High Performance Computing for Biology and Medicine

Srinivas Aluru
School of Computational Science and Engineering
Georgia Institute of Technology

Why High Performance Computing

- To solve compute-intensive problems in reasonable time limits
  - e.g. weather prediction, protein folding
- To solve data-intensive problems that don't fit in a single node
  - e.g. genomics, social network analysis
- To simulate what can't be done experimentally
  - e.g. simulating the galaxy

The Big Data Challenge

Then (2005)

ABI 3700

96 ~800 bp reads

76.8 X 10^3 bases

~$1 per kilo base

Now

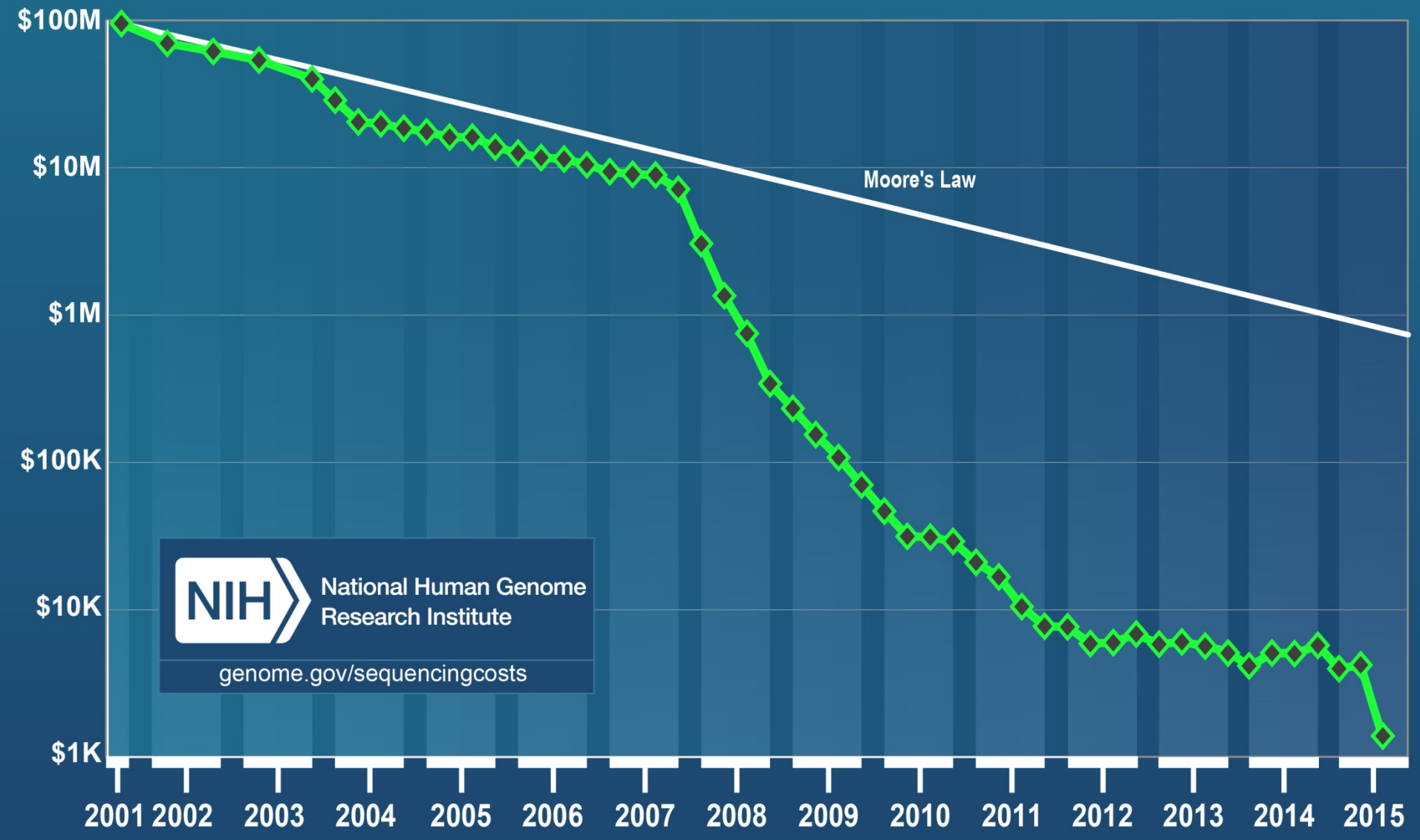
Illumina Hiseq 2500

6 billion 100 bp reads

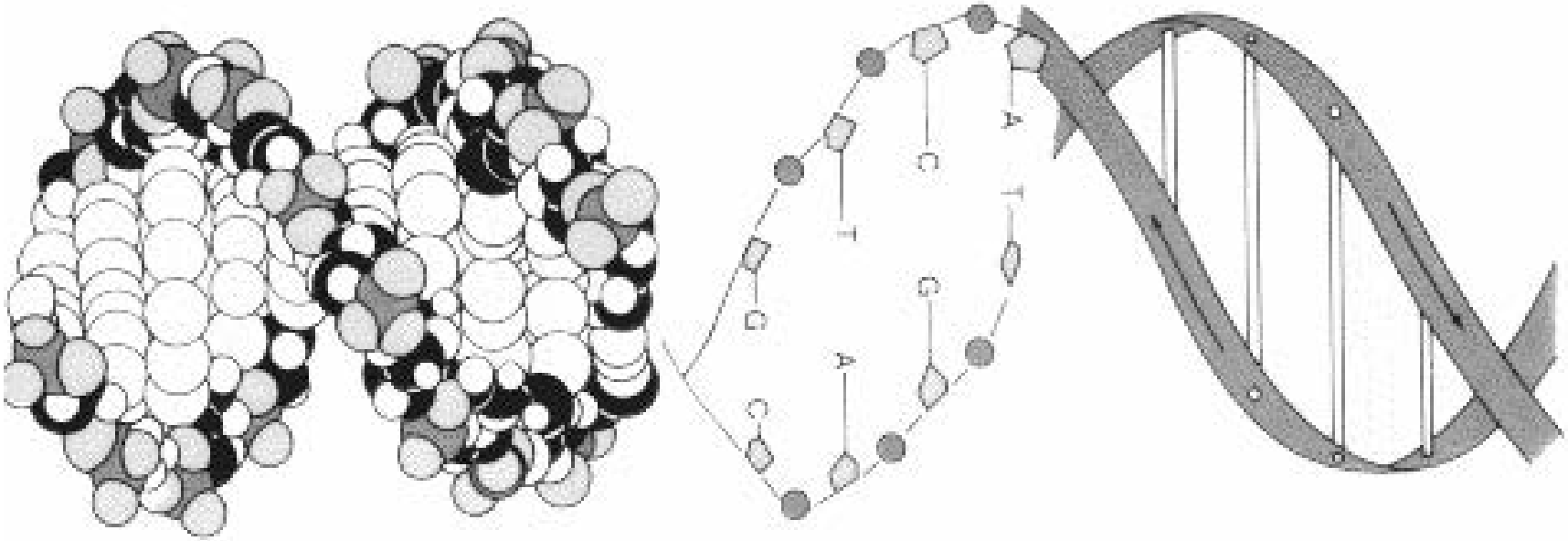
600 X 10^9 bases

~$1 per 200 million bases

