# Applied Comparative Genomics

Michael Schatz

January 31, 2017

Lecture 1: Course Overview

# Welcome!

***The primary goal of the course is for students to be grounded in theory and leave the course empowered to conduct independent genomic analyses.***

- We will study the leading computational and quantitative approaches for comparing and analyzing genomes starting from raw sequencing data.
- The course will focus on human genomics and human medical applications, but the techniques will be broadly applicable across the tree of life.
- The topics will include genome assembly & comparative genomics, variant identification & analysis, gene expression & regulation, personal genome analysis, and cancer genomics.

| | |
|---|---|
| ***Course Webpage:*** | https://github.com/schatzlab/appliedgenomics |
| ***Course Discussions:*** | http://piazza.com |
| ***Class Hours:*** | Tues + Thurs @ 1:30p – 2:45p, Shaffer 304 |
| ***Office Hours:*** | Tues + Thurs @ 3-4p and by appointment<br>Please try Piazza first! |

# Prerequisites and Resources

***Prerequisites***

- No formal course requirements
- Access to an Apple or Linux Machine, or Install VirtualBox
- Familiarity with the Unix command line for exercises
  - bash, ls, grep, sed, + install published genomics tools
- Familiarity with a major programming language for project
  - C/C++, Java, R, Perl, Python

***Primary Texts***

- None! We will be studying primary research papers

***Other Resources:***

- Google, SEQanswers, Biostars, StackOverflow

- Applied Computational Genomics Course at UU: Spring 2017
- https://github.com/quinlan-lab/applied-computational-genomics

- Ben Langmead's teaching materials:
- http://www.langmead-lab.org/teaching-materials/

# Grading Policies

***Assessments:***

- ~5 Assignments: 30% Due at 11:59pm a week later
  ***Practice using the tools we are discussing***

- 1 Exam: 30% In class (Tentatively 4/6)
  ***Assess your performance, focusing on the methods***

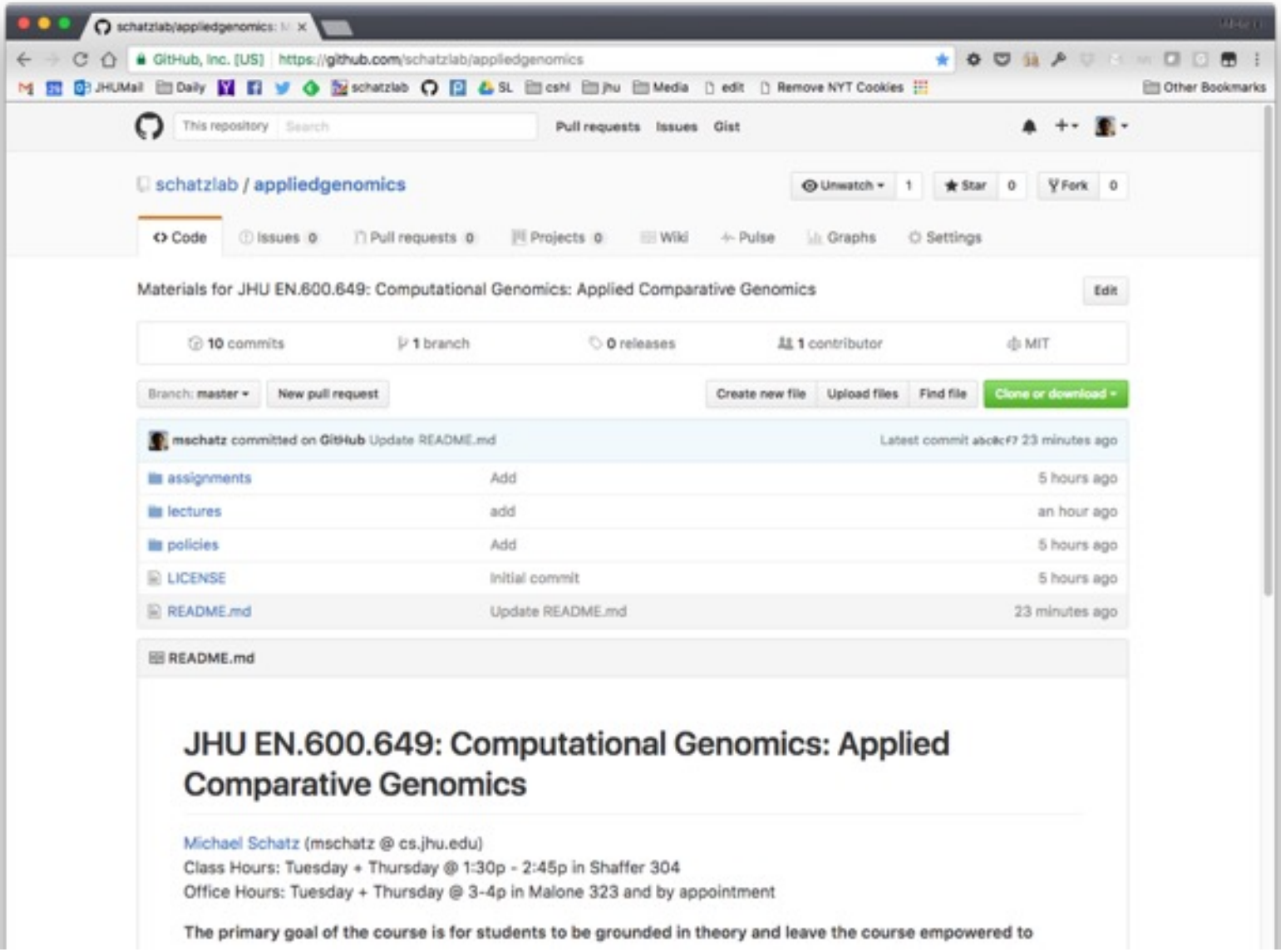
- 1 Class Project: 40% Presented last week of class
  ***Significant project developing a novel analysis/method***

- In-class Participation: Not graded, but there to help you!

***Policies:***

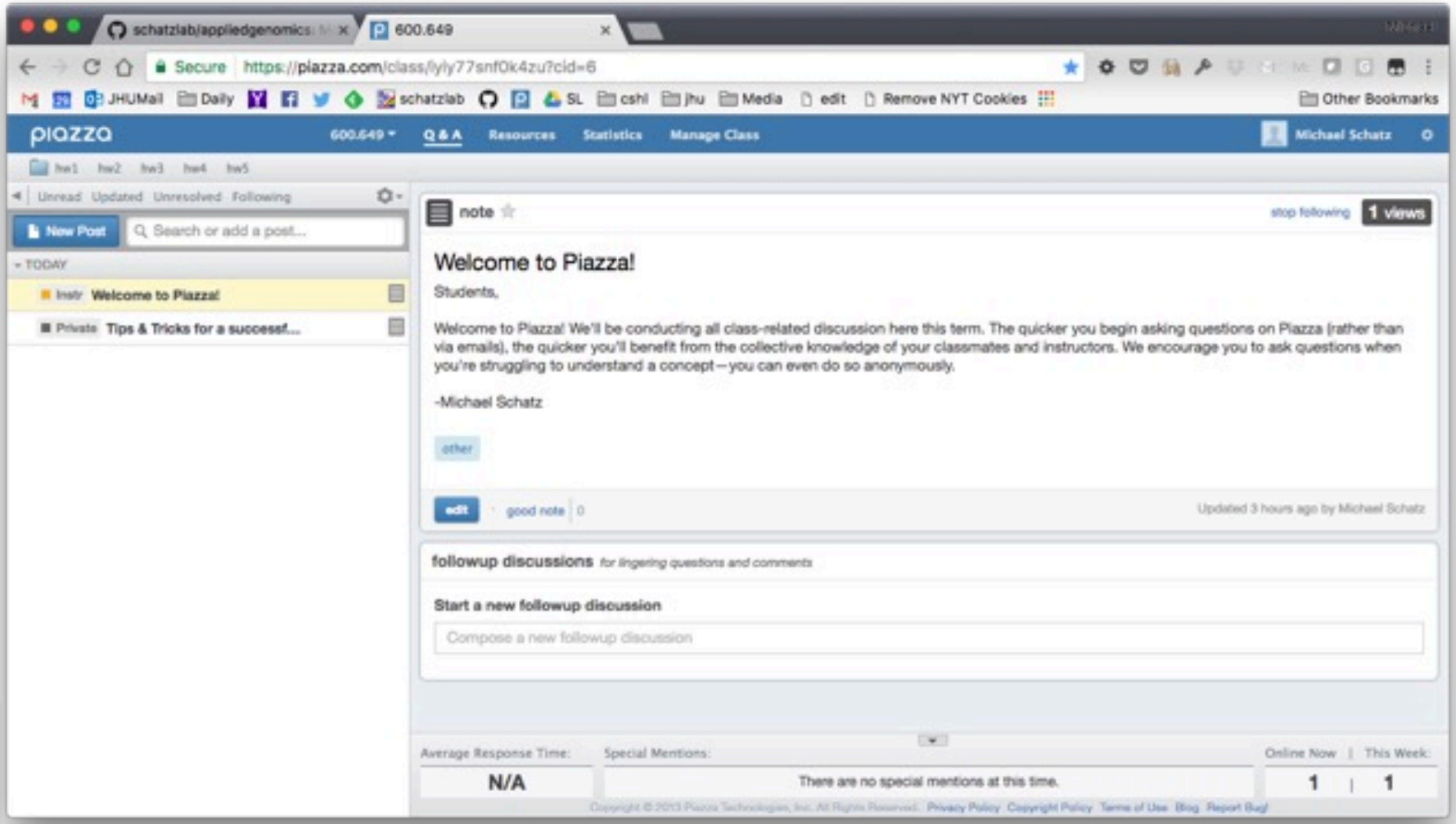
- Scores assigned relative to the highest points awarded
- Automated testing and grading of assignments
- ***Late Days:***
  - Three (3) chances to extend the deadline for assignments by 24 hours without any penalty

# Course Webpage



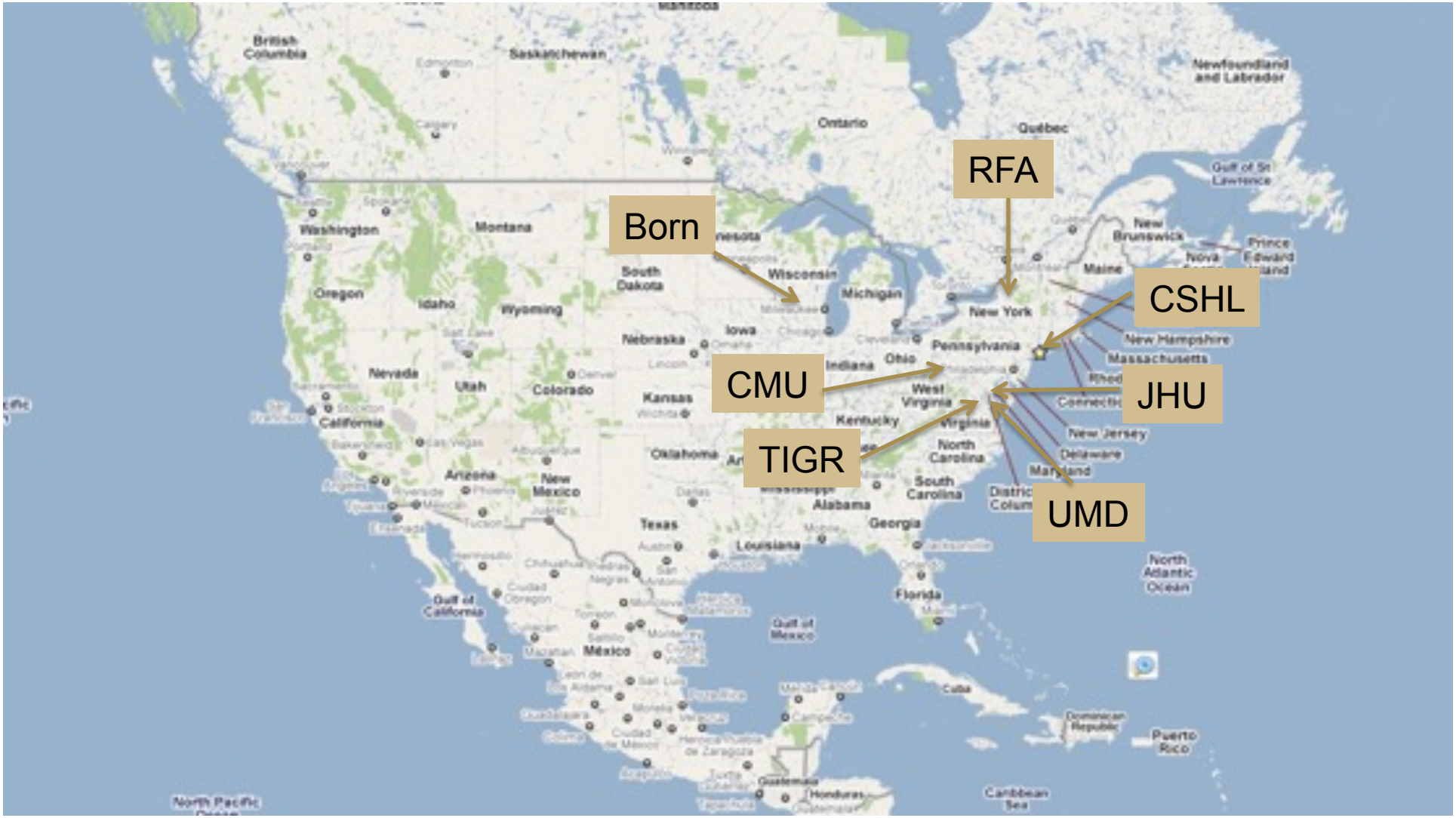

https://github.com/schatzlab/appliedgenomics

# Piazza




https://piazza.com/jhu/spring2017/600649/home

# A Little About Me



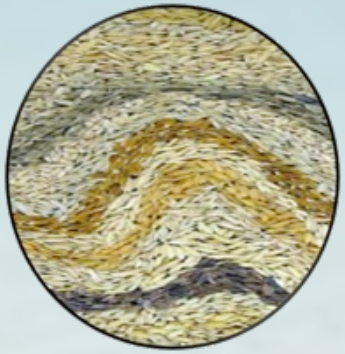
# A Little About Me

# Schatzlab Overview

## Human Genetics

Role of mutations in disease

Fang *et al.* (2016)
Narzisi *et al.* (2015)



