Big Data Meets DNA

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

Michael Schatz

June 18, 2014

CSHL Public Lecture Series

DNA: The secret of life

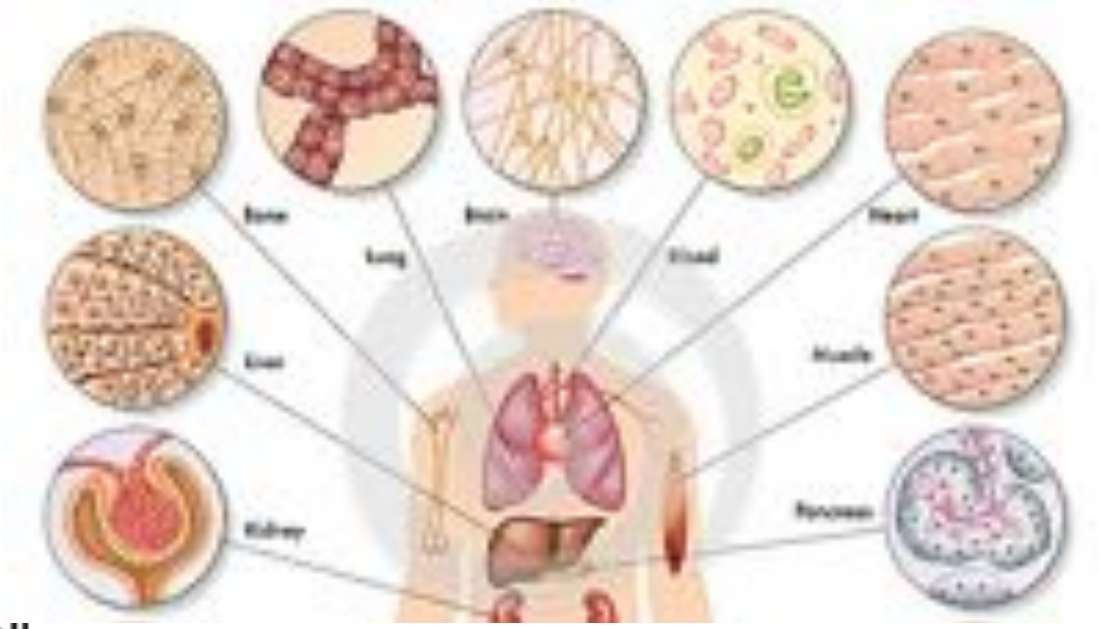
Your DNA, along with your environment and experiences, shapes who you are

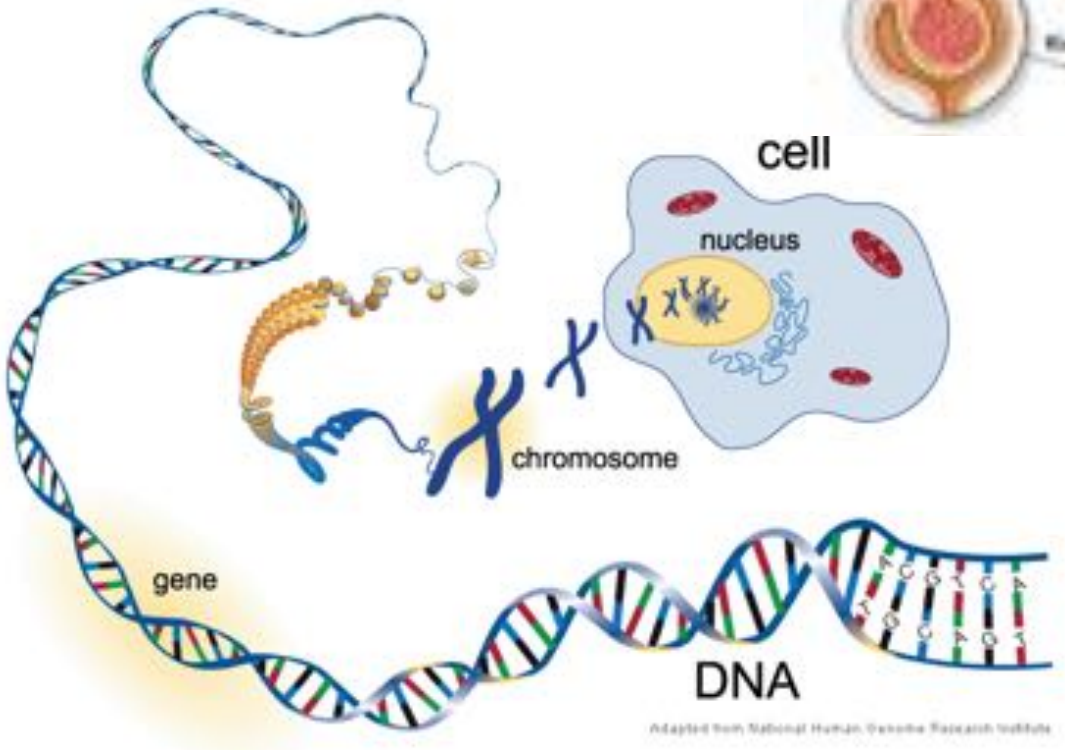
- Height
- Hair, eye, skin color
- Broad/narrow, small/large features
- Susceptibility to disease
- Response to drug treatments
- Longevity and cognition

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

Cells & DNA

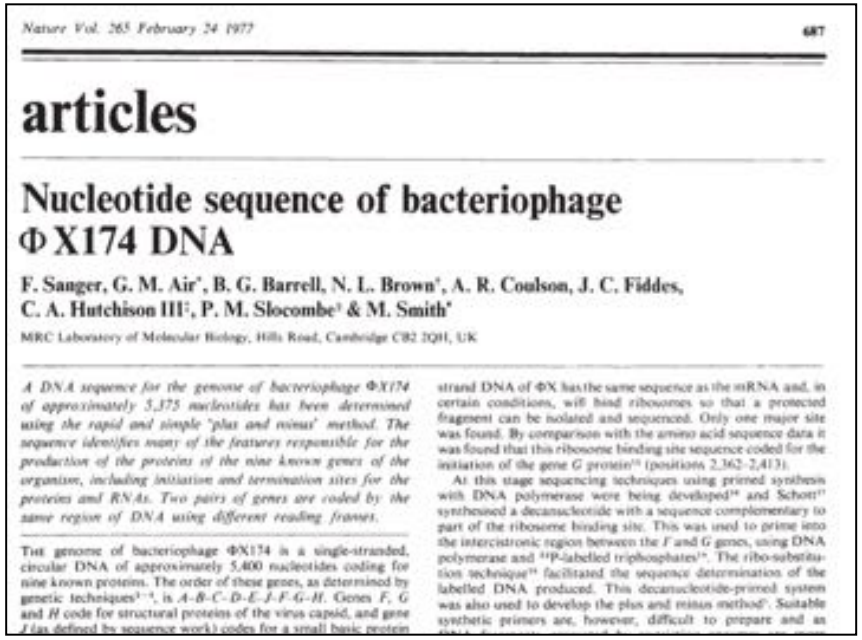
Each cell of your body contains an exact copy of your 3 billion base pair genome.





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.

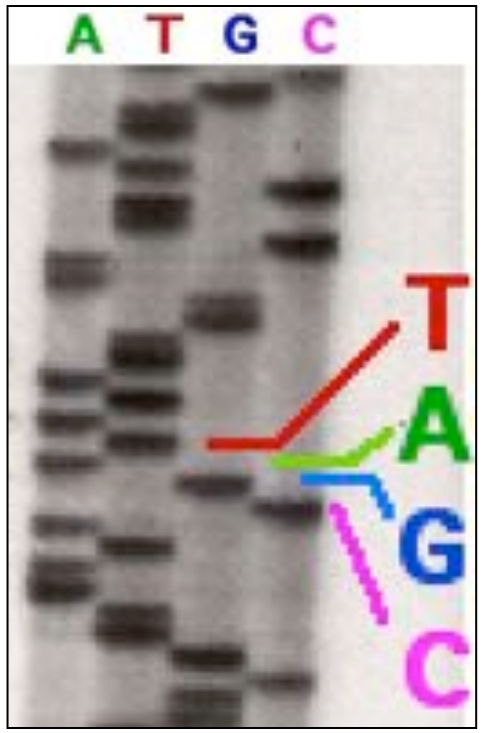
THE genome of bacteriophage ΦX174 is a single-stranded, circular DNA of approximately 5,400 nucleotides coding for nine known proteins. The order of these genes, as determined by genetic techniques[1-3], is *A–B–C–D–E–J–F–G–H*. Genes *F*, *G* and *H* code for structural proteins of the virus capsid, 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[4] (positions 2,362–2,413).

At this stage sequencing techniques using primed synthesis with DNA polymerase were being developed[5,6] and Schott[7] 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[5]. The ribo-substitution technique[8] 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



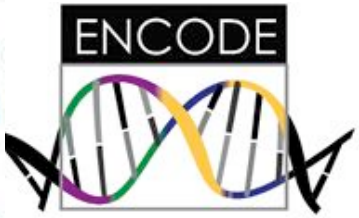
Radioactive Chain Termination
5000bp / week / person

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

Milestones in DNA Sequencing

(TIGR/Celera, 1995-2001)

Genomics across the tree of life



Unsolved Questions in Biology

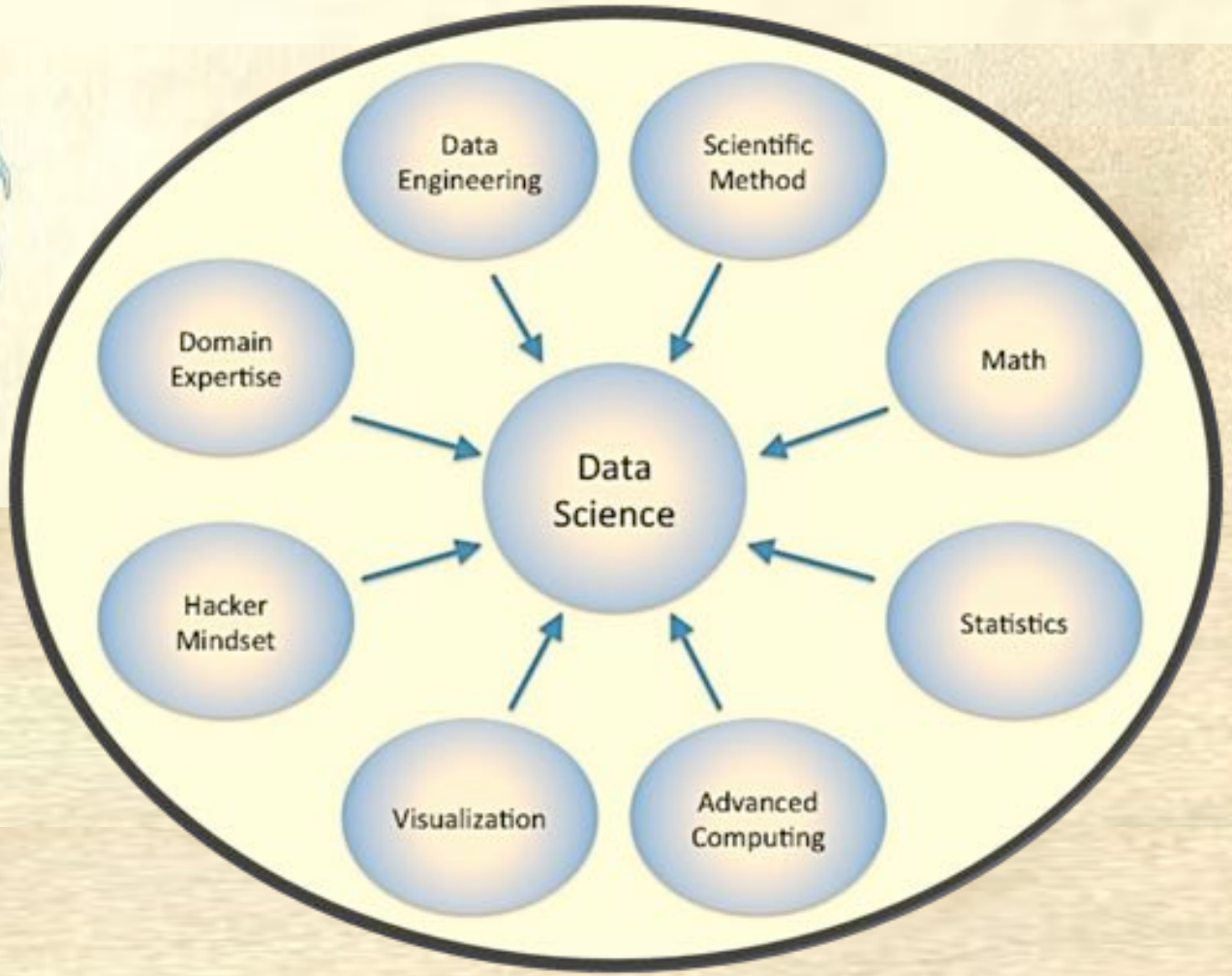
- What is your genome sequence?
-
-
-
-
-
-
-
-
-
-
-
-
- ***Plus hundreds and hundreds more***

The instruments provide the data, but none of the answers to any of these questions.

