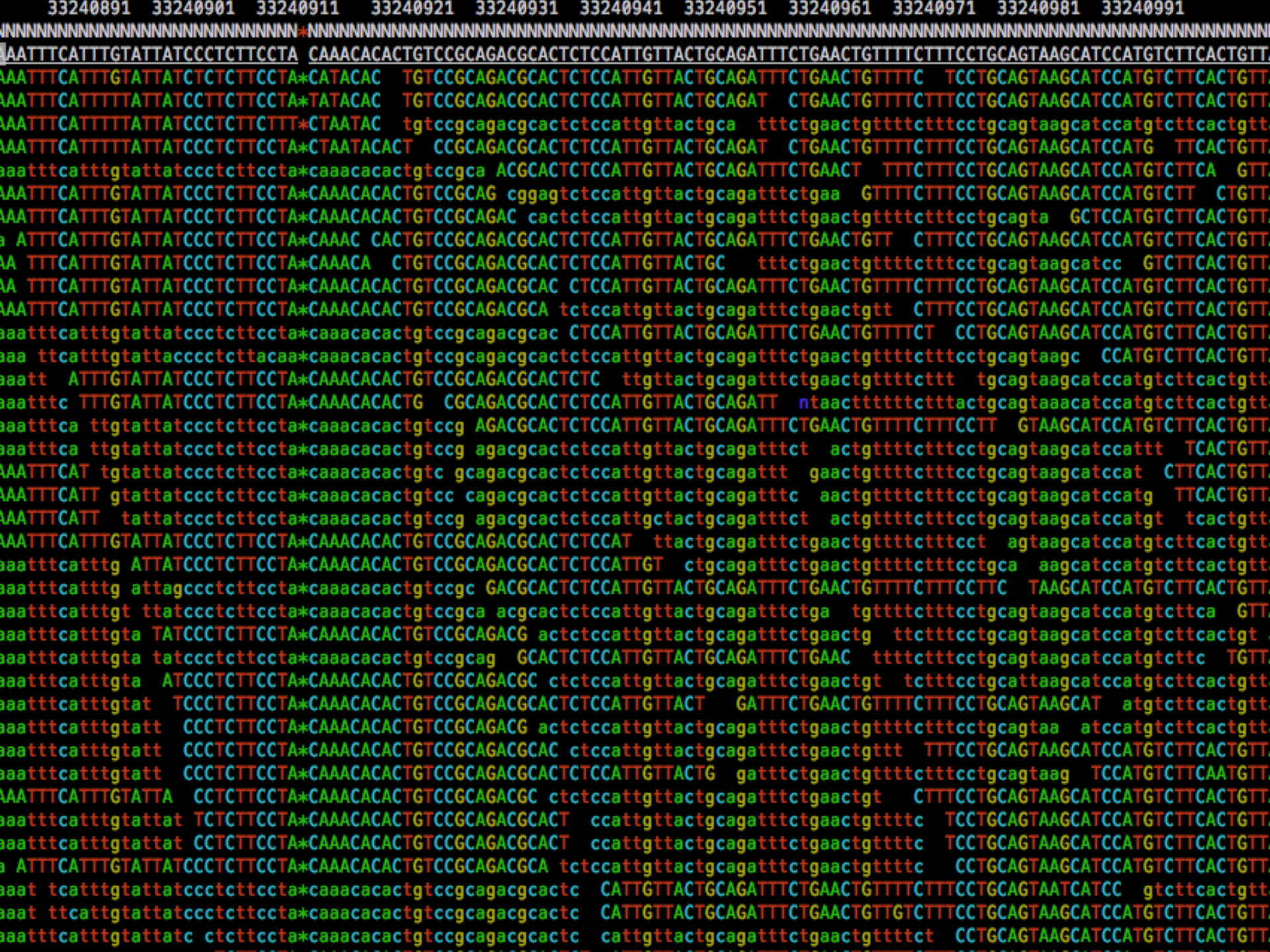


Clinical and Research Genomics
Spring 2022

Professor:

Christopher E. Mason, Ph.D.

Instructors:

Ebrahim Afshinnekoo, M.D.

Jaden Hastings, Ph.D.

TA:

Chandrima Bhattacharya, M.S.

Course Sessions:

I. **Sequencing Methods, Single-Cell Dynamics, and Molecular Detection Techniques (March 29th)**
II. **RNA Sequencing, Epitranscriptomes, and Single Cell / Spatial Omics (April 5th)**
III. **Epigenomes, DNA Modifications, and Chromatin Dynamics (April 12th)**
IV. **Metagenomes, BGCs, and Metabolomics (April 19th)**
V. **Complex Genome Re-arrangements, Transposons, and Tools for Genetic Variant Calling (April 26th)**
VI. **Multi-Omics, Spatial Omics, and Machine Learning (May 3rd)**
VII. **Synthetic Biology, Engineering Systems & Genome Ethics (May 10th)**
VIII. **COVID-19 Tracking and Pathophysiology (May 17th)**
IX. **Global Health and Beyond-Globe Health (Aerospace Medicine) (May 24th)**

All classes also on Zoom:

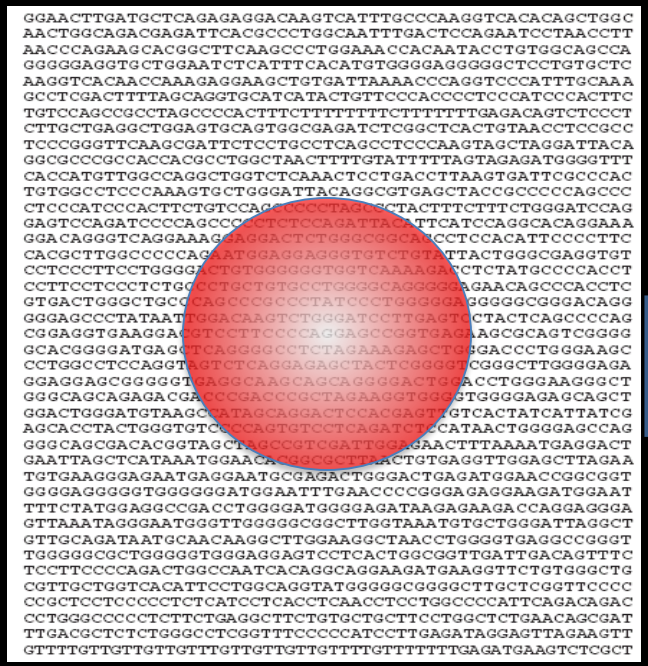
https://weillcornell.zoom.us/j/99565175034

Meeting ID: 995 6517 5034 (Passcode: ClinGen22)

Course webpage:

