Spark meets Genomics:

Helping Fight the Big C with the Big D

David Patterson and a Cast of Thousands
AMPLab, UC Berkeley

June 30, 2014

Opportunity?

- **Provocative Argument**:
  Given fast growing use of genomics, a CS group that learned some biology might be a big help with genetic diseases (e.g., cancer)
- "If you can help save 100 lives a year just in US, that could justify a whole career!"

or Obligation?

Computer Scientists May Have What It Takes to Help Cure Cancer

The New York Times

By DAVID PATTERSON, DECEMBER 6, 2011

...The night after we made that argument, I awoke in the middle of the night with this question etched into my mind:

Given that millions of people do have and will get cancer, if there is a chance that computer scientists may have the best skill set to fight cancer today, as moral people aren't we obligated to try?

The NEW ENGLAND JOURNAL of MEDICINE

HOME | ARTICLES & MULTIMEDIA | ISSUES | SPECIALTIES & TOPICS | FOR AUTHORS | CME

ORIGINAL ARTICLE
BRIEF REPORT June 4, 2014

Actionable Diagnosis of Neuroleptospirosis by Next-Generation Sequencing

Michael R. Wilson, M.D., Samia N. Naccache, Ph.D., Erik Samayoa, B.S., C.L.S., Mark Biagtan, M.D., Hiba Bashir, M.D., Guixia Yu, B.S., Shahriar M. Salamat, M.D., Ph.D., Sneha Somasekar, B.S., Scot Federman, B.A., Steve Miller, M.D., Ph.D., Robert Sokolic, M.D., Elizabeth Garabedian, R.N., M.S.L.S., Fabio Candotti, M.D., Rebecca H. Buckley, M.D., Kurt D. Reed, M.D., Teresa L. Meyer, R.N., M.S., Christine M. Seroogy, M.D., Renee Galloway, M.P.H., Sheryl L. Henderson, M.D., Ph.D., James E. Gern, M.D., Joseph L. DeRisi, Ph.D., and Charles Y. Chiu, M.D., Ph.D.

June 4, 2014 | DOI: 10.1056/NEJMoa1401268

Abstract | Article | References

More than half the cases of meningoencephalitis remain undiagnosed, despite extensive clinical laboratory testing.[1-4] Because more than 100 different infectious agents can cause encephalitis, establishing a diagnosis with the use of cultures, serologic tests, and pathogen-specific PCR assays can be difficult. Unbiased next-generation sequencing has the potential to revolutionize our ability to discover emerging pathogens, especially newly identified viruses.[5-8] However, the usefulness of next-generation sequencing for the diagnosis of infectious diseases in a clinically relevant timeframe is largely unexplored.[9] We used unbiased next-generation sequencing to identify a treatable, albeit rare, bacterial cause of meningoencephalitis. In this case, the results of next-generation sequencing contributed directly to a dramatic effect on the patient's care, resulting ultimately in a favorable outcome.

CASE REPORT

A 14-year-old boy with severe combined immunodeficiency (SCID) caused by adenosine deaminase deficiency and partial immune reconstitution after he had undergone two haploidentical bone marrow transplantations initially presented to the emergency department in early April 2013 after having had headache and fevers, with temperatures up to 39.4°C, for 6 days (Figure 1A). He was admitted to the hospital and discharged 1 day later after resolution of his fever and headache.

The patient's outpatient medications included monthly infusions of intravenous immune globulin for hypogammaglobulinemia and trimethoprim–sulfamethoxazole or atovaquone for prophylaxis against *Pneumocystis jirovecii* pneumonia. He had no known sick contacts but did have three pet cats. He had gone on a missionary trip to Puerto Rico during the first 2 weeks

FIGURE 1

Clinical Course of the

The New York Times

In a First, Test of DNA Finds Root of Illness

By CARL ZIMMER JUNE 4, 2014

Joshua Osborn, 14, lay in a coma at American Family Children's Hospital in Madison, Wis. For weeks his brain had been swelling with fluid, and a battery of tests had failed to reveal the cause.

The doctors told his parents, Clark and Julie, that they wanted to run one more test with an experimental new technology. Scientists would search Joshua's cerebrospinal fluid for pieces of DNA. Some of them might belong to the pathogen causing his encephalitis.

The Osborns agreed, although they were skeptical that the test would succeed where so many others had failed. But in the first procedure of its kind, researchers at the University of California, San Francisco, managed to pinpoint the cause of Joshua's problem — within 48 hours. He had been infected with an obscure species of bacteria. Once identified, it was eradicated within days.

The case, reported on Wednesday in The New England Journal of Medicine, signals an important advance in the science of diagnosis. For years, scientists have been sequencing DNA to identify pathogens. But until now, the process has been too cumbersome to yield useful information about an individual patient in a life-threatening emergency.

"This is an absolutely great story — it's a tremendous tour de force," said Tom Slezak, the leader of the pathogen informatics team at the Lawrence Livermore National Laboratory, who was not involved in the study.

Mr. Slezak and other experts noted that it would take years of further research before such a test might become approved for regular use. But it could be immensely useful: Not only might it provide speedy diagnoses to critically ill patients, they said, it could lead to more effective treatments for maladies that can be hard to identify, such as Lyme disease.

Diagnosis is a crucial step in medicine, but it can also be the most difficult. Doctors usually must guess the most likely causes of a medical problem and then order individual tests to see which is the right diagnosis.
The guessing game can waste precious time. The causes of some conditions, like encephalitis, can be so hard to diagnose that doctors often end up with no answer at all.

"About 60 percent of the time, we never make a diagnosis" in encephalitis, said Dr. Michael R. Wilson, a neurologist at the University of California, San Francisco, and an author of the new paper. "It's frustrating whenever someone is doing poorly, but it's especially frustrating when we can't even tell the parents what the hell is going on."

https://amplab.cs.berkeley.edu/2014/06/04/snap-helps-save-a-life/

HOME | ARTICLES &

ORIGINAL ARTICLE
BRIEF REPORT

Actionable D… Sequencing

Michael R. Wilson, M.D.
Guixia Yu, B.S., Shahria…
Ph.D., Robert Sokolic, M…
Kurt D. Reed, M.D., Ter…
Henderson, M.D., Ph.D.
June 4, 2014 | DOI: 10.1…

Abstract | Article

More than half the … laboratory testing.… establishing a diag… can be difficult. Un… discover emerging… generation sequer… largely unexplored… bacterial cause of … contributed directly… outcome.

CASE REPORT

A 14-year-old boy … deficiency and par… transplantations in… headache and feve… He was admitted t… his fever and head…

The patient's outpa… immune globulin f… sulfamethoxazole … *jirovecii* pneumoni… cats. He had gone…

…llness

…tal in Madison, …tests had failed to

…e more test with …brospinal fluid for …s encephalitis.

…ucceed where so …s at the University …s problem — …ria. Once

…icine, signals an …e been sequencing …nbersome to yield …rgency.

…aid Tom Slezak, … National

…search before …ensely useful: Not …d, it could lead to …n as Lyme disease.

…cult. Doctors …n order individual

…ions, like …no answer at all.

