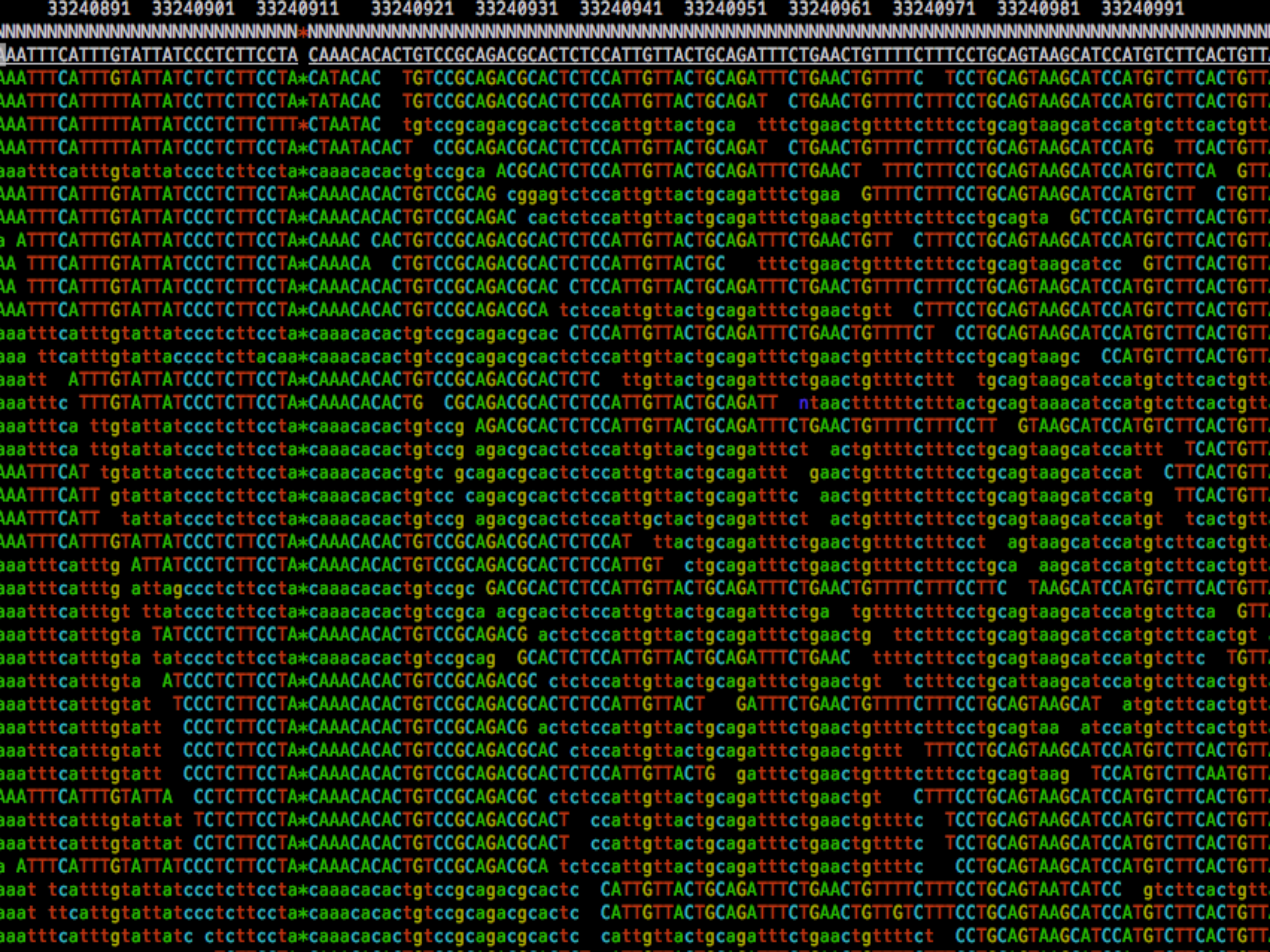


```
     33240891  33240901  33240911   33240921  33240931  33240941  33240951  33240961  33240971  33240981  33240991
NNNNNNNNNNNNNNNNNNNNNNNNNNNNN*NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
AAATTTCATTTGTATTATCCCTCTTCCTA CAAACACACTGTCCGCAGACGCACTCTCCATTGTTACTGCAGATTTCTGAACTGTTTTCTTTCCTGCAGTAAGCATCCATGTCTTCACTGTT
AAATTTCATTTGTATTATCTCTCTTCCTA*CATACAC  TGTCCGCAGACGCACTCTCCATTGTTACTGCAGATTTCTGAACTGTTTTC  TCCTGCAGTAAGCATCCATGTCTTCACTGTT
AAATTTCATTTTTATTATCCTTCTTCCTA*TATACAC  TGTCCGCAGACGCACTCTCCATTGTTACTGCAGAT  CTGAACTGTTTTCTTTCCTGCAGTAAGCATCCATGTCTTCACTGTT
AAATTTCATTTTTATTATCCCTCTTCTTT*CTAATAC  tgtccgcagacgcactctccattgttactgca  tttctgaactgttttctttcctgcagtaagcatccatgtcttcactgtt
AAATTTCATTTTTATTATCCCTCTTCCTA*CTAATACACT  CCGCAGACGCACTCTCCATTGTTACTGCAGAT  CTGAACTGTTTTCTTTCCTGCAGTAAGCATCCATG  TTCACTGTT
aaatttcatttgtattatccctcttccta*caaacacactgtccgca ACGCACTCTCCATTGTTACTGCAGATTTCTGAACT  TTTCTTTCCTGCAGTAAGCATCCATGTCTTCA  GTT
AAATTTCATTTGTATTATCCCTCTTCCTA*CAAACACACTGTCCGCAG cggagtctccattgttactgcagatttctgaa  GTTTTCTTTCCTGCAGTAAGCATCCATGTCTT  CTGTT
AAATTTCATTTGTATTATCCCTCTTCCTA*CAAACACACTGTCCGCAGAC cactctccattgttactgcagatttctgaactgttttctttcctgcagta  GCTCCATGTCTTCACTGTT
a ATTTCATTTGTATTATCCCTCTTCCTA*CAAAC CACTGTCCGCAGACGCACTCTCCATTGTTACTGCAGATTTCTGAACTGTT  CTTTCCTGCAGTAAGCATCCATGTCTTCACTGTT
AA TTTCATTTGTATTATCCCTCTTCCTA*CAAACA  CTGTCCGCAGACGCACTCTCCATTGTTACTGC   tttctgaactgttttctttcctgcagtaagcatcc  GTCTTCACTGTT
AA TTTCATTTGTATTATCCCTCTTCCTA*CAAACACACTGTCCGCAGACGCAC CTCCATTGTTACTGCAGATTTCTGAACTGTTTTCTTTCCTGCAGTAAGCATCCATGTCTTCACTGTT
AAATTTCATTTGTATTATCCCTCTTCCTA*CAAACACACTGTCCGCAGACGCA tctccattgttactgcagatttctgaactgtt  CTTTCCTGCAGTAAGCATCCATGTCTTCACTGTT
aaatttcatttgtattatccctcttccta*caaacacactgtccgcagacgcac CTCCATTGTTACTGCAGATTTCTGAACTGTTTTCT  CCTGCAGTAAGCATCCATGTCTTCACTGTT
aaa ttcatttgtattacccctcttacaa*caaacacactgtccgcagacgcactctccattgttactgcagatttctgaactgttttctttcctgcagtaagc  CCATGTCTTCACTGTT
aaatt  ATTTGTATTATCCCTCTTCCTA*CAAACACACTGTCCGCAGACGCACTCTC  ttgttactgcagatttctgaactgttttcttt  tgcagtaagcatccatgtcttcactgtt
aaatttc TTTGTATTATCCCTCTTCCTA*CAAACACACTG  CGCAGACGCACTCTCCATTGTTACTGCAGATT  ntaacttttttctttactgcagtaaacatccatgtcttcactgtt
aaatttca ttgtattatccctcttccta*caaacacactgtccg AGACGCACTCTCCATTGTTACTGCAGATTTCTGAACTGTTTTCTTTCCTT  GTAAGCATCCATGTCTTCACTGTT
aaatttca ttgtattatccctcttccta*caaacacactgtccg agacgcactctccattgttactgcagatttct  actgttttctttcctgcagtaagcatccattt  TCACTGTT
AAATTTCAT tgtattatccctcttccta*caaacacactgtc gcagacgcactctccattgttactgcagattt  gaactgttttctttcctgcagtaagcatccat  CTTCACTGTT
AAATTTCATT gtattatccctcttccta*caaacacactgtcc cagacgcactctccattgttactgcagatttc  aactgttttctttcctgcagtaagcatccatg  TTCACTGTT
AAATTTCATT  tattatccctcttccta*caaacacactgtccg agacgcactctccattgctactgcagatttct  actgttttctttcctgcagtaagcatccatgt  tcactgtt
AAATTTCATTTGTATTATCCCTCTTCCTA*CAAACACACTGTCCGCAGACGCACTCTCCAT  ttactgcagatttctgaactgttttctttcct  agtaagcatccatgtcttcactgtt
aaatttcatttg ATTATCCCTCTTCCTA*CAAACACACTGTCCGCAGACGCACTCTCCATTGT  ctgcagatttctgaactgttttctttcctgca  aagcatccatgtcttcactgtt
aaatttcatttg attagccctcttccta*caaacacactgtccgc GACGCACTCTCCATTGTTACTGCAGATTTCTGAACTGTTTTCTTTCCTTC  TAAGCATCCATGTCTTCACTGTT
aaatttcatttgt ttatccctcttccta*caaacacactgtccgca acgcactctccattgttactgcagatttctga  tgttttctttcctgcagtaagcatccatgtcttca  GTT
aaatttcatttgta TATCCCTCTTCCTA*CAAACACACTGTCCGCAGACG actctccattgttactgcagatttctgaactg  ttctttcctgcagtaagcatccatgtcttcactgt
aaatttcatttgta tatccctcttccta*caaacacactgtccgcag  GCACTCTCCATTGTTACTGCAGATTTCTGAAC  ttttctttcctgcagtaagcatccatgtcttc  TGTT
aaatttcatttgta  ATCCCTCTTCCTA*CAAACACACTGTCCGCAGACGC ctctccattgttactgcagatttctgaactgt  tctttcctgcattaagcatccatgtcttcactgtt
aaatttcatttgtat  TCCCTCTTCCTA*CAAACACACTGTCCGCAGACGCACTCTCCATTGTTACT   GATTTCTGAACTGTTTTCTTTCCTGCAGTAAGCAT  atgtcttcactgtt
aaatttcatttgtatt  CCCTCTTCCTA*CAAACACACTGTCCGCAGACG actctccattgttactgcagatttctgaactgttttctttcctgcagtaa  atccatgtcttcactgtt
aaatttcatttgtatt  CCCTCTTCCTA*CAAACACACTGTCCGCAGACGCAC ctccattgttactgcagatttctgaactgttt  TTTCCTGCAGTAAGCATCCATGTCTTCACTGTT
aaatttcatttgtatt  CCCTCTTCCTA*CAAACACACTGTCCGCAGACGCACTCTCCATTGTTACTG  gatttctgaactgttttctttcctgcagtaag  TCCATGTCTTCAATGTT
AAATTTCATTTGTATTA  CCTCTTCCTA*CAAACACACTGTCCGCAGACGC ctctccattgttactgcagatttctgaactgt   CTTTCCTGCAGTAAGCATCCATGTCTTCACTGTT
aaatttcatttgtattat TCTCTTCCTA*CAAACACACTGTCCGCAGACGCACT  ccattgttactgcagatttctgaactgttttc  TCCTGCAGTAAGCATCCATGTCTTCACTGTT
aaatttcatttgtattat CCTCTTCCTA*CAAACACACTGTCCGCAGACGCACT  ccattgttactgcagatttctgaactgttttc  TCCTGCAGTAAGCATCCATGTCTTCACTGTT
a ATTTCATTTGTATTATCCCTCTTCCTA*CAAACACACTGTCCGCAGACGCA tctccattgttactgcagatttctgaactgttttc   CCTGCAGTAAGCATCCATGTCTTCACTGTT
aaat tcatttgtattatccctcttccta*caaacacactgtccgcagacgcactc  CATTGTTACTGCAGATTTCTGAACTGTTTTCTTTCCTGCAGTAATCATCC  gtcttcactgtt
aaat ttcattgtattatccctcttccta*caaacacactgtccgcagacgcactc  CATTGTTACTGCAGATTTCTGAACTGTTGTCTTTCCTGCAGTAAGCATCCATGTCTTCACTGTT
aaatttcatttgtattatc ctcttccta*caaacacactgtccgcagacgcactc  cattgttactgcagatttctgaactgttttct  CCTGCAGTAAGCATCCATGTCTTCACTGTT
```

Clinical and Research Genomics
Spring 2021

Professor:
Christopher E. Mason, Ph.D.
Ebrahim Afshinnekoo, M.D.

TA:
Chandrima Bhattacharya, M.S.

Course Over Ten Sessions:

I. Sequencing Methods, Single-Cell Dynamics, and Molecular Detection Techniques (March 9th)

II. RNA Sequencing, Epitranscriptomes, and Single Cell / Spatial Omics (March 16th)

III. Epigenomes, DNA Modifications, and Chromatin Dynamics (March 23rd)

IV. Metagenomes, BGCs, and Metabolomics (March 30th)

V. Complex Genome Re-arrangements, Transposons, and Tools for Genetic Variant Calling (April 6th)

VI. Cancer Genomics, Non-coding Regulation and Variation(April 13th)

VII. Genome Ethics, Large Data, Small Data, and Disease Classification (April 20th)

VIII. Systems Biology, Synthetic Biology, & Genome Engineering (April 27th)

IX. COVID-19 Tracking and Pathophysiology (May 4th)

X. Global Health and Beyond-Globe Health (Aerospace Medicine) (May 11th)

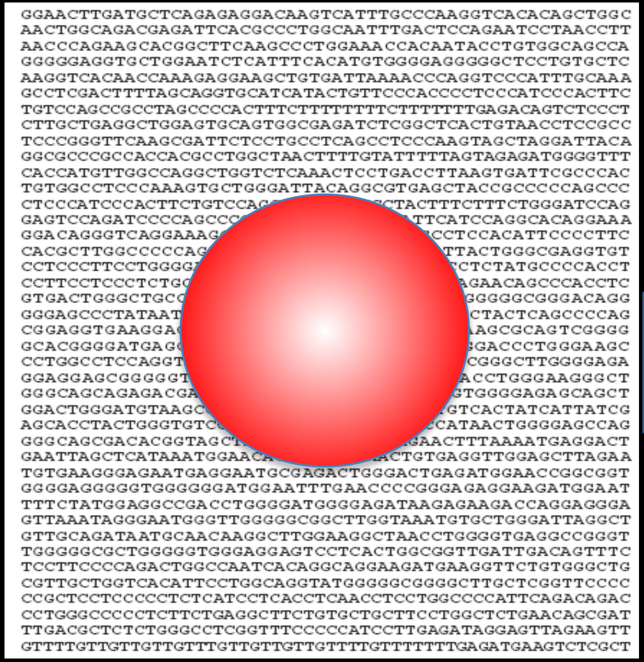
All classes on Zoom

Stay updated with the course webpage:

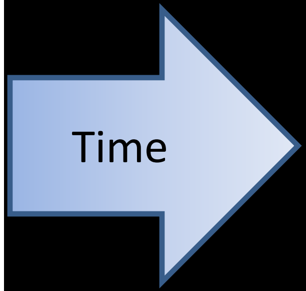
http://physiology.med.cornell.edu/faculty/mason/lab/clinicalgenomics/schedule.html

Start

Finish

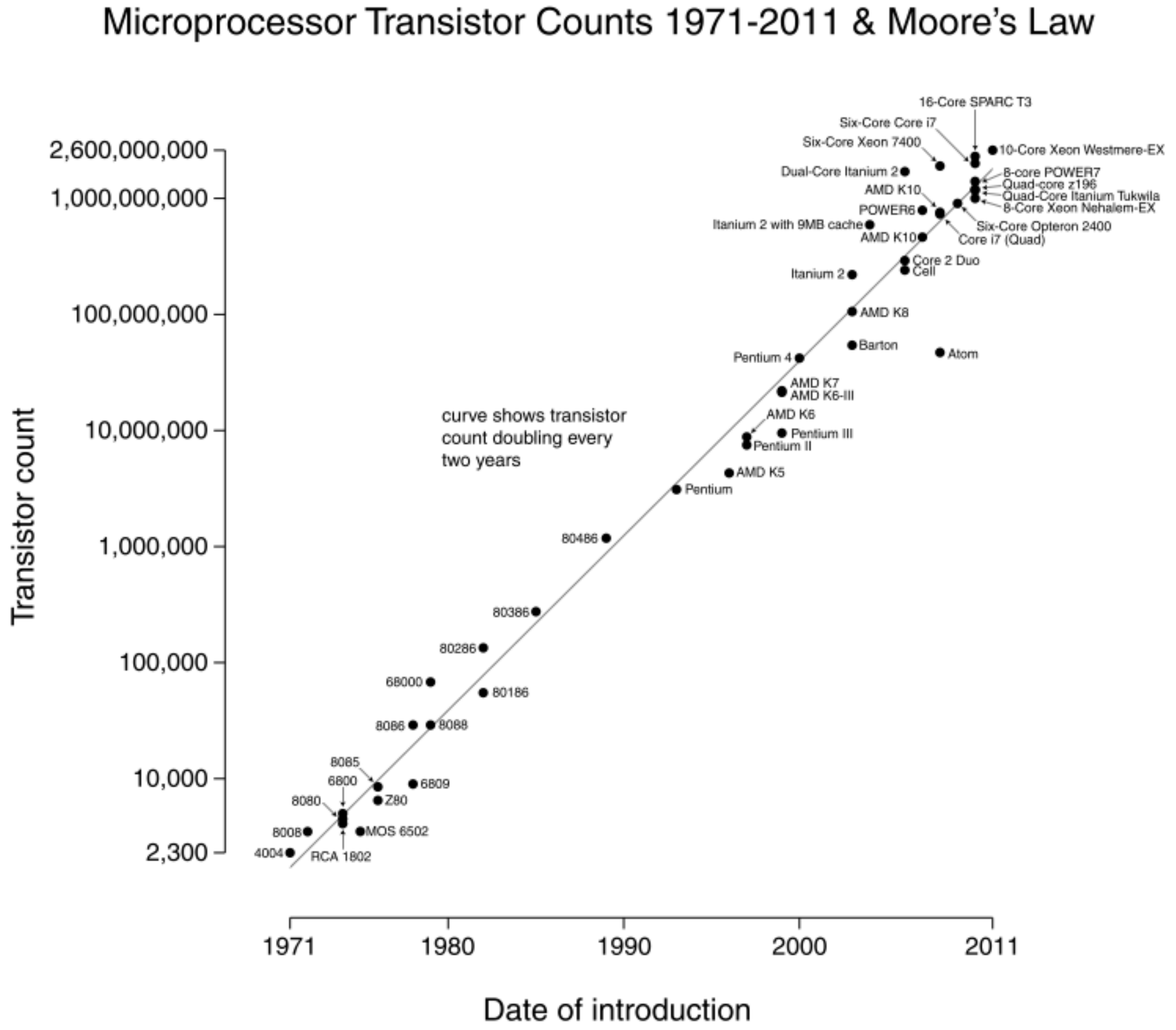


```
GGAACTTGATGCTCAGAGAGGACAAGTCATTTGCCCAAGGTCACACAGCTGGC
AACTGGCAGACGAGATTCACGCCCTGGCAATTTGACTCCAGAATCCTAACCTT
AACCCAGAAGCACGGCTTCAAGCCCTGGAAACCACAATACCTGTGGCAGCCA
GGGGGAGGTGCTGGAATCTCATTTCACATGTGGGGAGGGGGCTCCTGTGCTC
AAGGTCACAACCAAAGAGGAAGCTGTGATTAAAACCCAGGTCCCATTTGCAAA
GCCTCGACTTTTAGCAGGTGCATCATACTGTTCCCACCCCTCCCATCCCACTTC
TGTCCAGCCGCCTAGCCCCACTTTCTTTTTTTTCTTTTTTTGAGACAGTCTCCCT
CTTGCTGAGGCTGGAGTGCAGTGGCGAGATCTCGGCTCACTGTAACCTCCGCC
TCCCGGGTTCAAGCGATTCTCCTGCCTCAGCCTCCCAAGTAGCTAGGATTACA
GGCGCCCGCCACCACGCCTGGCTAACTTTTGTATTTTTAGTAGAGATGGGGTTT
CACCATGTTGGCCAGGCTGGTCTCAAACTCCTGACCTTAAGTGATTCGCCCAC
TGTGGCCTCCCAAAGTGCTGGGATTACAGGCGTGAGCTACCGCCCCCAGCCC
CTCCCATCCCACTTCTGTCCAG          CTACTTTCTTTCTGGATCCAG
GAGTCCAGATCCCCAGCCC               TTCATCCAGGCACAGGAAA
GGACAGGGTCAGGAAAG                   CCTCCACATTCCCCTTC
CACGCTTGGCCCCCAG                     TTACTGGGCGAGGTGT
CCTCCCTTCCTGGGG                       CTCTATGCCCCACCT
CCTTCCTCCCTCTGC                       AGAACAGCCCACCTC
GTGACTGGGCTGCC                         GGGGGCGGGACAGG
GGGAGCCCTATAAT                         CTACTCAGCCCCAG
CGGAGGTGAAGGA                           AGCGCAGTCGGGG
GCACGGGGATGAG                           GACCCTGGGAAGC
CCTGGCCTCCAGGT                         CGGGCTTGGGGAGA
GGAGGAGCGGGGGT                         ACCTGGGAAGGGCT
GGGCAGCAGAGACGA                       GTGGGGAGAGCAGCT
GGACTGGGATGTAAGC                     GTCACTATCATTATCG
AGCACCTACTGGGTGTC                   CATAACTGGGGAGCCAG
GGGCAGCGACACGGTAGCT               GAACTTTAAAATGAGGACT
GAATTAGCTCATAAATGGAAC          ACTGTGAGGTTGGAGCTTAGAA
TGTGAAGGGAGAATGAGGAATGCGAGACTGGGACTGAGATGGAACCGGCGGT
GGGGAGGGGGTGGGGGGATGGAATTTGAACCCCGGGAGAGGAAGATGGAAT
TTTCTATGGAGGCCGACCTGGGGATGGGGAGATAAGAGAAGACCAGGAGGGA
GTTAAATAGGGAATGGGTTGGGGGCGGCTTGGTAAATGTGCTGGGATTAGGCT
GTTGCAGATAATGCAACAAGGCTTGGAAGGCTAACCTGGGGTGAGGCCGGGT
TGGGGGCGCTGGGGGTGGGAGGAGTCCTCACTGGCGGTTGATTGACAGTTTC
TCCTTCCCCAGACTGGCCAATCACAGGCAGGAAGATGAAGGTTCTGTGGGCTG
CGTTGCTGGTCACATTCCTGGCAGGTATGGGGGCGGGGCTTGCTCGGTTCCCC
CCGCTCCTCCCCCTCTCATCCTCACCTCAACCTCCTGGCCCCATTCAGACAGAC
CCTGGGCCCCCTCTTCTGAGGCTTCTGTGCTGCTTCCTGGCTCTGAACAGCGAT
TTGACGCTCTCTGGGCCTCGGTTTCCCCCATCCTTGAGATAGGAGTTAGAAGTT
GTTTTGTTGTTGTTGTTTGTTGTTGTTGTTTTGTTTTTTTGAGATGAAGTCTCGCT
```



Time

The effects from Moore's Law ushered in a whole new era of technology



By Wgsimon

Initially we expected a $1K Genome in 2040

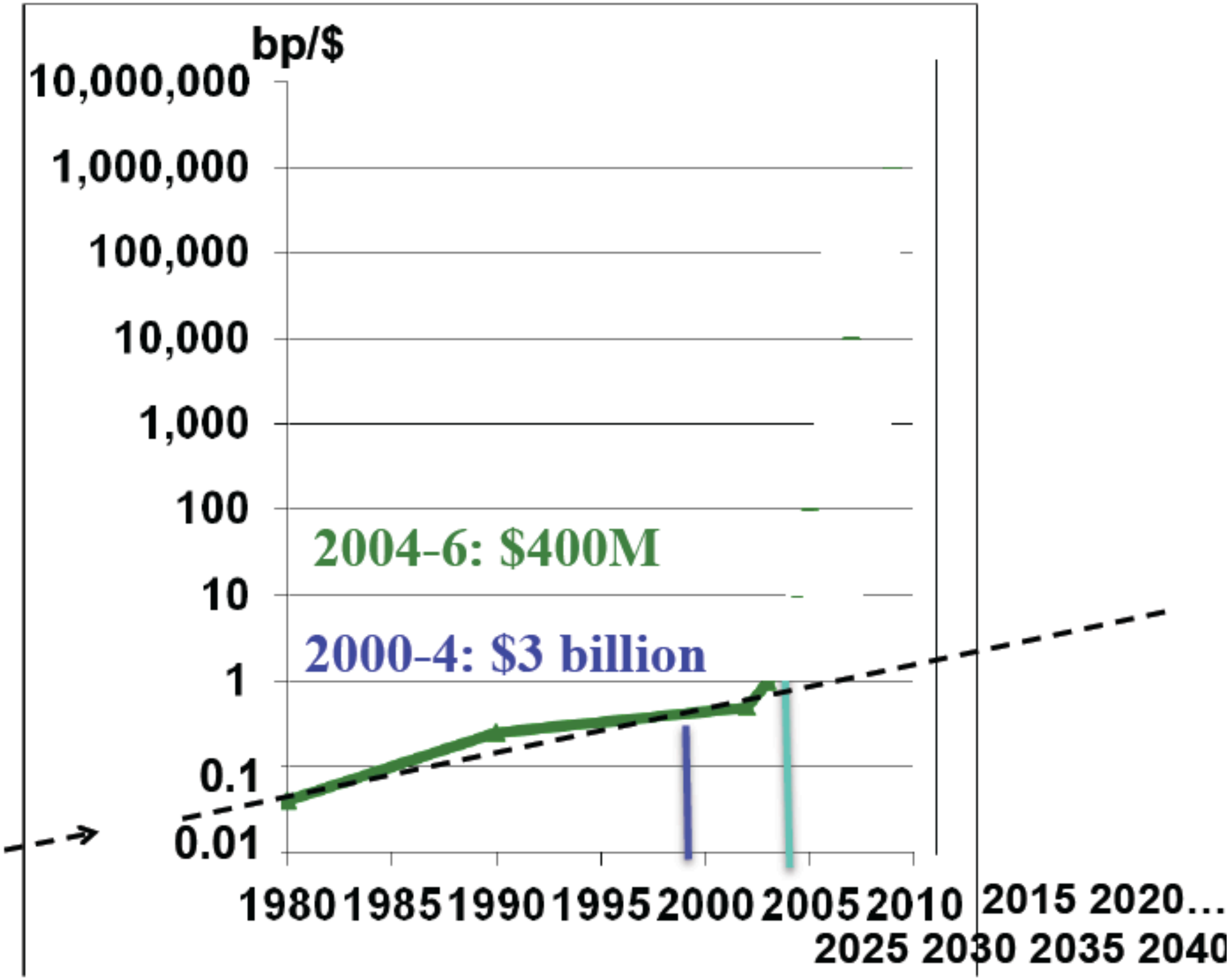
$1000
Genome

When?

2040

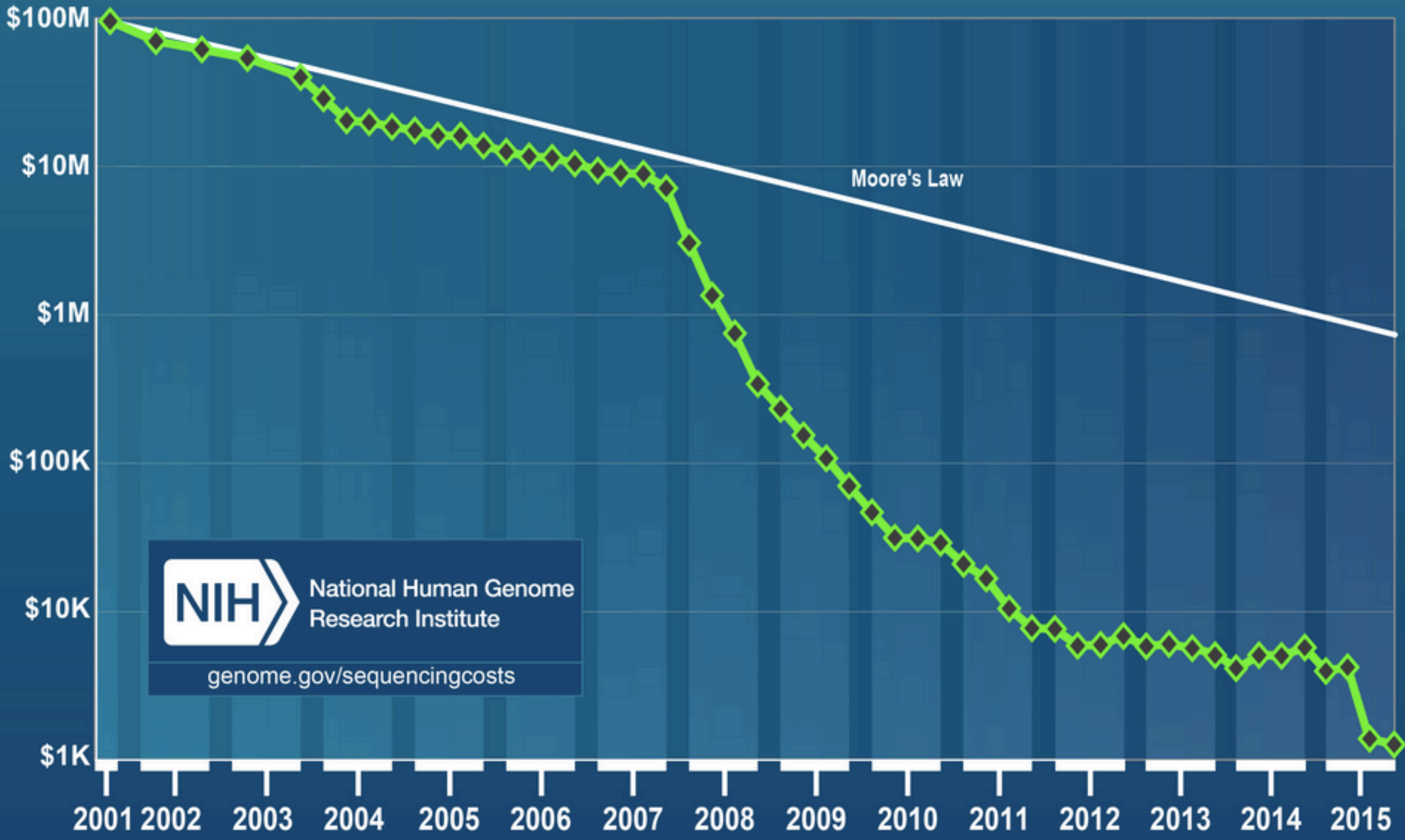
- - - - - -

Moore's law 1.5x/yr for electronics



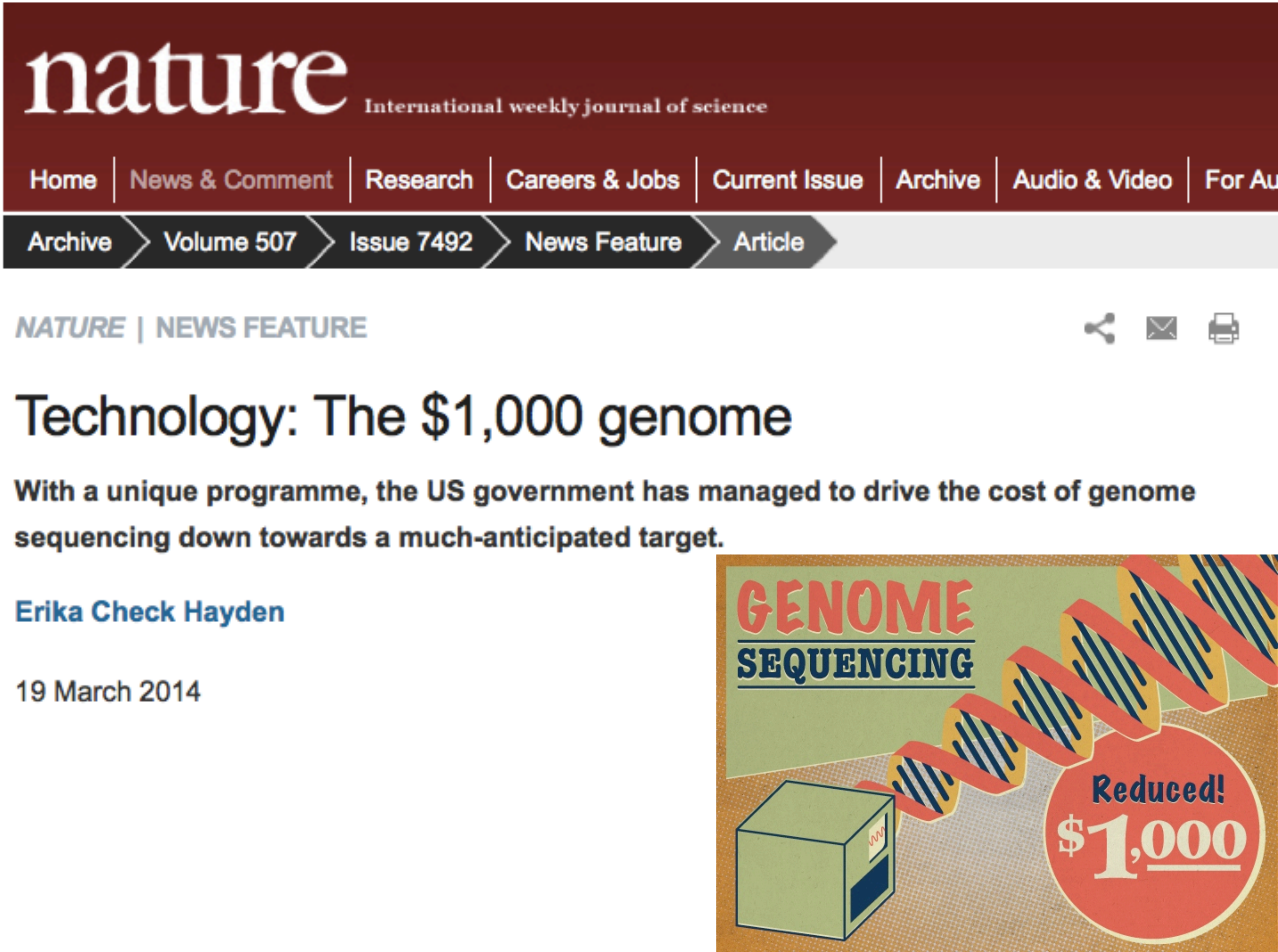
George Church

Cost per Genome



NIH National Human Genome Research Institute

genome.gov/sequencingcosts



NATURE | NEWS FEATURE

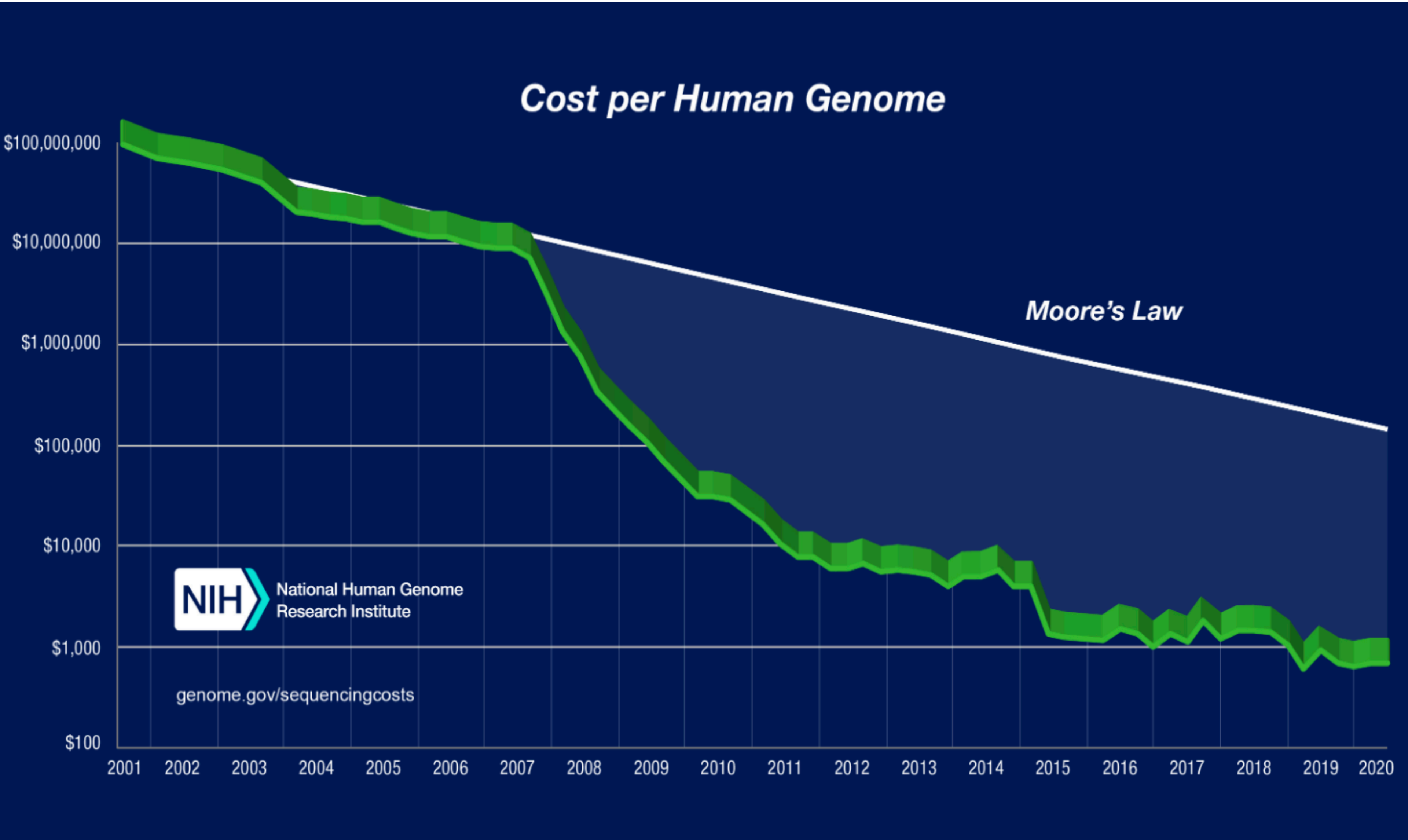
Technology: The $1,000 genome

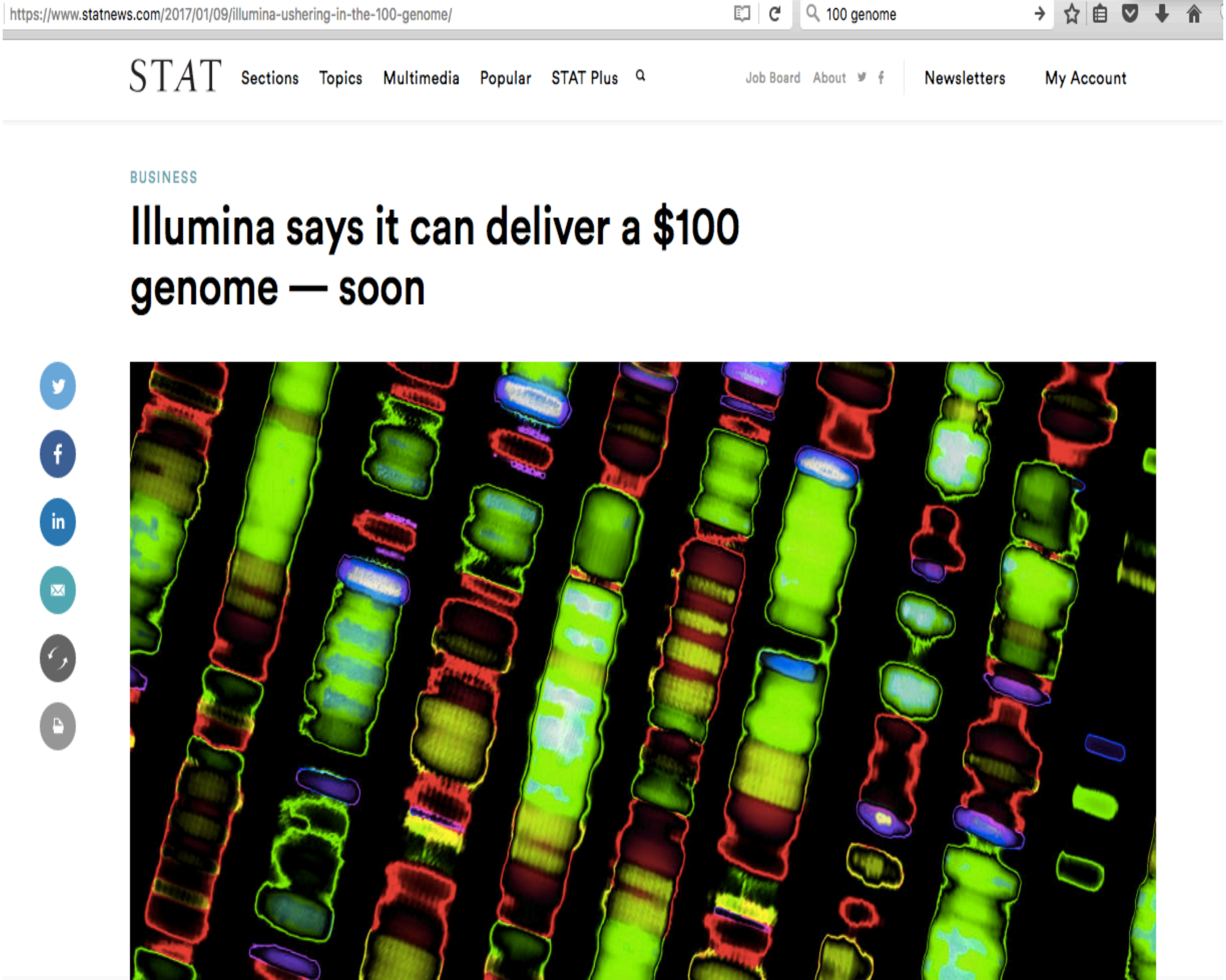
With a unique programme, the US government has managed to drive the cost of genome sequencing down towards a much-anticipated target.