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

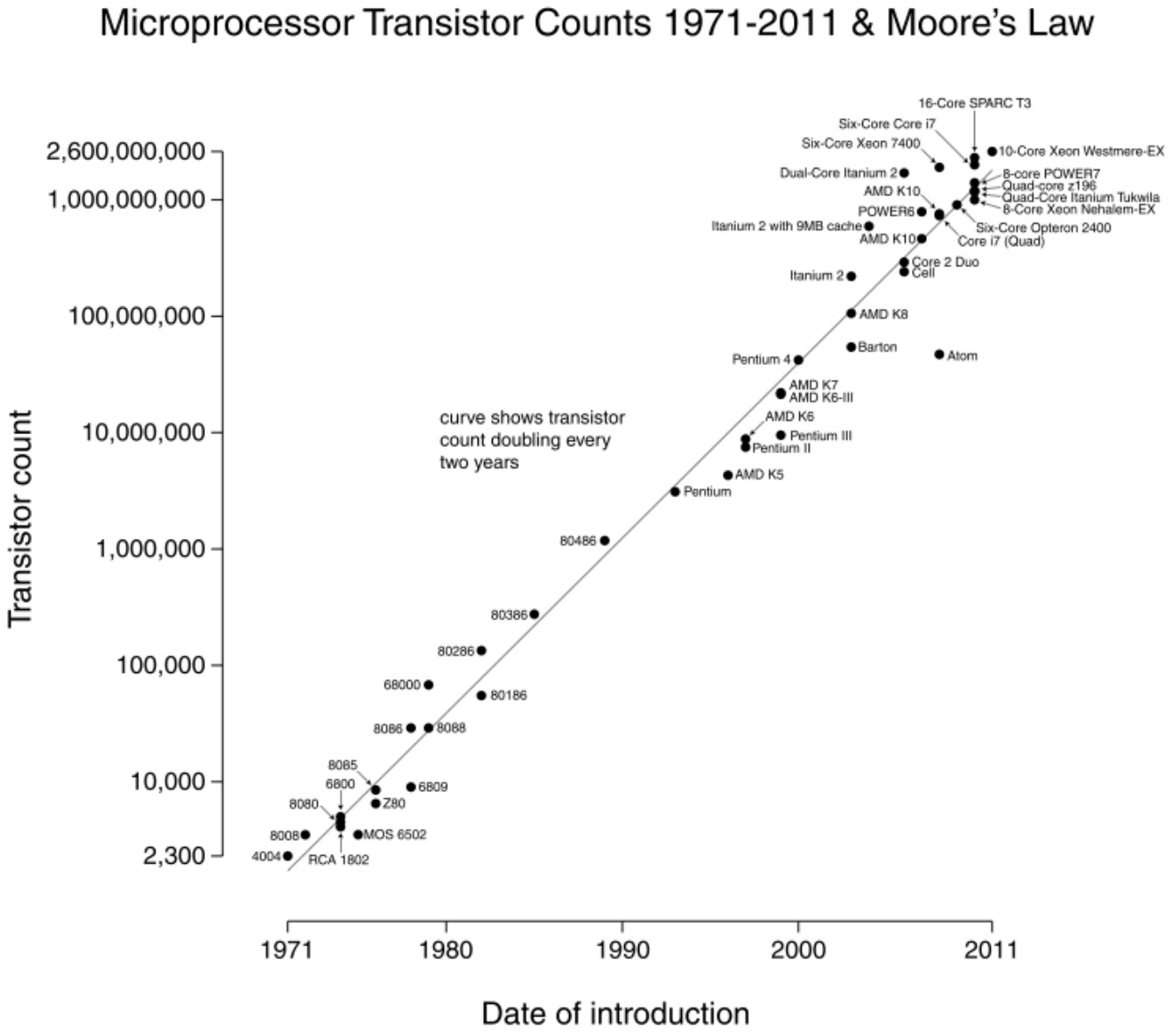
Start



Time

Finish

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



By Wgsimon

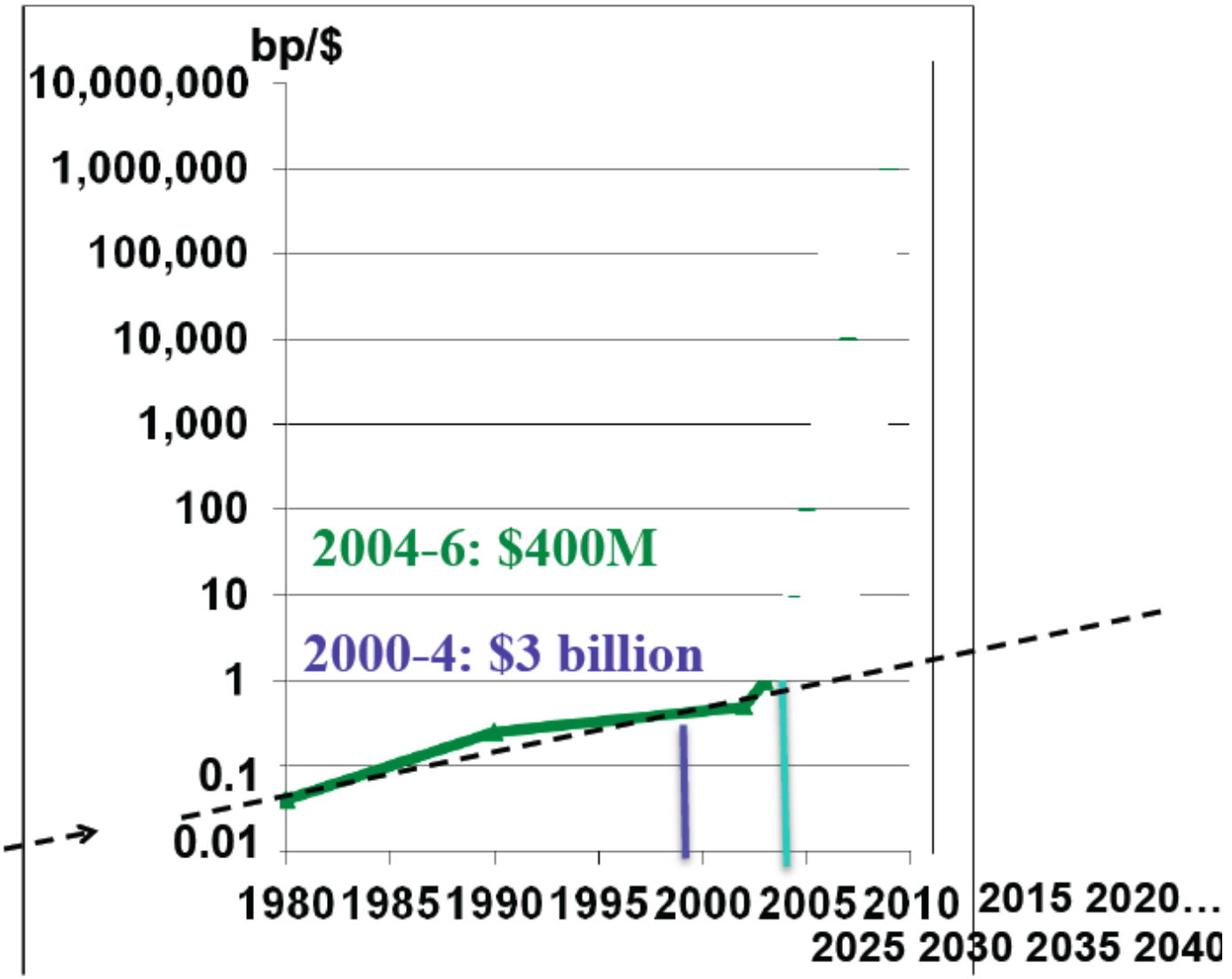
Initially we expected a $1K Genome in 2040

$1000 Genome

When?