…litis, said Dr. …rancisco, and an …oorly, but it's …is going on."

https:/…e-a-life/

Pathogen Identification

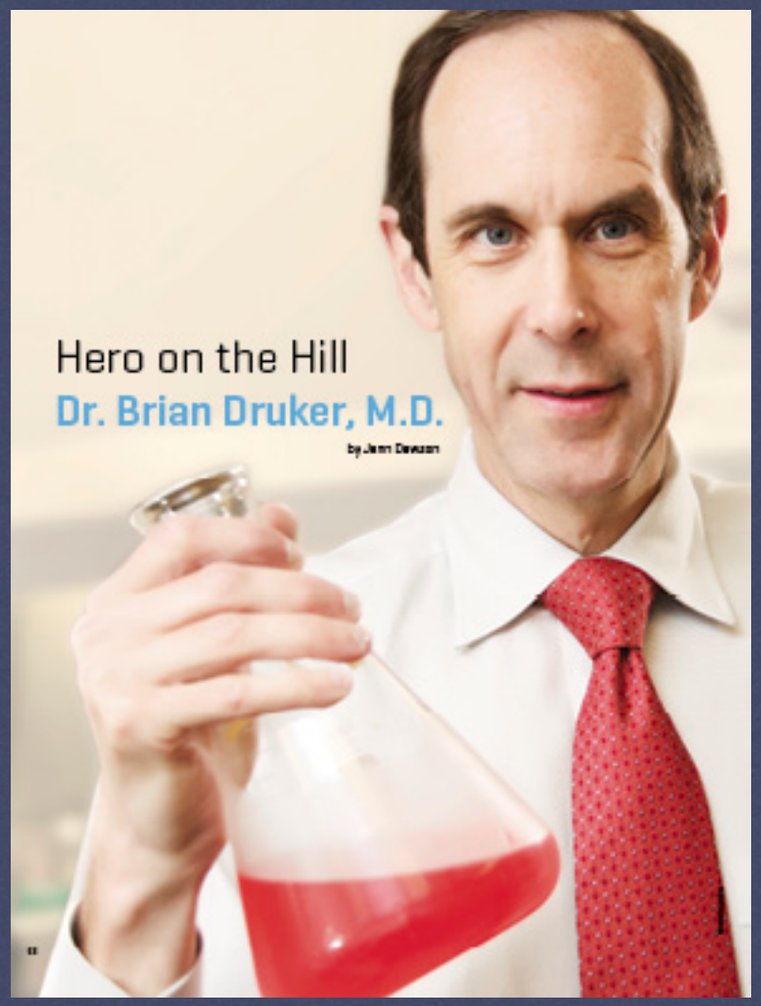
- Encephalitis (brain swelling) 15,000 cases/year in US
  2,000 deaths, >70% under diagnosed

- 0.1% - 1% *all* hospital patients under diagnosed
  35M patients/year in US: 35,000 - 350,000

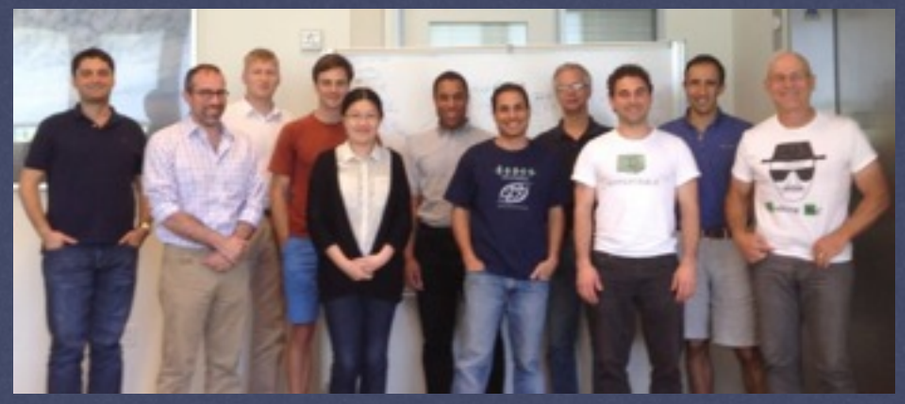
- UCSF use SNAP in Laboratory Test for (priority order)
  1) viral/bacterial cultures ID 2) pneumonia diagnosis
  3) meningitis/encephalitis diagnosis 4) undifferentiated fever ID
  5) diarrheal disease ID

- UCSF distribute SW for free to academic hospitals & public health agencies (Red Cross, CDC, Cal Pub Health, …)