## Agricultural Genomics

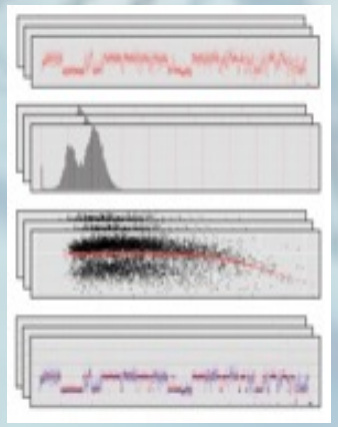
Genomes & Transcriptomes

Lemmon *et al.* (2016)
Ming *et al.* (2015)

## Algorithmics & Systems Research

Ultra-large scale biocomputing

Stevens *et al.* (2015)
Marcus *et al.* (2014)



## Biotechnology Development

Single Cell + Single Molecule Sequencing

Chin *et al.* (2016)
Garvin *et al.* (2015)

# DNA: The secret of life

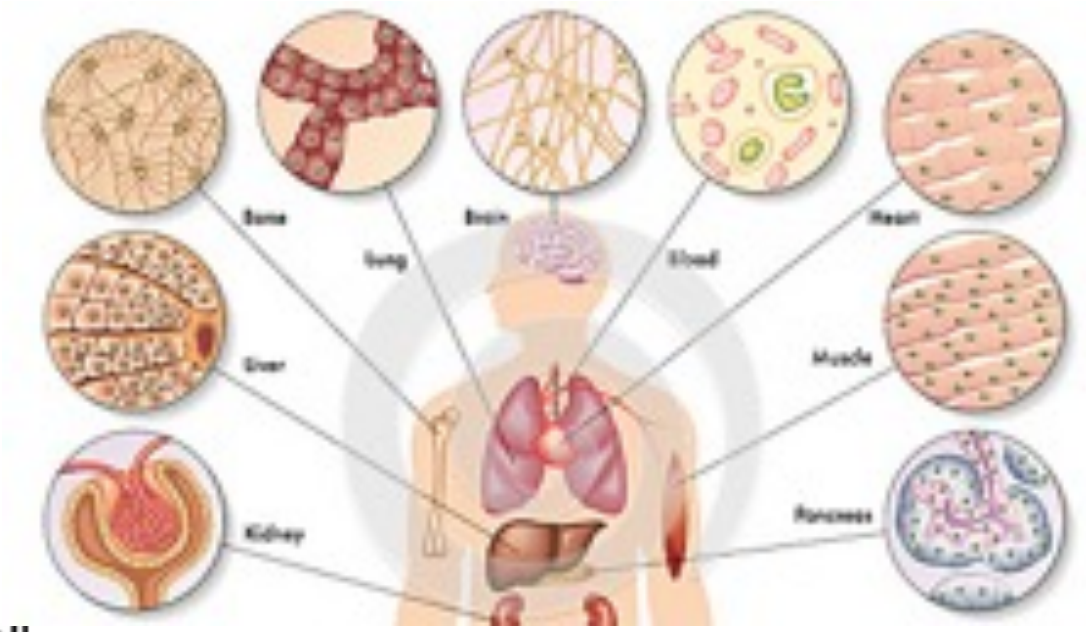
***Your DNA, along with your environment and experiences, shapes who you are***

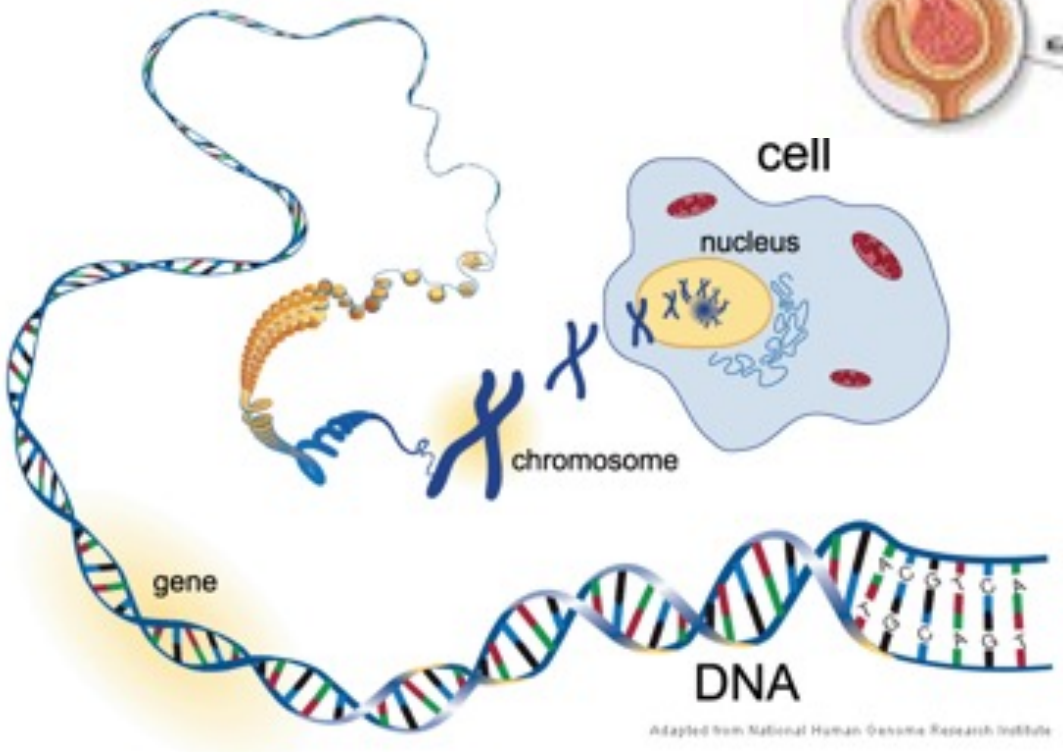
- Height
- Hair, eye, skin color
- Broad/narrow, small/large features
- Susceptibility to disease
- Response to drug treatments
- Longevity and cognition

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

# Cells & DNA

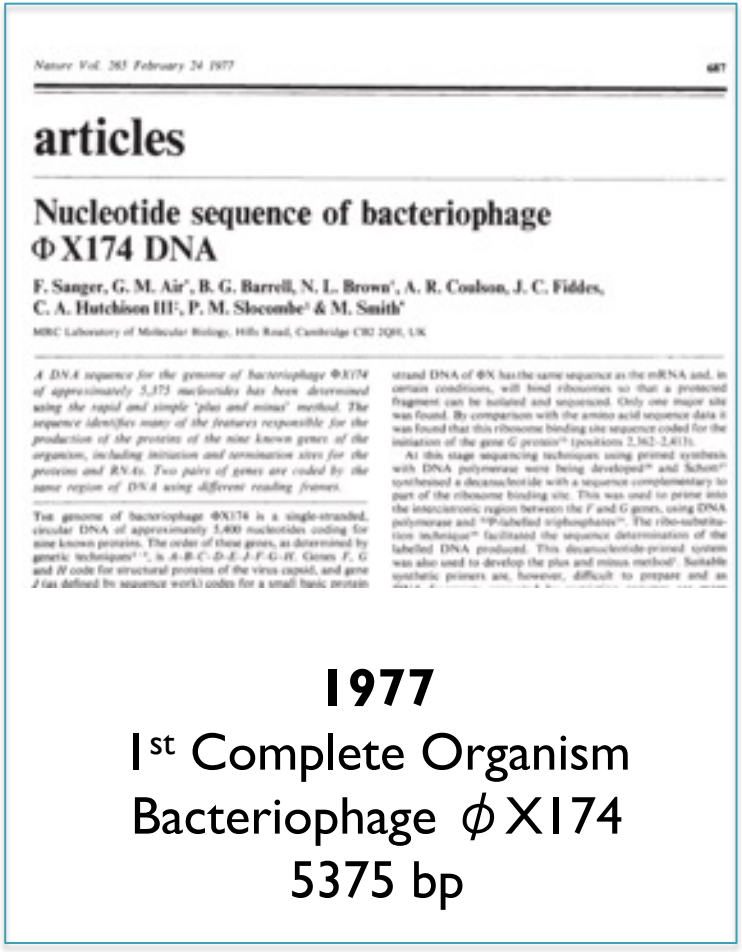
Each cell of your body contains an exact copy of your 3 billion base pair genome.






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).

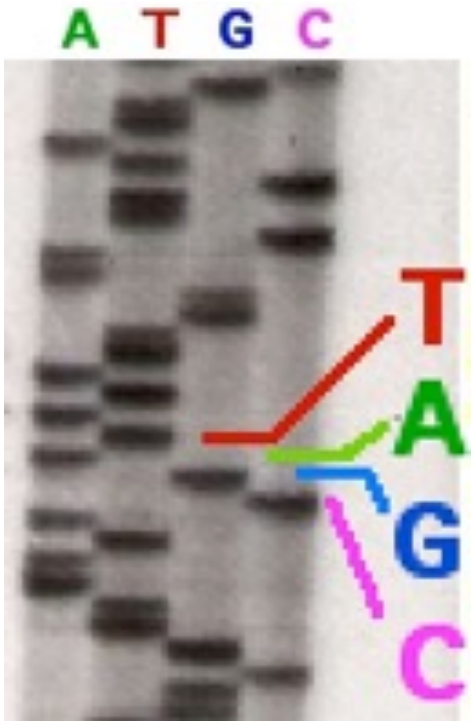
At this stage sequencing techniques using primed synthesis with DNA polymerase were being developed[5,6] and Schott[7] synthesized a decanucleotide with a sequence complementary to part of the ribosome binding site. This was used to prime into the intercistronic region between the J and G genes, using DNA polymerase and $^{32}$P-labelled triphosphates[8,9]. 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[1]. Suitable synthetic primers are, however, difficult to prepare and as

