Microsoft Research Faculty Summit 2014

15TH ANNUAL

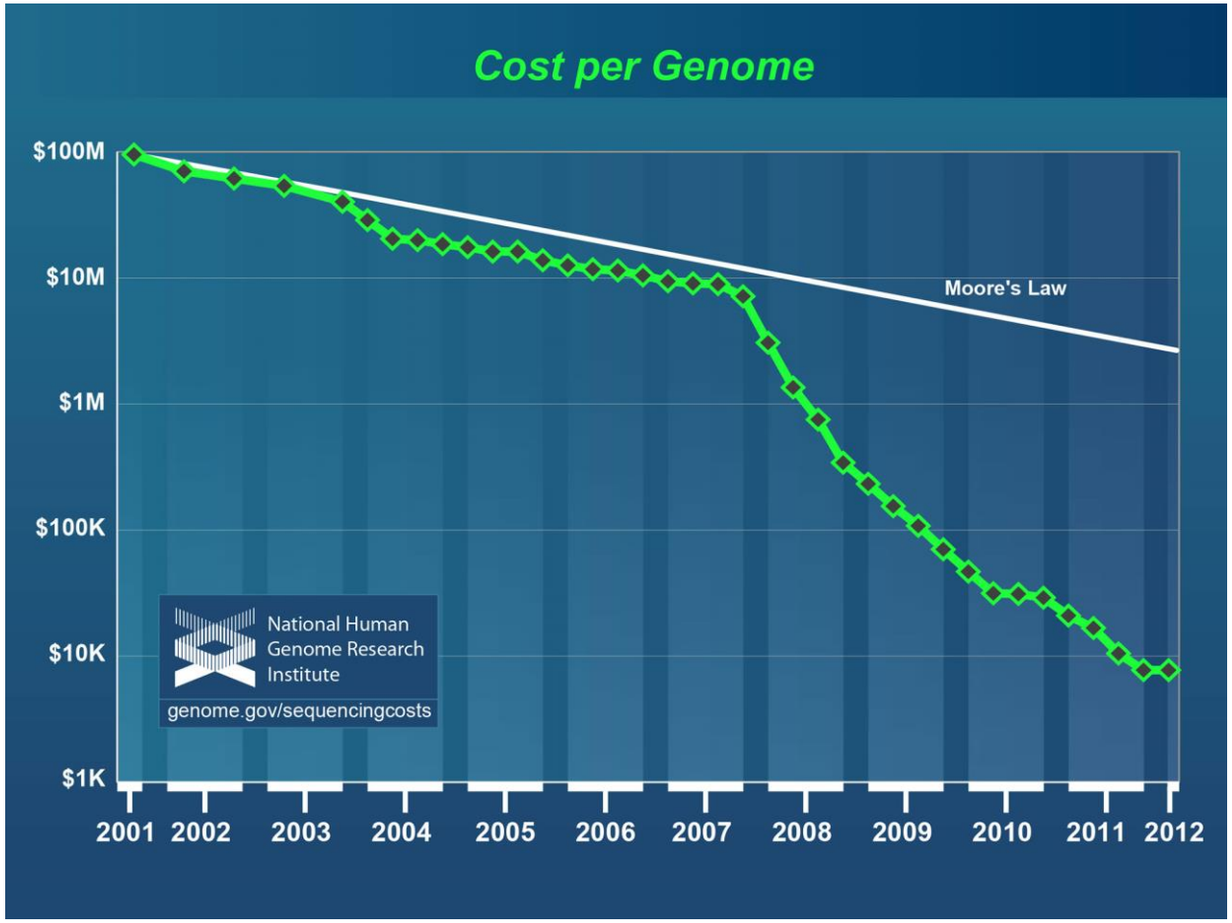
Microsoft Research
Faculty Summit
2014 15TH ANNUAL

