Big Data Meets DNA

How Biological Data Science is improving our health, foods, and energy needs

Michael Schatz

April 8, 2014

IEEE Fellows Night Syracuse

@mike_schatz

The secret of life

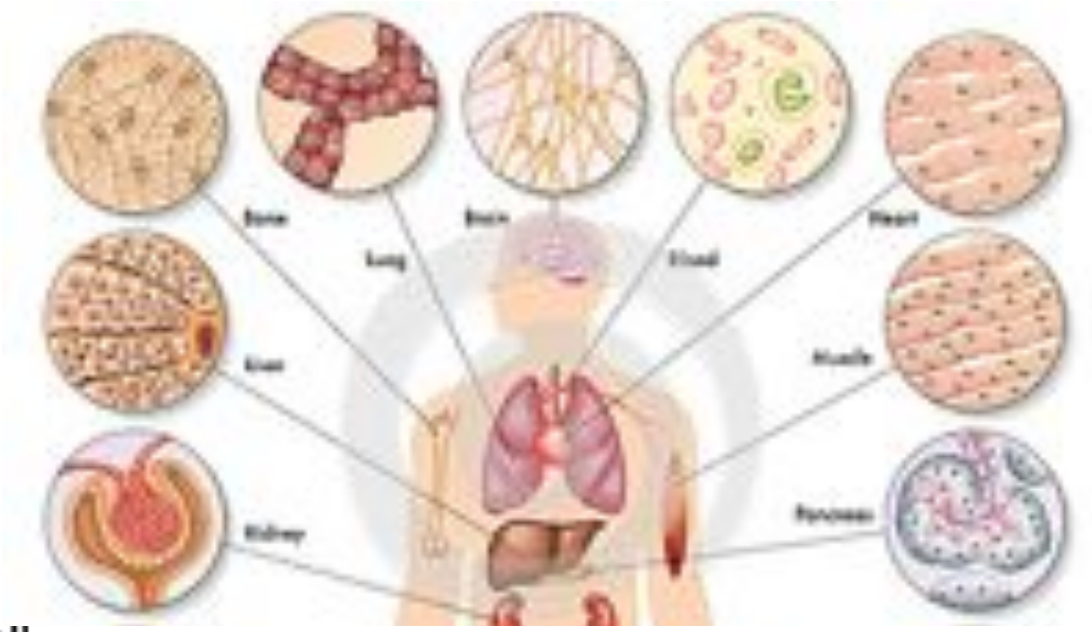
Your DNA, along with your environment and experiences, shapes who you are

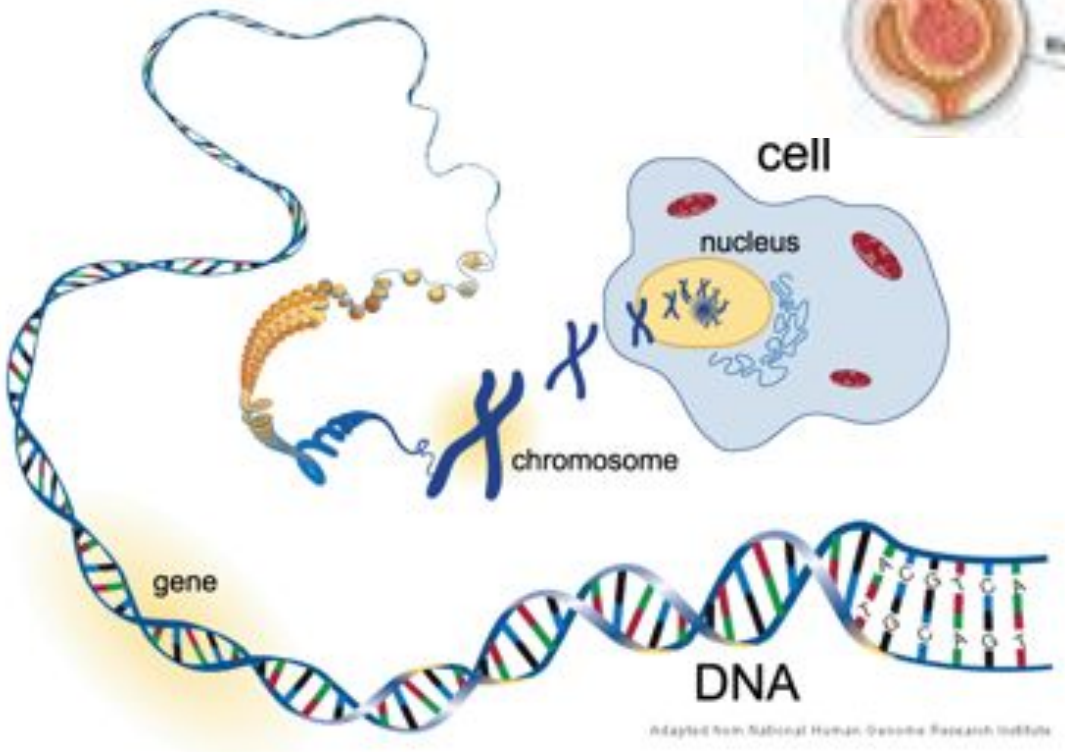
- Height
- Hair, eye, skin color
- Broad/narrow, small/large features
- Susceptibility to disease
- Response to drug treatments
- Longevity and Intelligence

Physical traits tend to be strongly genetic, social characteristics tend to be strongly environmental, and everything else is a combination

Cells & DNA

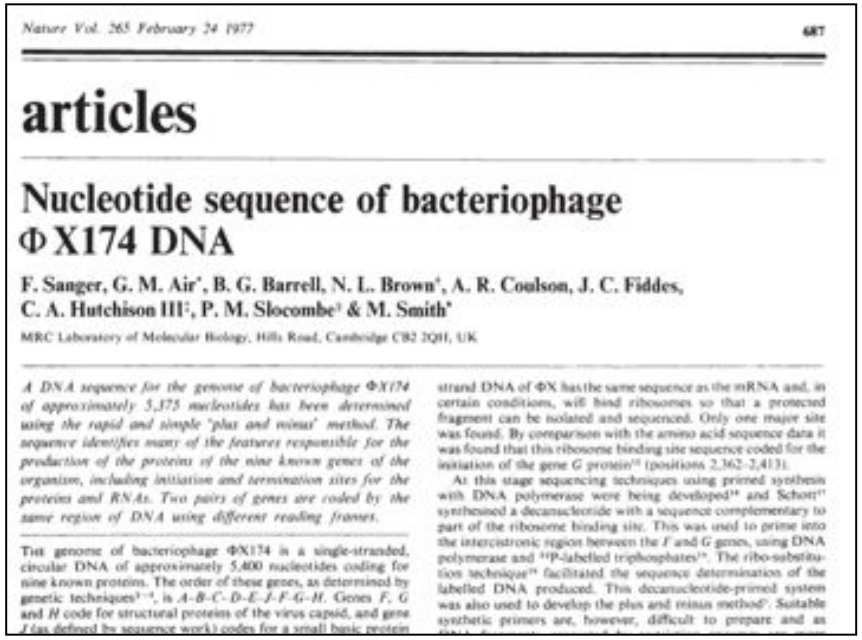
Each cell of your body contains an exact copy of your 3 billion base pair genome.





Your specific nucleotide sequence encodes the genetic program for your cells and ultimately your traits

The Origins of DNA Sequencing



Nature Vol. 265 February 24 1977 687

articles

Nucleotide sequence of bacteriophage ΦX174 DNA

F. Sanger, G. M. Air*, B. G. Barrell, N. L. Brown†, A. R. Coulson, J. C. Fiddes, C. A. Hutchison III‡, P. M. Slocombe‡ & M. Smith*

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

A DNA sequence for the genome of bacteriophage ΦX174 of approximately 5,375 nucleotides has been determined using the rapid and simple 'plus and minus' method. The sequence identifies many of the features responsible for the production of the proteins of the nine known genes of the organism, including initiation and termination sites for the proteins and RNAs. Two pairs of genes are coded by the same region of DNA using different reading frames.

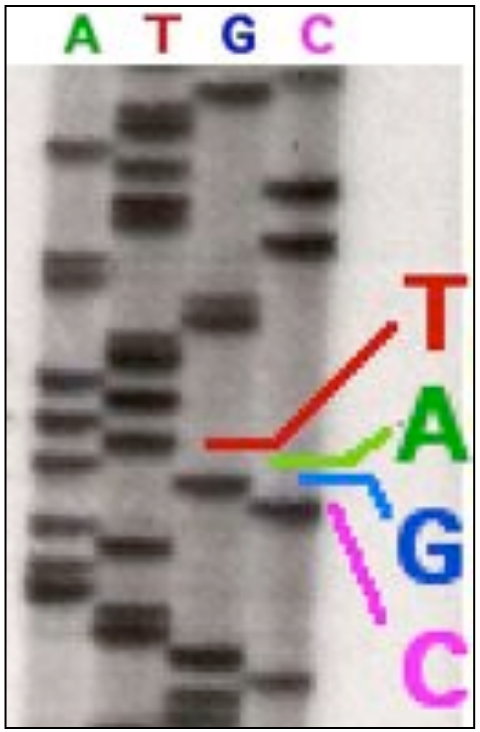
THE genome of bacteriophage ΦX174 is a single-stranded, circular DNA of approximately 5,400 nucleotides coding for nine known proteins. The order of these genes, as determined by genetic techniques[1-4], is *A–B–C–D–E–J–F–G–H*. Genes *F*, *G* and *H* code for structural proteins of the virus capsid, and gene *J* (as defined by sequence work) codes for a small basic protein

strand DNA of ΦX has the same sequence as the mRNA and, in certain conditions, will bind ribosomes so that a protected fragment can be isolated and sequenced. Only one major site was found. By comparison with the amino acid sequence data it was found that this ribosome binding site sequence coded for the initiation of the gene *G* protein[5] (positions 2,362–2,413).