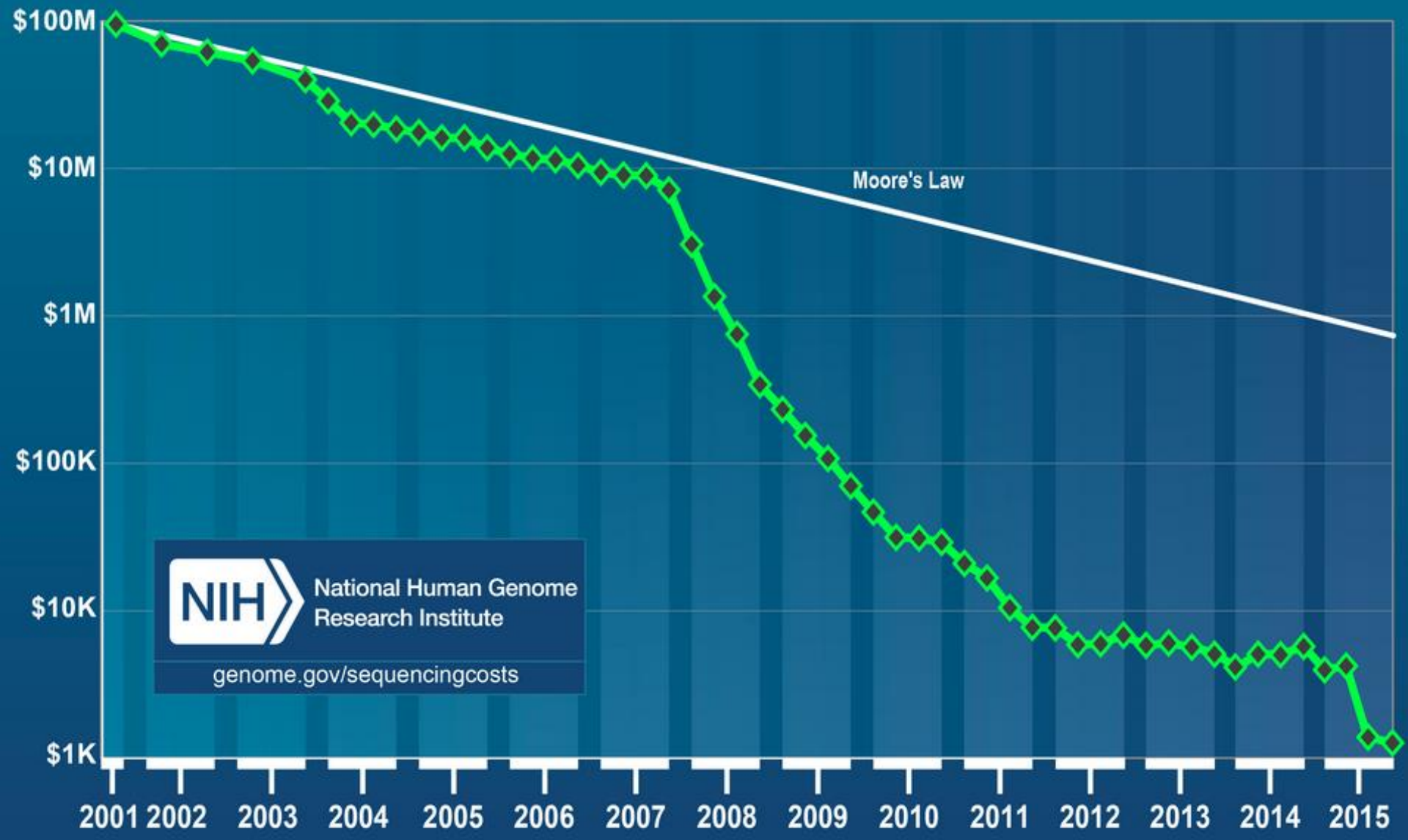
2040

Moore's law 1.5x/yr for electronics

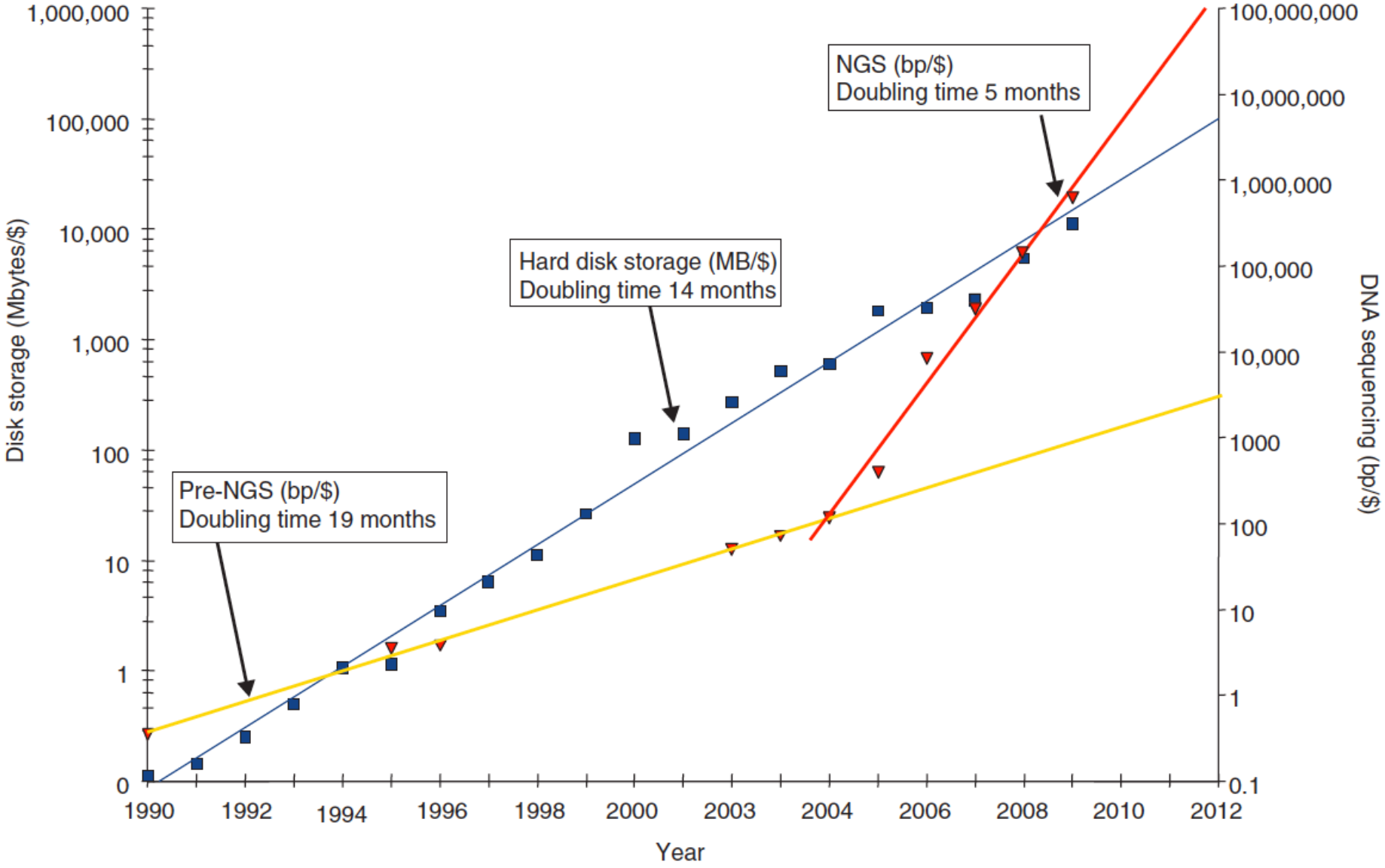


George Church

Cost per Genome



More and more data



nature International weekly journal of science

NATURE | NEWS FEATURE

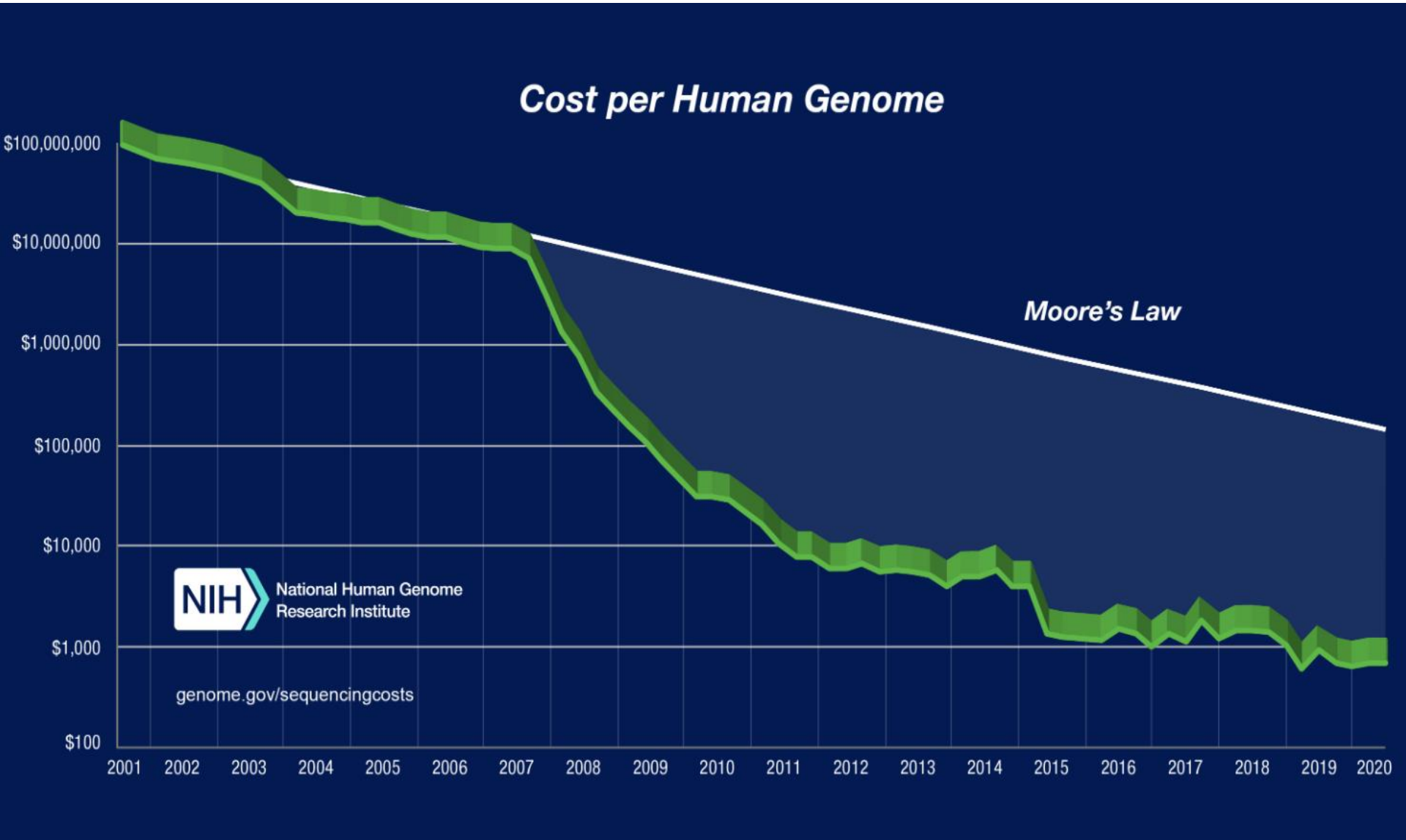
Technology: The $1,000 genome

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

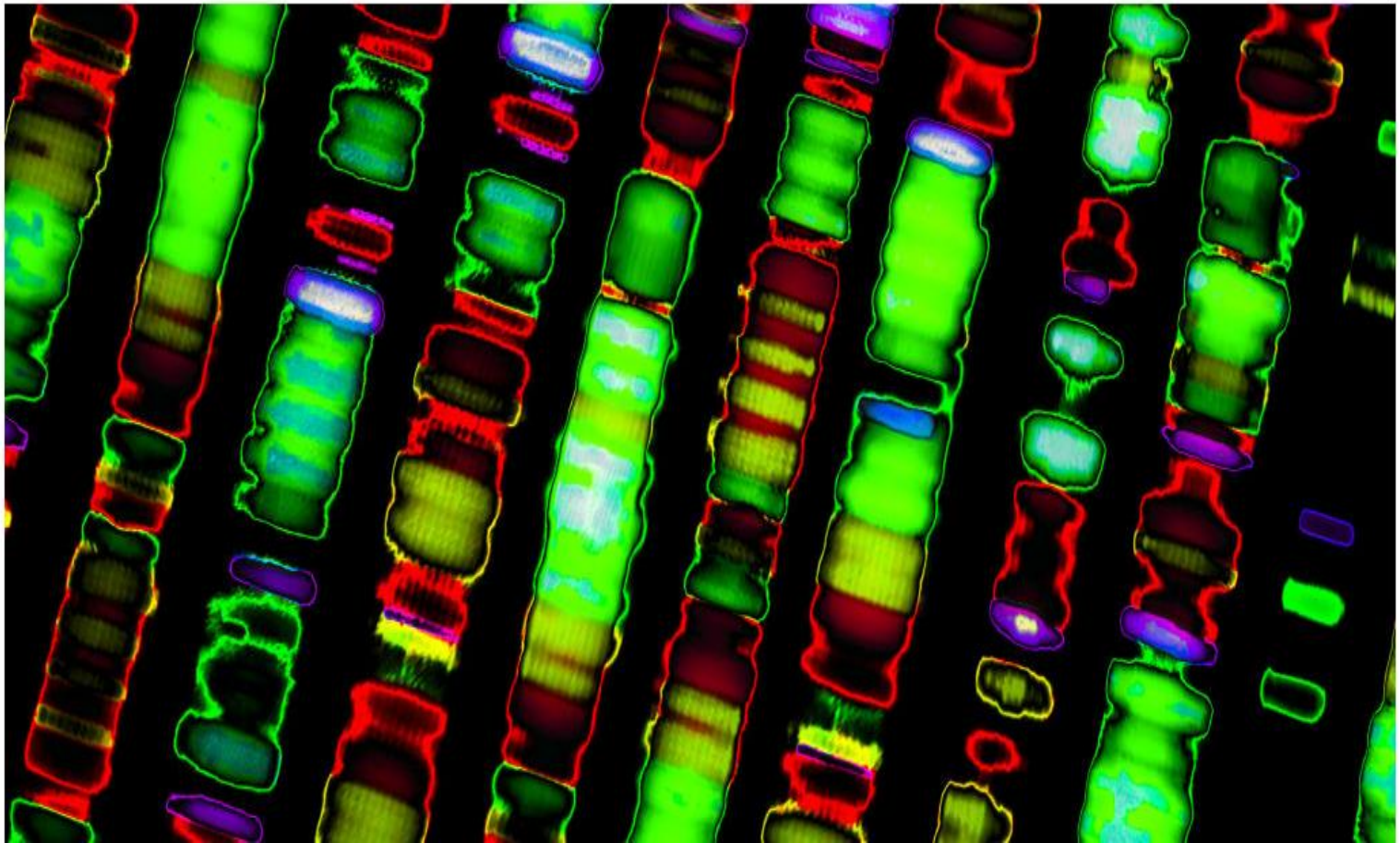
19 March 2014

Flatlined a little

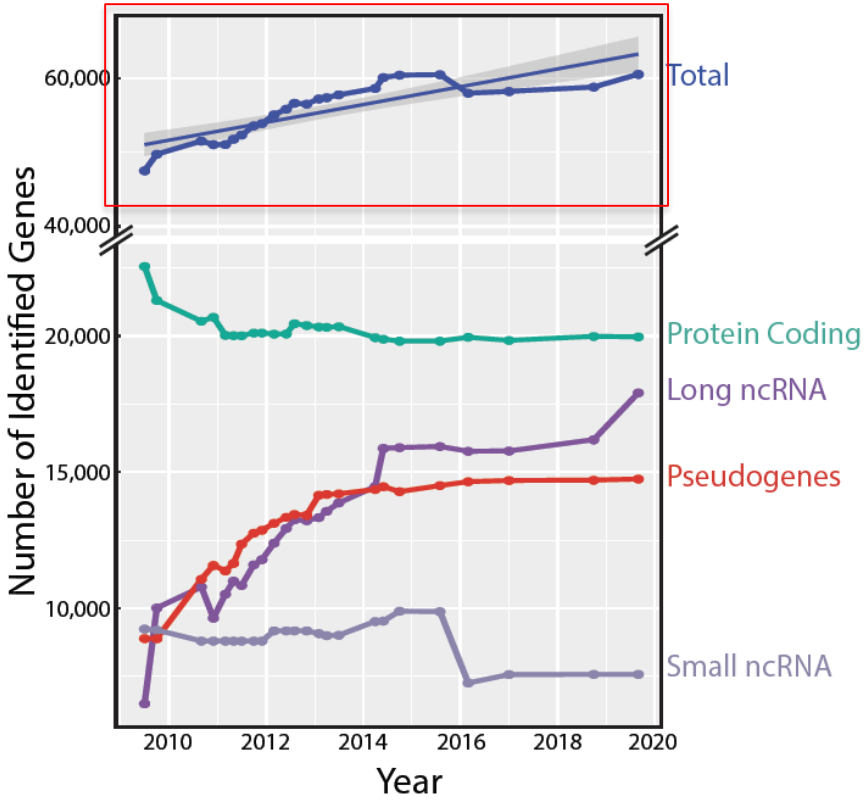


BUSINESS

Illumina says it can deliver a $100 genome — soon



New human genes are still being found



The Next 500 Years
https://www.gencodegenes.org/

gencodegenes.org

GENCODE

Human | Mouse | How to access data | FAQ | Documentation | About us

HUMAN

GENCODE 39 (09.12.21)

MOUSE

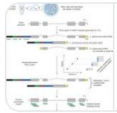
GENCODE M28 (09.12.21)