Cost per Genome



What is a Genome?



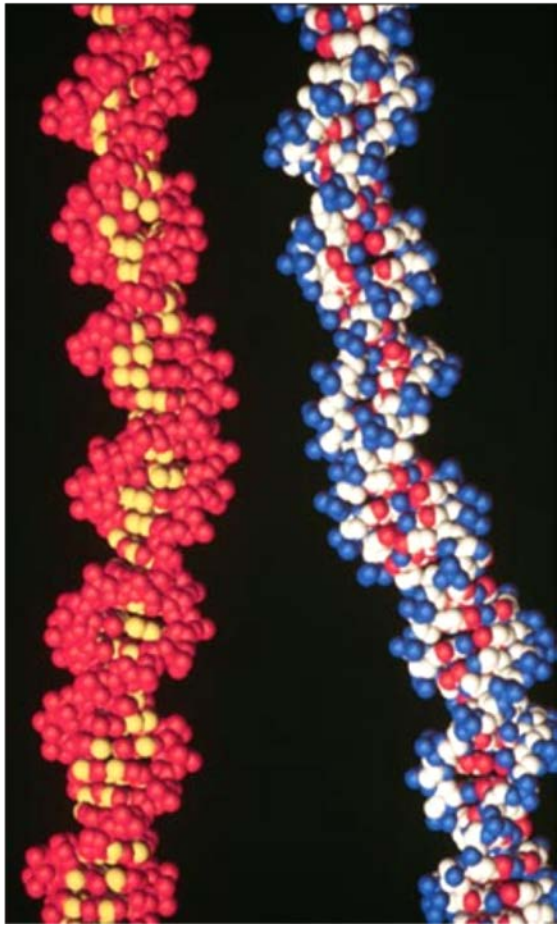
5′ A T T C G G G A A T G C A T G C C A 3′
3′ T A A G C C C T T A C G T A C G G T 5′

How Big are Genomes?

Viruses	50 kb+
Microbes	1 Mb+
Human/mouse/chimp	~3 Gb
Rice	450 Mb
Maize	2.5 Gb
Pine	19 Gb

DNA to be used for storage?

- *Regenesis: How synthetic biology will reinvent nature and ourselves*
- 53,000 words, 11 pictures, and some javascript
- 70 billion copies in a test tube
- 5.5 Pbits per mm^3, 10^6 times higher density then disk storage



Church *et al*, *Science 2013*

Human Genome Progress

- 1,000 genomes project
  - diverse anonymous individuals for study of human genetic variation
- Harvard's Personal Genome Project
  - 100,000 volunteers
  - no anonymity, limited to U.S.