What software and systems will?

And who will create them?

Who is a Data Scientist?



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

CSHL Quantitative Biology

Mickey Atwal
Population Genetics
Cancer, Fertility

Molly Hammel
Gene regulatory
Networks, RNA Biology

Ivan Iossifov
Human Genetics
Molecular Networks

Justin Kinney
Biophysics
Machine learning

Alexei Koulakov
Neurobiology
Cortical design, Memory

Alex Krasnitz
Genomics of Cancer
Machine Learning

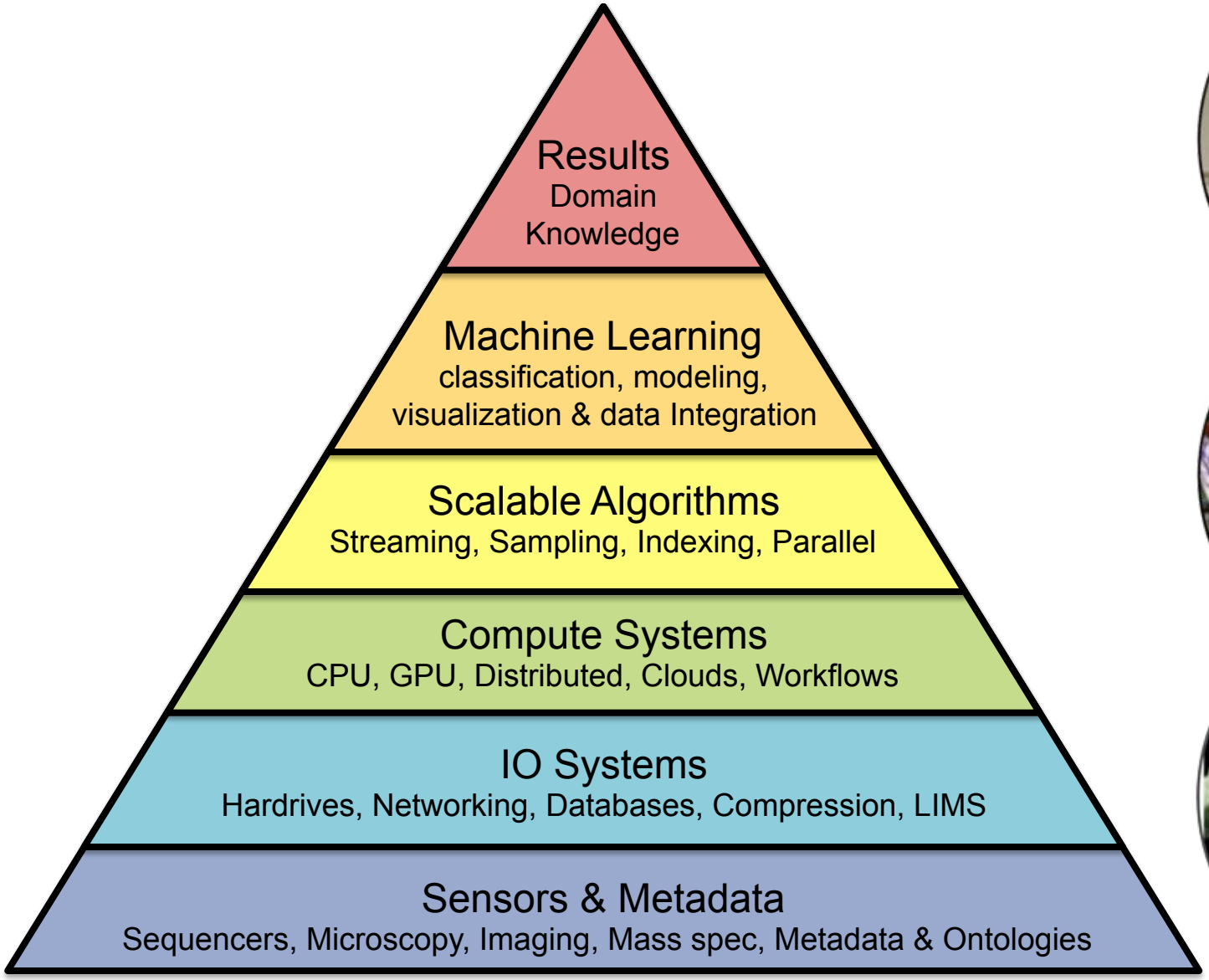
Dan Levy
Human Genetics
Phylogenetics, CNVs

Partha Mitra
Neuroscience
Neural Imaging & Disease

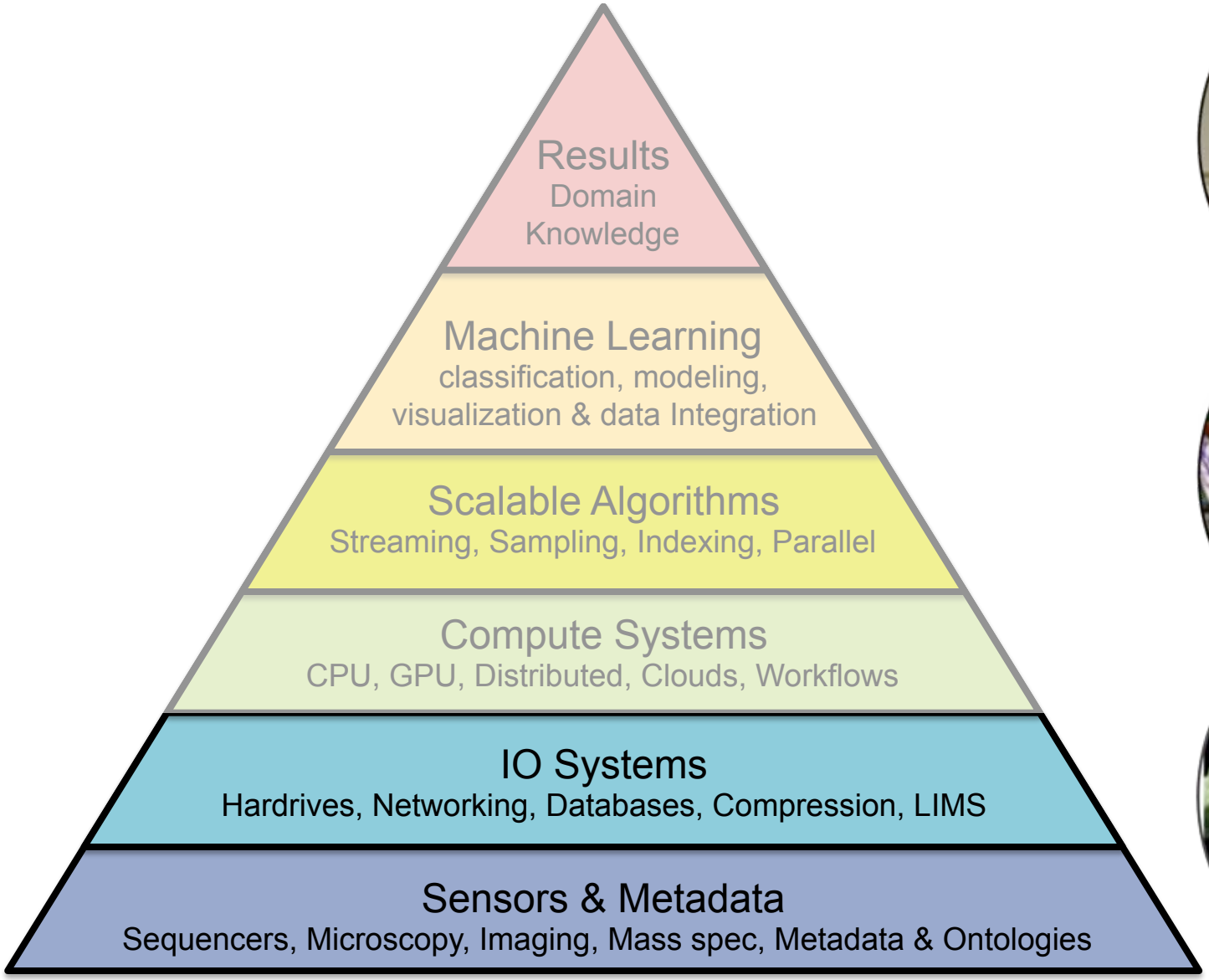
Adam Siepel
Evolution
Functional Annotation