The goal of the GENCODE project is to identify and classify all gene features in the human and mouse genomes with high accuracy based on biological evidence, and to release these annotations for the benefit of biomedical research and genome interpretation.

Tweets by @GencodeGenes

GencodeGenes
@GencodeGenes

New single-nuclei RNA sequencing method finds interesting alt splicing expression in brain. Still plenty of novel transcripts out there for us to add to the GENCODE geneset! nature.com/articles/s4158 …



Single-nuclei isoform RNA se…
Nature Biotechnology - Complet…
nature.com

Mar 9, 2022

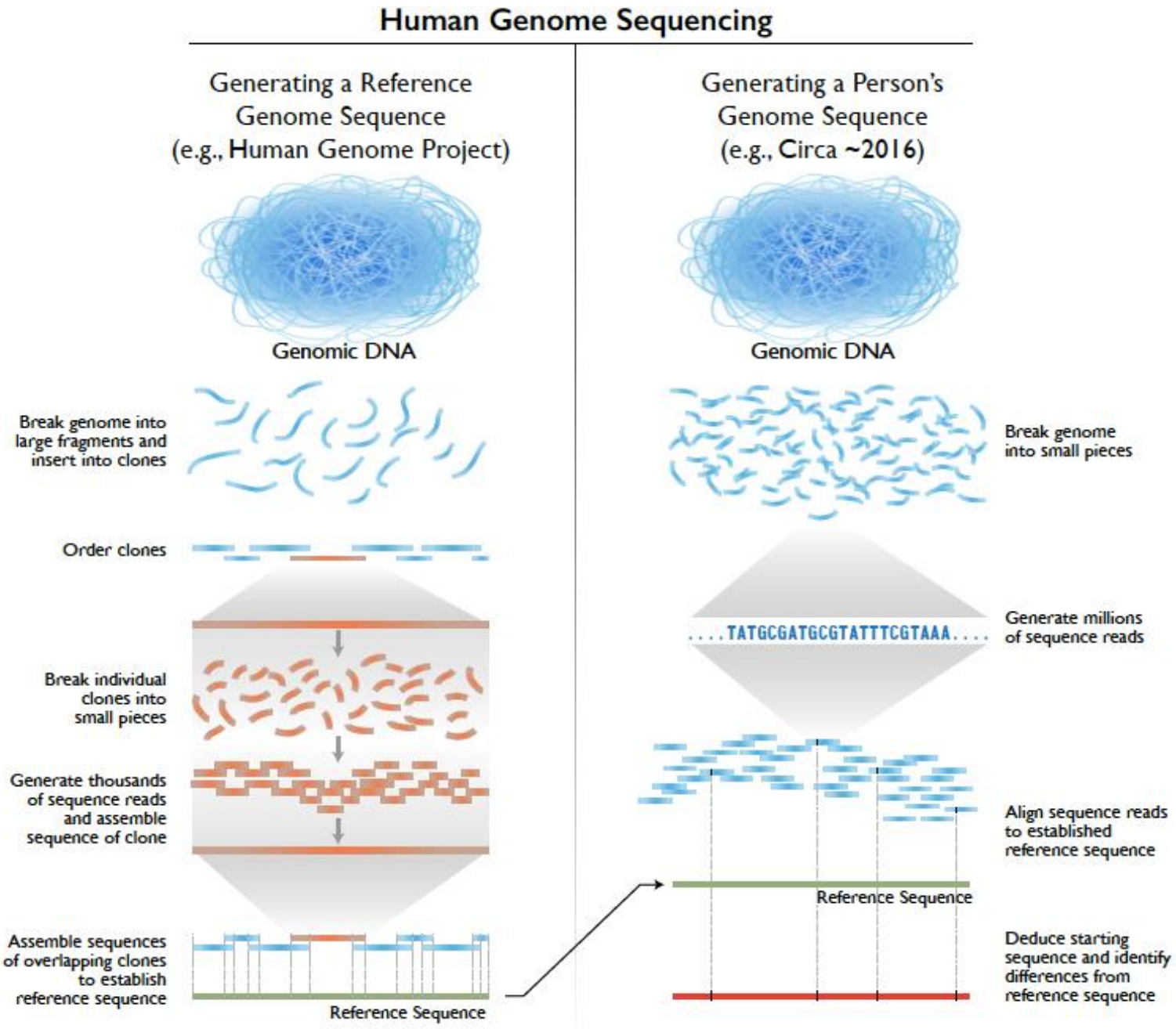
GencodeGenes Retweeted

UCSC Genome Browser
@GenomeBrowser

We have updated GENCODE Gene annotation tracks for human (V39 - hg19/hg38) and mouse (M28 - mm39) corresponding to Ensembl release 105.