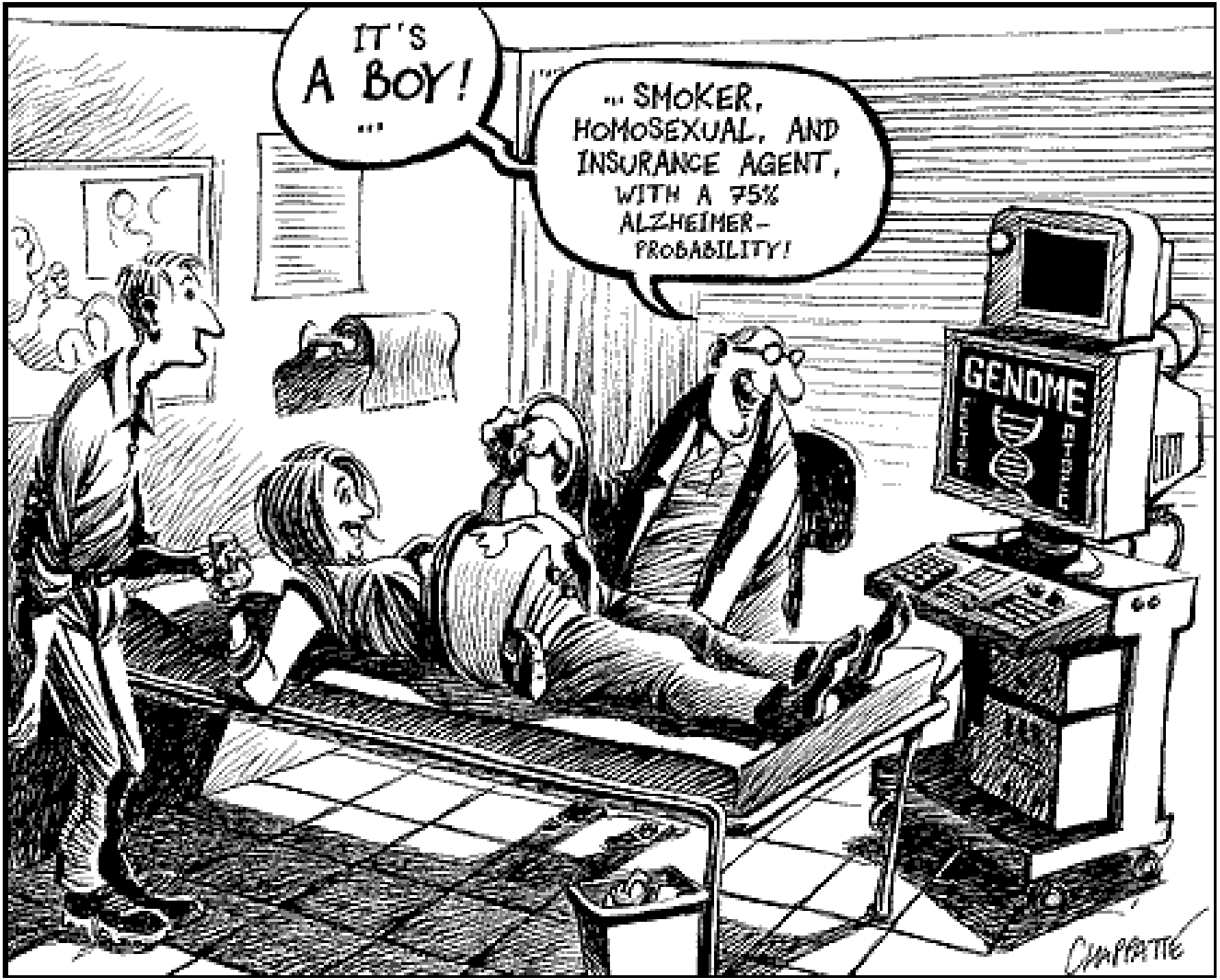
- What are we looking for?
  - SNPs, indels, CNVs, disease associations
  - Race, Ethnic, Gender specific analysis
  - Genome evolution during diseases such as cancer

