

Welcome to Biomedical Research!

Michael Schatz

August 31, 2017 – Lecture I

EN.601.452 Computational Biomedical Research

AS.020.415 Advanced Biomedical Research



Welcome!

The goal of this course is to prepare undergraduates to understand and perform state-of-the-art biomedical research. This will be accomplished through three main components:

1. **Lectures** on cross cutting techniques for biomedical research focusing on data visualization, statistical inference, and scientific computing
2. **Research presentations** from distinguished faculty on their active research projects
3. **A major research project** to be performed under the mentorship of a JHU professor.

Course Webpage: <https://github.com/schatzlab/biomedicalresearch>

Course Discussions: <http://piazza.com>

Class Hours: Mon + Wed @ 3p – 3:50p Malone 107

Office Hours: Wed @ 4-5p and by appointment
Please try Piazza first!

Prerequisites and Resources

Prerequisites

- No formal course requirements
- Access to an Apple or Linux Machine, or Install VirtualBox
- Familiarity with a major programming language is nice but not essential
 - C/C++, Java, R, Perl, Python, JavaScript, others?

Primary Texts

- None! We will be studying primary research papers

Other Resources:

- Google, SEQanswers, Biostars, StackOverflow
- Applied Computational Genomics Course
 - <https://github.com/schatzlab/appliedgenomics>
- Ben Langmead's teaching materials:
 - <http://www.langmead-lab.org/teaching-materials/>



Grading Policies

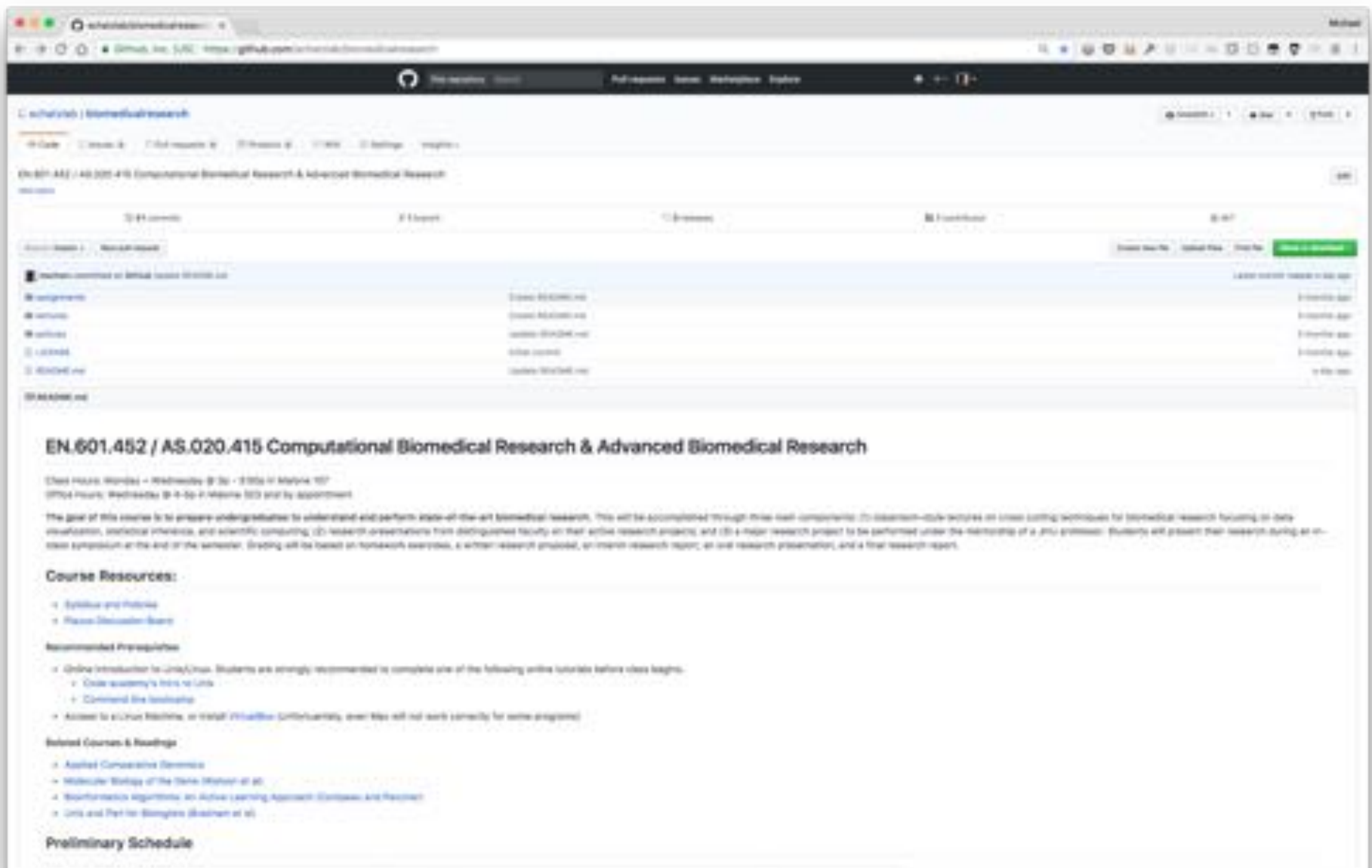
Assessments:

- ~6 HW Exercises: 10% Due at 11:59pm a week later
- Research Proposal: 10% ~1 page write up
- Interim Report: 10% ~3 page progress report
- Project Presentation: 30% Presented last week of class
- Final Report: 30% Due last week of semester
- In-class Participation: 10% Please ask questions!

Policies:

- Scores assigned relative to the highest points awarded
- Automated testing and grading of analysis assignments
- **Late Days:**
 - Four (4) chances to extend the deadline for assignments by 24 hours without any penalty, after that 25% deduction per day

Course Webpage



The screenshot shows the GitHub repository page for 'schatzlab/biomedicalresearch'. The repository is titled 'EN.601.452 / AS.020.415 Computational Biomedical Research & Advanced Biomedical Research'. It has 24 commits, 1 branch, 11 issues, and 1 pull request. The repository is a public project by schatzlab. The page displays a table of files and folders, including 'assignments', 'lectures', 'resources', 'scripts', and 'slides'. The main content area shows the repository description, course resources, recommended prerequisites, related courses, and a preliminary schedule.

EN.601.452 / AS.020.415 Computational Biomedical Research & Advanced Biomedical Research

Class Hours: Monday - Wednesday 9:30 - 11:00 in Marina 107
Office Hours: Wednesday 9:30 - 11:00 in Marina 107 and by appointment

The goal of this course is to prepare undergraduate to understand and perform state-of-the-art biomedical research. This will be accomplished through three main components: (1) classroom-style lectures on cross cutting techniques for biomedical research focusing on data visualization, statistical inference, and scientific computing; (2) research presentations from distinguished faculty on their active research projects; and (3) a major research project to be performed under the mentorship of a JHU professor. Students will present their research during an on-site symposium at the end of the semester. Grading will be based on homework exercises, a written research proposal, an interim research report, an oral research presentation, and a final research report.

Course Resources:

- Syllabus and Policies
- Piazza Discussion Board

Recommended Prerequisites

- Online Introduction to Linux/Unix. Students are strongly recommended to complete one of the following online tutorials before class begins:
 - Codecademy's Intro to Linux
 - Command Line Workshop
- Access to a Linux Machine, or install VirtualBox. Unfortunately, over time we will not work correctly for some programs.