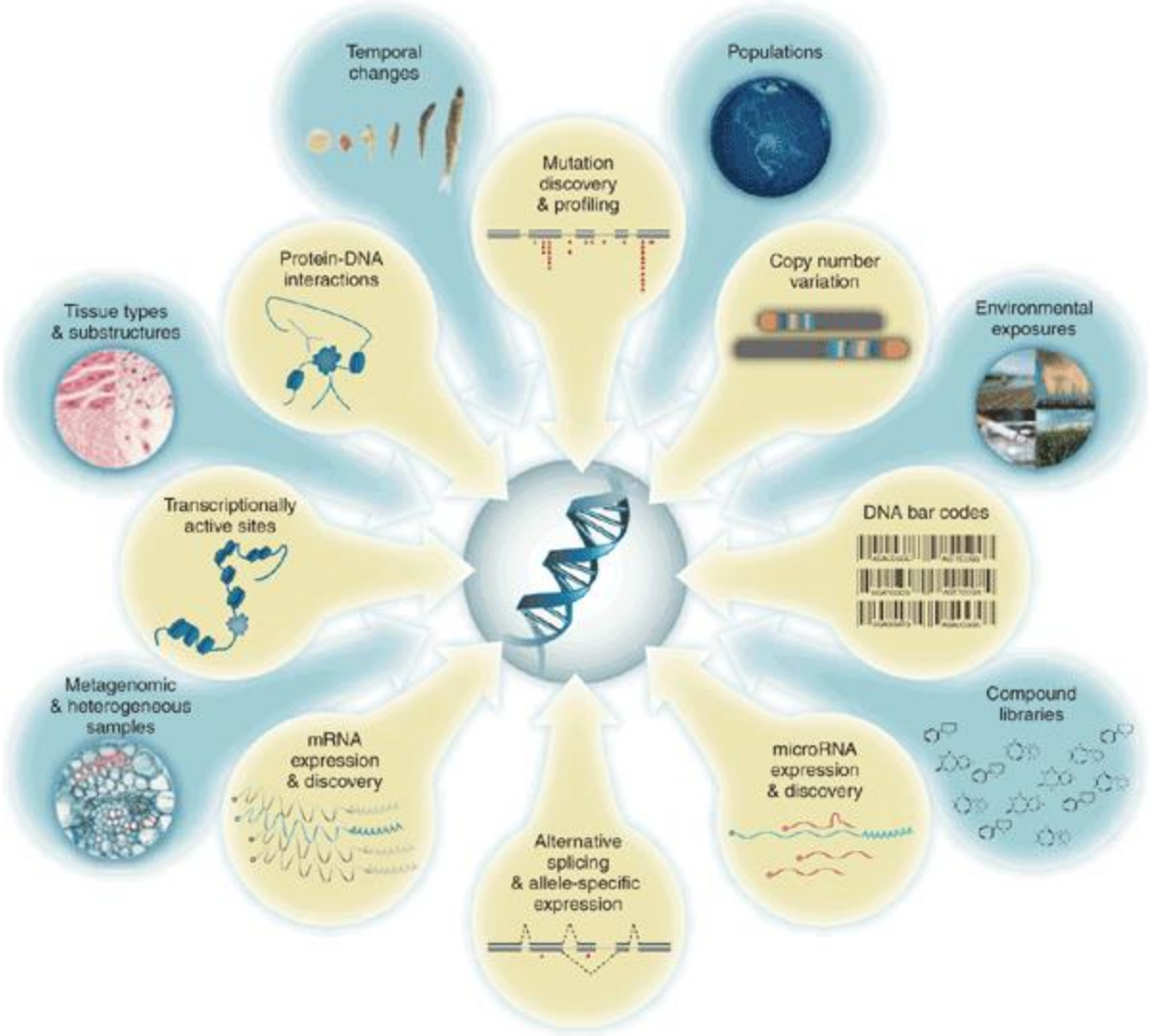
https://www.gencodegenes.org/

Every Day
is the
Best Day



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.



Kahvejian, 2008

Genome Medicine

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
Published: 26 November 2010

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

Genomics england

About Us | 100,000 Genomes Project | Taking Part | For Healthcare Professionals | Research | Industry Partnerships | News & Events

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/

ALL OF US℠ RESEARCH PROGRAM

All of Us Research Program

Scale and Scope
Participation
Program Components
Funding
FAQ
Advisory Groups
Events
Announcements
In the News
Multimedia

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:

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

1 million U.S. Veterans WGS

U.S. Department of Veterans Affairs

Search ORD

VA SITE MAP [A-Z]

Health | Benefits | Burials & Memorials | About VA | Resources | News Room | Locations | Contact Us

VA » Health Care » Office of R&D » Mvp » Million Veteran Program (MVP)

Print

Office of Research & Development

- ORD Home
- About Us
- Services
- Programs
  - See All Programs
  - Animal Research
  - Biosafety & Biosecurity
  - Cooperative Studies Program (CSP)
  - Health Disparities & Minority Health
  - Million Veteran Program (MVP)

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?

Text size: + -

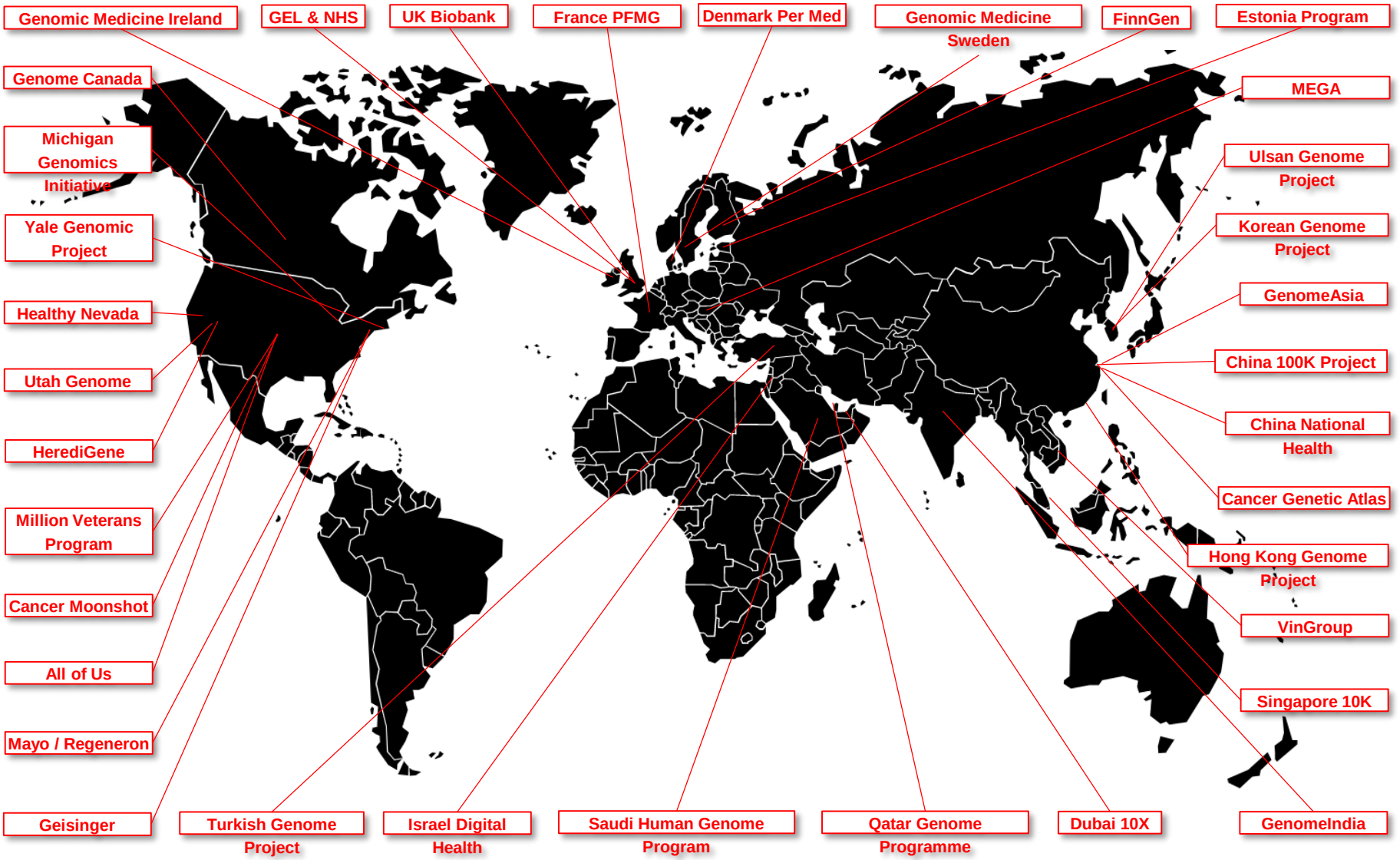
CONTACT MVP

Contact the MVP Information Center toll-free at:

866-441-6075

INFORMED CONSENT

POPULATION-SCALE NGS IS GLOBAL



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

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 onward;
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

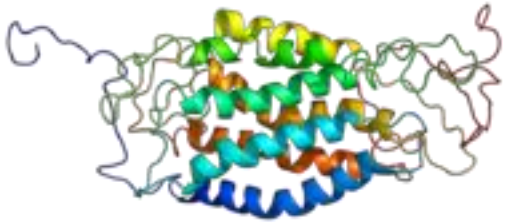
Specific genes can have significant impact

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

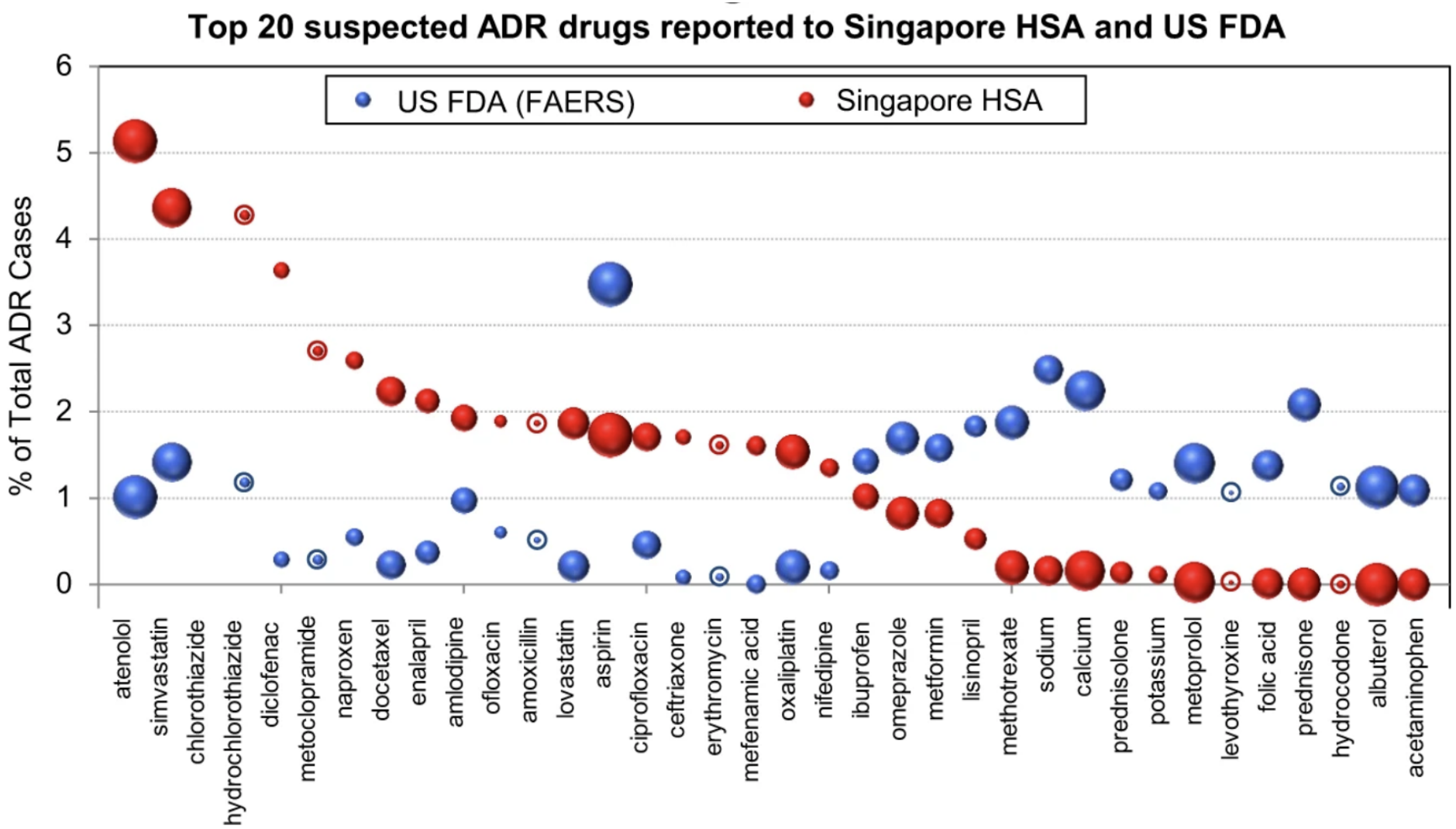
Towards precision medicine: interrogating the human genome to identify drug pathways associated with potentially functional, population-differentiated polymorphisms