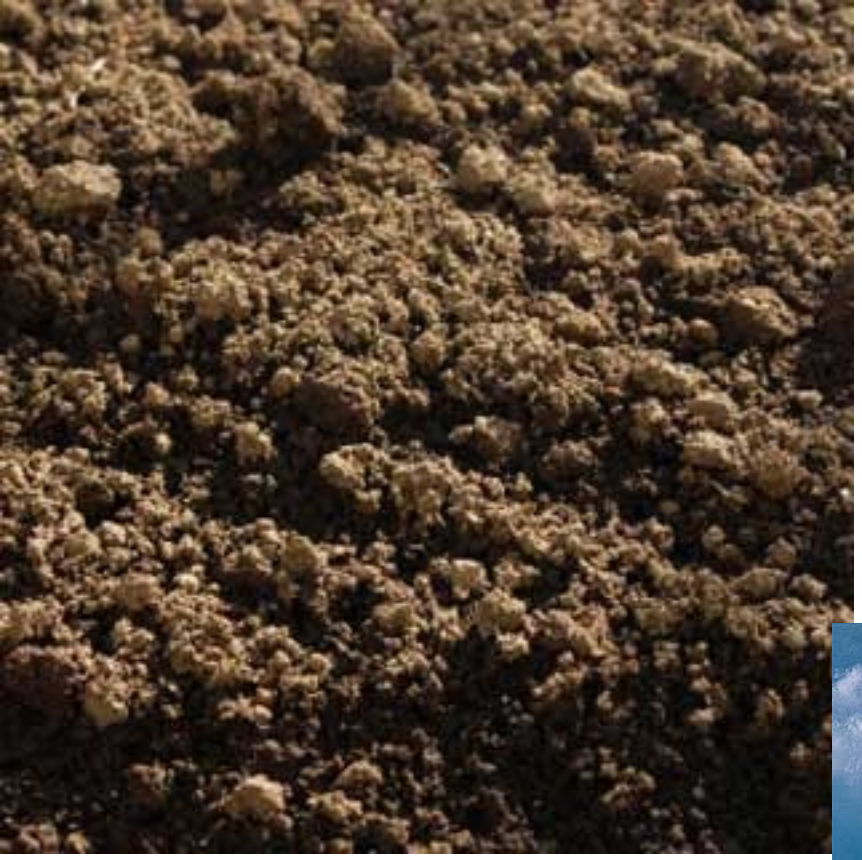
SNPs

Genome Resequencing

50X of 3GB human genome ≈ 1.5 billion 100bp reads

Reference genome


IT'S A BOY!
...
... SMOKER, HOMOSEXUAL, AND INSURANCE AGENT, WITH A 75% ALZHEIMER-PROBABILITY!