Michael Wigler
Genetic Disorders
Cancer, Autism

Quantitative Biology Technologies



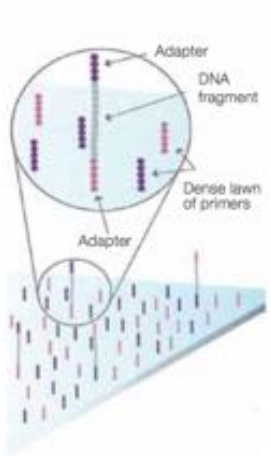
Quantitative Biology Technologies



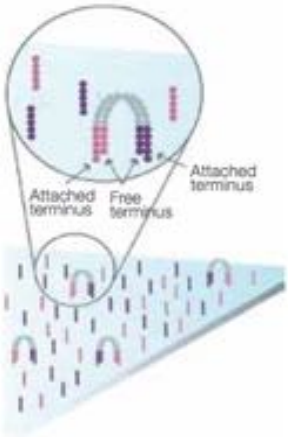
Massively Parallel Sequencing

Illumina HiSeq 2000
Sequencing by Synthesis

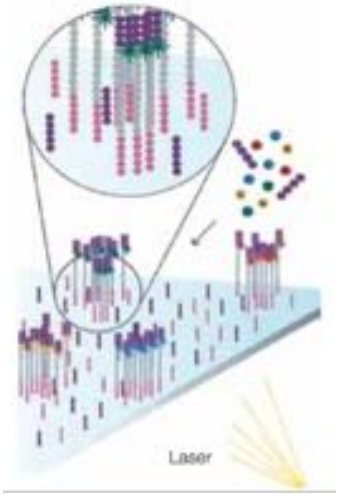
>60Gbp / day



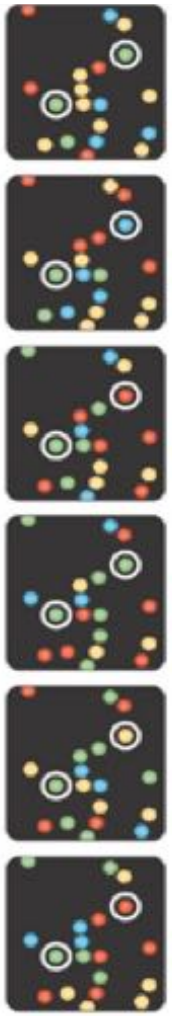
1. Attach



2. Amplify

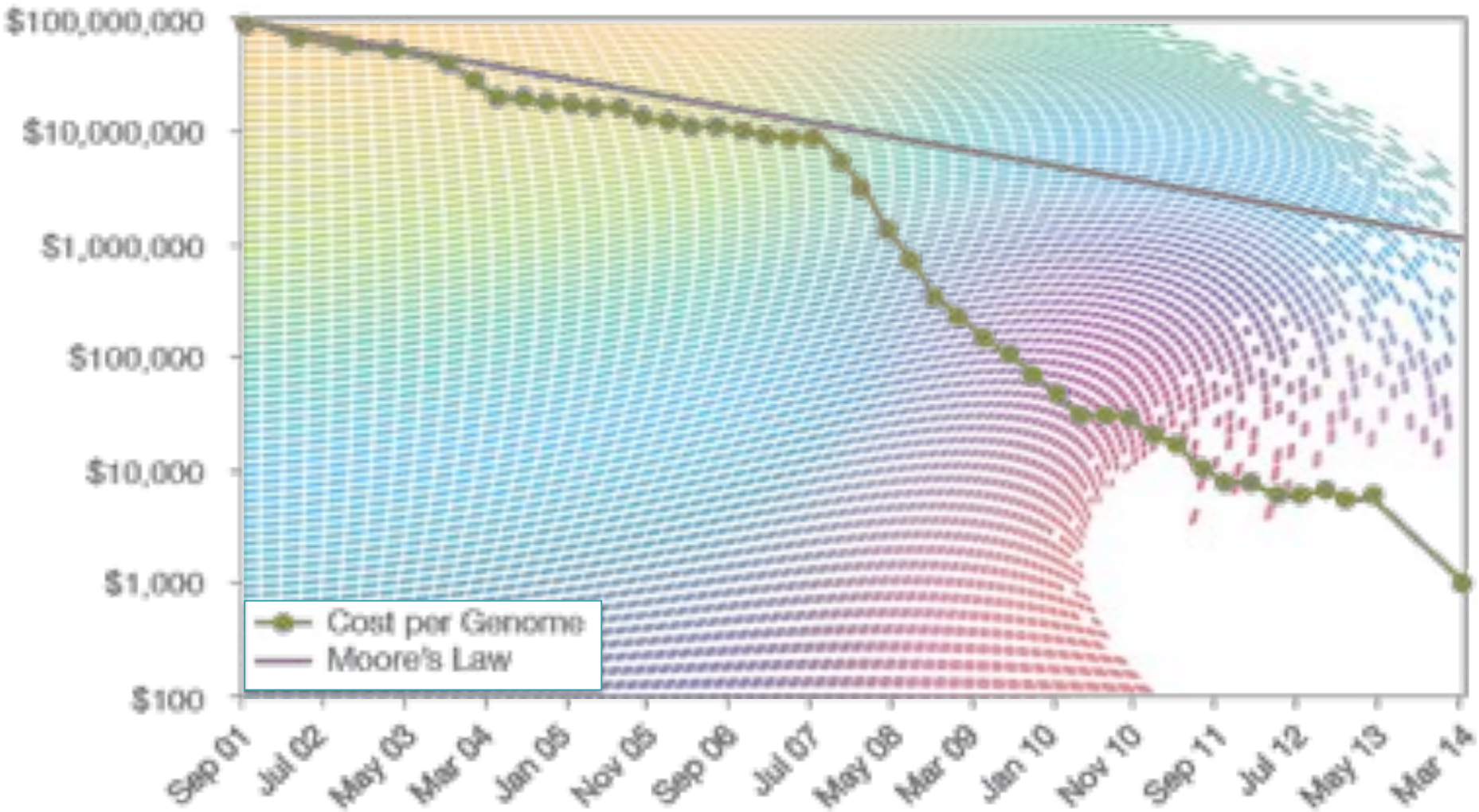


3. Image



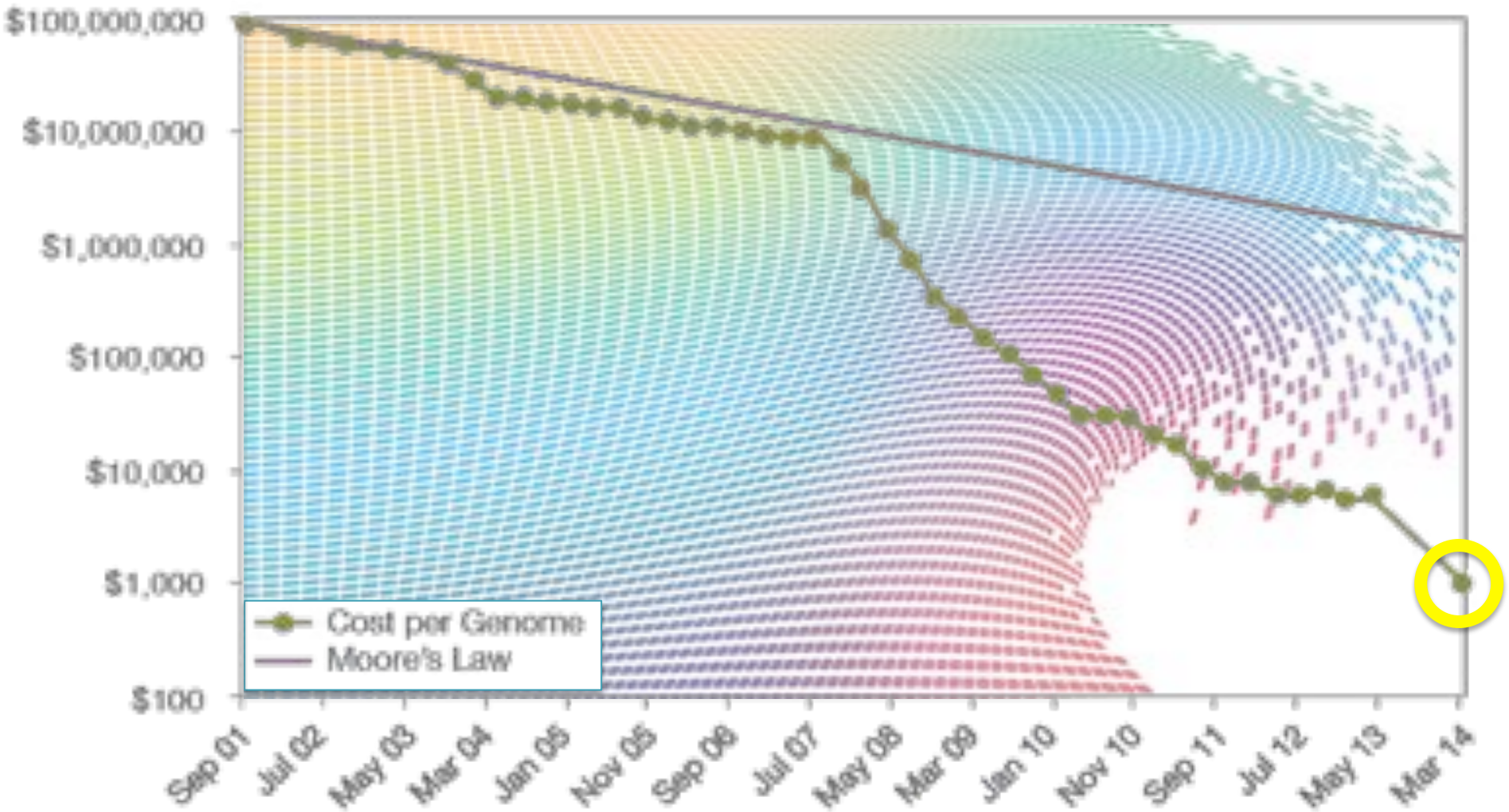
Metzker (2010) Nature Reviews Genetics 11:31-46
http://www.youtube.com/watch?v=l99aKKHcxC4

Cost per Genome



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

Cost per Genome



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

HiSeq X Ten

BROAD INSTITUTE

GARVAN INSTITUTE

NEW YORK GENOME CENTER

deCODE genetics

MACROGEN Axeq Technologies

Novogene 诺禾致源

药明康德 WuXi PharmaTech

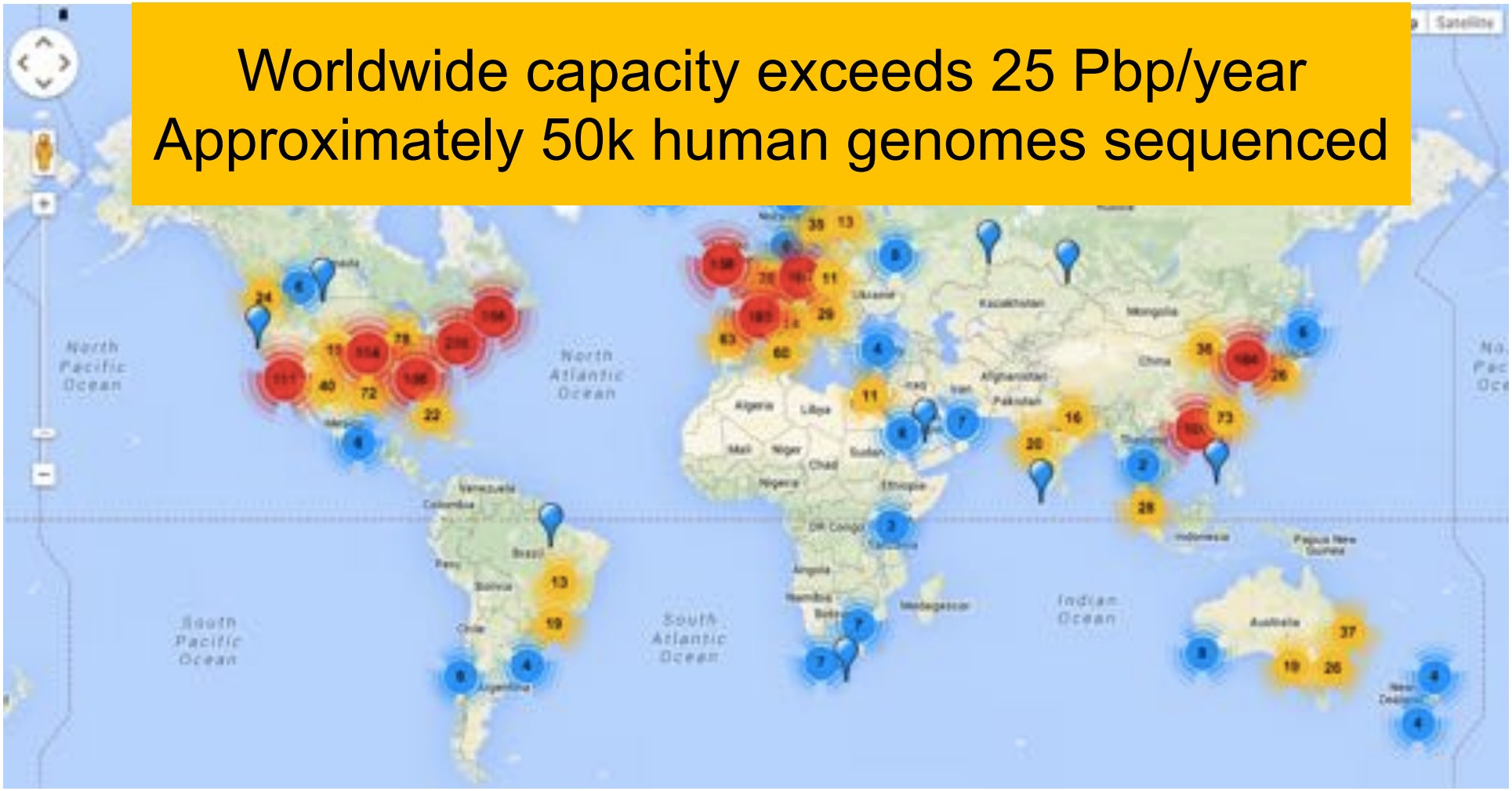
wellcome trust sanger institute

HUMAN LONGEVITY, INC.