At this stage sequencing techniques using primed synthesis with DNA polymerase were being developed[6,8] and Schott[11] synthesised a decanucleotide with a sequence complementary to part of the ribosome binding site. This was used to prime into the intercistronic region between the *F* and *G* genes, using DNA polymerase and ^{32}P-labelled triphosphates[7,8]. The ribo-substitution technique[10] facilitated the sequence determination of the labelled DNA produced. This decanucleotide-primed system was also used to develop the plus and minus method[9]. Suitable synthetic primers are, however, difficult to prepare and as

Sanger *et al. (1977) Nature*
1st Complete Organism
Bacteriophage ϕX174; 5375 bp

Awarded Nobel Prize in 1980



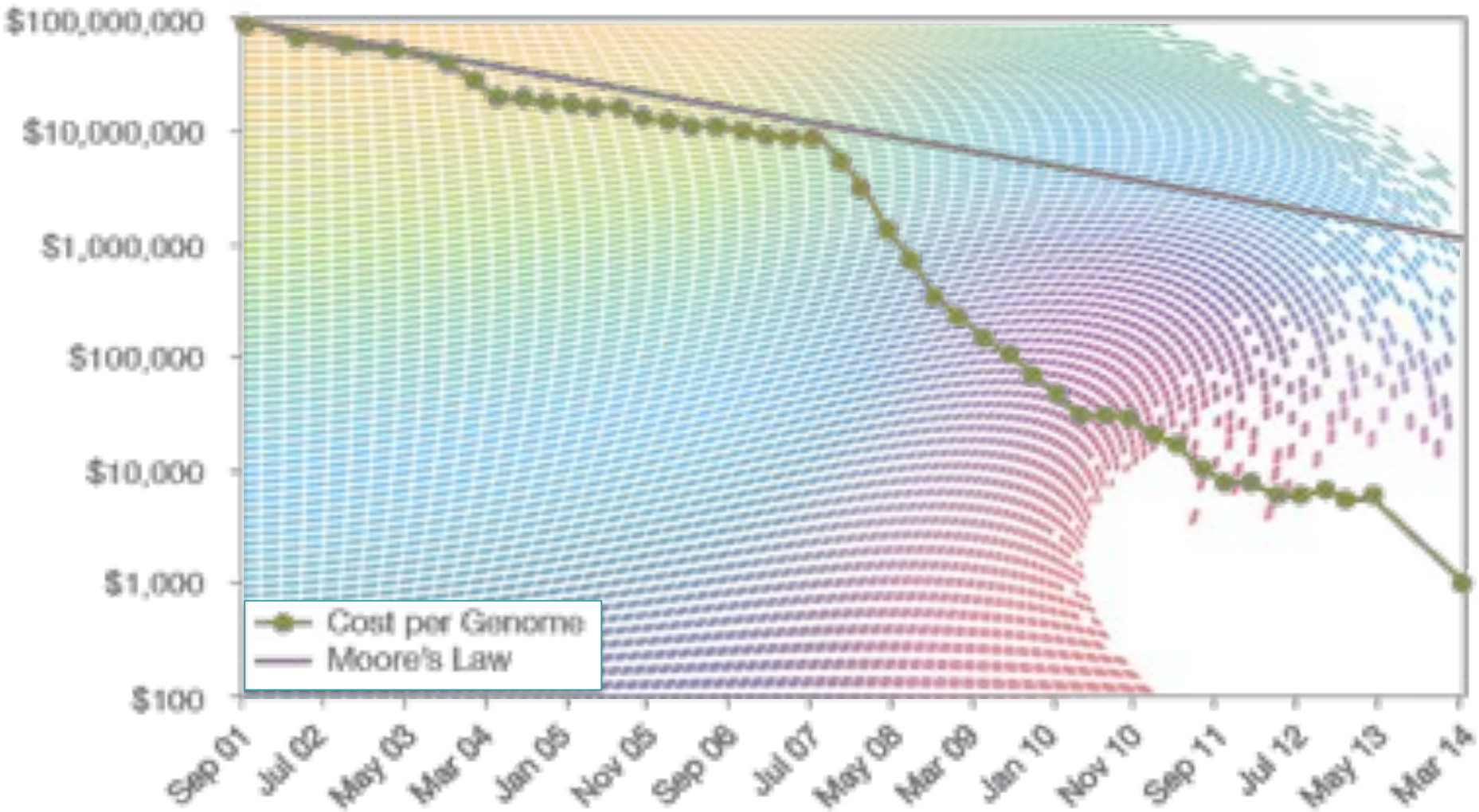
Radioactive Chain Termination
5000bp / week / person

http://en.wikipedia.org/wiki/File:Sequencing.jpg
http://www.answers.com/topic/automated-sequencer

Milestones in DNA Sequencing

(TIGR/Celera, 1995-2001)

Cost per Genome

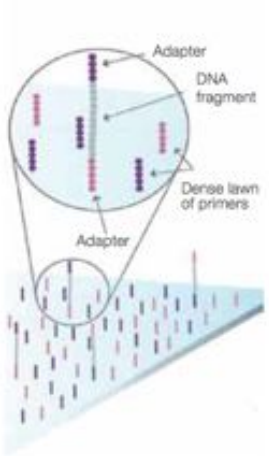


http://www.genome.gov/sequencingcosts/

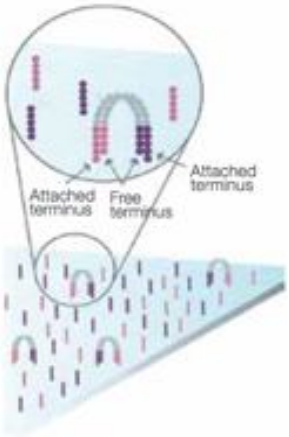
Massively Parallel Sequencing

Illumina HiSeq 2000
Sequencing by Synthesis

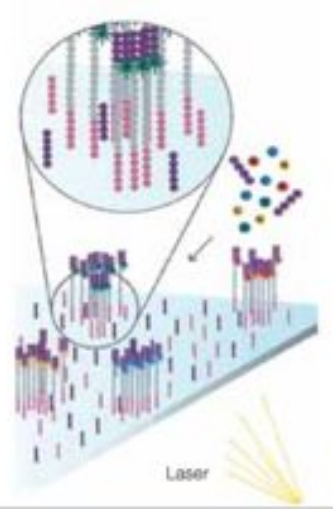
>60Gbp / day



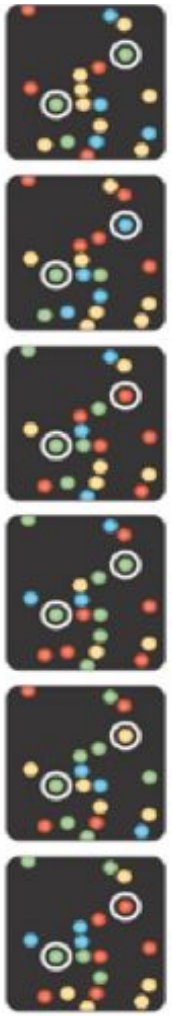
1. Attach



2. Amplify

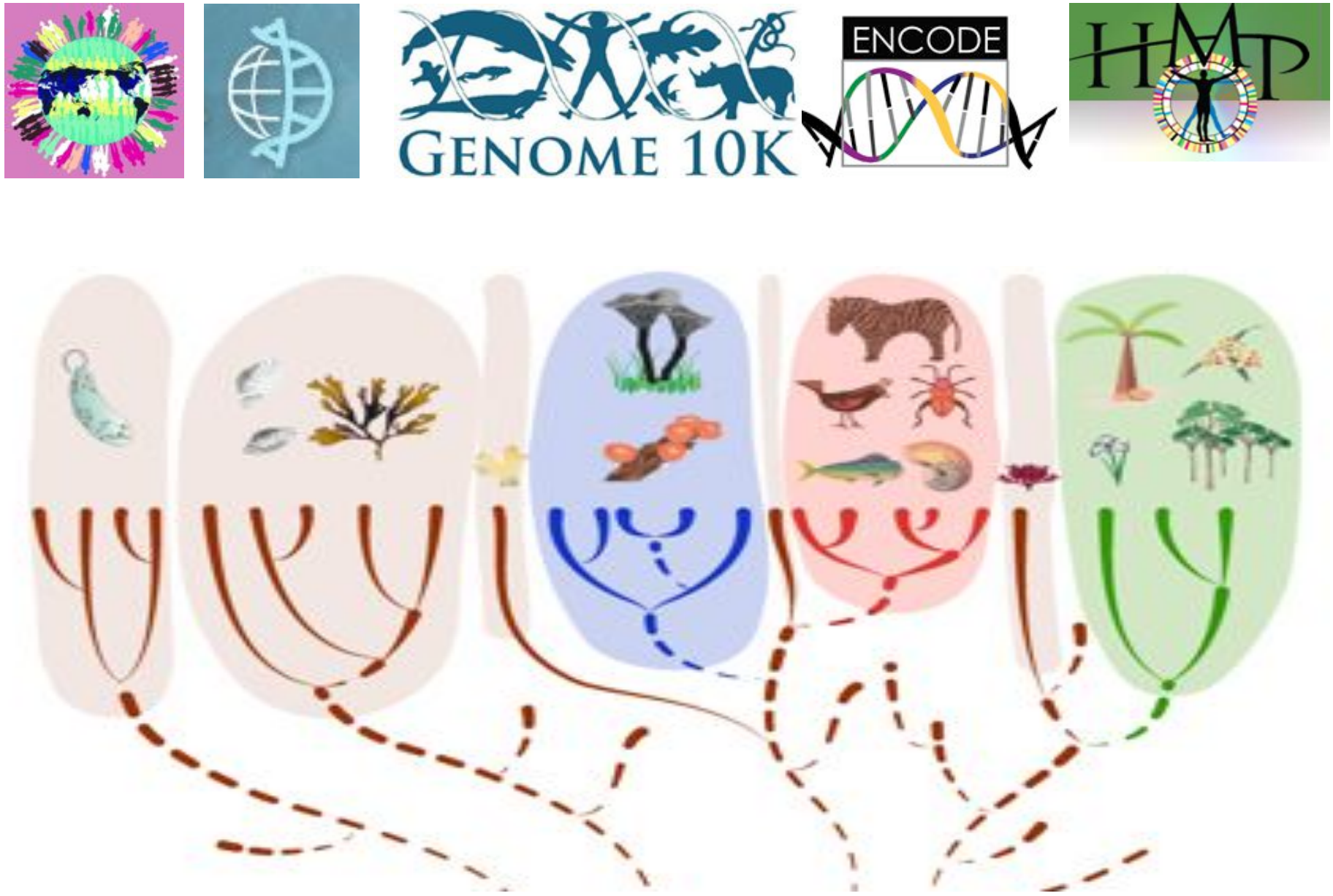


3. Image



Metzker (2010) Nature Reviews Genetics 11:31-46
http://www.youtube.com/watch?v=l99aKKHcxC4

Genomics across the tree of life



Unsolved Questions in Biology

- What is your genome sequence?