Erika Check Hayden

19 March 2014

Flatlined a little

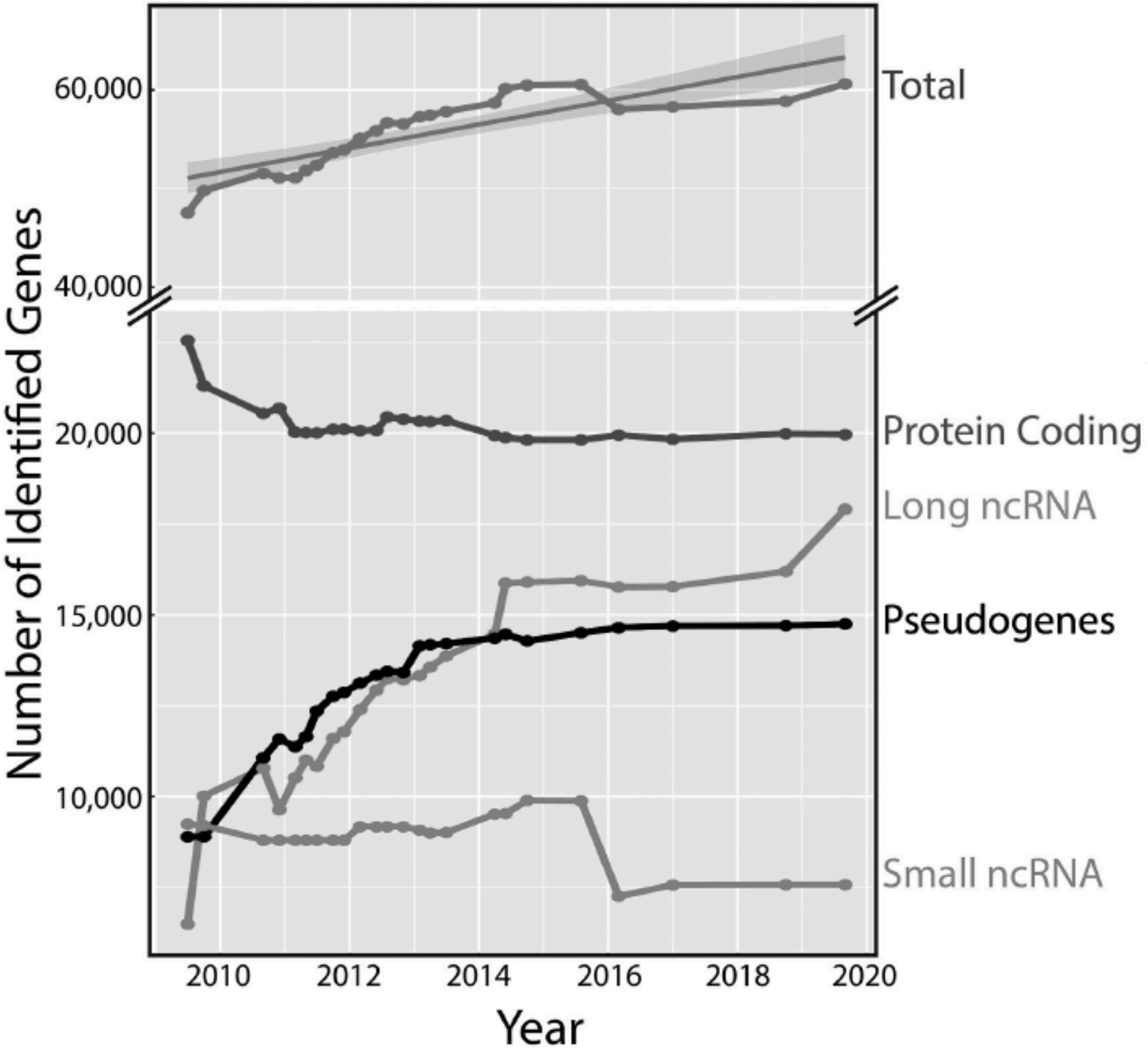




BUSINESS

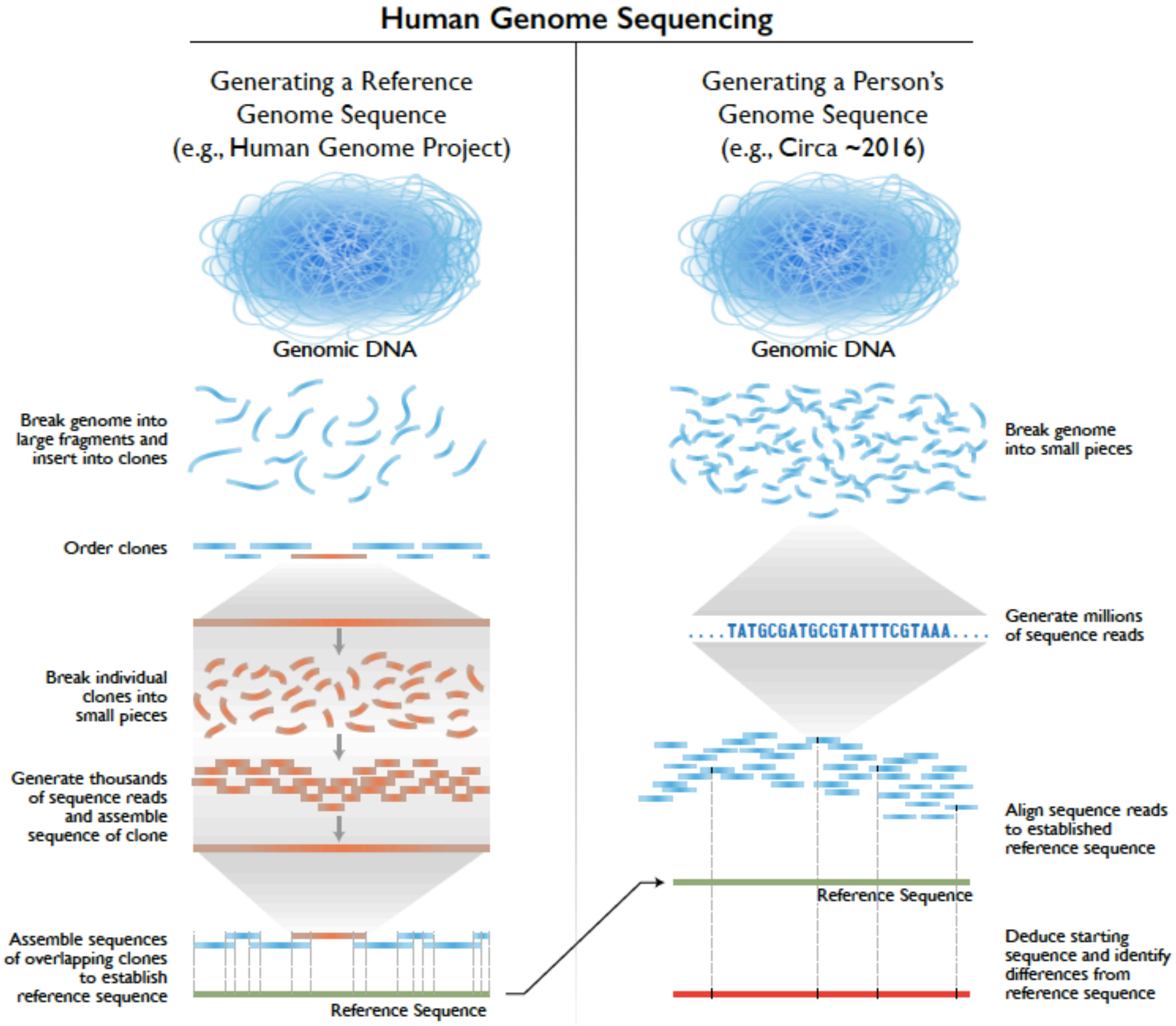
Illumina says it can deliver a $100 genome — soon

New genes still coming



https://mitpress.mit.edu/books/next-500-years

Every Day
is the
Best Day



Human Genome Sequencing

Generating a Reference Genome Sequence (e.g., Human Genome Project)

Break genome into large fragments and insert into clones

Order clones

Break individual clones into small pieces

Generate thousands of sequence reads and assemble sequence of clone

Assemble sequences of overlapping clones to establish reference sequence

Generating a Person's Genome Sequence (e.g., Circa ~2016)

Break genome into small pieces

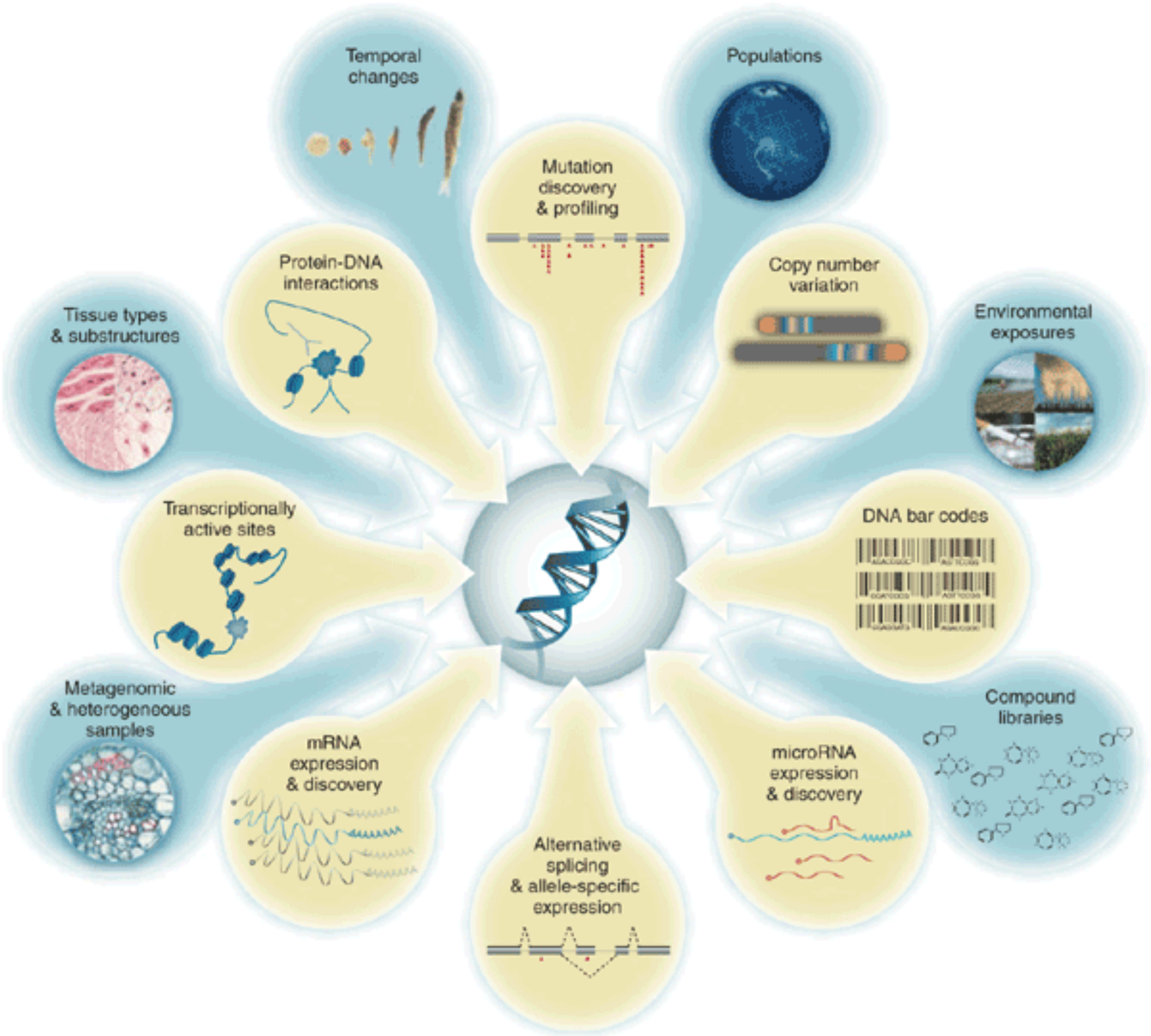
Generate millions of sequence reads

Align sequence reads to established reference sequence

Deduce starting sequence and identify differences from reference sequence

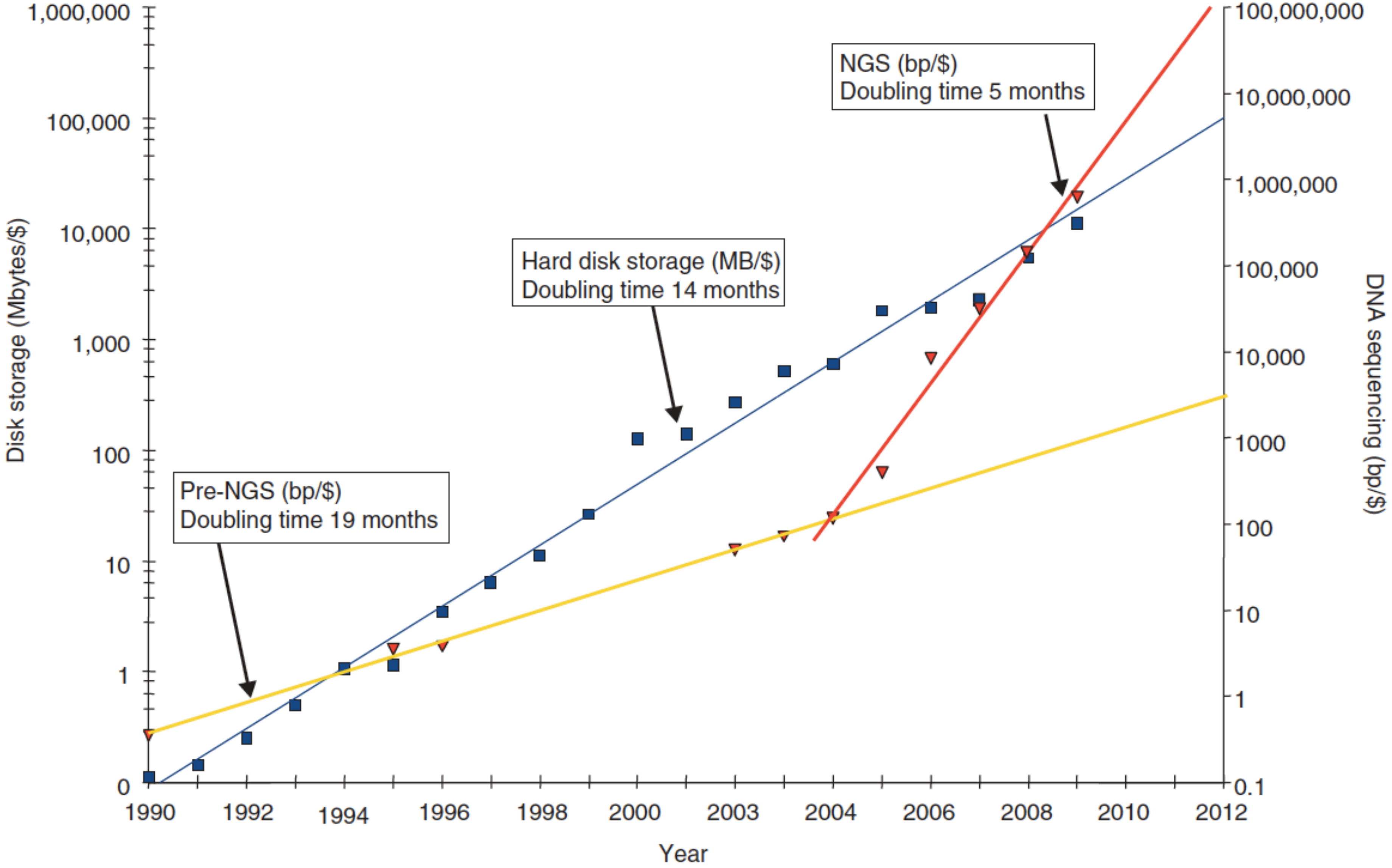
https://www.genome.gov/images/illustrations/sequencing.pdf

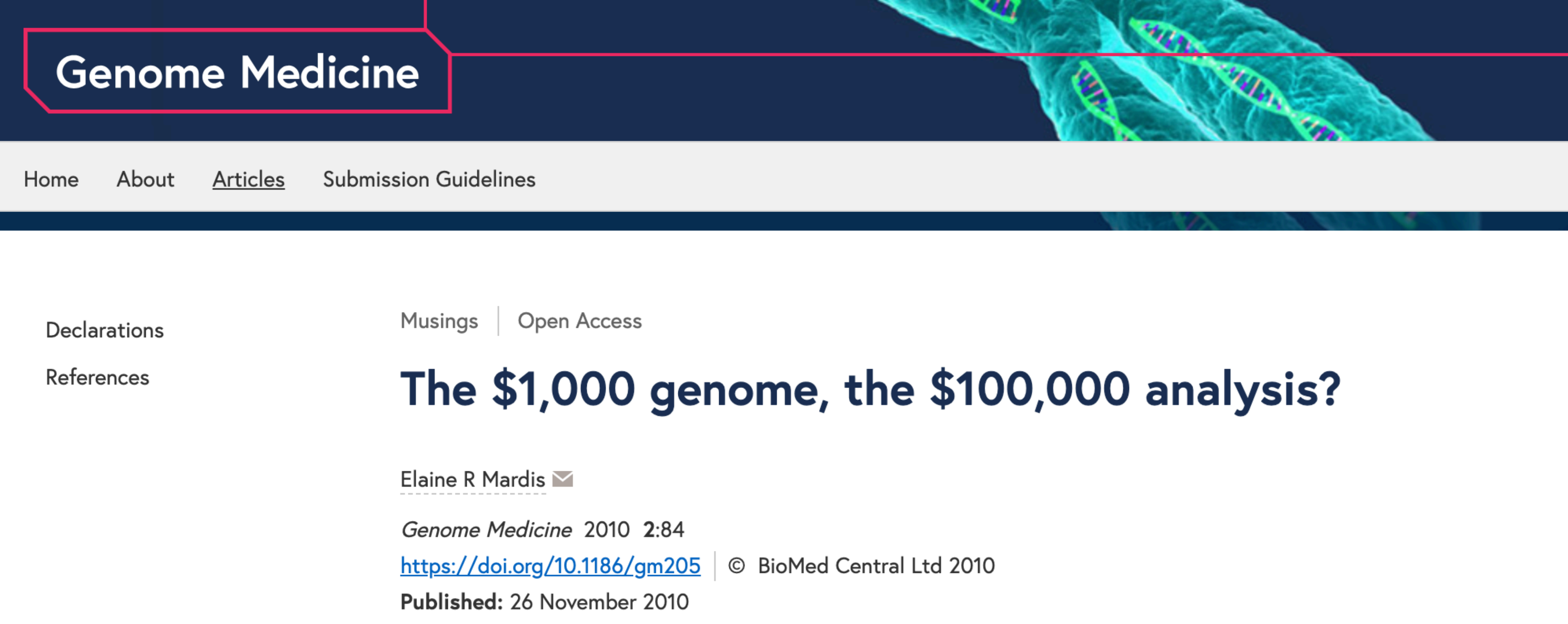
Since DNA defines the biochemical recipe for the genesis of organisms, sequencing allows us to create molecular portraits of development and disease at single-base resolution.


Temporal changes
Populations
Mutation discovery & profiling
Protein-DNA interactions
Copy number variation
Tissue types & substructures
Environmental exposures
Transcriptionally active sites
DNA bar codes
Metagenomic & heterogeneous samples
Compound libraries
mRNA expression & discovery
microRNA expression & discovery
Alternative splicing & allele-specific expression

Kahvejian, 2008

But, hard drive space is not keeping pace, creating a phalanx of companies aimed at the cloud





Musings | Open Access

The $1,000 genome, the $100,000 analysis?

Elaine R Mardis

Genome Medicine 2010 **2**:84

https://doi.org/10.1186/gm205 | © BioMed Central Ltd 2010

Published: 26 November 2010

https://genomemedicine.biomedcentral.com/articles/10.1186/gm205

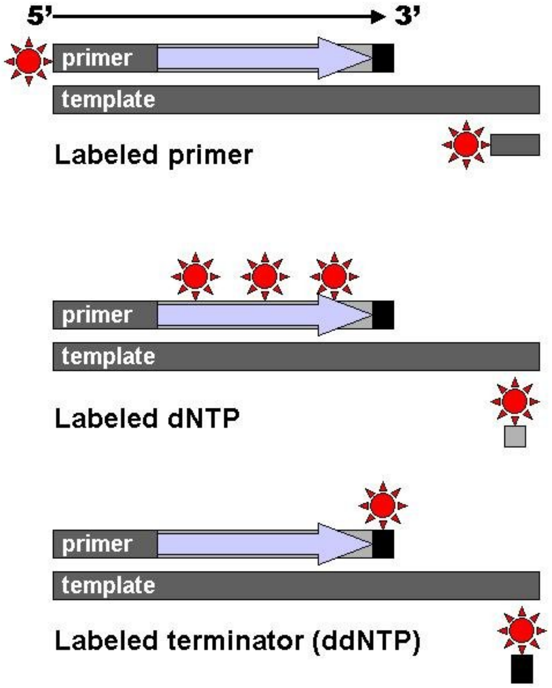
Sequencing Technologies

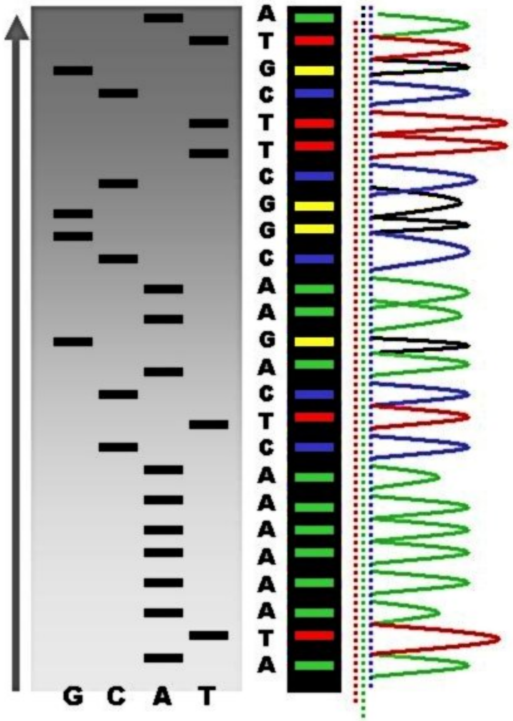
1. “Old School” dye-terminator sequencing (Sanger). 300-1000bp

2. “New School” methods
   a. Emulsion PCR Pyrosequencing
   b. Solid-phase amplification sequencing by synthesis (clonal or single molecule)
   c. Sequencing by ligation
   d. Single-molecule, real-time (SMRT) sequencing
   e. Electrical sequencing

Sequencing Technologies

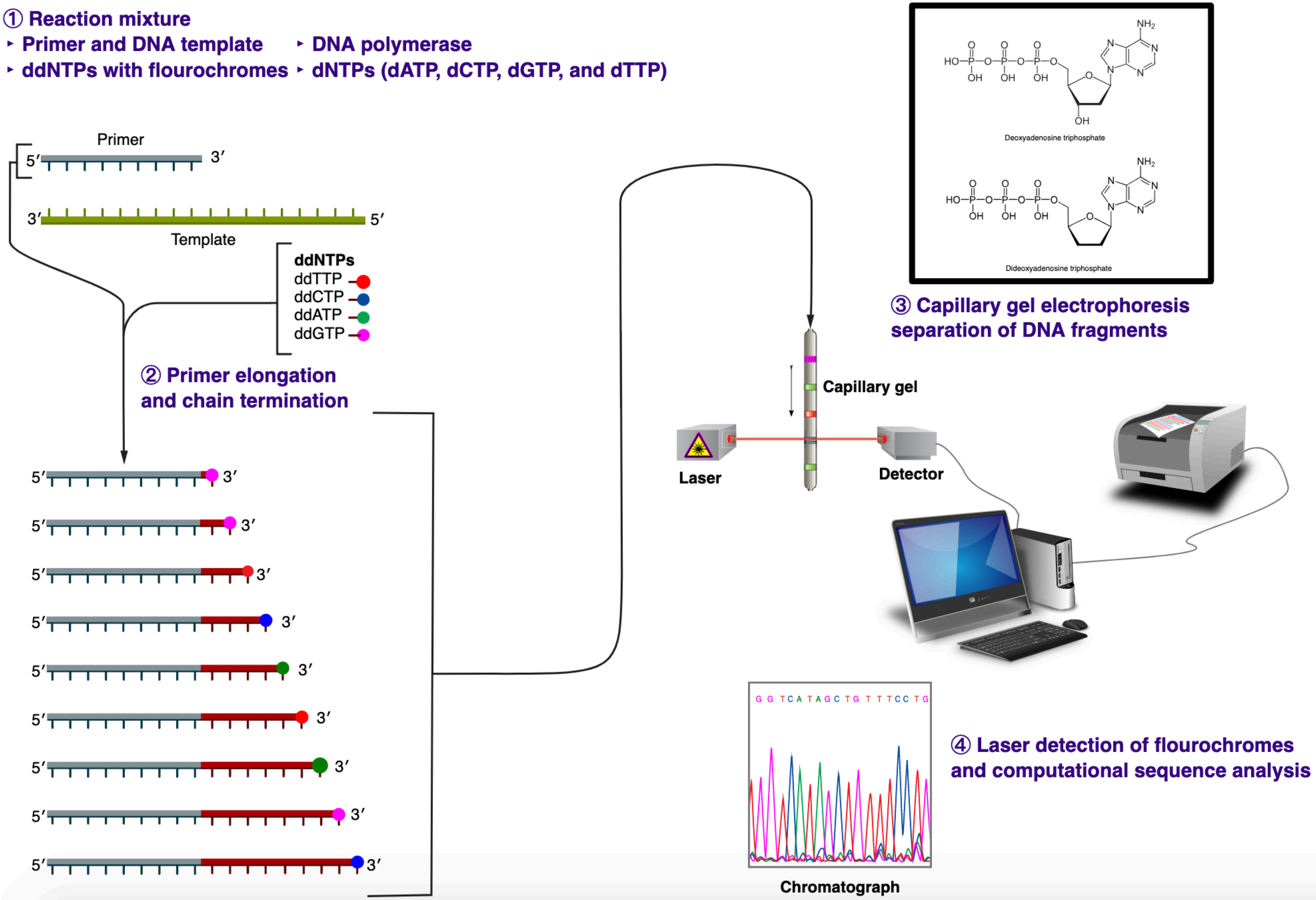
1. “Old School” dye-terminator sequencing (Sanger). 300-1000bp





① Reaction mixture

- Primer and DNA template
- ddNTPs with flourochromes
- DNA polymerase
- dNTPs (dATP, dCTP, dGTP, and dTTP)



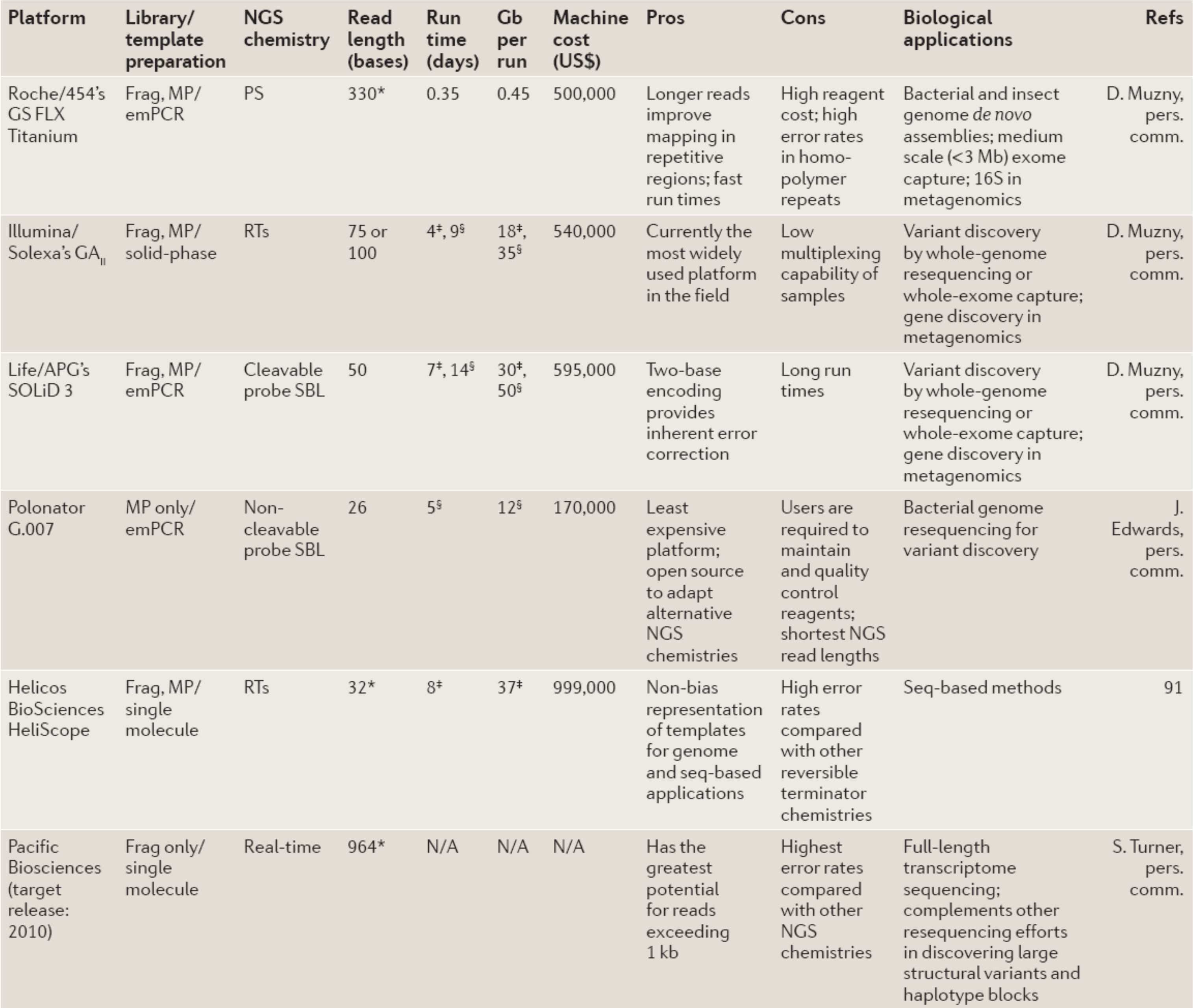
② Primer elongation and chain termination

③ Capillary gel electrophoresis separation of DNA fragments

④ Laser detection of flourochromes and computational sequence analysis

Estevezj

By 2009, many options emerged



Platform	Library/ template preparation	NGS chemistry	Read length (bases)	Run time (days)	Gb per run	Machine cost (US$)	Pros	Cons	Biological applications	Refs
Roche/454's GS FLX Titanium	Frag, MP/ emPCR	PS	330*	0.35	0.45	500,000	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homo-polymer repeats	Bacterial and insect genome *de novo* assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics	D. Muzny, pers. comm.
Illumina/ Solexa's GA_{II}	Frag, MP/ solid-phase	RTs	75 or 100	4‡, 9§	18‡, 35§	540,000	Currently the most widely used platform in the field	Low multiplexing capability of samples	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Life/APG's SOLiD 3	Frag, MP/ emPCR	Cleavable probe SBL	50	7‡, 14§	30‡, 50§	595,000	Two-base encoding provides inherent error correction	Long run times	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Polonator G.007	MP only/ emPCR	Non-cleavable probe SBL	26	5§	12§	170,000	Least expensive platform; open source to adapt alternative NGS chemistries	Users are required to maintain and quality control reagents; shortest NGS read lengths	Bacterial genome resequencing for variant discovery	J. Edwards, pers. comm.
Helicos BioSciences HeliScope	Frag, MP/ single molecule	RTs	32*	8‡	37‡	999,000	Non-bias representation of templates for genome and seq-based applications	High error rates compared with other reversible terminator chemistries	Seq-based methods	91
Pacific Biosciences (target release: 2010)	Frag only/ single molecule	Real-time	964*	N/A	N/A	N/A	Has the greatest potential for reads exceeding 1 kb	Highest error rates compared with other NGS chemistries	Full-length transcriptome sequencing; complements other resequencing efforts in discovering large structural variants and haplotype blocks	S. Turner, pers. comm.

Michael Metzker, 2010

Then, by 2014, an ecosystem of options erupted

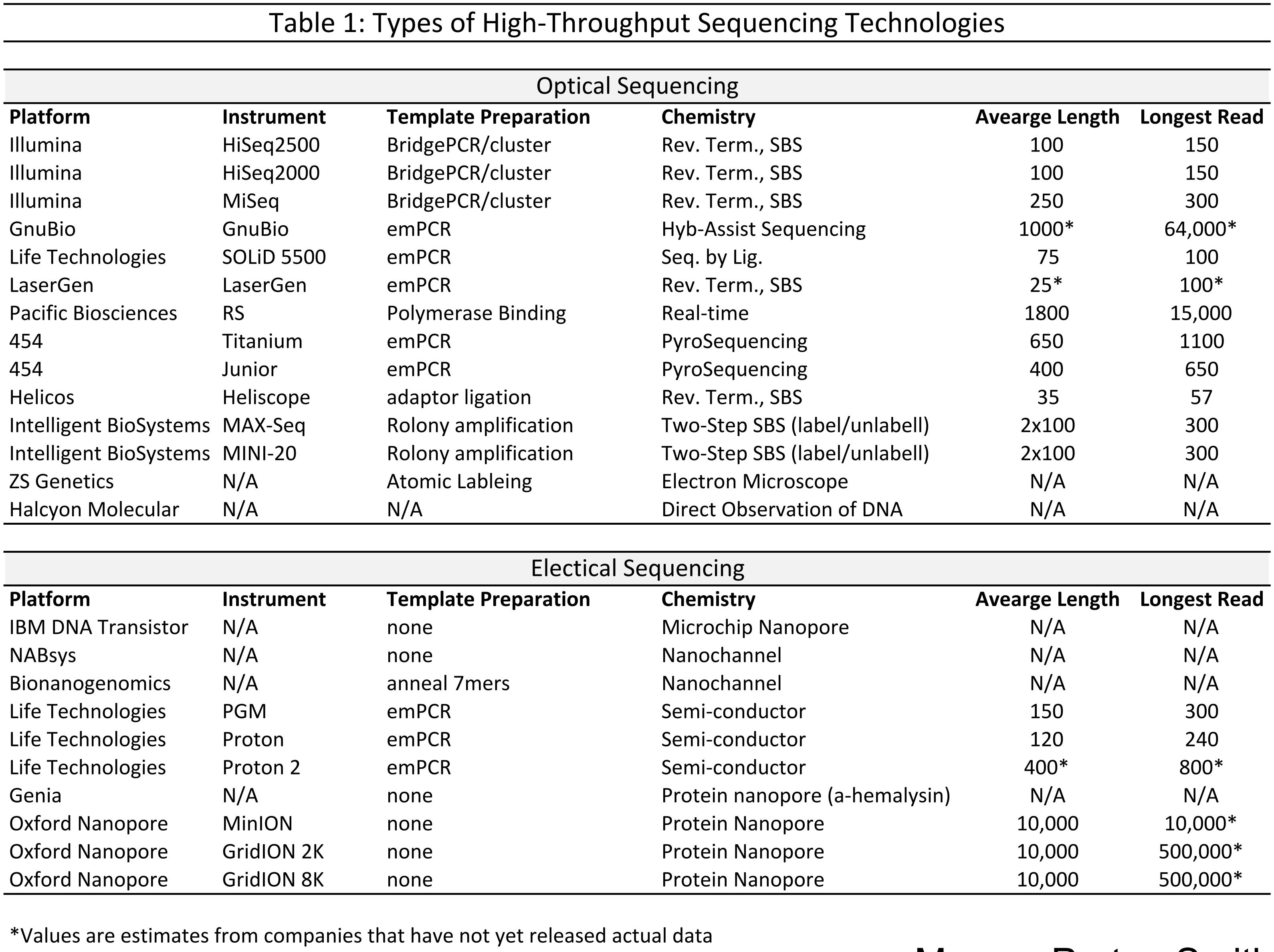


Table 1: Types of High-Throughput Sequencing Technologies

Optical Sequencing

Platform	Instrument	Template Preparation	Chemistry	Avearge Length	Longest Read
Illumina	HiSeq2500	BridgePCR/cluster	Rev. Term., SBS	100	150
Illumina	HiSeq2000	BridgePCR/cluster	Rev. Term., SBS	100	150
Illumina	MiSeq	BridgePCR/cluster	Rev. Term., SBS	250	300
GnuBio	GnuBio	emPCR	Hyb-Assist Sequencing	1000*	64,000*
Life Technologies	SOLiD 5500	emPCR	Seq. by Lig.	75	100
LaserGen	LaserGen	emPCR	Rev. Term., SBS	25*	100*
Pacific Biosciences	RS	Polymerase Binding	Real-time	1800	15,000
454	Titanium	emPCR	PyroSequencing	650	1100
454	Junior	emPCR	PyroSequencing	400	650
Helicos	Heliscope	adaptor ligation	Rev. Term., SBS	35	57
Intelligent BioSystems	MAX-Seq	Rolony amplification	Two-Step SBS (label/unlabell)	2x100	300
Intelligent BioSystems	MINI-20	Rolony amplification	Two-Step SBS (label/unlabell)	2x100	300
ZS Genetics	N/A	Atomic Lableing	Electron Microscope	N/A	N/A
Halcyon Molecular	N/A	N/A	Direct Observation of DNA	N/A	N/A

Electical Sequencing

Platform	Instrument	Template Preparation	Chemistry	Avearge Length	Longest Read
IBM DNA Transistor	N/A	none	Microchip Nanopore	N/A	N/A
NABsys	N/A	none	Nanochannel	N/A	N/A
Bionanogenomics	N/A	anneal 7mers	Nanochannel	N/A	N/A
Life Technologies	PGM	emPCR	Semi-conductor	150	300
Life Technologies	Proton	emPCR	Semi-conductor	120	240
Life Technologies	Proton 2	emPCR	Semi-conductor	400*	800*
Genia	N/A	none	Protein nanopore (a-hemalysin)	N/A	N/A
Oxford Nanopore	MinION	none	Protein Nanopore	10,000	10,000*
Oxford Nanopore	GridION 2K	none	Protein Nanopore	10,000	500,000*
Oxford Nanopore	GridION 8K	none	Protein Nanopore	10,000	500,000*

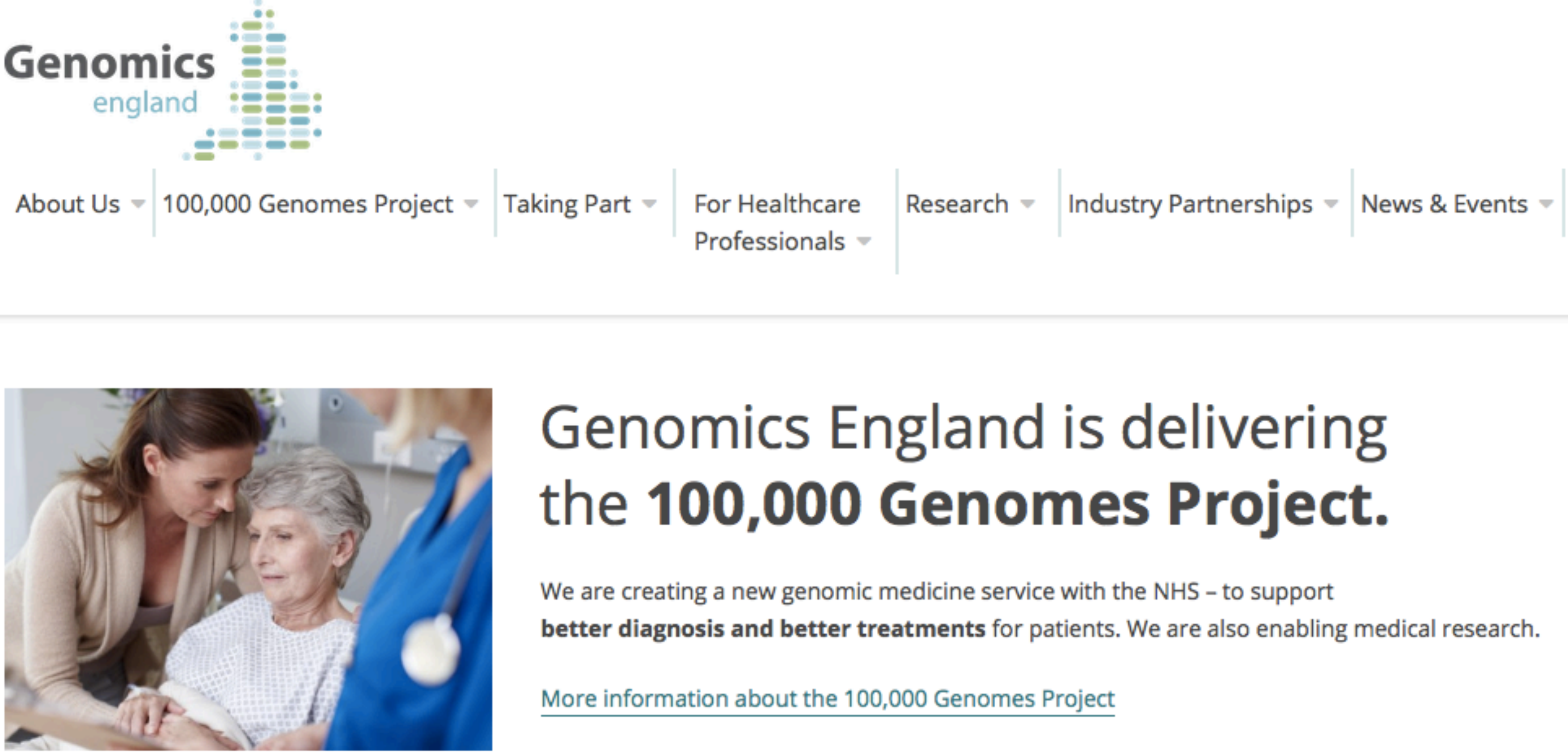
*Values are estimates from companies that have not yet released actual data

Mason, Porter, Smith, 2014

Coming of age: ten years of next-generation sequencing technologies

Sara Goodwin[1], John D. McPherson[2] and W. Richard McCombie[1]

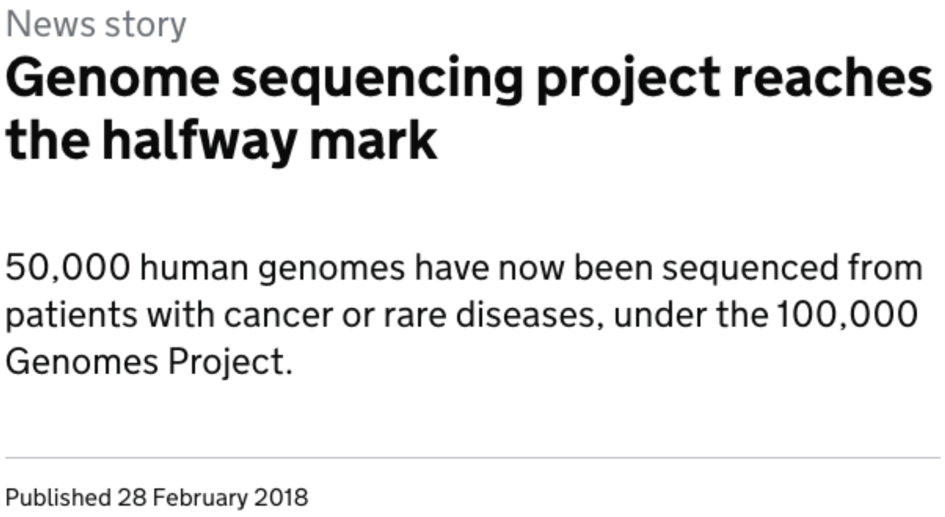
Abstract | Since the completion of the human genome project in 2003, extraordinary progress has been made in genome sequencing technologies, which has led to a decreased cost per megabase and an increase in the number and diversity of sequenced genomes. An astonishing complexity of genome architecture has been revealed, bringing these sequencing technologies to even greater advancements. Some approaches maximize the number of bases sequenced in the least amount of time, generating a wealth of data that can be used to understand increasingly complex phenotypes. Alternatively, other approaches now aim to sequence longer contiguous pieces of DNA, which are essential for resolving structurally complex regions. These and other strategies are providing researchers and clinicians a variety of tools to probe genomes in greater depth, leading to an enhanced understanding of how genome sequence variants underlie phenotype and disease.



Genomics England is delivering the **100,000 Genomes Project.**

We are creating a new genomic medicine service with the NHS – to support **better diagnosis and better treatments** for patients. We are also enabling medical research.