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

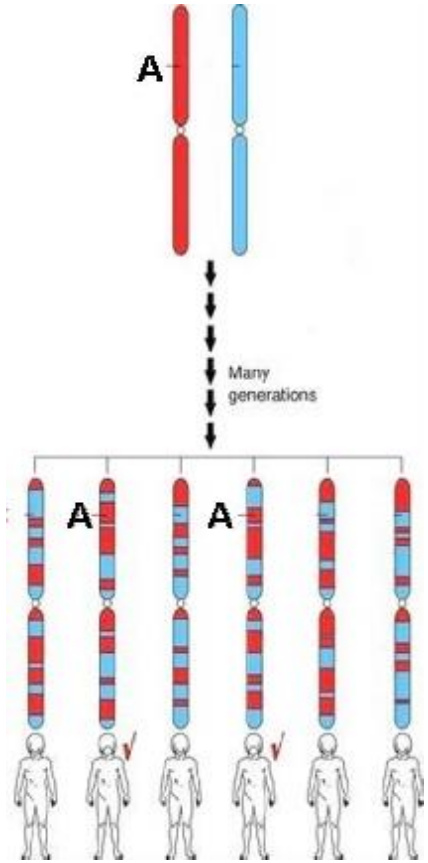
7 Accesses | **22** Altmetric | Metrics »

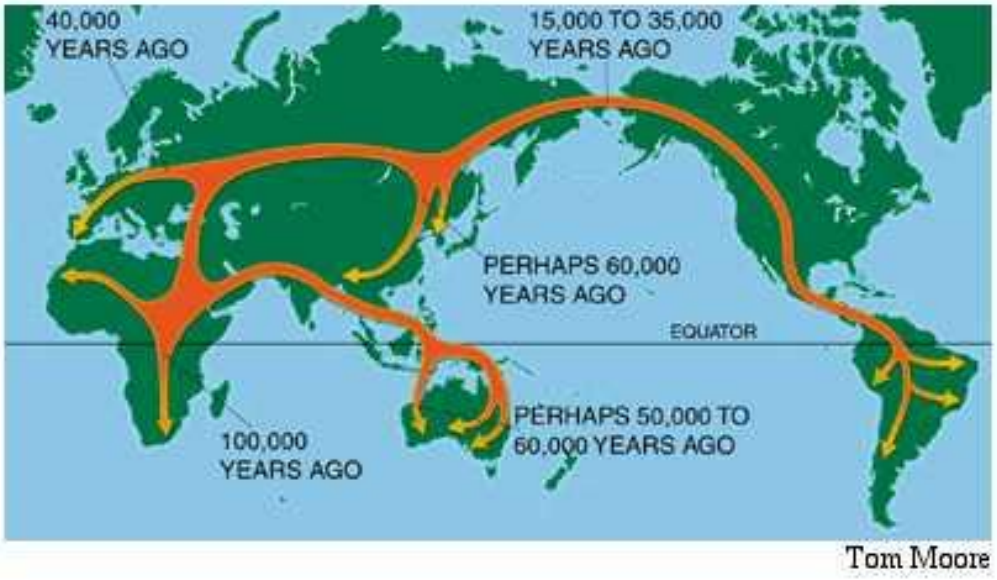
https://www.nature.com/articles/s41397-019-0096-y



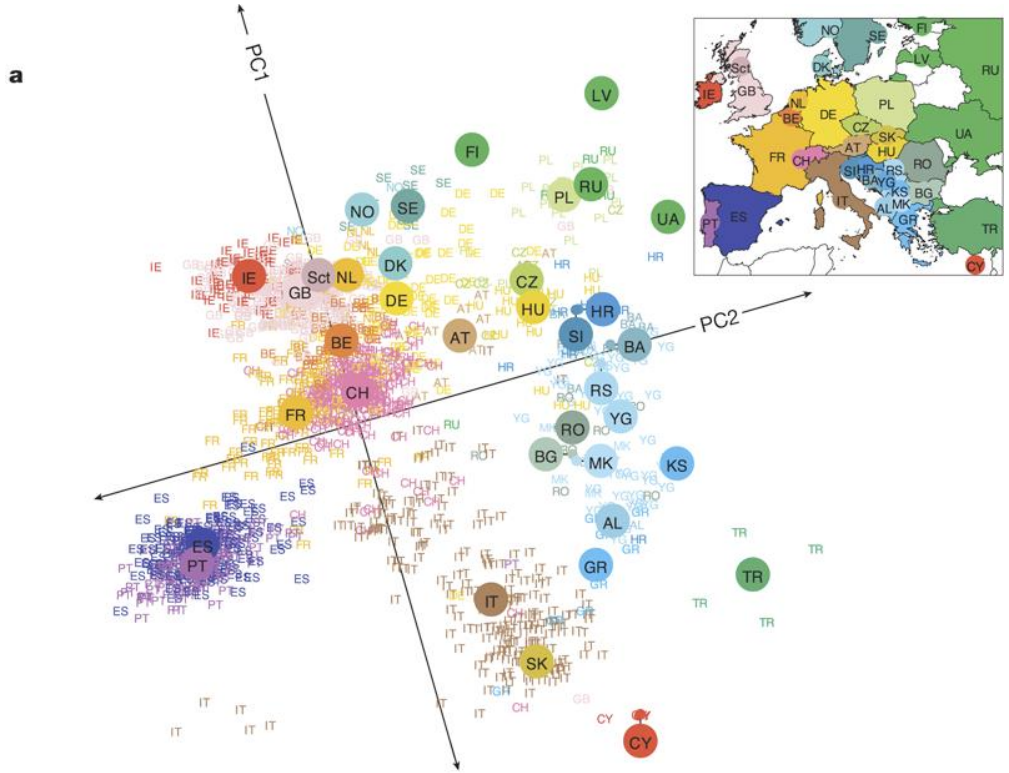
https://www.nature.com/articles/s41397-019-0096-y

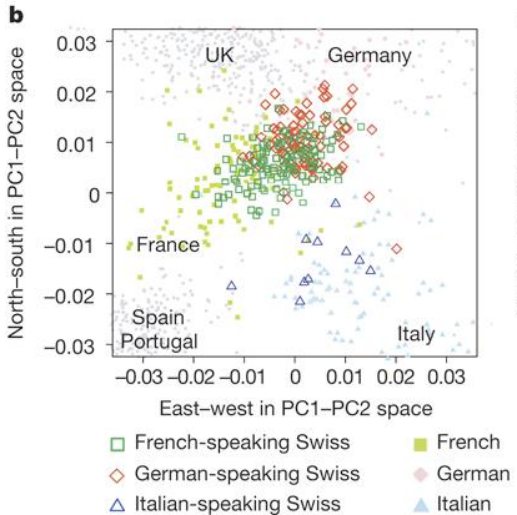
Our genes come from the migration patterns of haplotypes throughout human history ("Population Stratification")

