Big Data Challenge
Lecture 10

Amit Ghosh
IIT Kharagpur

http://www.energy.iitkgp.ac.in/~amitghosh/

Sequencing in 2001

PRODUCTION

Sample Preparations

35 People

3-4 weeks

SEQUENCING

74x Capillary sequencers

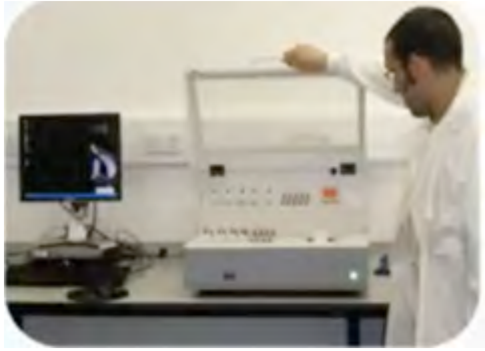
10 People

15-40 runs per day

1-2Mb per instrument per day

120Mb total capacity per day

Sequencing in 2007



PRODUCTION

1x Cluster Station

1 Person

1 Day

SEQUENCING

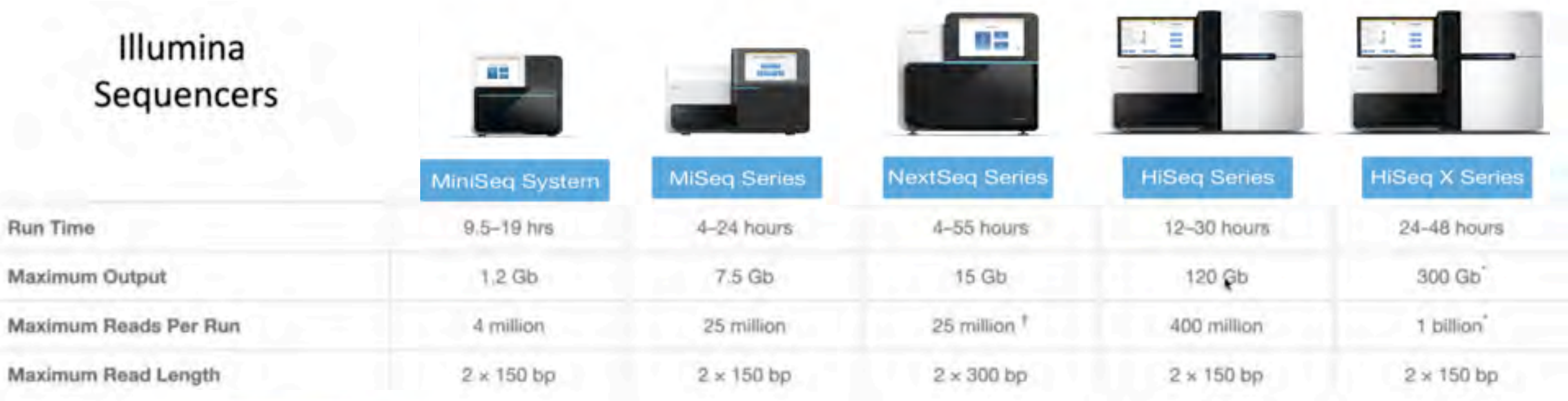
1x Genome Analyzer

Same person as above