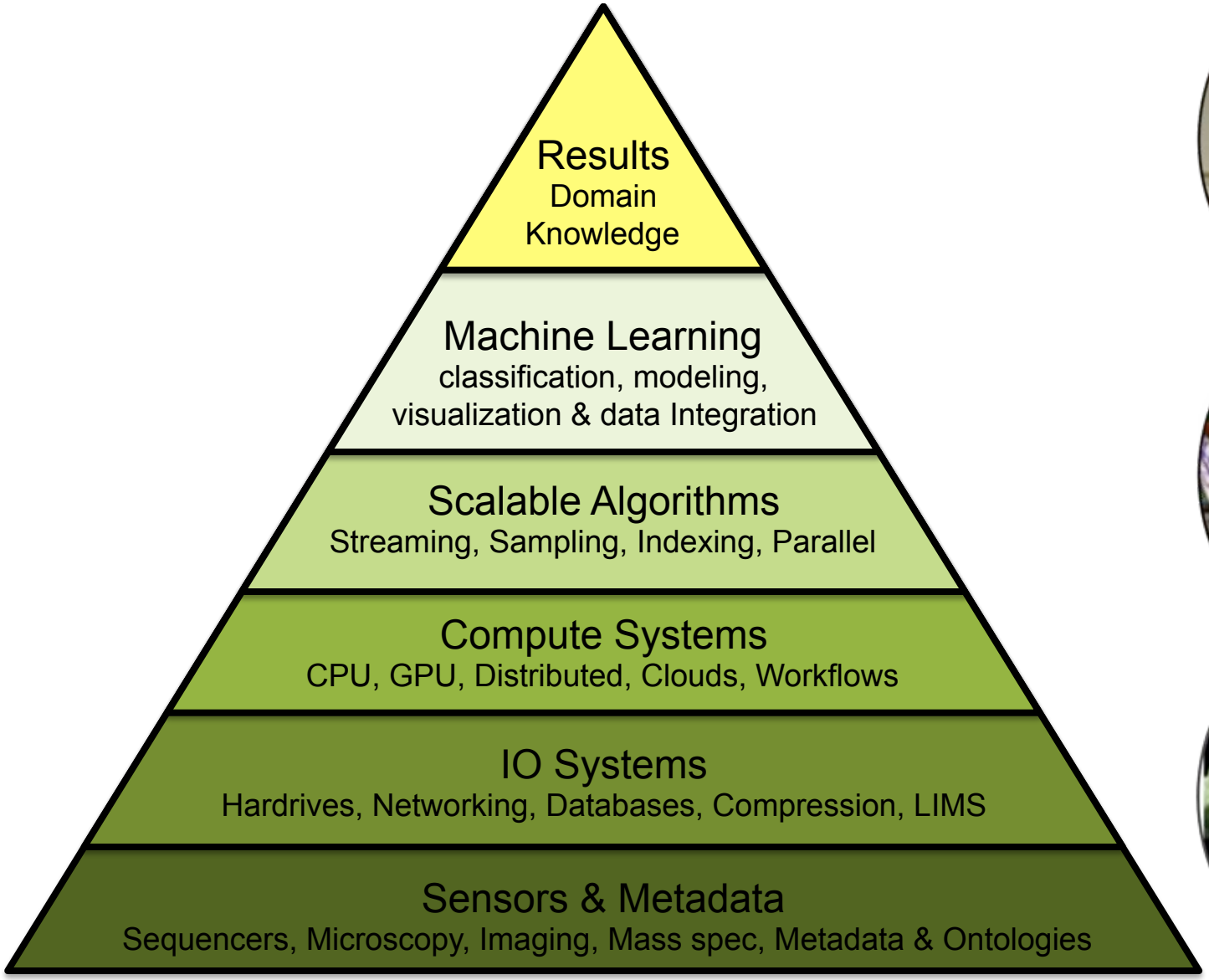
The instruments provide the data, but not the answers to any of these questions.

What software and systems will?

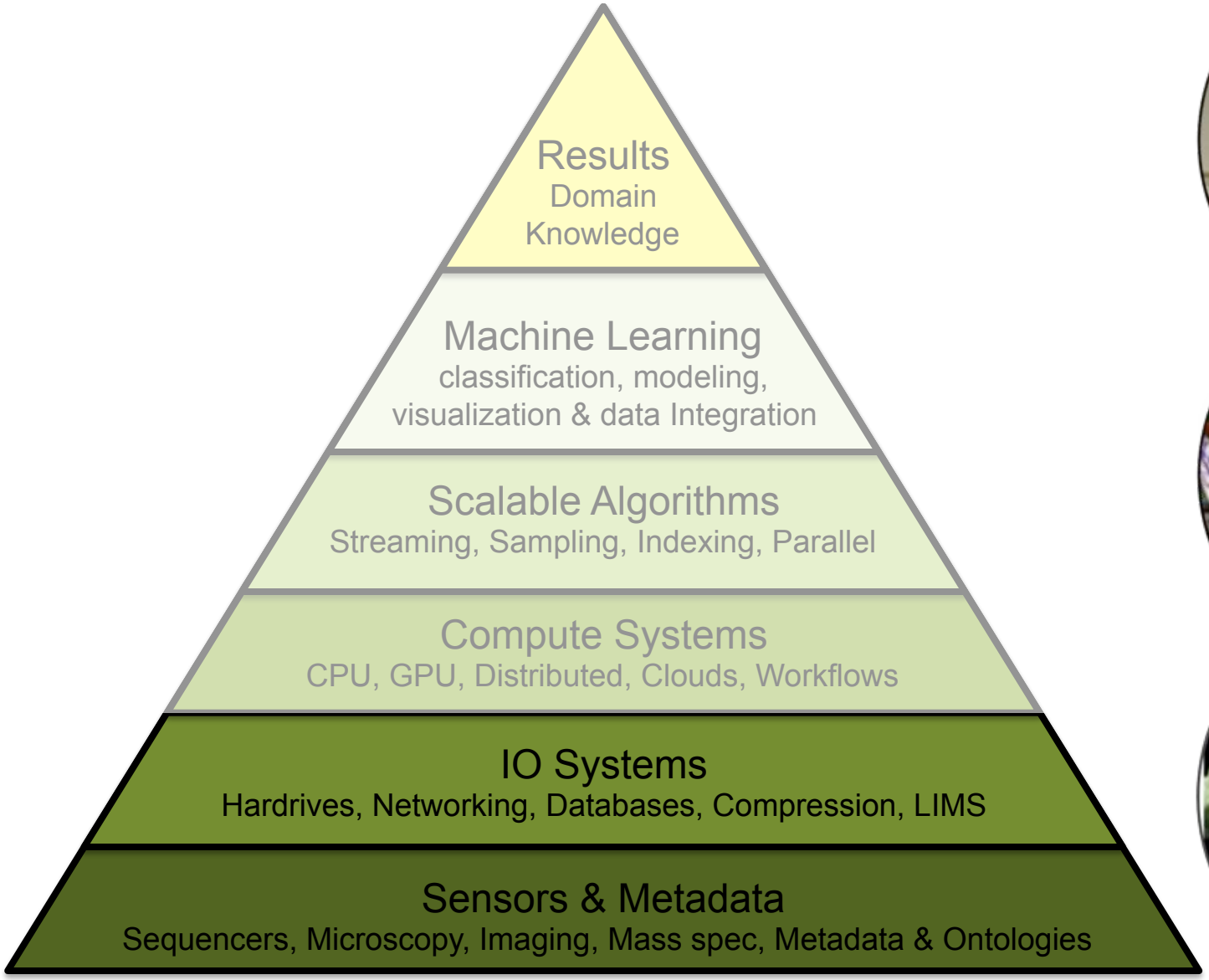
- What virus and microbes are living inside you?
- How do your mutations relate to disease?
- What drugs should we give you?

- Plus hundreds and hundreds more

Quantitative Biology Technologies

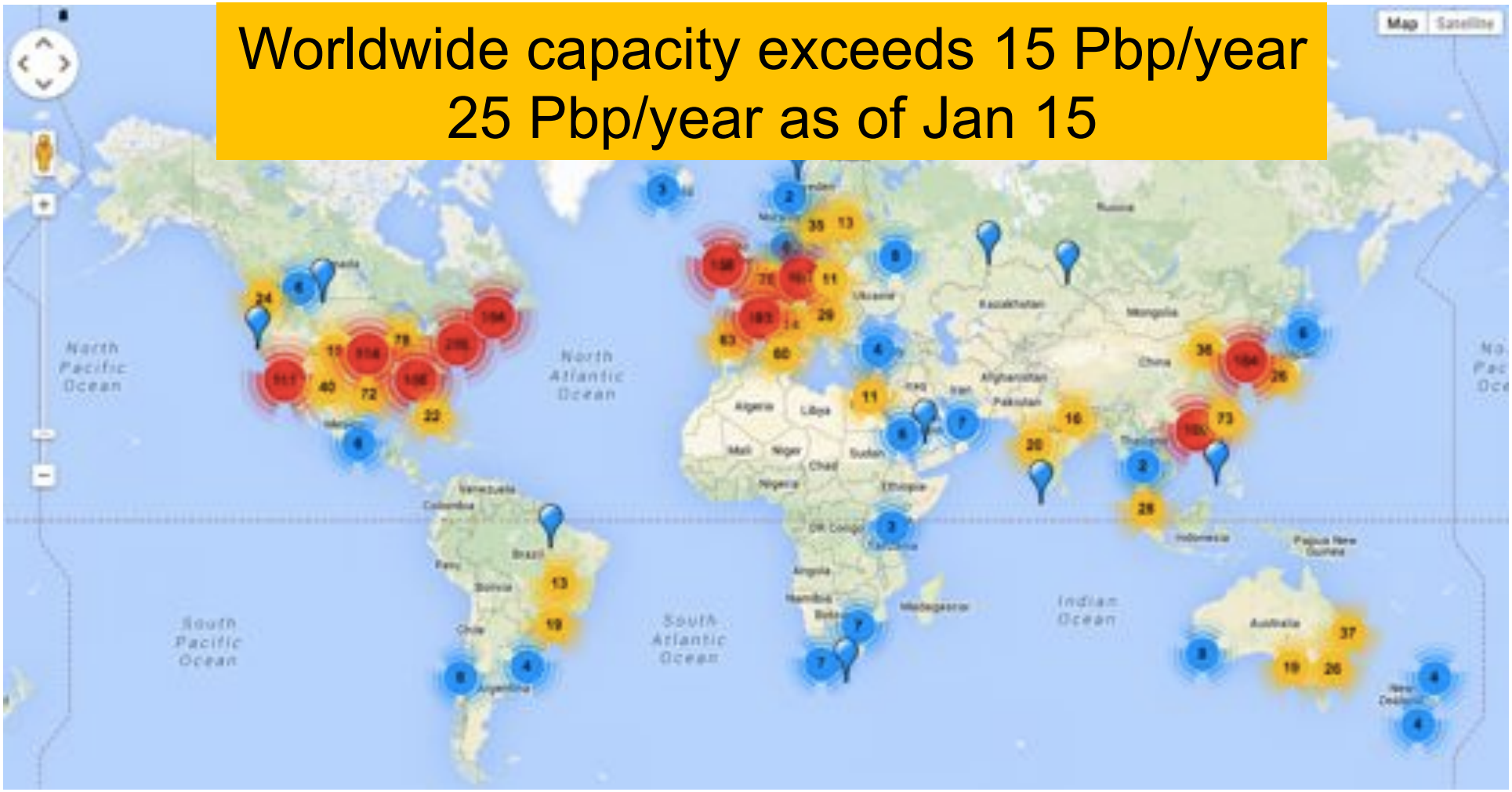


Quantitative Biology Technologies



Sequencing Centers

Worldwide capacity exceeds 15 Pbp/year
25 Pbp/year as of Jan 15



Next Generation Genomics: World Map of High-throughput Sequencers
http://omicsmaps.com

How much is a petabyte?

Unit	Size
Byte	1
Kilobyte	1,000
Megabyte	1,000,000
Gigabyte	1,000,000,000
Terabyte	1,000,000,000,000
Petabyte	1,000,000,000,000,000

*Technically a kilobyte is 2^{10} and a petabyte is 2^{50}

How much is a petabyte?