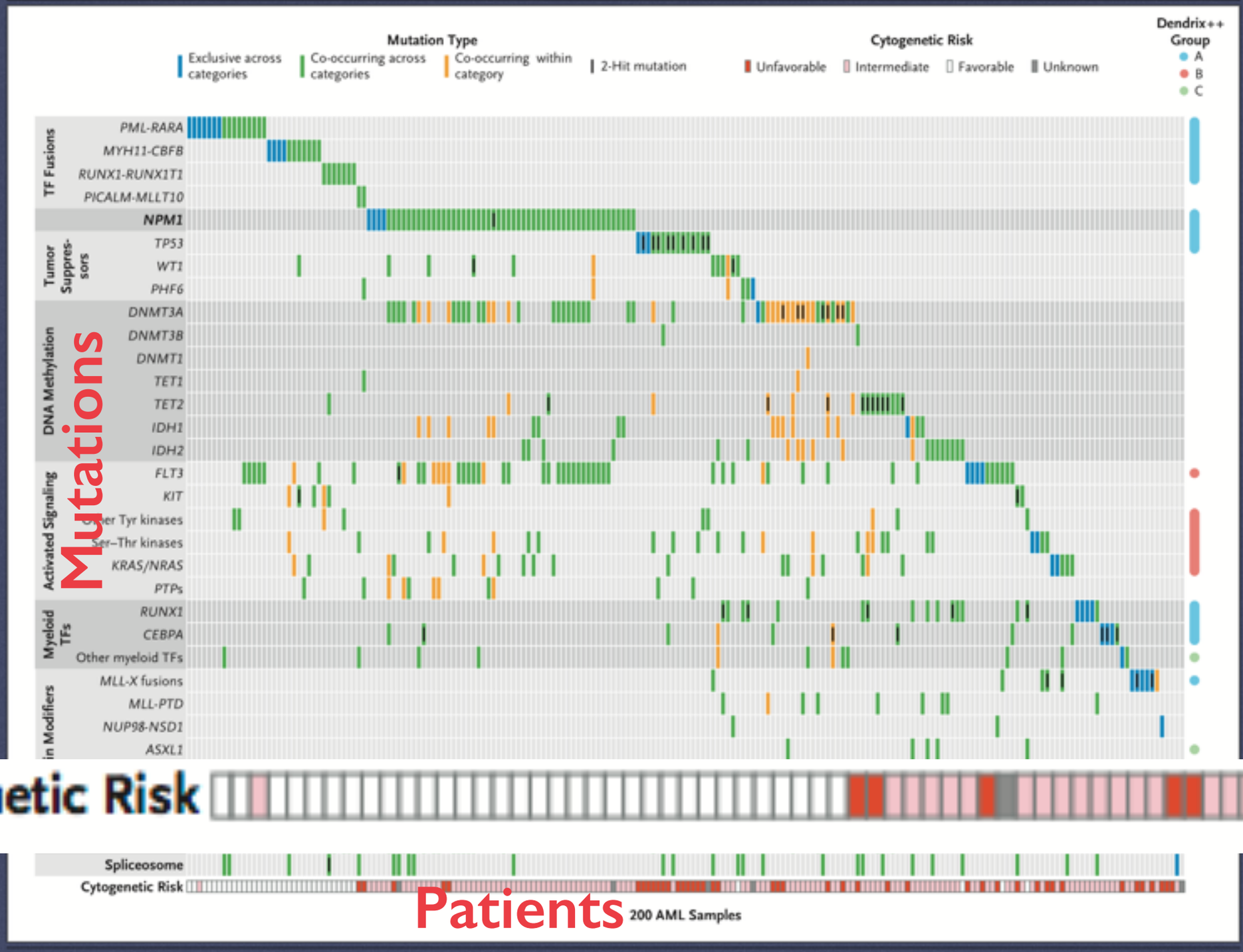
Beat AML



AML
9000

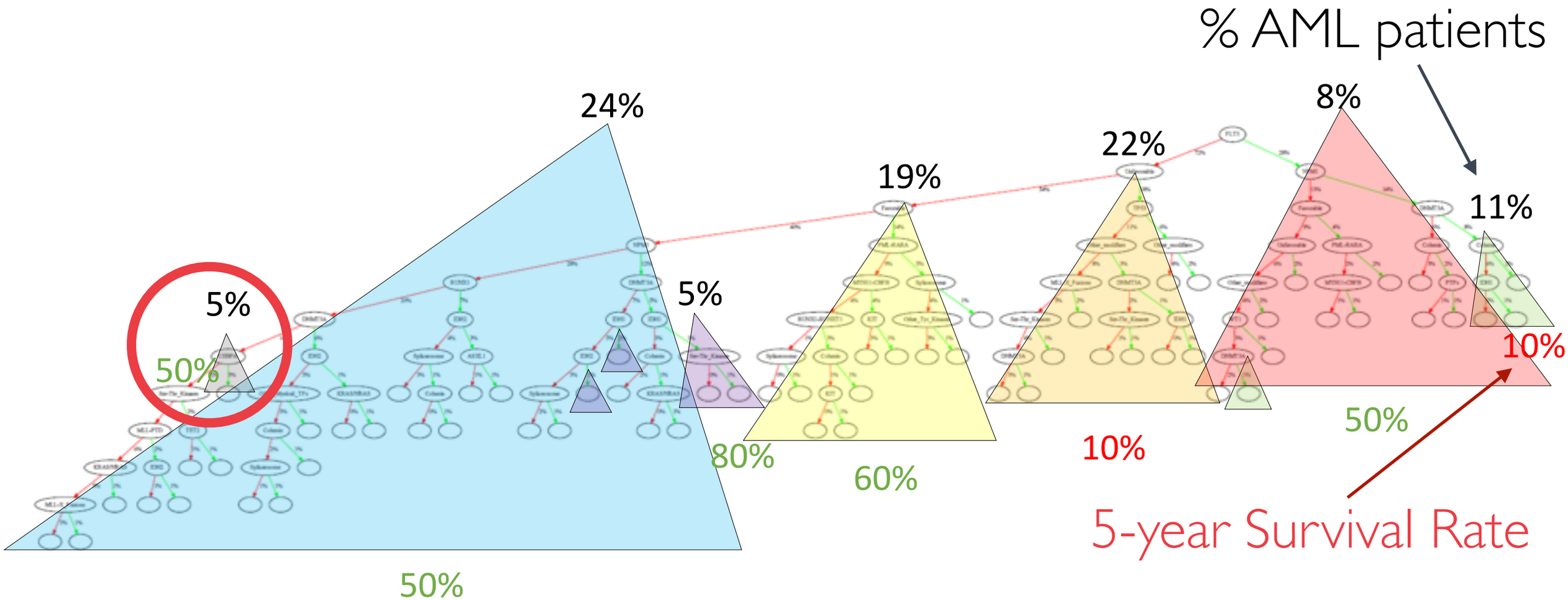


AML Mutations: Doctor's View

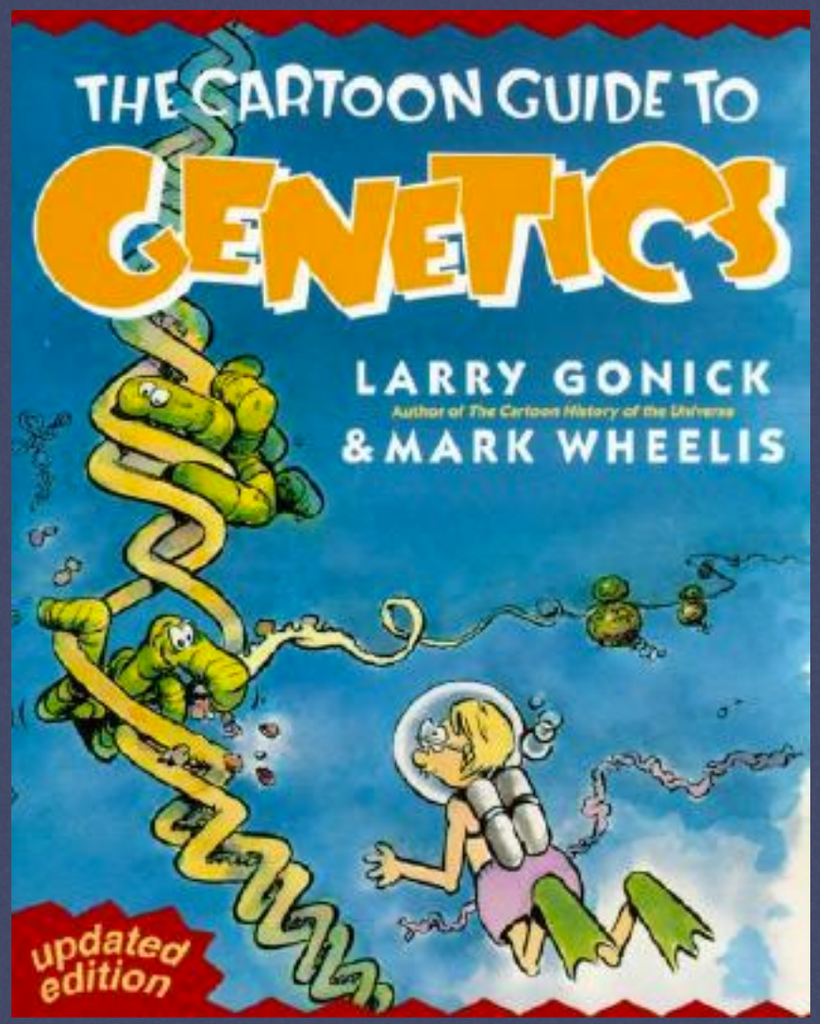


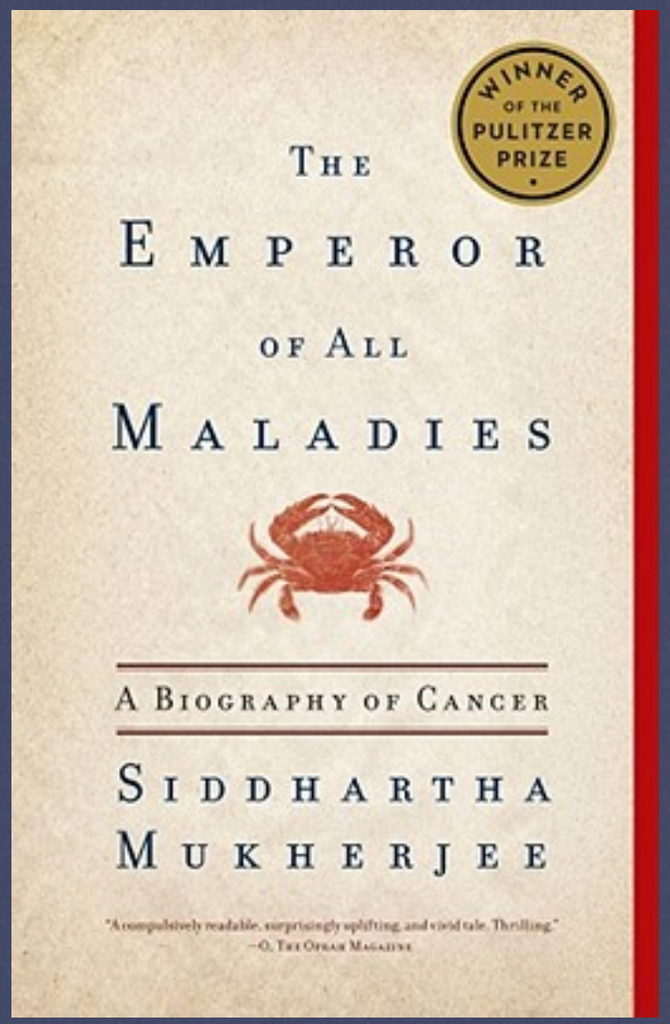
"Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia." *The New England Journal of Medicine* 368.22 (2013): 2059-2074.