Genomics in the Cloud

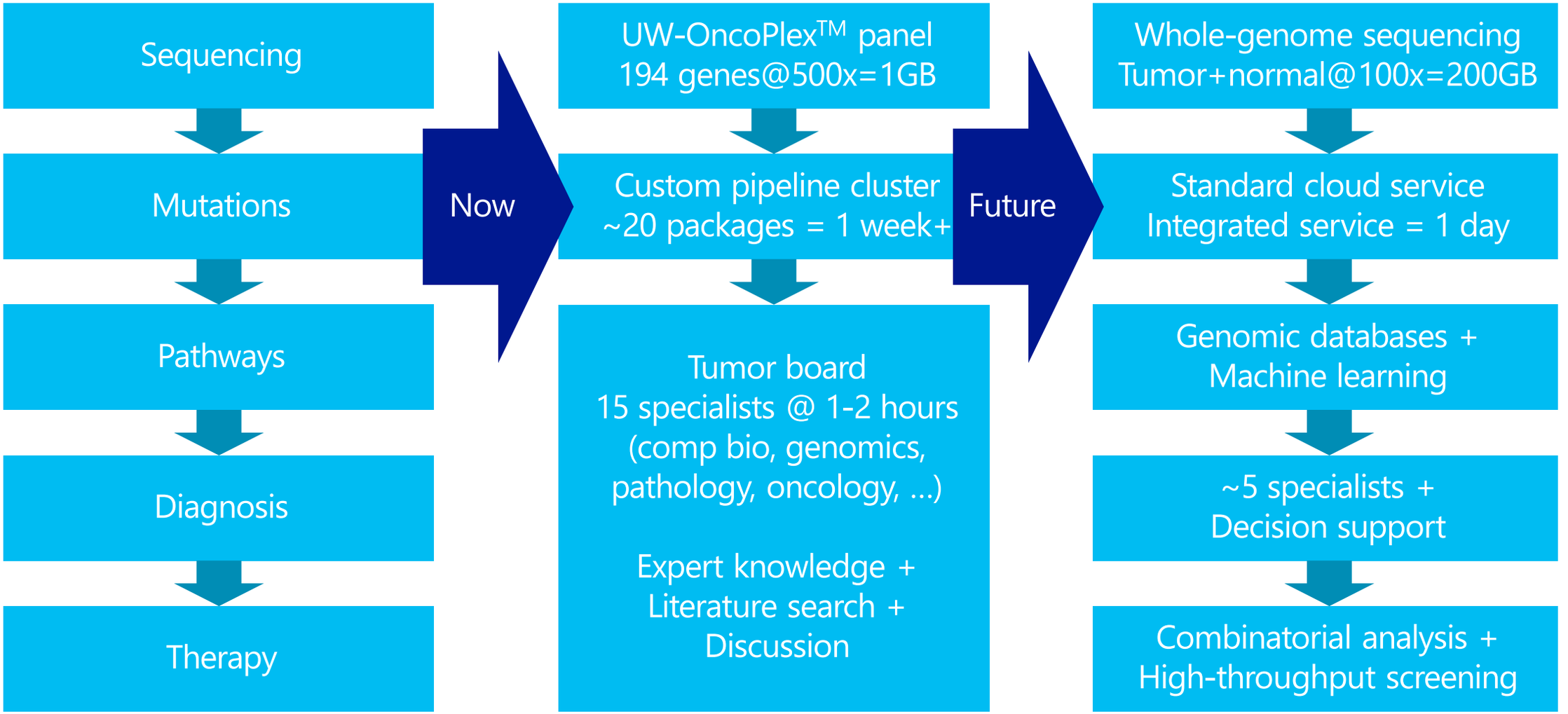
Ravi Pandya | MSR eScience

Microsoft

Genome sequencing cost



Computational medicine for cancer



Genome sequencing

sample genome
2 x 3B nucleotides
(A, G, C, T)

replication
fragmenting
sequencing

raw reads (300 Gb)