100 GB / Genome
4.7GB / DVD
~20 DVDs / Genome

X

10,000 Genomes

=

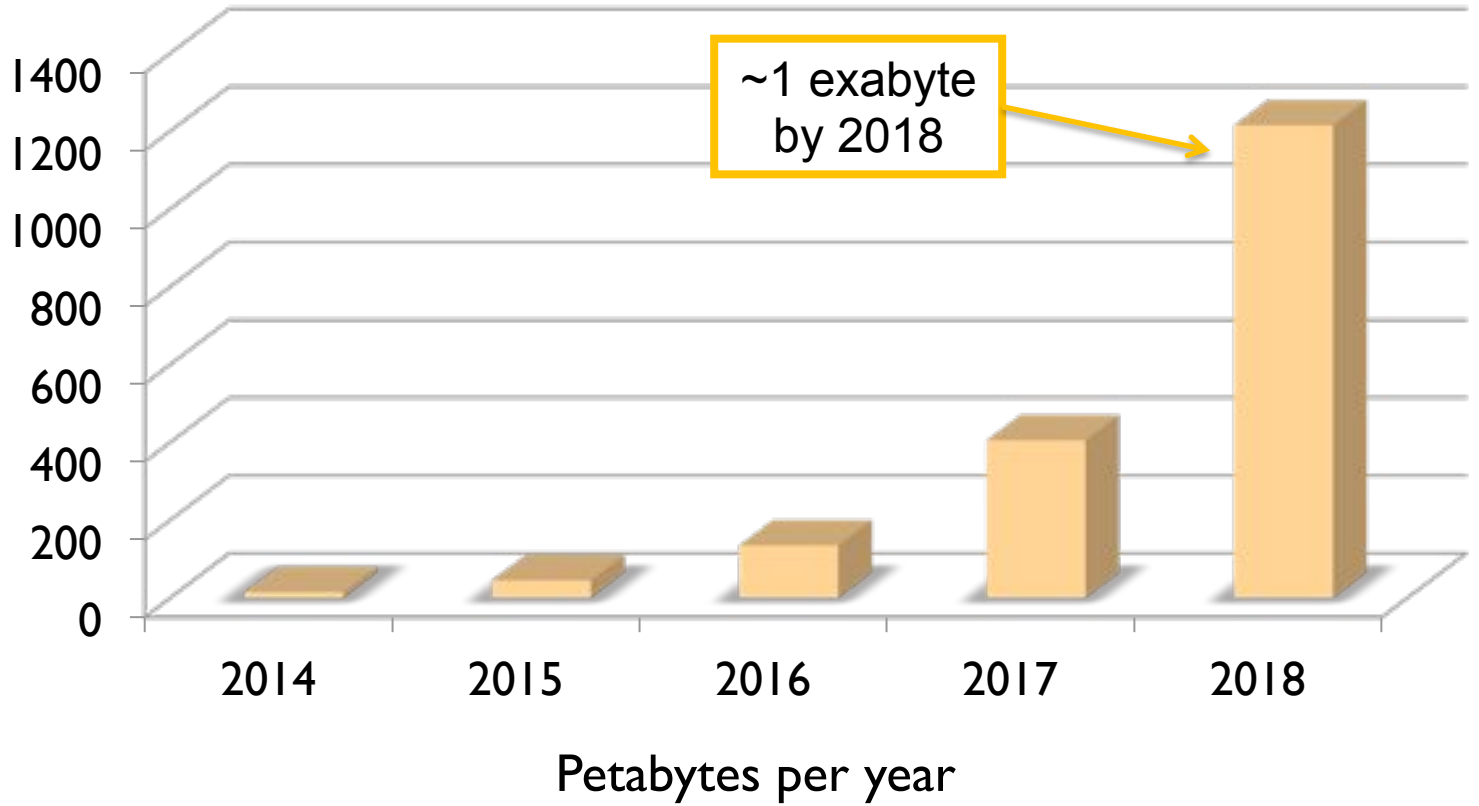
1PB Data
200,000 DVDs

787 feet of DVDs
~1/6 of a mile tall

500 2 TB drives
$500k

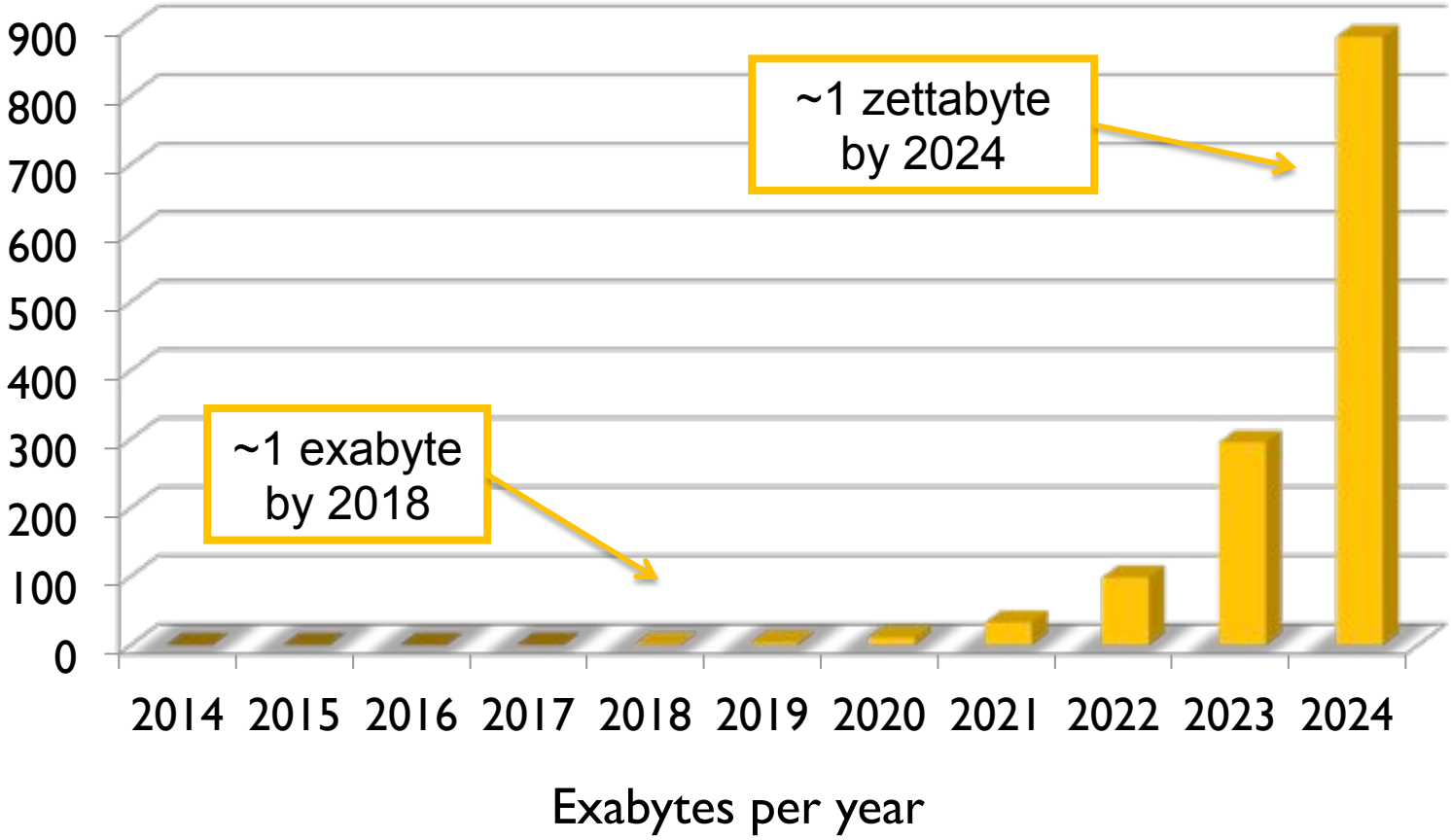
DNA Data Tsunami

Current world-wide sequencing capacity is growing at ~3x per year!



DNA Data Tsunami

Current world-wide sequencing capacity is growing at ~3x per year!



How much is a zettabyte?

Unit	Size
Byte	1
Kilobyte	1,000
Megabyte	1,000,000
Gigabyte	1,000,000,000
Terabyte	1,000,000,000,000
Petabyte	1,000,000,000,000,000
Exabyte	1,000,000,000,000,000,000
Zettabyte	1,000,000,000,000,000,000,000

How much is a zettabyte?

100 GB / Genome
4.7GB / DVD
~20 DVDs / Genome

X

10,000,000,000 Genomes

=

1ZB Data
200,000,000,000 DVDs

150,000 miles of DVDs
~ ½ distance to moon

YouTube

Both currently ~100Pb
But growing exponentially

Sequencing Centers 2014

Next Generation Genomics: World Map of High-throughput Sequencers
http://omicsmaps.com

Sequencing Centers 2024

Next Generation Genomics: World Map of High-throughput Sequencers
http://omicsmaps.com

Biological Sensor Network

Oxford Nanopore

DC Metro via the LA Times

The rise of a digital immune system
Schatz, MC, Phillippy, AM (2012) GigaScience 1:4

Data Production & Collection

Expect massive growth to sequencing and other biological sensor data over the next 10 years

- Exascale biology is certain, zettascale on the horizon
- Compression helps, but need to aggressively throw out data
- Requires careful consideration of the "preciousness" of the sample

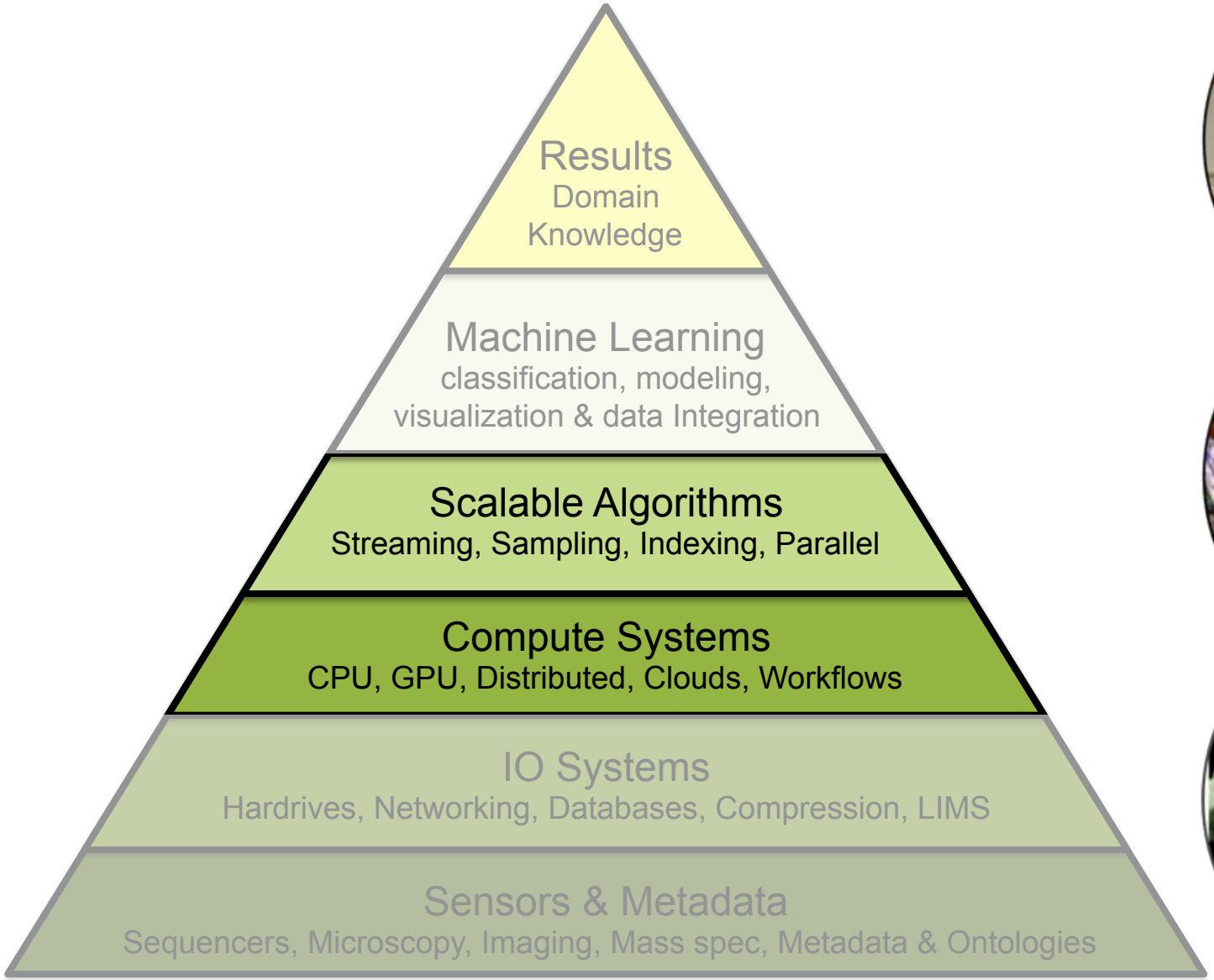
Major data producers concentrated in hospitals, universities, agricultural companies, research institutes

- Major efforts in human health and disease, agriculture, bioenergy

But also widely distributed mobile sensors

- Schools, offices, sports arenas, transportations centers, farms & food distribution centers
- Monitoring and surveillance, as ubiquitous as weather stations
- The rise of a digital immune system?