AML Mutations: Data Driven



Don't Need MD to Help Fight Cancer





- Free Massive Open Online Courses
  - MITx 7.00 "Introduction to Biology - The Secret of Life"
  - Harvardx PH525x "Data Analysis for Genomics"
  - …

Sequencing is a Puzzle

Shredded Book Analogy

- Dickens 1st printing of A Tale of Two Cities on 5 long spools

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, …

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness,

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, …

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, …

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, …

Slide credit to Michael Schatz http://schatzlab.cshl.edu/

Shredded Book Analogy

- Dickens 1st printing of A Tale of Two Cities on 5 long spools
- Dickens accidentally shreds the first printing!

It was the best of | times, it was the worst | of times, it was the | age of wisdom, it was | the age of foolishness,

It was the best | of times, it was the | worst of times, it was | the age of wisdom, it | was the age of foolishness,

It was the | best of times, it was | the worst of times, it | was the age of wisdom, | it was the age of | foolishness, …

It was | the best of times, it | was the worst of times, | it was the age of | wisdom, it was the age | of foolishness

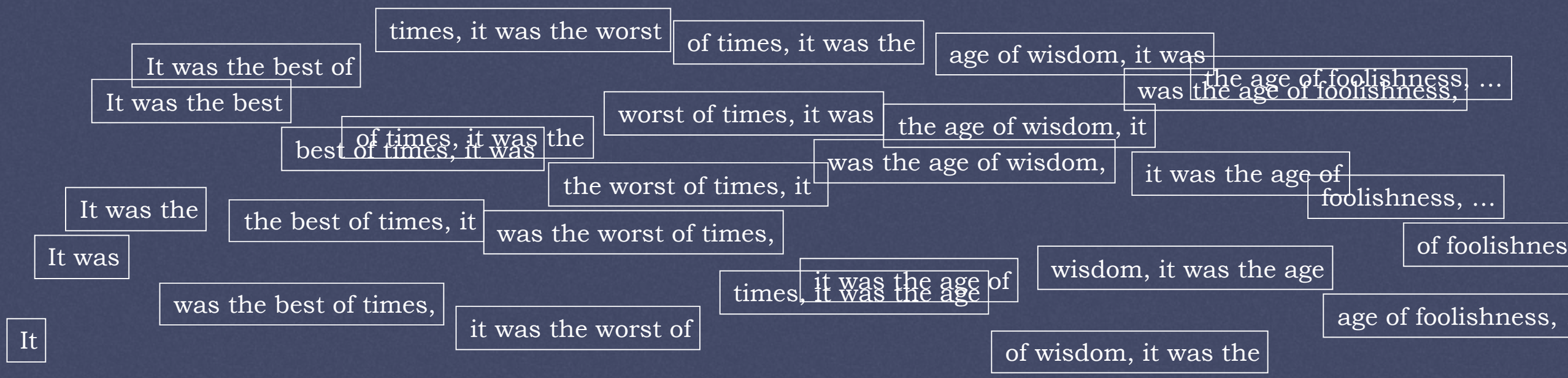
It | was the best of times, | it was the worst of | times, it was the age | of wisdom, it was the | age of foolishness

- How can he reconstruct the text?
  - 5 copies x 138, 656 words / 5 words per fragment = 138k fragments
  - Some fragments are identical

Slide credit to Michael Schatz http://schatzlab.cshl.edu/

Shredded Book Analogy

- Dickens 1st printing of A Tale of Two Cities on 5 long spools
- Dickens accidentally shreds the first printing!

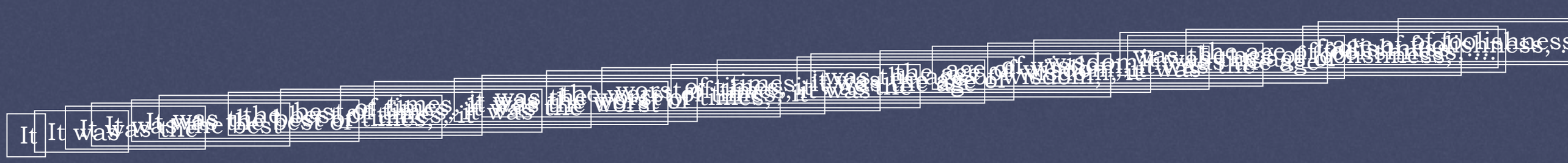


- How can he reconstruct the text?
  - 5 copies x 138, 656 words / 5 words per fragment = 138k fragments
  - Some fragments are identical
  - The short fragments from every copy are mixed together

Slide credit to Michael Schatz http://schatzlab.cshl.edu/

Shredded Book Analogy

- Dickens 1st printing of A Tale of Two Cities on 5 long spools
- Dickens accidentally shreds the first printing!