1B x 100bp
+ quality scores

alignment

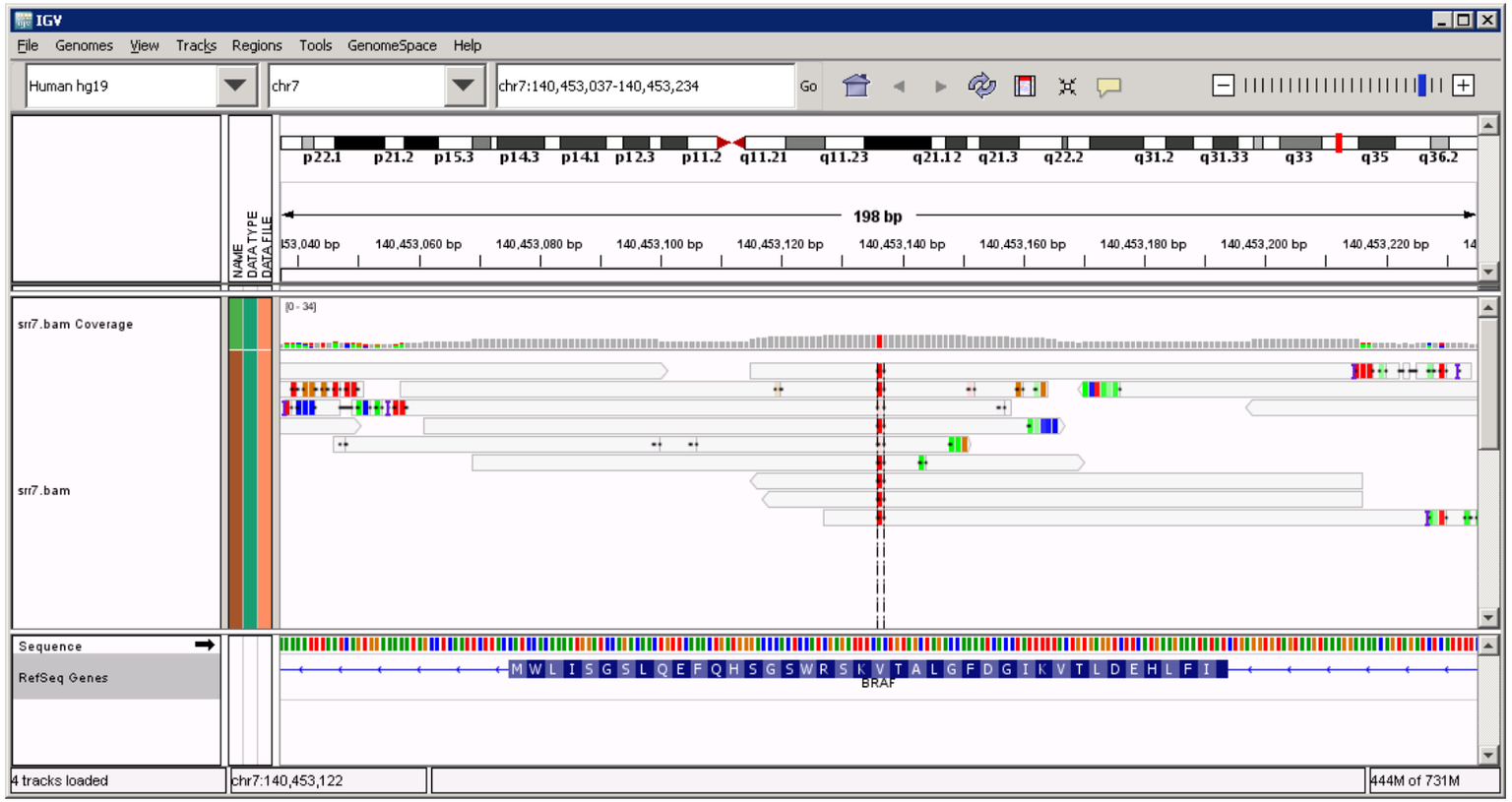
reference genome

variant calls (100 MB)

variant calling

aligned reads (300 GB)