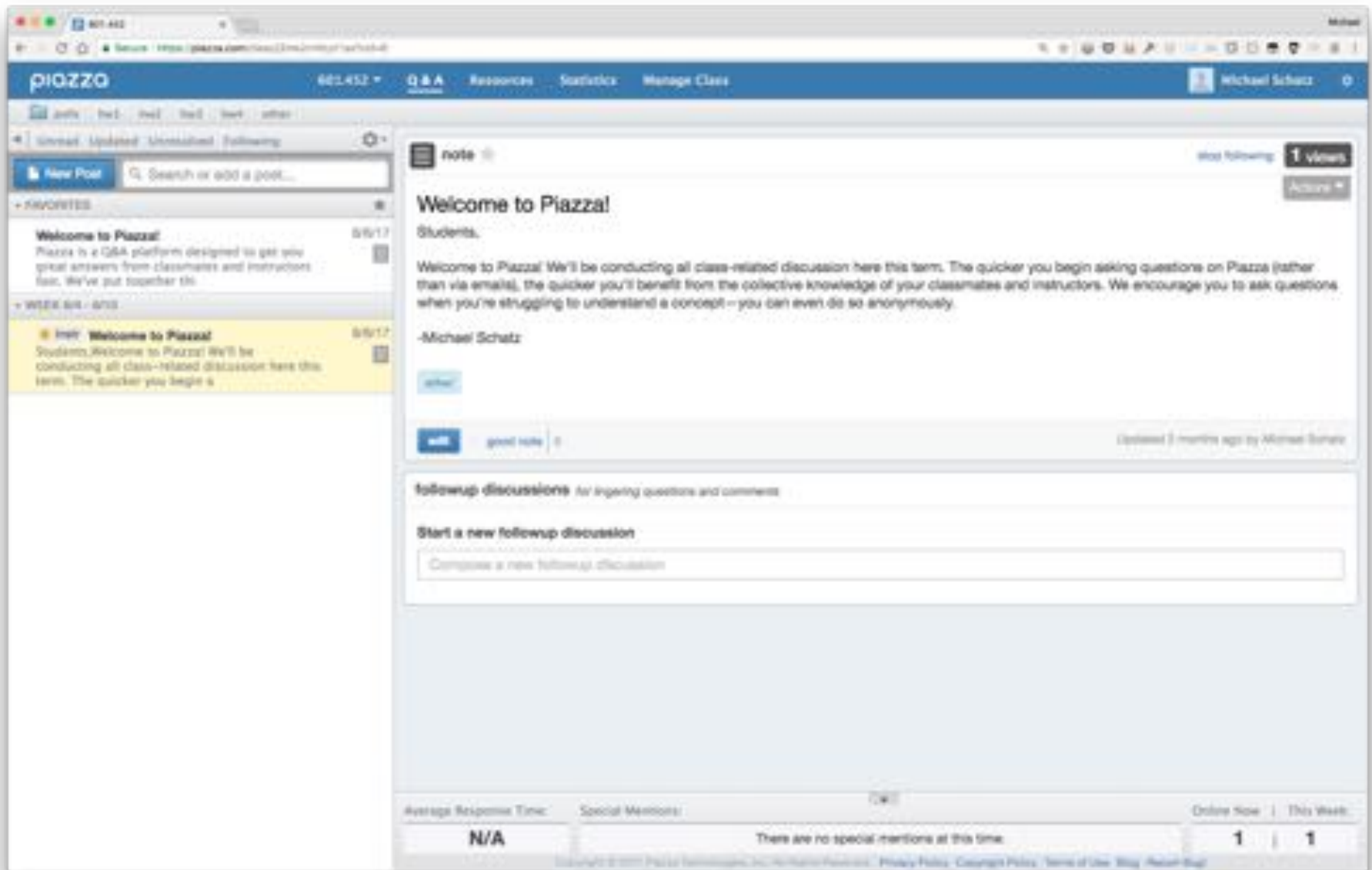
Related Courses & Readings

- Applied Comparative Genomics
- Molecular Biology of the Gene (Watson et al)
- Bioinformatics Algorithms: An Active Learning Approach (Compeau and Pevzner)
- Unix and Perl for Biologists (Bioinformatics)