320 genomes per week / 18,000 genomes per year
$1000 per genome / ~$10 M per instrument

Sequencing Centers

Worldwide capacity exceeds 25 Pbp/year
Approximately 50k human genomes sequenced



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

=

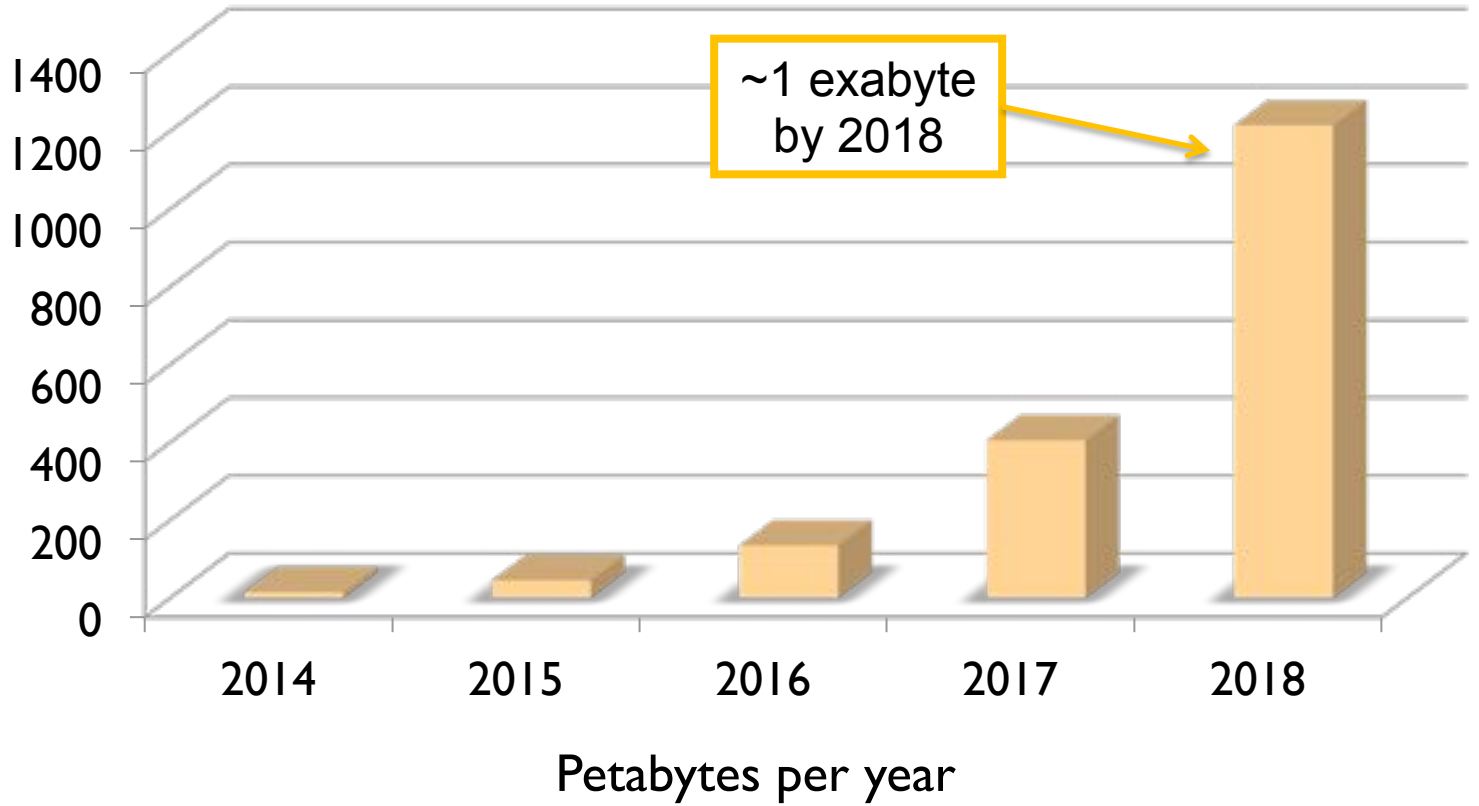
1PB Data
200,000 DVDs

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

500 2 TB drives
$500k

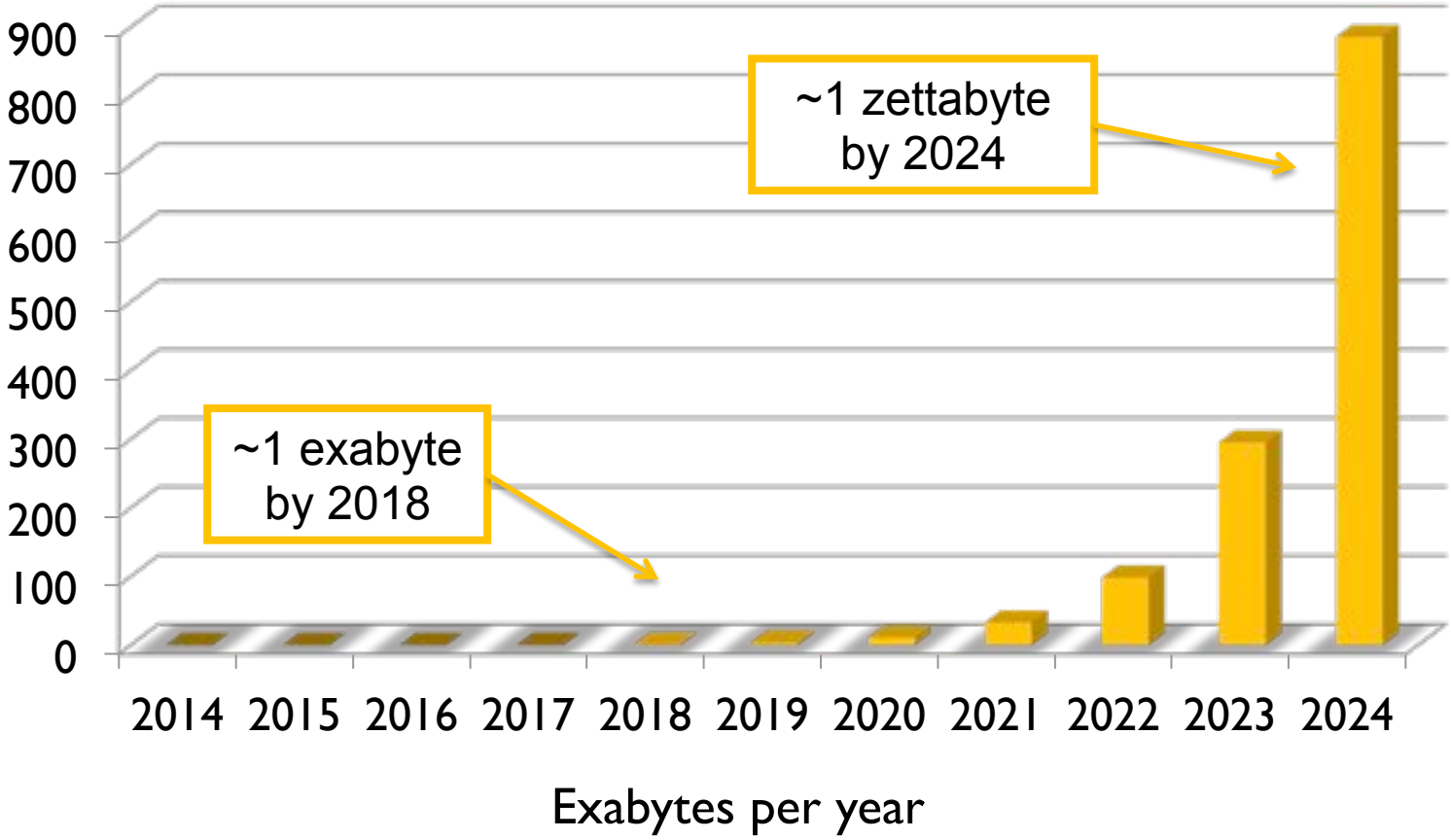
DNA Data Tsunami

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
And 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

Biological Sensor Network

@JasonWilliamsNY

Aspyn @ CSH High School

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

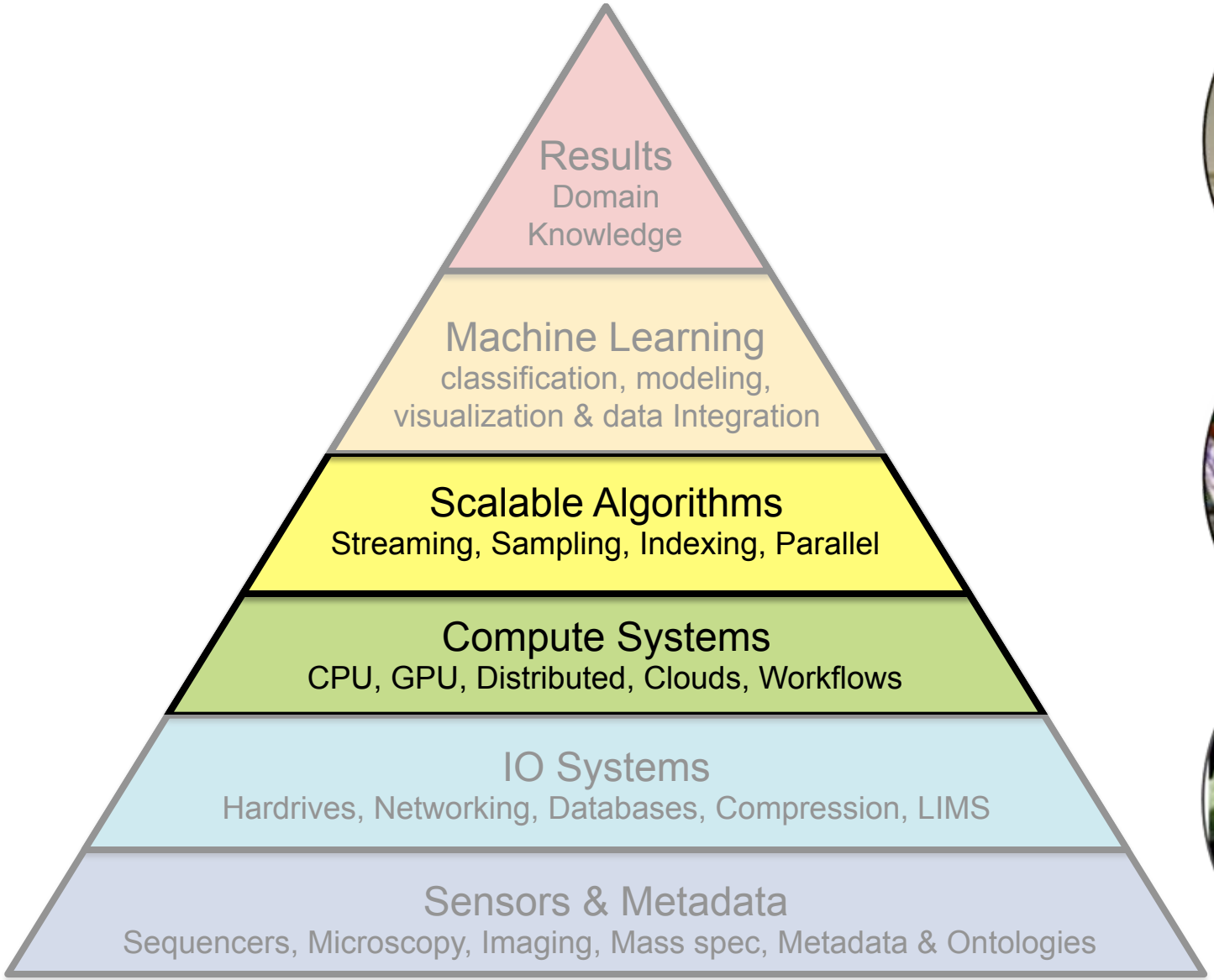
Major data producers concentrated in hospitals, universities, agricultural companies, research institutes