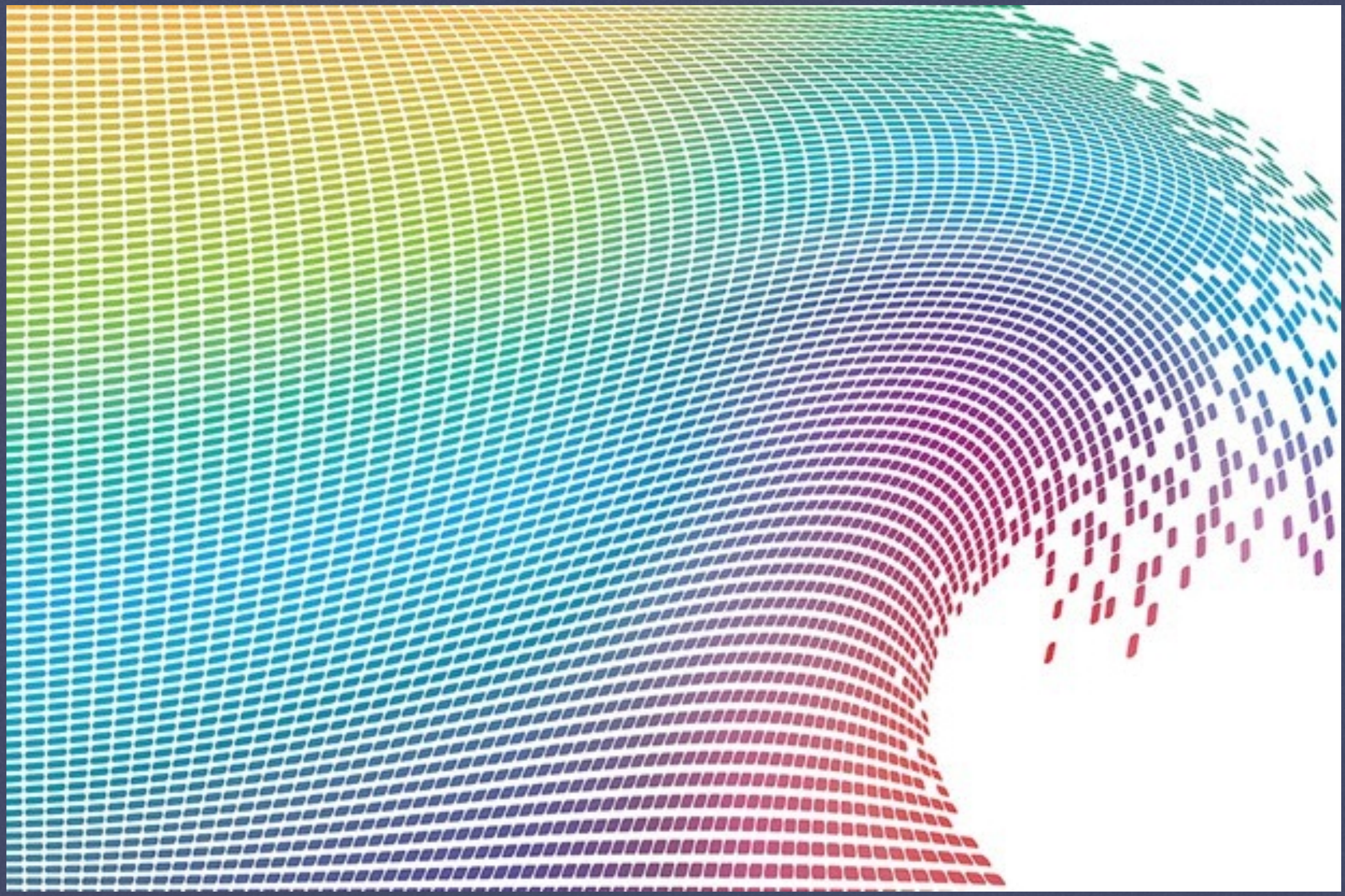
- How can he reconstruct the text?
  - 5 copies x 138, 656 words / 5 words per fragment = 138k fragments
  - Some fragments are identical
  - The short fragments from every copy are mixed together