1 run per 3-5 days

0.5Gb per instrument per day

Sequencing Now



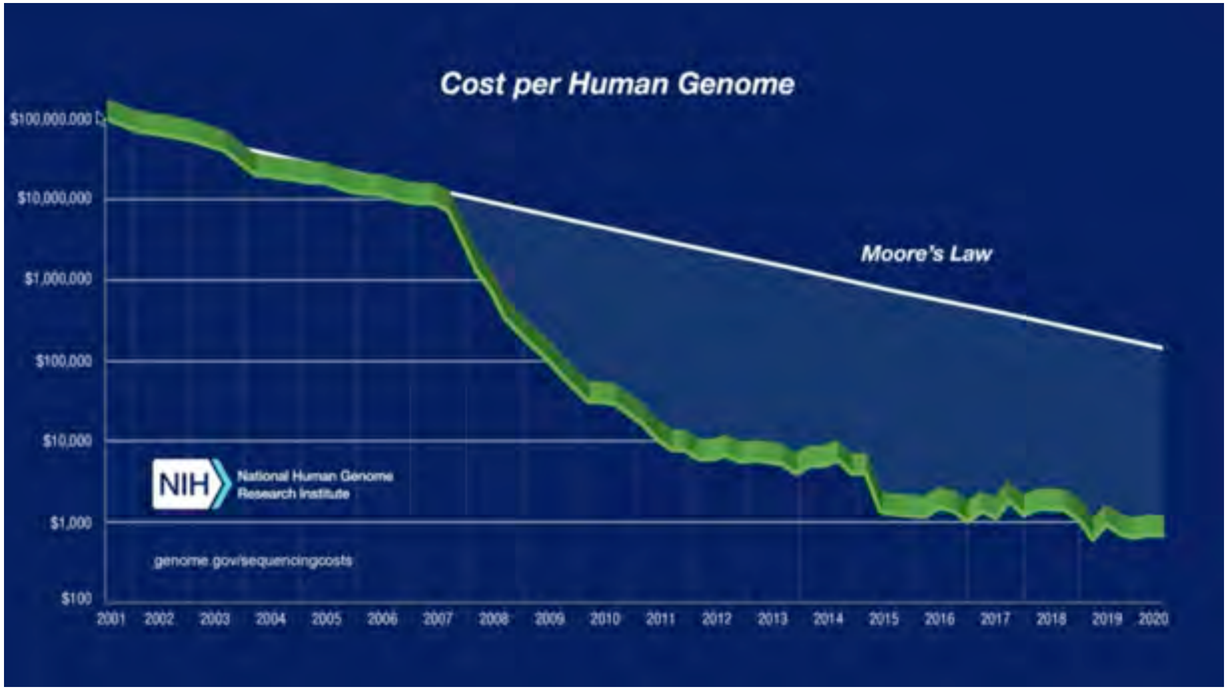
Illumina Sequencers

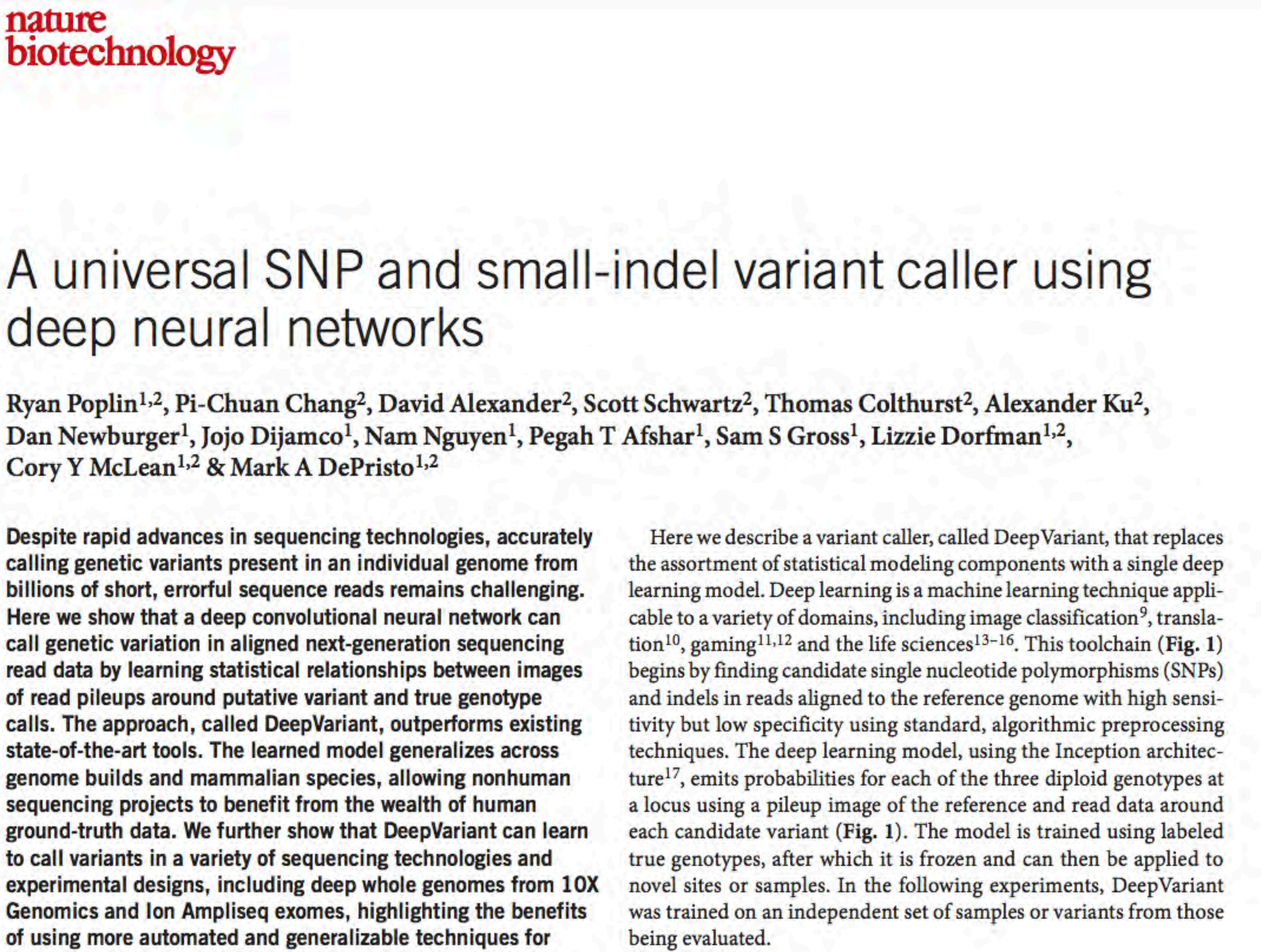
	MiniSeq System	MiSeq Series	NextSeq Series	HiSeq Series	HiSeq X Series
Run Time	9.5–19 hrs	4–24 hours	4–55 hours	12–30 hours	24–48 hours
Maximum Output*	1.2 Gb	7.5 Gb	15 Gb	120 Gb	300 Gb*
Maximum Reads Per Run	4 million	25 million	25 million†	400 million	1 billion*
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp