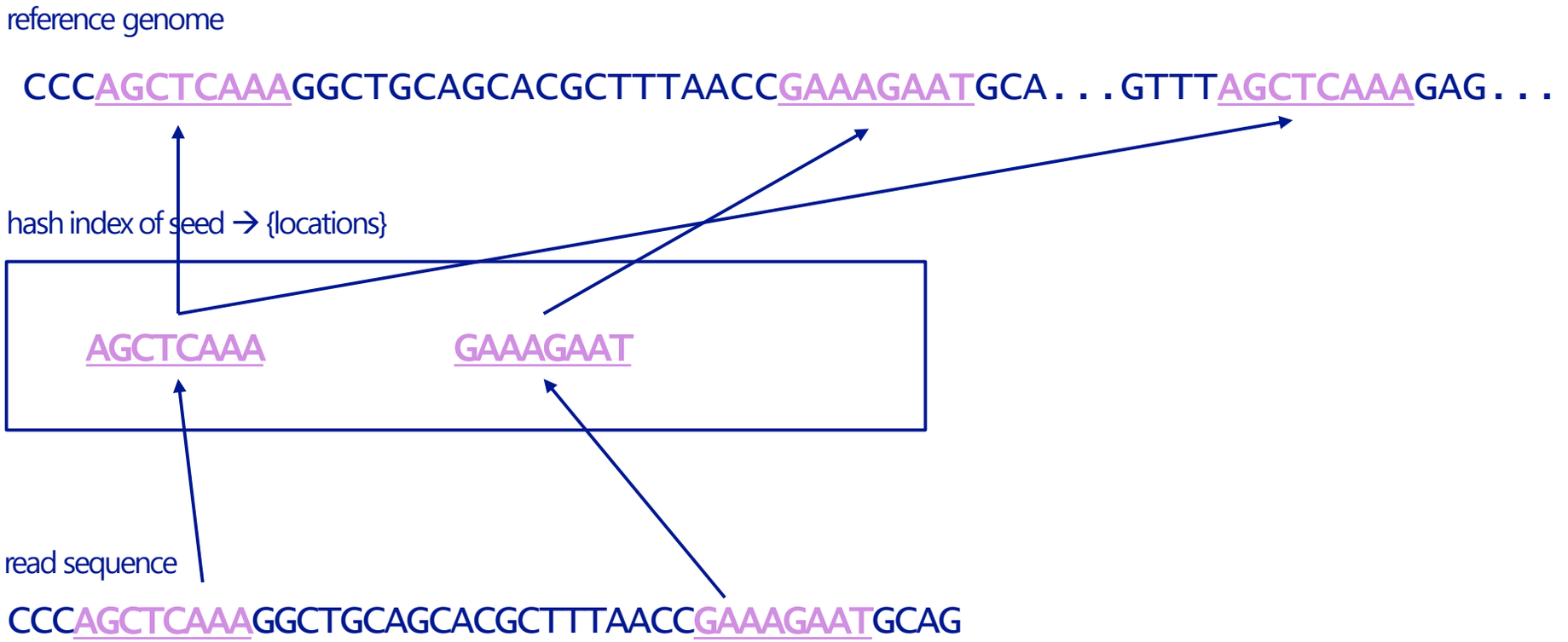
SNAP: Sequence alignment



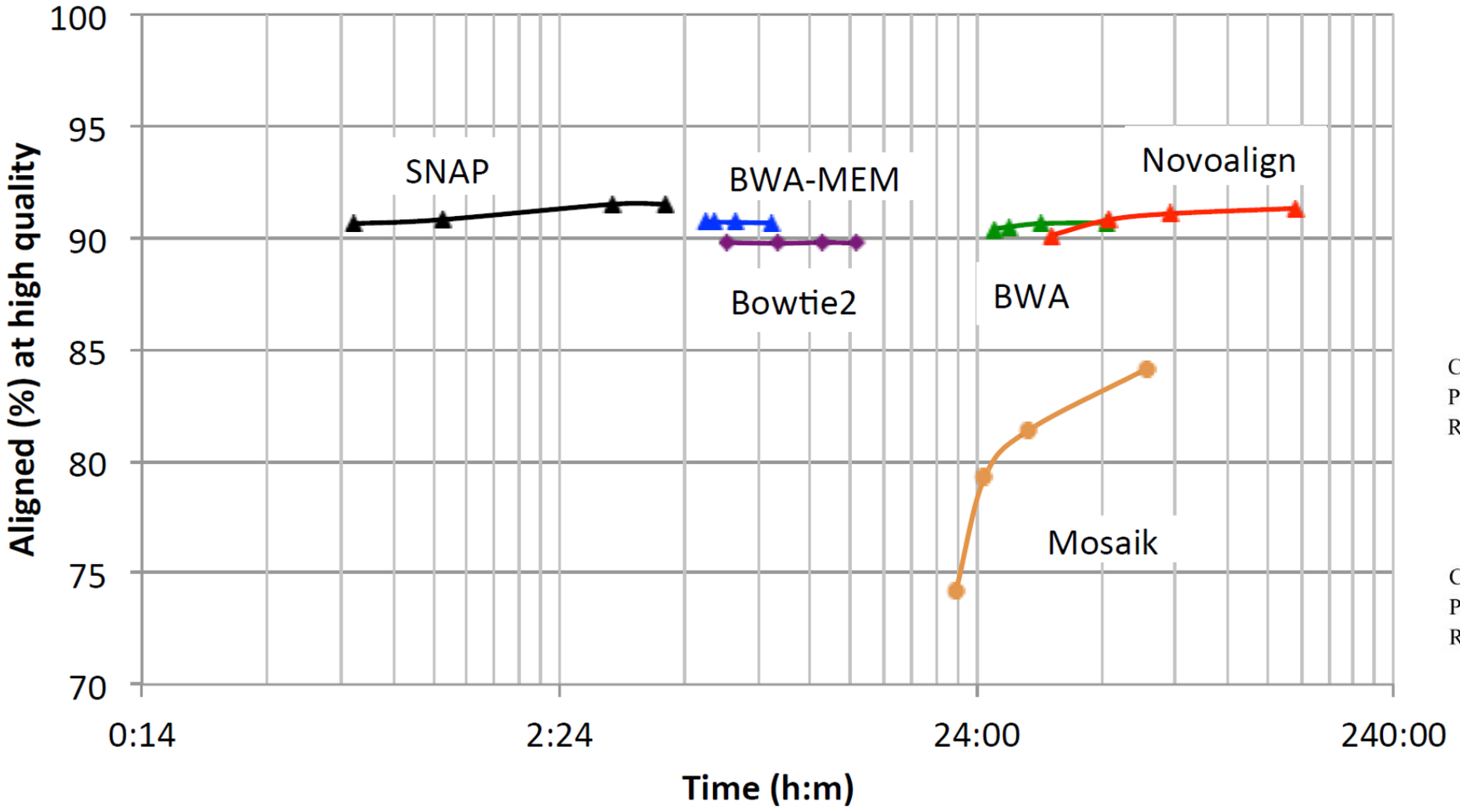
Bill Bolosky, Ravi Pandya (MSR); Matei Zaharia, Taylor Sittler, Kristal Curtis (UC Berkeley)