- Major efforts in human health and disease, agriculture, bioenergy
- Genomic information coupled with medical records and other medical data

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

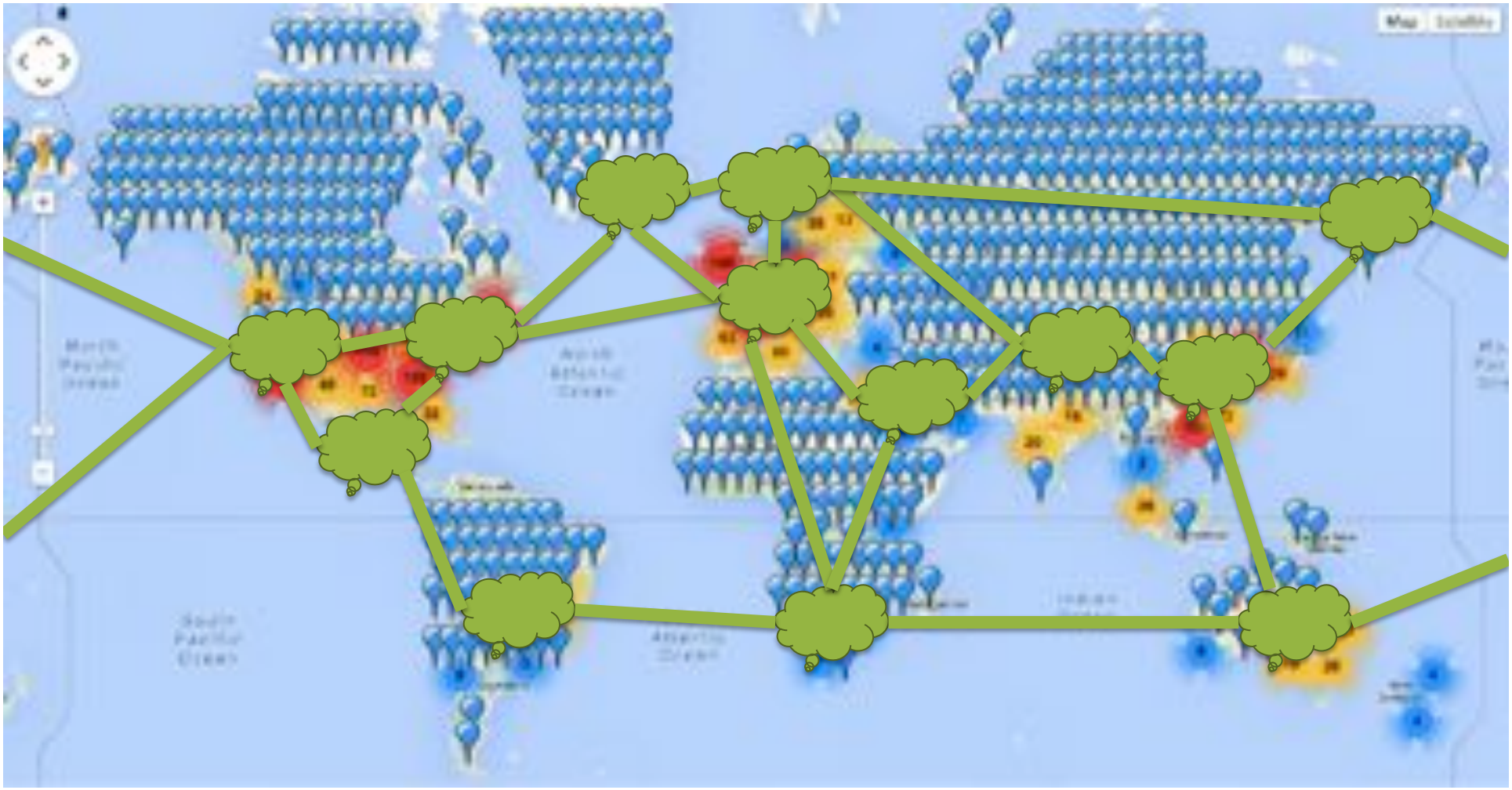


Sequencing Centers 2024

Informatics Centers 2024

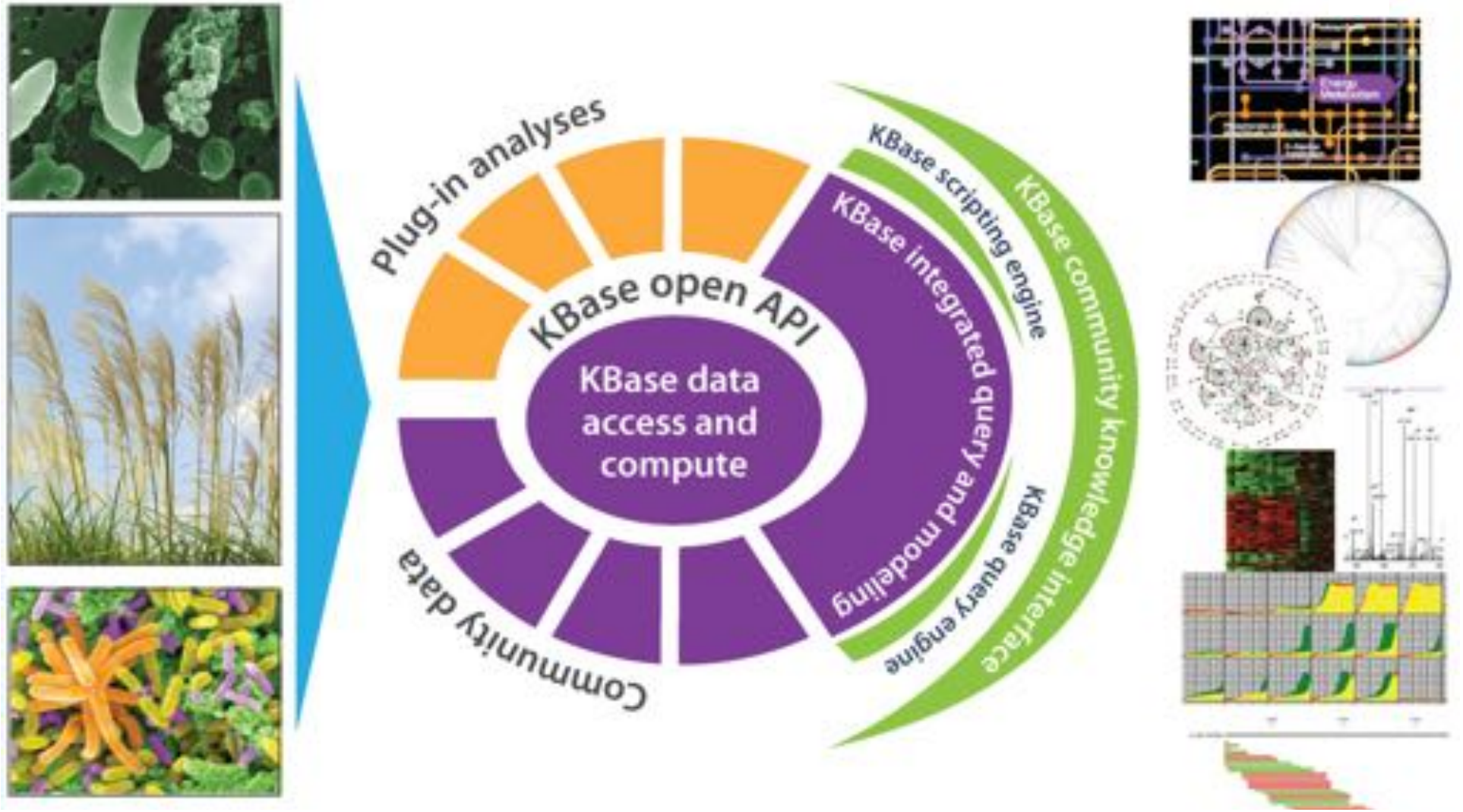
The DNA Data Deluge
Schatz, MC and Langmead, B (2013) *IEEE Spectrum*. July, 2013

Informatics Centers 2014



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

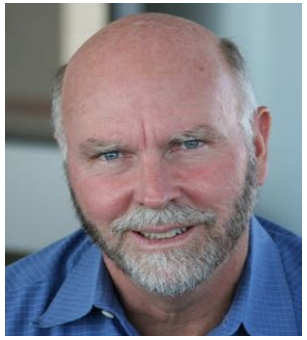
DOE Systems Biology Knowledgebase

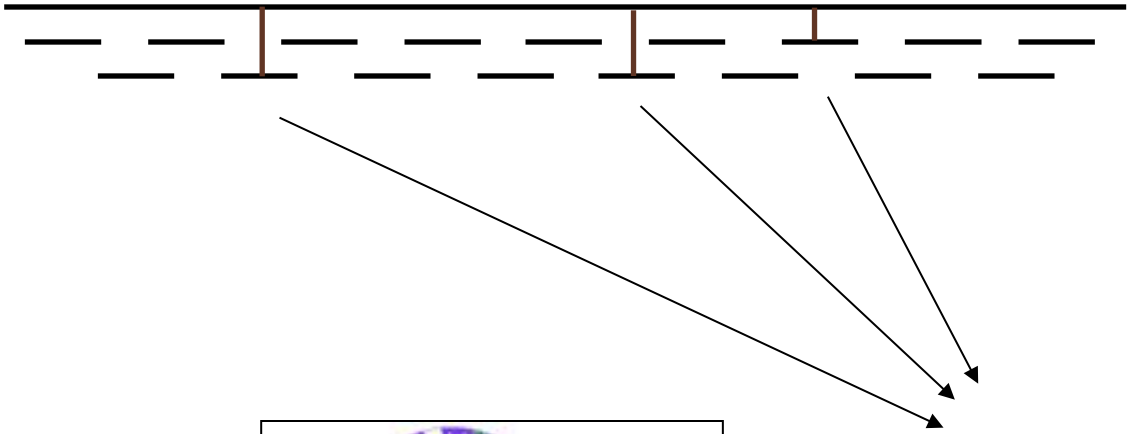


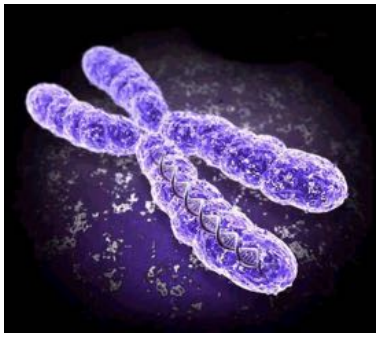
http://kbase.us: Predictive Biology in Microbes, Plants, and Meta-communities

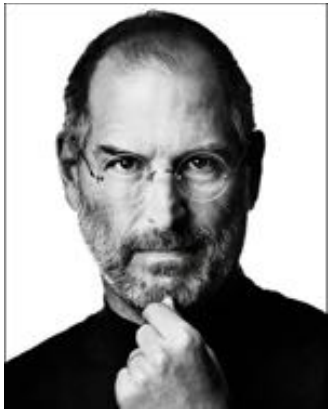
Personal Genomics

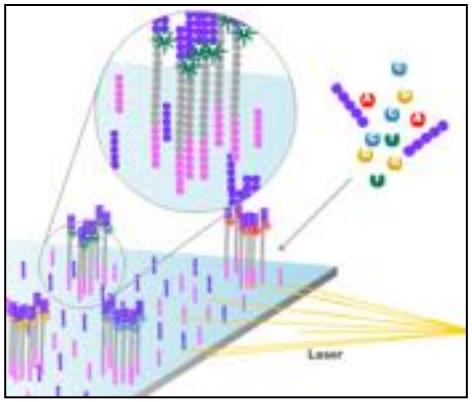
How does your genome compare to the reference?