More information about the 100,000 Genomes Project



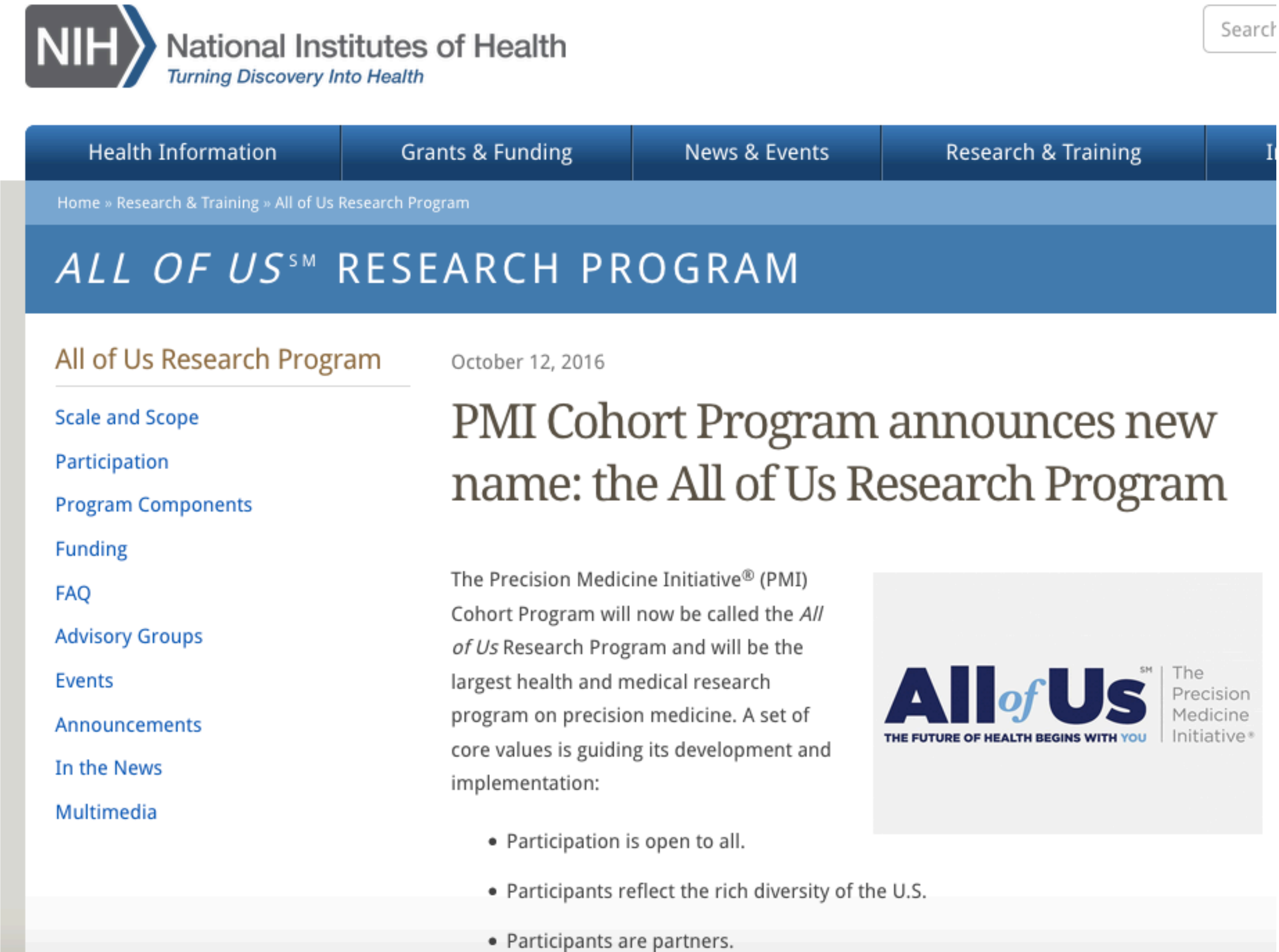
News story

Genome sequencing project reaches the halfway mark

50,000 human genomes have now been sequenced from patients with cancer or rare diseases, under the 100,000 Genomes Project.

Published 28 February 2018

https://www.genomicsengland.co.uk/


NIH
National Institutes of Health
Turning Discovery Into Health
Search
Health Information
Grants & Funding
News & Events
Research & Training
I
Home » Research & Training » All of Us Research Program

ALL OF US℠ RESEARCH PROGRAM

October 12, 2016

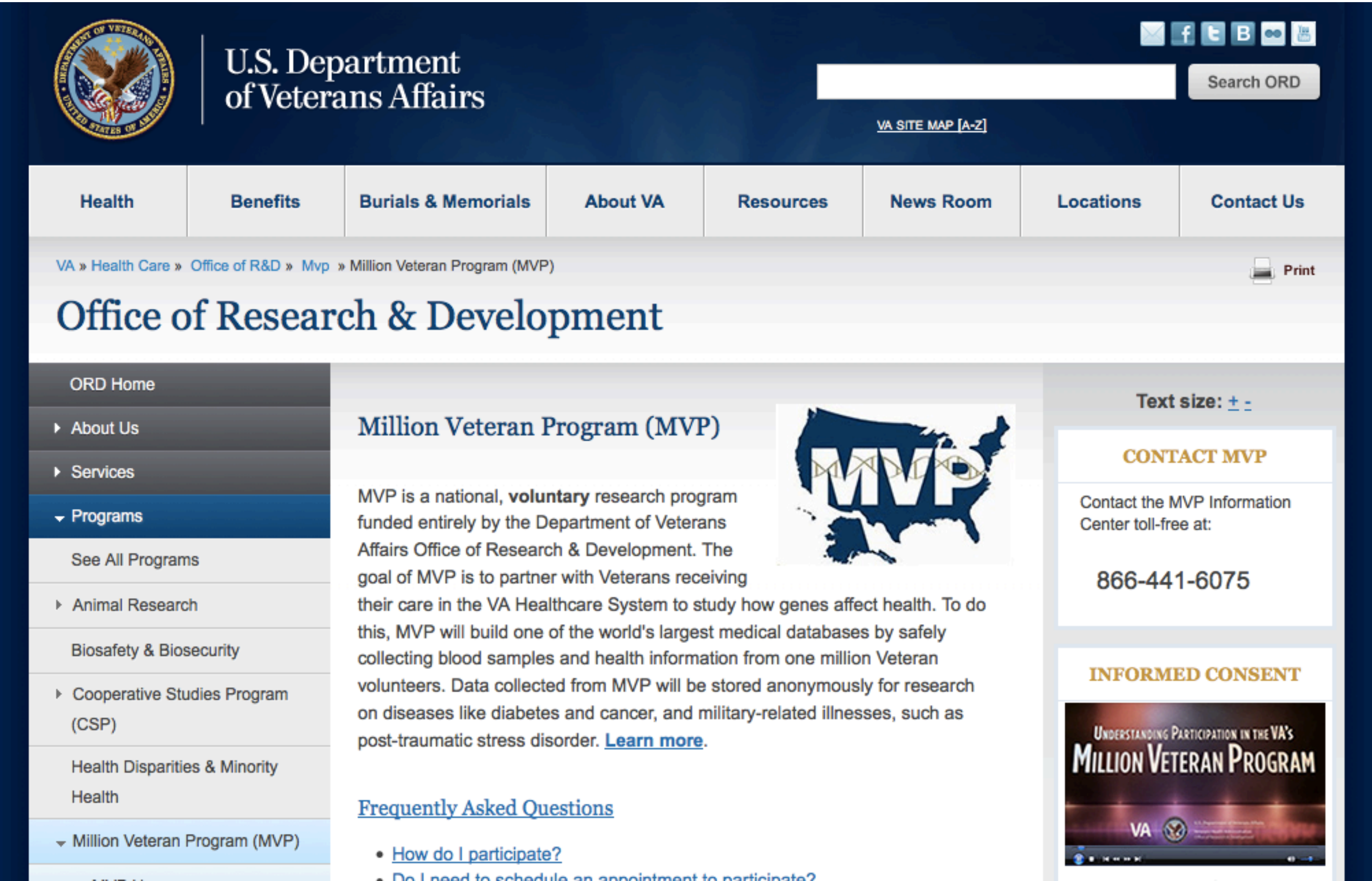
PMI Cohort Program announces new name: the All of Us Research Program

The Precision Medicine Initiative® (PMI) Cohort Program will now be called the *All of Us* Research Program and will be the largest health and medical research program on precision medicine. A set of core values is guiding its development and implementation:

All of Us℠
THE FUTURE OF HEALTH BEGINS WITH YOU
The Precision Medicine Initiative®

- Participation is open to all.
- Participants reflect the rich diversity of the U.S.
- Participants are partners.

1 million U.S. Veterans WGS



Office of Research & Development

Million Veteran Program (MVP)

MVP is a national, **voluntary** research program funded entirely by the Department of Veterans Affairs Office of Research & Development. The goal of MVP is to partner with Veterans receiving their care in the VA Healthcare System to study how genes affect health. To do this, MVP will build one of the world's largest medical databases by safely collecting blood samples and health information from one million Veteran volunteers. Data collected from MVP will be stored anonymously for research on diseases like diabetes and cancer, and military-related illnesses, such as post-traumatic stress disorder. Learn more.

Frequently Asked Questions

- How do I participate?
- Do I need to schedule an appointment to participate?

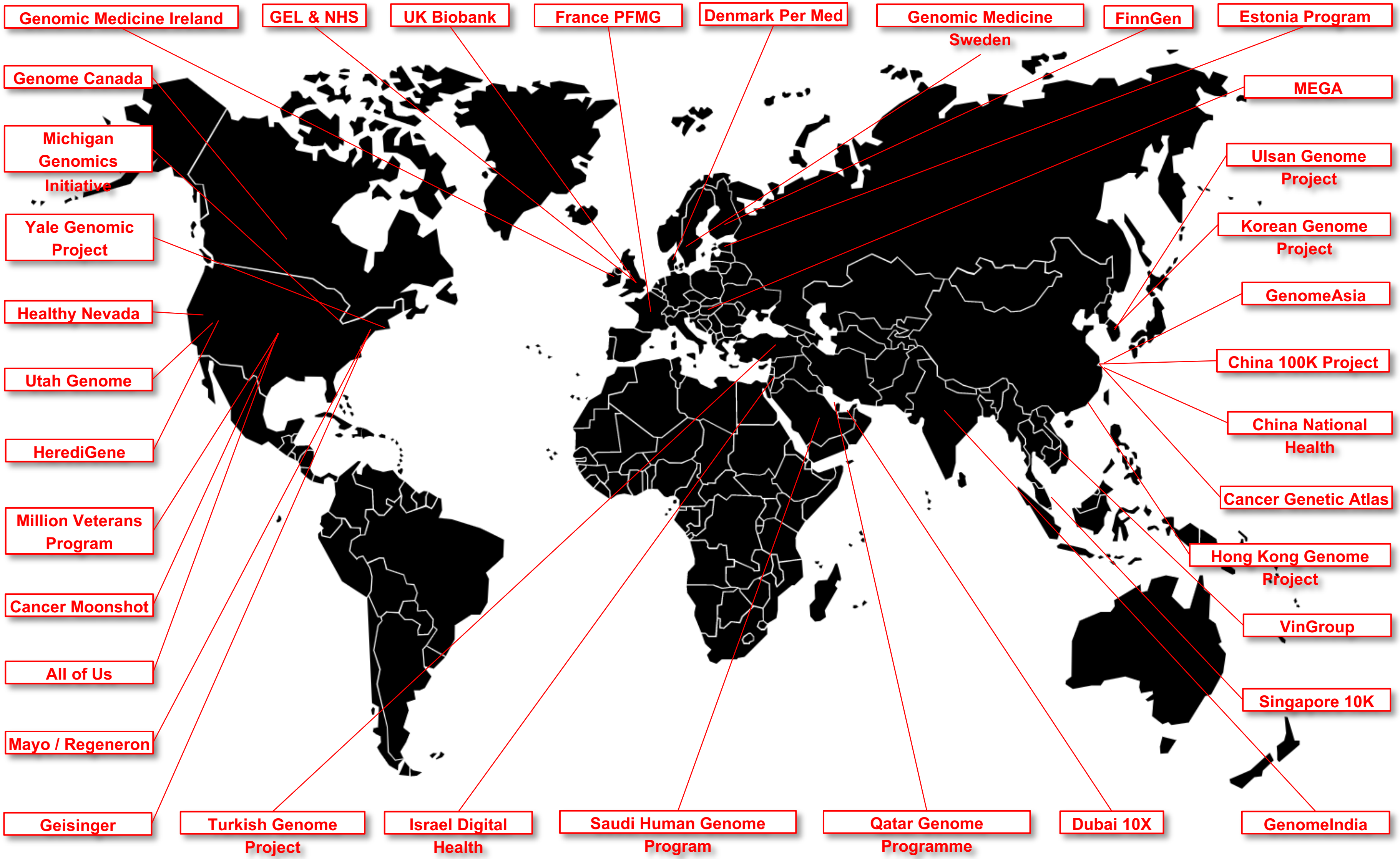
CONTACT MVP

Contact the MVP Information Center toll-free at:

866-441-6075

INFORMED CONSENT

POPULATION-SCALE NGS 2020 IS GLOBAL



Genomic Medicine Ireland

GEL & NHS

UK Biobank

France PFMG

Denmark Per Med

Genomic Medicine Sweden

FinnGen

Estonia Program

Genome Canada

Michigan Genomics Initiative

Yale Genomic Project

Healthy Nevada

Utah Genome

HerediGene

Million Veterans Program

Cancer Moonshot

All of Us

Mayo / Regeneron

Geisinger

MEGA

Ulsan Genome Project

Korean Genome Project

GenomeAsia

China 100K Project

China National Health

Cancer Genetic Atlas

Hong Kong Genome Project

VinGroup

Singapore 10K

GenomeIndia

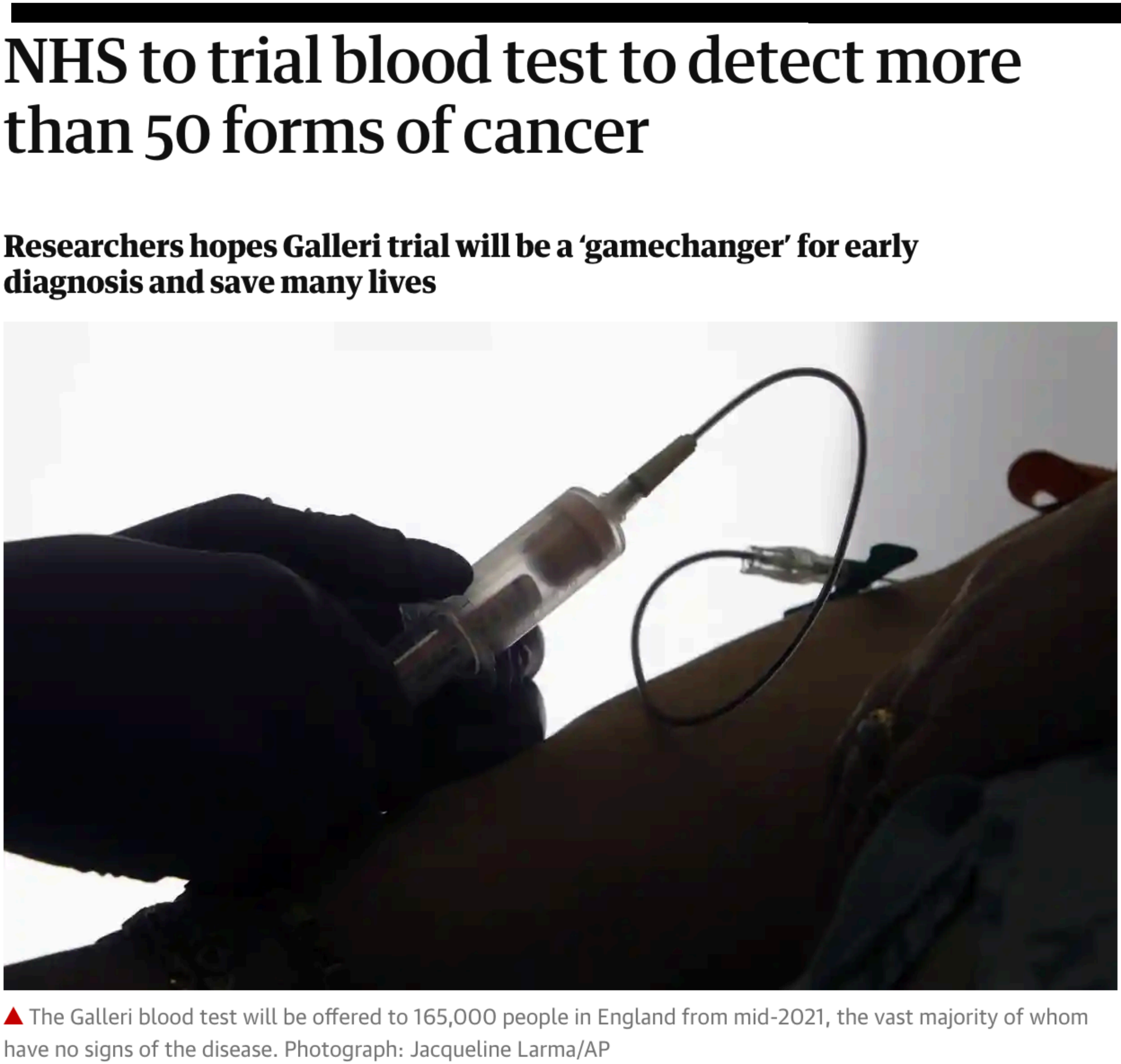
Turkish Genome Project

Israel Digital Health

Saudi Human Genome Program

Qatar Genome Programme

Dubai 10X



NHS to trial blood test to detect more than 50 forms of cancer

[redacted]

Researchers hopes Galleri trial will be a 'gamechanger' for early diagnosis and save many lives

▲ The Galleri blood test will be offered to 165,000 people in England from mid-2021, the vast majority of whom have no signs of the disease. Photograph: Jacqueline Larma/AP

Offered to 165,000 people in England from mid-2021,
with no signs of disease.

Followed through 2023;
If successful, move on to test 1M people in 2024-2025.

https://www.theguardian.com/science/2020/nov/27/nhs-to-trial-blood-test-to-detect-more-than-50-forms-of-cancer

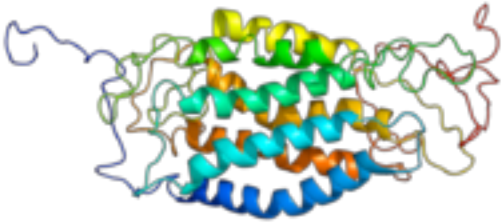
Specific genes can have significant impact

Myostatin (MSTN) homozygous nulls (-/-) give lean and large muscles

http://thevoiceofnetizen.blogspot.com

Low density lipoprotein receptor 5 (LRP5) heterozygotes (+/-) can have strong bones

C-C chemokine receptor type 5 (CCR5) homozygous nulls (-/-) have HIV protection



Carcinogenesis vol.29 no.4 pp.762–765, 2008
doi:10.1093/carcin/bgn044
Advance Access publication February 14, 2008

Constitutional CHEK2 mutations are associated with a decreased risk of lung and laryngeal cancers

Cezary Cybulski*, Bartłomiej Masojć, Dorota Oszutowska, Ewa Jaworowska[1], Tomasz Grodzki[2], Piotr Waloszczyk[2], Piotr Serwatowski[2], Juliusz Pankowski[2], Tomasz Huzarski, Tomasz Byrski, Bohdan Górski, Anna Jakubowska, Tadeusz Dębniak, Dominika Wokołorczyk, Jacek Gronwald, Czesława Tarnowska[1], Pablo Serrano-Fernández, Jan Lubiński and Steven A.Narod[3]

International Hereditary Cancer Center, Department of Genetics and Pathology, Pomeranian Medical University, ul. Połabska 4, 70-115 Szczecin, Poland, [1]Department of Otolaryngology and Laryngological Oncology, Pomeranian Medical University, ul.Unii Lubelskiej, 71–252 Szczecin, Poland, [2]Lung Diseases Hospital, ul. Sokołowskiego 11, 70–891 Szczecin, Poland and [3]Women's College Research Institute, Toronto, Ontario M5G IN8, Canada

*To whom correspondence should be addressed. Tel: +48 91 466 1532;
Fax: +48 91 466 1533;
Email: cezarycy@sci.pam.szczecin.pl

Mutations in the *CHEK2* gene have been associated with increased risks of breast, prostate and colon cancer. In contrast, a previous report suggests that individuals with the I157T missense variant of the *CHEK2* gene might be at decreased risk of lung cancer and upper aero-digestive cancers. To confirm this hypothesis, we genotyped 895 cases of lung cancer, 430 cases of laryngeal cancer and 6391 controls from Poland for four founder alleles in the *CHEK2* gene, each of which has been associated with an increased risk of cancer at several sites. The presence of a *CHEK2* mutation was protective against both lung cancer [odds ratio (OR) = 0.3; 95% confidence interval (CI) 0.2–0.5; $P = 3 \times 10^{-8}$] and laryngeal cancer (OR = 0.6; 95% CI 0.3–0.99; $P = 0.05$). The basis of the protective effect is unknown, but may relate to the reduced viability of lung cancer cells with a *CHEK2* mutation. Lung cancers frequently possess other defects in genes in the DNA damage response pathway (e.g. *p53* mutations) and have a high level of genotoxic DNA damage induced by tobacco smoke. We speculate that lung cancer cells with impaired *CHEK2* function undergo increased rates of cell death.

Introduction

Germ line mutations in *CHEK2* have been associated with a range of cancer types, in particular of the breast and the prostate, but cancers of

of Brennan *et al.* We have extended our series of lung cancer cases from 272 to 895 and our control sample from 4000 to 6391. We have also identified a fourth deleterious *CHEK2* allele (a large deletion of exons 9 and 10). Because smoking is the principal risk factor for lung cancer in Poland and elsewhere, we asked whether the protective effect of *CHEK2* might extend to laryngeal cancer patients as well.

Materials and methods

We studied 895 unselected cases of lung cancer (226 women and 669 men) diagnosed in the Lung Diseases Hospital in Szczecin, Poland, between 2004 and 2006. We also ascertained 430 consecutive, unselected patients with squamous cell carcinoma of the larynx (70 women and 360 men) at Department of Otolaryngology and Laryngological Oncology of the Pomeranian Medical University, Szczecin, Poland, during the period 2001–2004. Patients were recruited from the oncology services of the contributing hospitals and were unselected for age or family history. Patients were approached by a member of the study team during an outpatient visit to the oncology clinic and were asked if they wished to participate. Patient acceptance rates exceeded 80% for both cancer sites. Patients provided written informed consent. A blood sample of 10 cc was then drawn for DNA extraction. Two hundred and seventy-two of the lung cancer patients have been included in our previous study (5). The mean age of diagnosis of the lung cancer patients was 61.4 years (range 29–88 years) and of the laryngeal cancer patients was 58.2 years (range 30–84). Patients completed a questionnaire about their smoking habits at the time of cancer diagnosis. Smoking histories were available for 818 of 895 (91%) lung cancer cases and for 387 of 430 (90%) laryngeal cancer cases. The study was approved by the Ethics Committee of the Pomeranian Medical University in Szczecin.

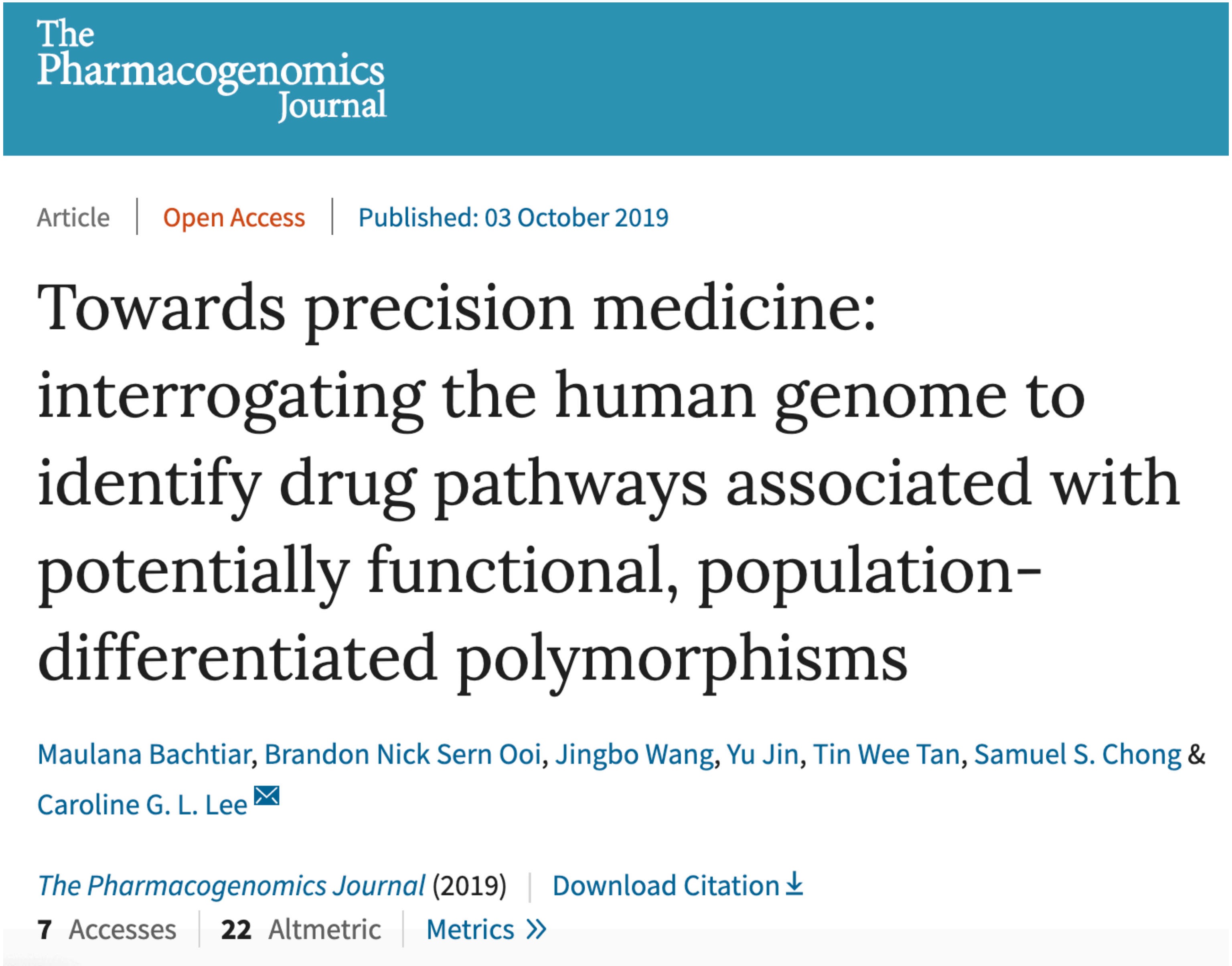
Unmatched analysis

In the unmatched analysis, four non-overlapping control groups were combined in order to maximize the number of controls.

The first control group of 1896 healthy adults, including 1079 women (age range 15–91, mean 58.3) and 817 men (age range 23–90, mean 59.4). These controls were selected at random from the computerized patient lists of five large family practices located in the region of Szczecin. These healthy adults were invited to participate by mail and participated in 2003 and 2004. Participation rates for this group exceeded 70%. During the interview, the goals of the study were explained, informed consent was obtained, genetic counselling was given and a blood sample was taken for DNA analysis. A detailed family history of cancer was taken (first- and second-degree relatives included). Probands were included regardless of their cancer family history status. Individuals affected with any malignancy were excluded from the study.

The second control group consisted of 1417 unselected young adults (705 women and 712 men; age range 18–35, mean 24.3) from Szczecin metropolitan region who submitted a blood sample for paternity testing between 1994 and 2001.

The third control group consisted of 2183 children from nine cities in Poland



The Pharmacogenomics Journal

Article | Open Access | Published: 03 October 2019

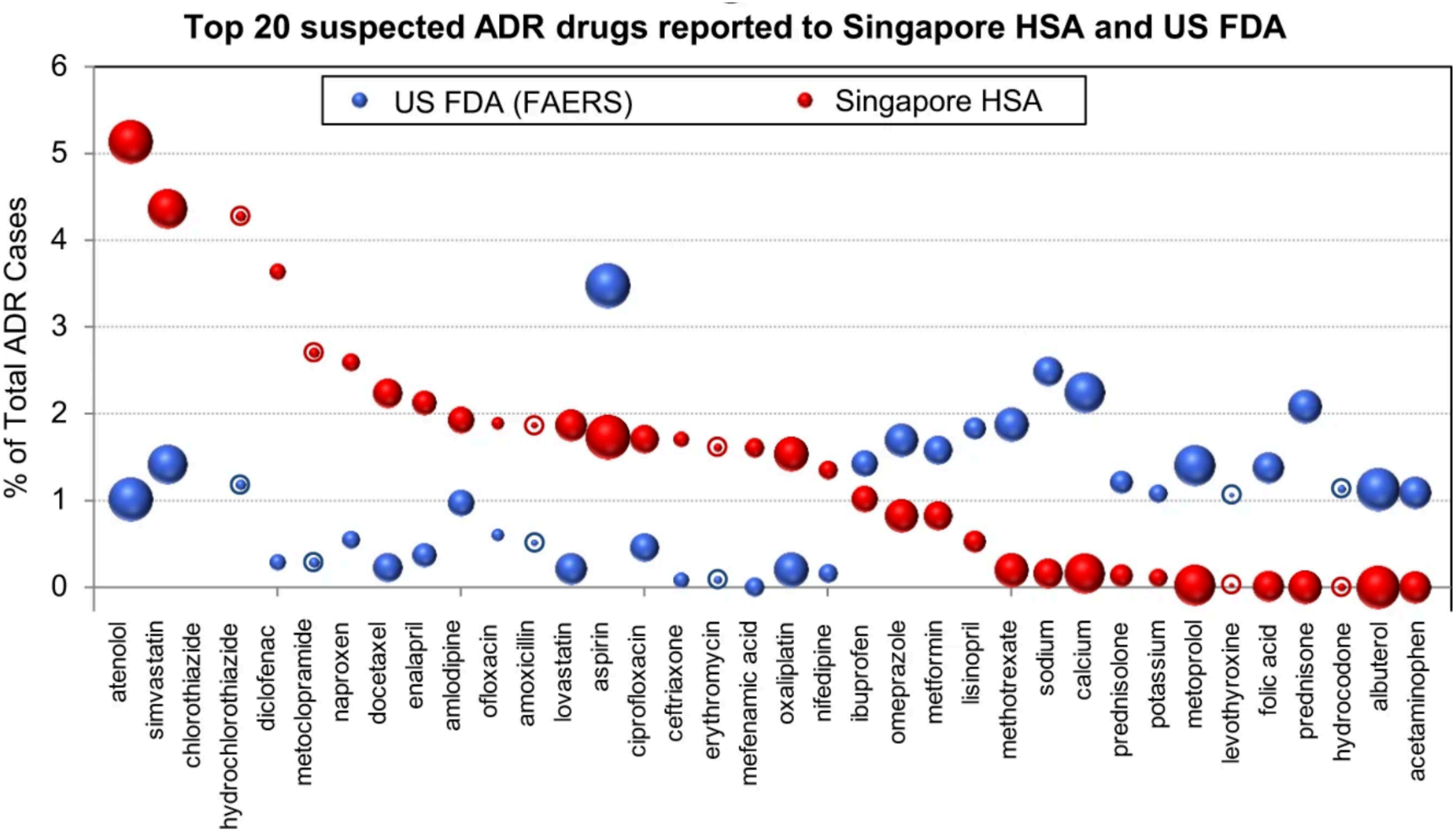
Towards precision medicine: interrogating the human genome to identify drug pathways associated with potentially functional, population-differentiated polymorphisms

Maulana Bachtiar, Brandon Nick Sern Ooi, Jingbo Wang, Yu Jin, Tin Wee Tan, Samuel S. Chong & Caroline G. L. Lee ✉

The Pharmacogenomics Journal (2019) | Download Citation ⤓

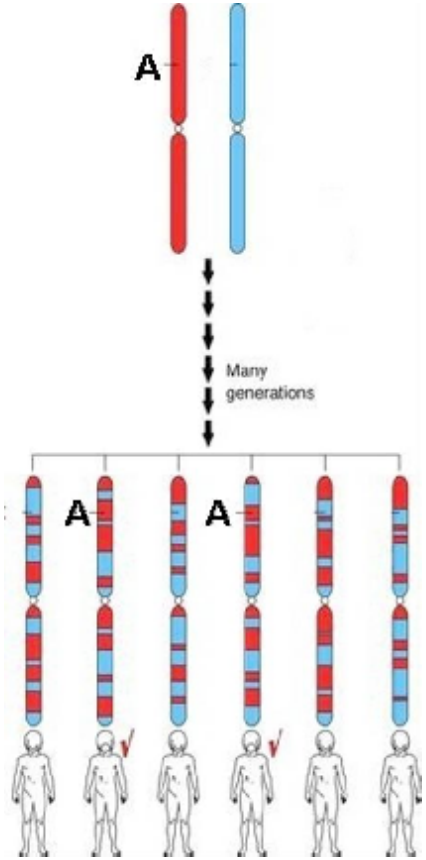
7 Accesses | 22 Altmetric | Metrics »

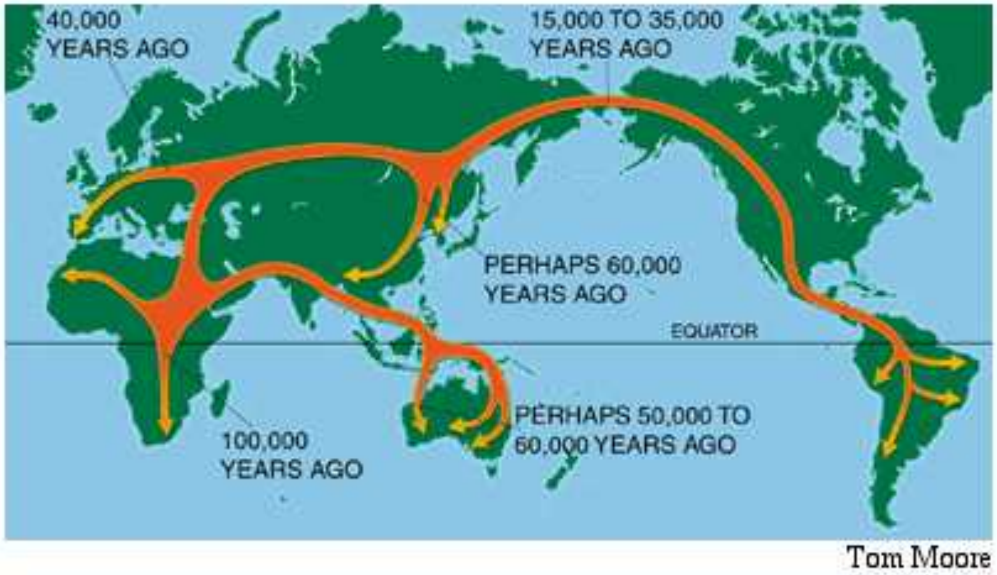
https://www.nature.com/articles/s41397-019-0096-y



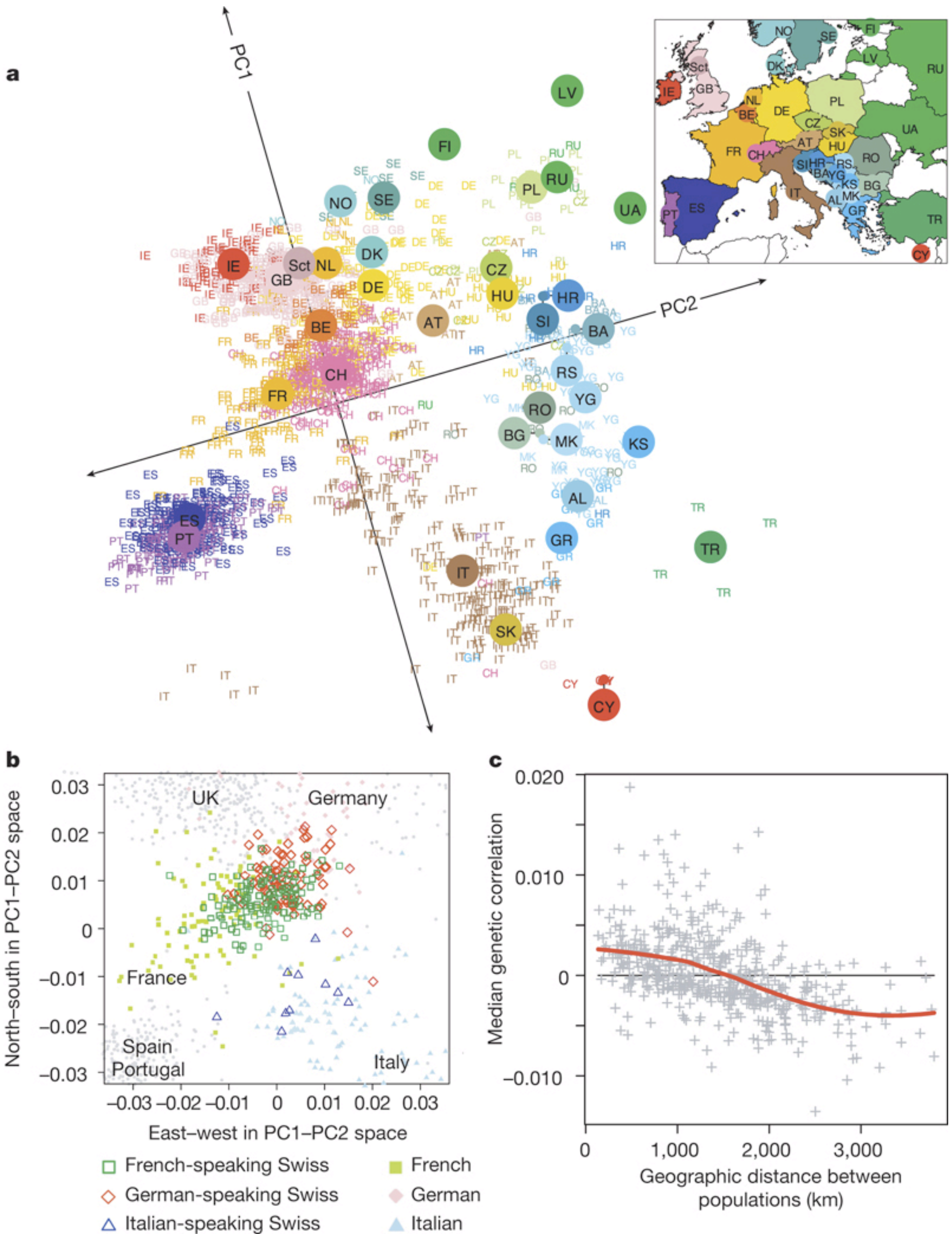
https://www.nature.com/articles/s41397-019-0096-y

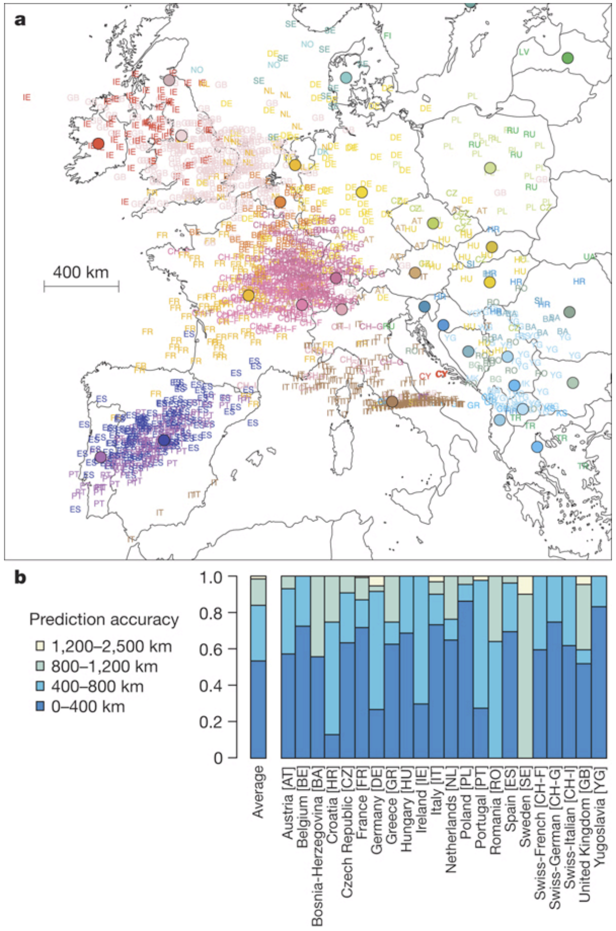
Our genes come from the migration patterns of haplotypes throughout human history ("Population Stratification")





Genotype data can even predict your birthplace



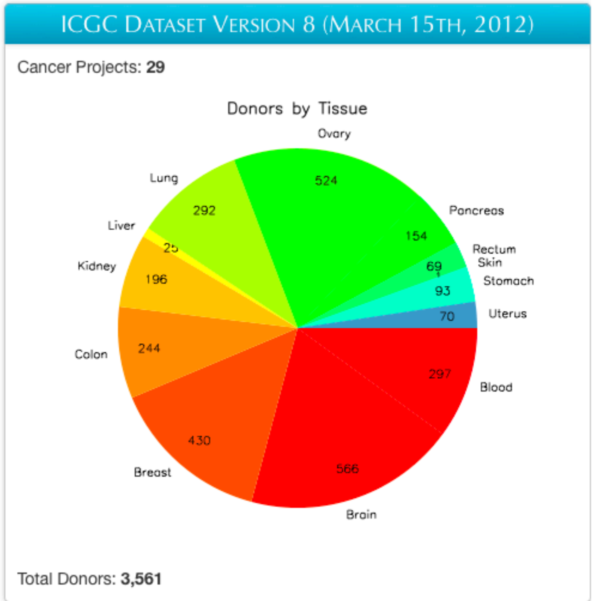


Genes mirror geography within Europe
Novembre *et al.*, 2008

Large impact for normal genomes and diseases, especially cancer

1000 Genomes
A Deep Catalog of Human Genetic Variation

The Cancer Genome Atlas
Data Portal
Understanding genomics to improve cancer care



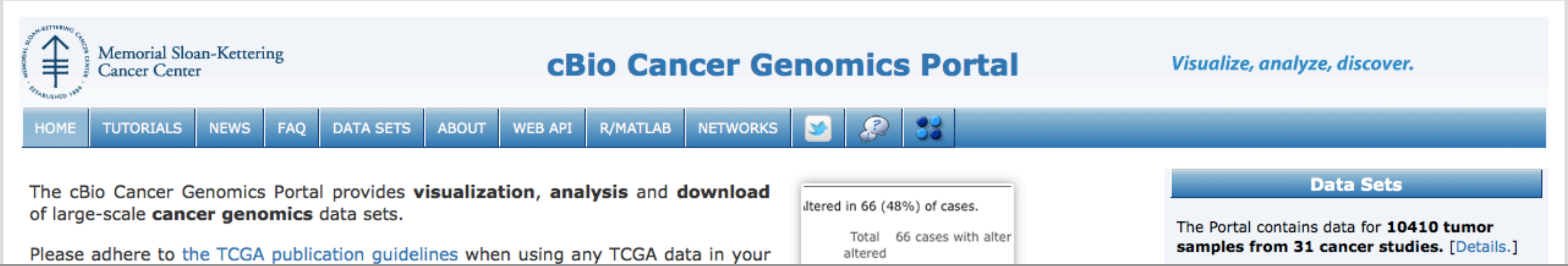
ICGC DATASET VERSION 8 (MARCH 15TH, 2012)

Cancer Projects: **29**

Total Donors: **3,561**

International Cancer Genome Consortium

ICGC Goal: To obtain a comprehensive description of genomic, epigenomic, and transcriptomic (GET) changes in 50 different tumor types and/or subtypes which are of clinical and societal importance across the globe.