GENOME
CHAPPATTE

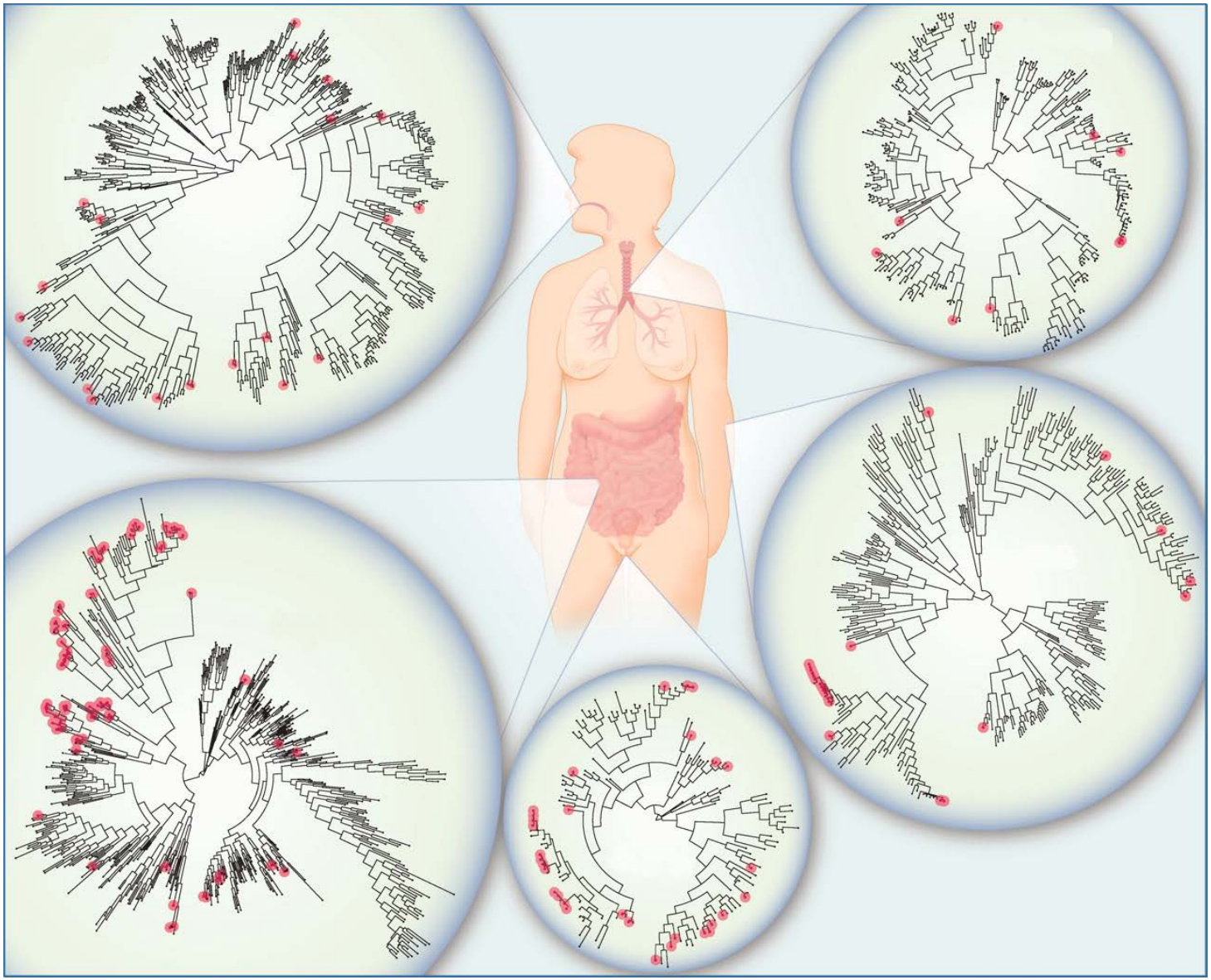


~40 million microbes/gram

~1 million/ml

$5X10^{30}$ estimated on the planet

Human Microbiome

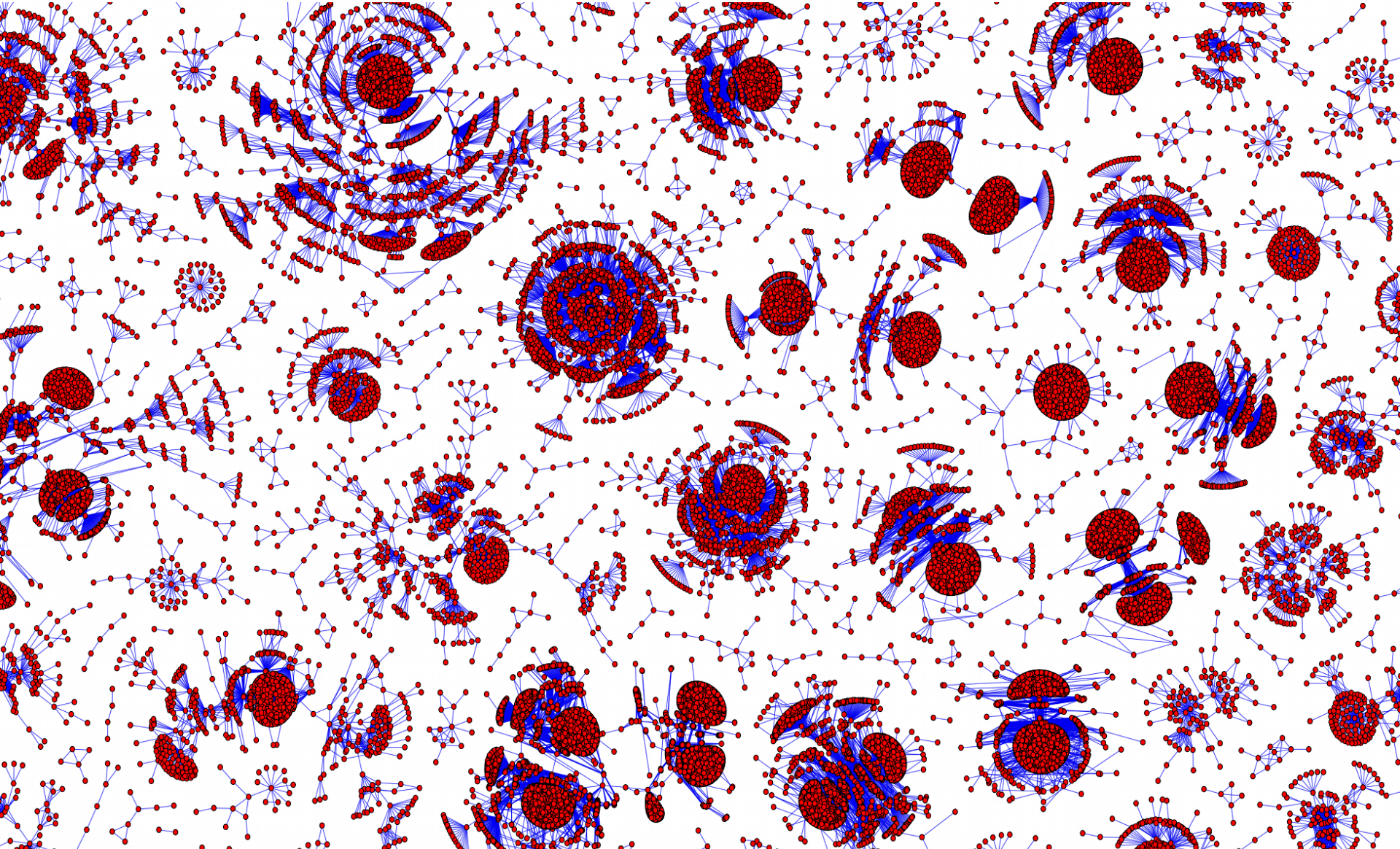