SNAP algorithm

Build index → Lookup seeds → Map locations → Score matches



SNAP performance



BWA		SNPs	SNAP	
Calls	3,785,822	Intersection: 3,683,920	Calls	3,826,960
Precision	99.9%		Precision	99.8%
Recall	98.3%		Recall	97.5%

BWA		Indels	SNAP	
Calls	772,957	Intersection: 732,867	Calls	796,185
Precision	99.7%		Precision	99.6%
Recall	91.4%		Recall	91.9%

NA12878 250bp reads

BWA-MEM		SNPs	SNAP	
Calls	3,722,901	Intersection: 3,641,350	Calls	3,694,426
Precision	99.9%		Precision	99.9%
Recall	93.3%		Recall	93.1%

BWA-MEM		Indels	SNAP	
Calls	898,943	Intersection: 810,284	Calls	830,122
Precision	99.7%		Precision	99.7%
Recall	91.6%		Recall	90.1%

SNAP applications

SECTIONS HOME SEARCH The New York Times SUBSCRIBE LOG IN

HEALTH

COMMENT

In First, Quick DNA Test Diagnoses a Boy's Illness

By CARL ZIMMER JUNE 4, 2014

EMAIL FACEBOOK TWITTER SAVE MORE

Top Stories
This article and

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

The NEW ENGLAND JOURNAL of MEDICINE

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

ORIGINAL ARTICLE

BRIEF REPORT

Actionable Diagnosis of Neuroleptospirosis by Next-Generation Sequencing

CSH PRESS GENOME RESEARCH epicentre an Illumina company

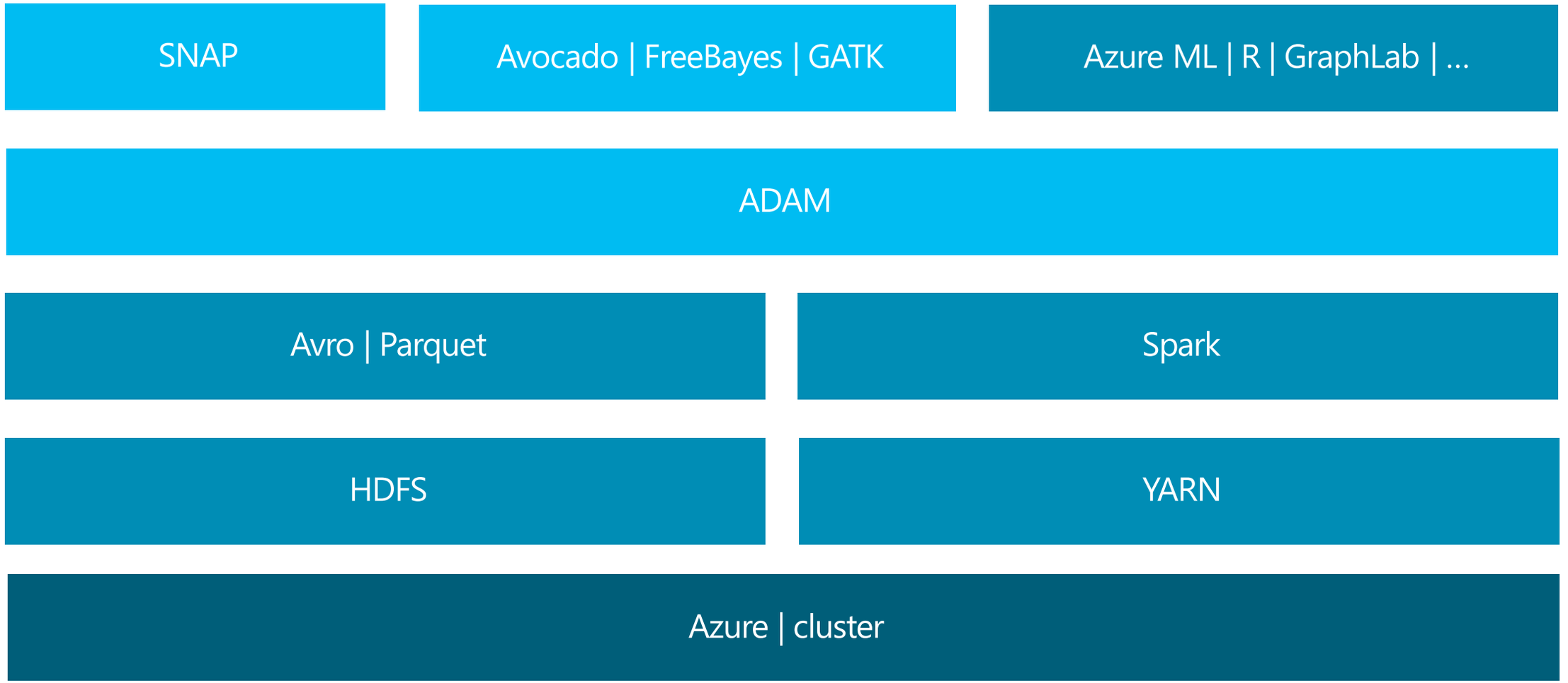
HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTHOR INFO | CONTAC