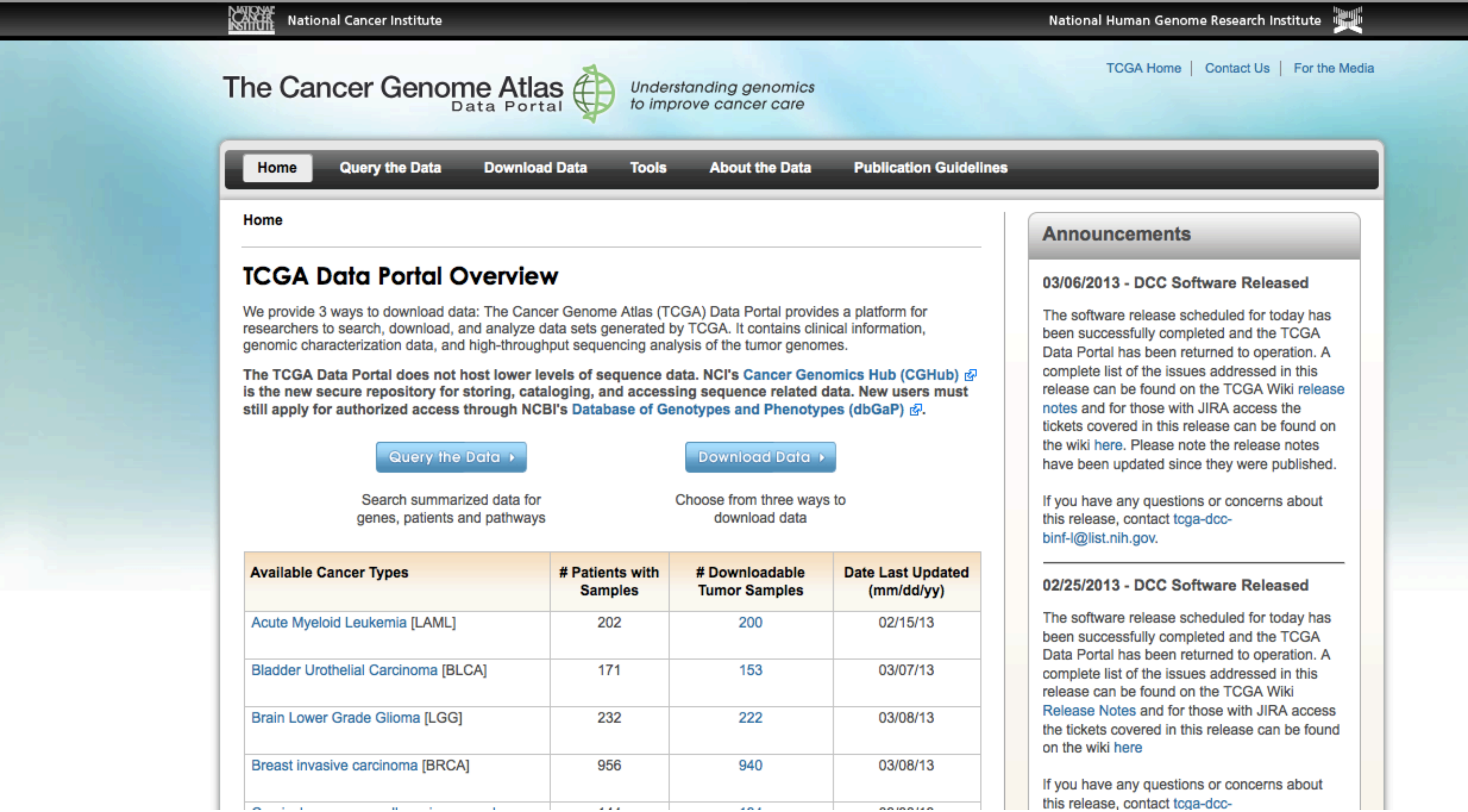

Memorial Sloan-Kettering Cancer Center
cBio Cancer Genomics Portal
Visualize, analyze, discover.
HOME
TUTORIALS
NEWS
FAQ
DATA SETS
ABOUT
WEB API
R/MATLAB
NETWORKS
ltered in 66 (48%) of cases.
Total altered
66 cases with alter

The cBio Cancer Genomics Portal provides **visualization**, **analysis** and **download** of large-scale **cancer genomics** data sets.

Please adhere to the TCGA publication guidelines when using any TCGA data in your

Data Sets

The Portal contains data for **10410 tumor samples from 31 cancer studies.** [Details.]


National Cancer Institute
National Human Genome Research Institute
TCGA Home | Contact Us | For the Media
The Cancer Genome Atlas Data Portal
Understanding genomics to improve cancer care
Home
Query the Data
Download Data
Tools
About the Data
Publication Guidelines
Home

TCGA Data Portal Overview

We provide 3 ways to download data: The Cancer Genome Atlas (TCGA) Data Portal provides a platform for researchers to search, download, and analyze data sets generated by TCGA. It contains clinical information, genomic characterization data, and high-throughput sequencing analysis of the tumor genomes.

The TCGA Data Portal does not host lower levels of sequence data. NCI's Cancer Genomics Hub (CGHub) is the new secure repository for storing, cataloging, and accessing sequence related data. New users must still apply for authorized access through NCBI's Database of Genotypes and Phenotypes (dbGaP).

Query the Data ›

Search summarized data for genes, patients and pathways

Download Data ›

Choose from three ways to download data

Available Cancer Types	# Patients with Samples	# Downloadable Tumor Samples	Date Last Updated (mm/dd/yy)
Acute Myeloid Leukemia [LAML]	202	200	02/15/13
Bladder Urothelial Carcinoma [BLCA]	171	153	03/07/13
Brain Lower Grade Glioma [LGG]	232	222	03/08/13
Breast invasive carcinoma [BRCA]	956	940	03/08/13

Announcements

03/06/2013 - DCC Software Released

The software release scheduled for today has been successfully completed and the TCGA Data Portal has been returned to operation. A complete list of the issues addressed in this release can be found on the TCGA Wiki release notes and for those with JIRA access the tickets covered in this release can be found on the wiki here. Please note the release notes have been updated since they were published.

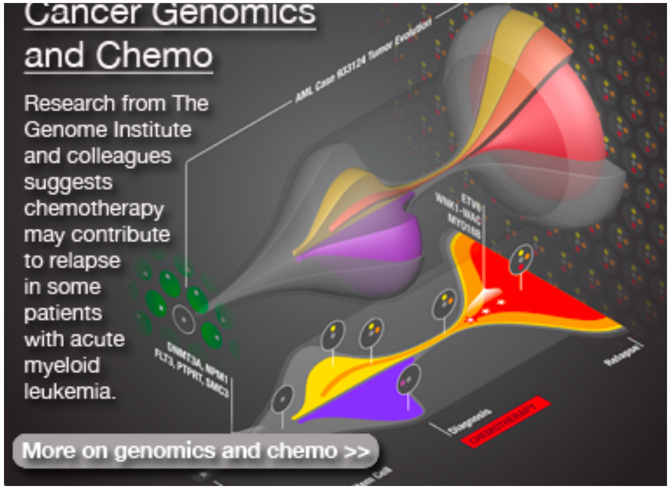
If you have any questions or concerns about this release, contact tcga-dcc-binf-l@list.nih.gov.

02/25/2013 - DCC Software Released

The software release scheduled for today has been successfully completed and the TCGA Data Portal has been returned to operation. A complete list of the issues addressed in this release can be found on the TCGA Wiki Release Notes and for those with JIRA access the tickets covered in this release can be found on the wiki here

If you have any questions or concerns about this release, contact tcga-dcc-

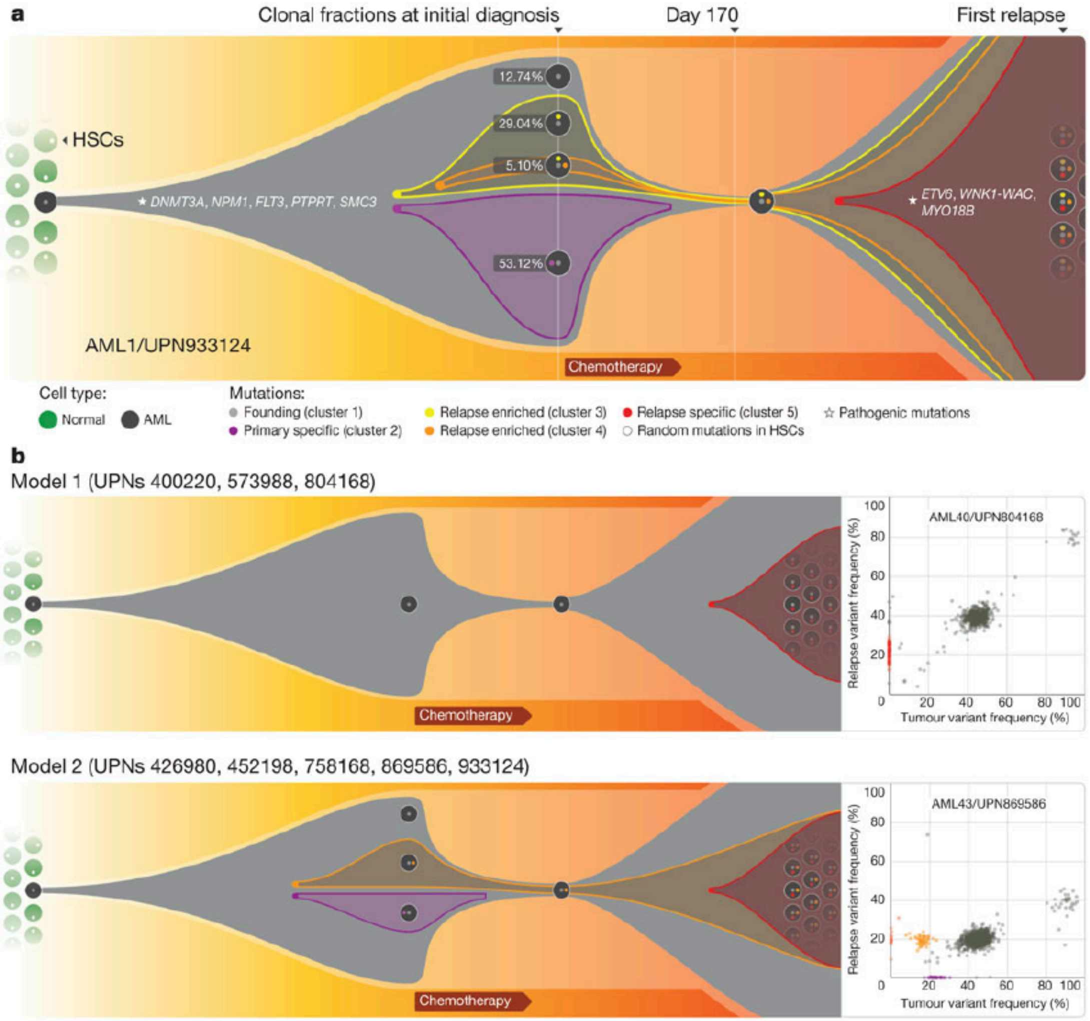
We can also observe the dynamics and evolution of cancers



Cancer Genomics and Chemo

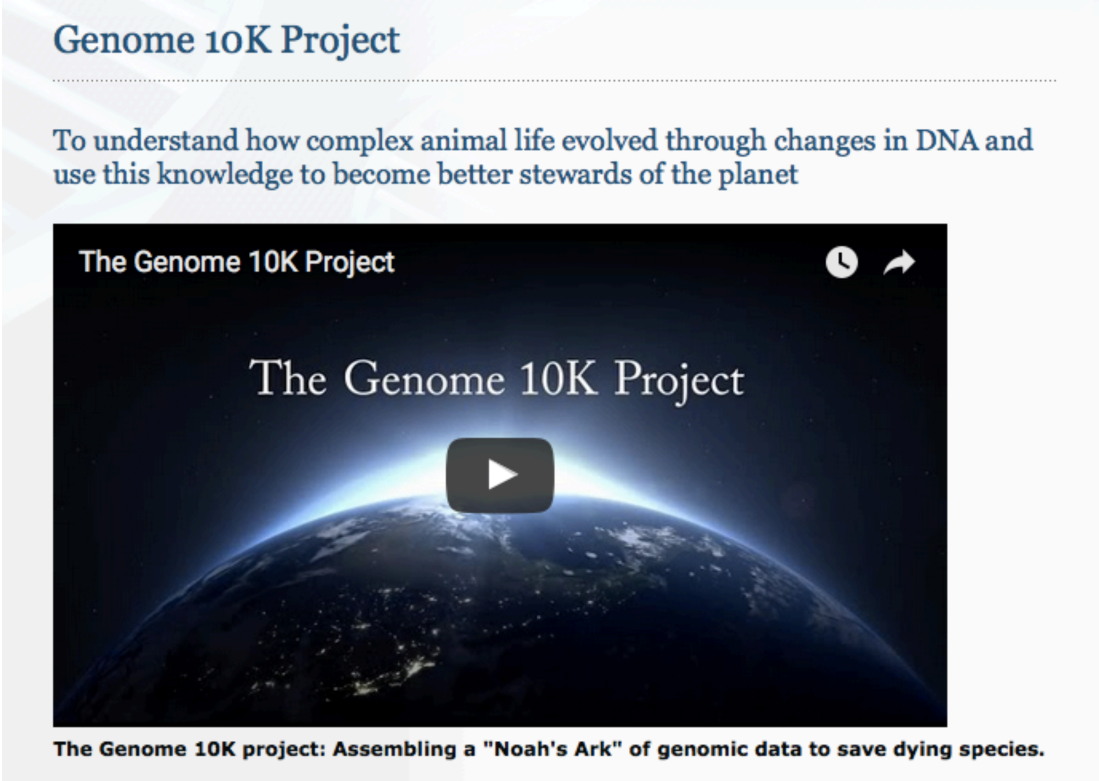
Research from The Genome Institute and colleagues suggests chemotherapy may contribute to relapse in some patients with acute myeloid leukemia.

More on genomics and chemo >>



Ding L, et.al, Clonal evolution in relapsed acute myeloid leukemia revealed by whole-genome sequencing. Nature. 2012 Jan 11;481(7382):506-10.

And look beyond just humans



Genome 10K Project

To understand how complex animal life evolved through changes in DNA and use this knowledge to become better stewards of the planet

The Genome 10K project: Assembling a "Noah's Ark" of genomic data to save dying species.

https://genome10k.soe.ucsc.edu/

https://www.hgsc.bcm.edu/i5k-pilot-project-summary

Plants as well!

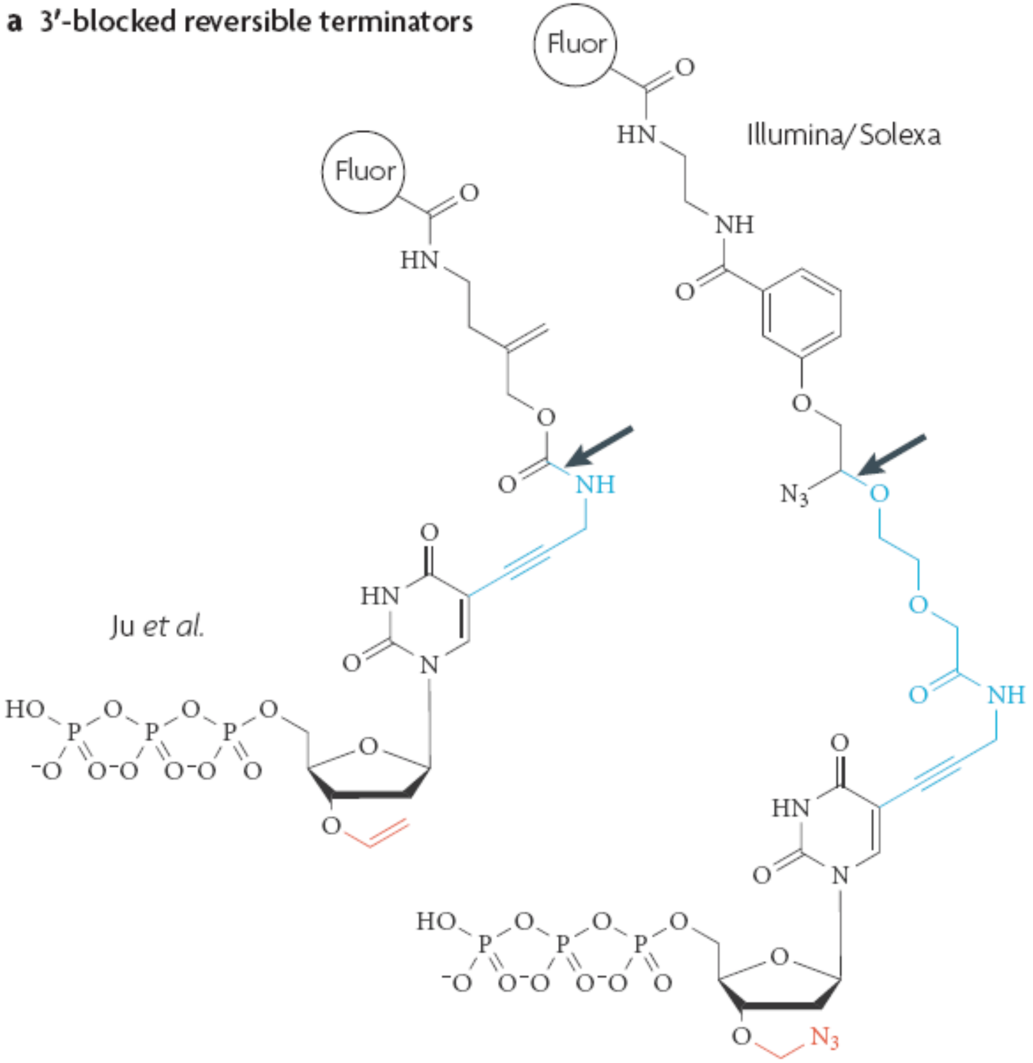
华大基因
BGI

Tree of life sequencing project in BGI

http://ldl.genomics.cn/page/pa-research.jsp

Consideration of WGS for each platform

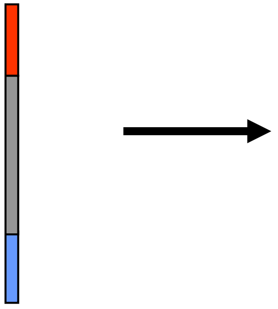
Reversible Terminator Bases are Essential Technology Used in Many Chemistries



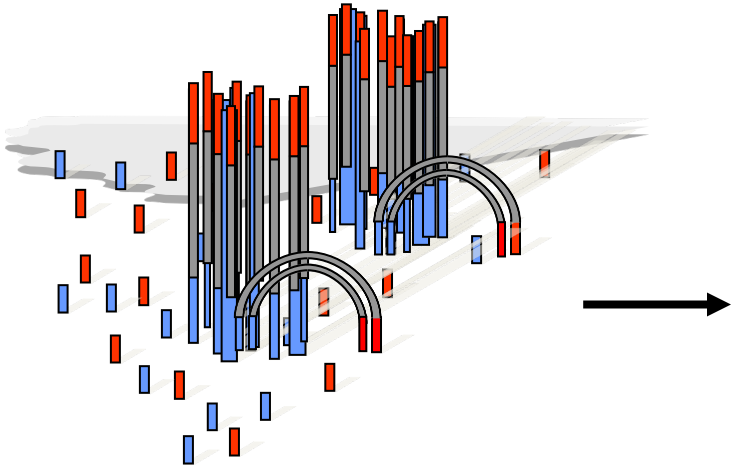
Illumina SBS Technology

Reversible Terminator Chemistry Foundation

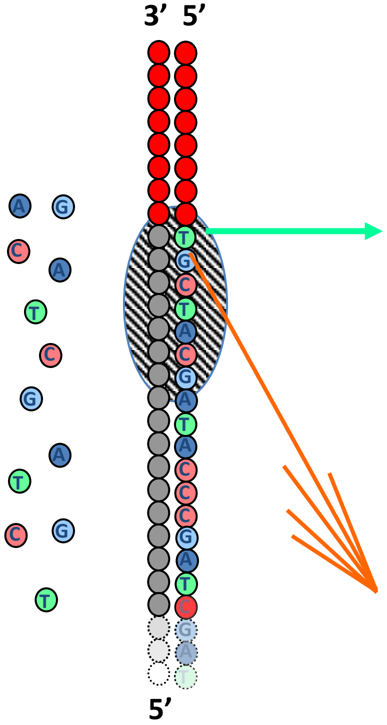
DNA
(0.1-1.0 ug)



Sample preparation



Cluster growth



Sequencing

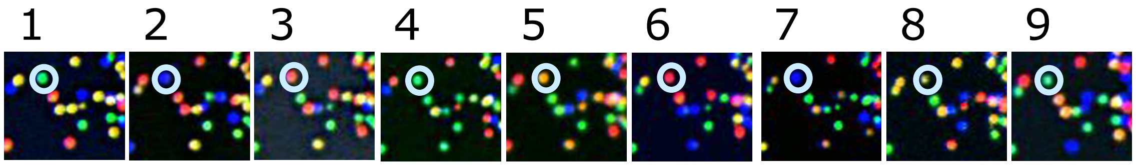
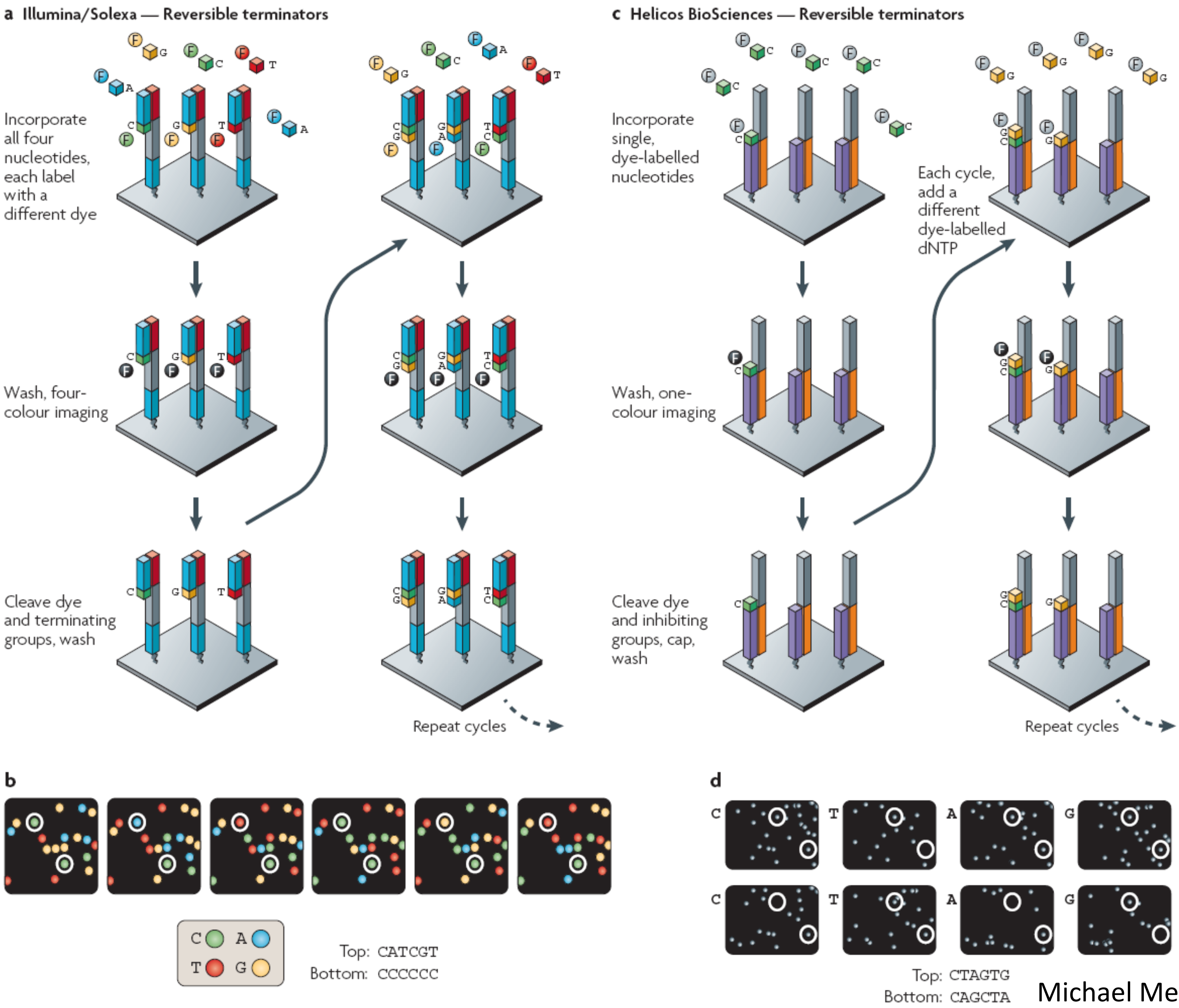


Image acquisition

Base calling

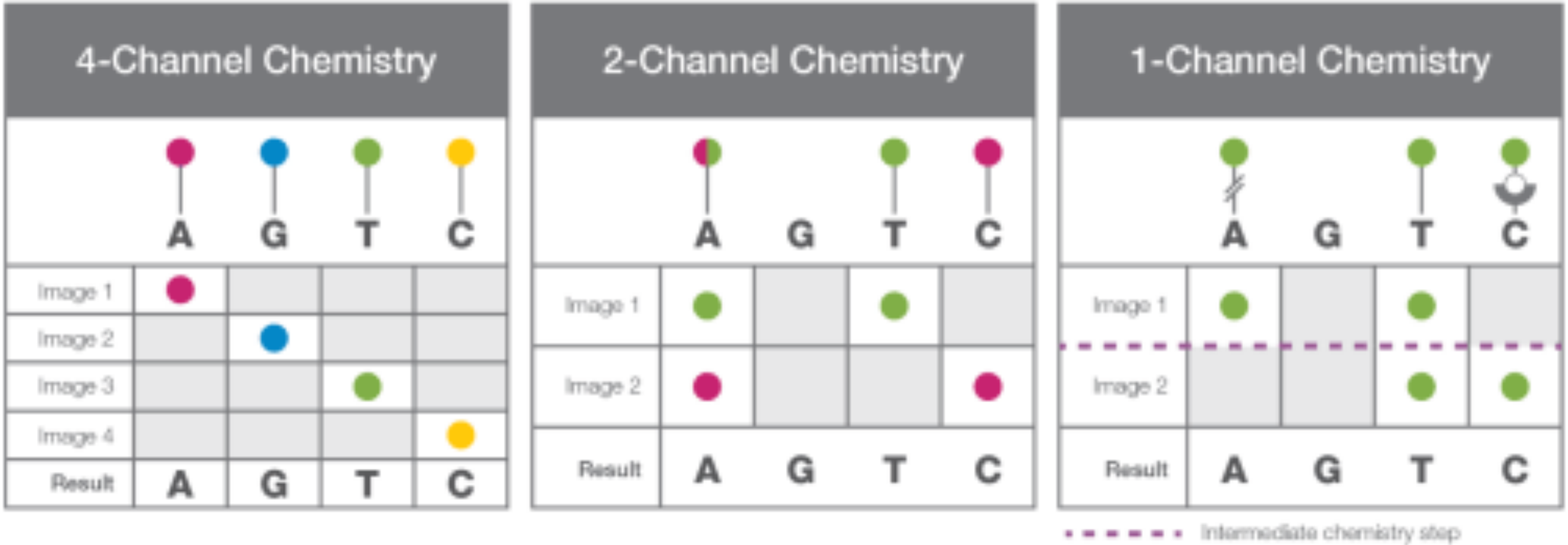
http://www.illumina.com/technology/sequencing_technology.ilmn

Sequencing by Synthesis (SBS)


a Illumina/Solexa — Reversible terminators
Incorporate all four nucleotides, each label with a different dye
Wash, four-colour imaging
Cleave dye and terminating groups, wash
Repeat cycles
F
A
C
G
T
c Helicos BioSciences — Reversible terminators
Incorporate single, dye-labelled nucleotides
Each cycle, add a different dye-labelled dNTP
Wash, one-colour imaging
Cleave dye and inhibiting groups, cap, wash
Repeat cycles
b
C A
T G
Top: CATCGT
Bottom: CCCCCC
d
C
T
A
G
C
T
A
G
Top: CTAGTG
Bottom: CAGCTA

Michael Metzker, 2010

Now three kinds of chemistry



4-Channel Chemistry

	A	G	T	C
Image 1	🔴			
Image 2		🔵		
Image 3			🟢	
Image 4				🟡
Result	A	G	T	C

2-Channel Chemistry

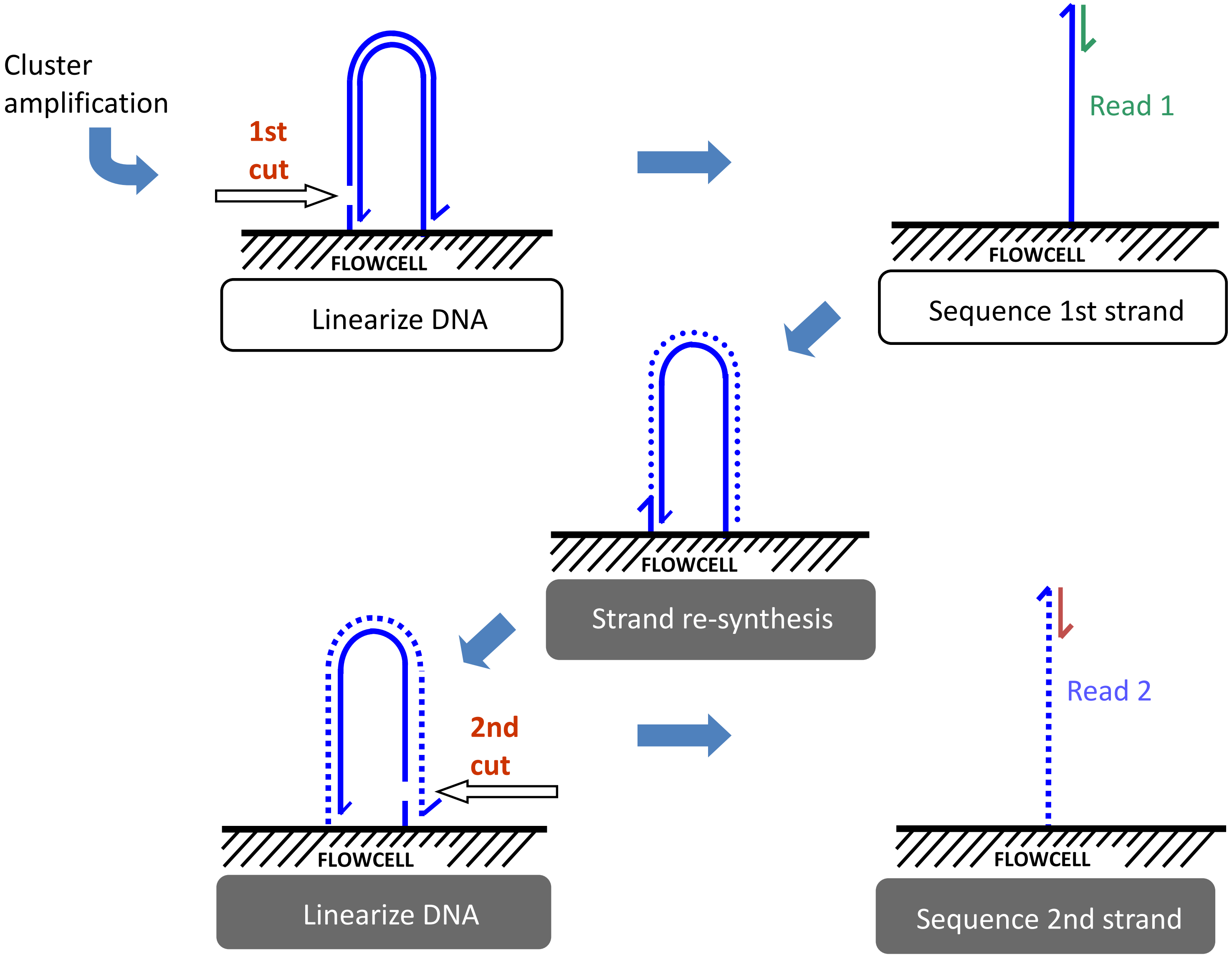
	A	G	T	C
Image 1	🟢		🟢	
Image 2	🔴			🔴
Result	A	G	T	C

1-Channel Chemistry

	A	G	T	C
Image 1	🟢		🟢	
Image 2			🟢	🟢
Result	A	G	T	C

Figure 2: Four-, Two-, and One-Channel Chemistry—Four-channel chemistry uses a mixture of nucleotides labeled with four different fluorescent dyes. Two-channel chemistry uses two different fluorescent dyes, and one-channel chemistry uses only one dye. The images are processed by image analysis software to determine nucleotide identity.

Paired-End Sequencing allows for two looks at a sequence



Cluster amplification

Linearize DNA

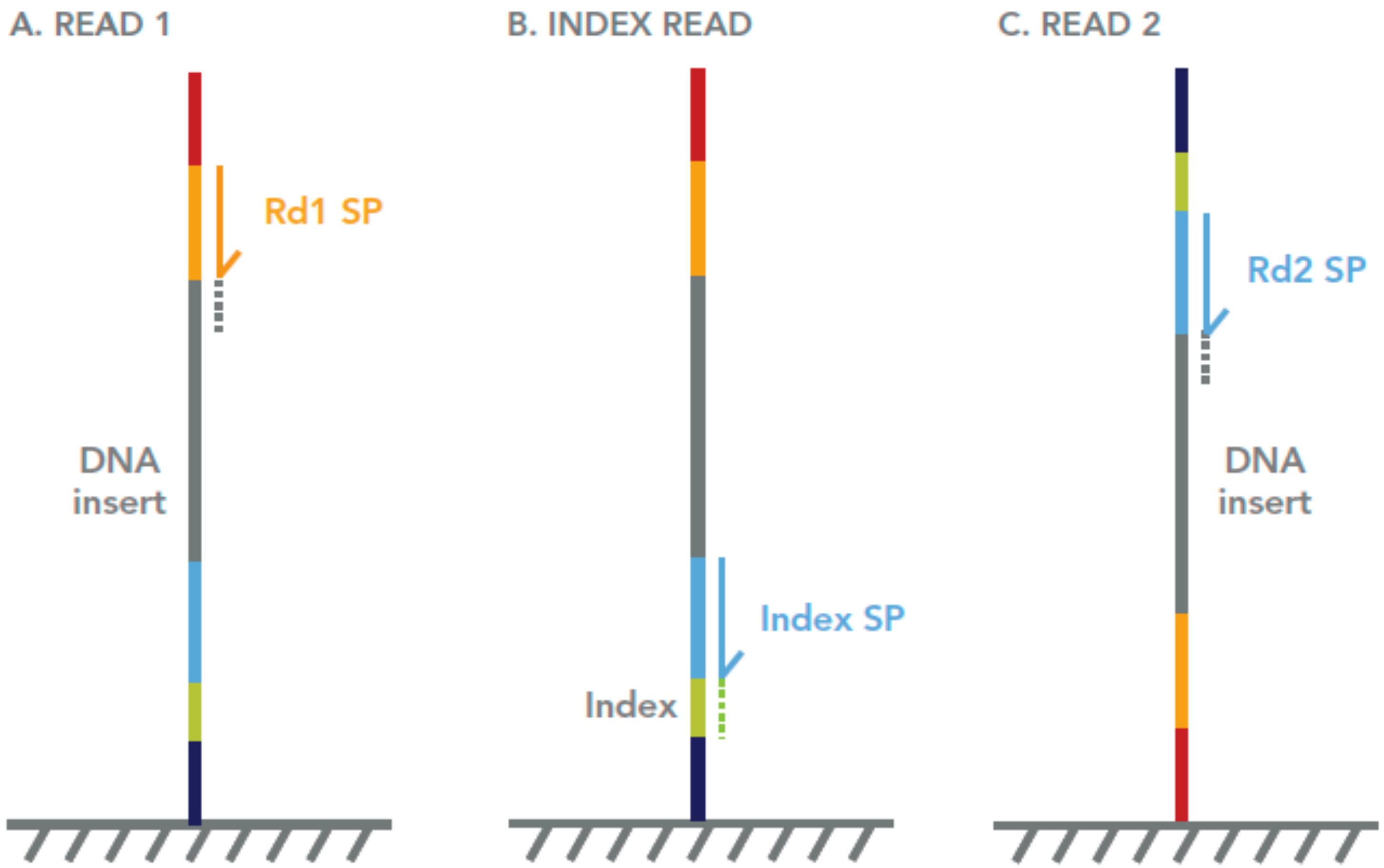
Sequence 1st strand

Strand re-synthesis

Linearize DNA

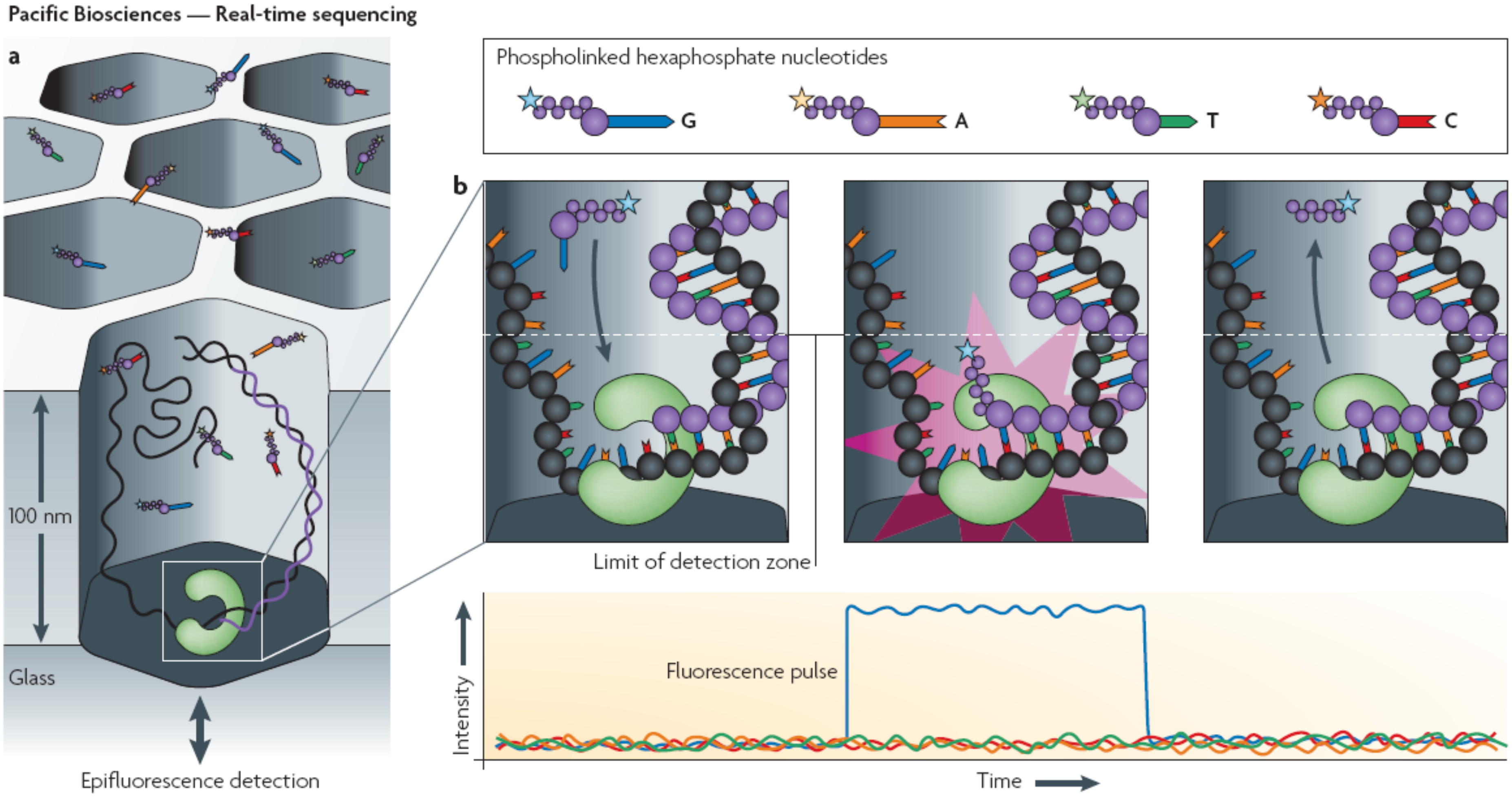
Sequence 2nd strand

Indexed sequencing method is now standard for single and paired reads



© Illumina, Inc.

