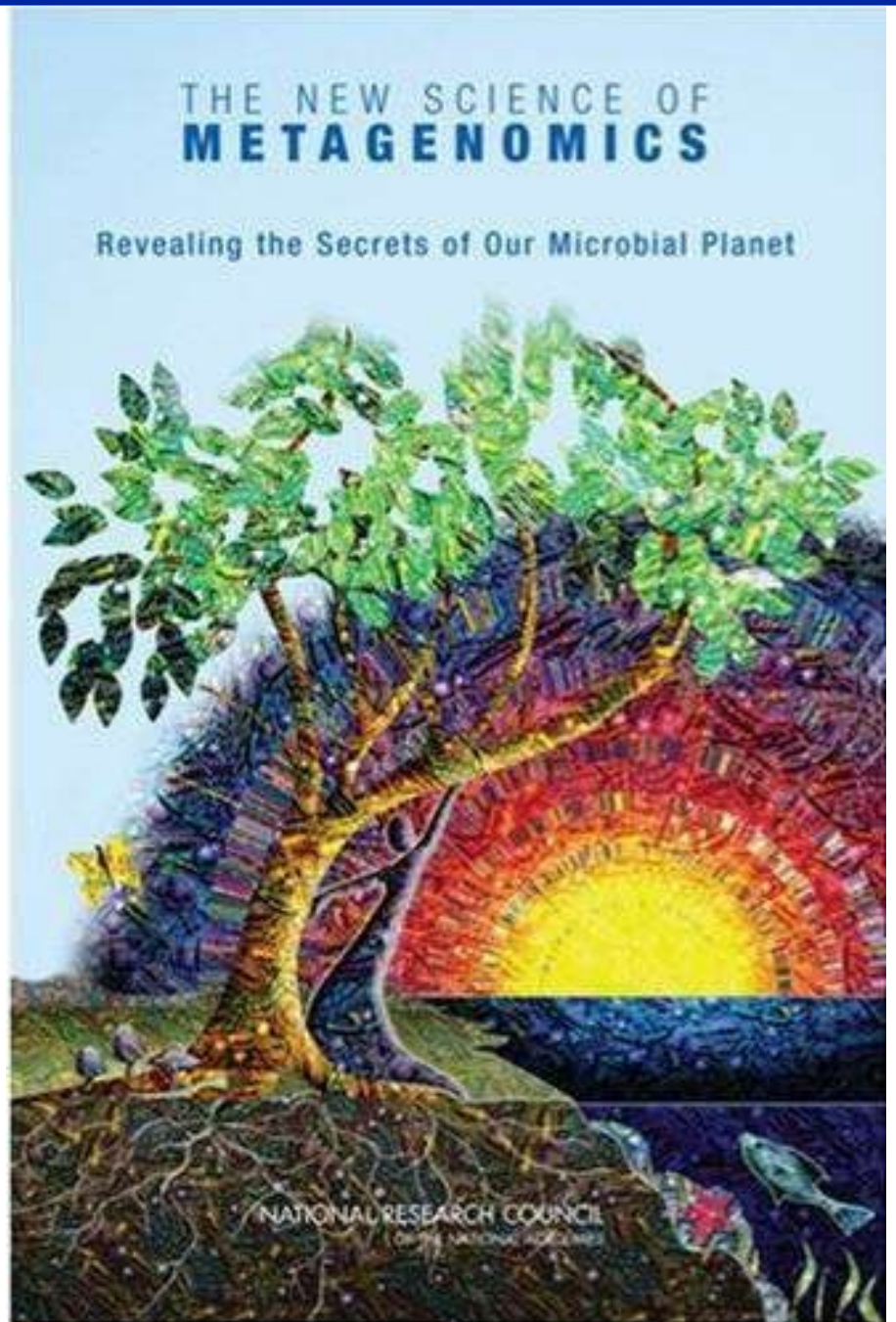
"Diversity of the Human Intestinal Microbial Flora"
Paul B. Eckburg, et al *Science* (10 June 2005)

it²

The New Science of Metagenomics

NRC Report:

Metagenomic data should be made publicly available in international archives as rapidly as possible.