A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples

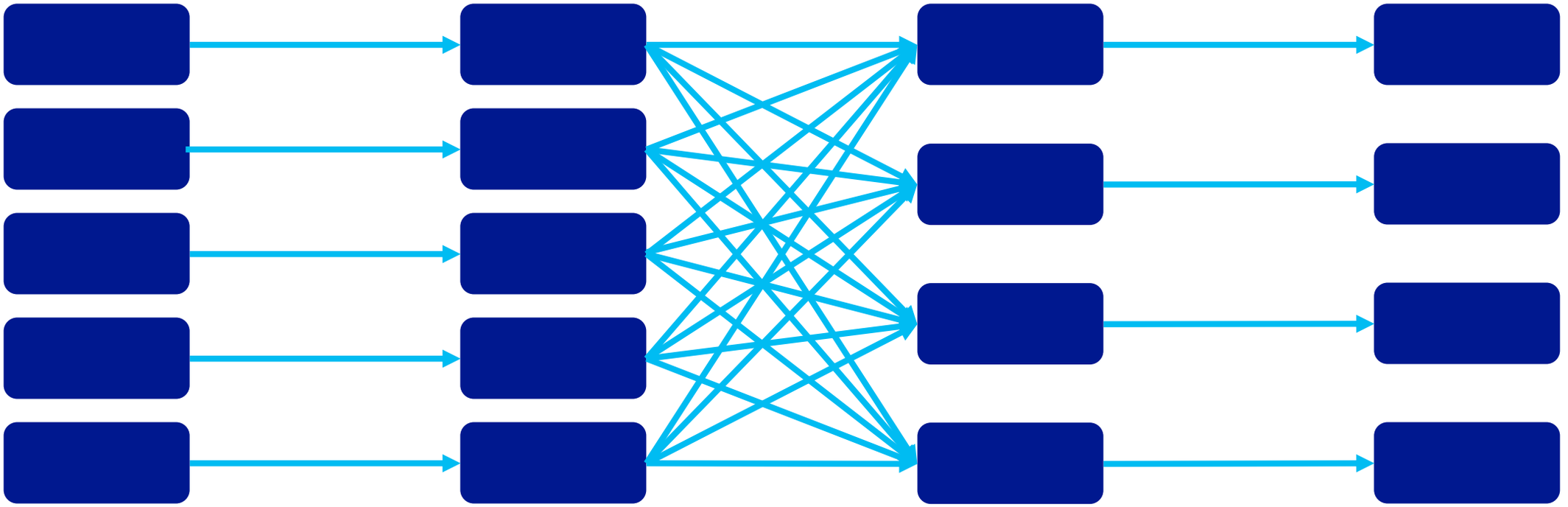
"This analysis of DNA sequences required just **96 minutes**. A similar analysis conducted with the use of previous generations of computational software on the same hardware platform would have taken **24 hours or more** to complete, Chiu said."

from complex metagenomic NGS data generated from clinical samples, and demonstrate use of the pipeline in the analysis of 237 clinical samples comprising more than 1.1 billion sequences. Deployable on both cloud-based and standalone servers, SURPI leverages two state-of-the-art aligners for accelerated analyses, SNAP and RAPSearch, which are as accurate as existing bioinformatics tools but orders of magnitude faster in performance. In *fast* mode, SURPI detects viruses and bacteria by scanning data sets of 7–500 million reads in 11 min to 5 h, while in *comprehensive* mode, all known microorganisms are identified, followed by de novo assembly and protein homology searches for divergent viruses in 50 min to 16 h. SURPI has also directly contributed to real-time microbial diagnosis in