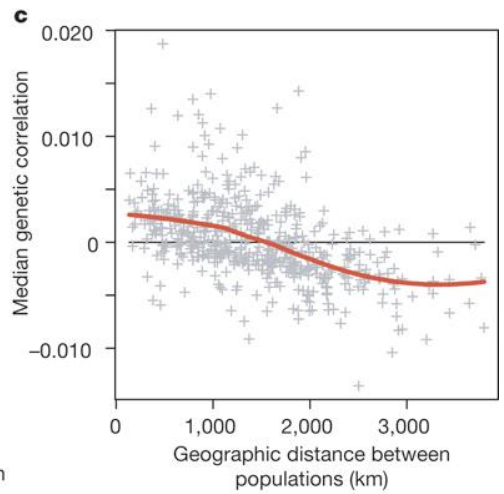


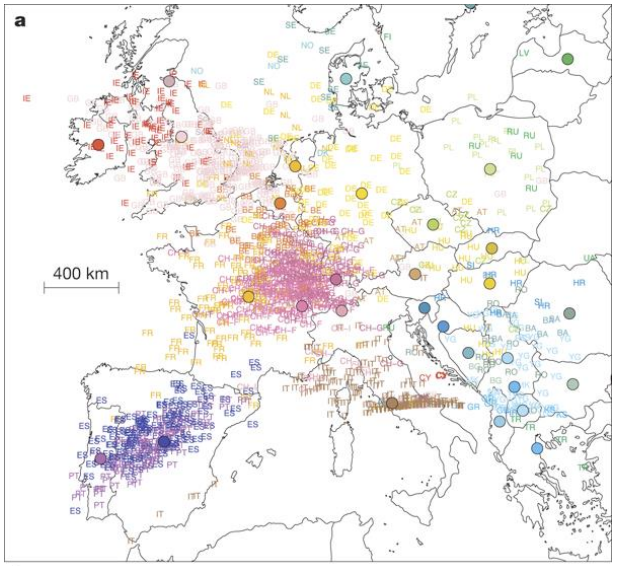


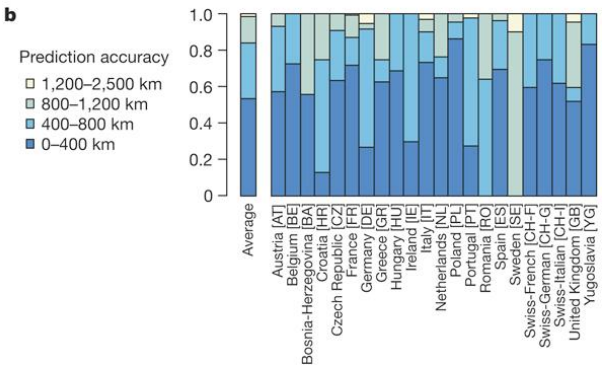
Genotype data can even predict your birthplace











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

Large impact for normal genomes and diseases, especially cancer

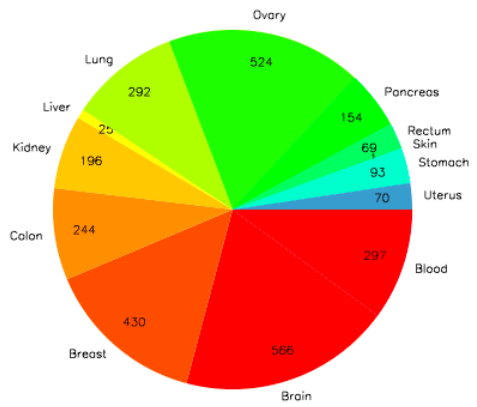
1000 Genomes
A Deep Catalog of Human Genetic Variation

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

ICGC Dataset Version 8 (March 15th, 2012)

Cancer Projects: **29**

Donors by Tissue



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

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

ltered in 66 (48%) of cases.

Total altered 66 cases with alter

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.

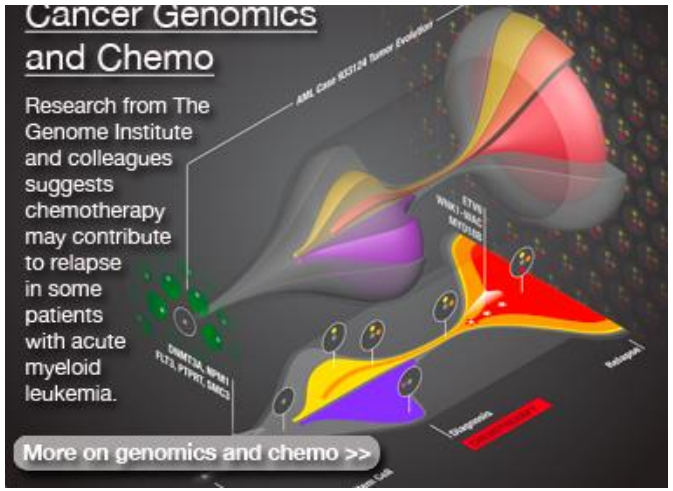
If you have any questions or concerns about this release, contact tcga-dcc-binf-l@list.nih.gov.

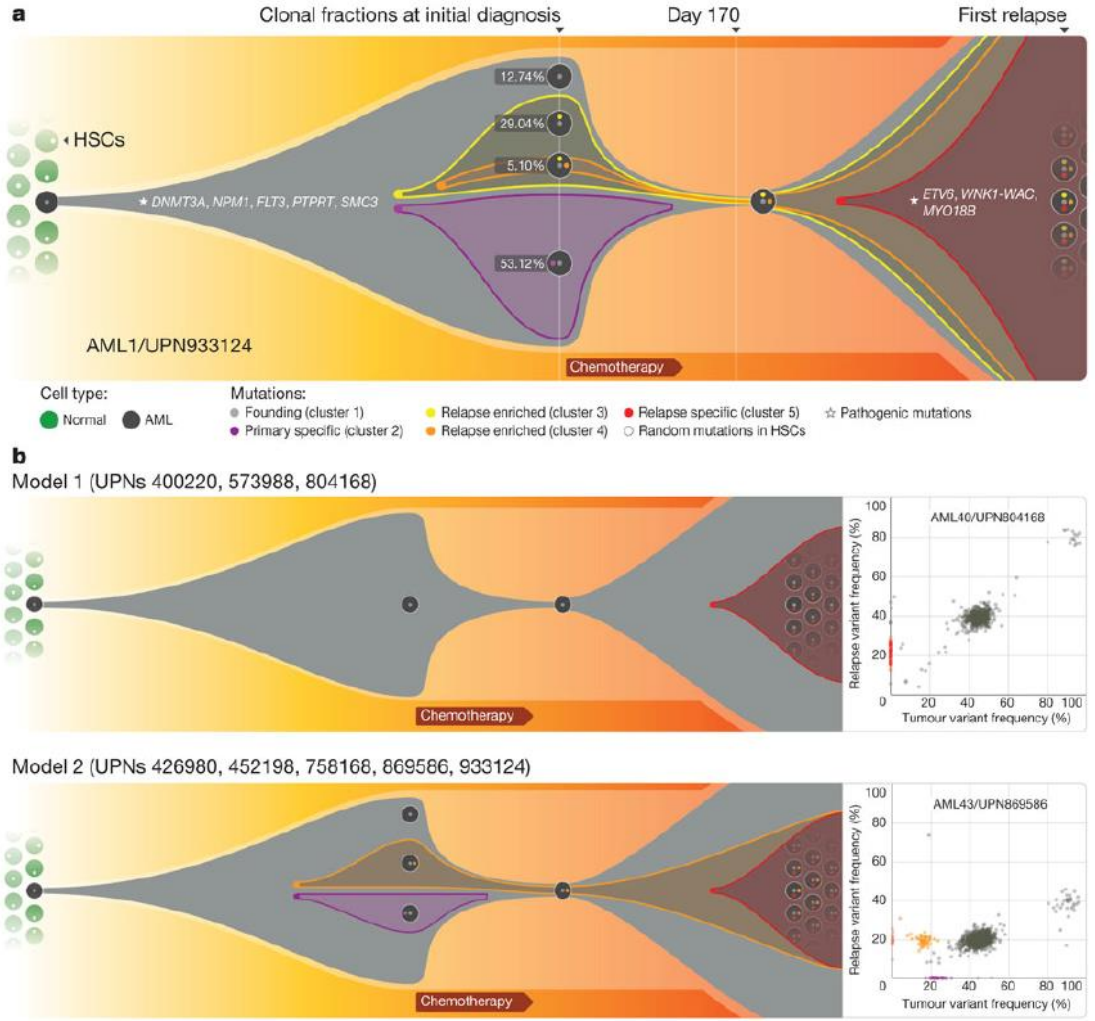
02/25/2013 - DCC Software Released

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





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.

GENOME 10K

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

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

i5k

Plants as well!

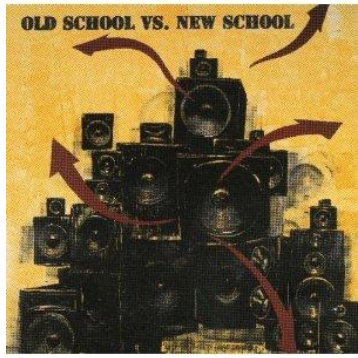
华大基因
BGI

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

The Tech

Sequencing Technologies

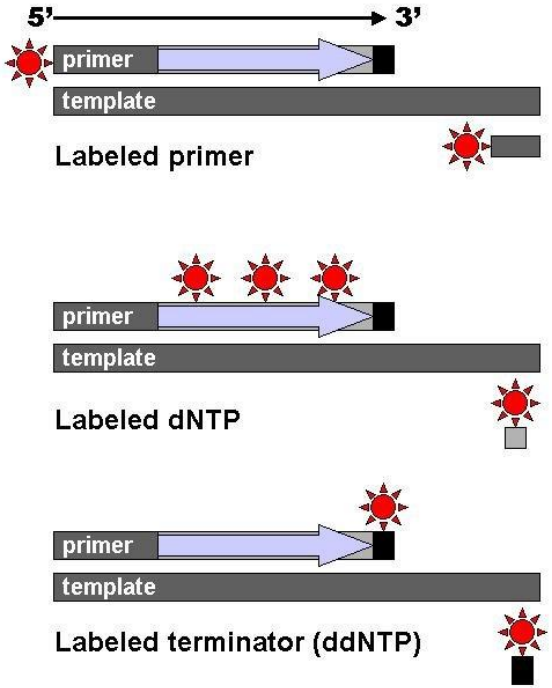
1. "Old School" dye-terminator sequencing (Sanger). 300-1000bp