Up to 6TB and 20B reads in less than 2 days

Functions

Lets look into another example





nature
biotechnology

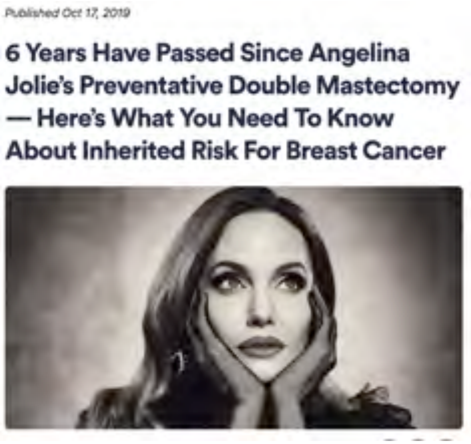
A universal SNP and small-indel variant caller using deep neural networks

Ryan Poplin[1,2], Pi-Chuan Chang[2], David Alexander[2], Scott Schwartz[2], Thomas Colthurst[2], Alexander Ku[2], Dan Newburger[1], Jojo Dijamco[1], Nam Nguyen[1], Pegah T Afshar[1], Sam S Gross[1], Lizzie Dorfman[1,2], Cory Y McLean[1,2] & Mark A DePristo[1,2]

Despite rapid advances in sequencing technologies, accurately calling genetic variants present in an individual genome from billions of short, errorful sequence reads remains challenging. Here we show that a deep convolutional neural network can call genetic variation in aligned next-generation sequencing read data by learning statistical relationships between images of read pileups around putative variant and true genotype calls. The approach, called DeepVariant, outperforms existing state-of-the-art tools. The learned model generalizes across genome builds and mammalian species, allowing nonhuman sequencing projects to benefit from the wealth of human ground-truth data. We further show that DeepVariant can learn to call variants in a variety of sequencing technologies and experimental designs, including deep whole genomes from 10X Genomics and Ion Ampliseq exomes, highlighting the benefits of using more automated and generalizable techniques for