Pacific Biosciences
Single Molecule Real-Time (SMRT) Sequencing


Pacific Biosciences — Real-time sequencing
a
100 nm
Glass
Epifluorescence detection
Phospholinked hexaphosphate nucleotides
G
A
T
C
b
Limit of detection zone
Fluorescence pulse
Intensity
Time

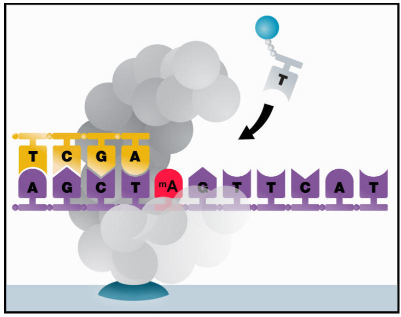
https://www.pacb.com/videos/video-overview-of-smrt-technology/

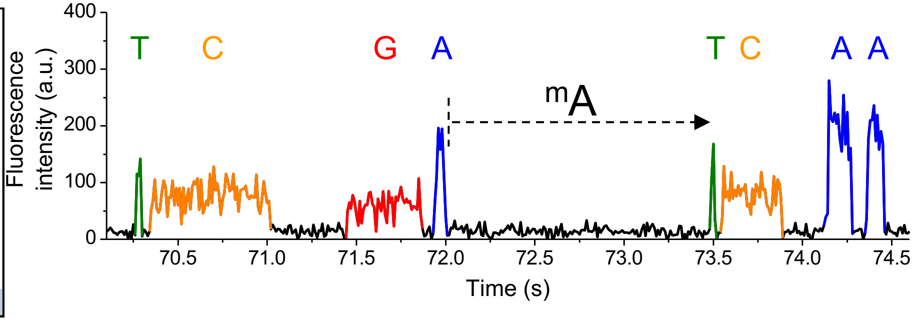
Metzker, 2010

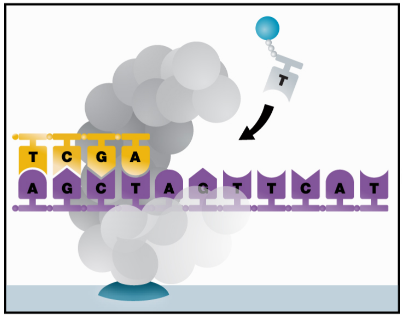
Single Molecule Kinetics Allow for the Direct Detection of Methylation

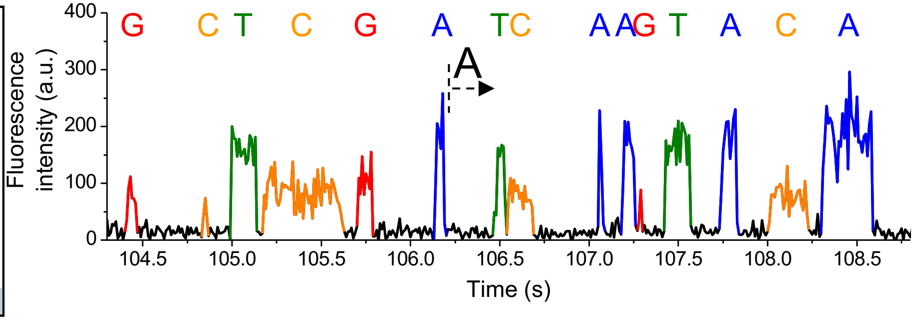
Approach: Kinetic detection of methylated bases during SMRT DNA sequencing

Example: N^6-methyladenosine (^{m}A)



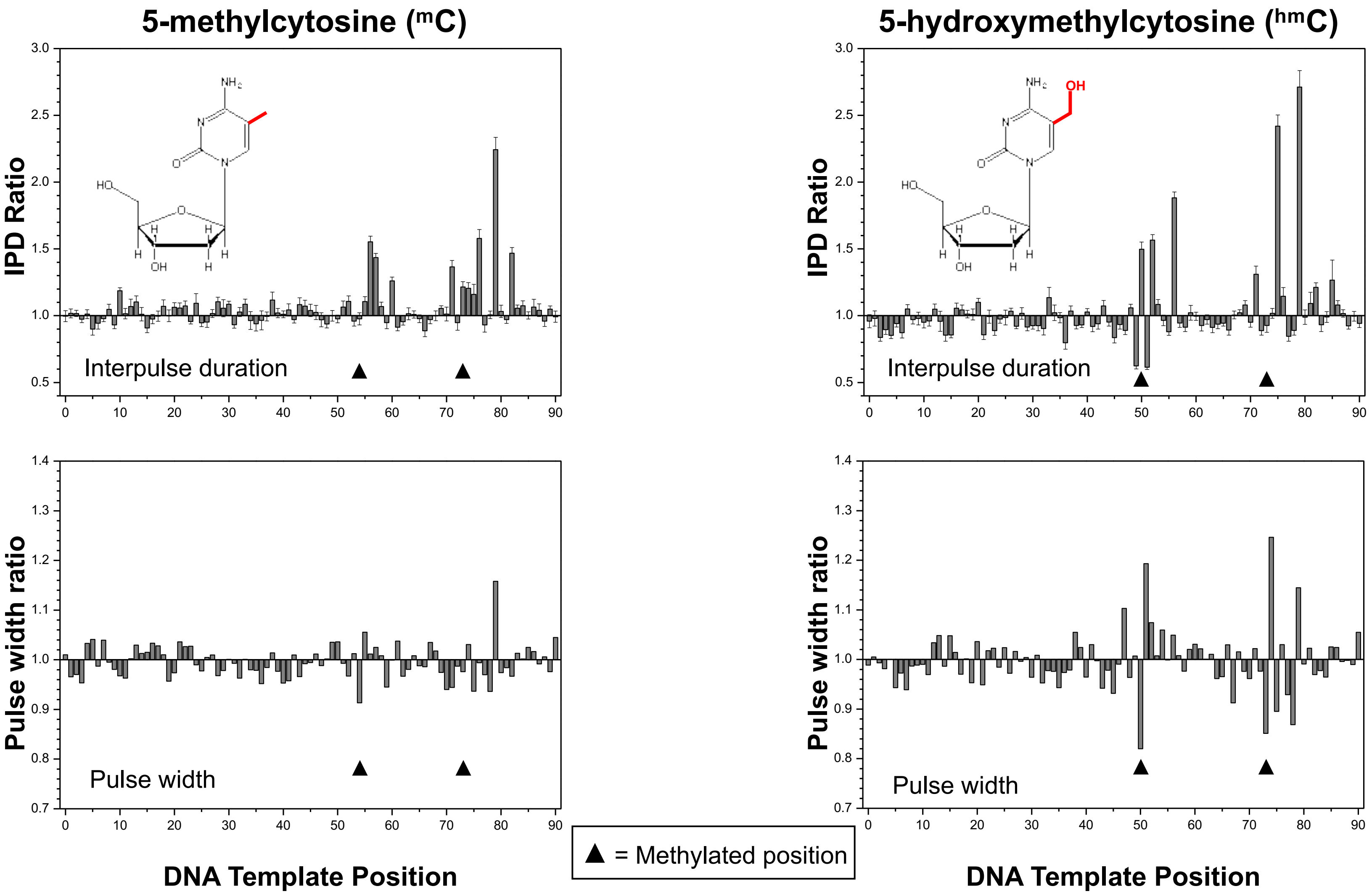




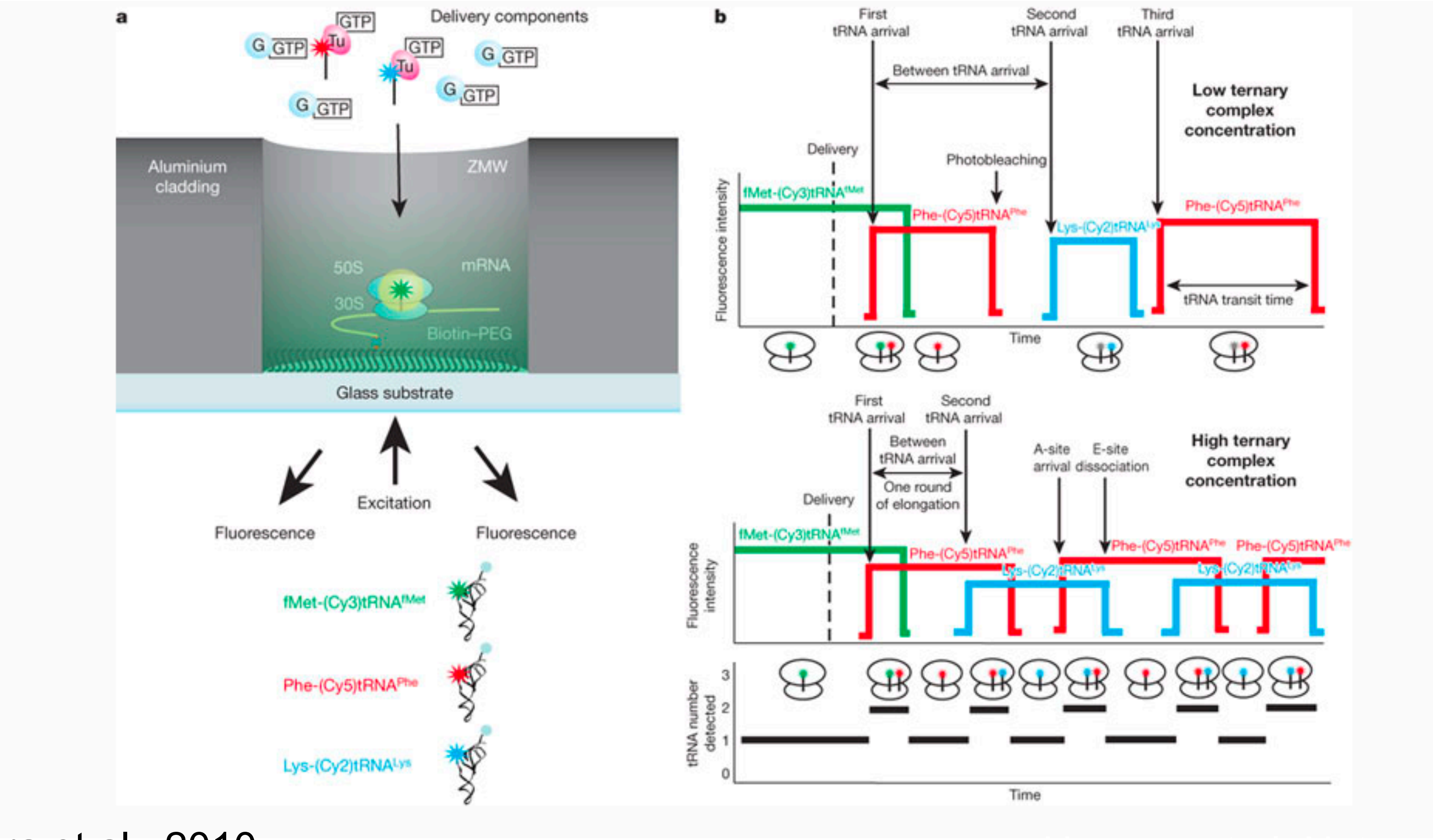


Flusberg et al., 2010.

Kinetics can detect other base modifications



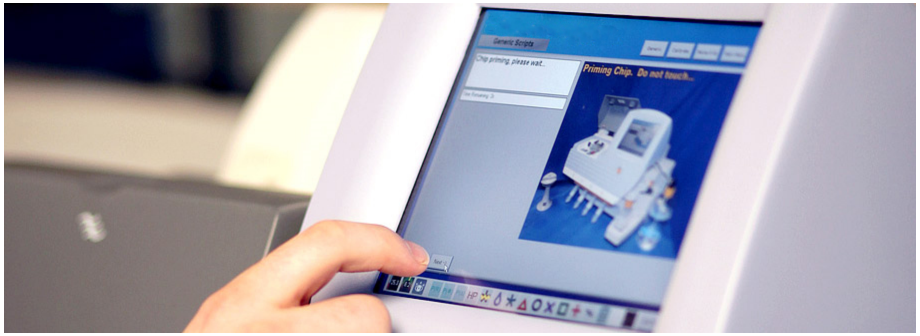
Kinetics allow one to watch protein translation as it occurs

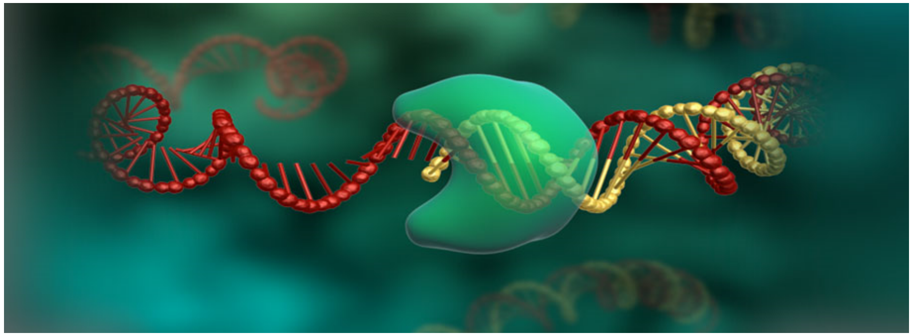


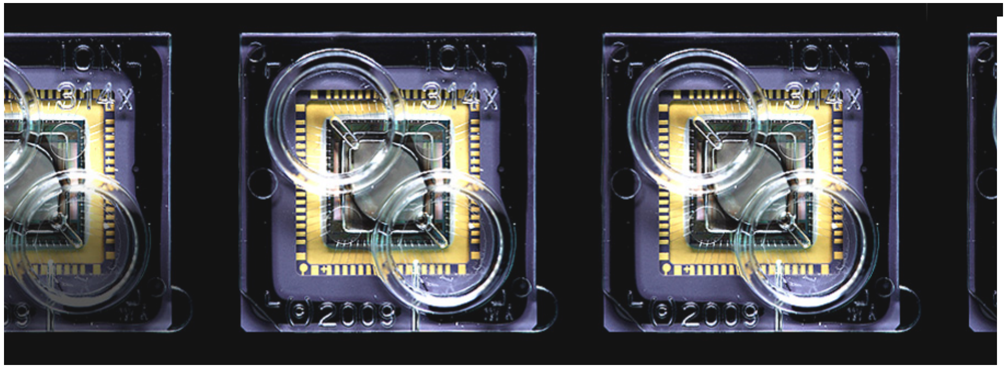
Uemura et al., 2010

"Post-Light," Semi-Conductor Sequencing:

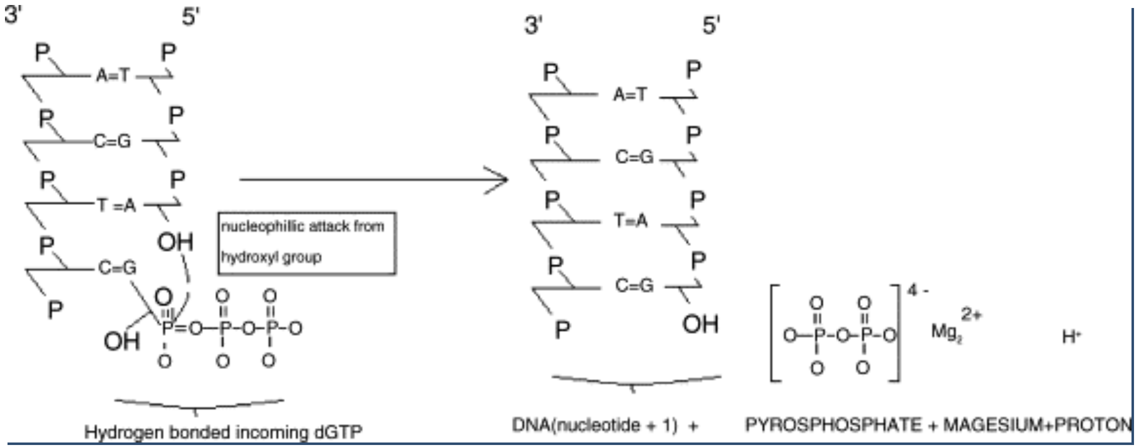
Thermo Fisher's Personal Genome Machine (PGM), the Proton I and Proton II, and S5







Simple chip loading
Intuitive graphical user interface
State-of-the-art electronics
Integrated reagent delivery


3'
5'
A=T
C=G
T=A
C=G
OH
nucleophillic attack from hydroxyl group
Hydrogen bonded incoming dGTP
DNA(nucleotide + 1) +
PYROPHOSPHATE + MAGESIUM+PROTON
Mg
H+

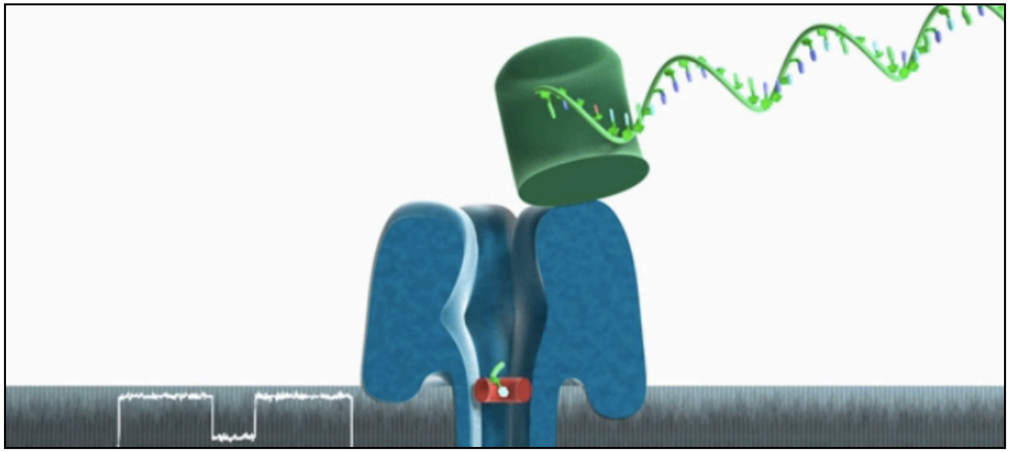
Essentially, Millions of very small pH meters

Purushothaman *et al*, 2005
IonTorrent, Inc.

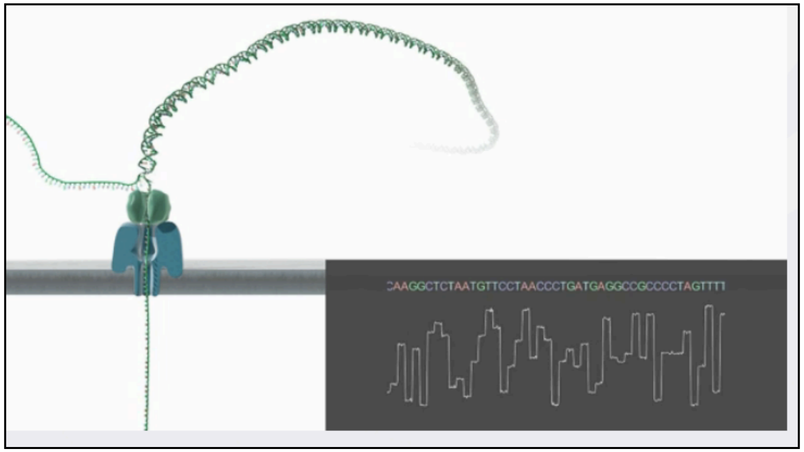
Latest Ion Platforms
Thermo Fisher's Ion S5 & S5 XL

Oxford NANOPORE Technologies®

2014:Sequencing with a protein nanopore

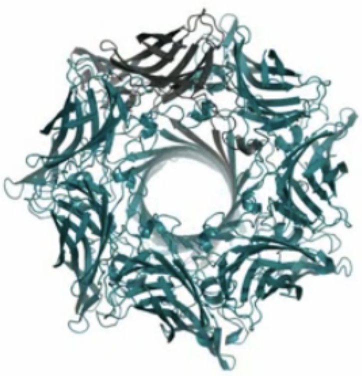


Exonuclease-Seq



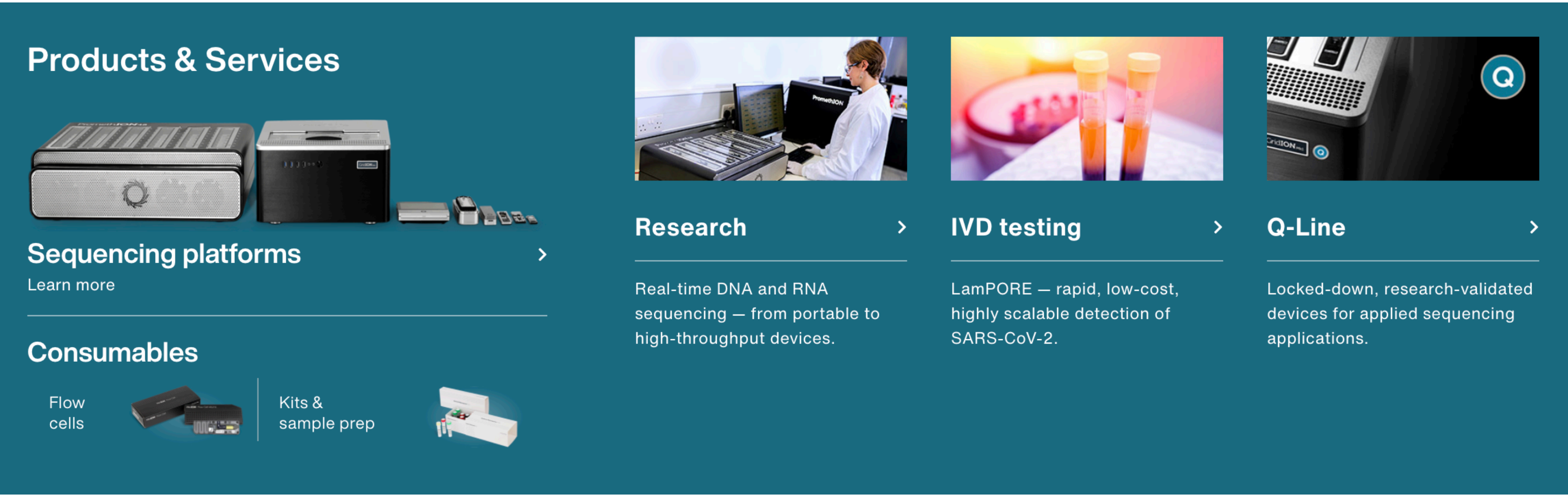
Strand-Seq

MinION



PromethION

2021



Products & Services

Sequencing platforms >

Learn more

Consumables

Flow cells

Kits & sample prep

Research >

Real-time DNA and RNA sequencing — from portable to high-throughput devices.

IVD testing >

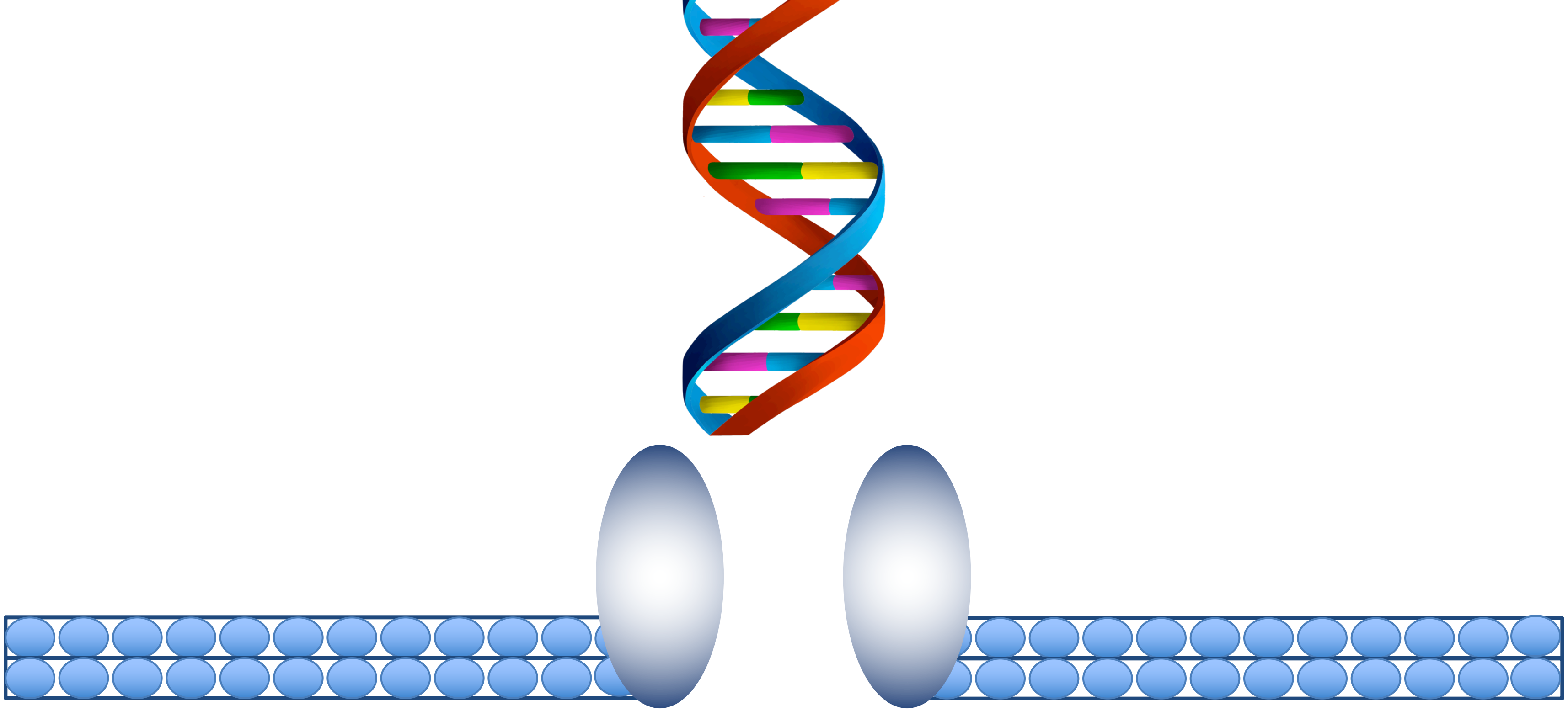
LamPORE — rapid, low-cost, highly scalable detection of SARS-CoV-2.

Q-Line >

Locked-down, research-validated devices for applied sequencing applications.

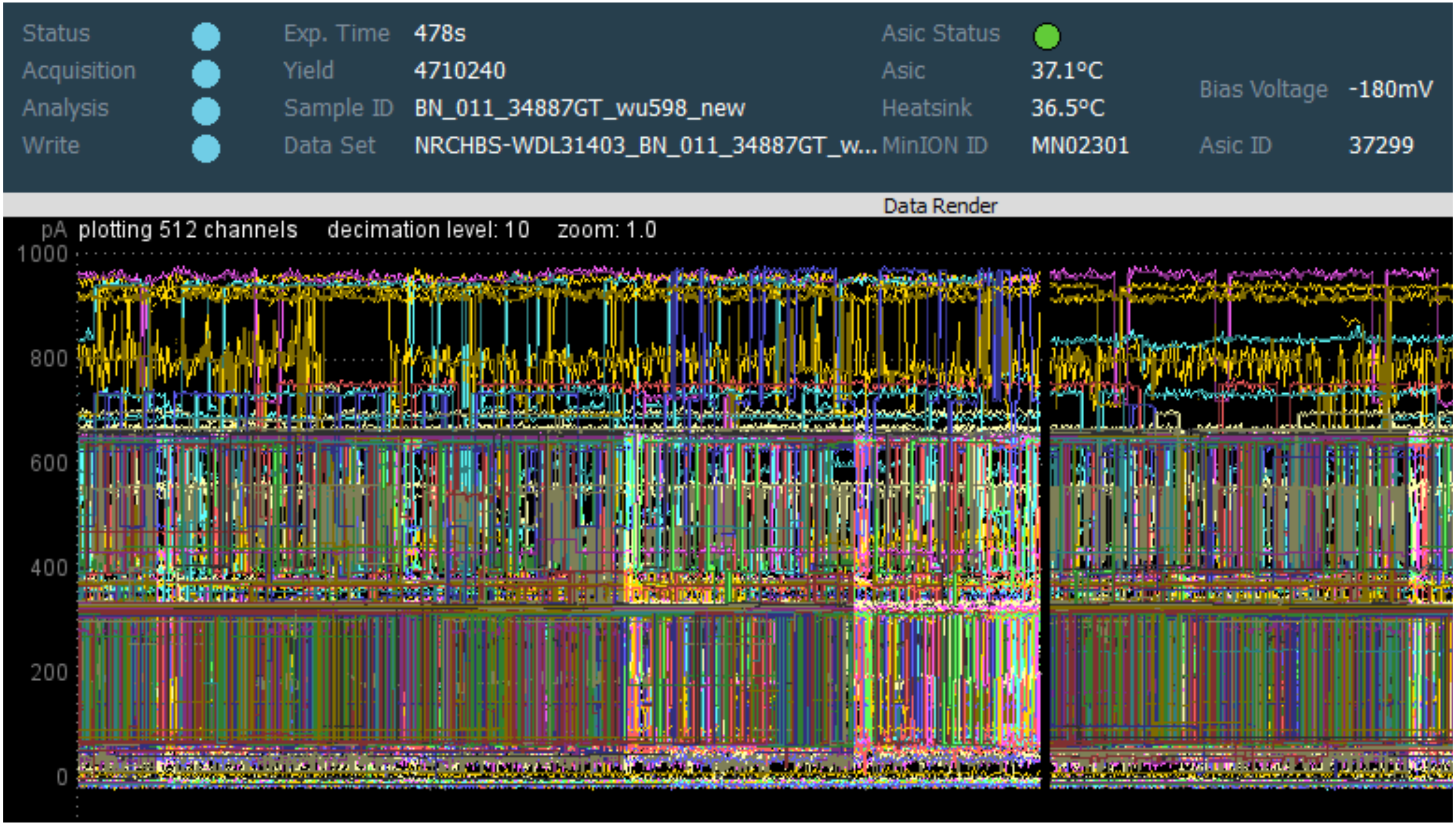
https://nanoporetech.com/

They are small



Meyer *et al.*, *Cell*, 2012 | Saletore *et al.*, *Genome Biology*, 2012 | McIntyre *et al.*, 2015

Base space is now “squiggle space”



Status ●
Acquisition ●
Analysis ●
Write ●

Exp. Time 478s
Yield 4710240
Sample ID BN_011_34887GT_wu598_new
Data Set NRCHBS-WDL31403_BN_011_34887GT_w...

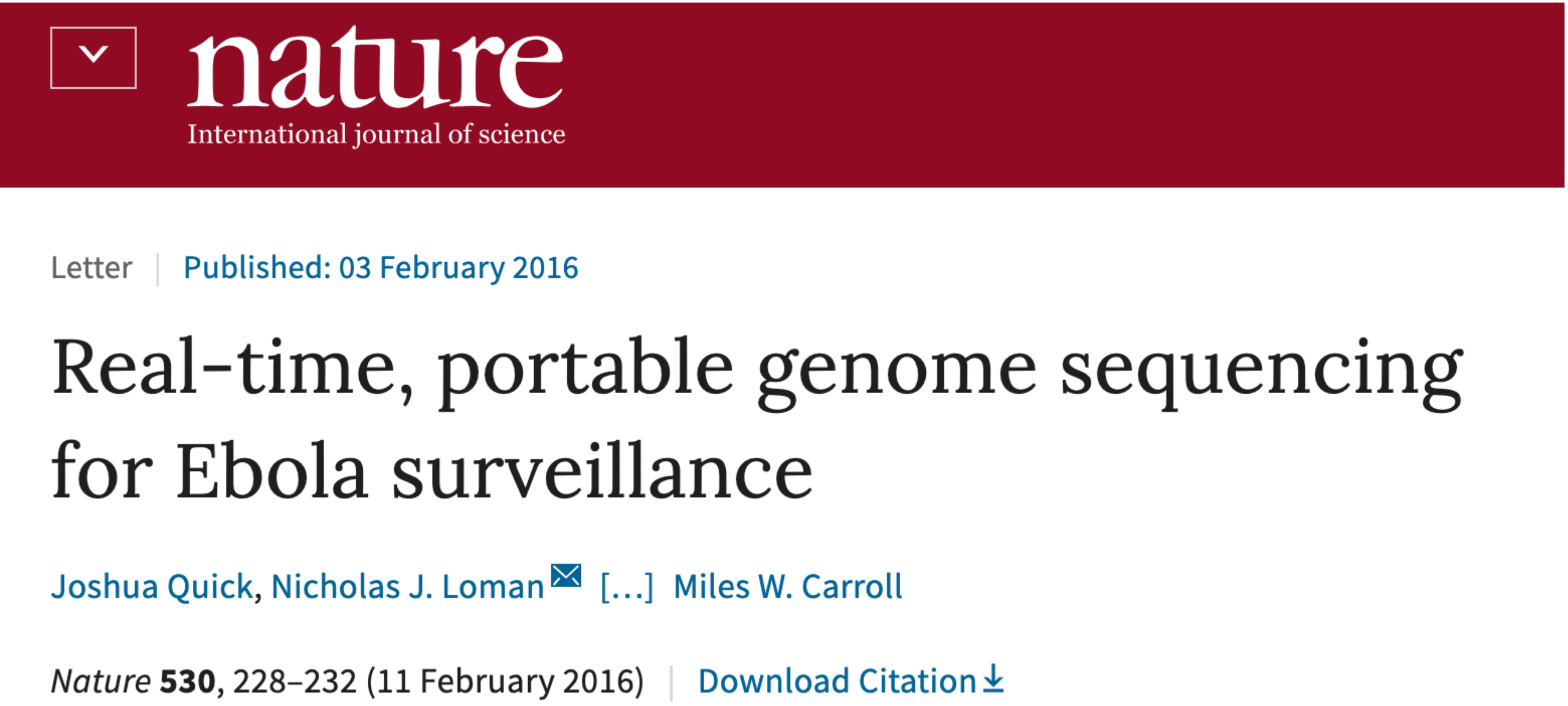
Asic Status ●
Asic 37.1°C
Heatsink 36.5°C
MinION ID MN02301

Bias Voltage -180mV
Asic ID 37299

Data Render

plotting 512 channels decimation level: 10 zoom: 1.0

You can do it anywhere



nature
International journal of science

Letter | Published: 03 February 2016

Real-time, portable genome sequencing for Ebola surveillance

Joshua Quick, Nicholas J. Loman […] Miles W. Carroll

Nature **530**, 228–232 (11 February 2016) | Download Citation

https://www.nature.com/articles/nature16996

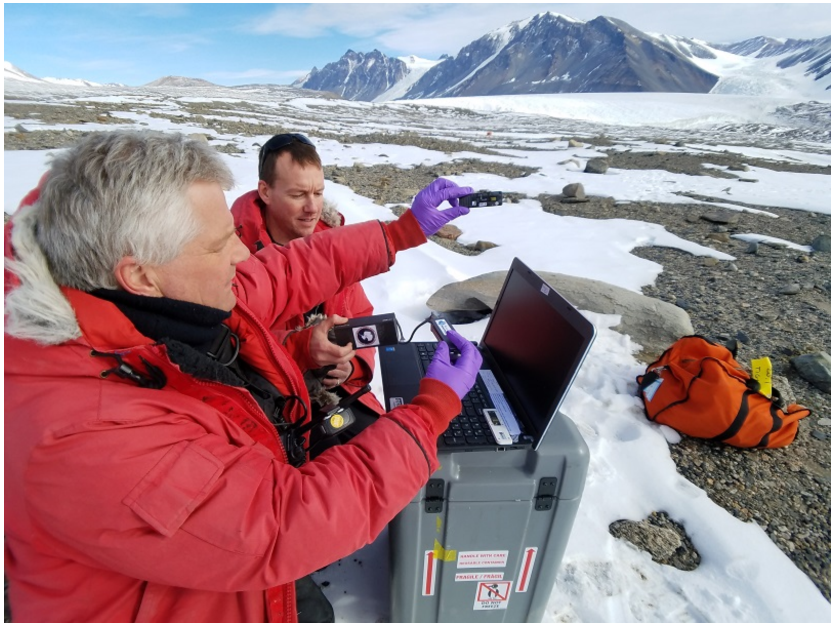
N403IL
NATIONAL SCIENCE
FOUNDATION
THE NORTH FACE

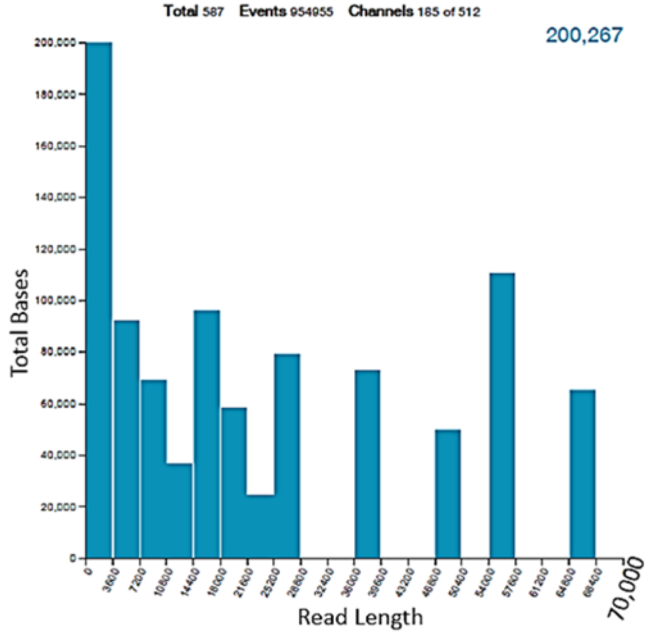
HANDLE WITH CARE
REUSABLE CONTAINER
FRAGILE / FRÁGIL
DO NOT FREEZE

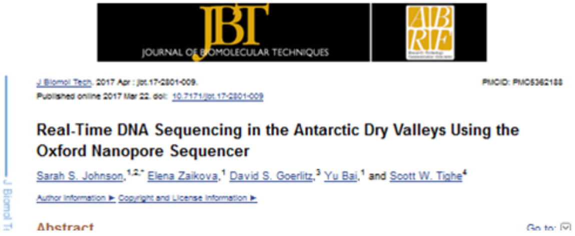
Scott Tighe

Lake Fryxell, Antarctica
Scott Tighe

Sequencing HW DNA in the field with the Oxford Nanopore
Sarah Johnson (PI) expedition G062 team







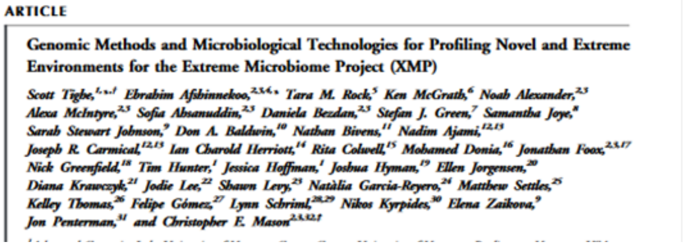
J Biomol Tech. 2017 Apr : jbt.17-2801-009. PMCID: PMC5362188
Published online 2017 Mar 22. doi: 10.7171/jbt.17-2801-009

Real-Time DNA Sequencing in the Antarctic Dry Valleys Using the Oxford Nanopore Sequencer

Sarah S. Johnson,[1,2,*] Elena Zaikova,[1] David S. Goerlitz,[3] Yu Bai,[1] and Scott W. Tighe[4]

Author Information ► Copyright and License Information ►

Abstract Go to: ☐

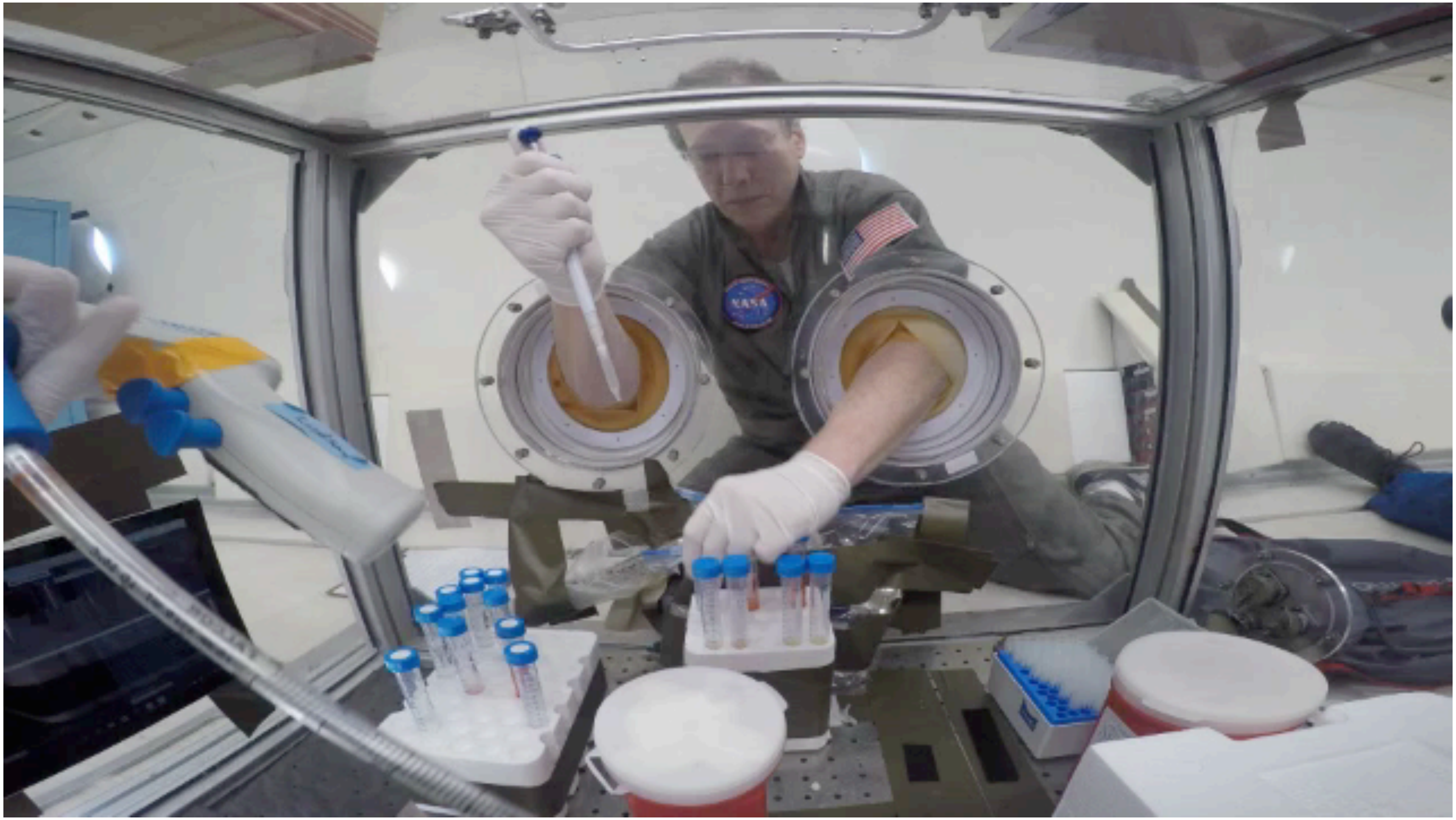


ARTICLE

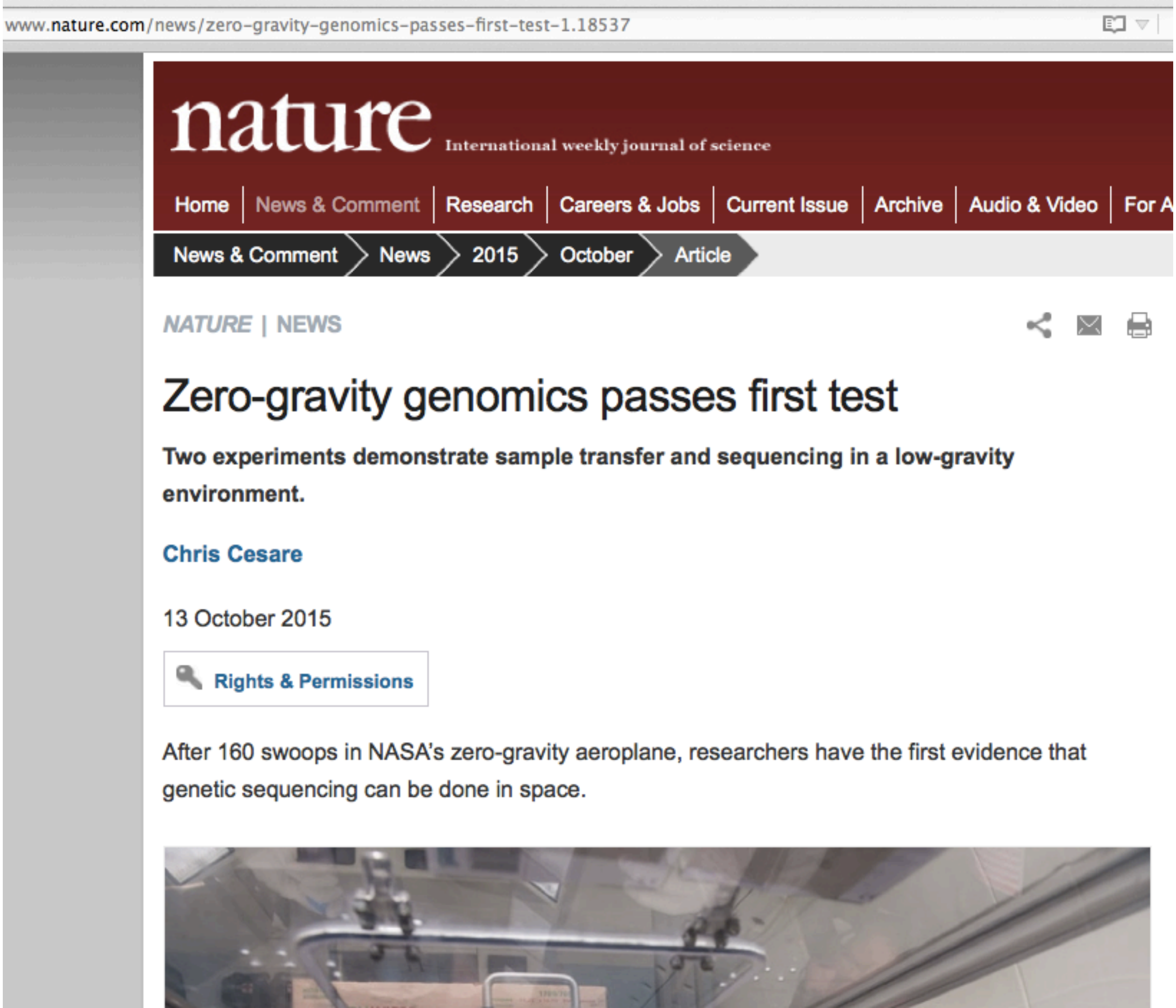
Genomic Methods and Microbiological Technologies for Profiling Novel and Extreme Environments for the Extreme Microbiome Project (XMP)

Scott Tighe,[1,,†] Ebrahim Afshinnekoo,[2,3,4,*] Tara M. Rock,[5] Ken McGrath,[6] Noah Alexander,[2,3] Alexa McIntyre,[2,3] Sofia Ahsanuddin,[2,3] Daniela Bezdan,[2,3] Stefan J. Green,[7] Samantha Joye,[8] Sarah Stewart Johnson,[9] Don A. Baldwin,[10] Nathan Bivens,[11] Nadim Ajami,[12,13] Joseph R. Carmical,[12,13] Ian Charold Herriott,[14] Rita Colwell,[15] Mohamed Donia,[16] Jonathan Foox,[2,3,17] Nick Greenfield,[18] Tim Hunter,[1] Jessica Hoffman,[1] Joshua Hyman,[19] Ellen Jorgensen,[20] Diana Krawczyk,[21] Jodie Lee,[22] Shawn Levy,[23] Natàlia Garcia-Reyero,[24] Matthew Settles,[25] Kelley Thomas,[26] Felipe Gómez,[27] Lynn Schriml,[28,29] Nikos Kyrpides,[30] Elena Zaikova,[9] Jon Penterman,[31] and Christopher E. Mason[2,3,32,‡]*

Zero-G Pipetting: Hardest Lab Job Ever



Dr. Andrew Feinberg



NATURE | NEWS

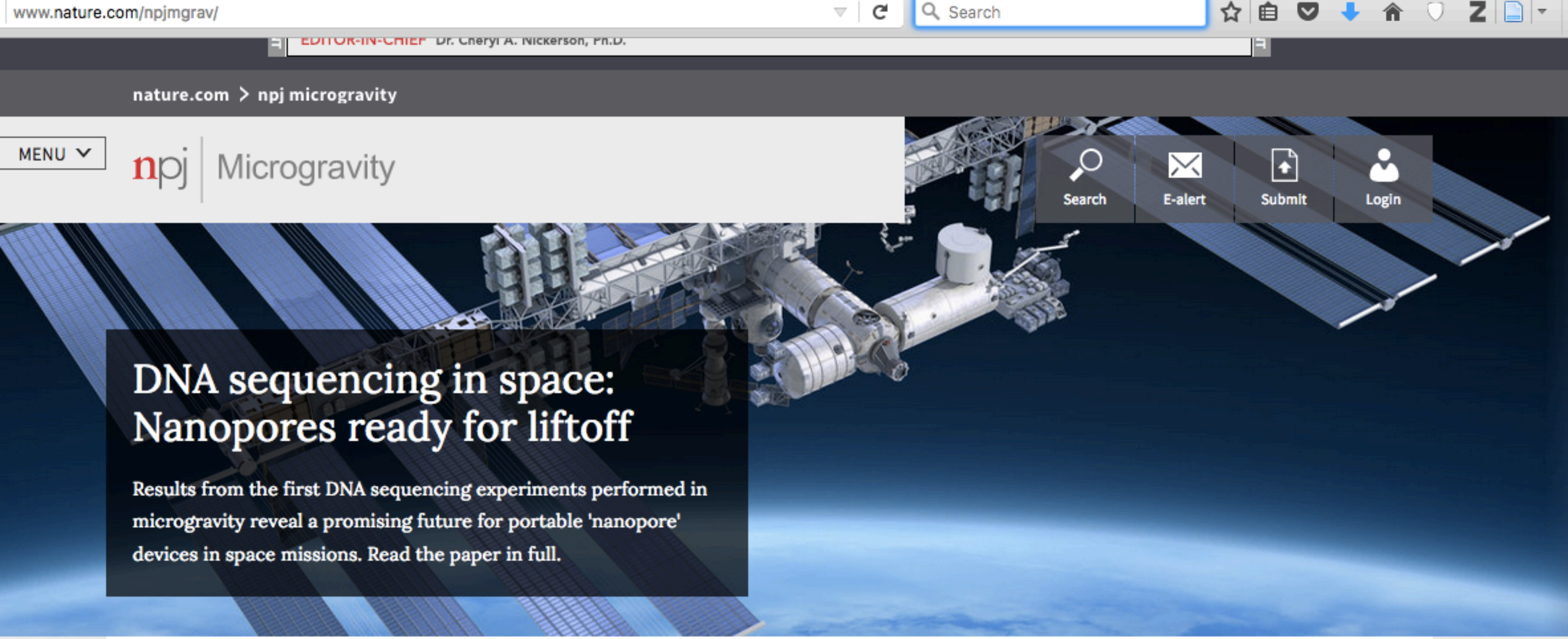
Zero-gravity genomics passes first test

Two experiments demonstrate sample transfer and sequencing in a low-gravity environment.