"The emerging field of metagenomics, where the DNA of entire communities of microbes is studied simultaneously, presents the greatest opportunity -- perhaps since the invention of the microscope – to revolutionize understanding of the microbial world." –

National Research Council
March 27, 2007

UCSD
UCIrvine

it²

Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis



SDSC

Gordon and Betty MOORE FOUNDATION

it²

Calit2 CAMERA: 0ver 4000 Registered Users From Over 80 Countries

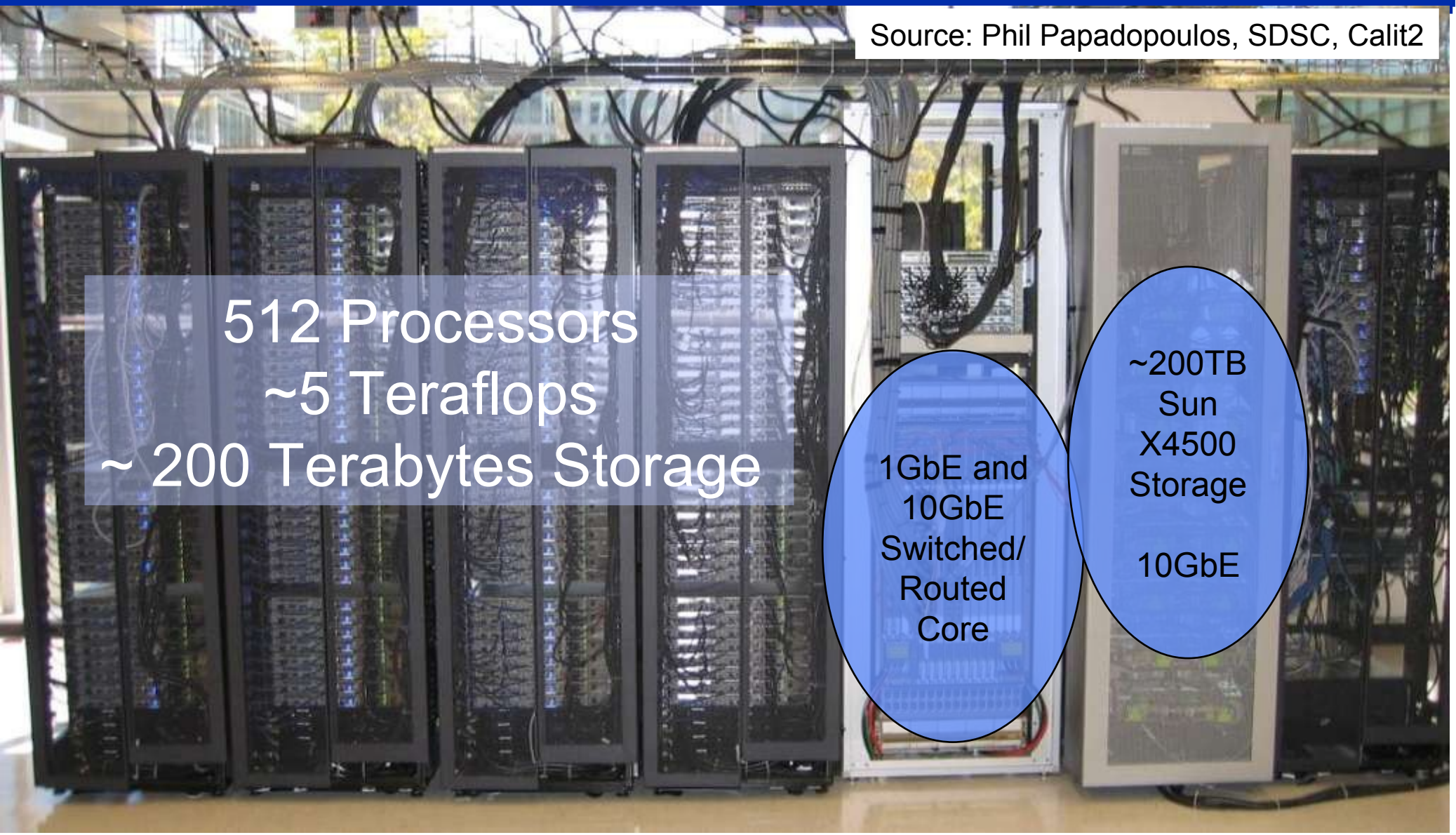


UCSD
UCIrvine

Gordon and Betty MOORE FOUNDATION

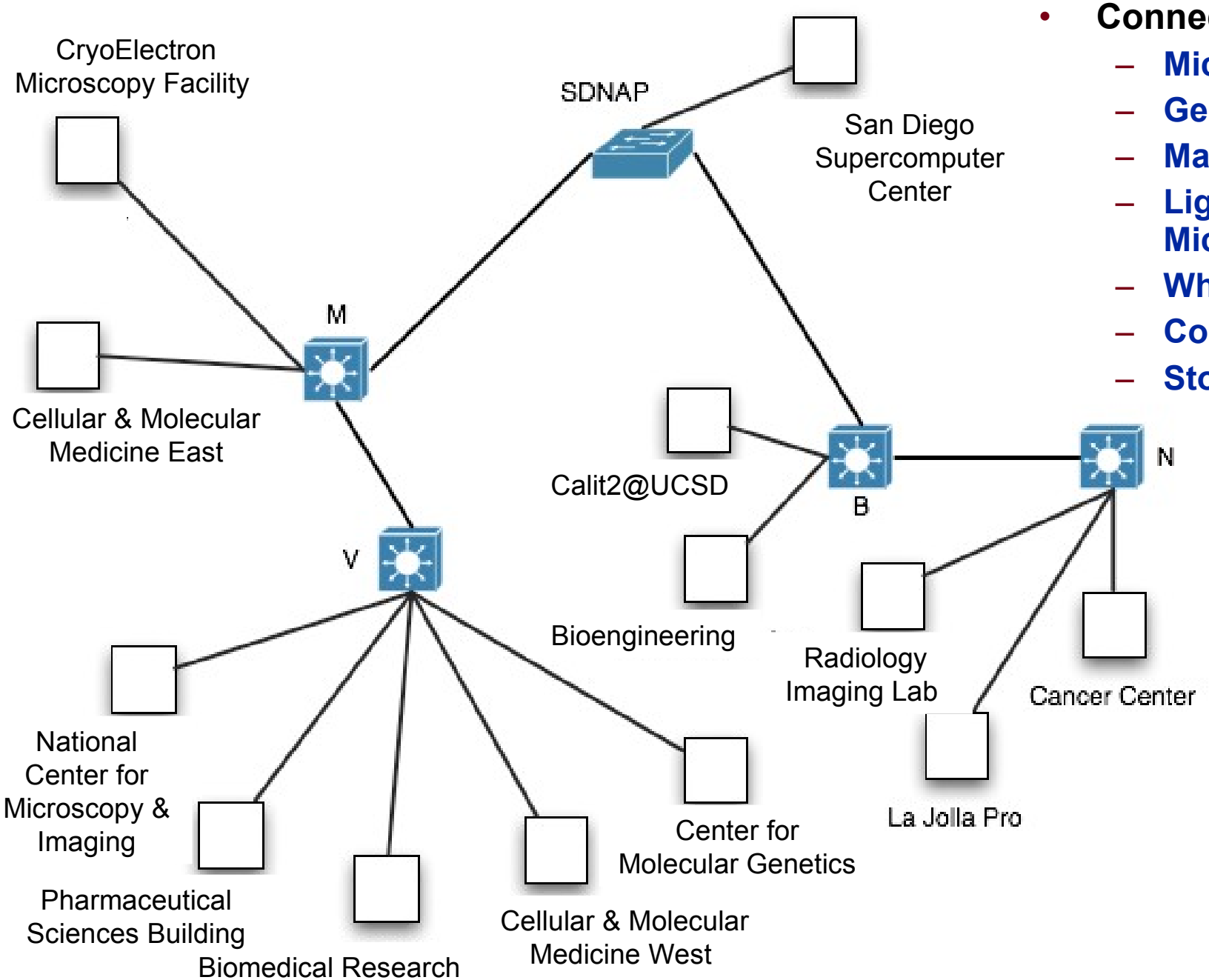
it²

Calit2 Microbial Metagenomics Cluster- Next Generation Optically Linked Science Data Server



4000 Users
From 90 Countries

UCSD Planned Optical Networked Biomedical Researchers and Instruments



- **Connects at 10 Gbps :**
  - **Microarrays**
  - **Genome Sequencers**
  - **Mass Spectrometry**
  - **Light and Electron Microscopes**
  - **Whole Body Imagers**
  - **Computing**
  - **Storage**

UCSD Campus Investment in Fiber Enables Big Data Science

SDSC

N x 10Gb/s

WAN 10Gb:
CENIC, NLR, I2
GLIF

Gordon – HPD System

Cluster Condo

Triton – Petascale Data Analysis

Scientific Instruments

DataOasis (Central) Storage

Extensible Disk Farm

Lustre | GridFTP | NFS | WebDAV | ...