You are host to …

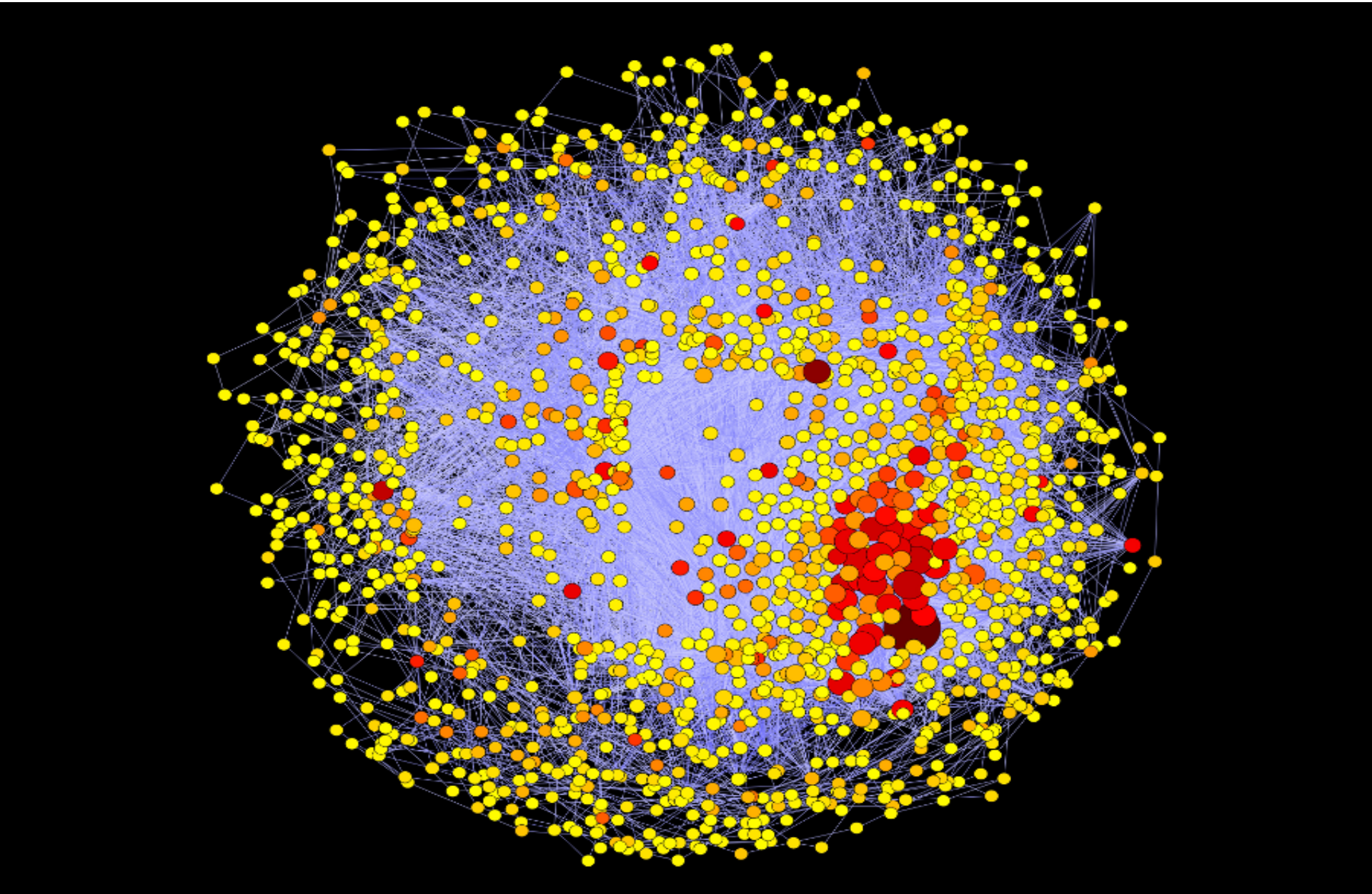
- 10,000 microbial species
- Outnumbering your cells 10 to 1
- About 5 lbs. of your body weight
- Over 8 million genes to your ~30,000