Chris Cesare

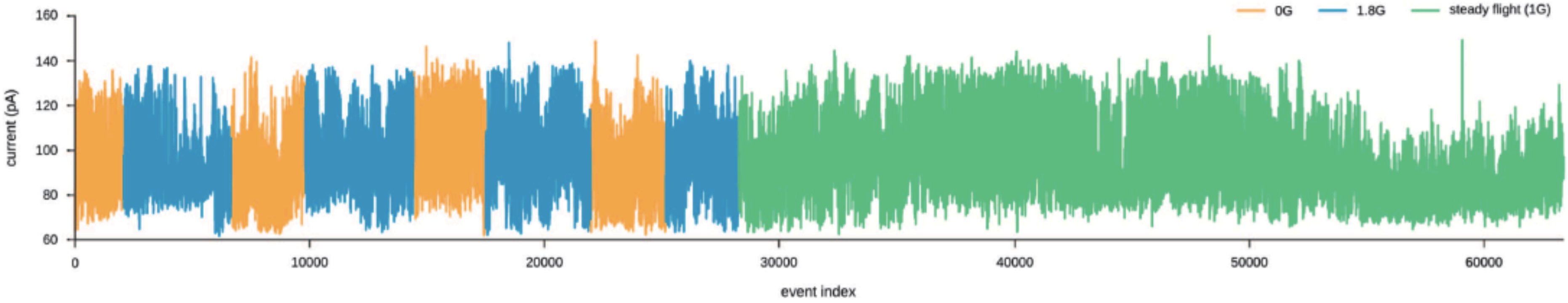
13 October 2015

After 160 swoops in NASA's zero-gravity aeroplane, researchers have the first evidence that genetic sequencing can be done in space.

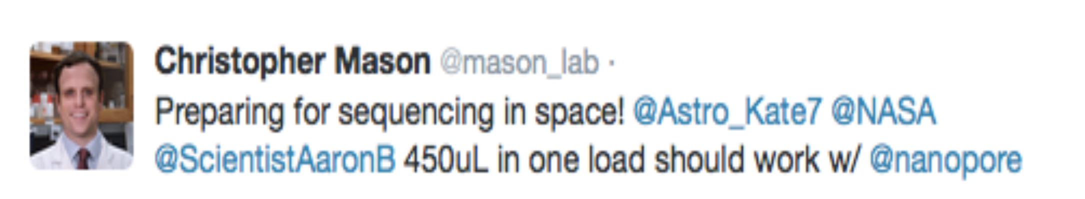

www.nature.com/npjmgrav/
Search
EDITOR-IN-CHIEF Dr. Cheryl A. Nickerson, Ph.D.
nature.com > npj microgravity
MENU
npj Microgravity
Search
E-alert
Submit
Login

DNA sequencing in space: Nanopores ready for liftoff

Results from the first DNA sequencing experiments performed in microgravity reveal a promising future for portable 'nanopore' devices in space missions. Read the paper in full.



McIntyre ABR et al., *Nature Microgravity, 2016.*



Christopher Mason @mason_lab ·

Preparing for sequencing in space! @Astro_Kate7 @NASA @ScientistAaronB 450uL in one load should work w/ @nanopore

SpaceX CRS-7 blows up

National Aeronautics and Space Administration

Office of the Administrator
Washington, DC 20546-0001

Dr. Christopher Mason
Weill Cornell Medical College
1300 York Ave.
New York, NY 10065

Dear Dr. Mason:

As NASA astronaut Scott Kelley tweeted on Sunday, June 28, 2015, "space is hard."

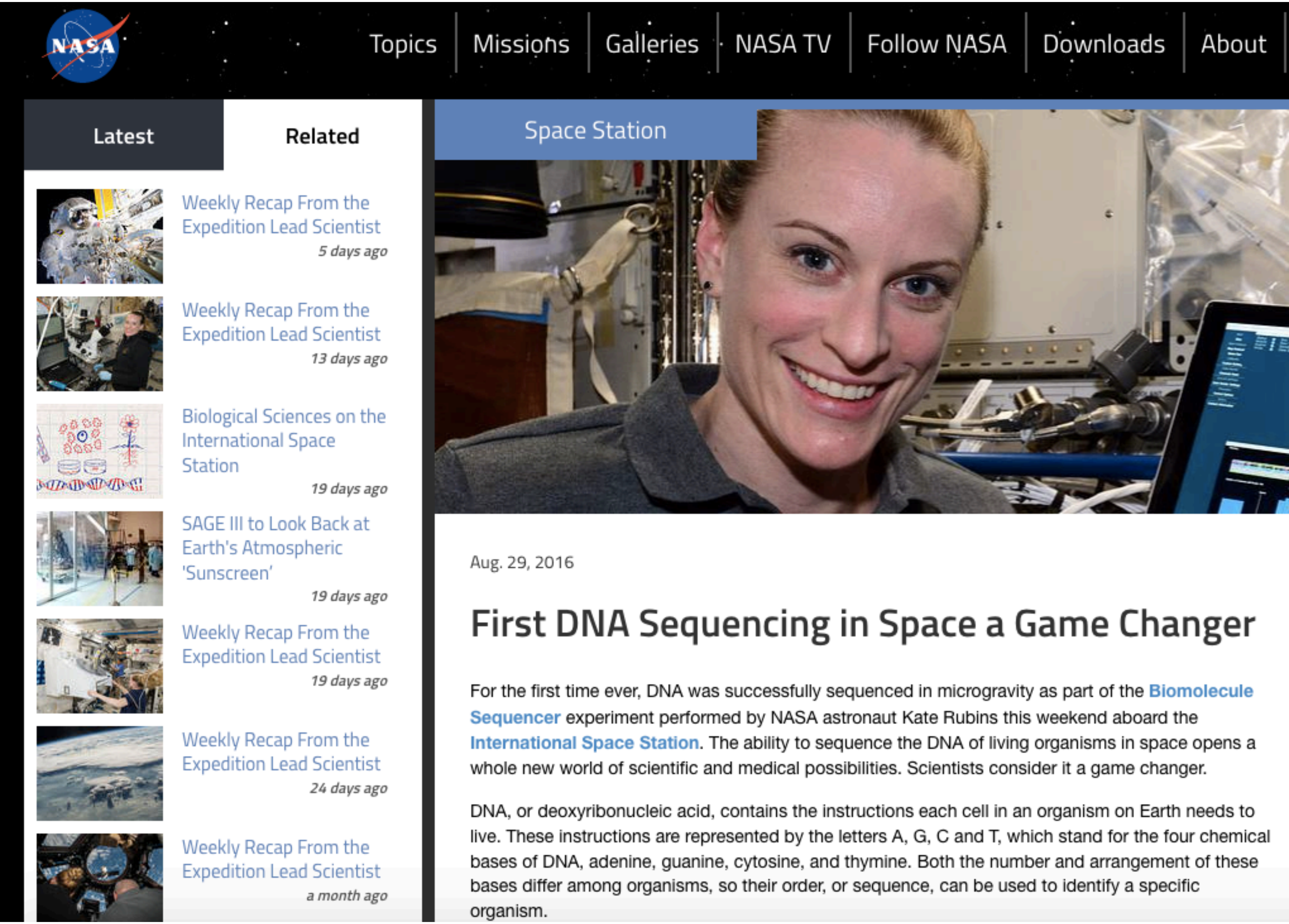
Speaking as a fellow researcher, I can only imagine how devastated you must be feeling right now with the loss of SpaceX's CRS-7. I am saddened and disappointed too. I am sure that the tremendous honor of being selected to have your experiment flown on the International Space Station is of little solace after the loss of months, and perhaps even years, of hard work.

I am writing to encourage you – and in fact, to urge you – to continue your inquiry. The story of space exploration is the story of people just like you who meet adversity, head on, with determination and scientific and technological advancement. If you think about it, virtually every major innovation and technological breakthrough in human history has been the product of many different stops and starts; learning and being better because of failures and setbacks and, ultimately, enhanced knowledge and moving forward.

ВНИМАНИ
ЗОНА ПЕРЕМЕ
НЕ ДОПУС

SpaceX CRS-9: perfect launch and booster return
July 18, 2016

TO NOD 2



Weekly Recap From the Expedition Lead Scientist
5 days ago

Weekly Recap From the Expedition Lead Scientist
13 days ago

Biological Sciences on the International Space Station
19 days ago

SAGE III to Look Back at Earth's Atmospheric 'Sunscreen'
19 days ago

Weekly Recap From the Expedition Lead Scientist
19 days ago

Weekly Recap From the Expedition Lead Scientist
24 days ago

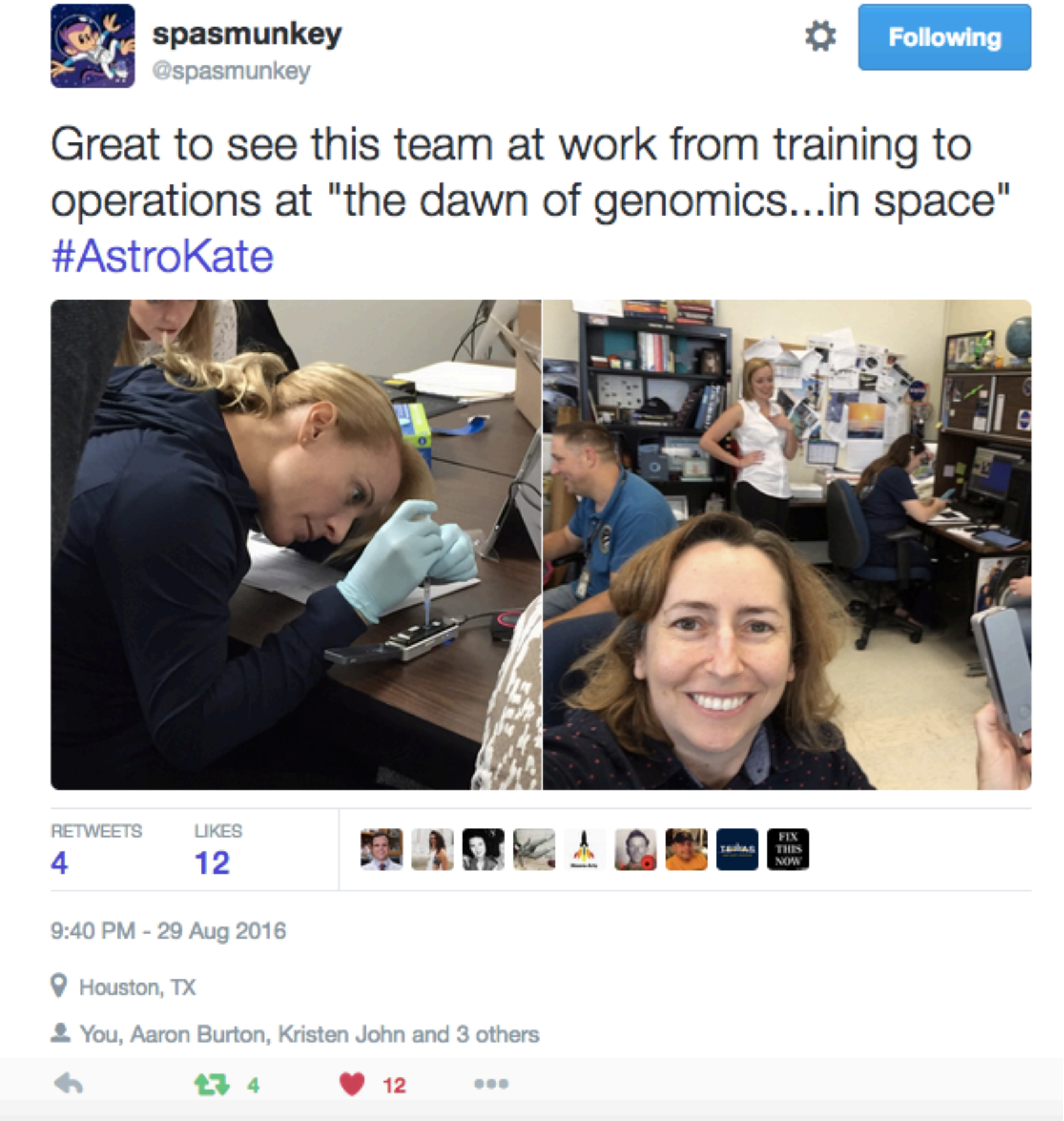
Weekly Recap From the Expedition Lead Scientist
a month ago

Aug. 29, 2016

First DNA Sequencing in Space a Game Changer

For the first time ever, DNA was successfully sequenced in microgravity as part of the **Biomolecule Sequencer** experiment performed by NASA astronaut Kate Rubins this weekend aboard the **International Space Station**. The ability to sequence the DNA of living organisms in space opens a whole new world of scientific and medical possibilities. Scientists consider it a game changer.

DNA, or deoxyribonucleic acid, contains the instructions each cell in an organism on Earth needs to live. These instructions are represented by the letters A, G, C and T, which stand for the four chemical bases of DNA, adenine, guanine, cytosine, and thymine. Both the number and arrangement of these bases differ among organisms, so their order, or sequence, can be used to identify a specific organism.



spasmunkey
@spasmunkey

Great to see this team at work from training to operations at "the dawn of genomics...in space" #AstroKate

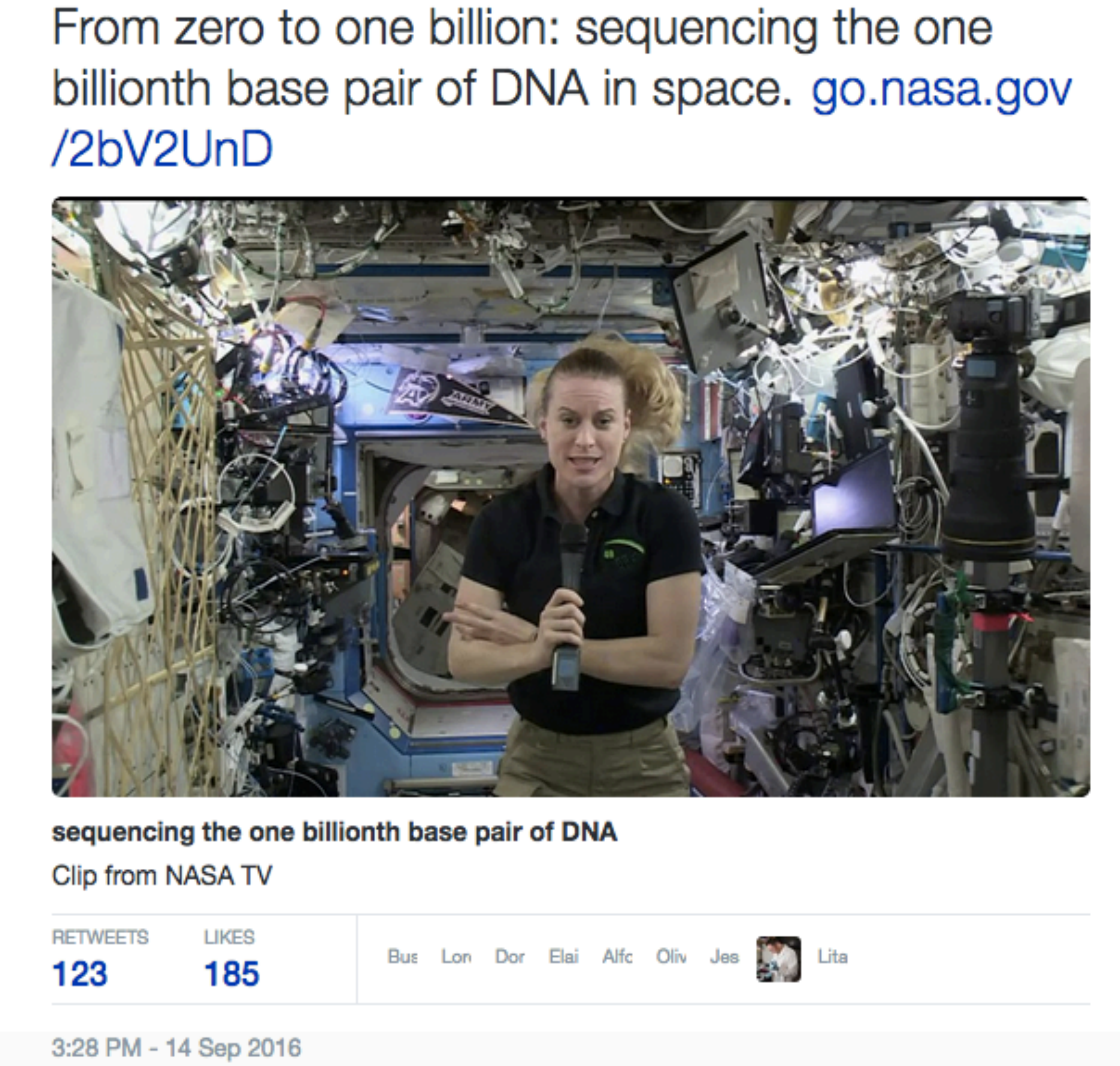
RETWEETS 4

LIKES 12

9:40 PM - 29 Aug 2016

Houston, TX

You, Aaron Burton, Kristen John and 3 others



From zero to one billion: sequencing the one billionth base pair of DNA in space. go.nasa.gov/2bV2UnD

sequencing the one billionth base pair of DNA

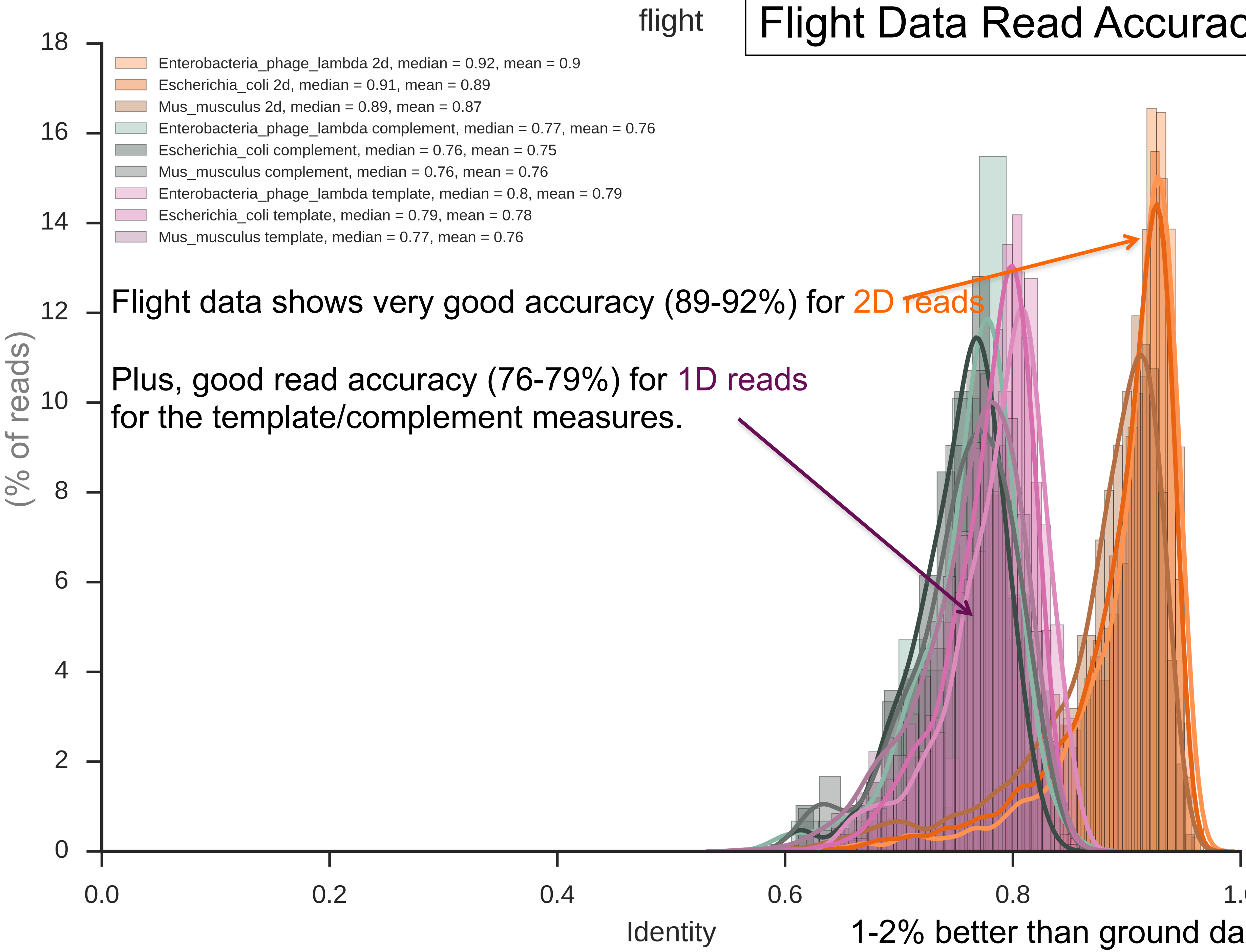
Clip from NASA TV

RETWEETS 123 LIKES 185

Bus Lon Dor Elai Alfc Oliv Jes Lita

3:28 PM - 14 Sep 2016

Flight Data Read Accuracy

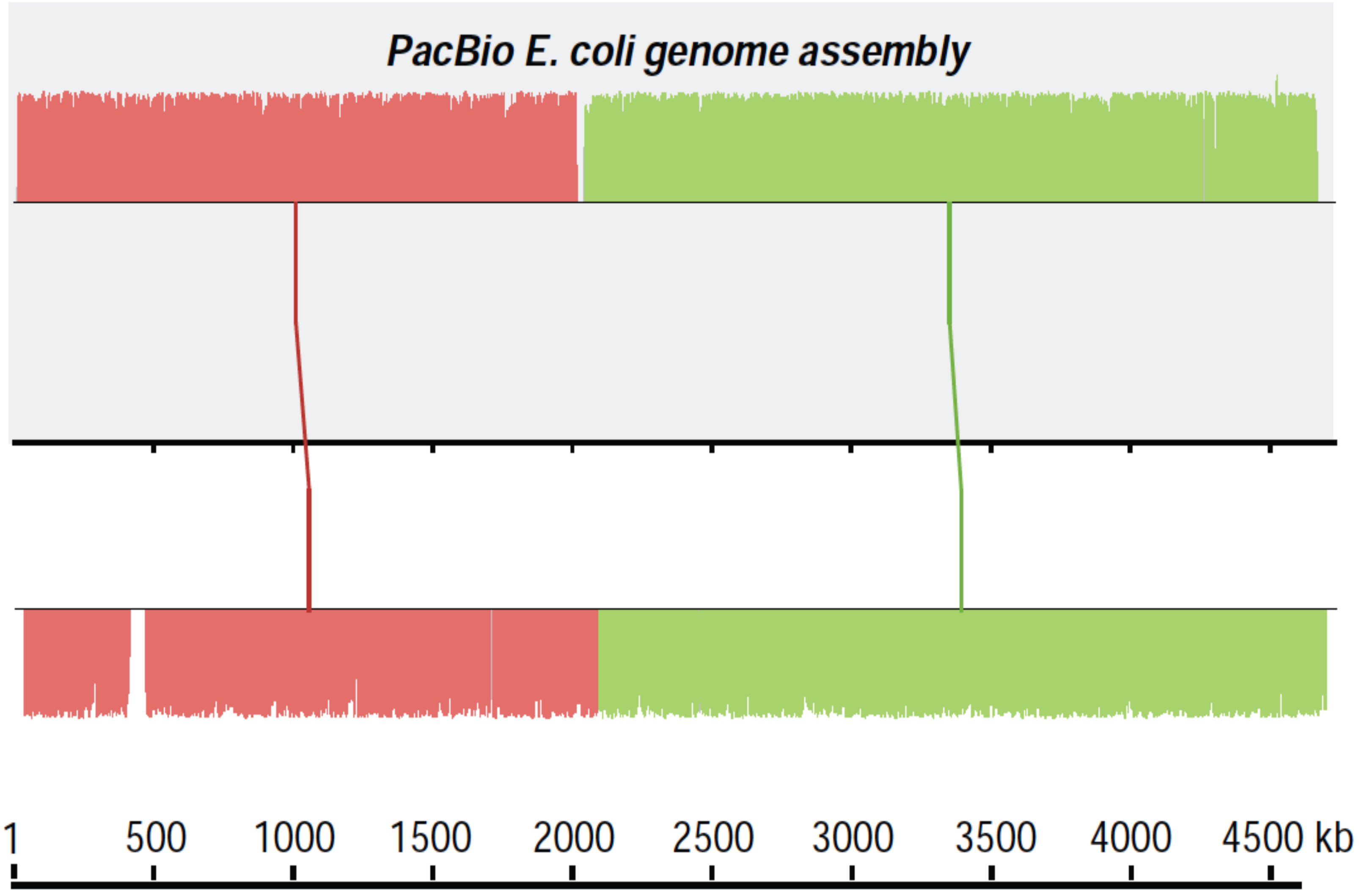


Flight data shows very good accuracy (89-92%) for 2D reads

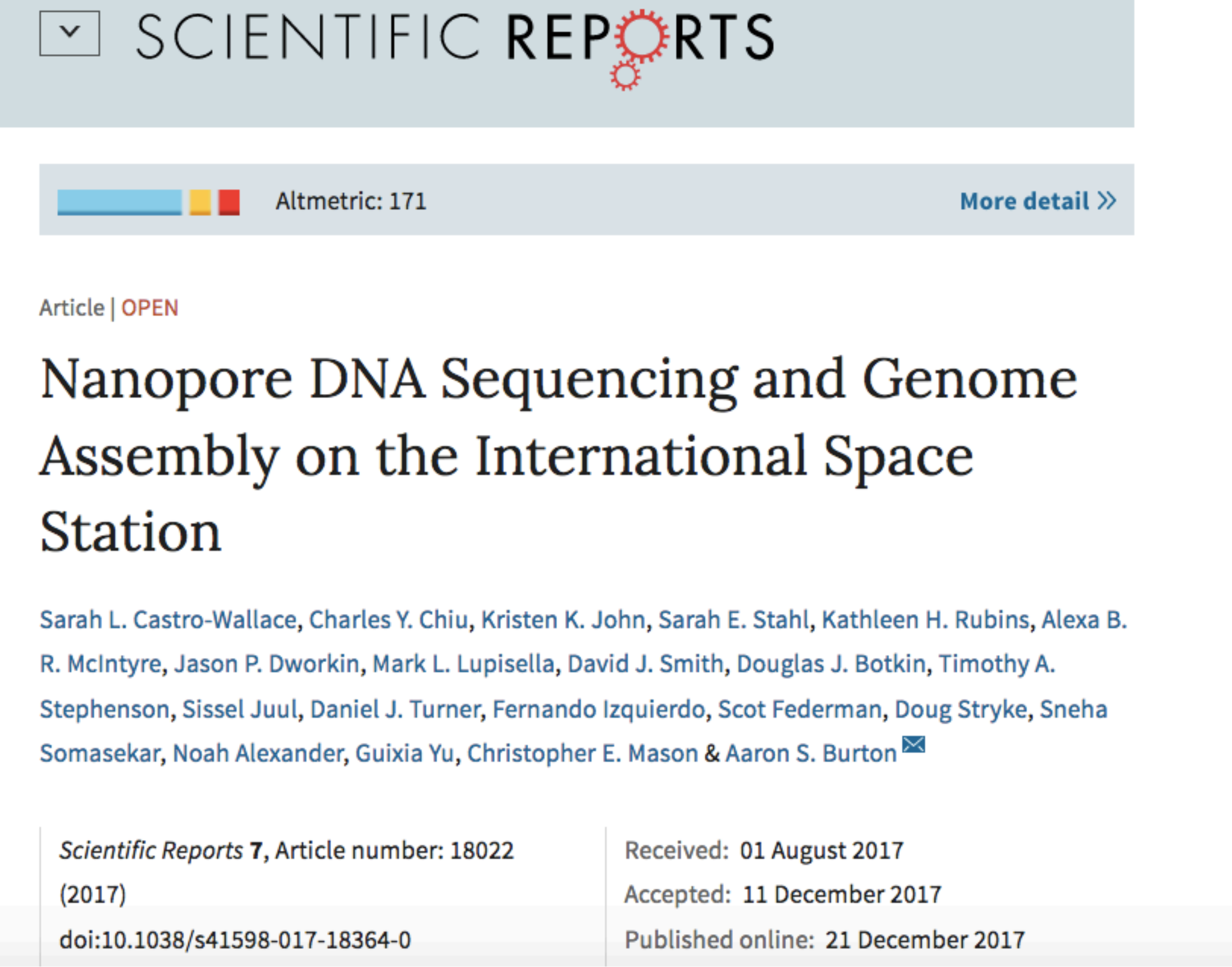
Plus, good read accuracy (76-79%) for 1D reads
for the template/complement measures.

1-2% better than ground data

Almost perfect when compared to PacBio



The first genome sequence, assembly, and AMR detection off Earth



SCIENTIFIC REPORTS

Altmetric: 171

More detail »

Article | OPEN

Nanopore DNA Sequencing and Genome Assembly on the International Space Station

Sarah L. Castro-Wallace, Charles Y. Chiu, Kristen K. John, Sarah E. Stahl, Kathleen H. Rubins, Alexa B. R. McIntyre, Jason P. Dworkin, Mark L. Lupisella, David J. Smith, Douglas J. Botkin, Timothy A. Stephenson, Sissel Juul, Daniel J. Turner, Fernando Izquierdo, Scot Federman, Doug Stryke, Sneha Somasekar, Noah Alexander, Guixia Yu, Christopher E. Mason & Aaron S. Burton ✉

Scientific Reports **7**, Article number: 18022 (2017)

doi:10.1038/s41598-017-18364-0

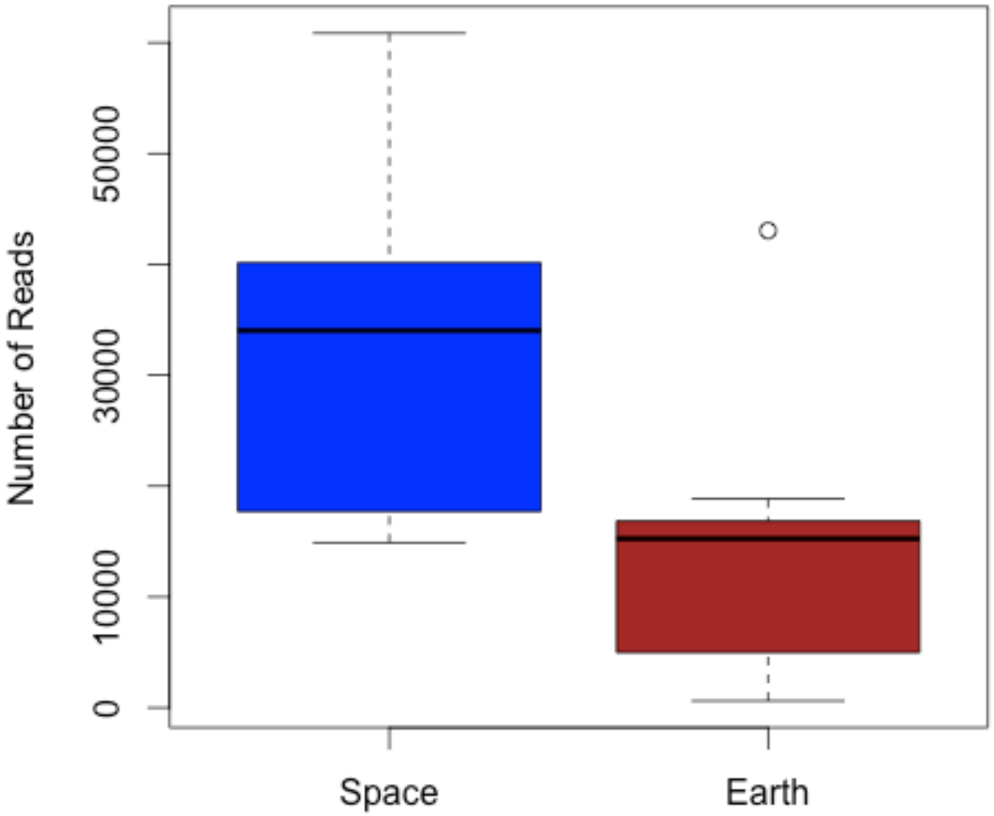
Received: 01 August 2017

Accepted: 11 December 2017

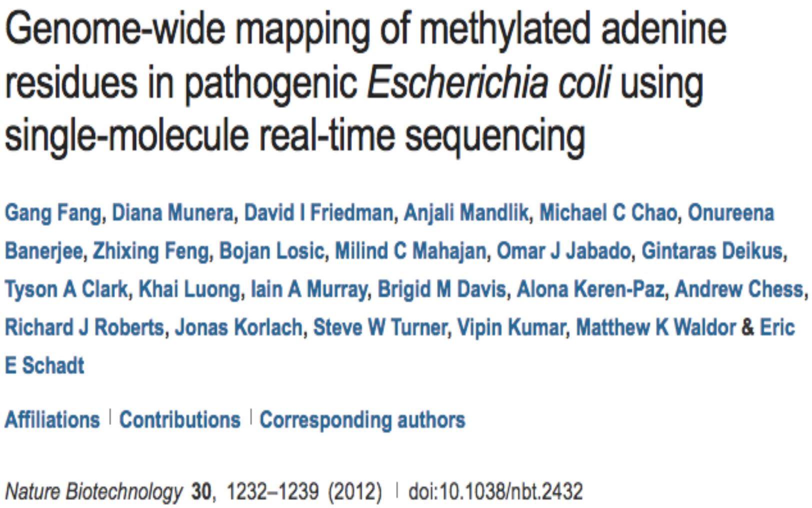
Published online: 21 December 2017

https://www.nature.com/articles/s41598-017-18364-0

As good, or better (8/9) data in space



Bacteria are splattered with epigenetic marks

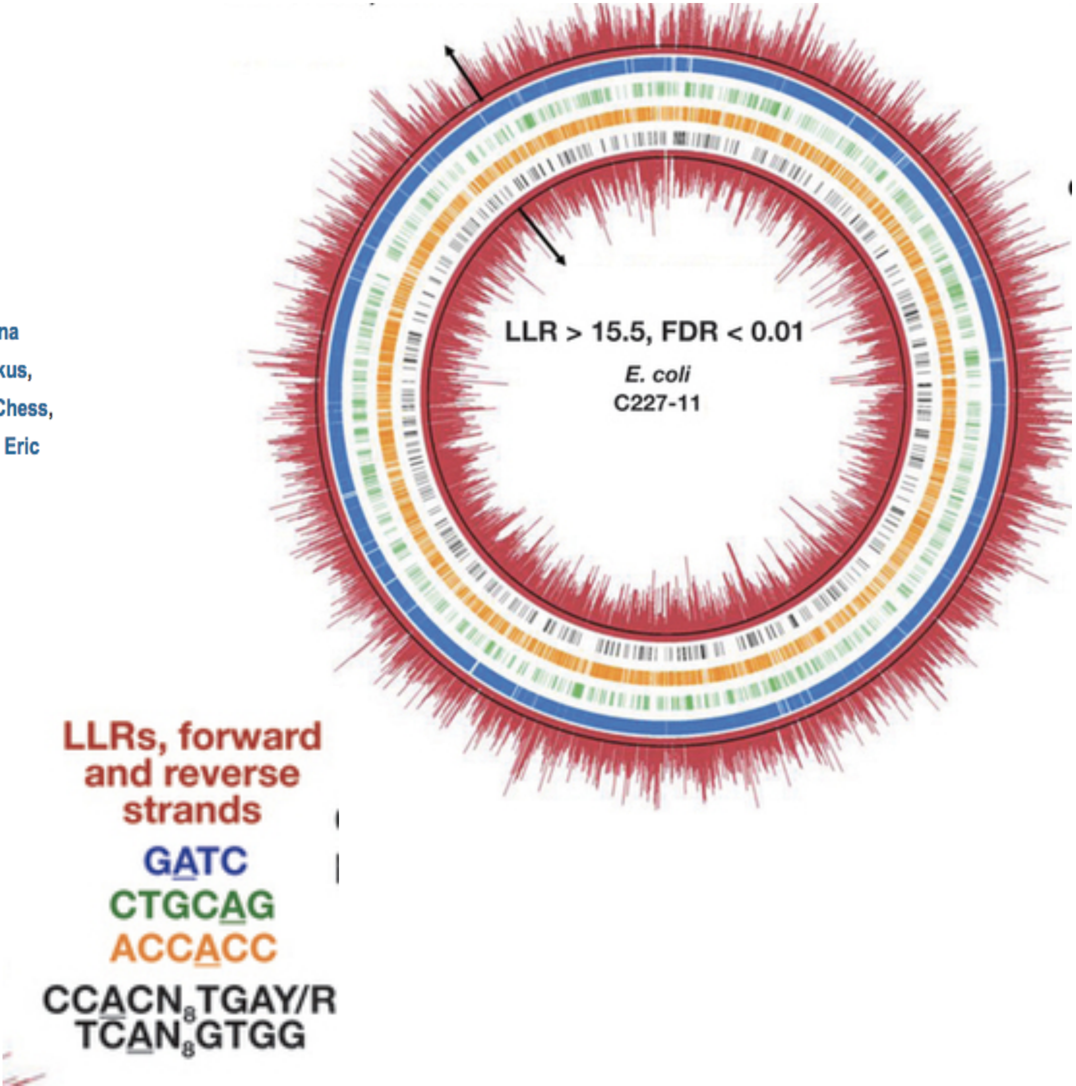


Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing

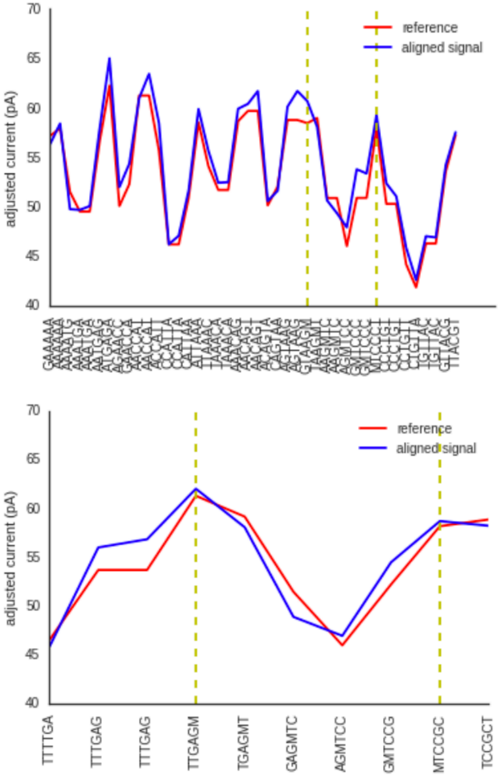
Gang Fang, Diana Munera, David I Friedman, Anjali Mandlik, Michael C Chao, Onureena Banerjee, Zhixing Feng, Bojan Losic, Milind C Mahajan, Omar J Jabado, Gintaras Deikus, Tyson A Clark, Khai Luong, Iain A Murray, Brigid M Davis, Alona Keren-Paz, Andrew Chess, Richard J Roberts, Jonas Korlach, Steve W Turner, Vipin Kumar, Matthew K Waldor & Eric E Schadt