Slide credit to Michael Schatz http://schatzlab.cshl.edu/

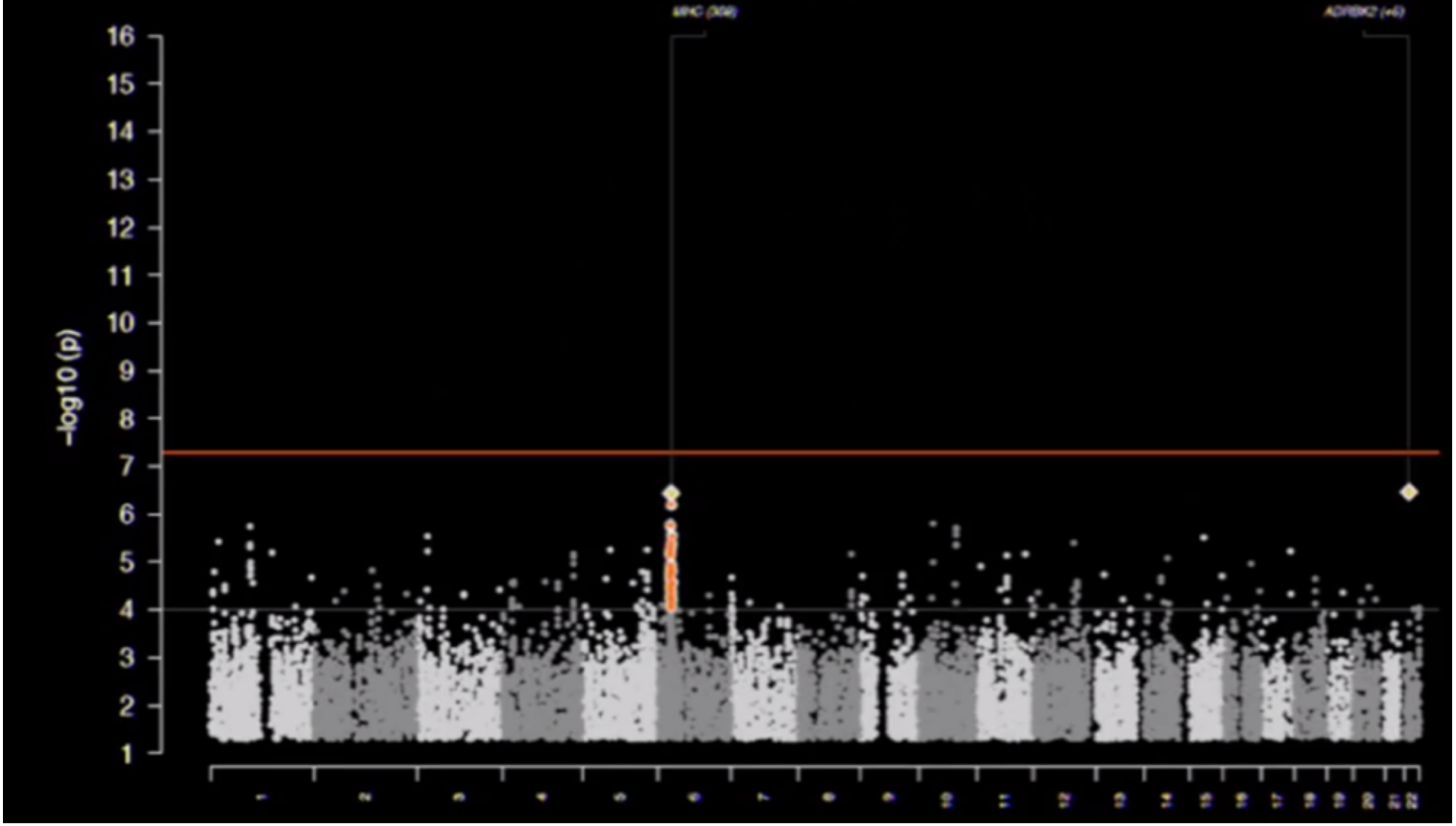
DNA Analogy

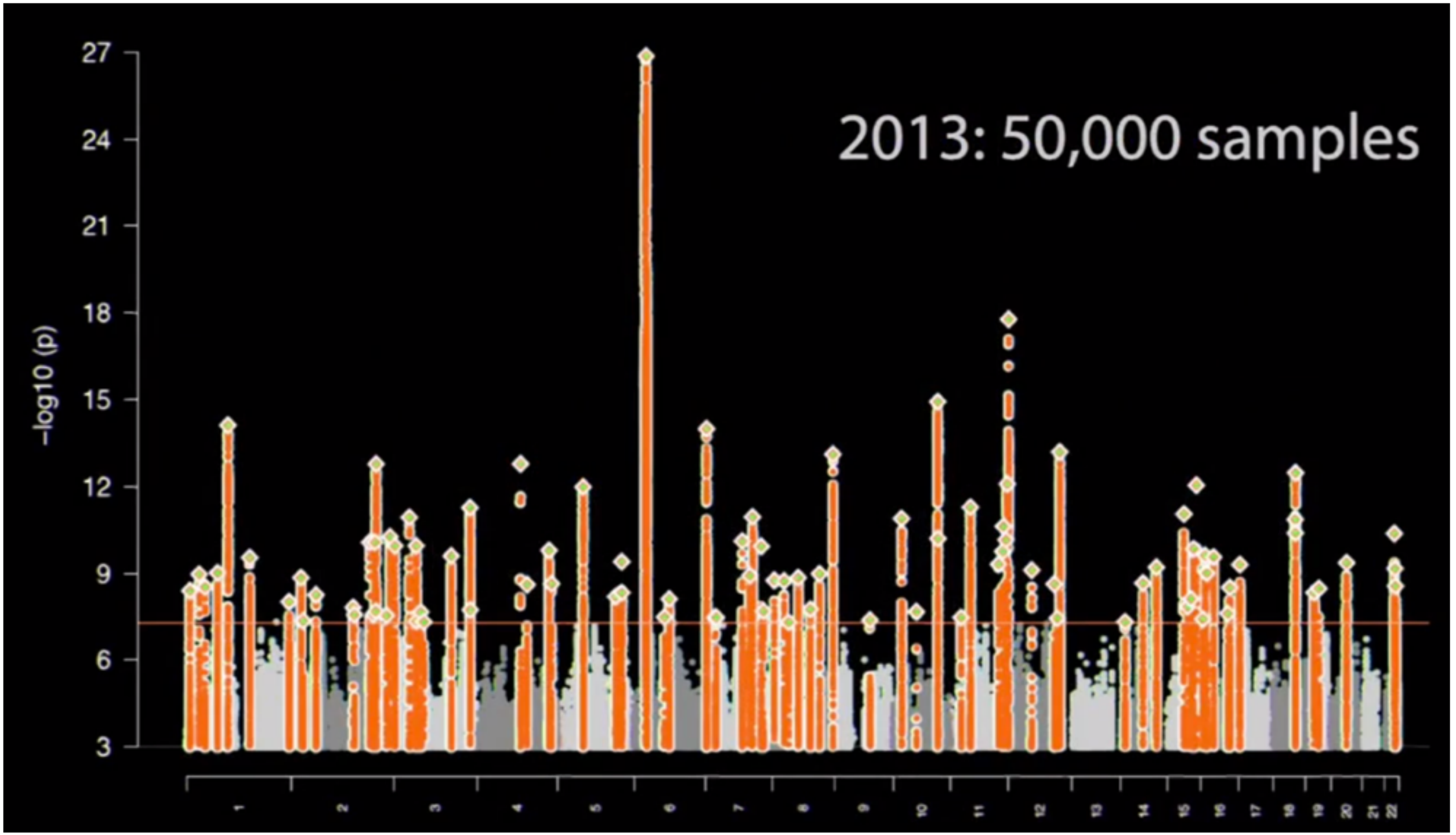
- *T, C, G, A*: 4 letters of DNA Alphabet (*bases, 3B in human genome*)
- *Reads:* DNA Fragments (~150 letters or bases, 20M reads/spool)
- *Coverage*: Number of spools (~60X, 600M reads/genome)
- *Reference:* Generic human sequence (puzzle box cover)
- *Alignment:* Trying to match reads (fragments) to reference
- *Variant Calling:* Finding differences (*variants*) between the reference and a specific DNA

Why Do We Need Big Genetics Data?

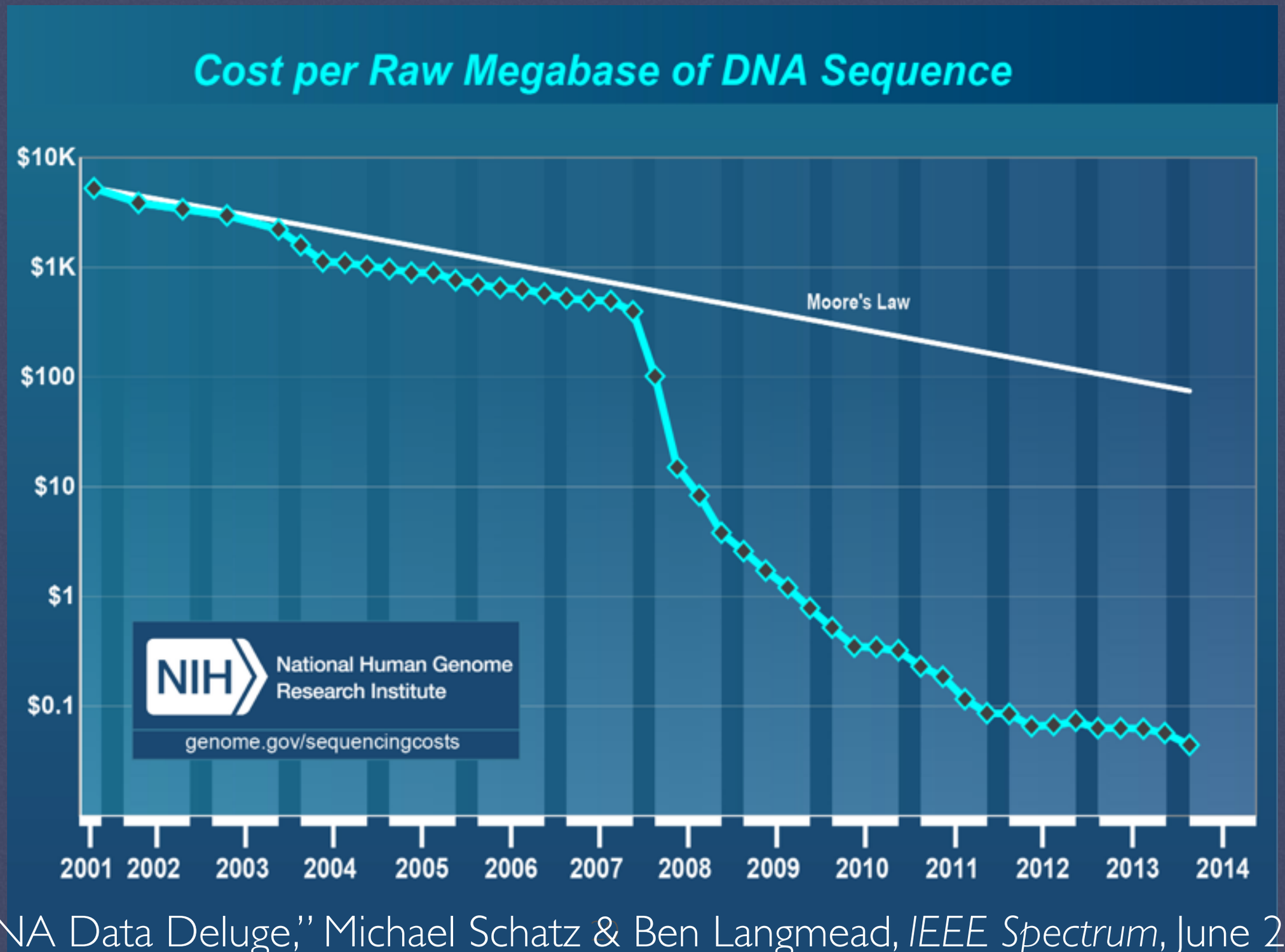


2009: 7000 samples




2013: 50,000 samples
−log10 (p)
27
24
21
18
15
12
9
6
3

Why Will Genetics be Big Data?