Source: Human microbiome project

Mouse Gut Microbiome



Gene Interaction Networks

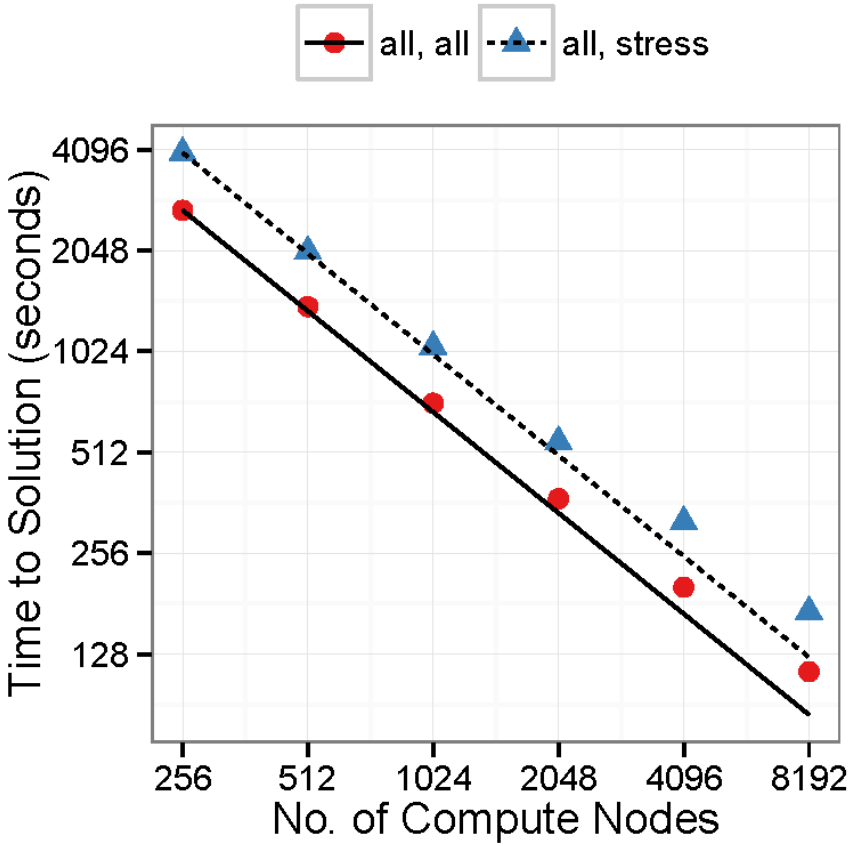


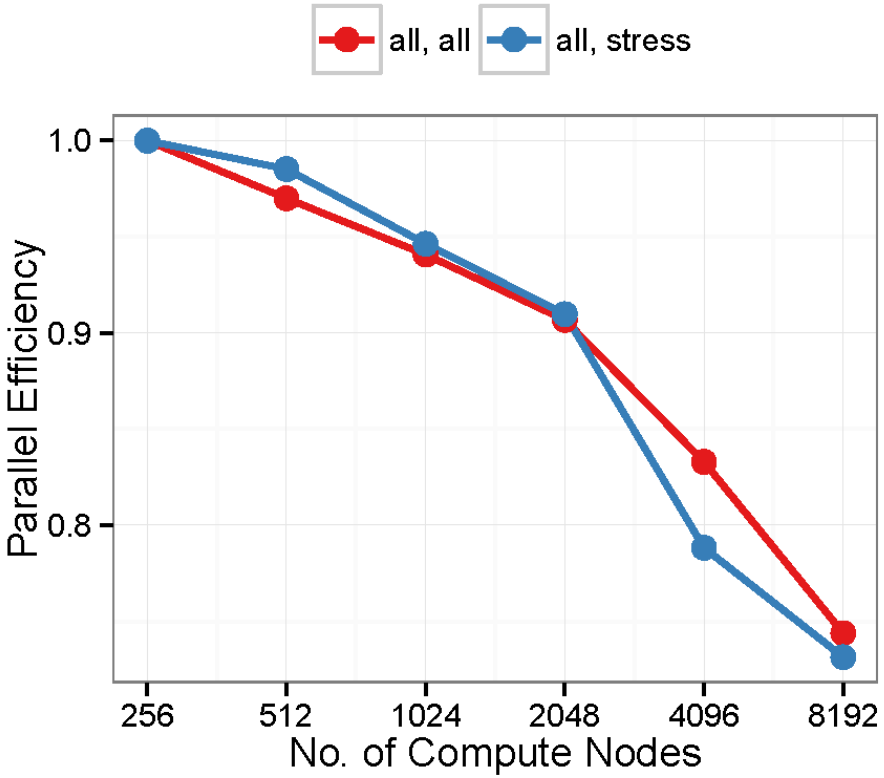
Target Supercomputers

- Tianhe-2, National University of Defense Technology, Changsha.
- Stampede, Texas Advanced Computing Center, Austin

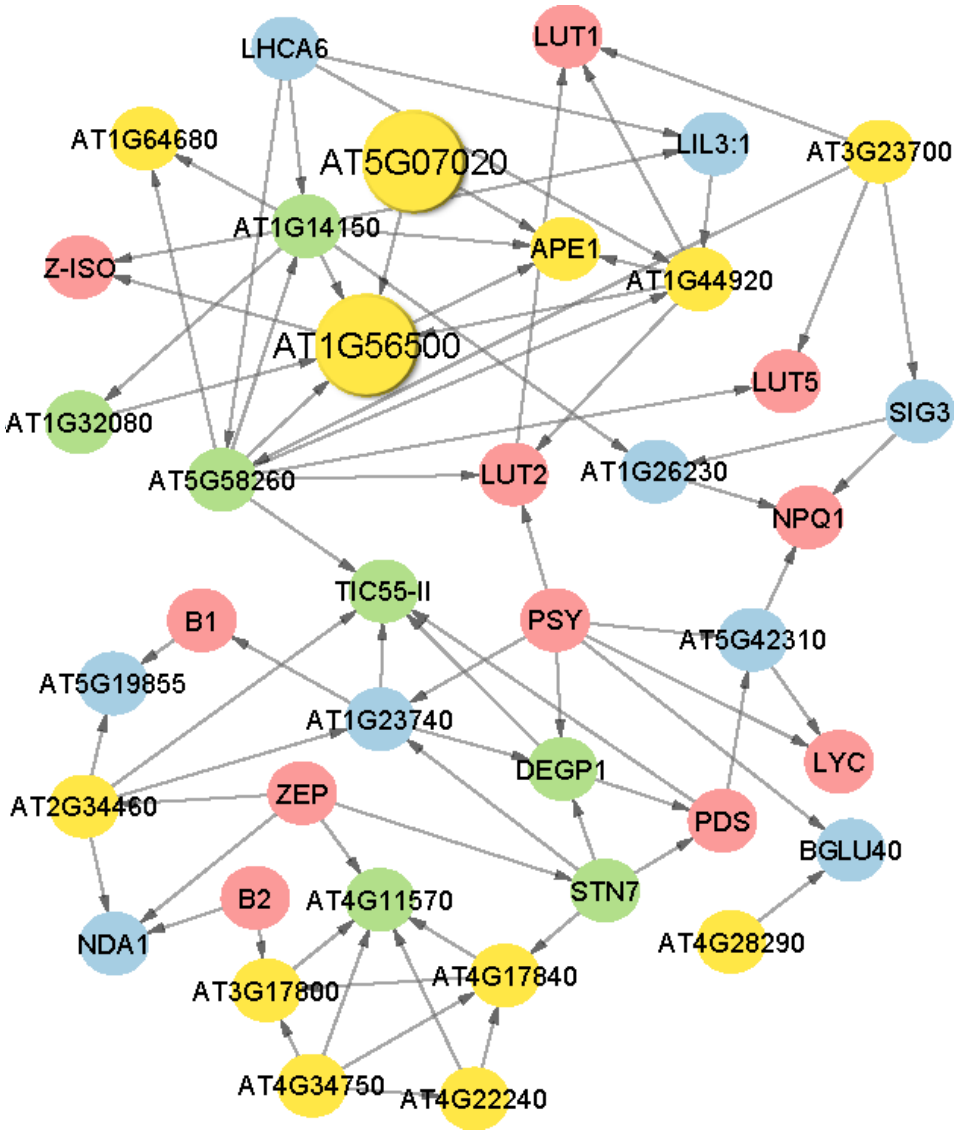
	Tianhe-2(54.9 PF)	Stampede(8.5 PF)
CPU	Intel Xeon E5-2600	Intel Xeon E5-2680
CPU Frequency	2.2 GHz	2.7 GHz
No. of CPUs	2	2
DRAM	64 GB	32 GB
Coprocessors	Intel Xeon Phi 31 S1P	Intel Xeon Phi SE10P
Coprocessors frequency	1.09 GHz	1.09 GHz
No. of Coprocessors	3	1
Coprocesssor Memory	8 GB	8 GB
Cores per node	192 (2 * 12 + 3*56)	76 * (2*8 + 60)
Threads per node	696	256