Affiliations | Contributions | Corresponding authors

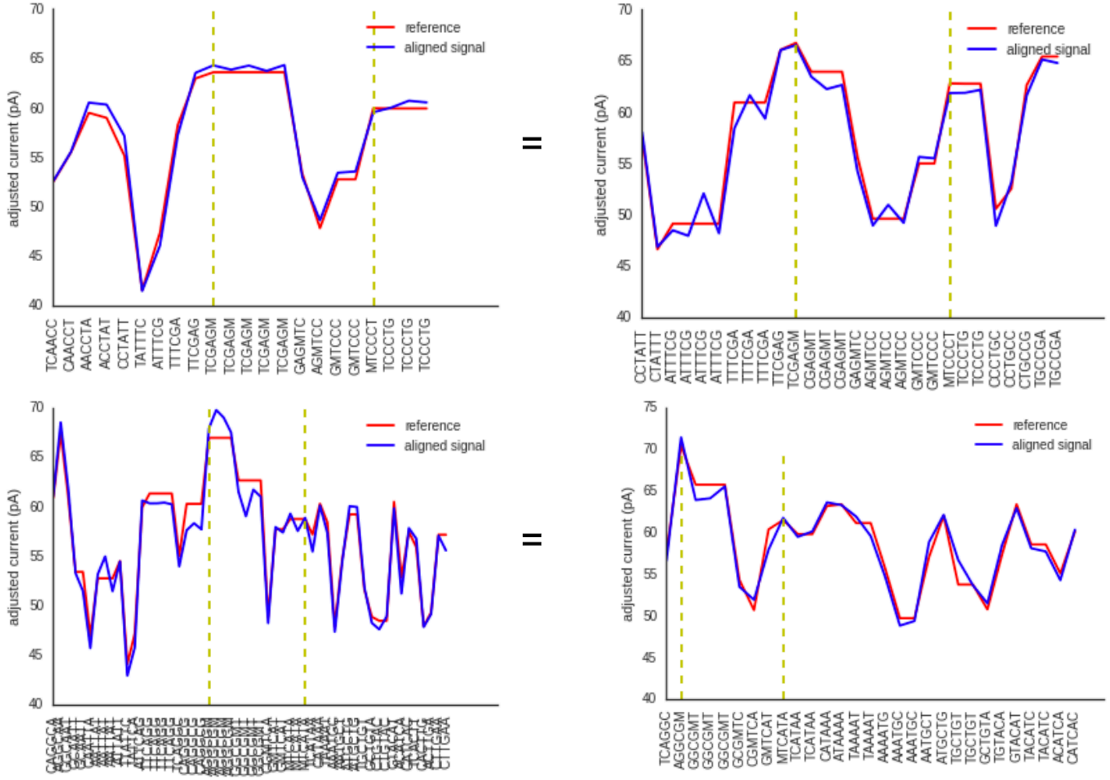
Nature Biotechnology **30**, 1232–1239 (2012) | doi:10.1038/nbt.2432



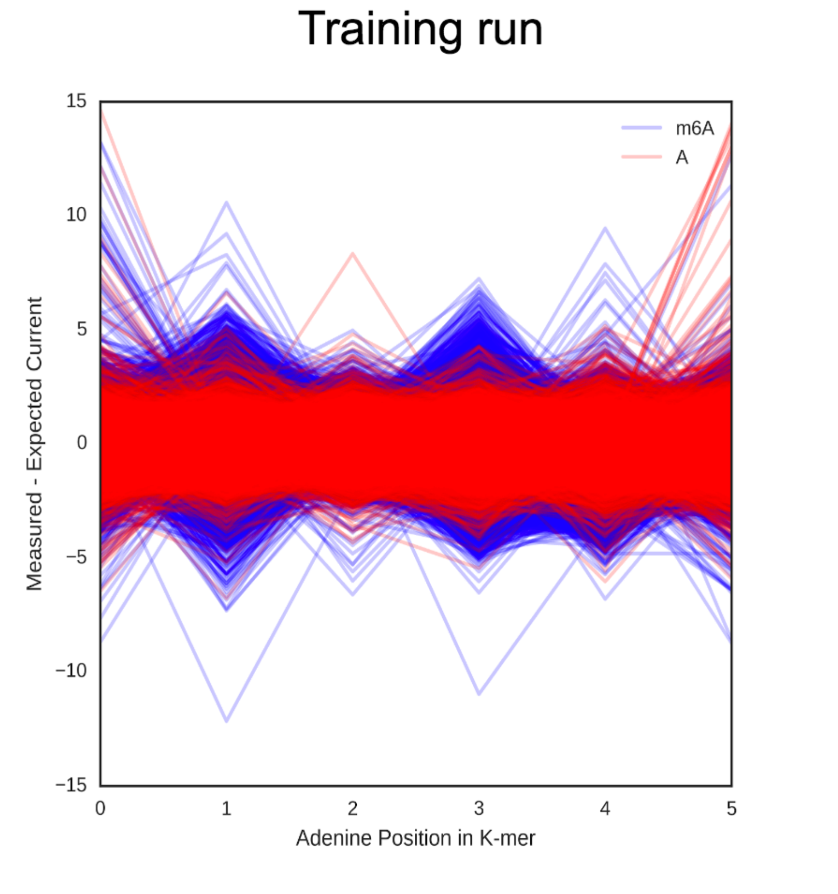
Calling current (pA) differences, similar to PacBio

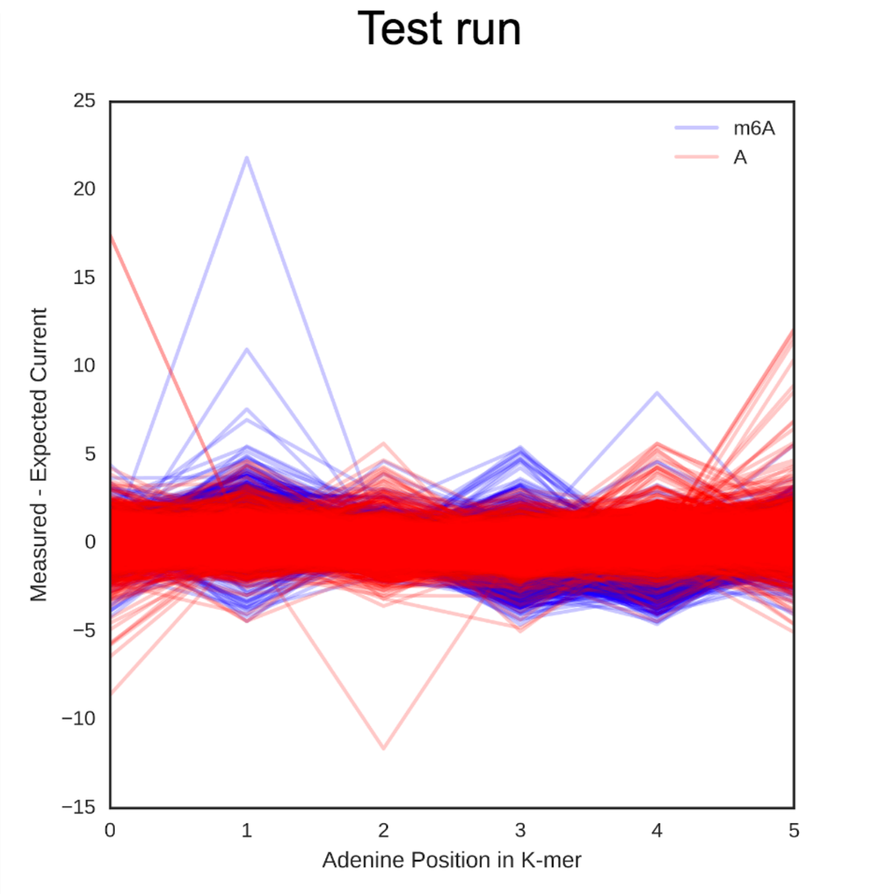


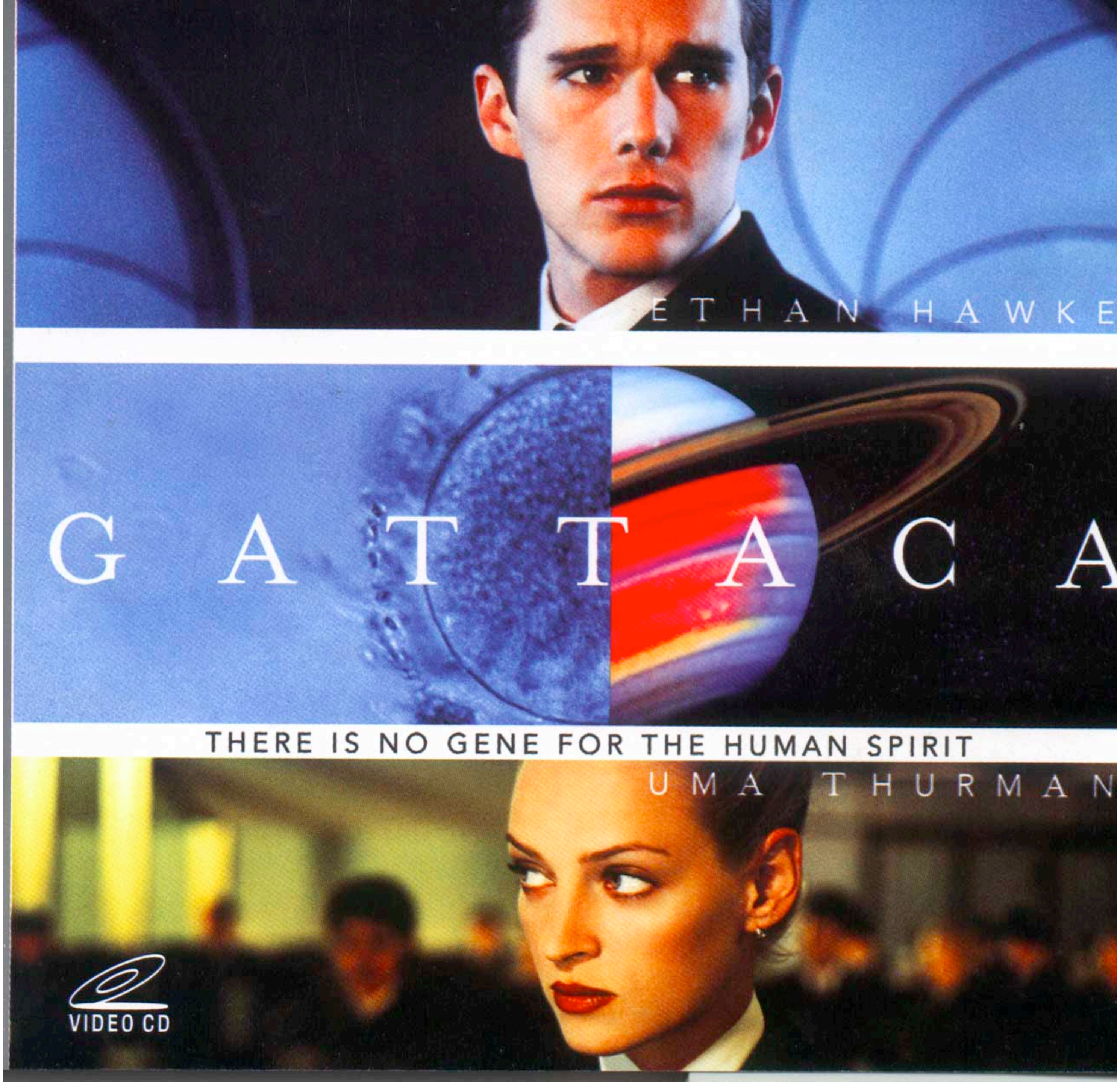
Reads aligned to same positions



Certain positions of the pore and more informative then others






ETHAN HAWKE
GATTACA
THERE IS NO GENE FOR THE HUMAN SPIRIT
UMA THURMAN
VIDEO CD

Is a 2.6 minute genome possible?
No today, but if the physics holds up…

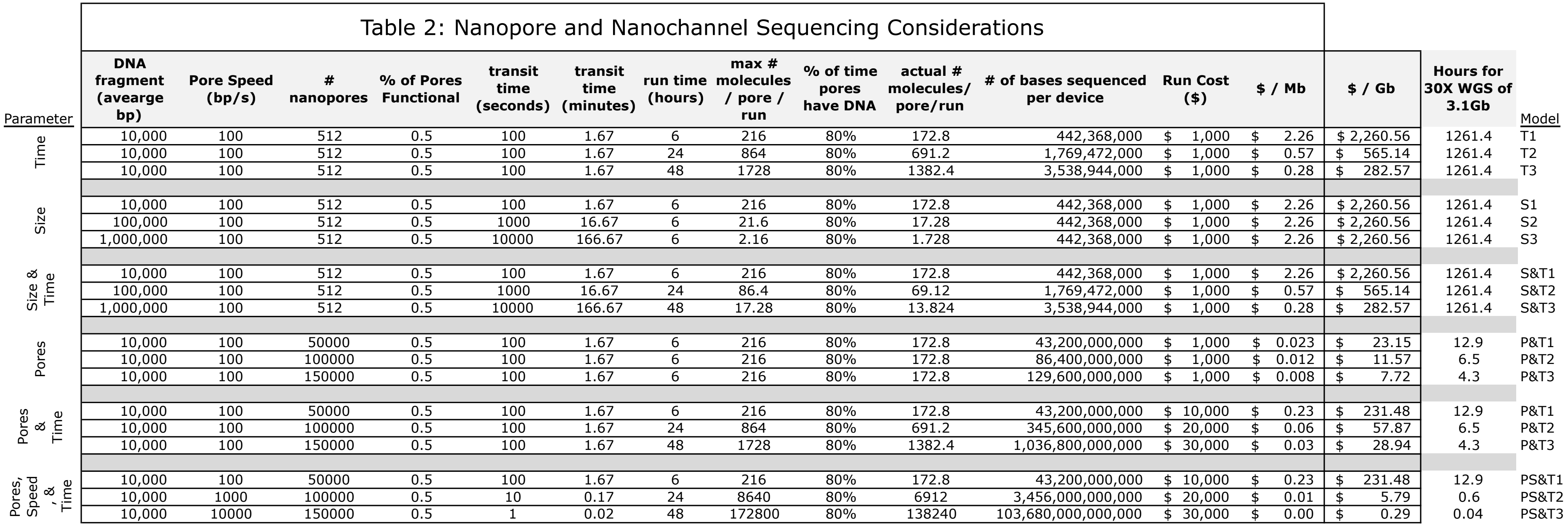
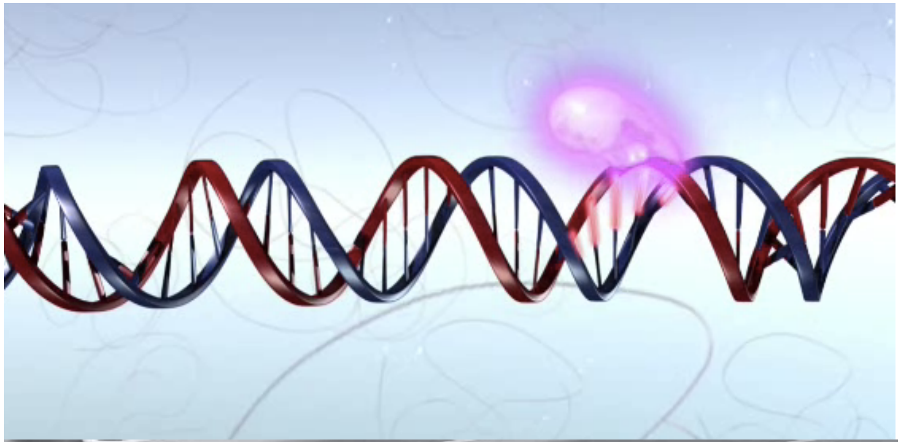


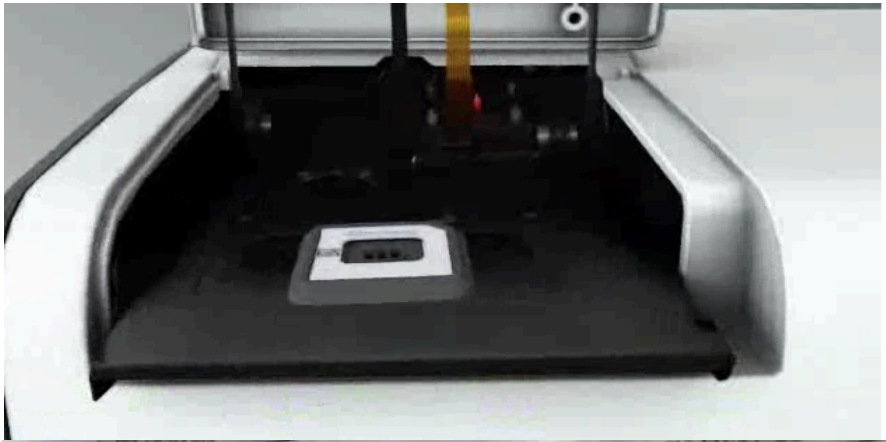
Table 2: Nanopore and Nanochannel Sequencing Considerations

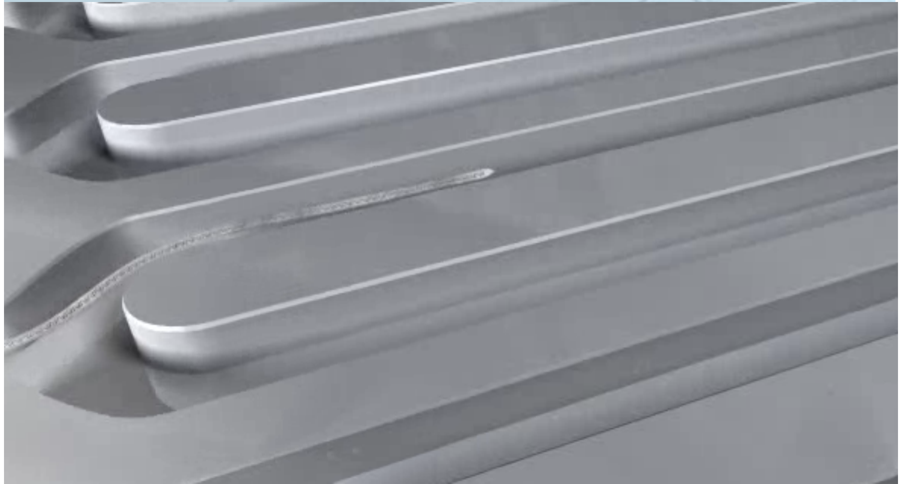
Parameter	DNA fragment (avearge bp)	Pore Speed (bp/s)	# nanopores	% of Pores Functional	transit time (seconds)	transit time (minutes)	run time (hours)	max # molecules / pore / run	% of time pores have DNA	actual # molecules/ pore/run	# of bases sequenced per device	Run Cost ($)	$ / Mb	$ / Gb	Hours for 30X WGS of 3.1Gb	Model
Time	10,000	100	512	0.5	100	1.67	6	216	80%	172.8	442,368,000	$ 1,000	$ 2.26	$ 2,260.56	1261.4	T1
	10,000	100	512	0.5	100	1.67	24	864	80%	691.2	1,769,472,000	$ 1,000	$ 0.57	$ 565.14	1261.4	T2
	10,000	100	512	0.5	100	1.67	48	1728	80%	1382.4	3,538,944,000	$ 1,000	$ 0.28	$ 282.57	1261.4	T3
Size	10,000	100	512	0.5	100	1.67	6	216	80%	172.8	442,368,000	$ 1,000	$ 2.26	$ 2,260.56	1261.4	S1
	100,000	100	512	0.5	1000	16.67	6	21.6	80%	17.28	442,368,000	$ 1,000	$ 2.26	$ 2,260.56	1261.4	S2
	1,000,000	100	512	0.5	10000	166.67	6	2.16	80%	1.728	442,368,000	$ 1,000	$ 2.26	$ 2,260.56	1261.4	S3
Size & Time	10,000	100	512	0.5	100	1.67	6	216	80%	172.8	442,368,000	$ 1,000	$ 2.26	$ 2,260.56	1261.4	S&T1
	100,000	100	512	0.5	1000	16.67	24	86.4	80%	69.12	1,769,472,000	$ 1,000	$ 0.57	$ 565.14	1261.4	S&T2
	1,000,000	100	512	0.5	10000	166.67	48	17.28	80%	13.824	3,538,944,000	$ 1,000	$ 0.28	$ 282.57	1261.4	S&T3
Pores	10,000	100	50000	0.5	100	1.67	6	216	80%	172.8	43,200,000,000	$ 1,000	$ 0.023	$ 23.15	12.9	P&T1
	10,000	100	100000	0.5	100	1.67	6	216	80%	172.8	86,400,000,000	$ 1,000	$ 0.012	$ 11.57	6.5	P&T2
	10,000	100	150000	0.5	100	1.67	6	216	80%	172.8	129,600,000,000	$ 1,000	$ 0.008	$ 7.72	4.3	P&T3
Pores & Time	10,000	100	50000	0.5	100	1.67	6	216	80%	172.8	43,200,000,000	$ 10,000	$ 0.23	$ 231.48	12.9	P&T1
	10,000	100	100000	0.5	100	1.67	24	864	80%	691.2	345,600,000,000	$ 20,000	$ 0.06	$ 57.87	6.5	P&T2
	10,000	100	150000	0.5	100	1.67	48	1728	80%	1382.4	1,036,800,000,000	$ 30,000	$ 0.03	$ 28.94	4.3	P&T3
Pores, Speed , & Time	10,000	100	50000	0.5	100	1.67	6	216	80%	172.8	43,200,000,000	$ 10,000	$ 0.23	$ 231.48	12.9	PS&T1
	10,000	1000	100000	0.5	10	0.17	24	8640	80%	6912	3,456,000,000,000	$ 20,000	$ 0.01	$ 5.79	0.6	PS&T2
	10,000	10000	150000	0.5	1	0.02	48	172800	80%	138240	103,680,000,000,000	$ 30,000	$ 0.00	$ 0.29	0.04	PS&T3

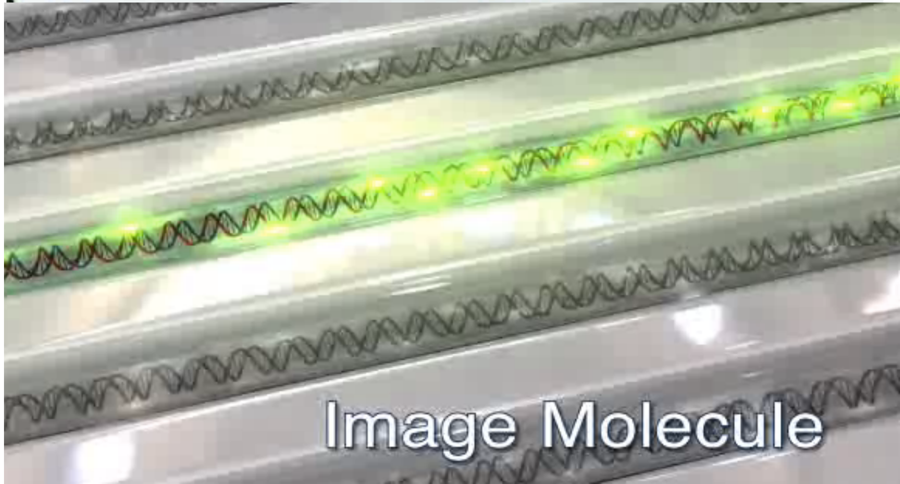
Bionanogenomics - Irys System

BIONANO
GENOMICS





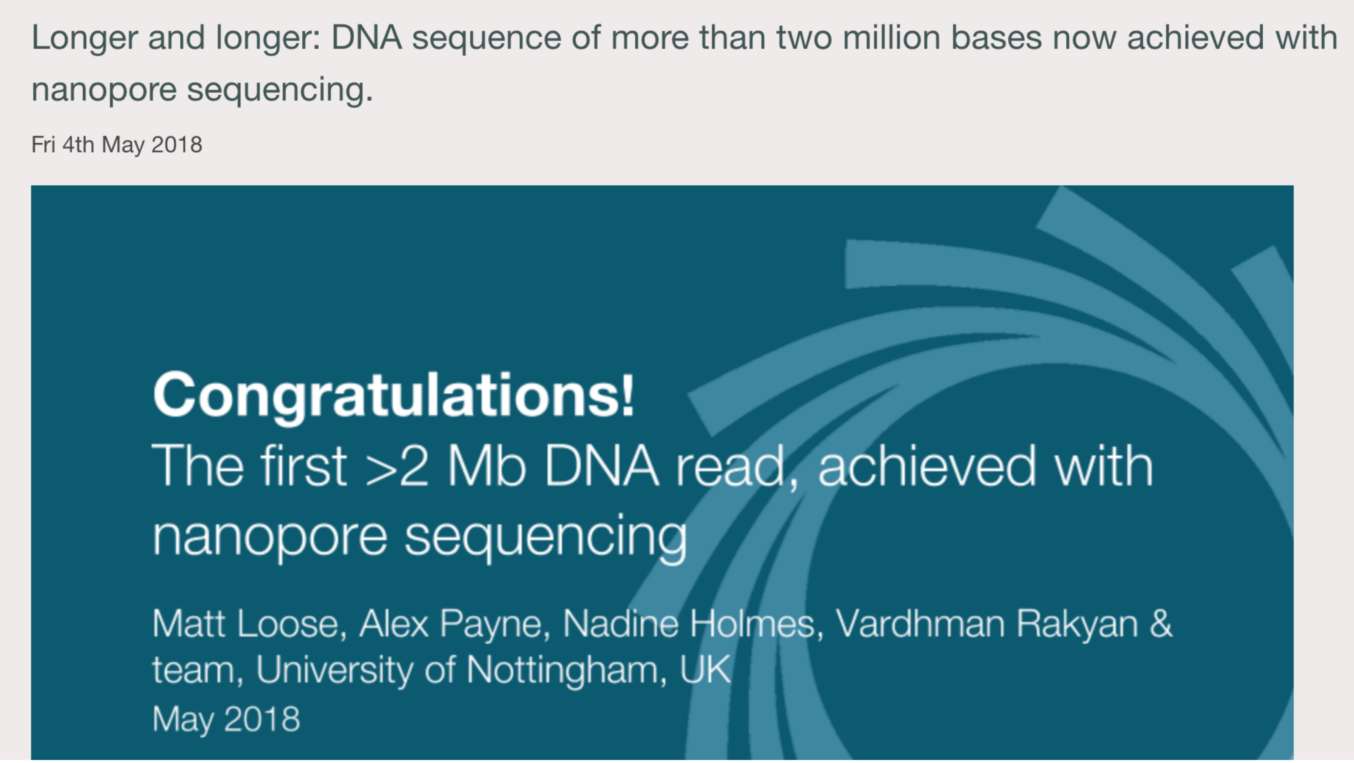




QIAGEN GeneReader

Emerging Technologies

The race for long is on



Longer and longer: DNA sequence of more than two million bases now achieved with nanopore sequencing.

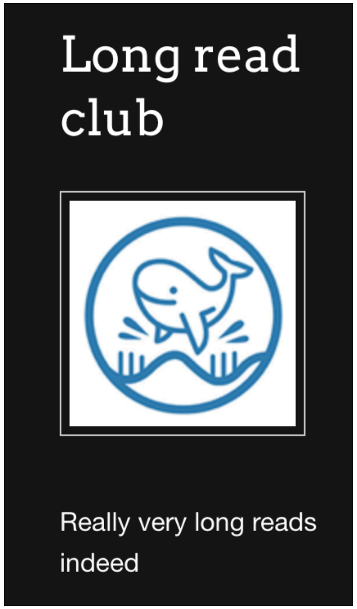
Fri 4th May 2018

Congratulations!

The first >2 Mb DNA read, achieved with nanopore sequencing

Matt Loose, Alex Payne, Nadine Holmes, Vardhman Rakyan & team, University of Nottingham, UK

May 2018

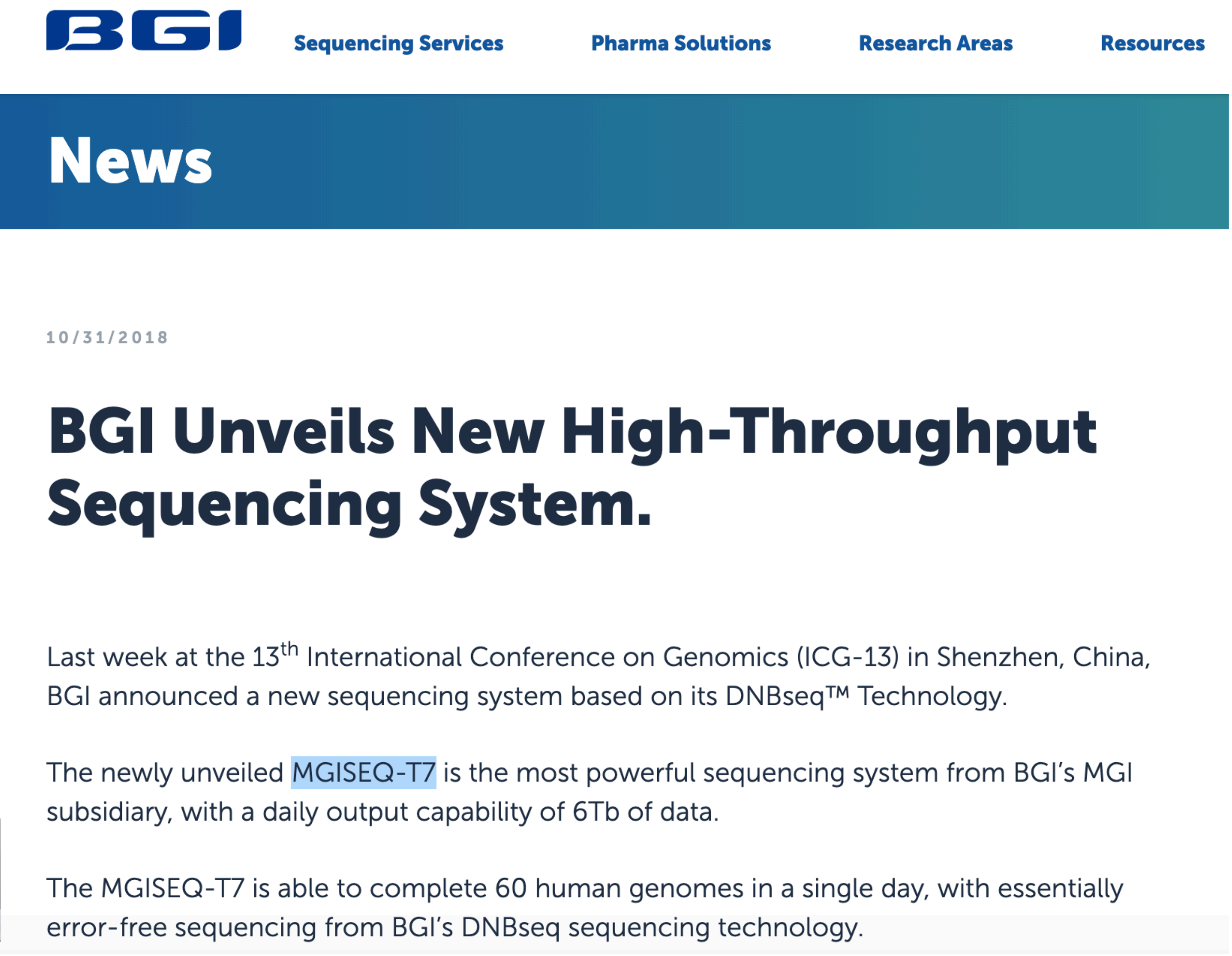


Long read club

Really very long reads indeed

http://longreadclub.org/

https://nanoporetech.com/about-us/news/longer-and-longer-dna-sequence-more-two-million-bases-now-achieved-nanopore



10/31/2018

BGI Unveils New High-Throughput Sequencing System.

Last week at the 13th International Conference on Genomics (ICG-13) in Shenzhen, China, BGI announced a new sequencing system based on its DNBseq™ Technology.

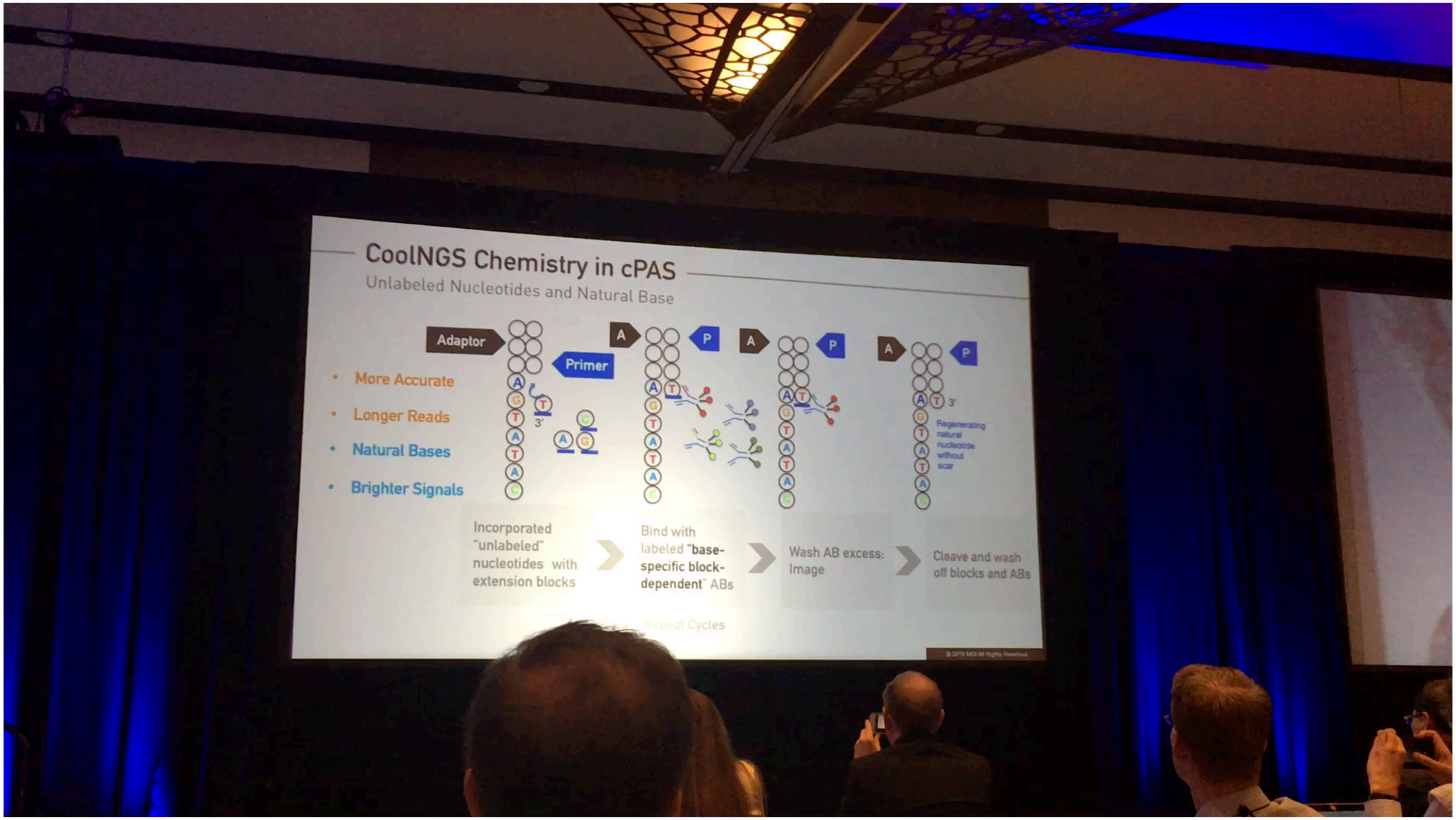
The newly unveiled MGISEQ-T7 is the most powerful sequencing system from BGI's MGI subsidiary, with a daily output capability of 6Tb of data.

The MGISEQ-T7 is able to complete 60 human genomes in a single day, with essentially error-free sequencing from BGI's DNBseq sequencing technology.

https://www.bgi.com/us/company/news/bgi-unveils-new-high-throughput-sequencing-system/

T-1000?

BGI - CoolNGS



CoolNGS Chemistry in cPAS

Unlabeled Nucleotides and Natural Base

- More Accurate
- Longer Reads
- Natural Bases
- Brighter Signals

Incorporated "unlabeled" nucleotides with extension blocks

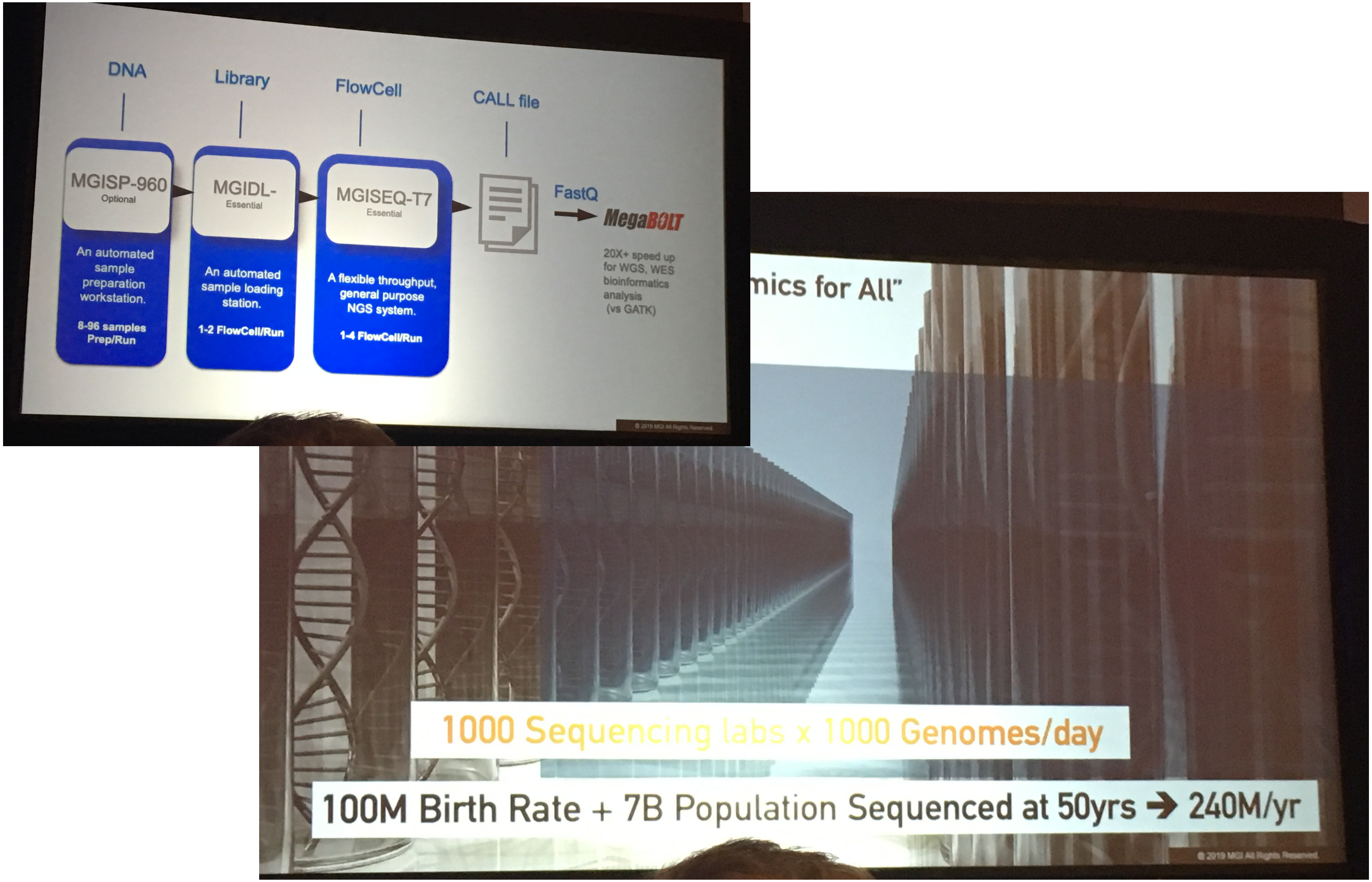
Bind with labeled "base-specific block-dependent" ABs

Wash AB excess; Image

Cleave and wash off blocks and ABs

Repeat Cycles

BGI – NGS streets



DNA

Library

FlowCell

CALL file

MGISP-960
Optional

An automated sample preparation workstation.

8-96 samples Prep/Run

MGIDL-
Essential

An automated sample loading station.

1-2 FlowCell/Run

MGISEQ-T7
Essential

A flexible throughput, general purpose NGS system.

1-4 FlowCell/Run

FastQ

MegaBOLT

20X+ speed up for WGS, WES bioinformatics analysis (vs GATK)

© 2019 MGI All Rights Reserved.

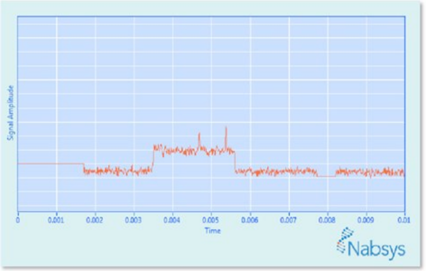
mics for All"

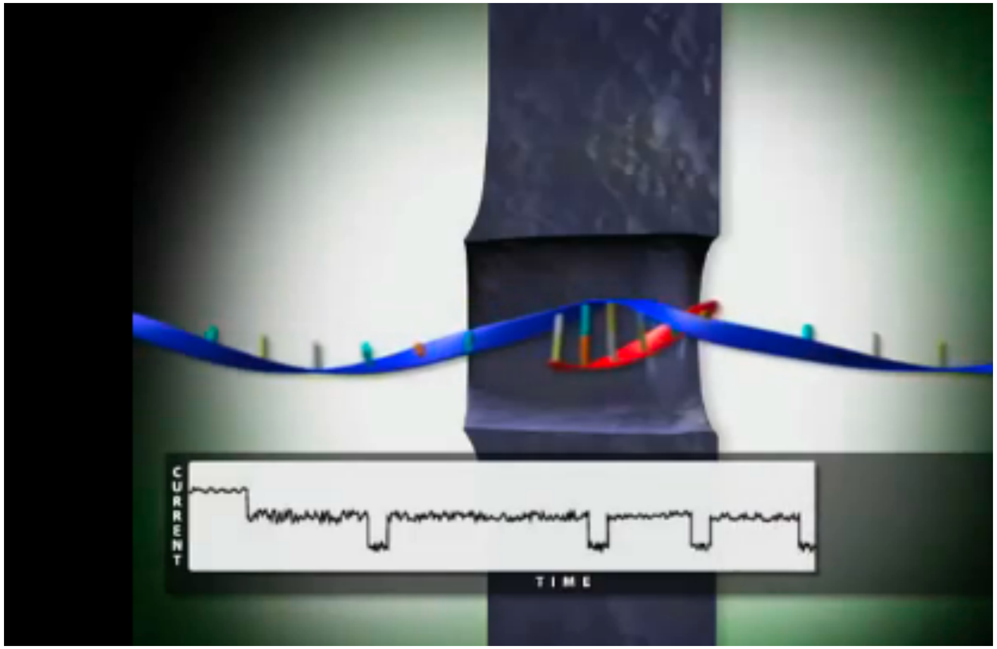
1000 Sequencing labs x 1000 Genomes/day

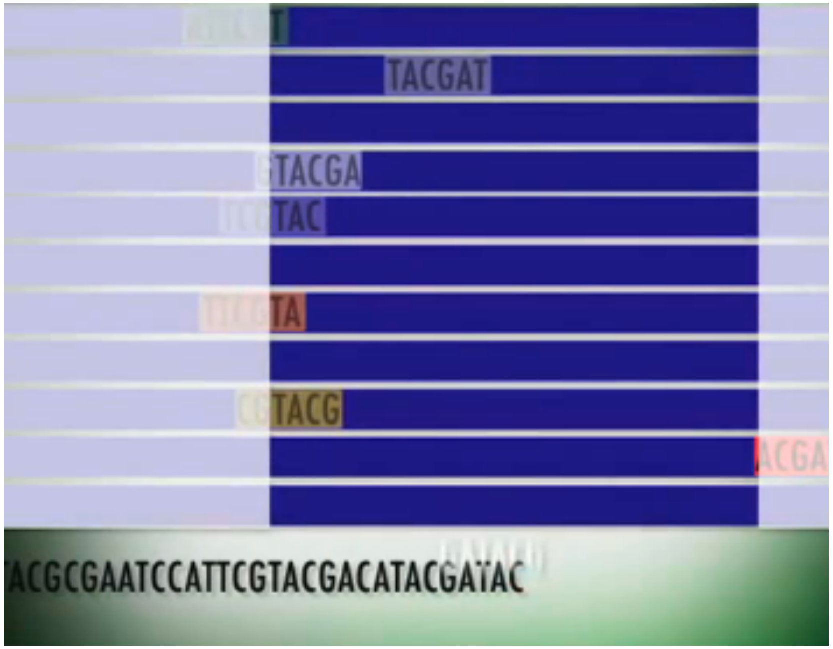
100M Birth Rate + 7B Population Sequenced at 50yrs → 240M/yr

© 2019 MGI All Rights Reserved.