**1977**

1st Complete Organism

Bacteriophage $\phi$X174

5375 bp




Radioactive Chain Termination

5000bp / week / person

http://en.wikipedia.org/wiki/File:Sequencing.jpg

http://www.answers.com/topic/automated-sequencer

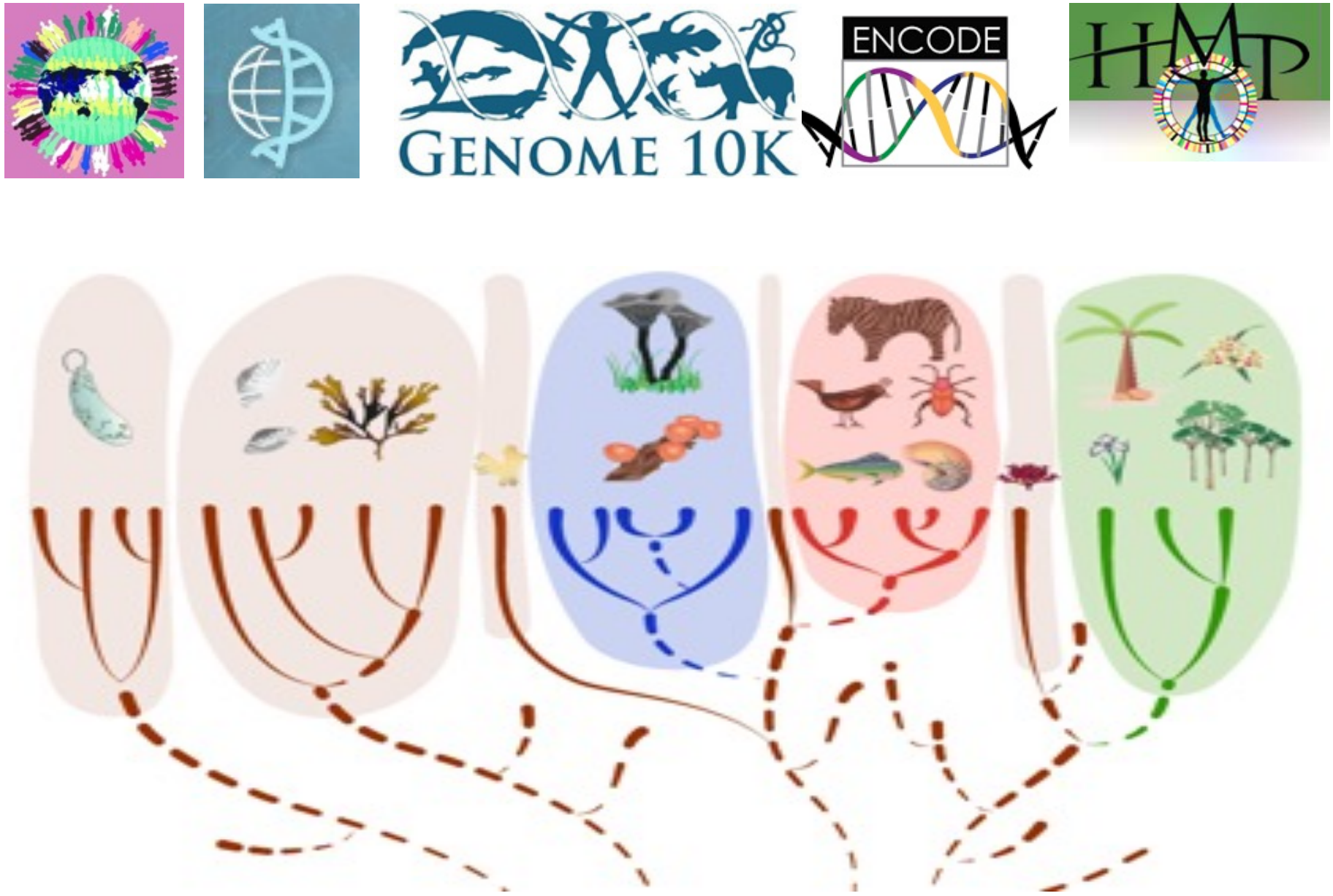
***Nucleotide sequence of bacteriophage $\phi$X174 DNA***

Sanger, F. et al. (1977) *Nature. 265: 687 - 695*

# Milestones in DNA Sequencing


(TIGR/Celera, 1995-2001)

# Genomics across the tree of life



# Unsolved Questions in Biology

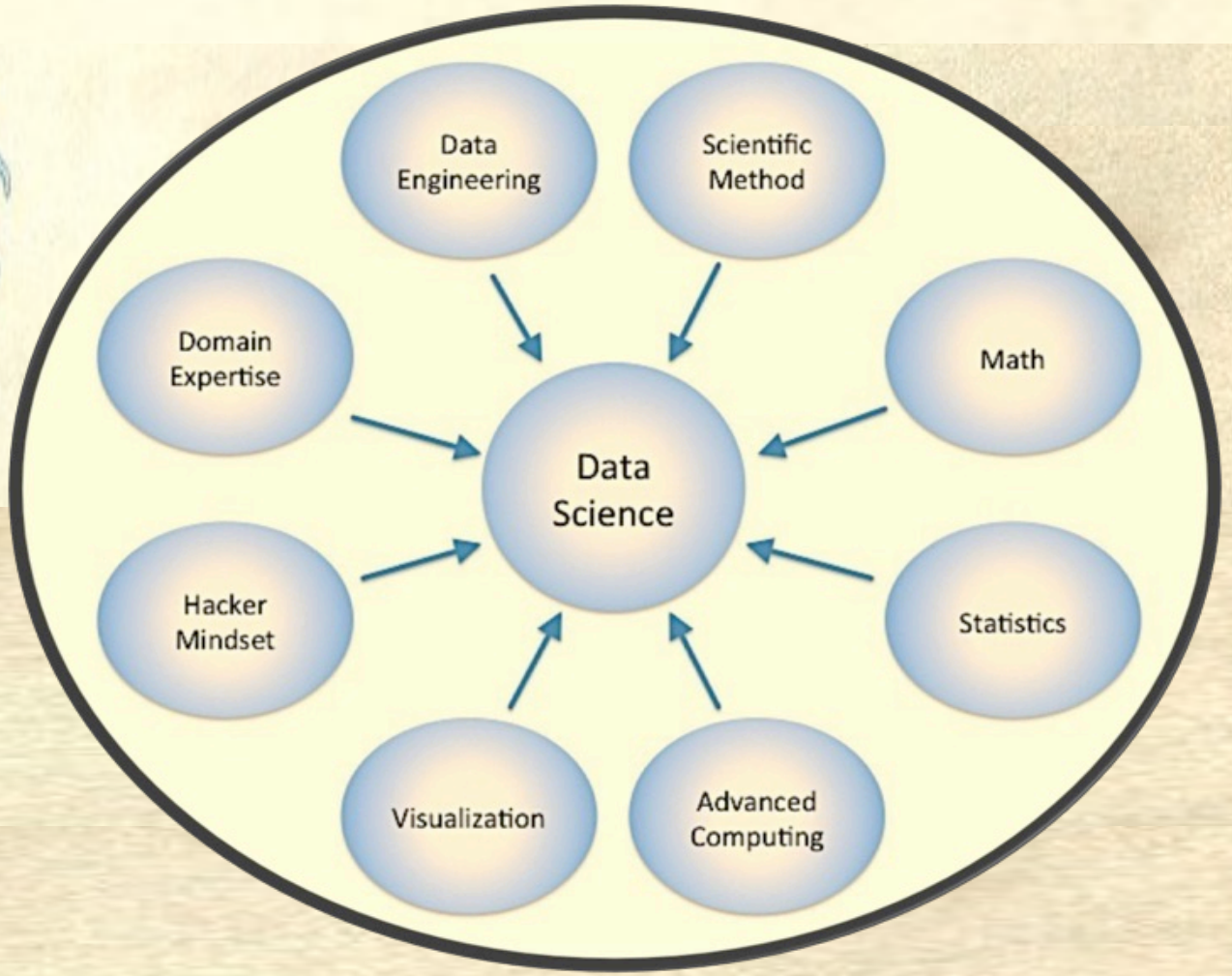
- What is your genome sequence?
- 
- 
- 
- 
- 
- 
- 
- 
- 
- 
- 
- 
- ***Plus thousands and thousands more***