Remote Access Interfaces

GreenLight Data Center

Digital Data Collections

Campus Lab Cluster

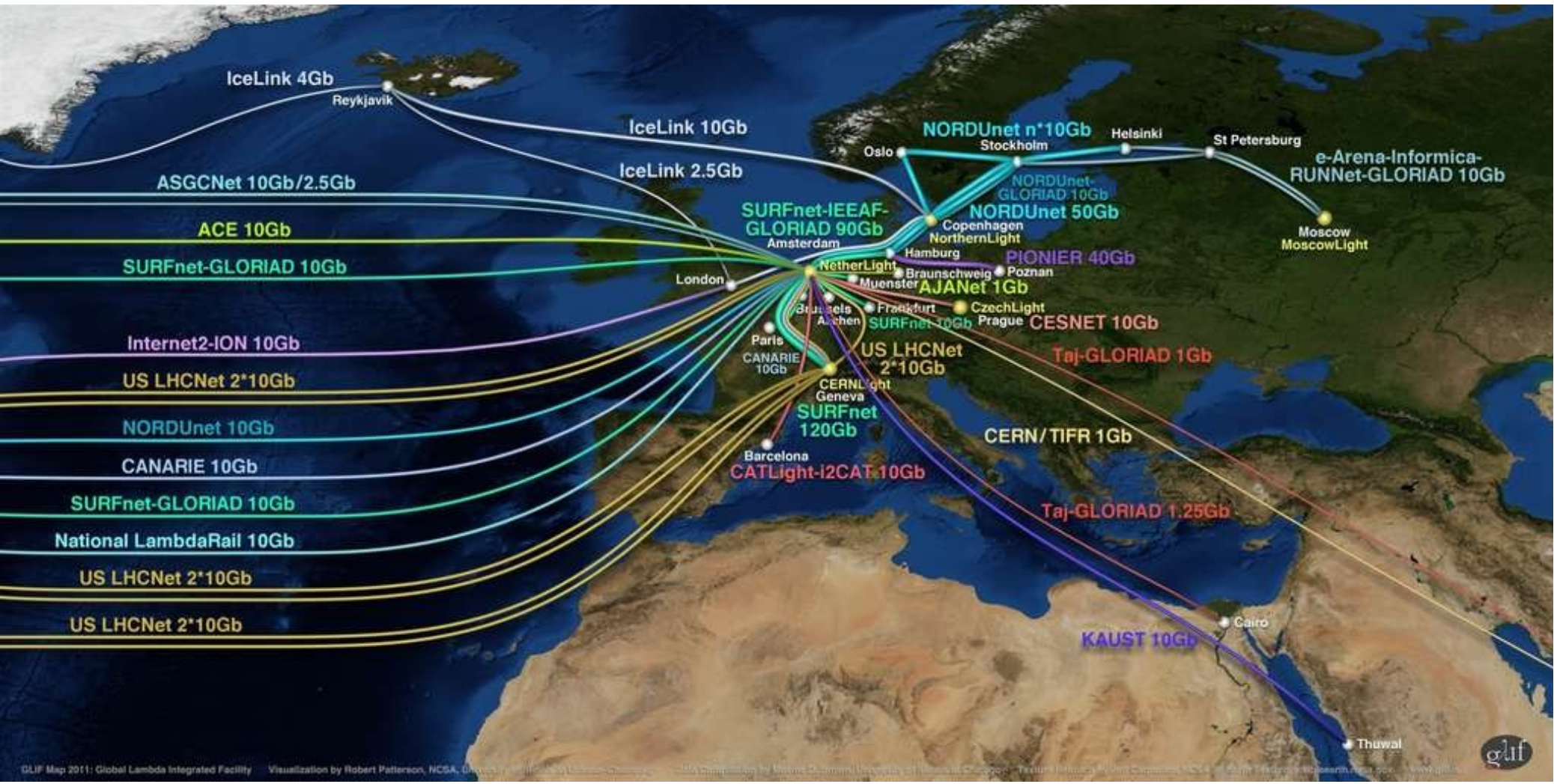
OptIPortal Tiled Display Wall

UCSD
UCIrvine

Source: Philip Papadopoulos, SDSC, UCSD

it²

SURFnet – a Global SuperNetwork Connecting to the Global Lambda Integrated Facility



Visualization courtesy of
Donna Cox, Bob Patterson, NCSA.