Nabsys







Hybridization -Assisted Nanopore Sequencing (HANS):

-1 million bases per second
-Variable probe length can be used for HANS
-Long Reads (100kb)
-Single molecule

ZS Genetics, Inc.

Working At The Scale Of Life

Single-atom labeling and then visualization with EM

-Long Reads (20kb)
-Single molecule

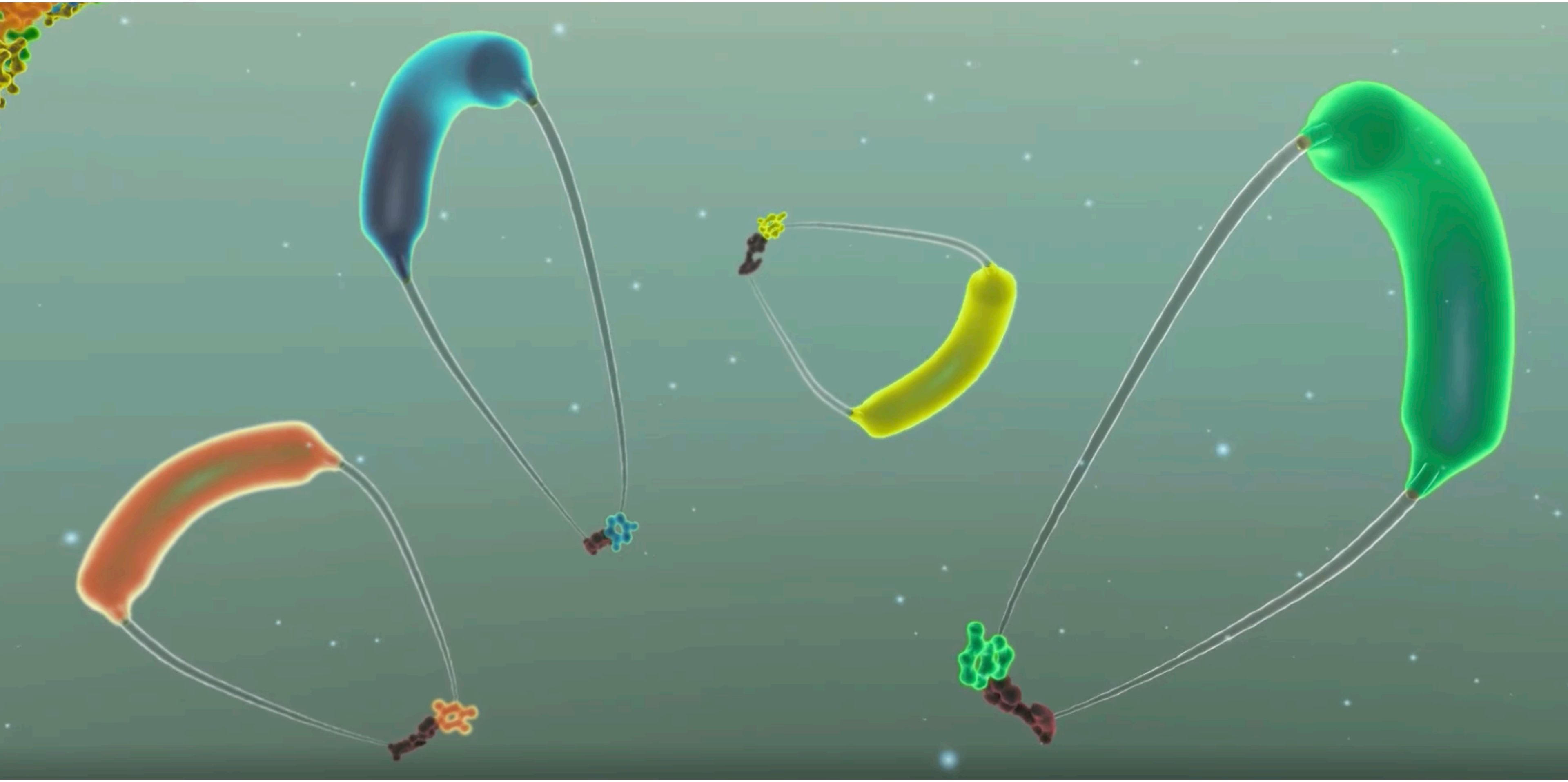
The new Illumina Firefly (iSeq100) can sequence in <6h.

GenapSys

(1M, 16M or 144M)

GenapSys

Roche’s nanopore tech



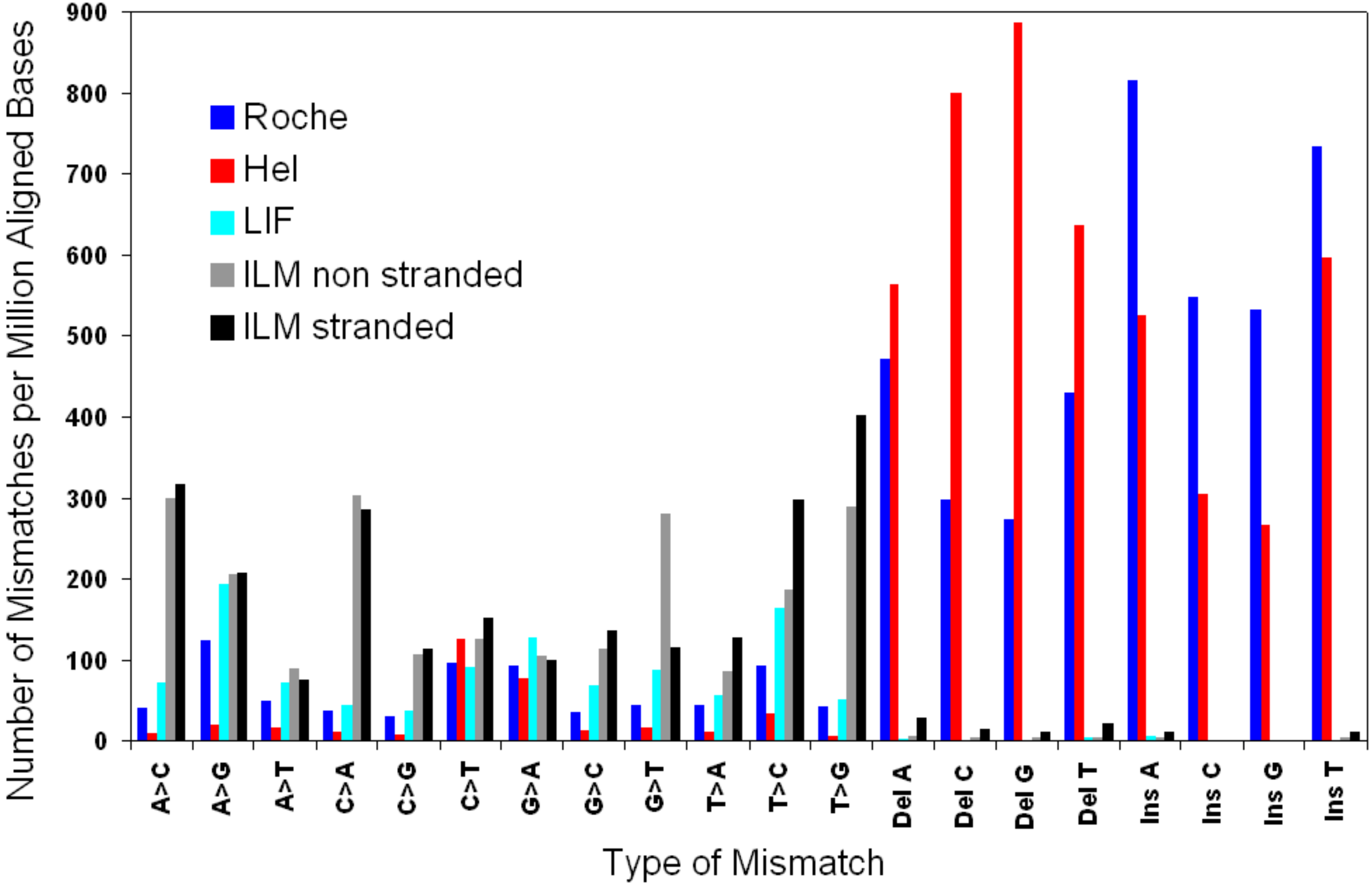
Sequencing by eXpansion (SBX)

https://sequencing.roche.com/en/science-education/technology/nanopore-sequencing.html

Each Platform has various sources of noise, and thus Error

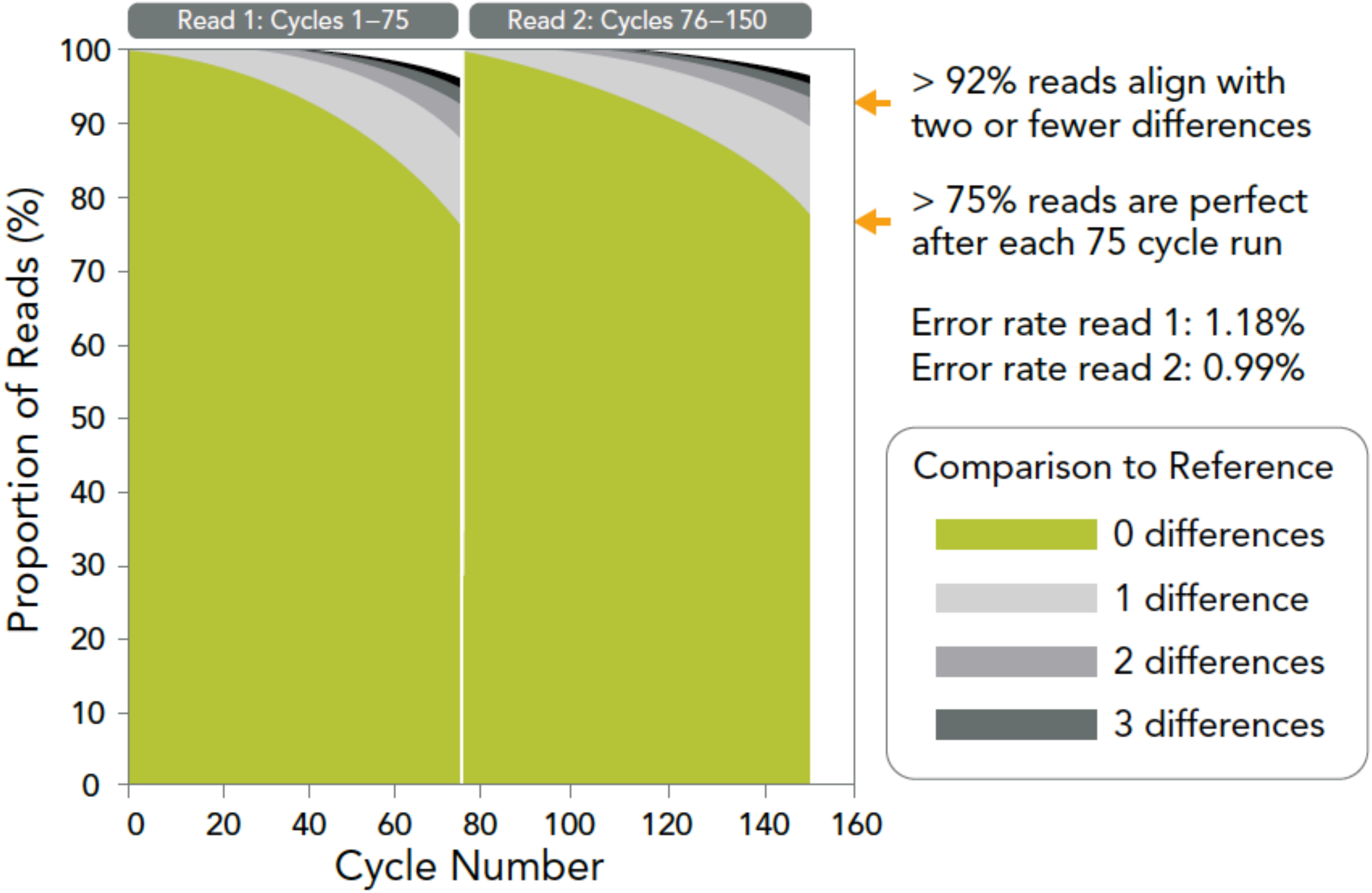
- De-Phasing
  - Lagging strand dephasing from incomplete extension
  - Leading strand dephasing from over-extension
- Dark Nucleotides
- Polymerase errors (10^{-5} to 10^{-7})
- Single molecule challenges
  - High noise
  - Polymerase “wiggling” from tail
- Platform-specific errors
  - Illumina more likely to have error after ‘G’
  - PCR-based methods miss GC- and AT-rich regions

Each platform is slightly different, and so intrinic errors are different



SeQC Consortium

Many platforms are cycle-dependent on error rate - ILMN



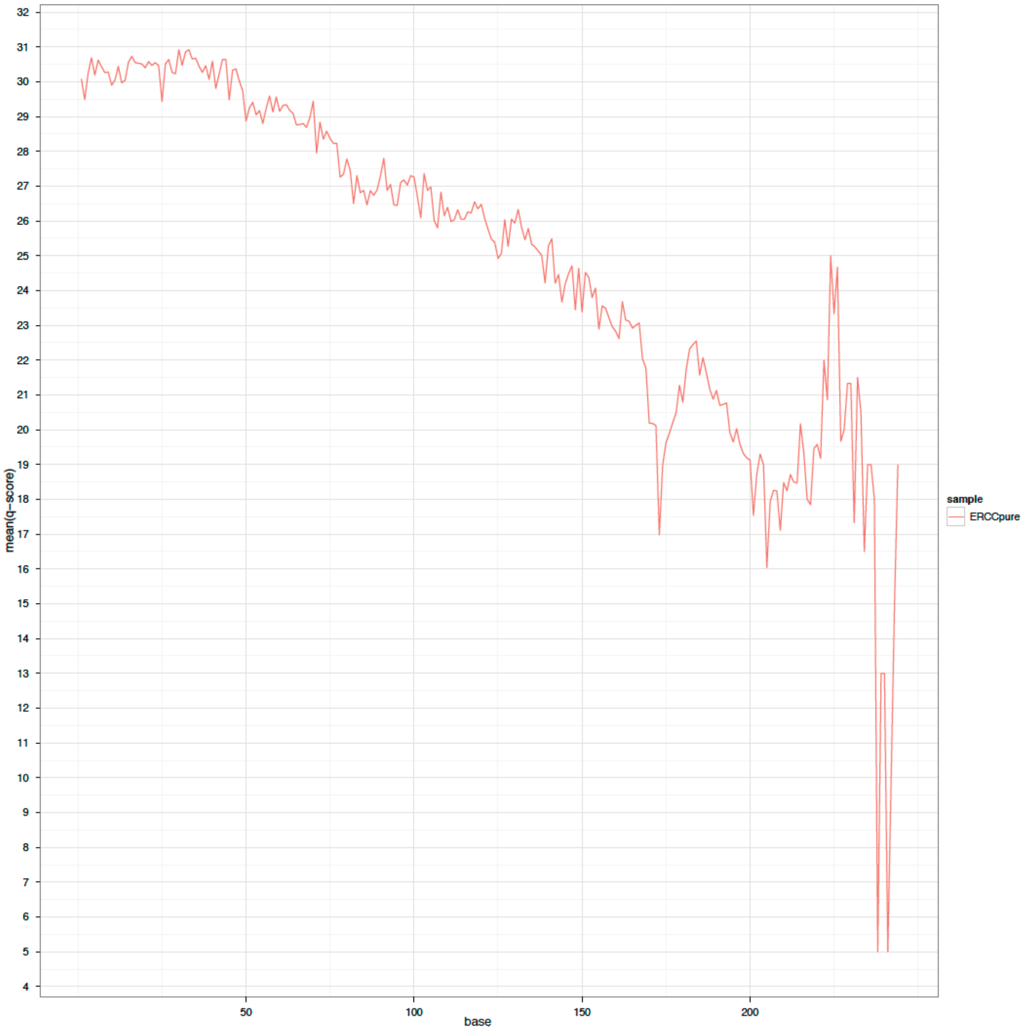
> 92% reads align with two or fewer differences

> 75% reads are perfect after each 75 cycle run

Error rate read 1: 1.18%
Error rate read 2: 0.99%

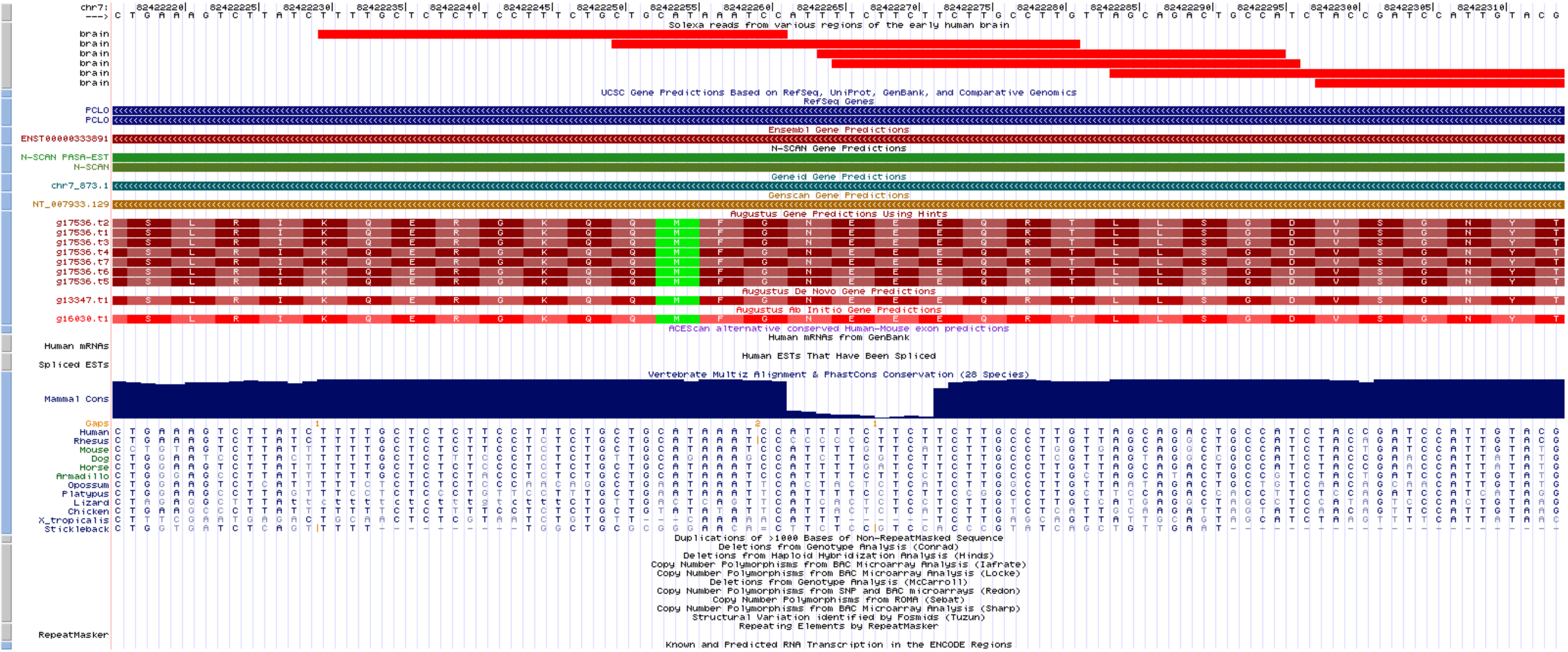
© Illumina, Inc.

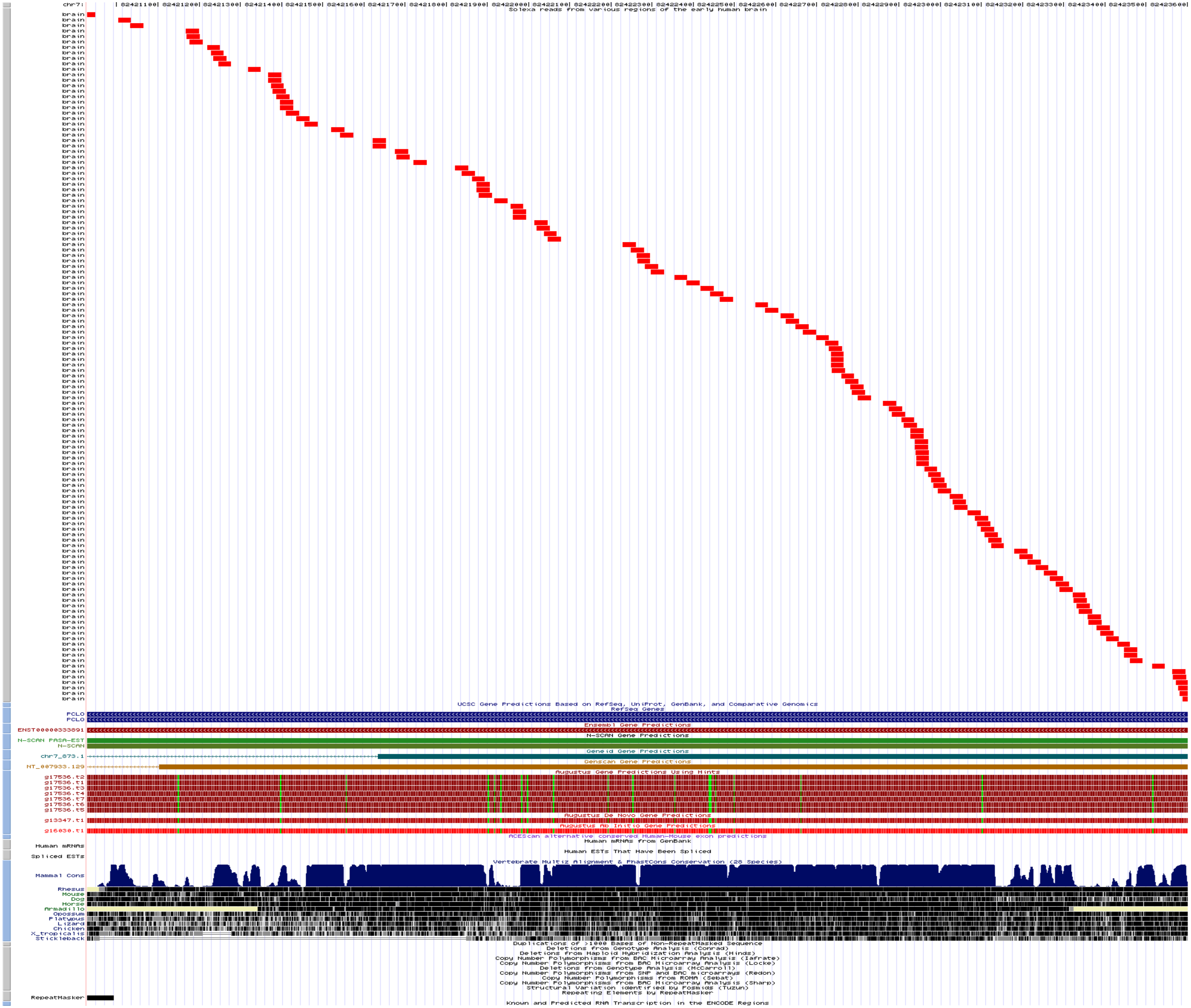
Many platforms are cycle-dependent on error rate - ION



What do you do with the reads?

Alignment to the genome


chr7: 82422220 82422225 82422230 82422235 82422240 82422245 82422250 82422255 82422260 82422265 82422270 82422275 82422280 82422285 82422290 82422295 82422300 82422305 82422310
---> C T G A A A G T C T T A T C T T T T G C T C T C T T C C T T T C T G C T G C A T A A A T C C A T T T T C T T C T T C T T G C C T T G T T A G C A G A C T G C C A T C T A C C G A T C C A T T G T A C G
Solexa reads from various regions of the early human brain
brain
brain
brain
brain
brain
brain
UCSC Gene Predictions Based on RefSeq, UniProt, GenBank, and Comparative Genomics
RefSeq Genes
PCLO
PCLO
Ensembl Gene Predictions
ENST00000333891
N-SCAN Gene Predictions
N-SCAN PASA-EST
N-SCAN
Geneid Gene Predictions
chr7_873.1
Genscan Gene Predictions
NT_007933.129
Augustus Gene Predictions Using Hints
g17536.t2
g17536.t1
g17536.t3
g17536.t4
g17536.t7
g17536.t6
g17536.t5
S L R I K Q E R G K Q Q M F G N E E E Q R T L L S G D V S G N Y T
Augustus De Novo Gene Predictions
g13347.t1
Augustus Ab Initio Gene Predictions
g16030.t1
ACEScan alternative conserved Human-Mouse exon predictions
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
Vertebrate Multiz Alignment & PhastCons Conservation (28 Species)
Mammal Cons
Gaps
Human
Rhesus
Mouse
Dog
Horse
Armadillo
Opossum
Platypus
Lizard
Chicken
X_tropicalis
Stickleback
Duplications of >1000 Bases of Non-RepeatMasked Sequence
Deletions from Genotype Analysis (Conrad)
Deletions from Haploid Hybridization Analysis (Hinds)
Copy Number Polymorphisms from BAC Microarray Analysis (Iafrate)
Copy Number Polymorphisms from BAC Microarray Analysis (Locke)
Deletions from Genotype Analysis (McCarroll)
Copy Number Polymorphisms from SNP and BAC microarrays (Redon)
Copy Number Polymorphisms from ROMA (Sebat)
Copy Number Polymorphisms from BAC Microarray Analysis (Sharp)
Structural Variation identified by Fosmids (Tuzun)
Repeating Elements by RepeatMasker
RepeatMasker
Known and Predicted RNA Transcription in the ENCODE Regions


chr7:
| 82421100| 82421200| 82421300| 82421400| 82421500| 82421600| 82421700| 82421800| 82421900| 82422000| 82422100| 82422200| 82422300| 82422400| 82422500| 82422600| 82422700| 82422800| 82422900| 82423000| 82423100| 82423200| 82423300| 82423400| 82423500| 82423600|
Solexa reads from various regions of the early human brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
brain
UCSC Gene Predictions Based on RefSeq, UniProt, GenBank, and Comparative Genomics
RefSeq Genes
PCLO
PCLO
Ensembl Gene Predictions
ENST00000333891
N-SCAN Gene Predictions
N-SCAN PASA-EST
N-SCAN
Geneid Gene Predictions
chr7_873.1
Genscan Gene Predictions
NT_007933.129
Augustus Gene Predictions Using Hints
g17536.t2
g17536.t1
g17536.t3
g17536.t4
g17536.t7
g17536.t6
g17536.t5
Augustus De Novo Gene Predictions
g13347.t1
Augustus Ab Initio Gene Predictions
g16030.t1
ACEScan alternative conserved Human-Mouse exon predictions
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
Vertebrate Multiz Alignment & PhastCons Conservation (28 Species)
Mammal Cons
Rhesus
Mouse
Dog
Horse
Armadillo
Opossum
Platypus
Lizard
Chicken
X_tropicalis
Stickleback
Duplications of >1000 Bases of Non-RepeatMasked Sequence
Deletions from Genotype Analysis (Conrad)
Deletions from Haploid Hybridization Analysis (Hinds)
Copy Number Polymorphisms from BAC Microarray Analysis (Iafrate)
Copy Number Polymorphisms from BAC Microarray Analysis (Locke)
Deletions from Genotype Analysis (McCarroll)
Copy Number Polymorphisms from SNP and BAC microarrays (Redon)
Copy Number Polymorphisms from ROMA (Sebat)
Copy Number Polymorphisms from BAC Microarray Analysis (Sharp)
Structural Variation identified by Fosmids (Tuzun)
Repeating Elements by RepeatMasker
RepeatMasker
Known and Predicted RNA Transcription in the ENCODE Regions

The reads: FASTQ

The most common format is FASTQ, based off the FASTA data format:

```
>SequenceID
CGTAGTCTATATATGCGCGAATGCGTA
```

But….

FASTQ also includes quality information:

```
@Sample_Info
CCTTGCTGCC
+
3.6;#$!>><
```

Understanding FASTQ

For Illumina, sequences have an ID:

@HWUSI-EAS100R:6:73:941:1973#0/1

HWUSI-EAS100R	the unique instrument name
6	flowcell lane
73	tile number within the flowcell lane
941	'x'-coordinate of the cluster within the tile
1973	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 (paired-end or mate-pair reads only)

Understanding Quality Scores

Q-values are the probability (p) of a base being incorrect. From Sanger sequencing:

Q_{value}=-10$\log_{10}$p

So, if your p=0.1, then $Q_{value} = (-10\log_{10}(0.1))$

$= (-10(-1)) = 10$

If your p=0.01, then $Q_{value} = (-10\log_{10}(0.01))$

$= (-10(-2)) = 20$

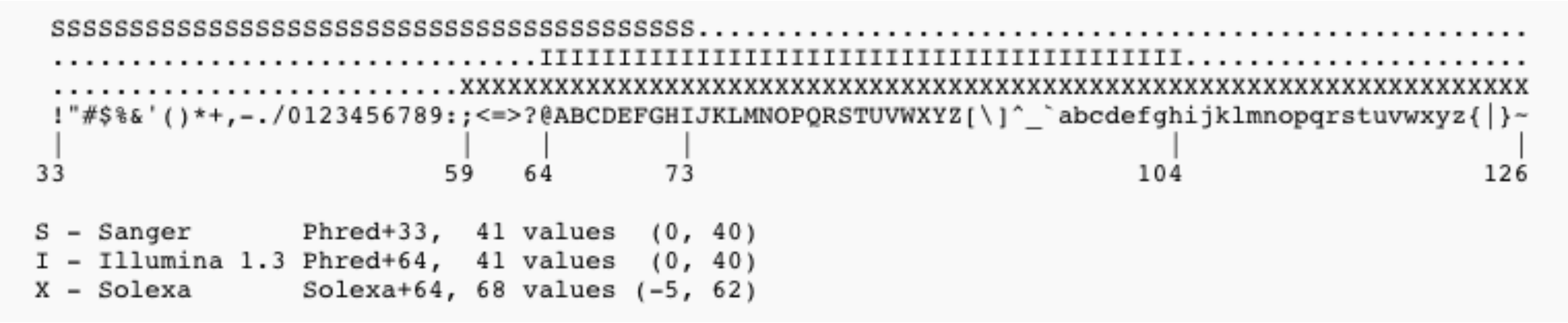
If p=0.001, then $Q_{value} = (-10\log_{10}(0.001))$

$= (-10(-3)) = 30$

Understanding Quality Scores

Q-values are the probability (p) of a base being incorrect, but it is most efficient to represent this with a single bit in ASCII (American Standard Code for Information Interchange) format.

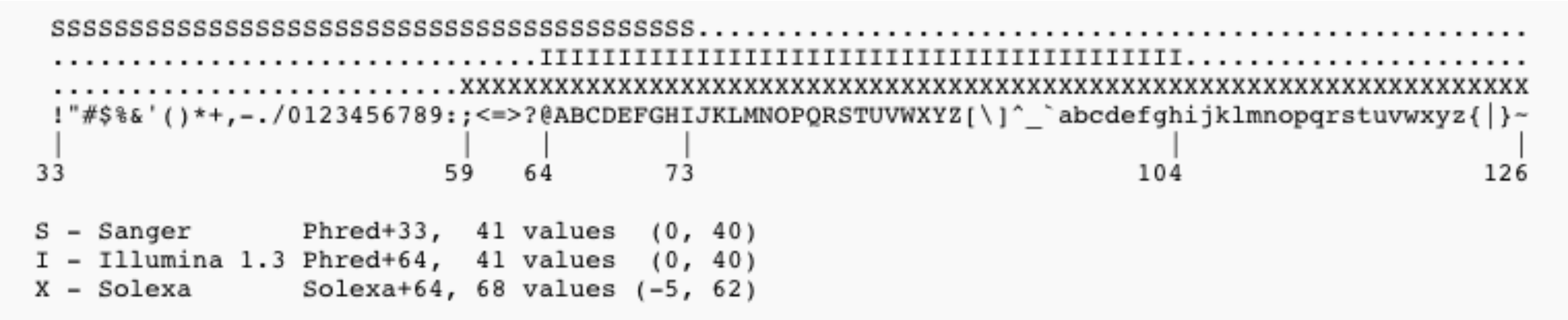
The first 32 symbols in ASCII are control characters, so we start at 33.



```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126

S - Sanger        Phred+33,  41 values  (0, 40)
I - Illumina 1.3  Phred+64,  41 values  (0, 40)
X - Solexa        Solexa+64, 68 values (-5, 62)
```

Phred-Based Base Quality



```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                             104                  126

S - Sanger        Phred+33,  41 values  (0, 40)
I - Illumina 1.3  Phred+64,  41 values  (0, 40)
X - Solexa        Solexa+64, 68 values (-5, 62)
```

If your ASCII character is ‘B’, then 66-64=2, so

$P=10^{-Q/10}$

$-0.2 = \log_{10}p$

$10^{-0.2}$ =p, so p=0.63, or 63% change of an incorrect base.

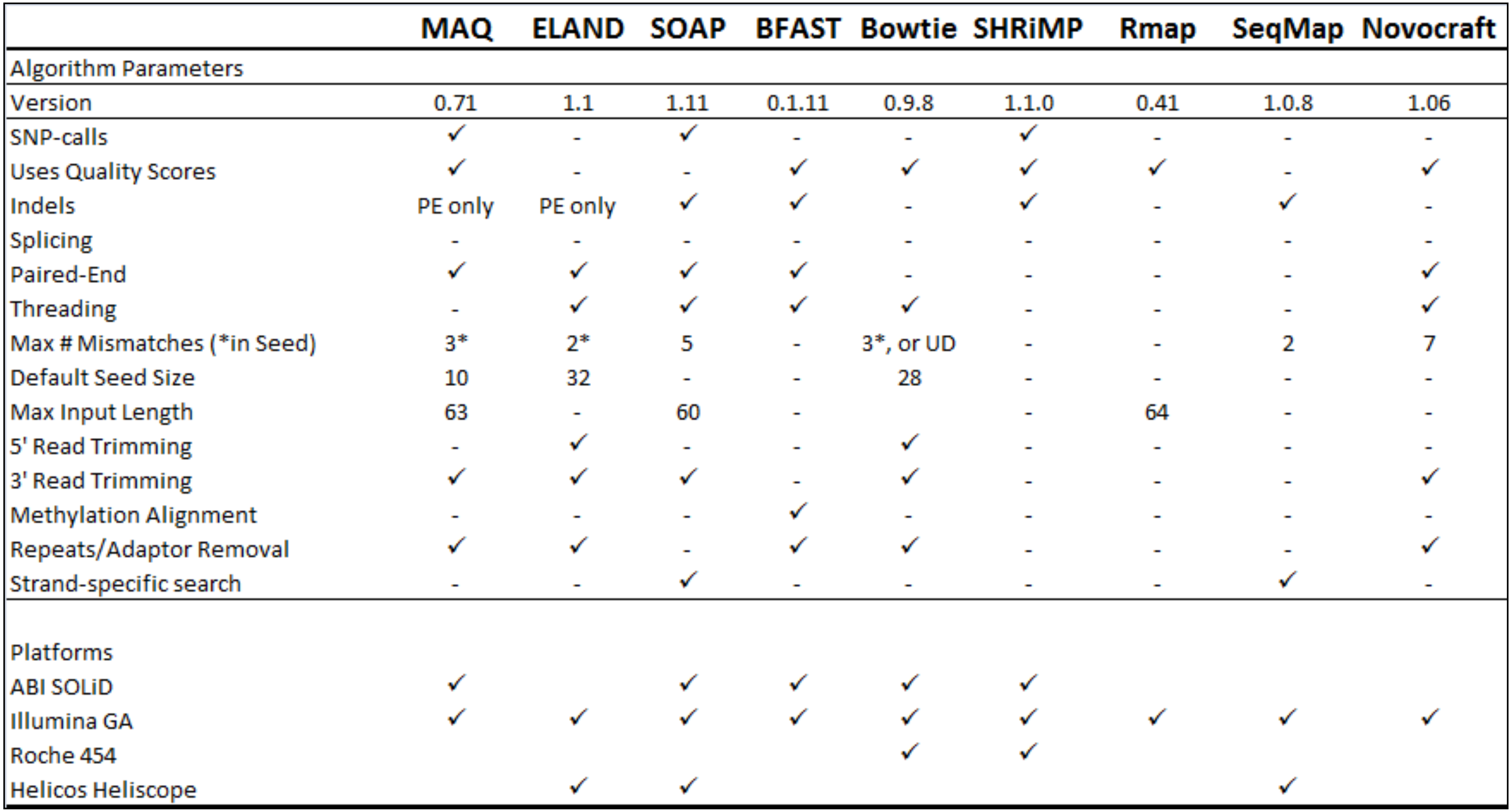
If your ASCII character is ‘h’, then 104-64=40, so

$40 \quad = (-10\log_{10}p)$

$-4.0 \quad = \log_{10}p$

10^{-4} =p, so p=0.0001, or 0.01% change of an incorrect base.

Many Options for Alignment - 2009

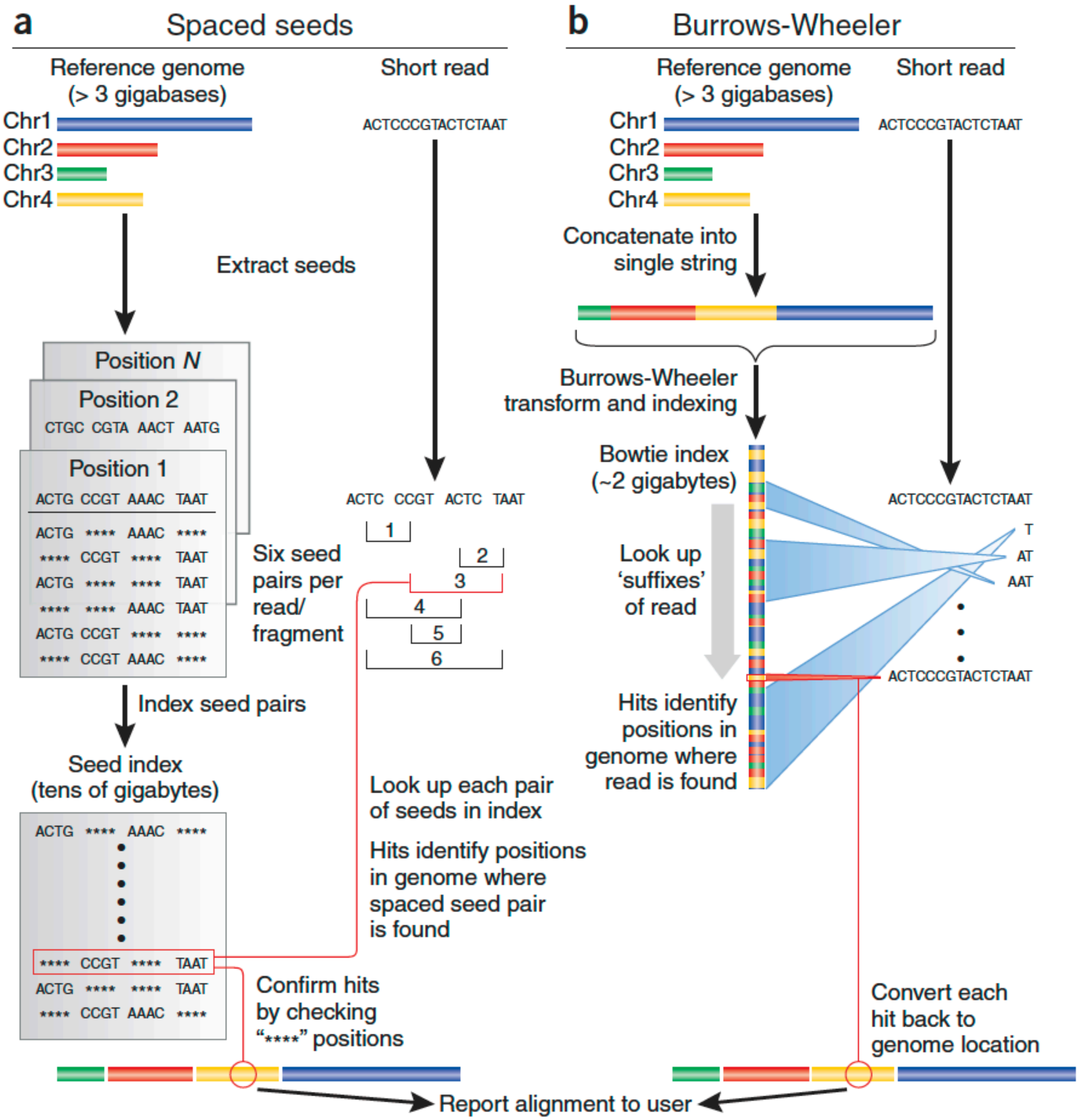


	MAQ	ELAND	SOAP	BFAST	Bowtie	SHRiMP	Rmap	SeqMap	Novocraft
Algorithm Parameters									
Version	0.71	1.1	1.11	0.1.11	0.9.8	1.1.0	0.41	1.0.8	1.06
SNP-calls	✓	-	✓	-	-	✓	-	-	-
Uses Quality Scores	✓	-	-	✓	✓	✓	✓	-	✓
Indels	PE only	PE only	✓	✓	-	✓	-	✓	-
Splicing	-	-	-	-	-	-	-	-	-
Paired-End	✓	✓	✓	✓	-	-	-	-	✓
Threading	-	✓	✓	✓	✓	-	-	-	✓
Max # Mismatches (*in Seed)	3*	2*	5	-	3*, or UD	-	-	2	7
Default Seed Size	10	32	-	-	28	-	-	-	-
Max Input Length	63	-	60	-		-	64	-	-
5' Read Trimming	-	✓	-	-	✓	-	-	-	-
3' Read Trimming	✓	✓	✓	-	✓	-	-	-	✓
Methylation Alignment	-	-	-	✓	-	-	-	-	-
Repeats/Adaptor Removal	✓	✓	-	✓	✓	-	-	-	✓
Strand-specific search	-	-	✓	-	-	-	-	✓	-
Platforms									
ABI SOLiD	✓		✓	✓	✓	✓			
Illumina GA	✓	✓	✓	✓	✓	✓	✓	✓	✓
Roche 454					✓	✓			
Helicos Heliscope		✓	✓					✓	

Many Options for Alignment - 2021

- Bfast
- BioScope
- Bowtie
- BWA
- CLC bio
- CloudBurst
- Eland/Eland2
- GenomeMapper
- GnuMap
- Karma
- MAQ
- MOM
- Mosaik
- MrFAST/MrsFAST
- NovoAlign
- PASS
- PerM
- RazerS
- RMAP
- SSAHA2
- Segemehl
- SeqMap
- SHRiMP
- Slider/SliderII
- SOAP/SOAP2
- Srprism
- Stampy
- vmatch
- ZOOM
- ......

Many common methods are BW-based



Trapnell and Salzberg, 2010

Burrows-Wheeler Transformation (BWT)

- First discovered in 1983 by Wheeler at AT&T Bell Labs
- Used for compression in 1994.
- First implemented for aligners with “Bowtie”
  Ben Langmead, Cole Trapnell, Mihai Pop,
  and Steven Salzberg
- Allows for fast searching with a small memory footprint

http://bio-bwa.sourceforge.net/

Li H. and Durbin R. “Fast and accurate short read alignment with Burrows-Wheeler transform.” (2009) *Bioinformatics*, 25, 1754-60.

Burrows M, Wheeler DJ. “A Block Sorting Lossless Data Compression Algorithm.” Technical Report 124. Palo Alto, CA: Digital Equipment Corporation; 1994.

Questions?