Preliminary Schedule

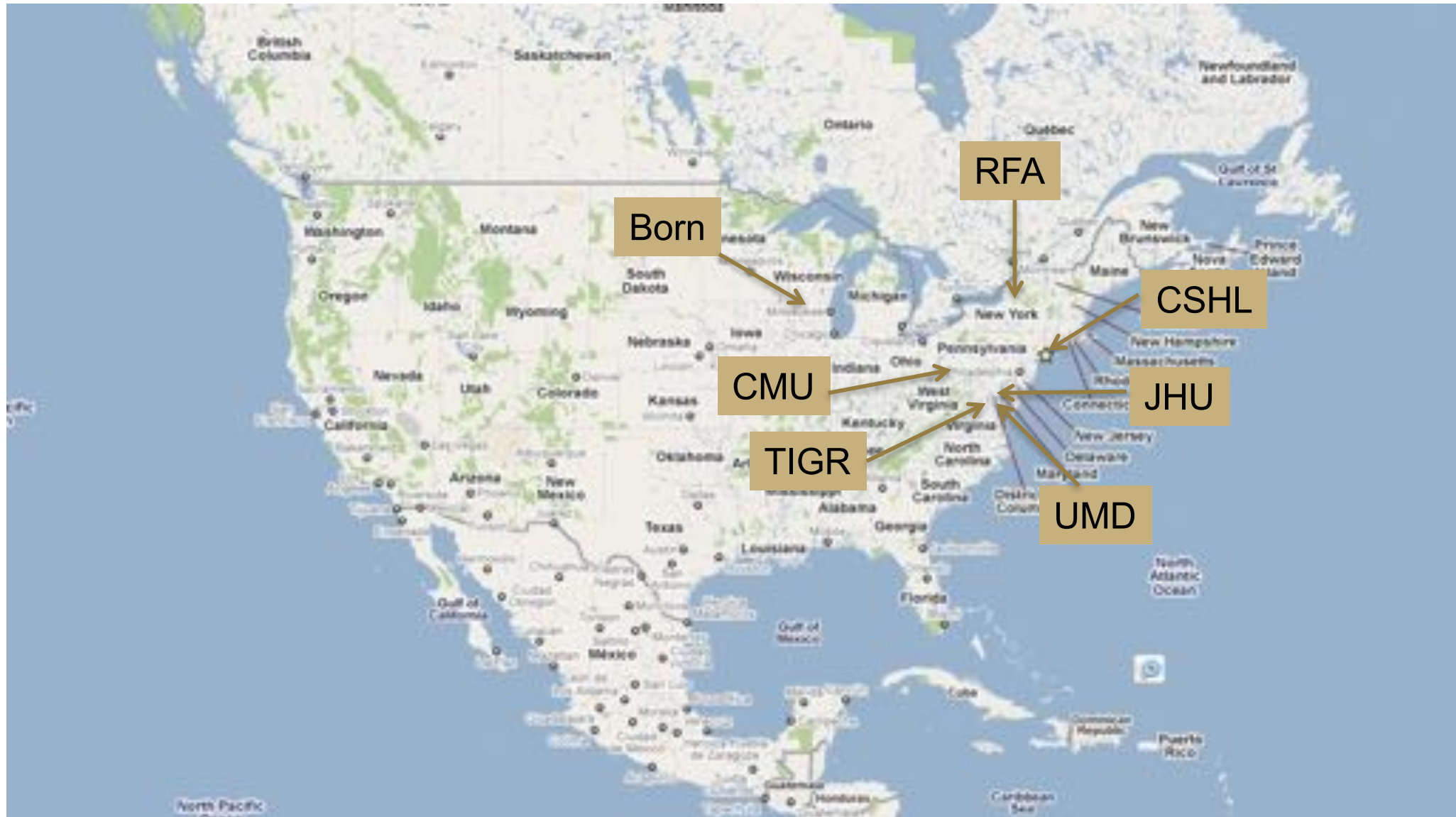
<https://github.com/schatzlab/biomedicalresearch>