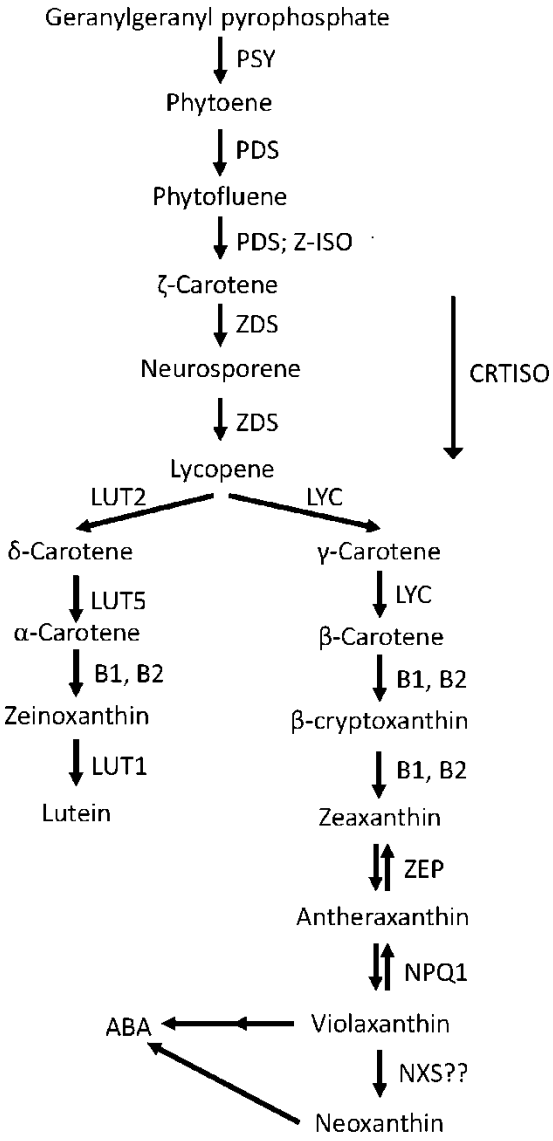
Parallel Efficiency





Carotenoid Subnetwork and Pathway





Arabidopsis T-DNA Mutants

Wild-Type **AT1G56500** **AT5G07020**

Faculty: Interdisciplinary Innovators

Georgia Tech School of Computational Science and Engineering

Srinivas Aluru
Professor

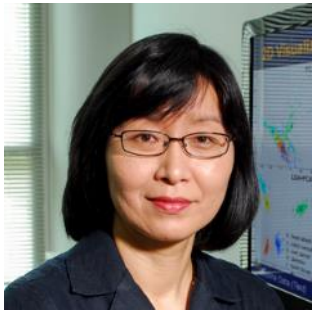
David Bader
Professor and Chair

Polo Chau
Assistant Professor

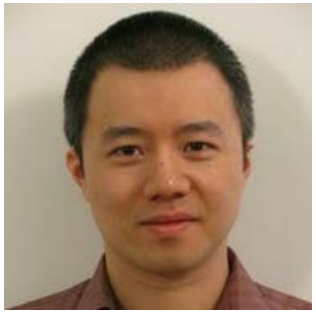
Edmond Chow
Associate Professor

Bistra Dilkina
Assistant Professor

Richard Fujimoto
Regents' Professor

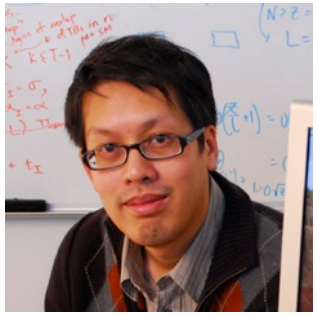


Haesun Park
Professor

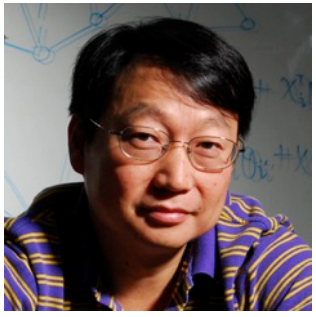


Le Song
Assistant Professor

Jimeng Sun
Associate Professor



Richard Vuduc
Associate Professor



Hongyuan Zha
Professor

Kenneth Brown
Chemistry

Mark Borodovsky
BME

David Sherrill
Chemistry

Surya Kalidindi
Mech. Engr.