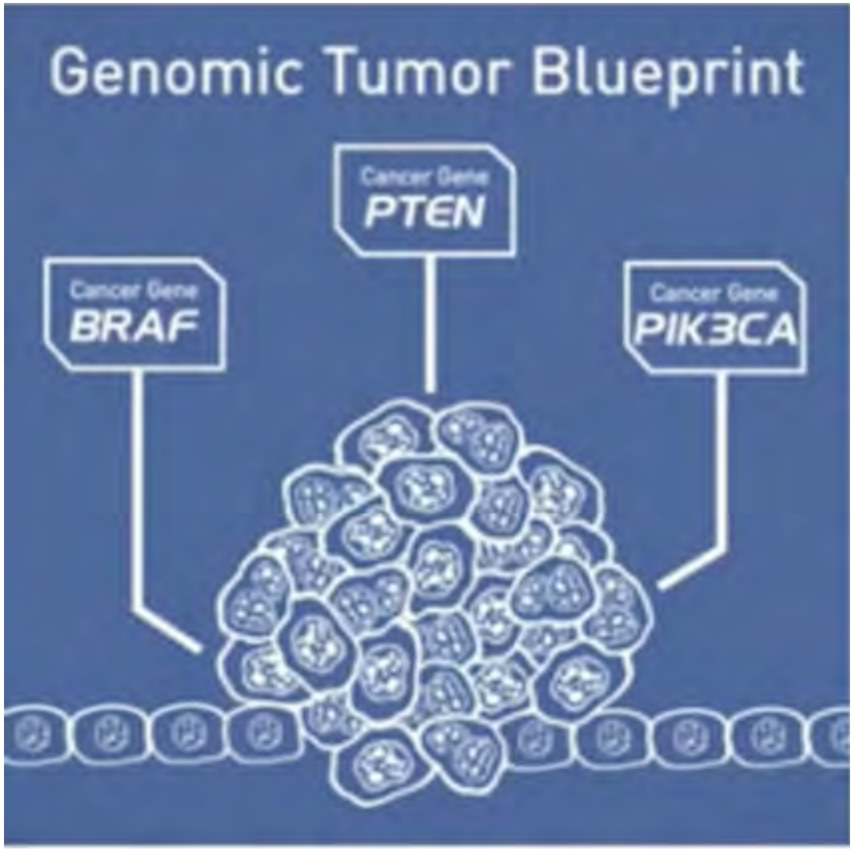
Here we describe a variant caller, called DeepVariant, that replaces the assortment of statistical modeling components with a single deep learning model. Deep learning is a machine learning technique applicable to a variety of domains, including image classification[9], translation[10], gaming[11,12] and the life sciences[13–16]. This toolchain (**Fig. 1**) begins by finding candidate single nucleotide polymorphisms (SNPs) and indels in reads aligned to the reference genome with high sensitivity but low specificity using standard, algorithmic preprocessing techniques. The deep learning model, using the Inception architecture[17], emits probabilities for each of the three diploid genotypes at a locus using a pileup image of the reference and read data around each candidate variant (**Fig. 1**). The model is trained using labeled true genotypes, after which it is frozen and can then be applied to novel sites or samples. In the following experiments, DeepVariant was trained on an independent set of samples or variants from those being evaluated.

Personalized Disease Prevention and Treatment



Published Oct 17, 2019

6 Years Have Passed Since Angelina Jolie's Preventative Double Mastectomy — Here's What You Need To Know About Inherited Risk For Breast Cancer



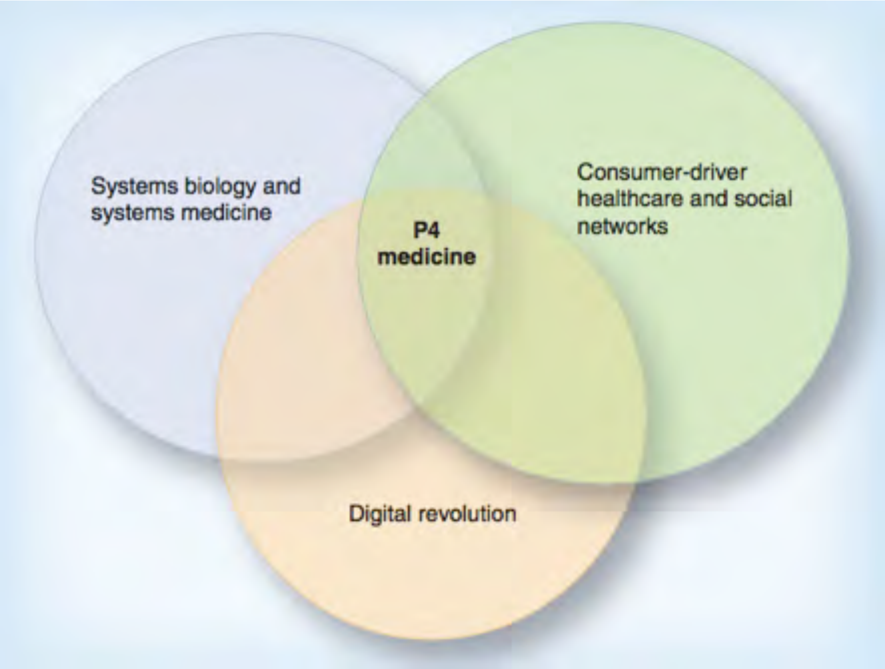
Personalized Medicine

A new healthcare system is emerging: predictive, preventive, personalized and participatoryhealthcare that encompasses:

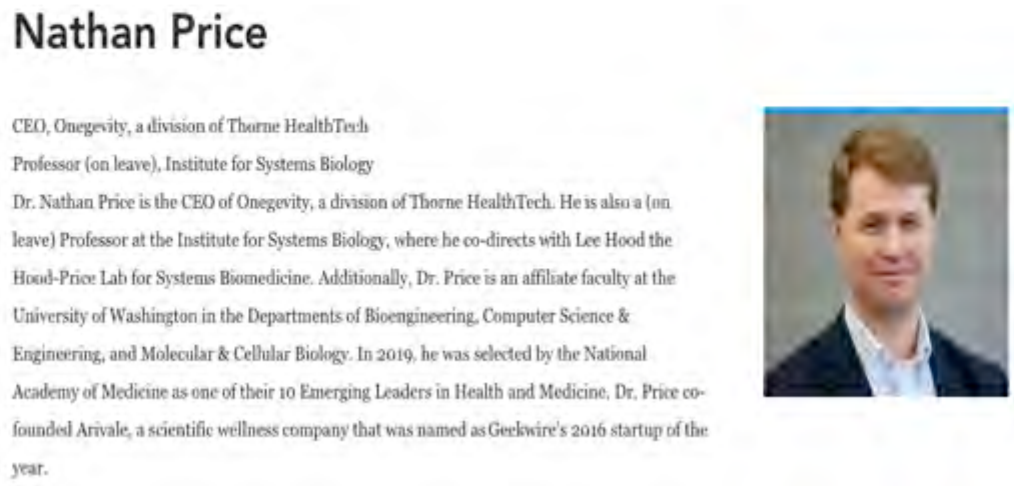
– Systems approaches to biology and medicine.

– The digital revolution’s radically enhanced capabilities for collecting, integrating, storing, analyzing and communicating data and information.

– Increasing numbers of networked and activated patients and consumers


Systems biology and systems medicine
P4 medicine
Consumer-driver healthcare and social networks
Digital revolution

Onegevity: Next-Generation Multi-Omic Platform For Precision Health



Nathan Price

CEO, Onegevity, a division of Thorne HealthTech

Professor (on leave), Institute for Systems Biology

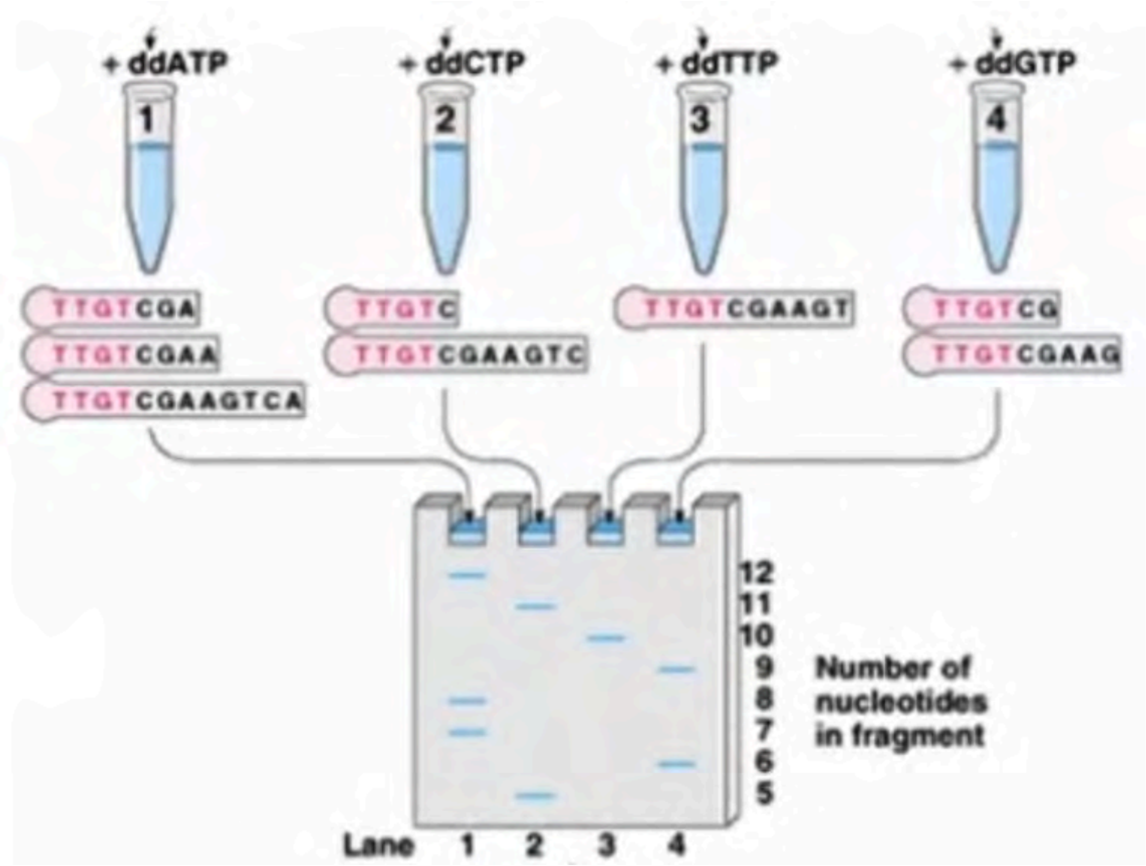
Dr. Nathan Price is the CEO of Onegevity, a division of Thorne HealthTech. He is also a (on leave) Professor at the Institute for Systems Biology, where he co-directs with Lee Hood the Hood-Price Lab for Systems Biomedicine. Additionally, Dr. Price is an affiliate faculty at the University of Washington in the Departments of Bioengineering, Computer Science & Engineering, and Molecular & Cellular Biology. In 2019, he was selected by the National Academy of Medicine as one of their 10 Emerging Leaders in Health and Medicine. Dr. Price co-founded Arivale, a scientific wellness company that was named as Geekwire's 2016 startup of the year.