Quantitative Biology Technologies

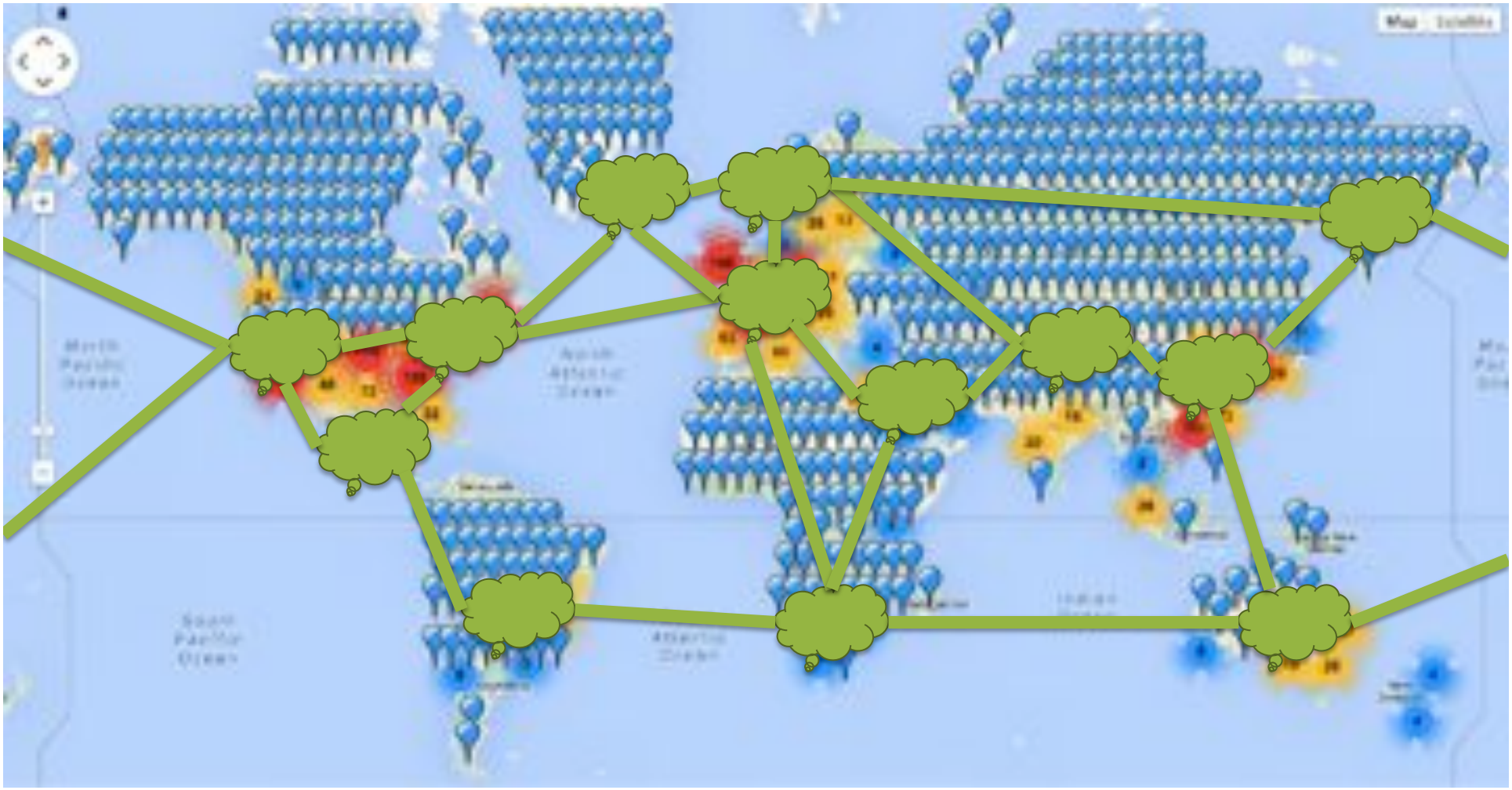


Sequencing Centers 2024

Informatics Centers 2024

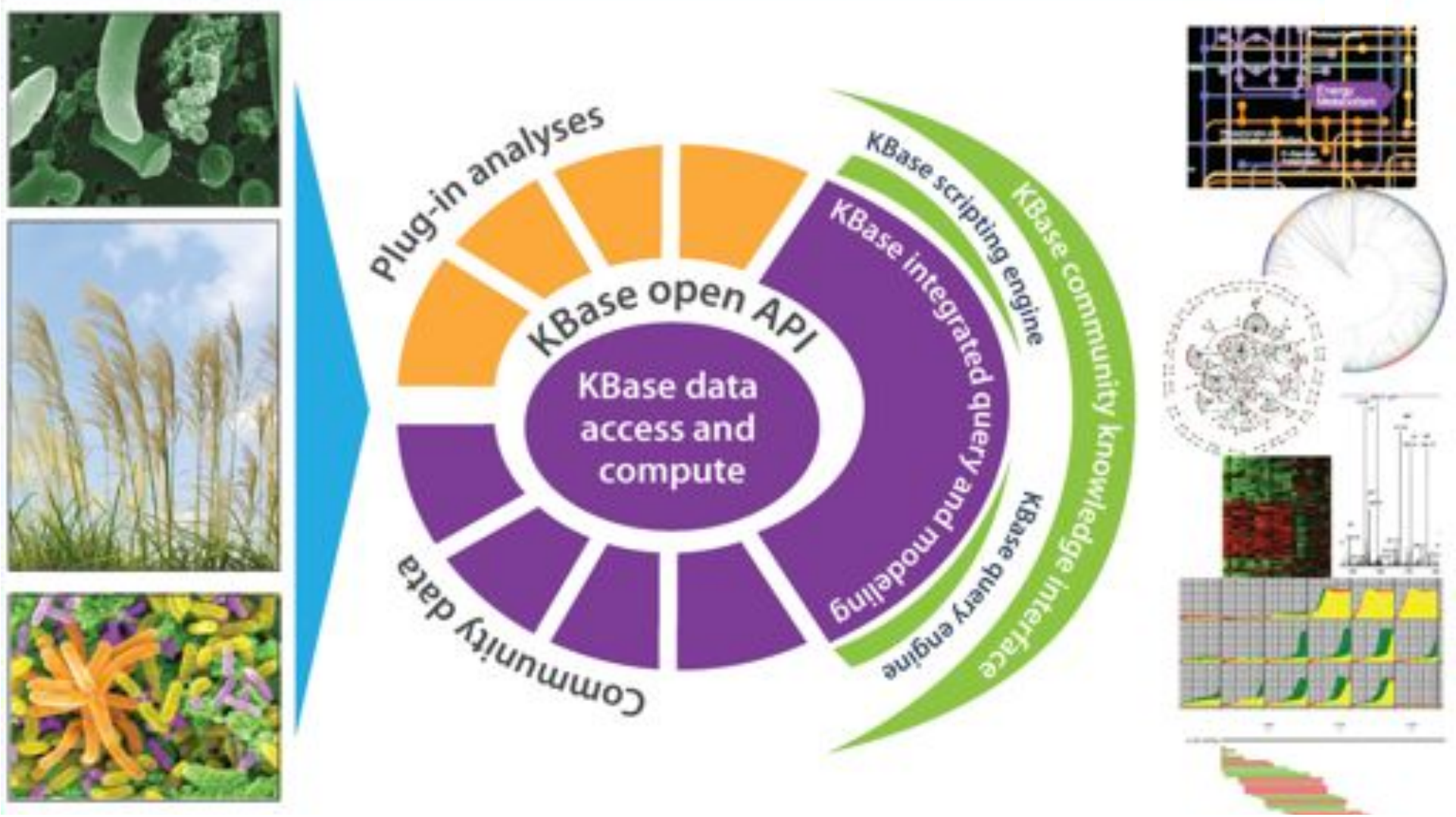
The DNA Data Deluge
Schatz, MC and Langmead, B (2013) *IEEE Spectrum*. July, 2013

Informatics Centers 2014



The DNA Data Deluge
Schatz, MC and Langmead, B (2013) *IEEE Spectrum*. July, 2013

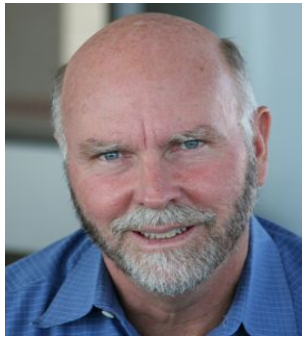
DOE Systems Biology Knowledgebase

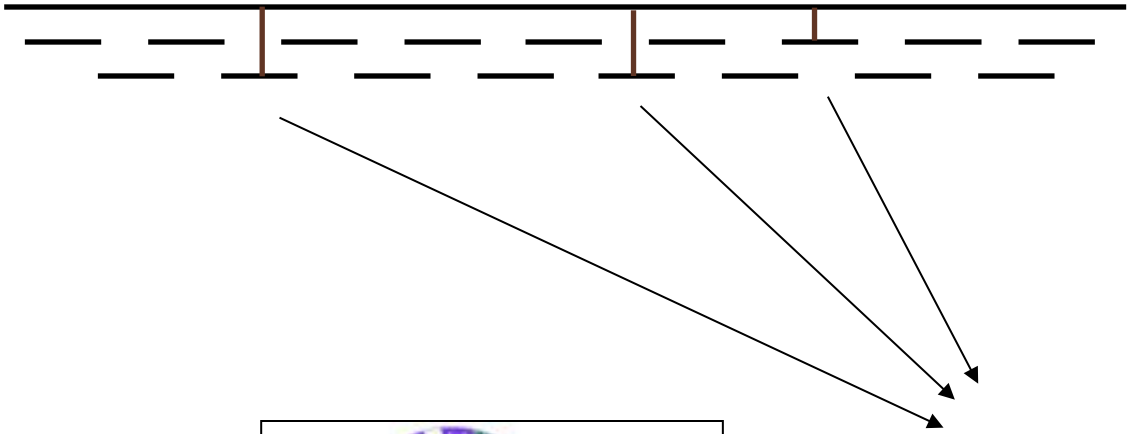


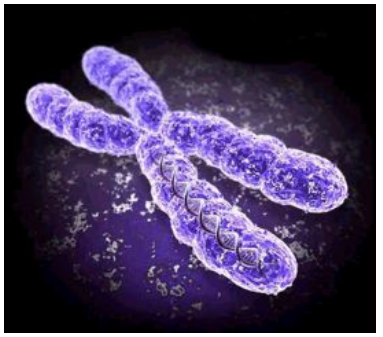
http://kbase.us: Predictive Biology in Microbes, Plants, and Meta-communities