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

***What software and systems will?***

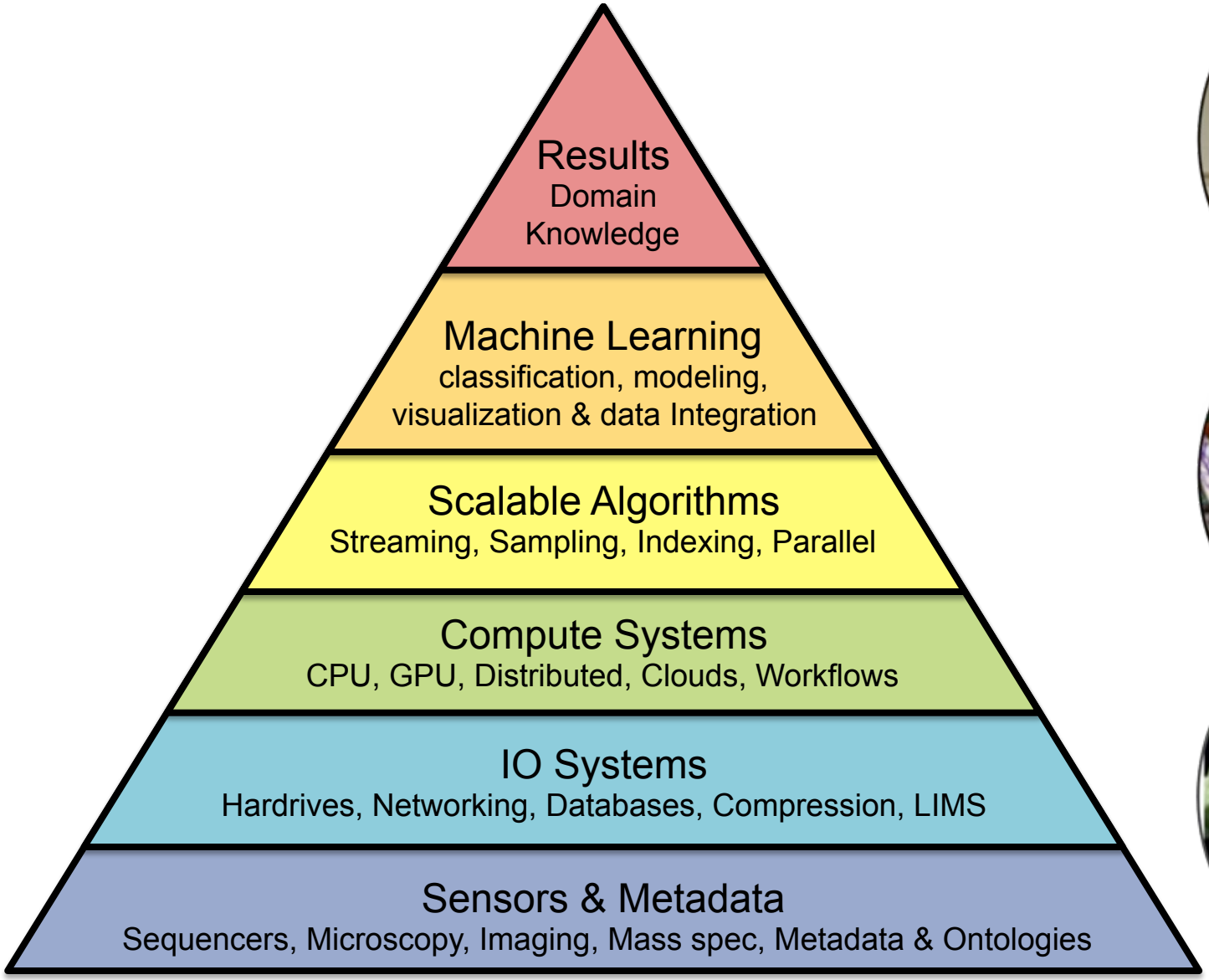
***And who will create them?***

# Who is a Data Scientist?

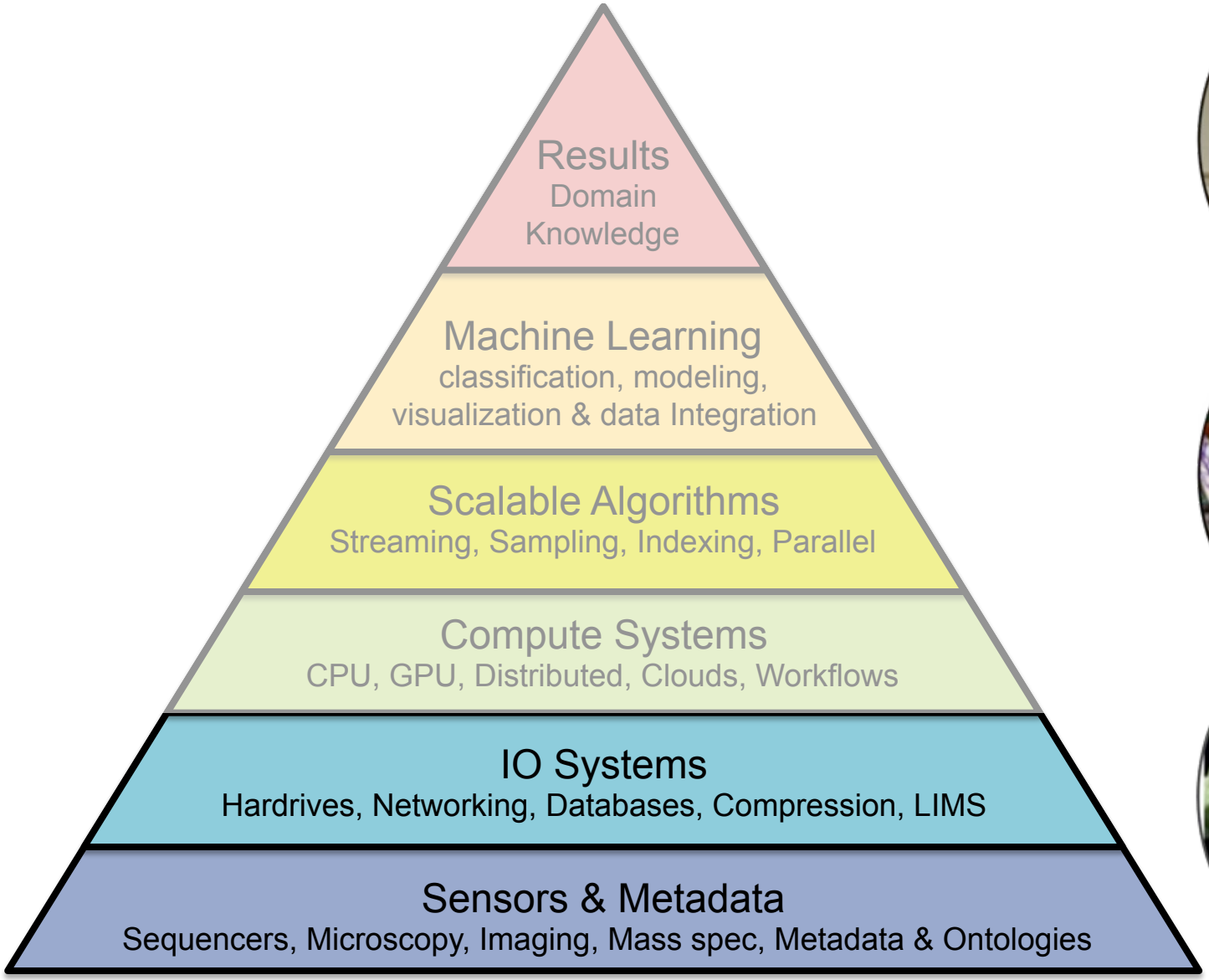



http://en.wikipedia.org/wiki/Data_science

# Comparative Genomics Technologies



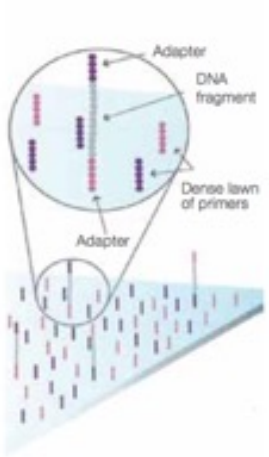

# Comparative Genomics Technologies



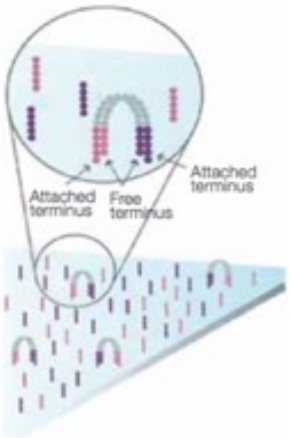

# Massively Parallel Sequencing

**Illumina HiSeq 2000**
*Sequencing by Synthesis*

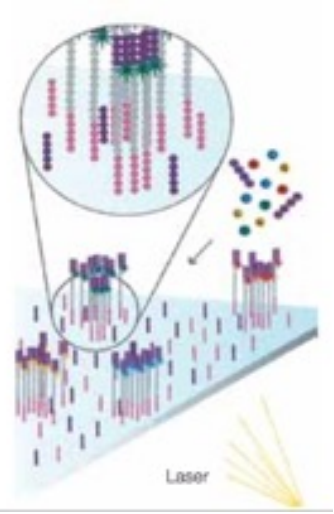
>60Gbp / day



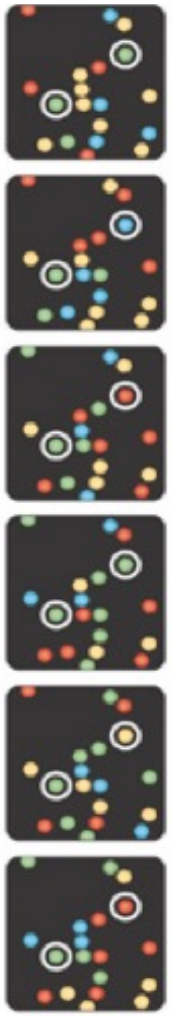

1. Attach




2. Amplify

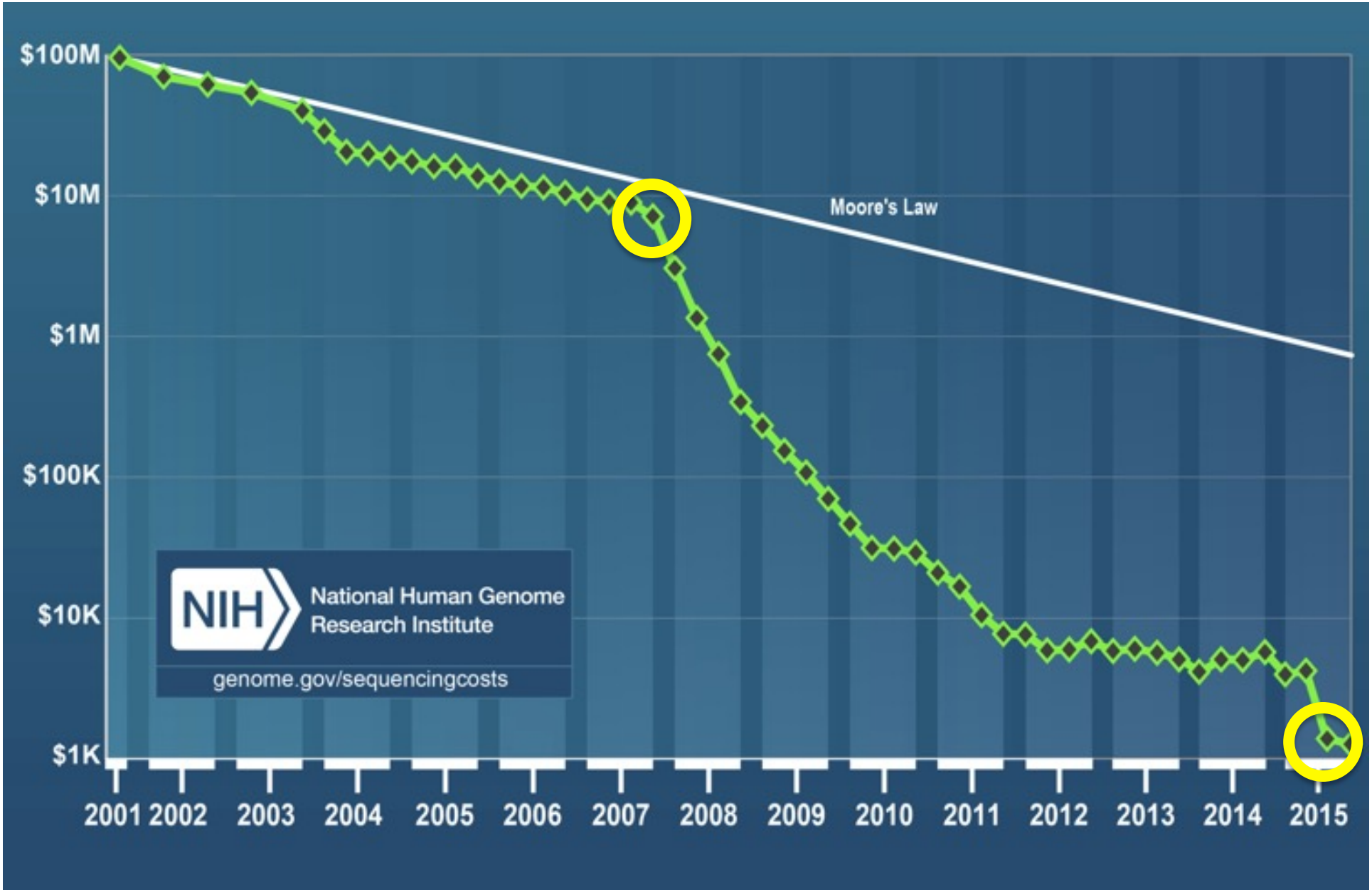



3. Image



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

# Cost per Genome



# HiSeq X Ten

BROAD INSTITUTE

GARVAN INSTITUTE

NEW YORK GENOME CENTER

deCODE genetics

MACROGEN Axeq Technologies

Novogene 诺禾致源

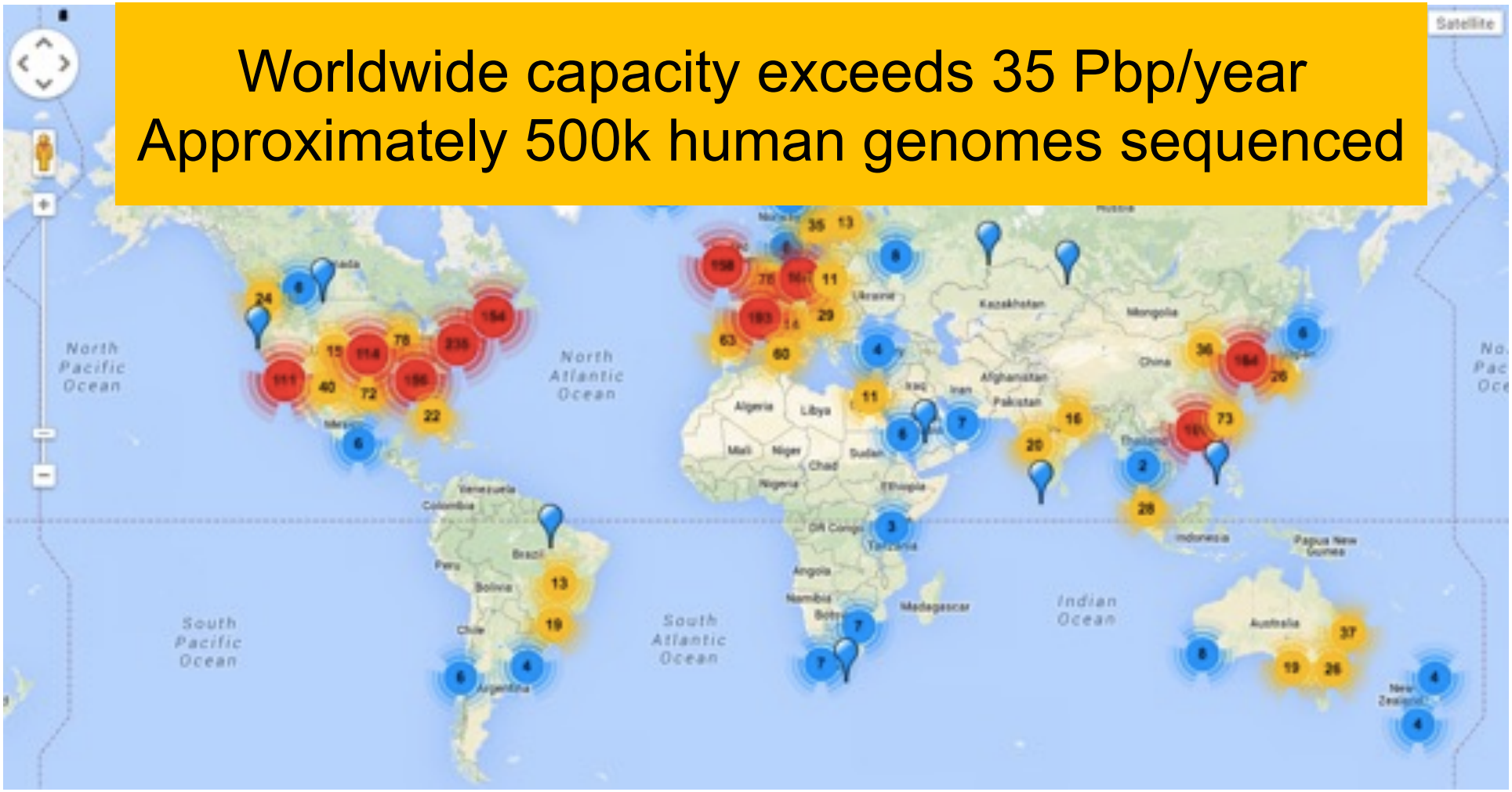
药明康德 WuXi PharmaTech

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