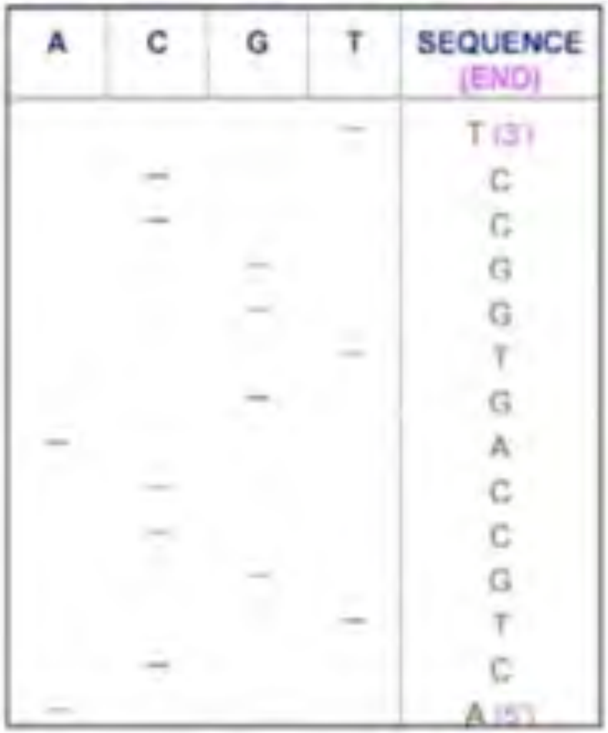
Big Data Challenge

1st Generation: Sanger Sequencing

Add single-stranded DNA sequence to four test tubes

Each tube contains all dNTPs + one ddNTP

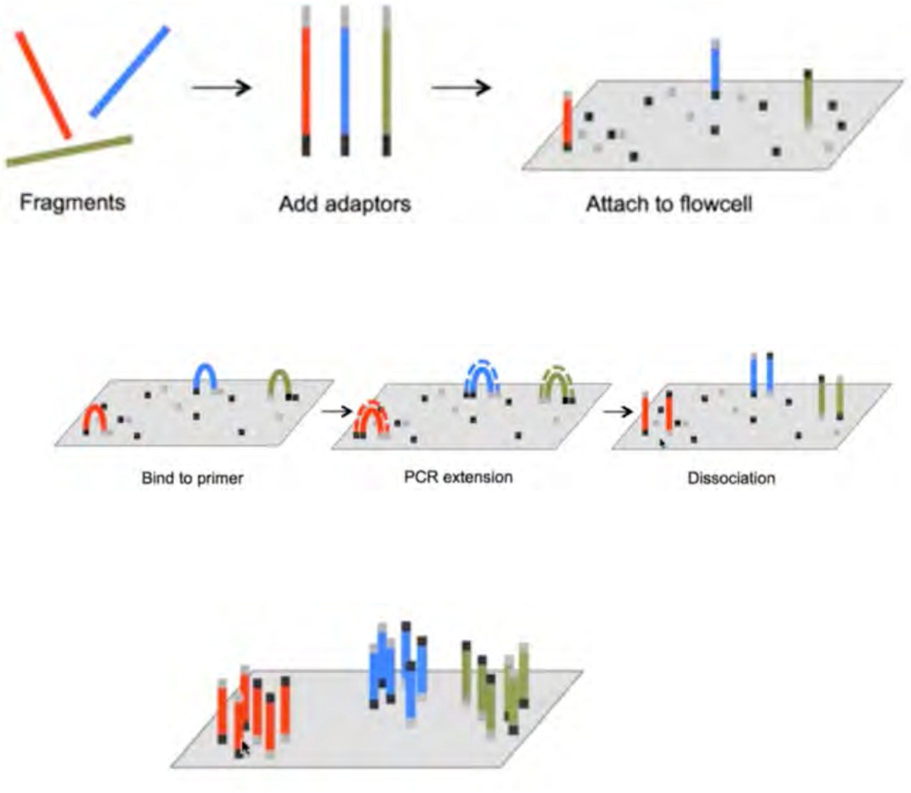




A	C	G	T	SEQUENCE (END)
			—	T (3')
	—			C
	—			C
		—		G
		—		G
			—	T
		—		G
—				A
	—			C
	—			C
		—		G
			—	T
	—			C
—				A (5')