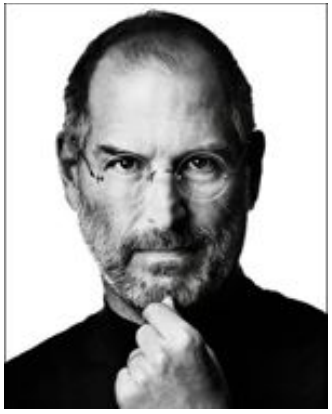
Personal Genomics

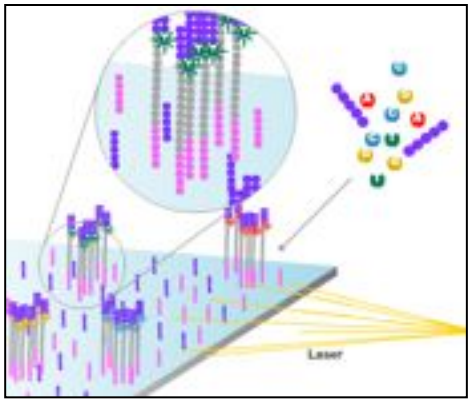
How does your genome compare to the reference?











Heart Disease

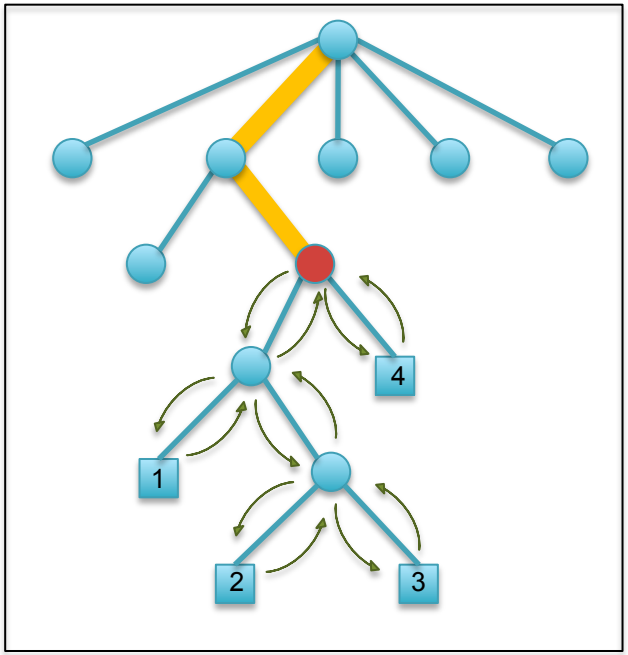
Cancer

Creates magical technology

MUMmerGPU

http://mummergpu.sourceforge.net

- Map many reads simultaneously on GPU
  - Find matches by walking the tree
  - Find coordinates with depth first search



- Performance on nVidia GTX 8800
  - Match kernel was ~10x faster than CPU
  - Search kernel was ~4x faster than CPU
  - End-to-end runtime ~4x faster than CPU

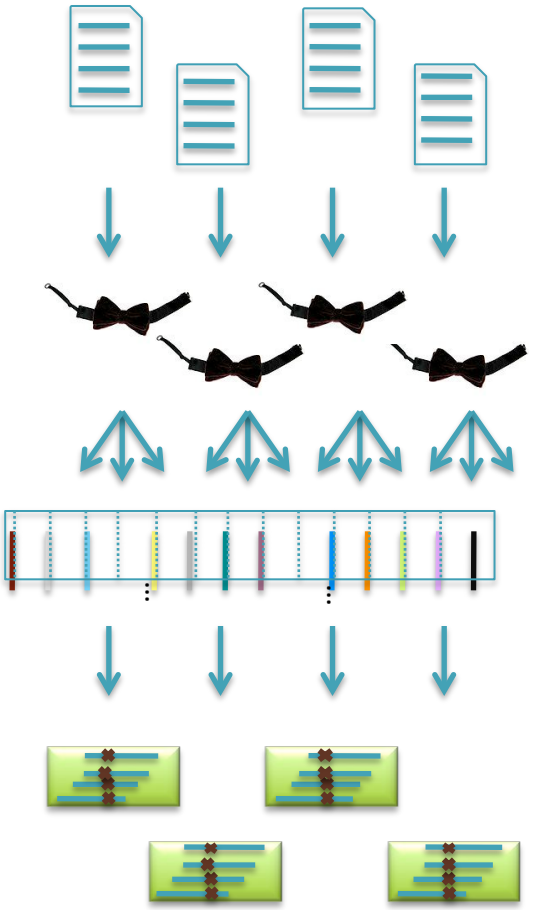
- Cores are only part of the solution.
- Need fast storage & IO
- Locality is king

High-throughput sequence alignment using Graphics Processing Units.
Schatz, MC, Trapnell, C, Delcher, AL, Varshney, A. (2007) BMC Bioinformatics 8:474.

Crossbow

http://bowtie-bio.sourceforge.net/crossbow

- Align billions of reads and find SNPs
  - Reuse software components: Hadoop Streaming
  - Mapping with Bowtie, SNP calling with SOAPsnp
- 4 hour end-to-end runtime including upload
  - Costs $85; Todays costs <$10

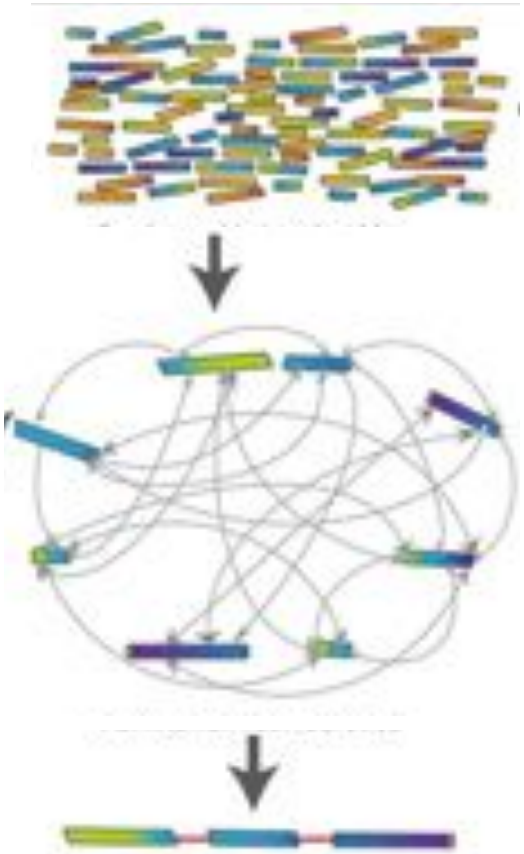


- Very compelling example of cloud computing in genomics
- Commercial vendors probably have better security than your institution
- Need more applications!

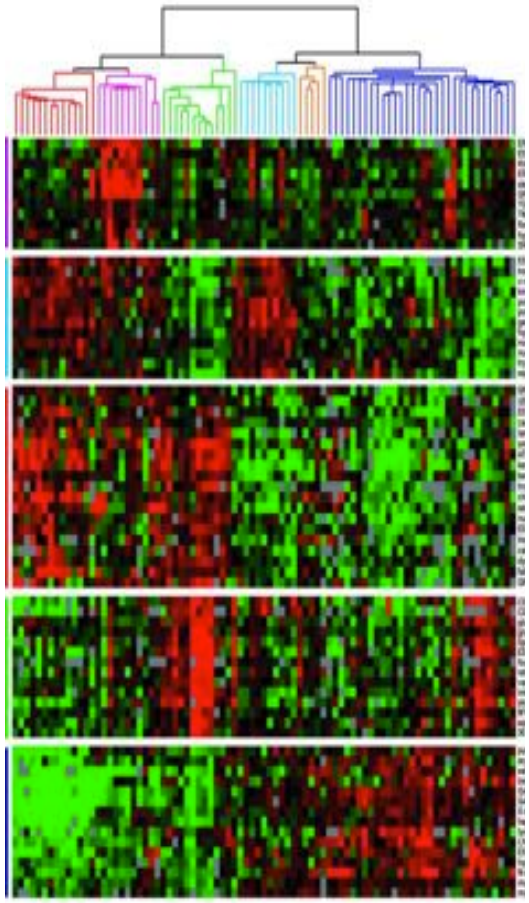
Searching for SNPs with Cloud Computing.
Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL (2009) *Genome Biology*. **10**:R134

Genomics Algorithms

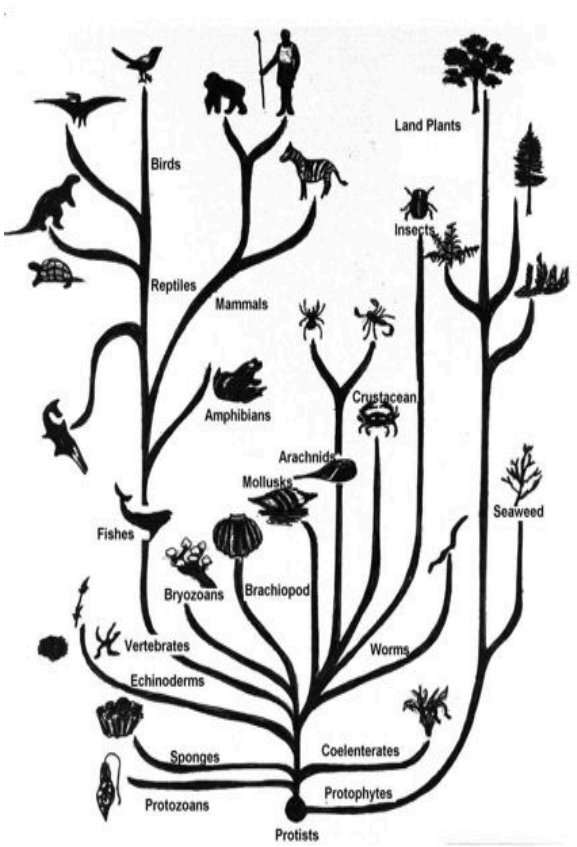
De novo Assembly



Differential Analysis



Phylogeny, Evolution, and Modeling



Compute & Algorithmic Challenges

Expect to see many dozens of major informatics centers that consolidate regional / topical information

- Clouds for Cancer, Autism, Heart Disease, etc
- Plus many smaller warehouses down to individuals
- Move the code to the data