***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

=

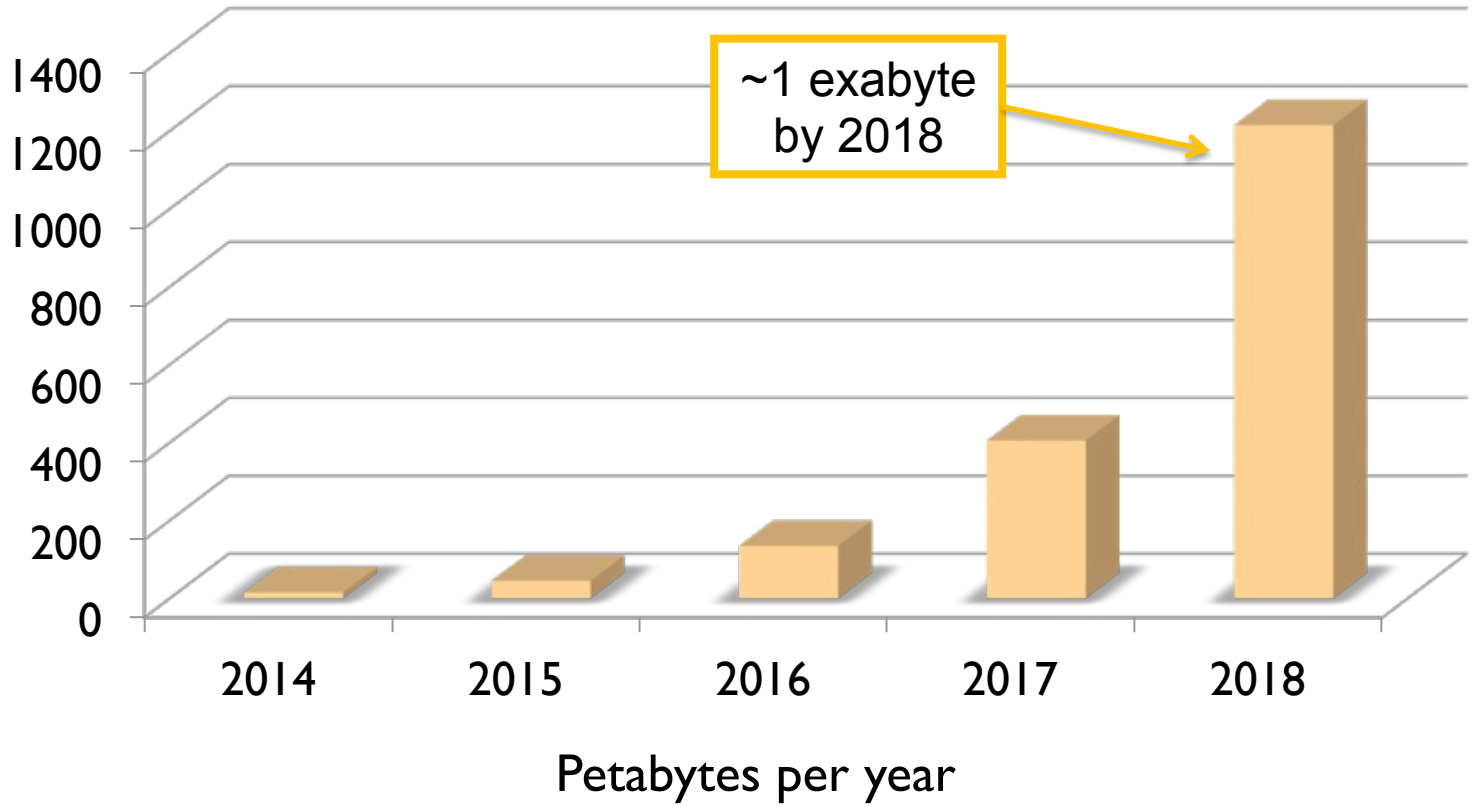
1PB Data
200,000 DVDs

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

500 2 TB drives
$500k

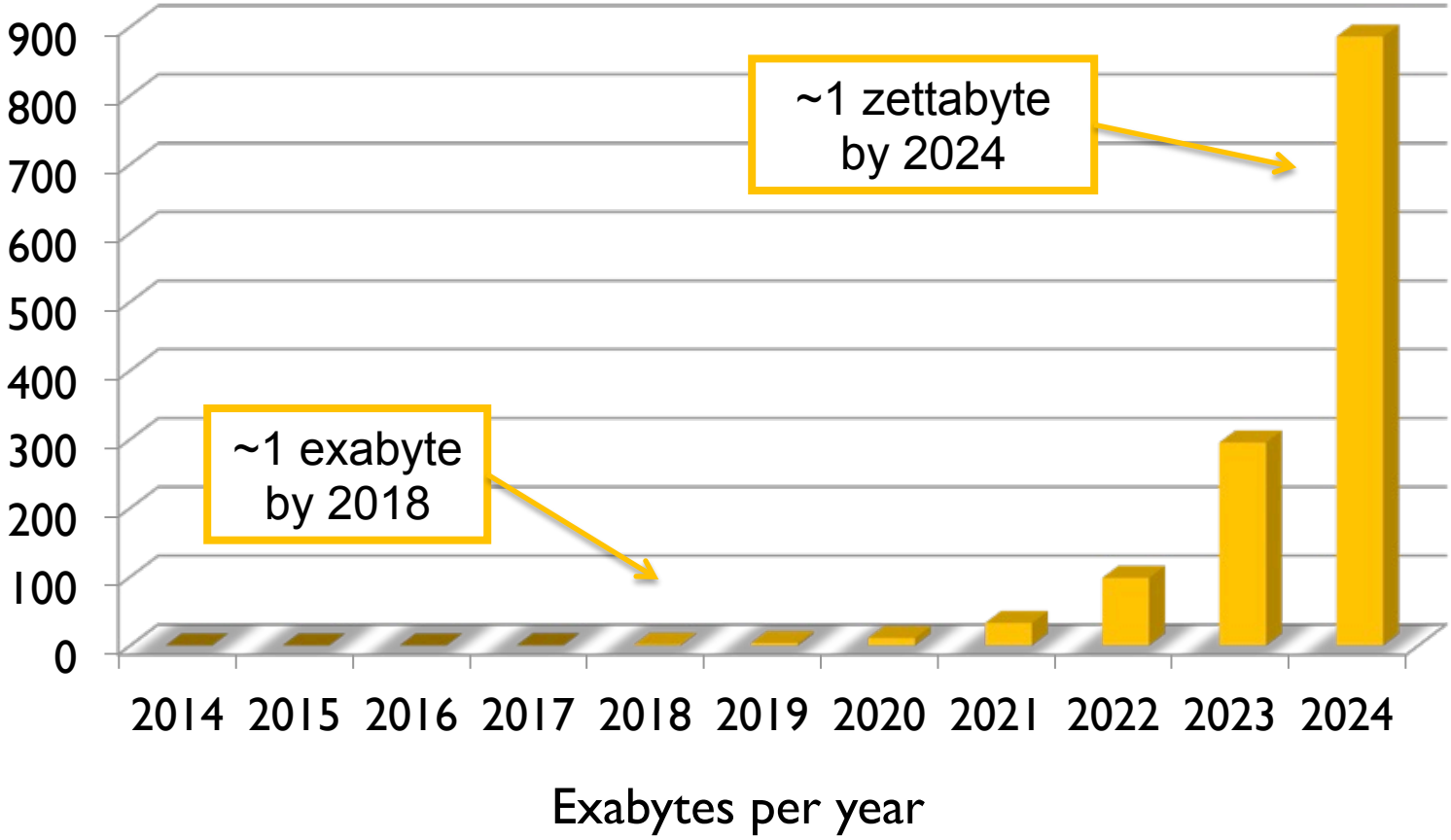
# DNA Data Tsunami

*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

=

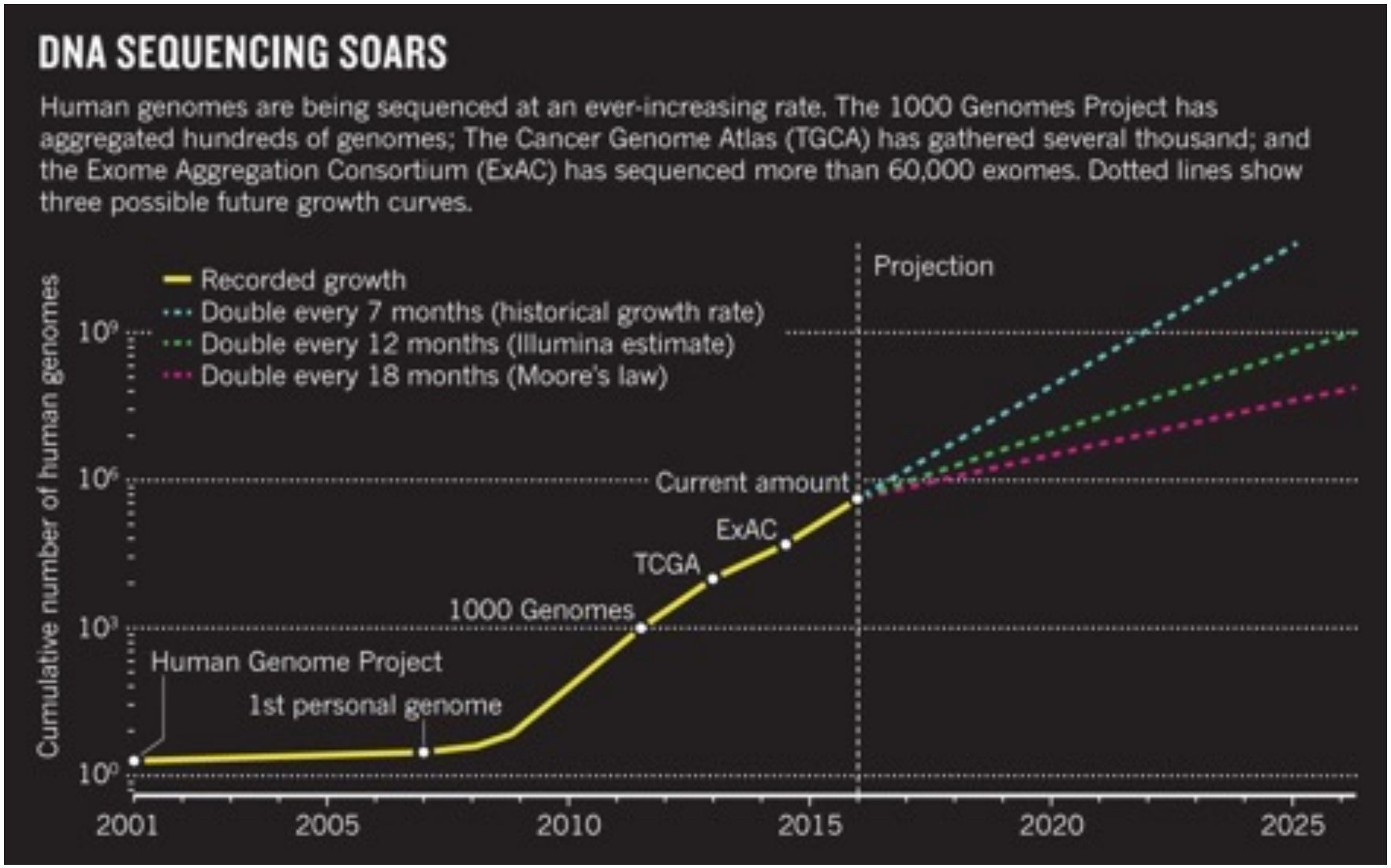
1ZB Data
200,000,000,000 DVDs

150,000 miles of DVDs
~ ½ distance to moon

YouTube

Both currently ~100Pb
And growing exponentially

# Sequencing Capacity




***Big Data: Astronomical or Genomical?***
*Stephens, Z, et al. (2015) PLOS Biology DOI: 10.1371/journal.pbio.1002195*

# Sequencing Centers 2017


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

# Sequencing Centers 2027

***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

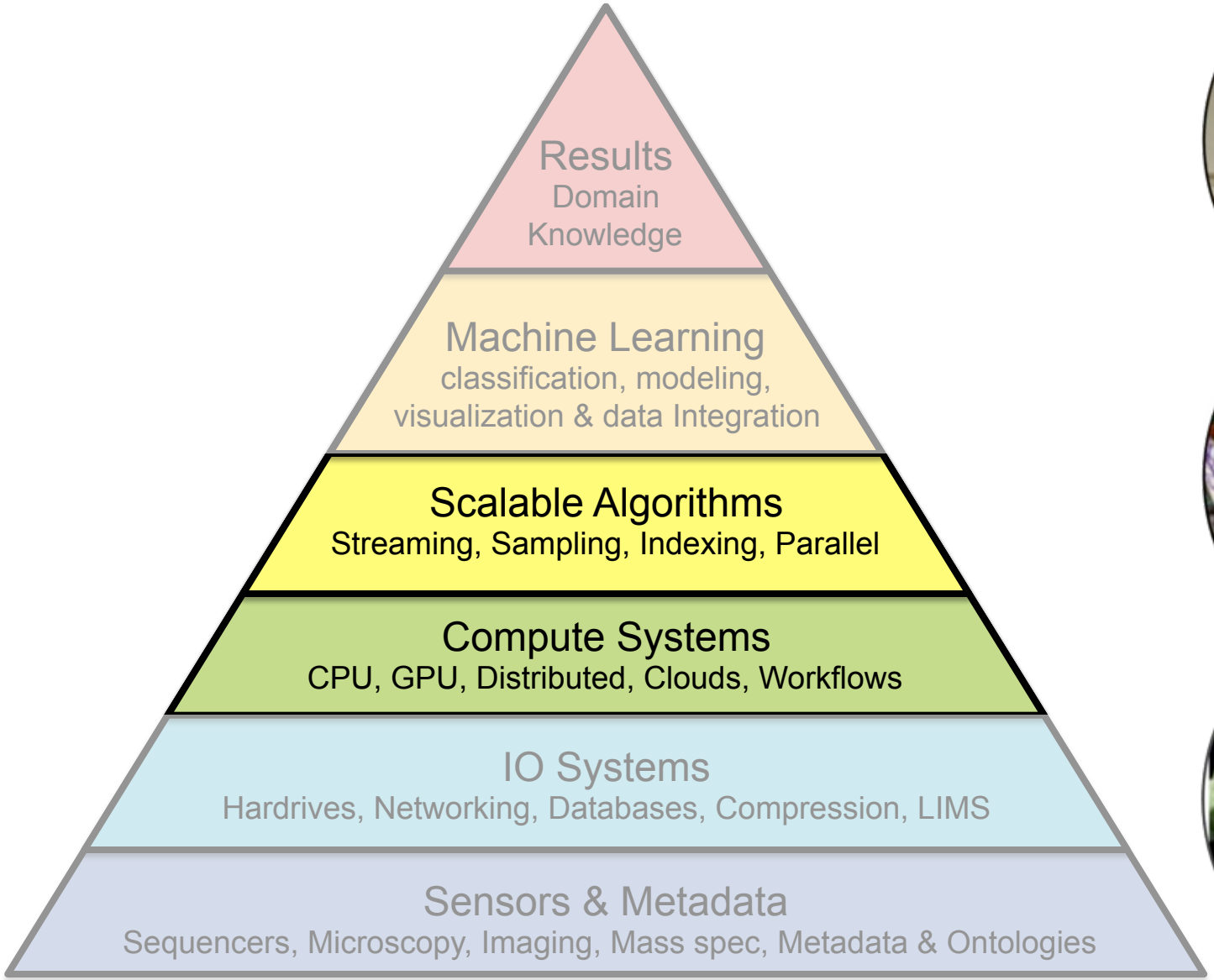
**Major data producers concentrated in hospitals, universities, agricultural companies, research institutes**

- 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?