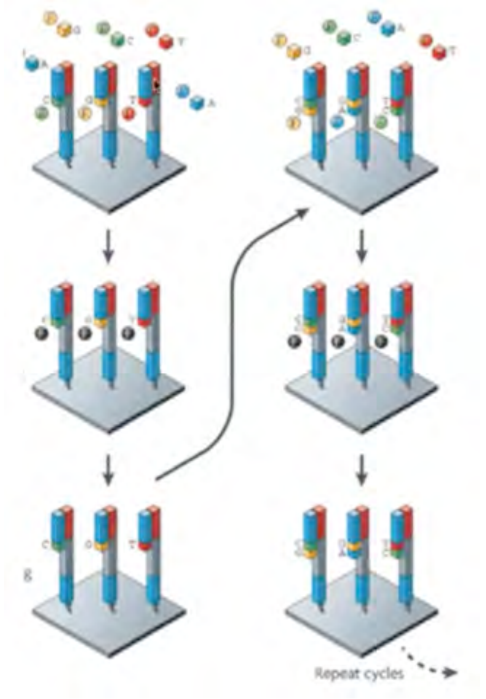
2nd Generation: Illumina Sequencing

Cluster Generation



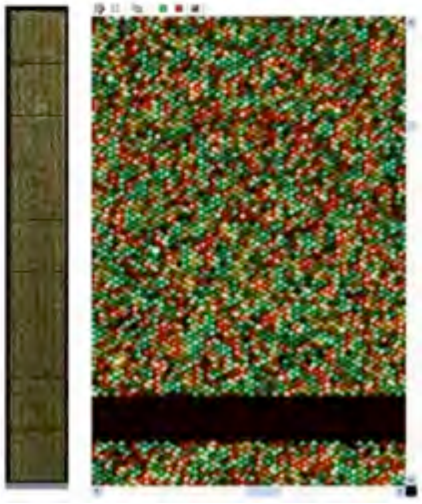
2nd Generation: Illumina Sequencing
Sequencing by Synthesis

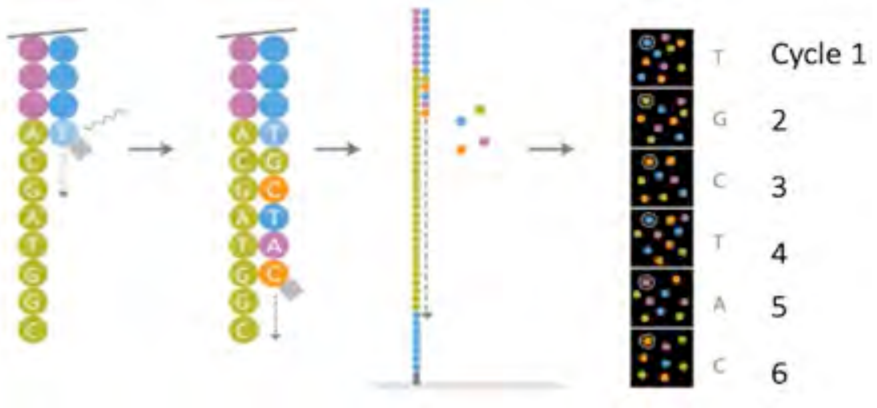
- Incorporate all 4-nucleotide, each label with a different dye
- Wash 4-color imaging
- Cleave dye and terminating groups, wash
- Repeat cycles