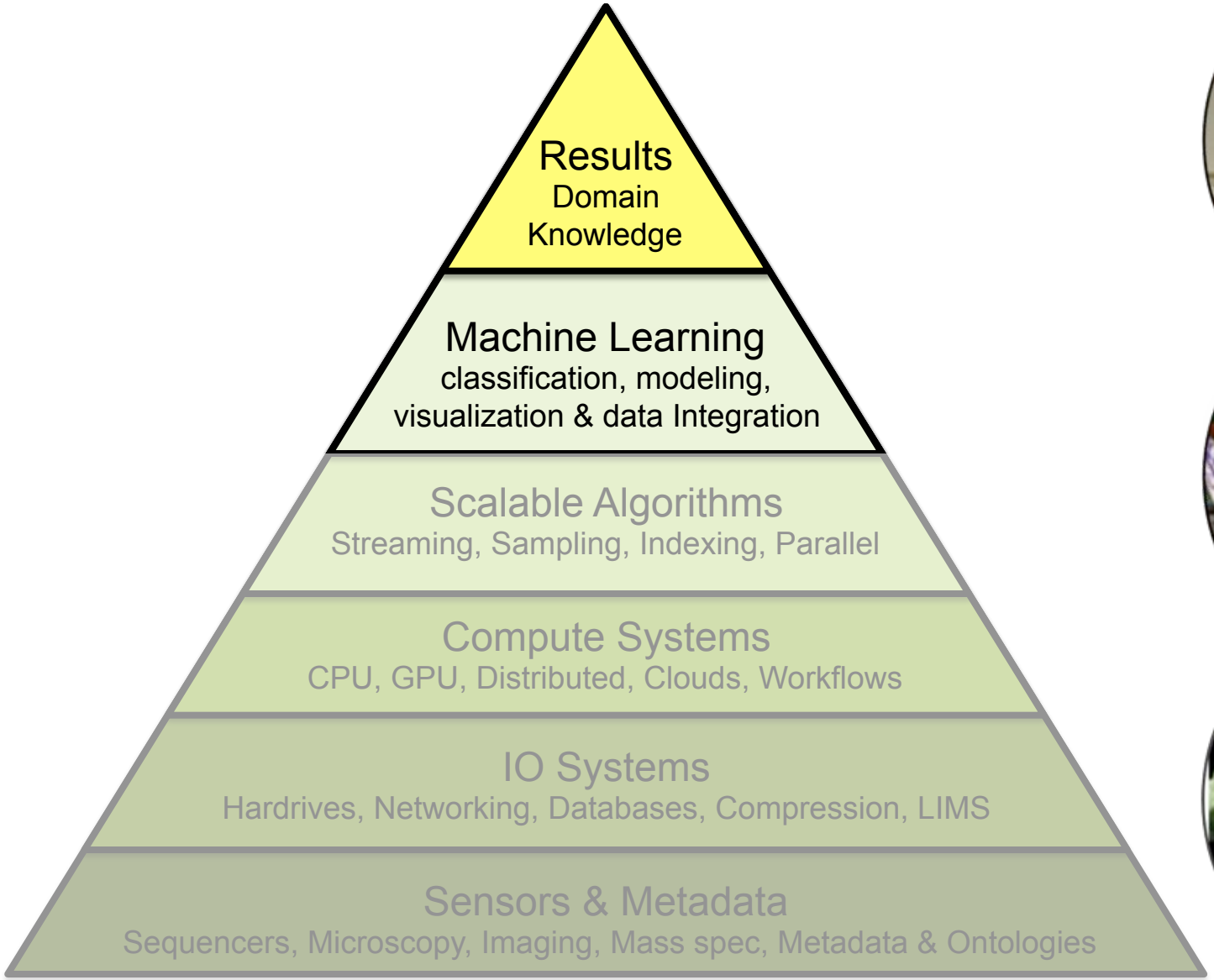
Parallel hardware and algorithms are required

- Expect to see >1000 cores in a single computer
- Compute & IO needs to be considered together
- Rewriting efficient parallel software is complex and expensive

Applications will shift from individuals to populations

- Read mapping & assembly fade out
- Population analysis and time series analysis fade in
- Need for network analysis, probabilistic techniques

Quantitative Biology Technologies



Genetic Basis of Autism Spectrum Disorders

Complex disorders of brain development

- Characterized by difficulties in social interaction, verbal and nonverbal communication and repetitive behaviors.
- Have their roots in very early brain development, and the most obvious signs of autism and symptoms of autism tend to emerge between 2 and 3 years of age.

U.S. CDC identify around 1 in 68 American children as on the autism spectrum

- Ten-fold increase in prevalence in 40 years, only partly explained by improved diagnosis and awareness.
- Studies also show that autism is four to five times more common among boys than girls.
- Specific causes remain elusive

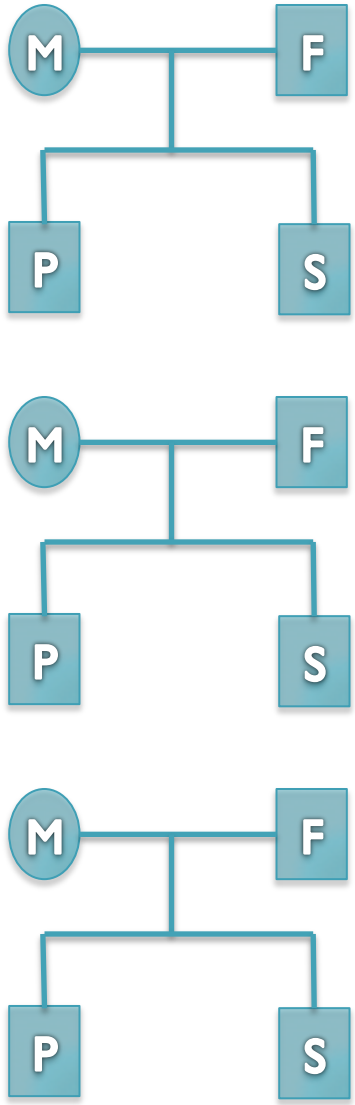
What is Autism?
http://www.autismspeaks.org/what-autism

Searching for the genetic risk factors

Search Strategy

- Thousands of families identified from a dozen hospitals around the United States
- Large scale genome sequencing of "simplex" families: mother, father, affected child, unaffected sibling
- Unaffected siblings provide a natural control for environmental factors

Are there any genetic variants present in affected children, that are not in their parents or unaffected siblings?



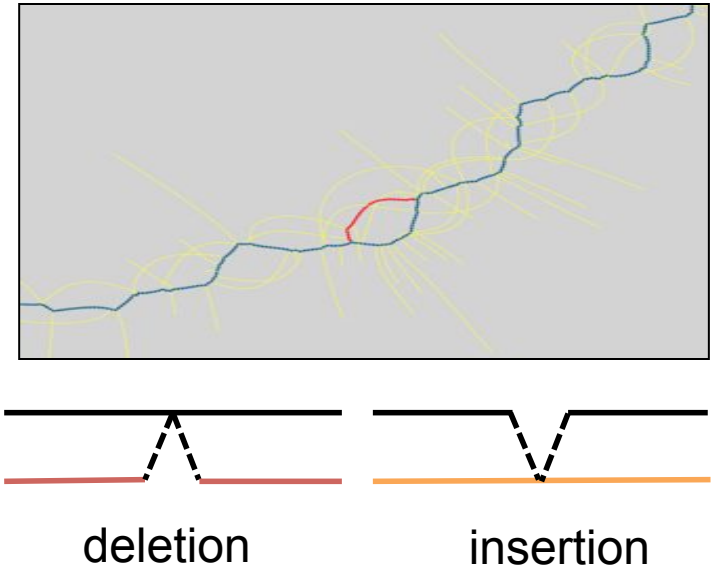
SFARI SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

Scalpel: Haplotype Microassembly

DNA sequence **micro-assembly** pipeline for accurate detection and validation of *de novo* mutations (SNPs, indels) within exome-capture data.

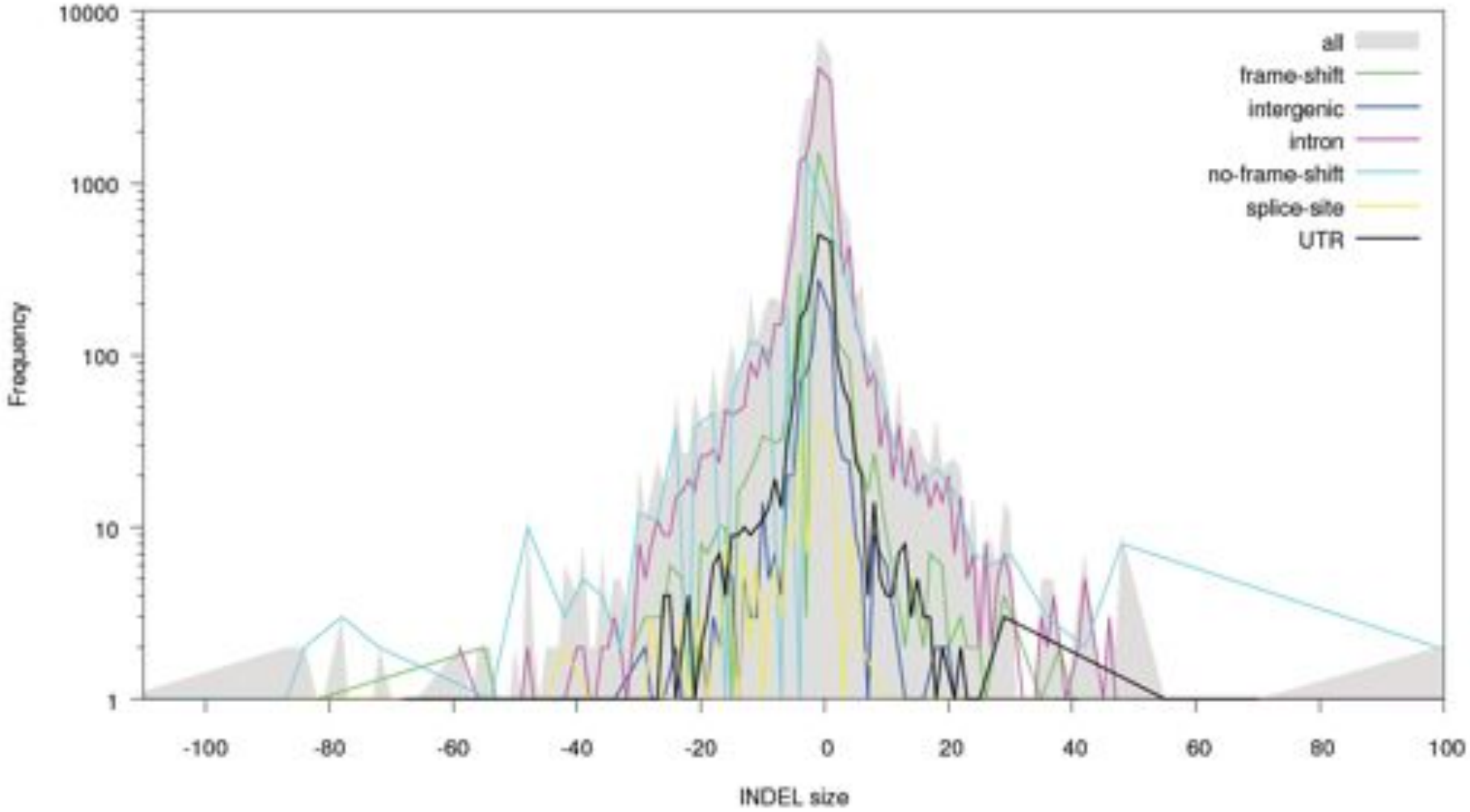
Features

1. Combine mapping and assembly
2. Exhaustive search of haplotypes
3. De novo mutations



Accurate detection of de novo and transmitted INDELs within exome-capture data using micro-assembly
Narzisi, G, O'Rawe, J, Iossifov, I, Lee, Y, Wang, Z, Wu, Y, Lyon, G, Wigler, M, Schatz, MC (2014) *Under review.*

Population Analysis of the SSC

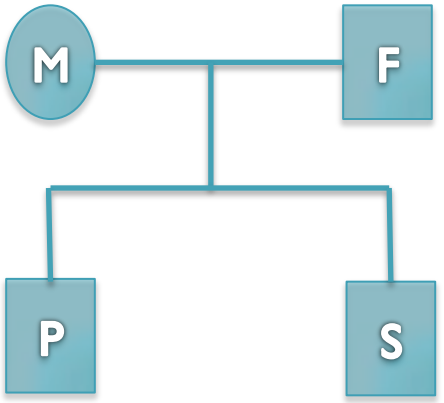


Constructed database of >1M transmitted and de novo indels

De novo mutation discovery and validation

Concept: Identify mutations not present in parents.