# Comparative Genomics Technologies

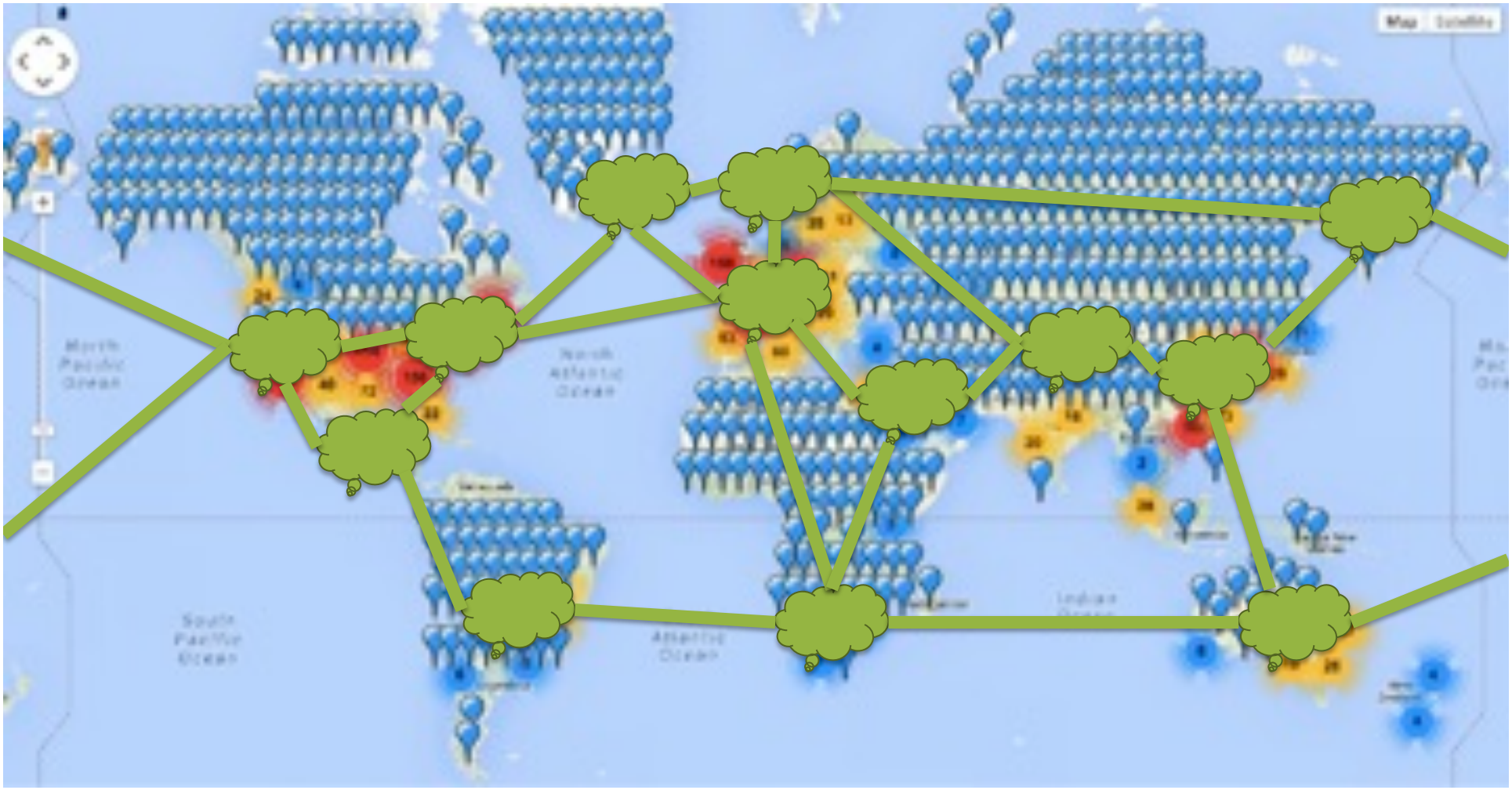



# Sequencing Centers 2027

# Informatics Centers 2027

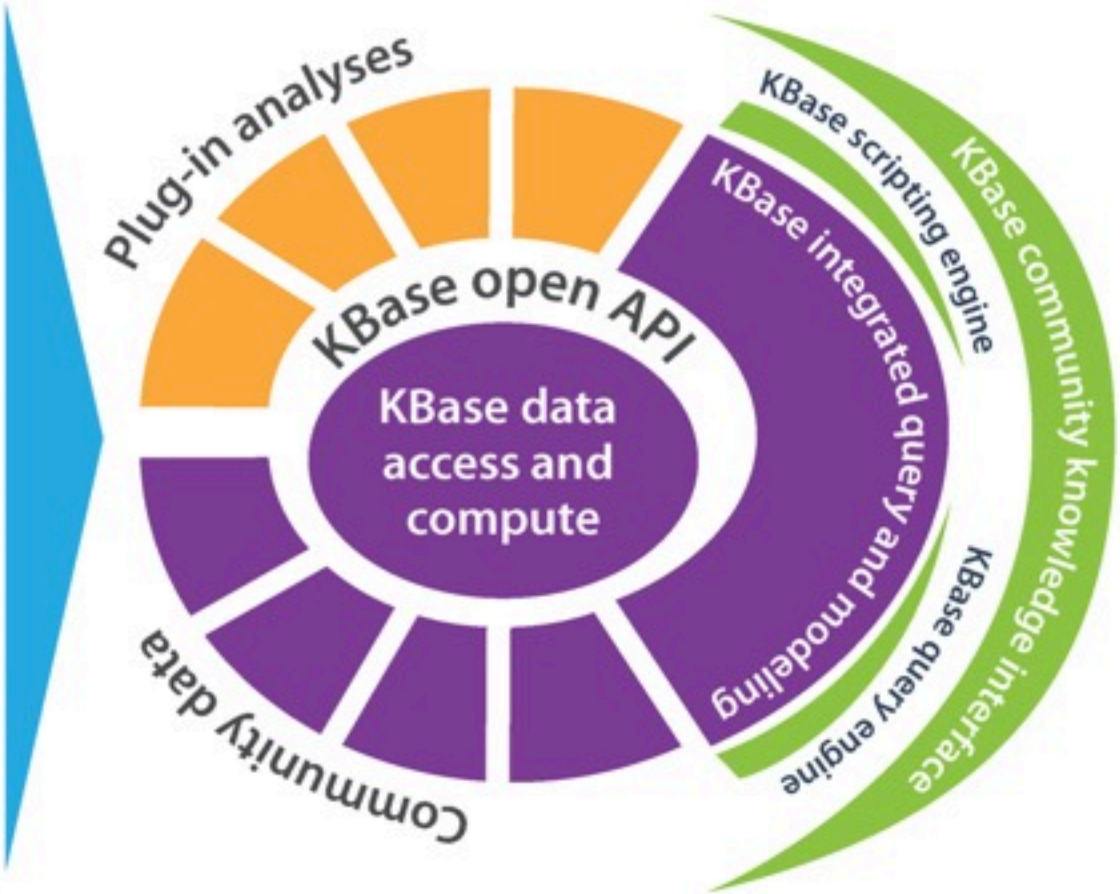

***The DNA Data Deluge***
Schatz, MC and Langmead, B (2013) *IEEE Spectrum*. July, 2013

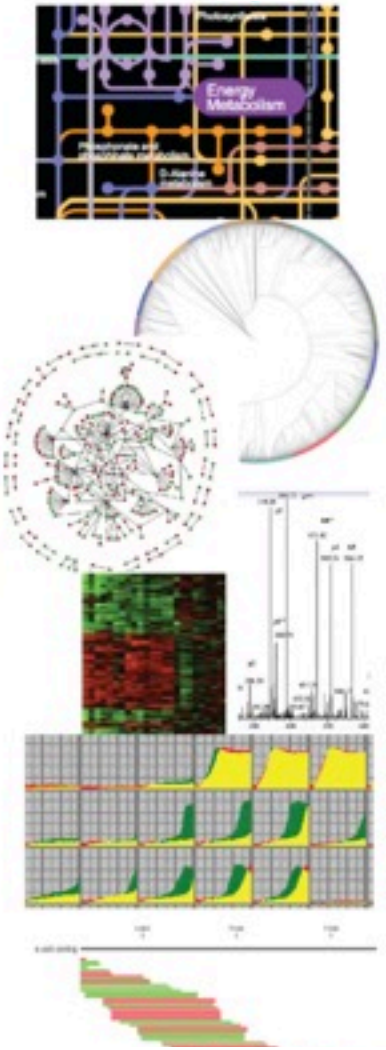
# Informatics Centers 2017



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

# DOE Systems Biology Knowledgebase

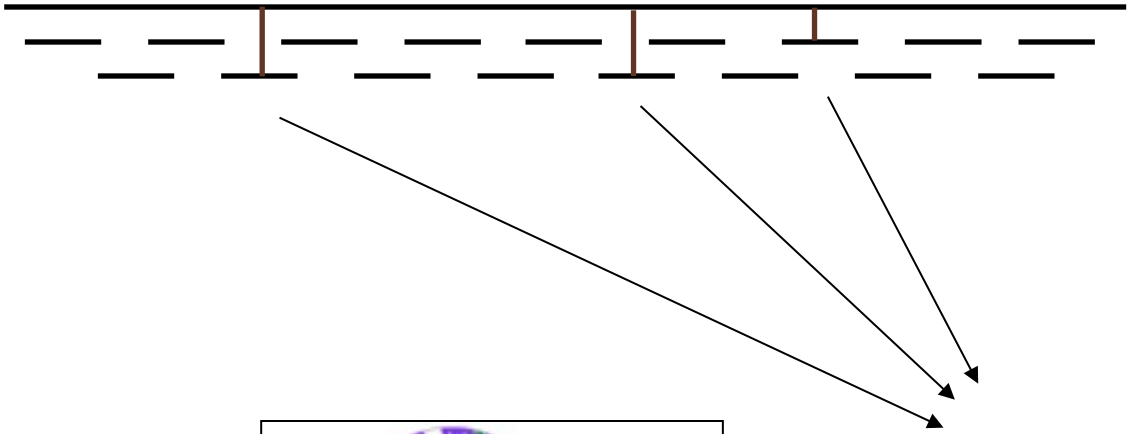


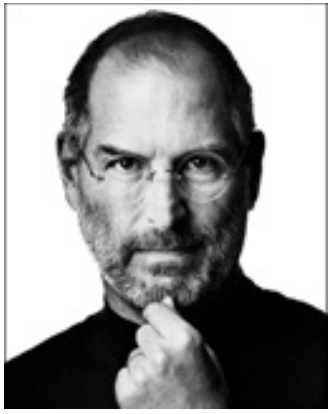



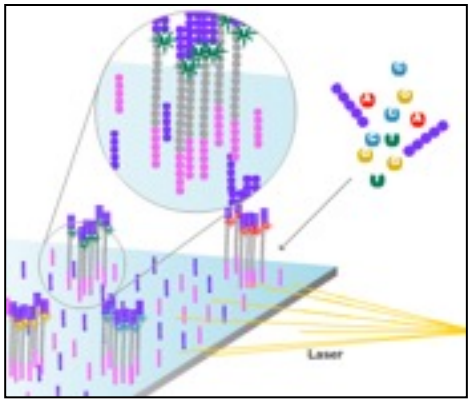
**http://kbase.us: Predictive Biology in Microbes, Plants, and Meta-communities**

# Personal Genomics

How does your genome compare to the reference?








Heart Disease

Cancer

Creates magical technology

# Personal Genomics

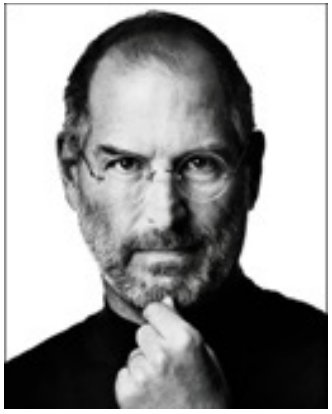
## How does your genome compare to the reference?

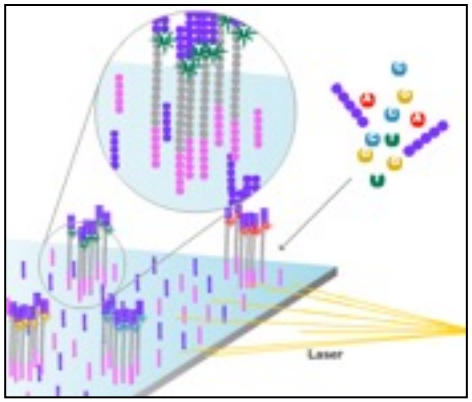