Piazza



<http://piazza.com/jhu/fall2017/601452/home>

A Little About Me



A Little About Me



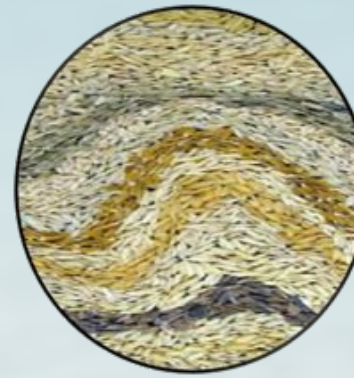
Schatzlab Overview



Human Genetics

Role of mutations in
disease

Feigin *et al.* (2017)
Fang *et al.* (2016)



Agricultural Genomics

Genomes &
Transcriptomes

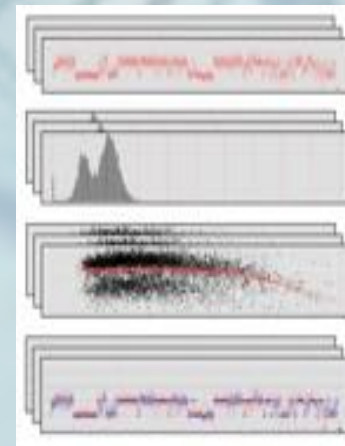
Lemmon *et al.* (2016)
Ming *et al.* (2015)



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

The Origins of DNA Sequencing

Nature Vol. 265 February 24 1977

articles

Nucleotide sequence of bacteriophage Φ X174 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 2QF, UK

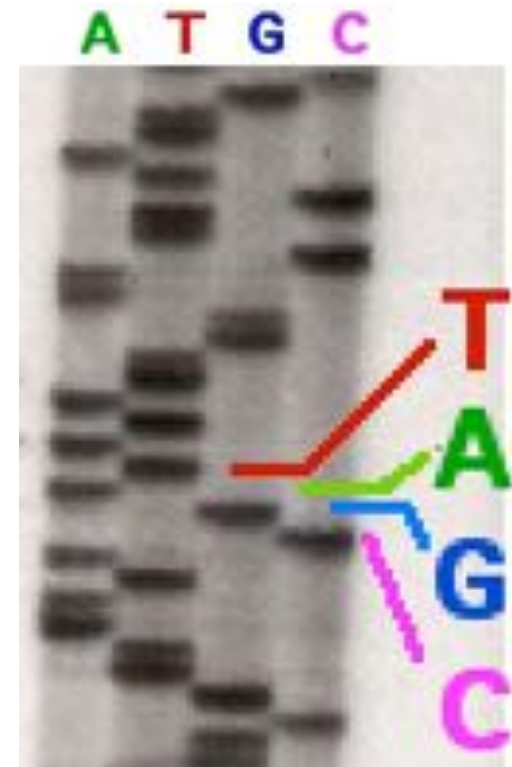
A DNA sequence for the genome of bacteriophage Φ X174 of approximately 5,375 nucleotides has been determined using the rapid and simple 'plus and minus' method. The sequence identifies many of the features responsible for the production of the proteins of the nine known genes of the organism, including 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 Φ X174 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^{1,2}, is A-B-C-D-E-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 Φ X 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 protein³ (positions 2,562-2,611).

At this stage sequencing techniques using primer extension with DNA polymerase were being developed⁴ and Schmitt⁵ synthesized a deoxynucleotide with a sequence complementary to part of the ribosome binding site. This was used to prime into the intergenic region between the F and G genes, using DNA polymerase and ³²P-labelled dNTPs. The ribosome-inhibition technique⁶ facilitated the sequence determination of the labelled DNA produced. This deoxynucleotide-primed system was also used to develop the plus and minus method⁷. Suitable synthetic primers are, however, difficult to prepare and so

1977
1st Complete Organism
Bacteriophage ϕ 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 ϕ X174 DNA