ADAM: Cloud genomics framework



ADAM pipeline



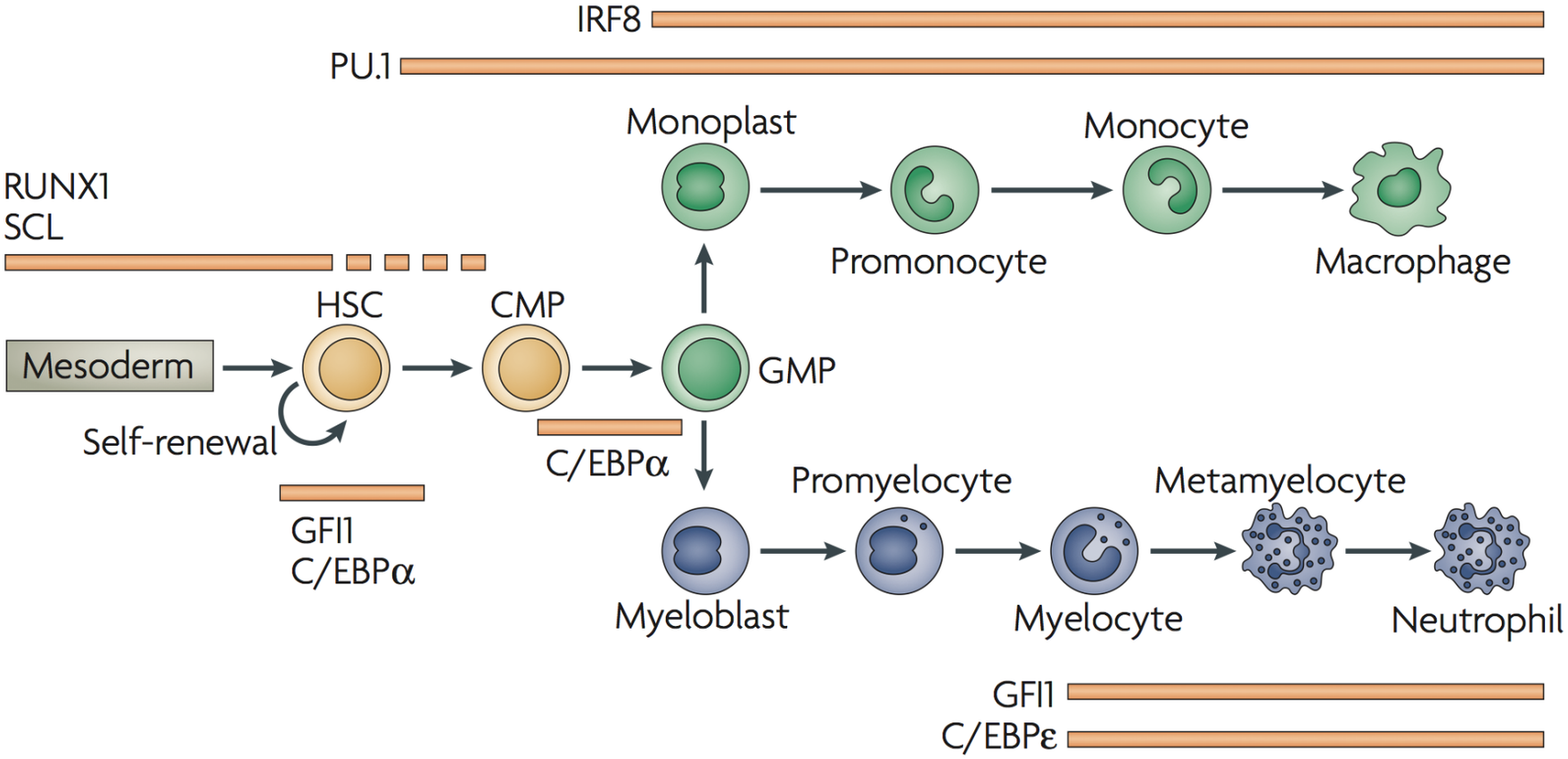
Unaligned reads (sequencer)

Align to reference (SNAP)

Sort, index, dups, BQSR (ADAM)

Variant calls (Avocado/FreeBayes)