Databases & Assays

Variant Calling

Interpretation

Assembly & Mapping



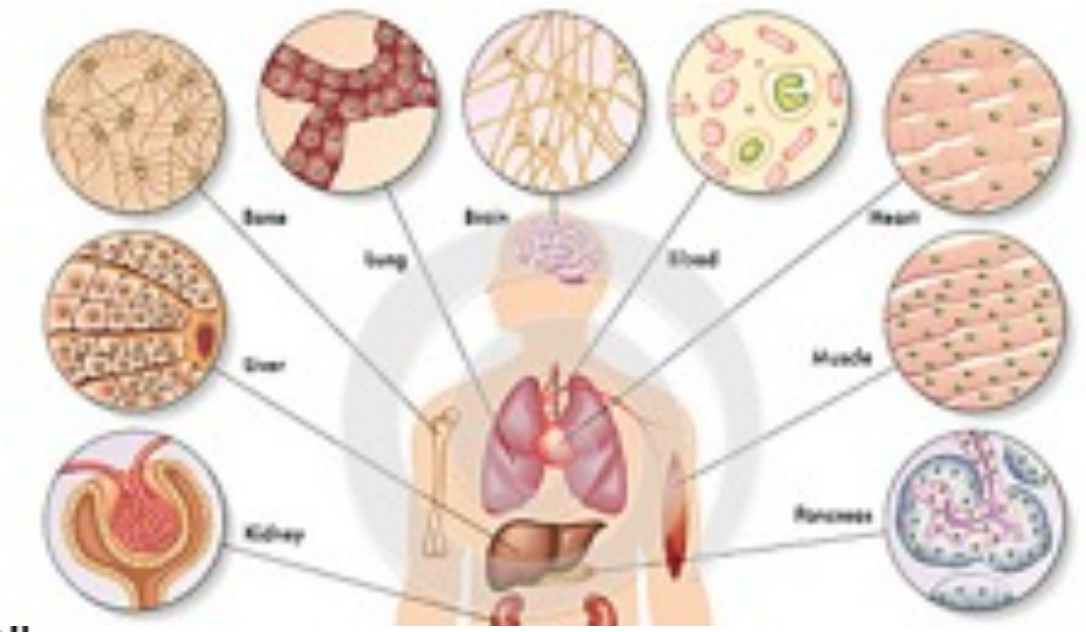


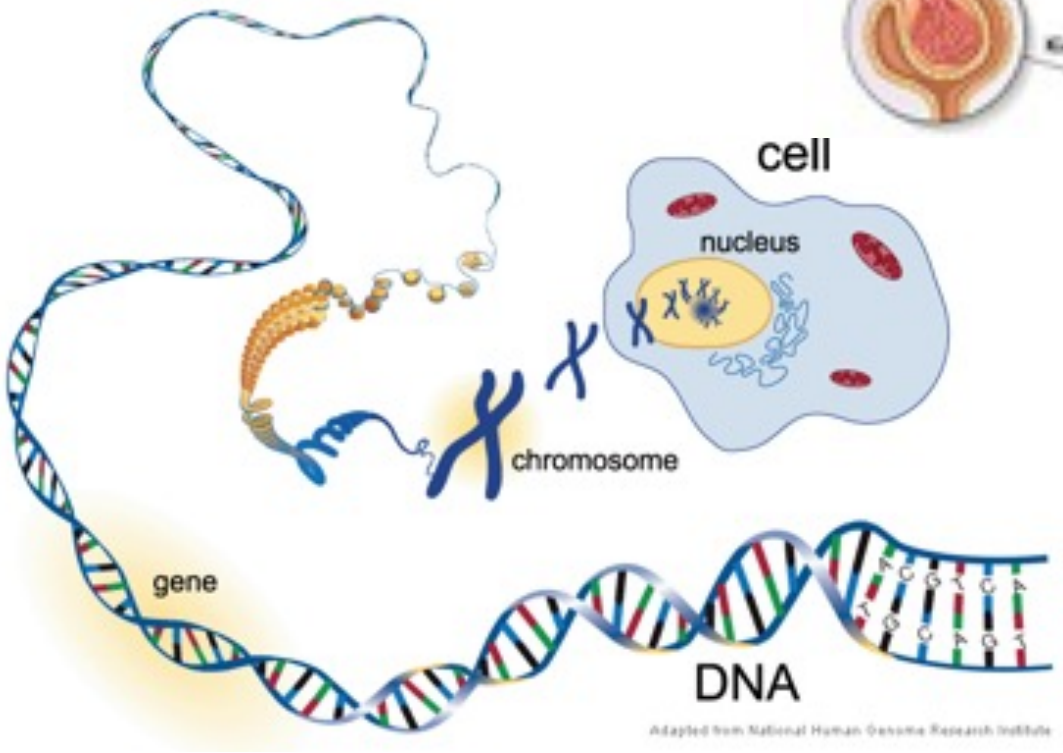

Association

Basecalling

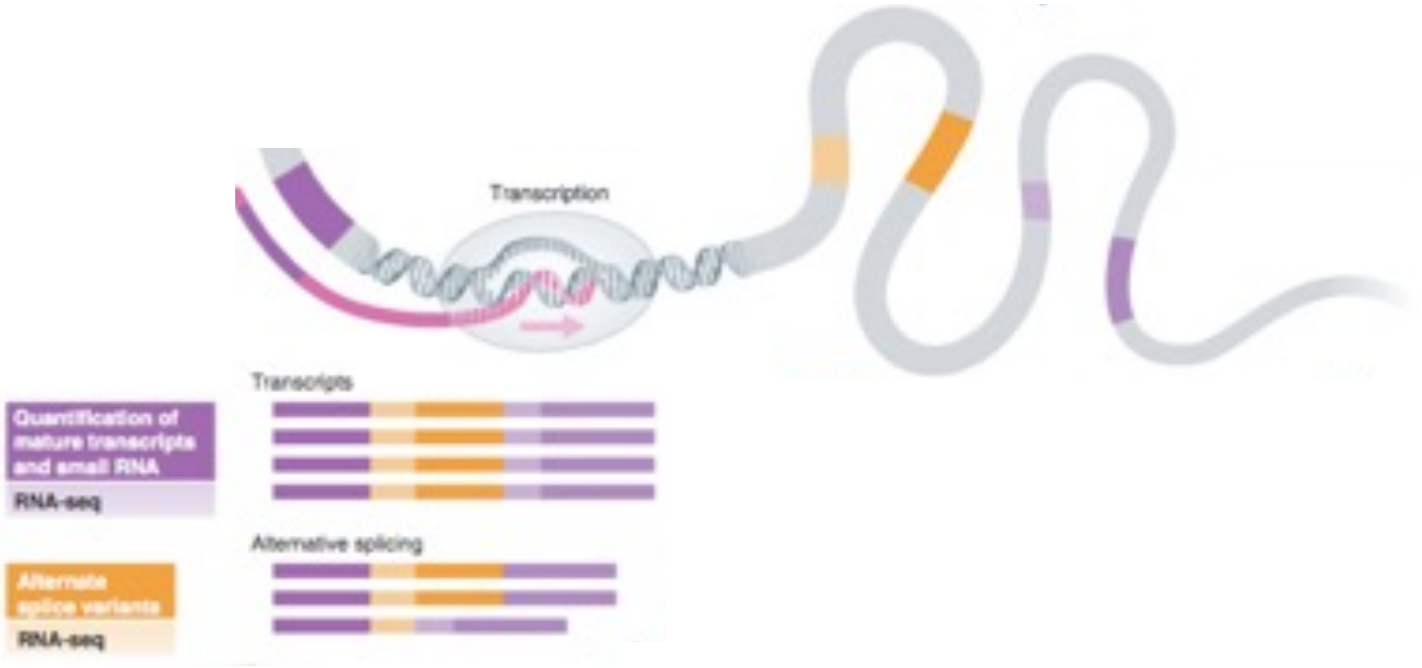
# Cells & DNA

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



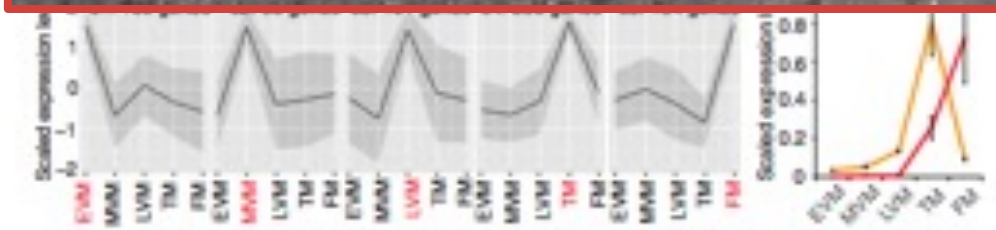

Soon et al., Molecular Systems Biology, 2013

# Rate of meristem maturation determines inflorescence architecture in tomato

Soon Ju Park[1], Ke Jiang[1], Michael C. Schatz, and Zachary B. Lippman[2]

Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724

Edited by Maarten Koornneef, Wageningen University and Research Centre, Cologne, Germany, and approved November 28, 2011 (received for review



- When those genes are delayed or interupted, tomato mutants take on very different branching patterns.