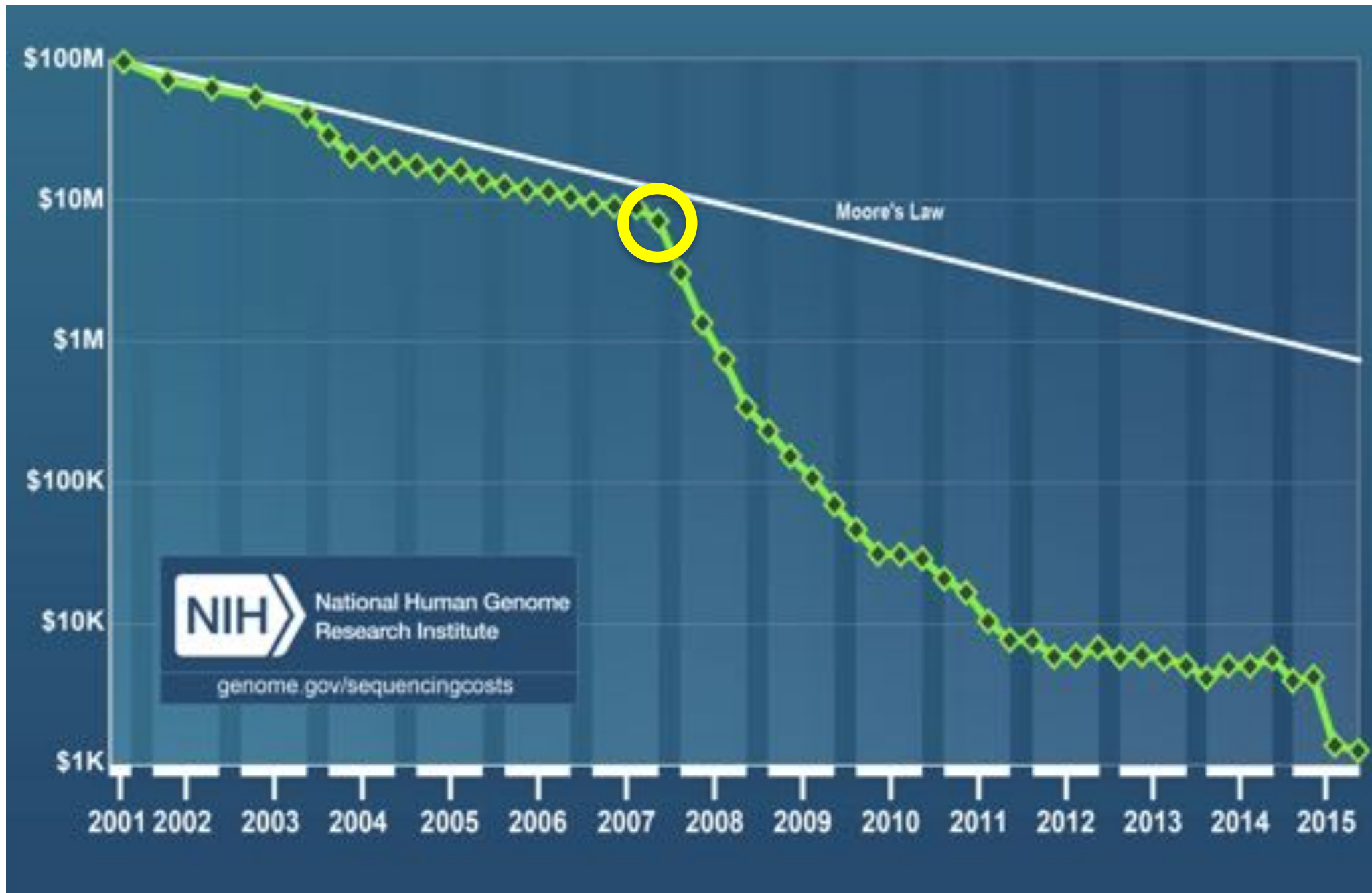
Sanger, F. et al. (1977) *Nature*. 265: 687 - 695