Heart Disease

Cancer

Creates magical technology

Crossbow

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

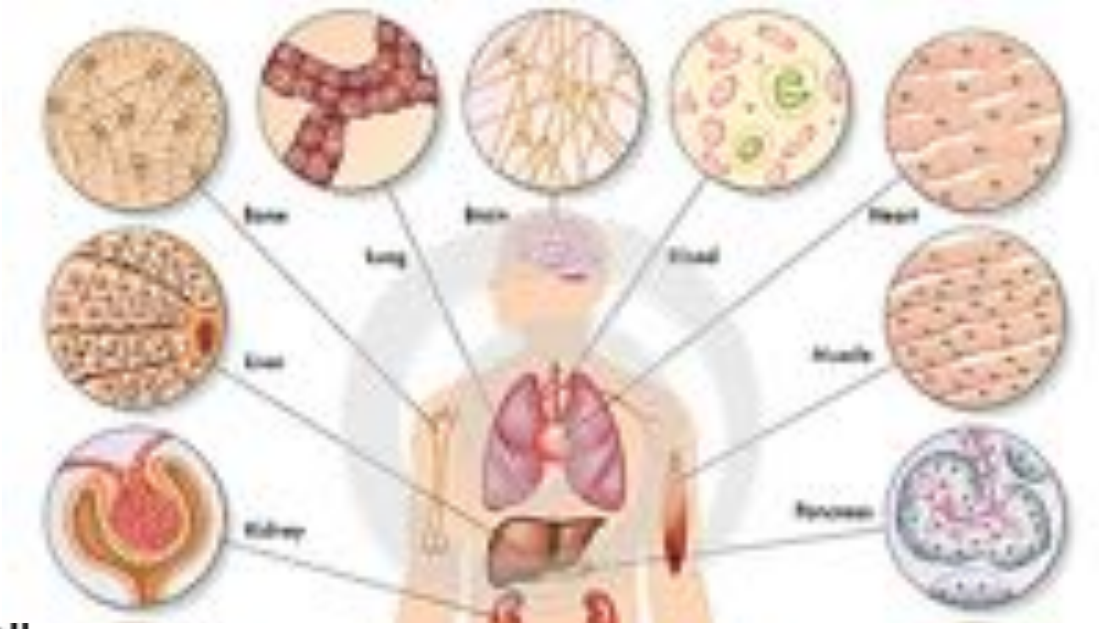
- Align billions of reads and find SNPs
  - Reuse software components: Hadoop Streaming
  - Mapping with Bowtie, SNP calling with SOAPsnp
- 4 hour end-to-end runtime including upload
  - Costs $85; Todays costs <$10

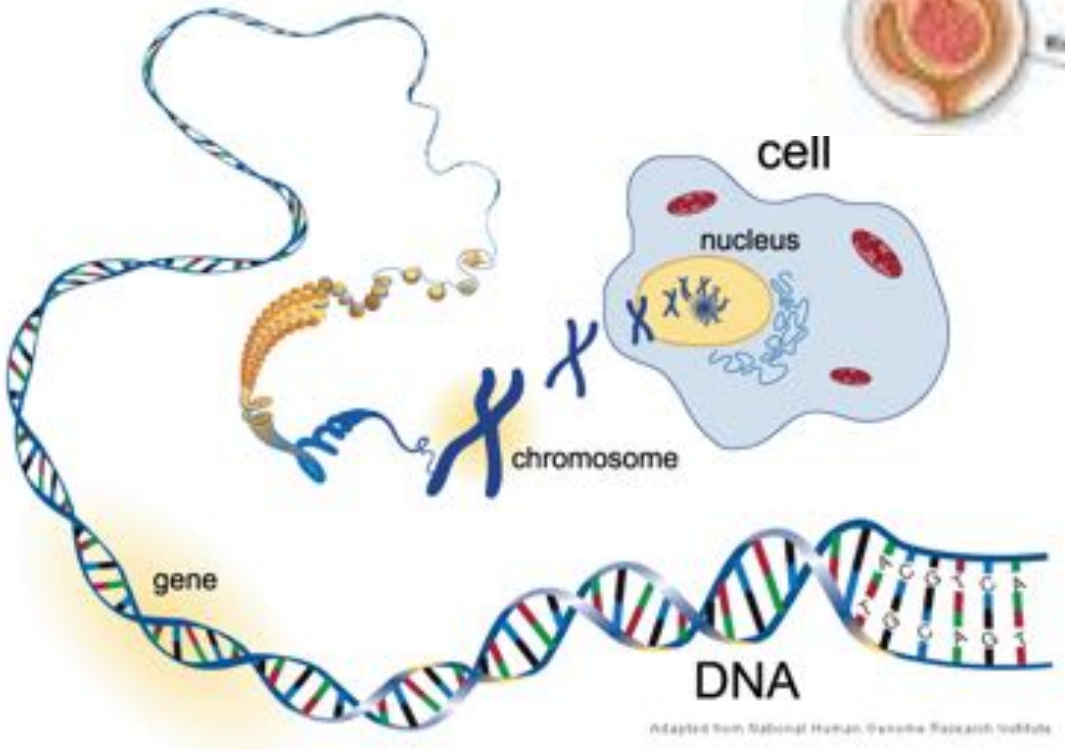
- Very compelling example of cloud computing in genomics
- Commercial vendors probably have better security than your institution
- Need more applications!