2nd Generation: Illumina Sequencing Image Analysis

https://www.youtube.com/watch?v=fCd6B5HRaZ8





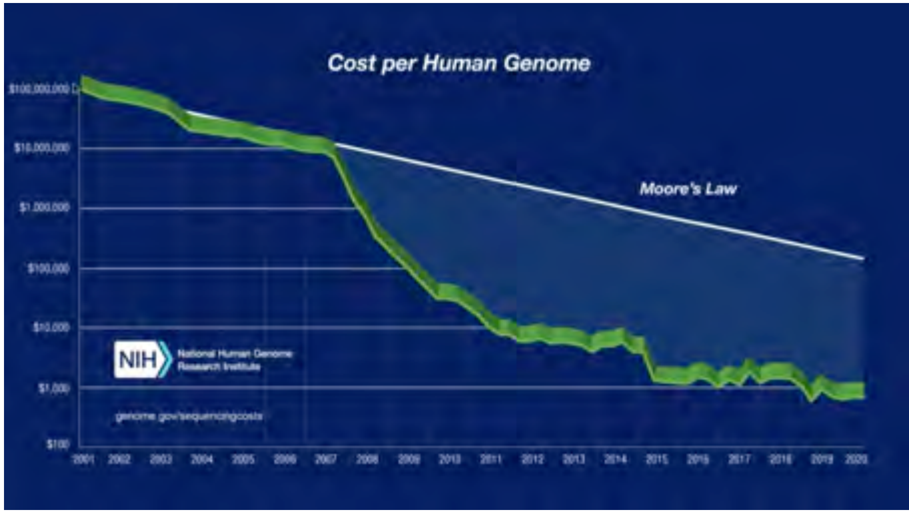
2nd Generation: Illumina Sequencing

Lets look into another example

NovaSeq 6000

System Specifications*

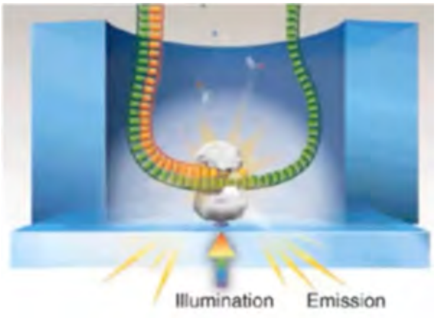
⟷	▥	⇡
4800-6000 Gb	32-40 B	2 × 250 bp
OUTPUT RANGE	READS PER RUN	MAX READ LENGTH

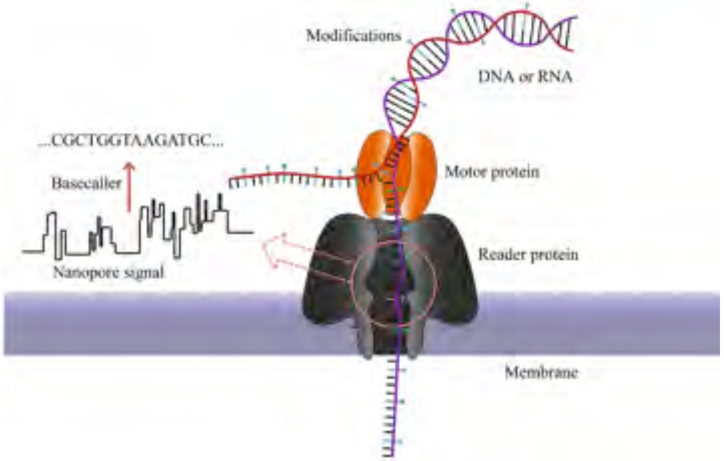


3rd Generation Sequencing

- Single molecule sequencing: use polymerase for sequencing without DNA amplification
- Fewer but much longer reads: good for long sequences but not good for read count applications
- Technology still under active development :

https://www.youtube.com/watch?v=v8p4ph2MAvI
https://www.youtube.com/watch?v=E9-Rm5AoZGw


Illumination
Emission


Modifications
DNA or RNA
...CGCTGGTAAGATGC...
Basecaller
Motor protein
Nanopore signal
Reader protein
Membrane