Milestones in DNA Sequencing



(TIGR/Celera, 1995-2001)

Cost per Genome

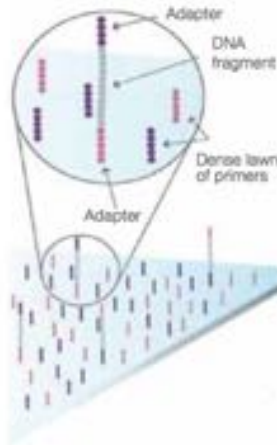


Massively Parallel Sequencing

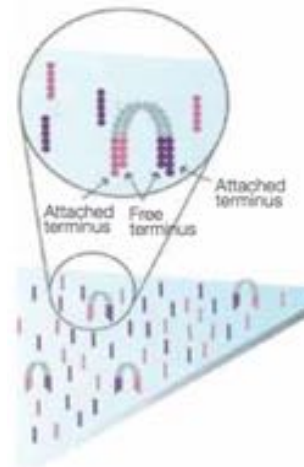


Illumina HiSeq 2000
Sequencing by Synthesis

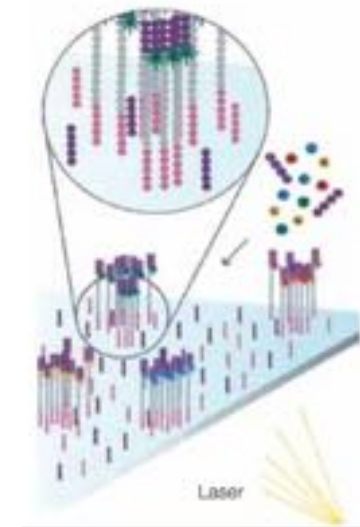
>60Gbp / day



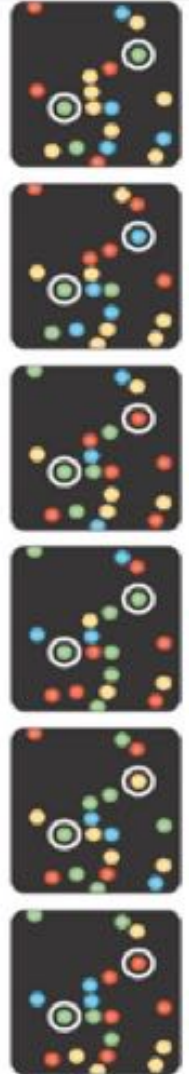
1. Attach



2. Amplify

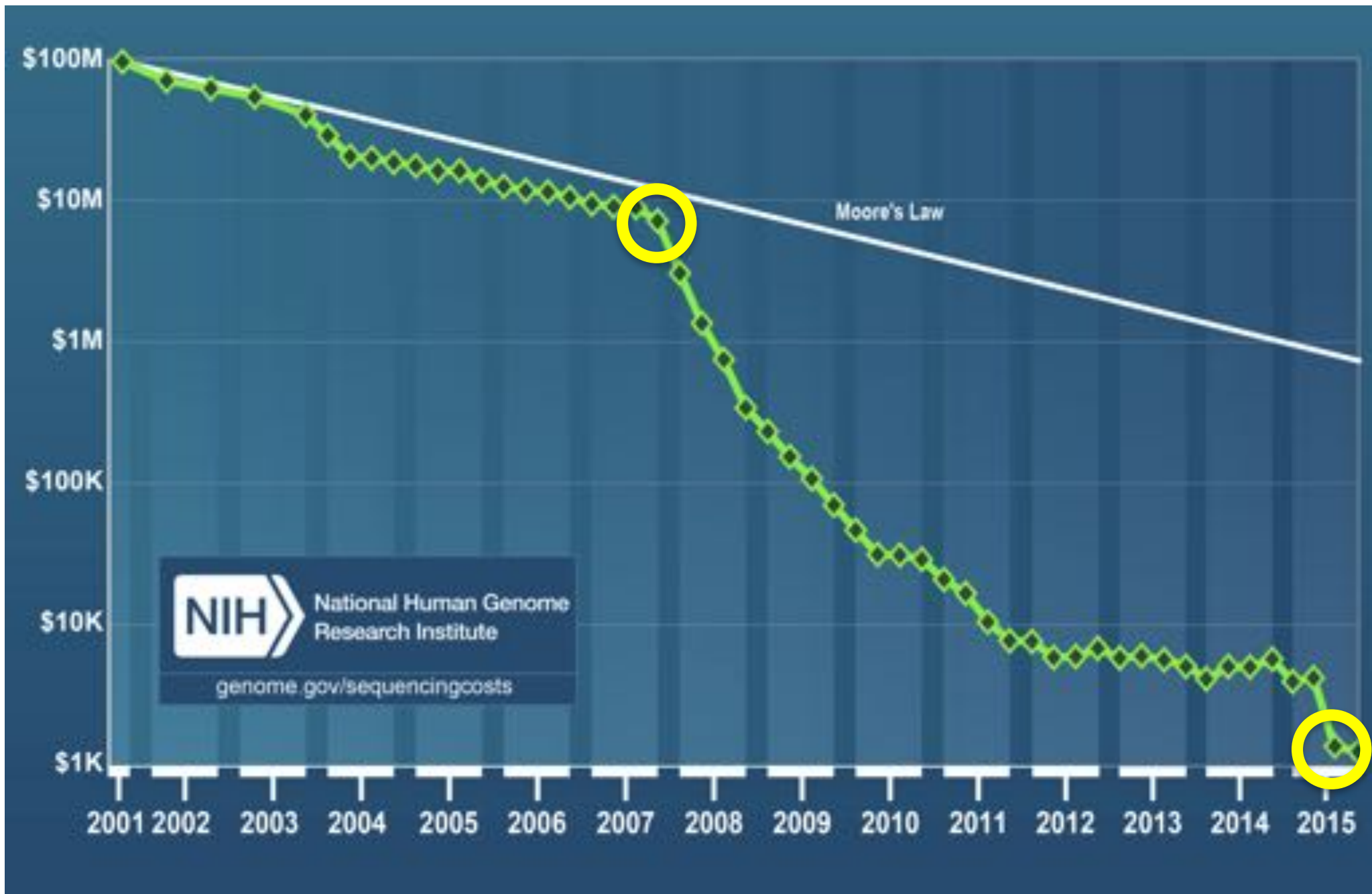


3. Image



Metzker (2010) Nature Reviews Genetics 11:31-46
<http://www.youtube.com/watch?v=l99aKKHcx4>

Cost per Genome



HiSeq X Ten



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

Sequencing Centers

A world map with blue pins indicating sequencing centers. Each pin has a number next to it, representing the number of sequencing centers in that region. The numbers are: North America (38, 13), Europe (11, 7, 16, 20, 28, 73), Africa (22, 13, 19, 4, 8), Asia (10, 13, 19, 4, 8), Australia (37, 19, 20, 4, 4), and South America (13, 19, 4, 8).

Worldwide capacity exceeds 50 Pbp/year
Over 500k human genomes sequenced

On track to exceed over 1 Zbp / year by 2024

A zetta-what?

Next Generation Genomics: World Map of High-throughput Sequencers

<http://omicsmaps.com>

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



Both currently ~100Pb
And growing exponentially

Unsolved Questions in Biology

- What is your genome sequence?

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?

- ***Plus thousands and thousands more***





Next Steps

1. Reflect on the magic and power of DNA 😊
2. Check out the course webpage
3. Register on Piazza & say hello!
4. Set up Dropbox for yourself!



Welcome to Advanced Biomedical Research

<https://github.com/schatzlab/biomedicalresearch>

Questions?