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

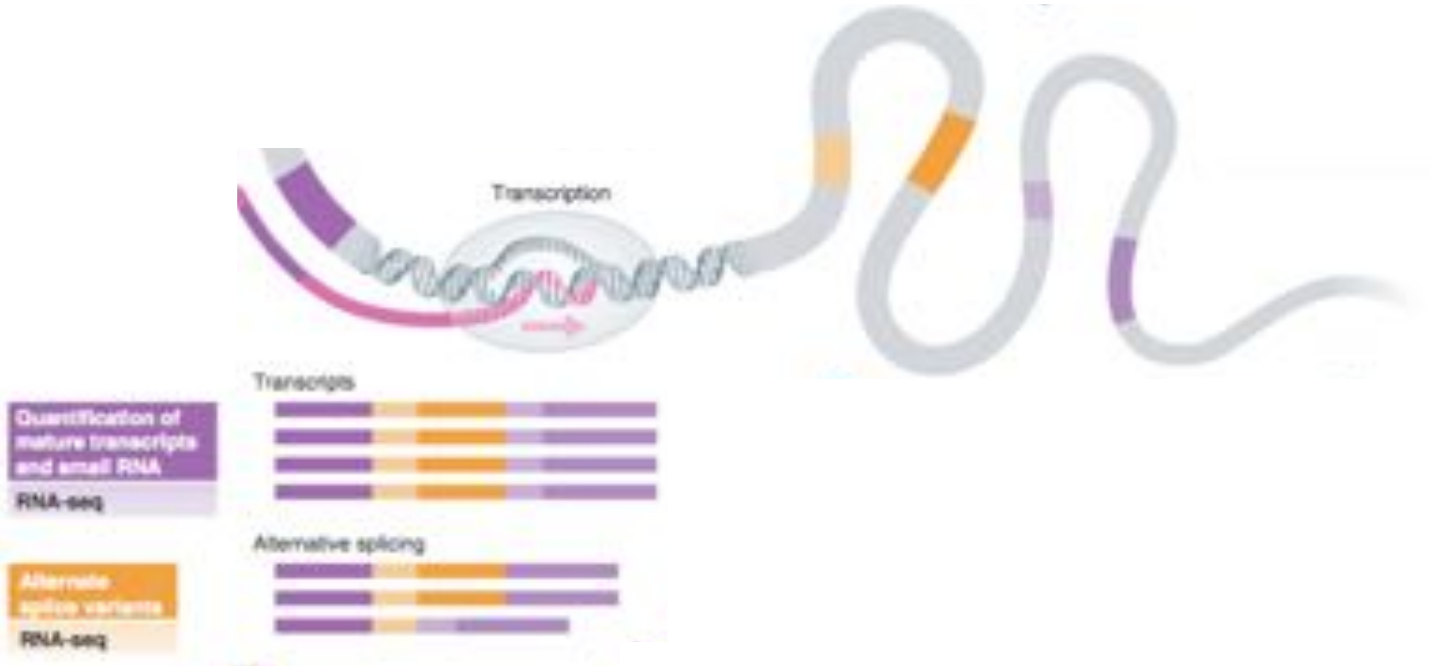
Cells & DNA

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



Soon et al., Molecular Systems Biology, 2013

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

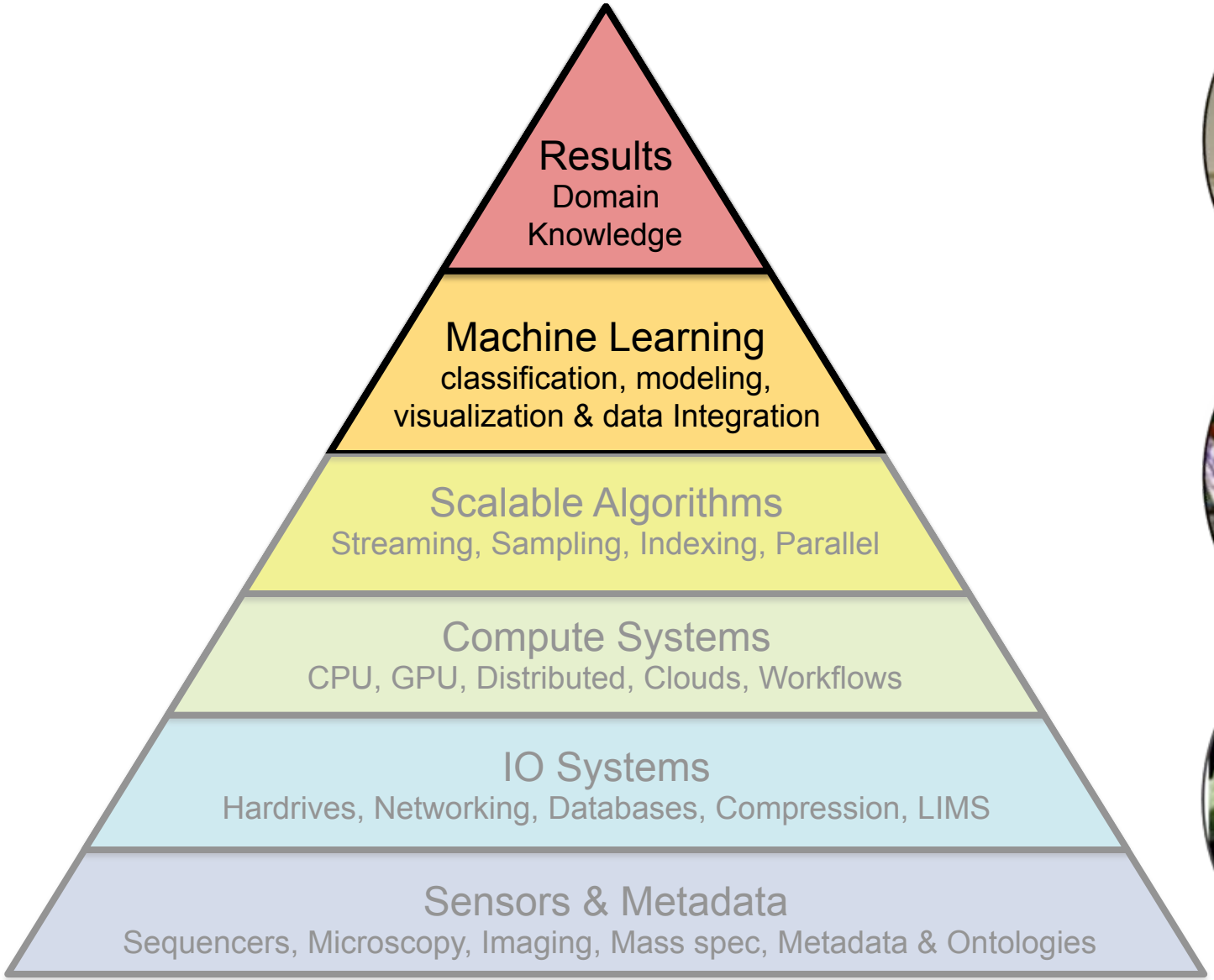
Parallel hardware and algorithms are required

- 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

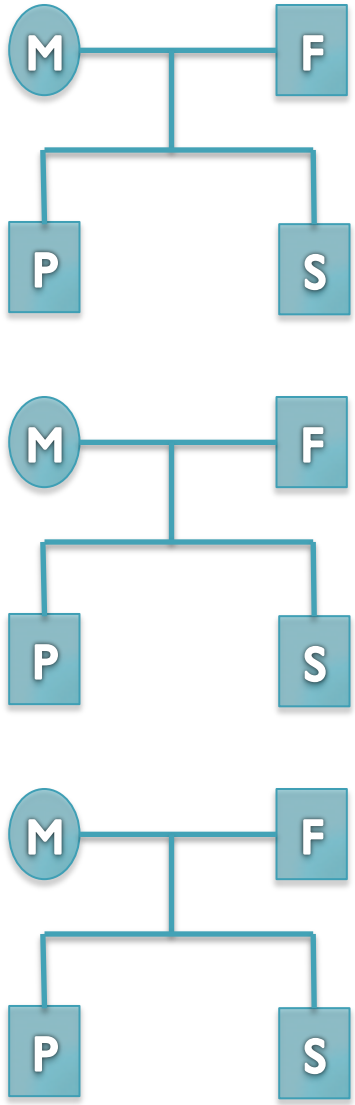
What is Autism?
http://www.autismspeaks.org/what-autism

Searching for the genetic risk factors

Search Strategy

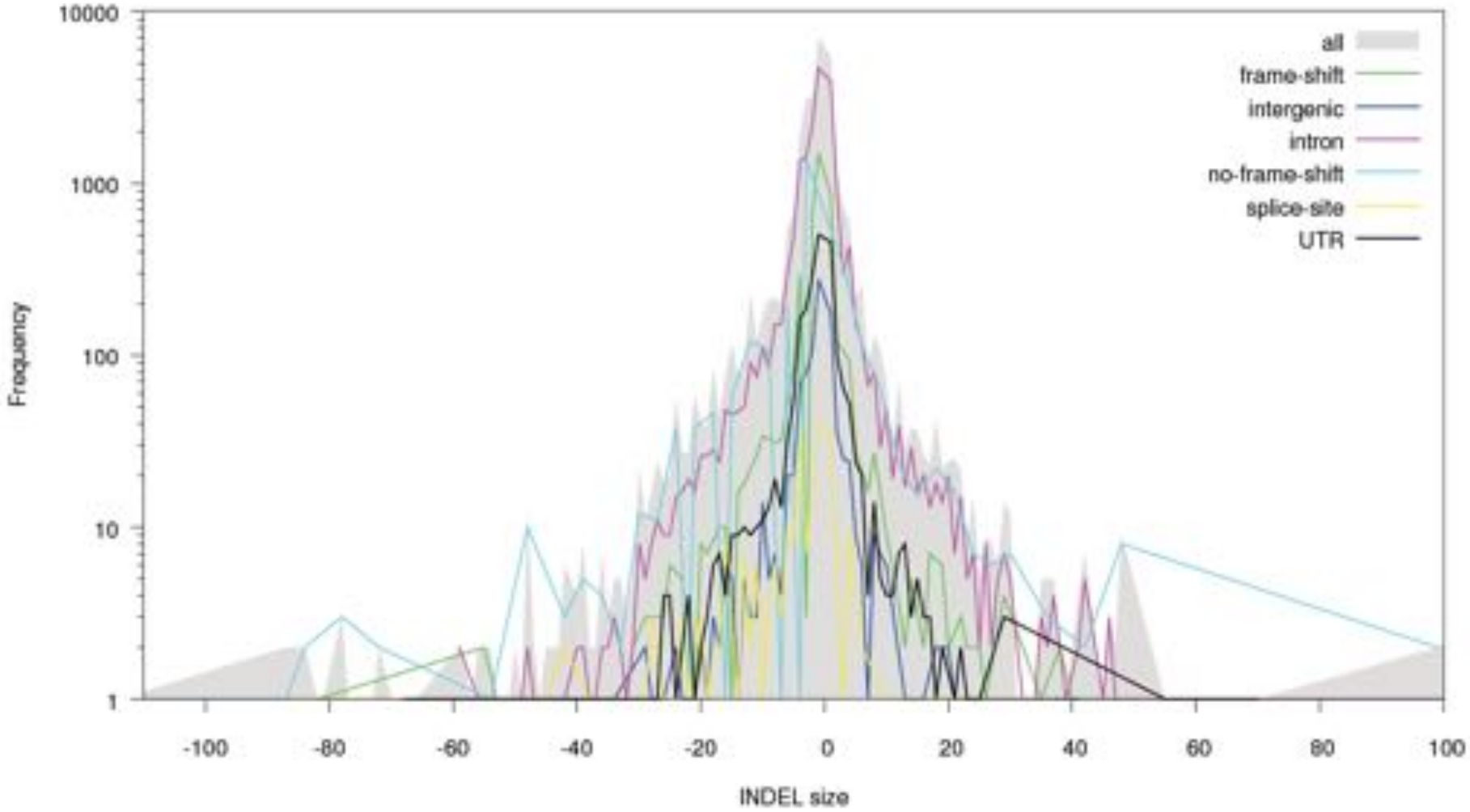
- Thousands of families identified from a dozen hospitals around the United States
- Large scale genome sequencing of "simplex" families: mother, father, affected child, unaffected sibling
- Unaffected siblings provide a natural control for environmental factors

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



SFARI SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

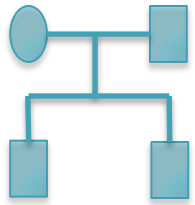
Population Analysis of the SSC



Constructed database of >1M transmitted and de novo genetic mutations

De novo mutation discovery and validation

De novo mutations:
Sequences not inherited from your parents.



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

Father(1):    ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Father(2):    ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...

Mother(1):    ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Mother(2):    ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...

Sibling(1):   ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Sibling(2):   ...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

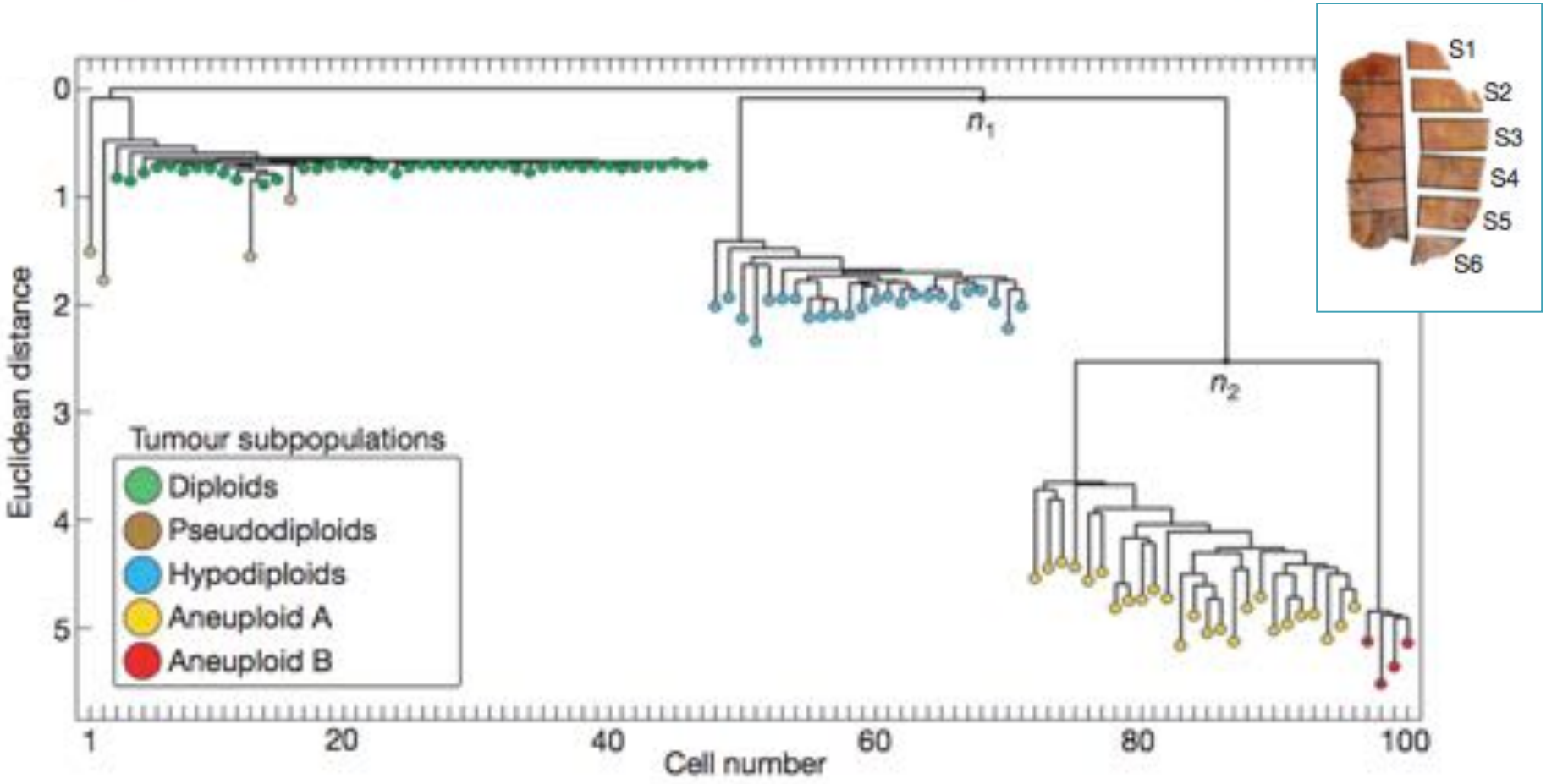
Accurate detection of de novo and transmitted INDELs within exome-capture data using micro-assembly
Narzisi, G, O'Rawe, J, Iossifov, I, Lee, Y, Wang, Z, Wu, Y, Lyon, G, Wigler, M, Schatz, MC (2014) *In press.*