BeatAML



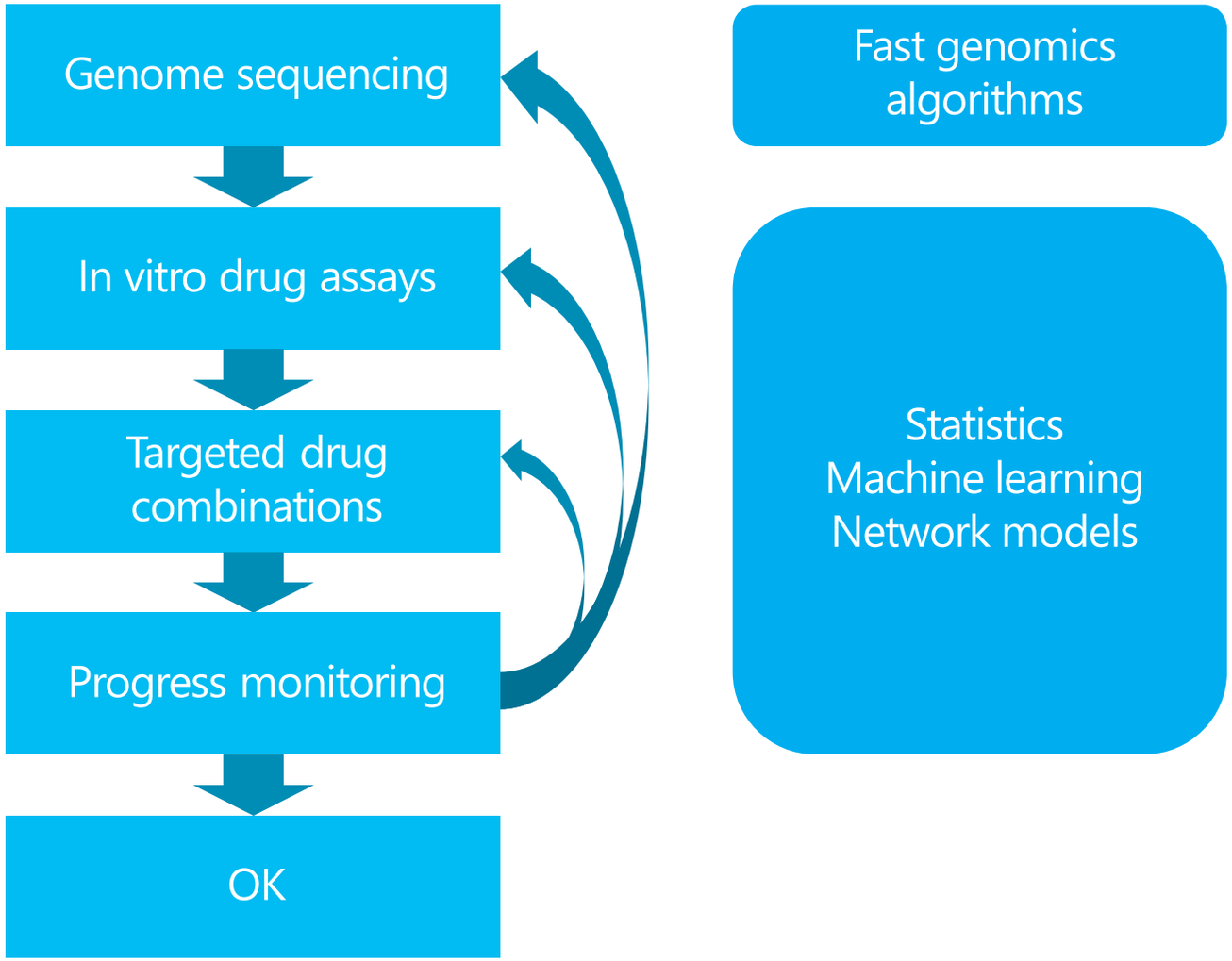
Rosenbauer Nature 2007

Microsoft Research (eScience, Systems, NLP); UC Berkeley (AMPLab); OHSU (Druker Lab)

BeatAML patient timeline

Days	1	2	3	4	5	6	7	8	9	10	11	12	13
Emer-gency Room Data	Bone Marrow Biopsy	Test for APL		Patients who do not have APL (and are under 70) get standard chemotherapy									
		BMB Drug panels: SiRNA, single agents, combination agents (384/panel)							20 ml Blood Sample	Blood drug panel for patients over 70: ~12 combination agents			Show Results to Oncologist (& Patient) for consideration
		BMB Quantum Dot test of 13 combinations							Predict per patient combinations based on outcome of panels, quantum dot tests, & genetics	Blood Quantum Dot test of 13 combinations			
		Sequencing Assay of 76 genes (e.g., Gene Trails)											
		Long term sequencing (whole genome or high coverage exhome); takes 2(?) months ….											

Computational medicine



Save the planet and return
your name badge before you
leave (on Tuesday)