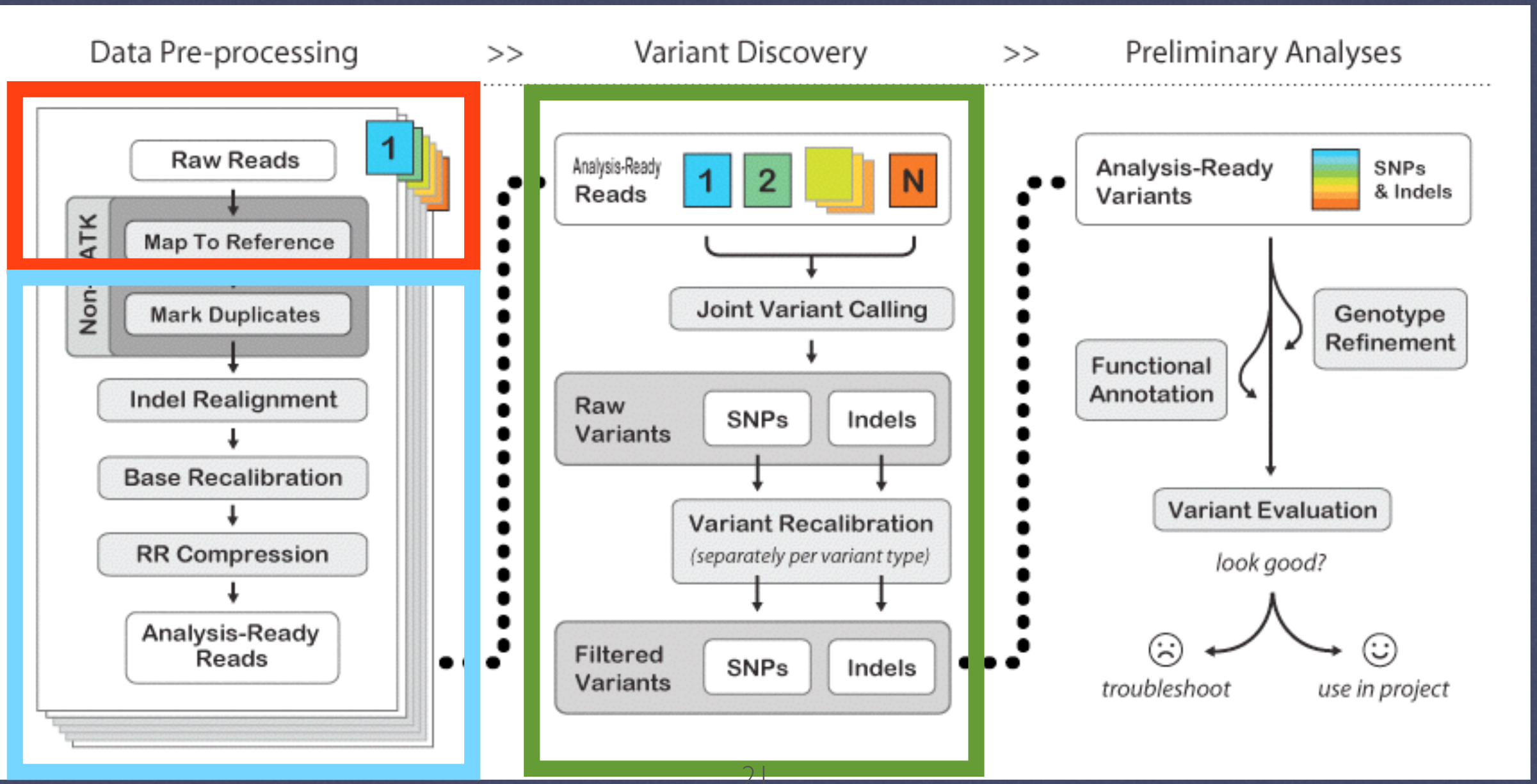
"The DNA Data Deluge," Michael Schatz & Ben Langmead, *IEEE Spectrum*, June 2013

Broad Institute "Best Practices" Pipeline

SNAP
ADAM

SiRen CAGE
Avocado

MLBase
GraphX



Single Whole Genome Data Sizes

	Input	Pipeline Stage	Output
SNAP	3GB Fasta 200GB Fastq	Alignment	100GB BAM
ADAM	250GB BAM	Pre-processing	200GB ADAM
Avocado	200GB ADAM	Variant Calling	10MB ADAM

Variants found at about 1 in 1,000 loci

SNAP: Scalable Nucleotide Alignment Program

- 4 innovations to improve alignment =>
  3X to 10X faster, as accurate

1. A complete index with larger overlapping seeds

2. Faster string matching – Ukkonen: O(n x min(d,k))) vs. O(nd) n is string length, d is edit distance

3. Using index lookup misses as a way of eliminating poor candidates without scoring them

4. For paired-end reads, algorithm based on intersection of seed hit sets that find places where seed hits for both ends of read occur nearby

ADAM Pipeline Stack

RDD	Transform records using Apache Spark Query with SQL using Shark/SparkSQL Graph processing with GraphX Machine learning using MLBase
Record/Split	Schema-driven records w/ Apache Avro Store and retrieve records using Parquet Read BAM Files using Hadoop-BAM
File/Block	Hadoop Distributed Filesystem Local Filesystem
Physical	Commodity Hardware Cloud Systems - Amazon, GCE, Azure

ADAM Implementation

- Work began April 2013
- 20K lines of Scala code
- 100% Apache-licensed open-source
- 17 contributors from Mt. Sinai, GenomeBridge, The Broad Institute, Microsoft Research, BC Cancer Agency, UC Berkeley, & others
- Proofs of concept at The Broad Institute, Duke, Harvard, & UC Santa Cruz

"*Before seeing your data, I would not think sorting could be done that fast.*"

- research scientist at the Broad Institute

ADAM: Genomics Formats and Processing Patterns for Cloud Scale Computing

Matt Massie[1], Frank Austin Nothaft[1], Christopher Hartl[1,2], Christos Kozanitis[1], André Schumacher[3], Anthony D. Joseph[1], and David Patterson[1]

[1]Department of Electrical Engineering and Computer Science, University of California, Berkeley
[2]The Broad Institute of MIT and Harvard
[3]International Computer Science Institute (ICSI), University of California, Berkeley