Challenge: Sequencing errors in the child or low coverage in parents lead to false positive de novos




```
Reference:   ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...

Father:      ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Mother:      ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Sibling:     ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Proband(1):  ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Proband(2):  ...TCAAATCCTTTTAAT****AAGAGCTGACA...
```

4bp heterozygous deletion at chr15:93524061 CHD2

De novo Genetics of Autism

- In 593 family quads so far, we see significant enrichment in de novo ***likely gene killers*** in the autistic kids
  - Overall rate basically 1:1
  - 2:1 enrichment in nonsense mutations
  - 2:1 enrichment in frameshift indels
  - 4:1 enrichment in splice-site mutations
  - Most de novo originate in the paternal line in an age-dependent manner (56:18 of the mutations that we could determine)

- Observe strong overlap with the 842 genes known to be associated with fragile X protein FMPR
  - Related to neuron development and synaptic plasticity
  - Also strong overlap with chromatin remodelers

Accurate detection of de novo and transmitted INDELs within exome-capture data using micro-assembly
Narzisi, G, O'Rawe, J, Iossifov, I, Lee, Y, Wang, Z, Wu, Y, Lyon, G, Wigler, M, Schatz, MC (2014) Under review.

Learning and Translation

Tremendous power from data aggregation

- Observe the dynamics of biological systems
- Breakthroughs in medicine and biology of profound significance

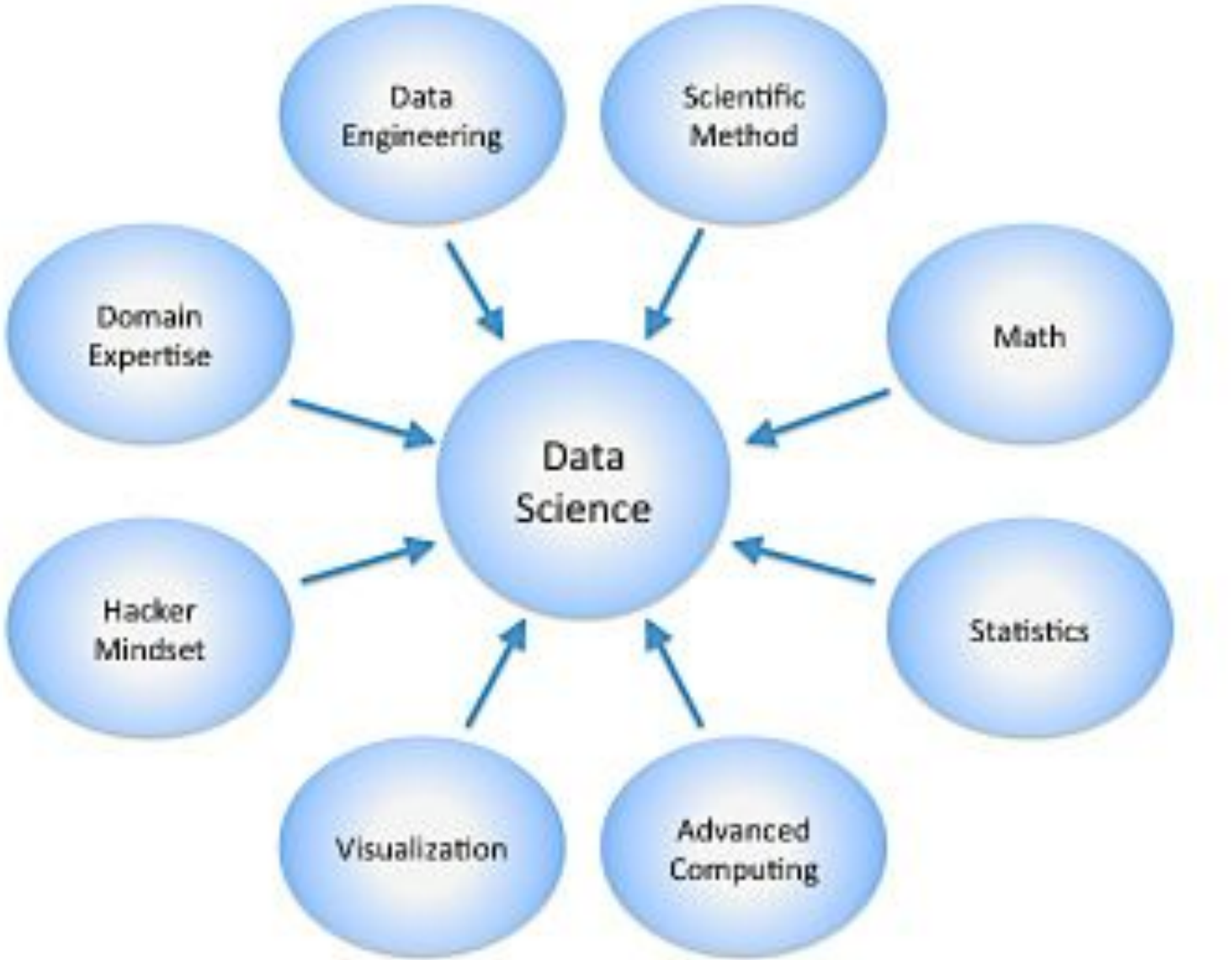
Be mindful of the risks

- The potential for over-fitting grows with the complexity of the data, statistical significance is a statement about the sample size
- Reproducible workflows, APIs are a must
- Caution is prudent for personal data

The foundations of biology will continue to be observation, experimentation, and interpretation

- Technology will continue to push the frontier
- Feedback loop from the results of one project into experimental design for the next

Who is a Data Scientist?



http://en.wikipedia.org/wiki/Data_science

Acknowledgements

Schatz Lab

Giuseppe Narzisi

Shoshana Marcus

James Gurtowski

Srividya Ramakrishnan

Hayan Lee

Rob Aboukhalil

Mitch Bekritsky

Charles Underwood

Tyler Gavin

Alejandro Wences

Greg Vurture

Eric Biggers

Aspyn Palatnick

CSHL

Hannon Lab

Gingeras Lab

Jackson Lab

Iossifov Lab

Levy Lab

Lippman Lab

Lyon Lab

Martienssen Lab

McCombie Lab

Tuveson Lab

Ware Lab

Wigler Lab

IT Department

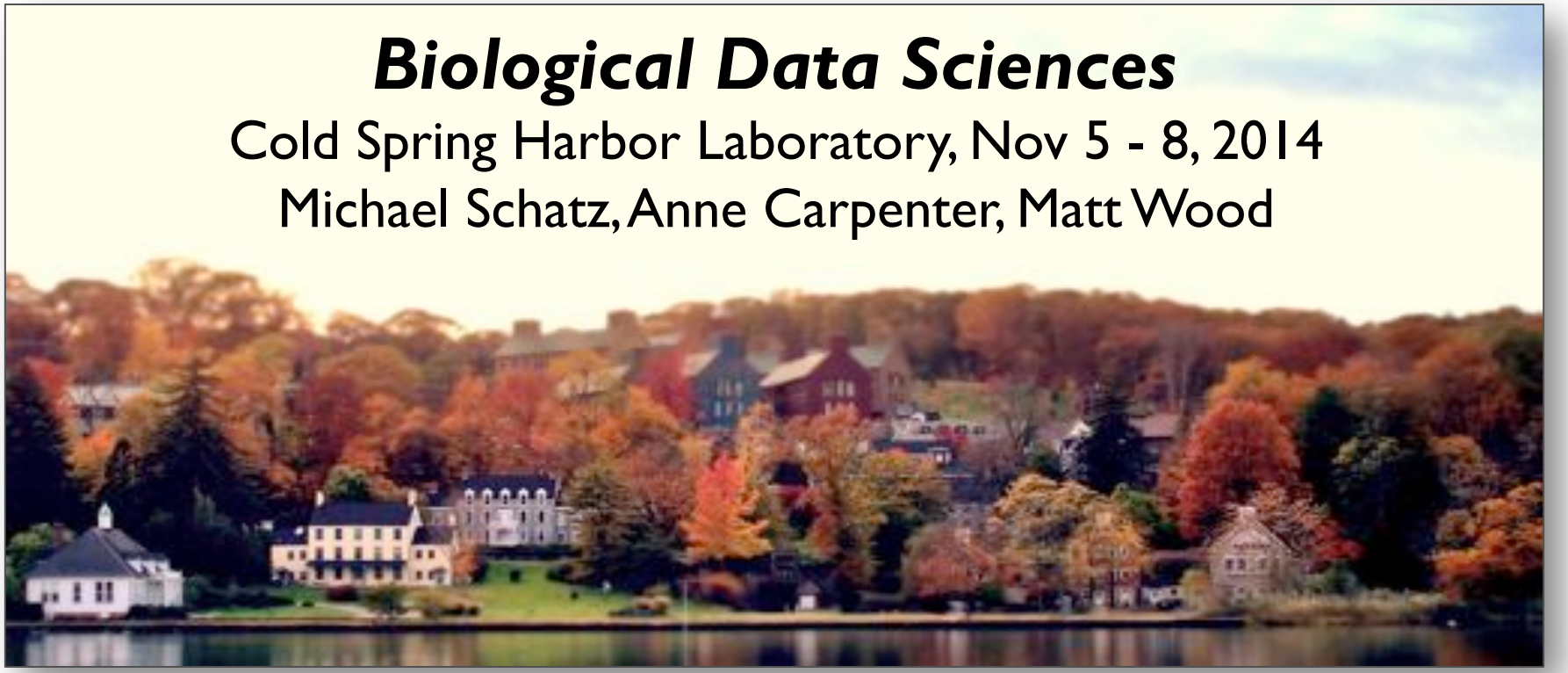
SFARI

SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

National Human Genome Research Institute

U.S. DEPARTMENT OF ENERGY

NSF



Thank you

http://schatzlab.cshl.edu

@mike_schatz