# 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

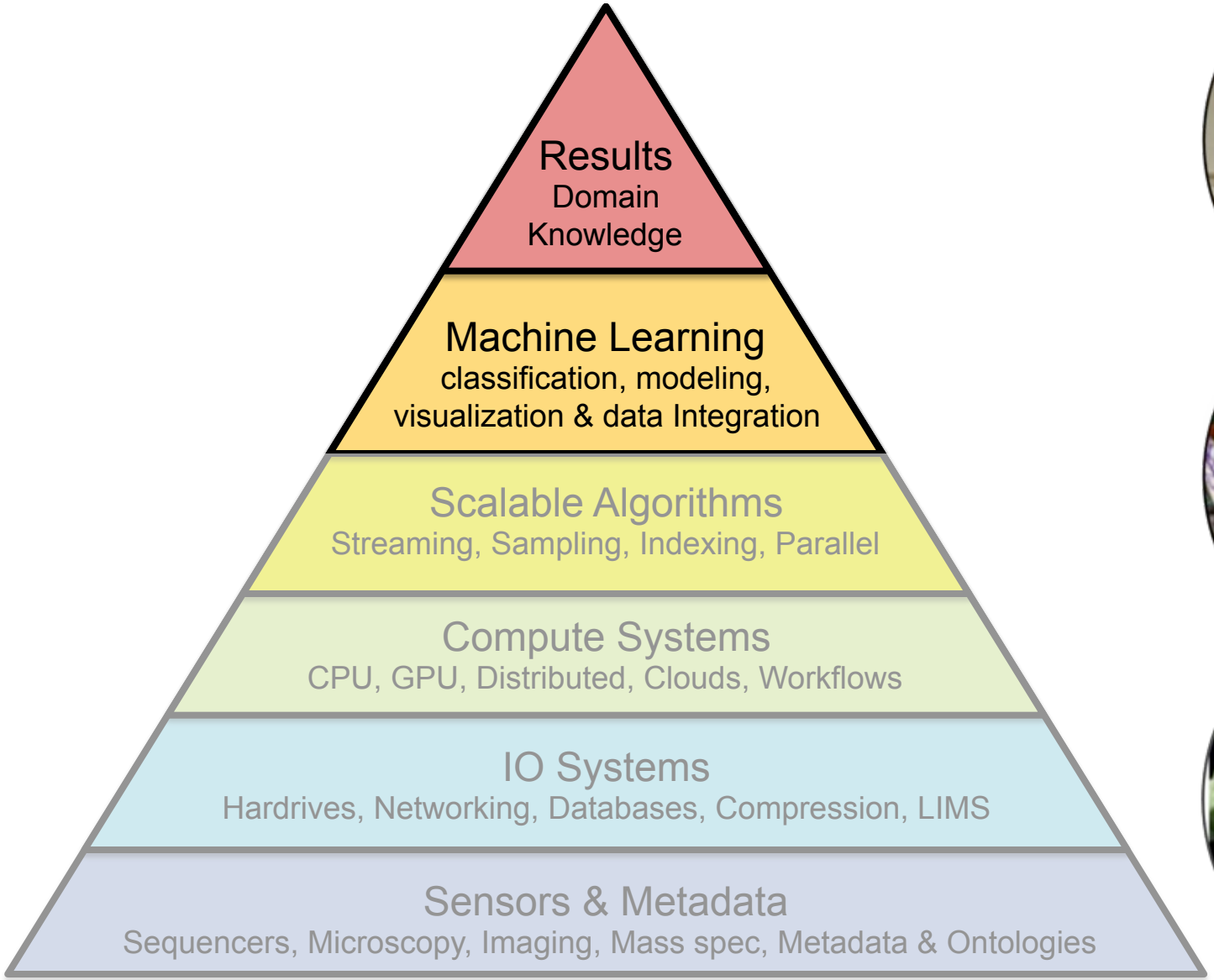
**Parallel hardware and algorithms are required**

- 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

# Comparative Genomics 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

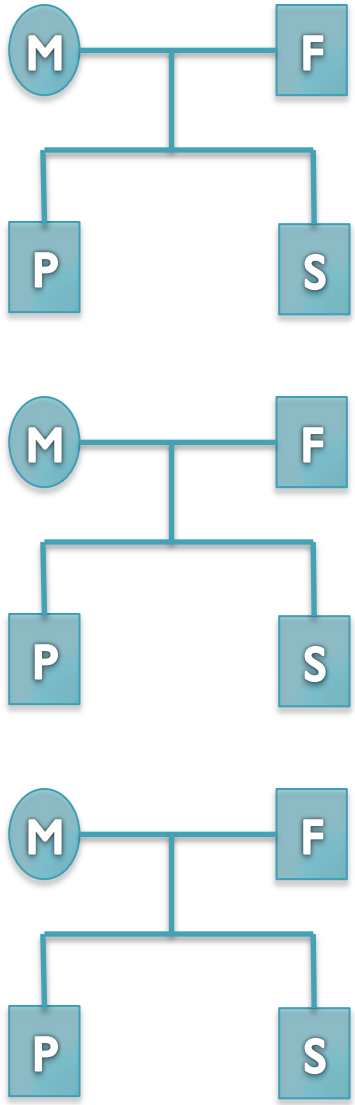
**What is Autism?**
http://www.autismspeaks.org/what-autism

# 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

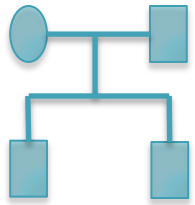
***Are there any genetic variants present in affected children, that are not in their parents or unaffected siblings?***




SFARI SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

# 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

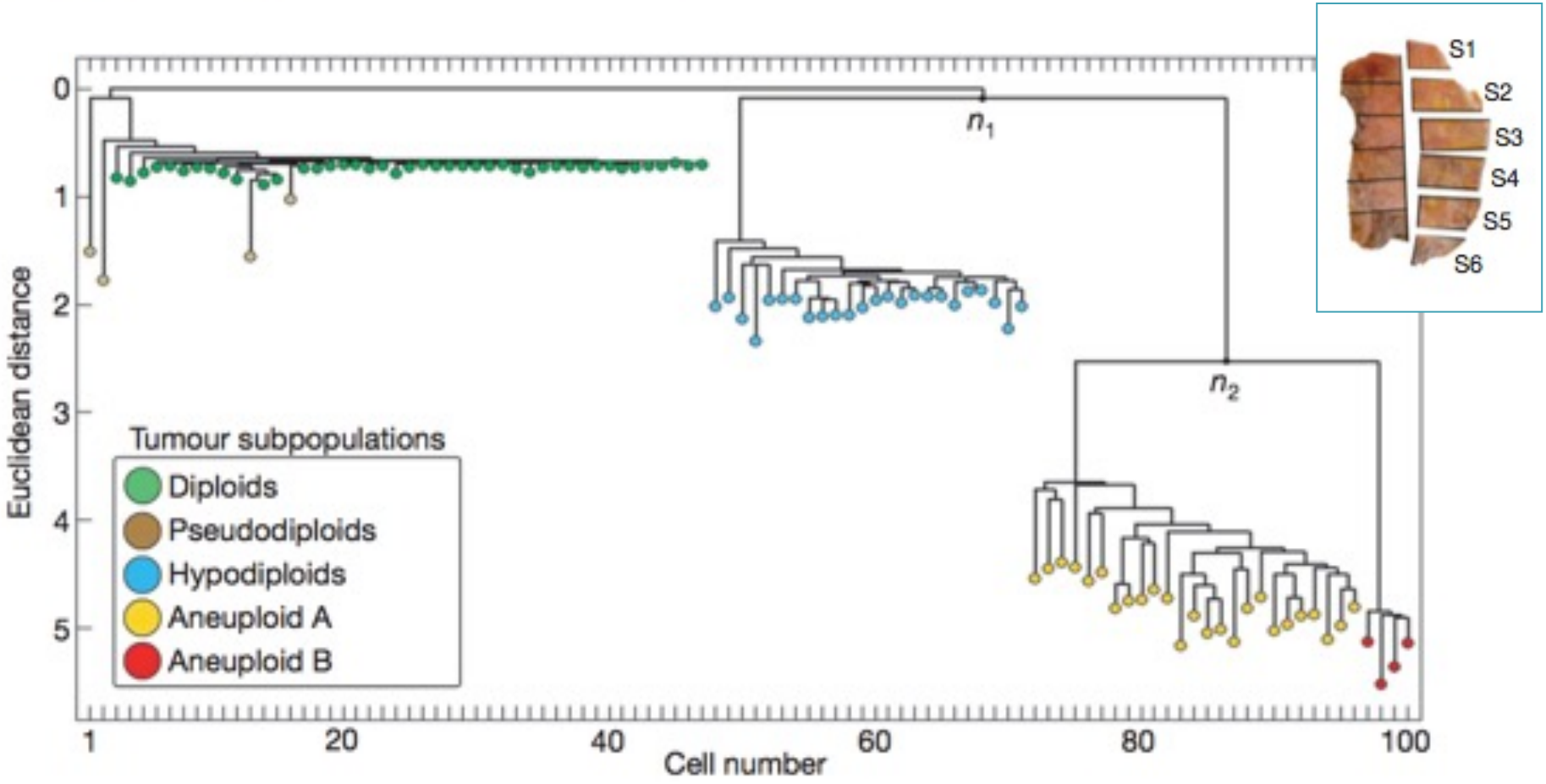
**Accurate de novo and transmitted indel detection in exome-capture data using microassembly.**
Narzisi *et al* (2014) Nature Methods doi:10.1038/nmeth.3069

# LETTER

doi:10.1038/nature09807

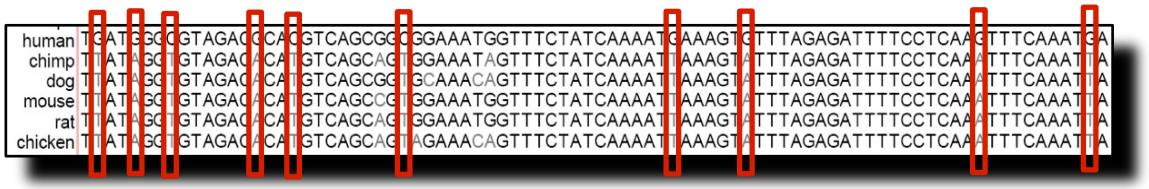
# Tumour evolution inferred by single-cell sequencing

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]



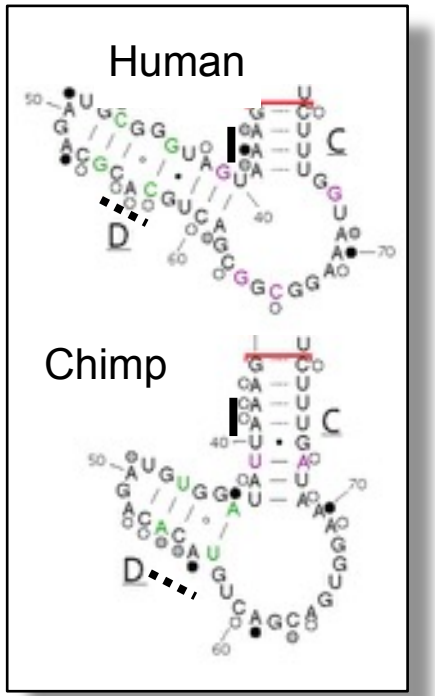
# What makes us human?

## "*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 haploty
(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 I
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**

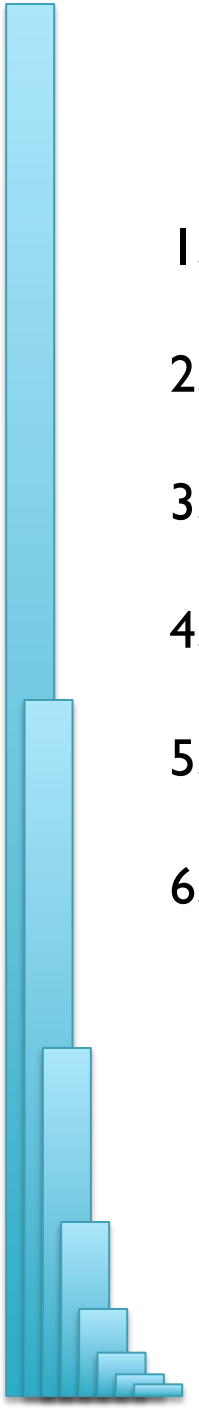
- The potential for over-fitting grows with the complexity of the data, statistical significance is a statement about the sample size
- Reproducible workflows, APIs are a must
- Caution is prudent for personal data

**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

# Next Steps

1. Reflect on the magic and power of DNA ☺
2. Check out the course webpage
3. Register on Piazza
4. Set up Dropbox for yourself!
5. Set up Linux, set up Virtual Machine
6. Get comfortable on the command line




*Welcome to Applied Comparative Genomics*
https://github.com/schatzlab/appliedgenomics

# Questions?