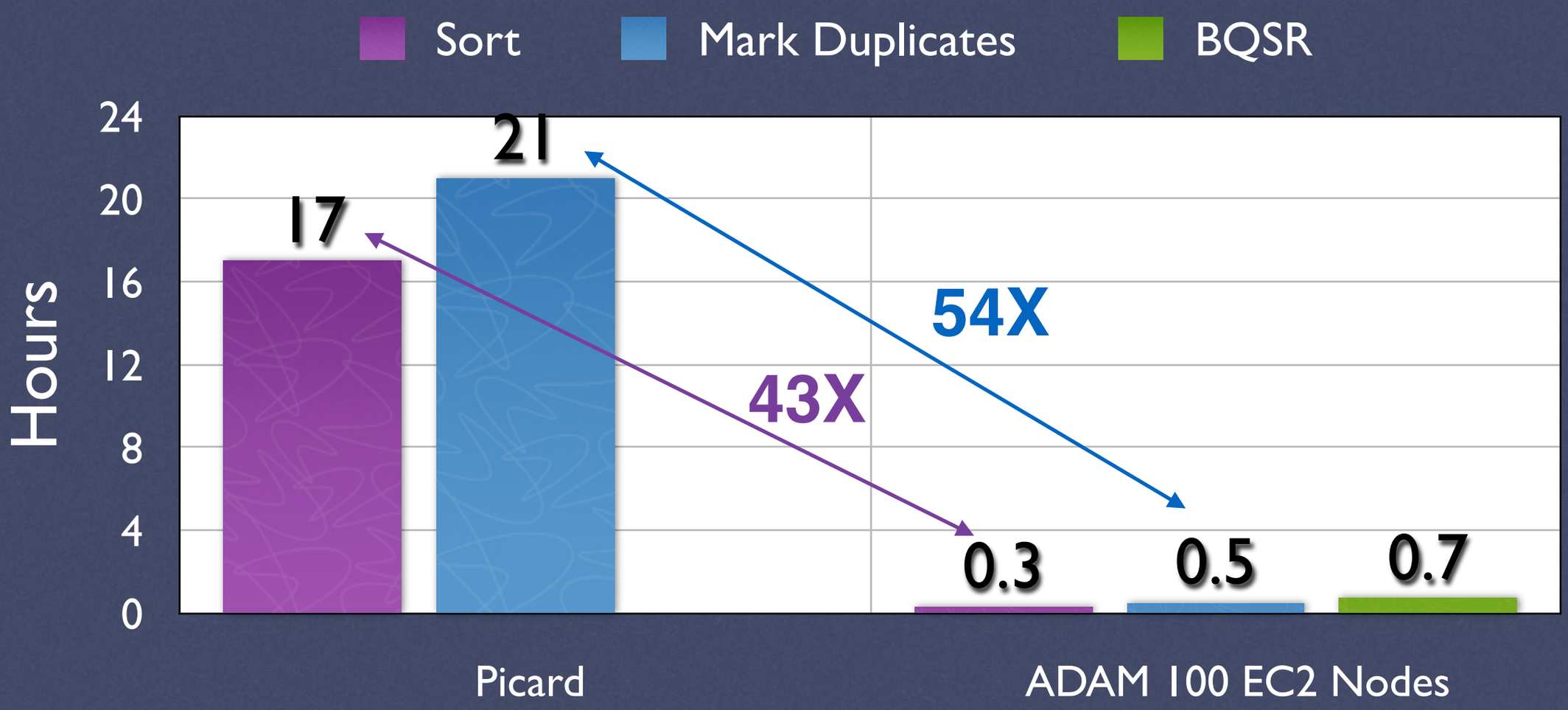
Executive Summary

Current genomics data formats and processing pipelines are not designed to scale well to large datasets. The current Sequence/Binary Alignment/Map (SAM/BAM) formats were intended for single node processing [18]. There have been attempts to

mented in Apache Avro—a cross-platform/language serialization format—they eliminate the need for the development of language-specific libraries for format decoding/encoding, which eliminates the possibility of library incompatibilities.

A key feature of ADAM is that any application that

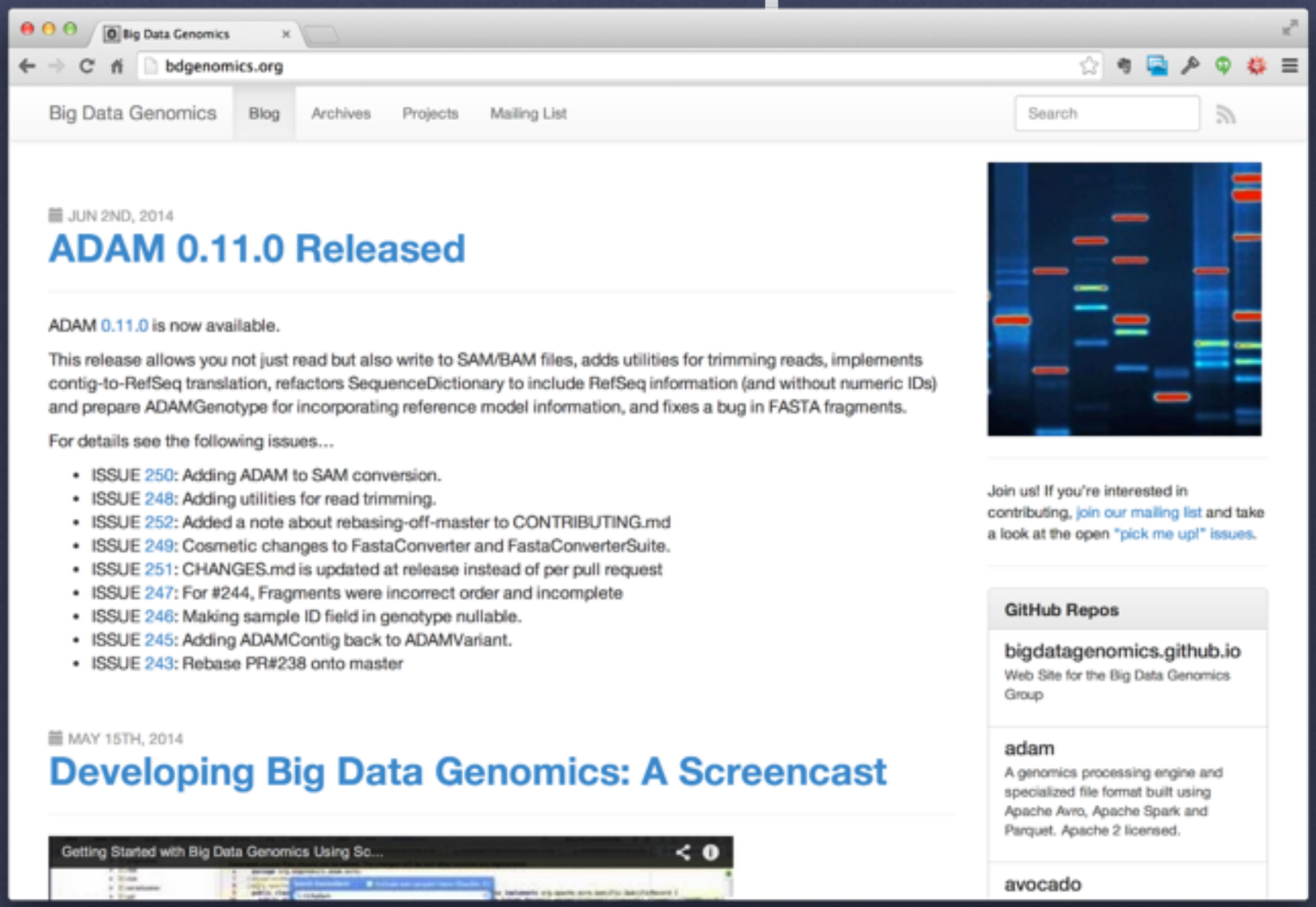
ADAM Performance



1000g NA12878 Whole Genome, 60x Coverage

For comparison, Bina Technologies quotes .94 hours for BQSR at only 37x coverage

You can help now!



- See to bdgenomics.org: ADAM mailing list, IRC channel, and "pick me up!" issues:

How Can You Get Paid To Help?

- If want to help make rich people richer,
  join a company like Lehman Brothers
- If want to help save lives of very sick people,
  join UC Berkeley
- AMPLab has 2 openings for programmers
- Work with bright people on university campus
  at top rated department in computer systems
- Open source
- Highly visible in Spark community
- Competitive salaries (including signing bonuses)
- Won't go out of business
- Contact me pattrsn@berkeley.edu or
  Matt Massie massie@berkeley.edu

Matt Massie

Thanks to Collaborators

- Microsoft Research: Bill Bolosky, Ravi Padya, Jeremy Ellson, David Heckerman
- Oregon Health Sciences University: Brian Druker, Jeff Tyner, Mark Loriaux, Cristina Tognon
- UC Berkeley: Anthony Joseph, Christos Kozanitis, Matt Massie, Frank Nothaft, Jonathan Terhorst, Nir Yosef
- UC Santa Cruz: David Hausller, Benedict Paten
- UC San Francisco: Taylor Sittler, Charles Chiu