LETTER

doi:10.1038/nature09807

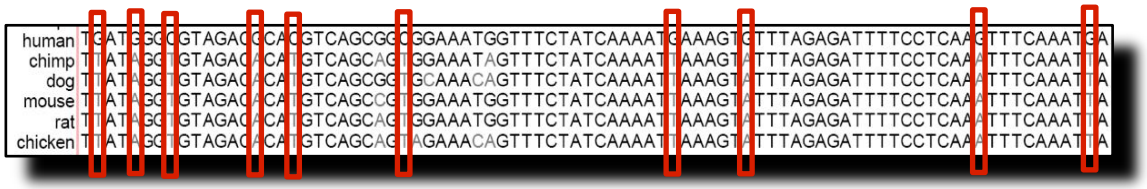
Tumour evolution inferred by single-cell sequencing

Nicholas Navin[1,2], Jude Kendall[1], Jennifer Troge[1], Peter Andrews[1], Linda Rodgers[1], Jeanne McIndoo[1], Kerry Cook[1], Asya Stepansky[1], Dan Levy[1], Diane Esposito[1], Lakshmi Muthuswamy[3], Alex Krasnitz[1], W. Richard McCombie[1], James Hicks[1] & Michael Wigler[1]



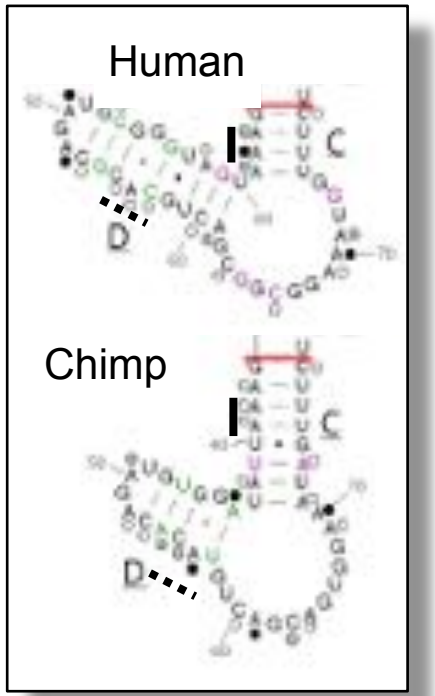
What makes us human?

"*Human Accelerated Regions*"



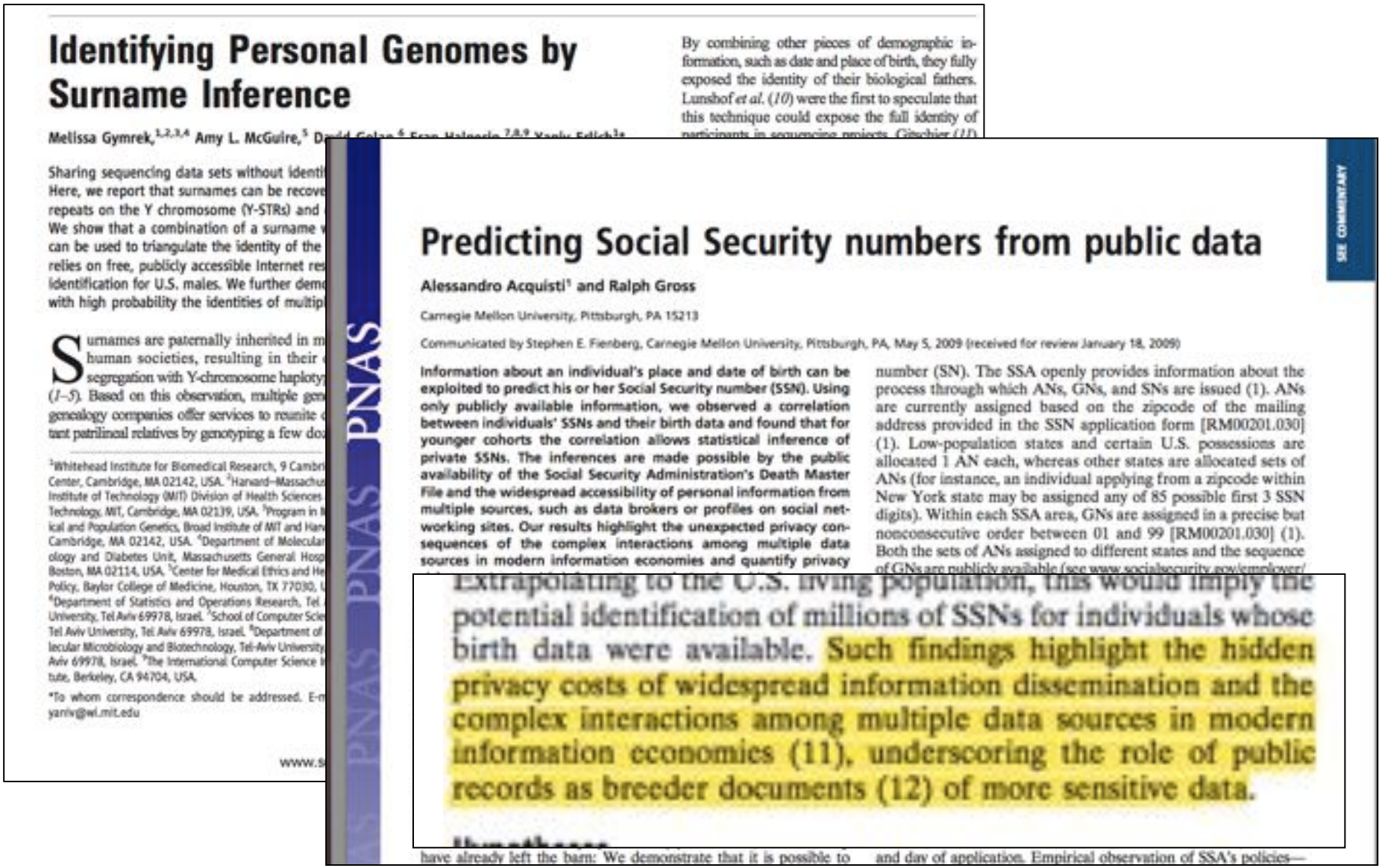
Systematic scan of recent human evolution identified the gene *HAR1F* as the most dramatic "human accelerated region".

Follow up analysis found it was specifically expressed in Cajal-Retzius neurons in the human brain from 6 to 19 gestational weeks.



(Pollard et al., *Nature*, 2006)

Genetic Privacy



Identifying Personal Genomes by Surname Inference

Melissa Gymrek,[1,2,3,4] Amy L. McGuire,[5] David Golan,[6] Eran Halperin,[7,8,9] Yaniv Erlich[3,*]

Sharing sequencing data sets without identi
Here, we report that surnames can be recove
repeats on the Y chromosome (Y-STRs) and
We show that a combination of a surname
can be used to triangulate the identity of the
relies on free, publicly accessible Internet re
identification for U.S. males. We further dem
with high probability the identities of multip

Surnames are paternally inherited in m
human societies, resulting in their
segregation with Y-chromosome haplotyp
(1–5). Based on this observation, multiple ge
genealogy companies offer services to reunite
tant patrilineal relatives by genotyping a few do

[1]Whitehead Institute for Biomedical Research, 9 Cambri
Center, Cambridge, MA 02142, USA. [2]Harvard–Massachu
Institute of Technology (MIT) Division of Health Sciences
Technology, MIT, Cambridge, MA 02139, USA. [3]Program in
ical and Population Genetics, Broad Institute of MIT and Har
Cambridge, MA 02142, USA. [4]Department of Molecular
ology and Diabetes Unit, Massachusetts General Hosp
Boston, MA 02114, USA. [5]Center for Medical Ethics and He
Policy, Baylor College of Medicine, Houston, TX 77030, U
[6]Department of Statistics and Operations Research, Tel
University, Tel Aviv 69978, Israel. [7]School of Computer Scie
Tel Aviv University, Tel Aviv 69978, Israel. [8]Department of
lecular Microbiology and Biotechnology, Tel-Aviv University
Aviv 69978, Israel. [9]The International Computer Science
tute, Berkeley, CA 94704, USA.

*To whom correspondence should be addressed. E-m
yaniv@wi.mit.edu

www.s

By combining other pieces of demographic information, such as date and place of birth, they fully exposed the identity of their biological fathers. Lunshof *et al.* (*10*) were the first to speculate that this technique could expose the full identity of participants in sequencing projects. Gitschier (*11*)

PNAS PNAS PNAS

SEE COMMENTARY

Predicting Social Security numbers from public data

Alessandro Acquisti[1] and Ralph Gross

Carnegie Mellon University, Pittsburgh, PA 15213

Communicated by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, May 5, 2009 (received for review January 18, 2009)

Information about an individual's place and date of birth can be exploited to predict his or her Social Security number (SSN). Using only publicly available information, we observed a correlation between individuals' SSNs and their birth data and found that for younger cohorts the correlation allows statistical inference of private SSNs. The inferences are made possible by the public availability of the Social Security Administration's Death Master File and the widespread accessibility of personal information from multiple sources, such as data brokers or profiles on social networking sites. Our results highlight the unexpected privacy consequences of the complex interactions among multiple data sources in modern information economies and quantify privacy

number (SN). The SSA openly provides information about the process through which ANs, GNs, and SNs are issued (1). ANs are currently assigned based on the zipcode of the mailing address provided in the SSN application form [RM00201.030] (1). Low-population states and certain U.S. possessions are allocated 1 AN each, whereas other states are allocated sets of ANs (for instance, an individual applying from a zipcode within New York state may be assigned any of 85 possible first 3 SSN digits). Within each SSA area, GNs are assigned in a precise but nonconsecutive order between 01 and 99 [RM00201.030] (1). Both the sets of ANs assigned to different states and the sequence of GNs are publicly available (see www.socialsecurity.gov/employer/

Extrapolating to the U.S. living population, this would imply the potential identification of millions of SSNs for individuals whose birth data were available. Such findings highlight the hidden privacy costs of widespread information dissemination and the complex interactions among multiple data sources in modern information economies (11), underscoring the role of public records as breeder documents (12) of more sensitive data.

have already left the barn: We demonstrate that it is possible to

and day of application. Empirical observation of SSA's policies—

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

How can you participate?

Students

- Learn python!
- Study math & statistics & computer science
- Visit the DNA Learning Center

Individuals

- Personal Genome Project
  Harvard Medical School
  http://www.personalgenomes.org

- 23andMe
  Genetic testing and ancestry
  http://www.23andme.com

- CSHL Public Lectures & Events
  http://www.cshl.edu

Acknowledgements

Schatz Lab

Giuseppe Narzisi
Shoshana Marcus
James Gurtowski
Srividya Ramakrishnan
Hayan Lee
Rob Aboukhalil
Mitch Bekritsky
Charles Underwood
Tyler Gavin
Maria Nattestad
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