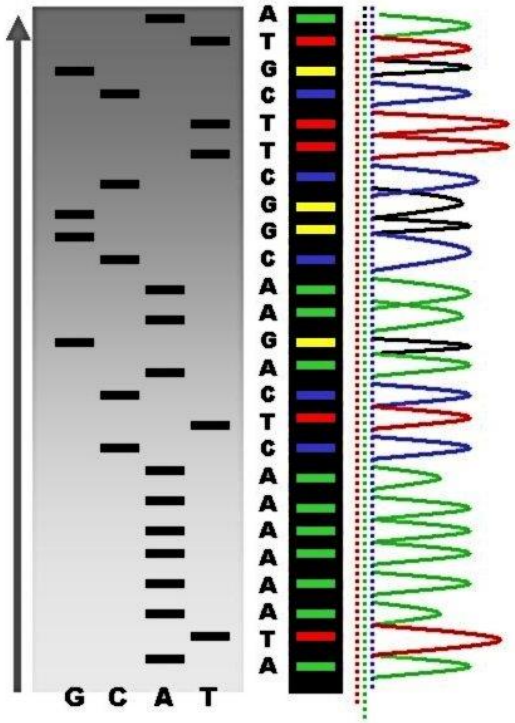


2. "New School" methods
   a. Emulsion PCR Pyrosequencing
   b. Solid-phase amplification sequencing by synthesis (clonal or single molecule)
   c. Sequencing by ligation
   d. Single-molecule, real-time (SMRT) sequencing
   e. Electrical sequencing

Sequencing Technologies

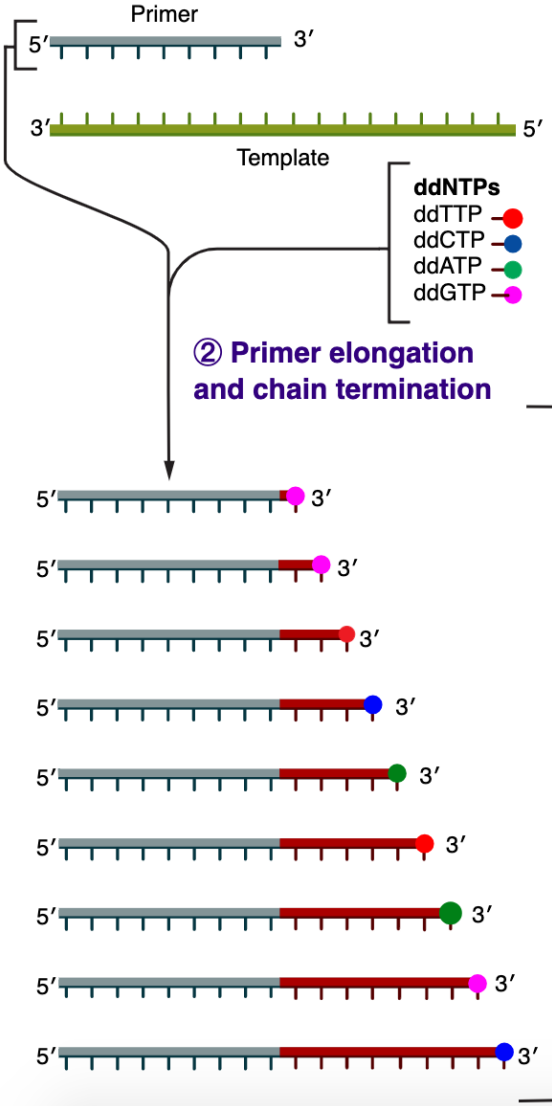
1. "Old School" dye-terminator sequencing (Sanger). 300-1000bp



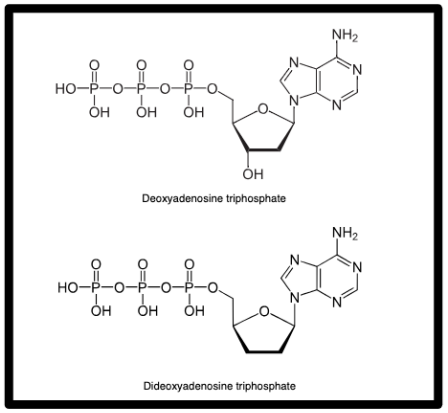


① Reaction mixture

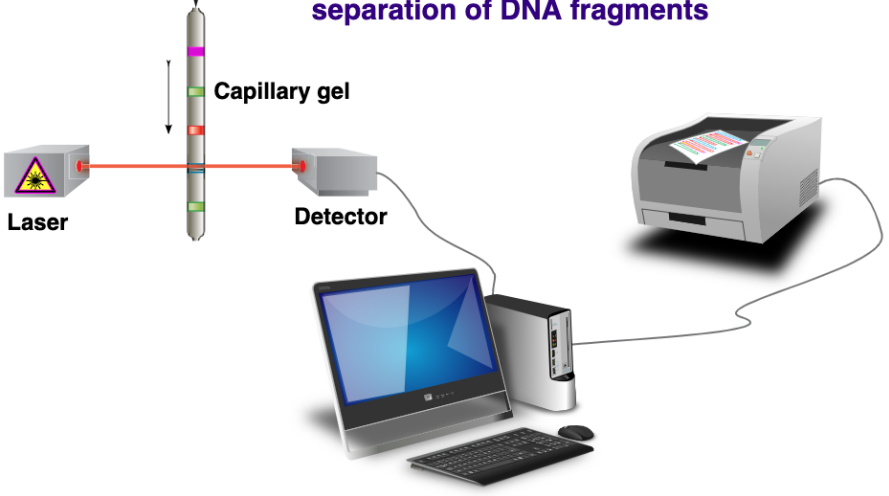
- Primer and DNA template
- DNA polymerase
- ddNTPs with flourochromes
- dNTPs (dATP, dCTP, dGTP, and dTTP)

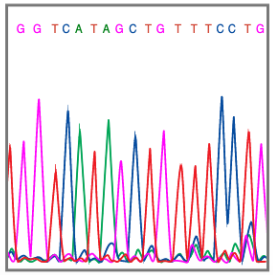


② Primer elongation and chain termination



③ Capillary gel electrophoresis separation of DNA fragments





Chromatograph

④ Laser detection of flourochromes and computational sequence analysis

Estevezj

By 2009, many MPS 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**	**Averege 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 Labeling	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**	**Averege 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

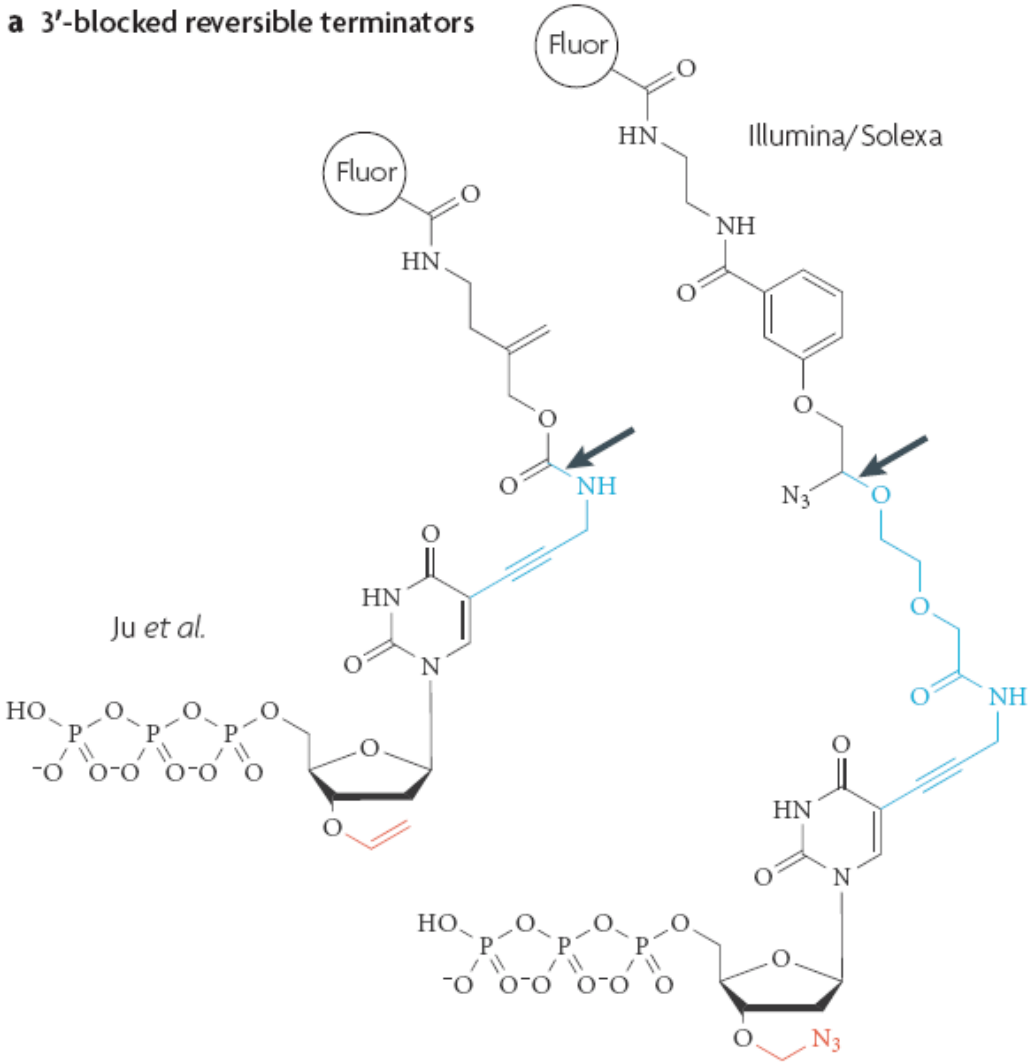
Coming of age: ten years of next-generation sequencing technologies

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.

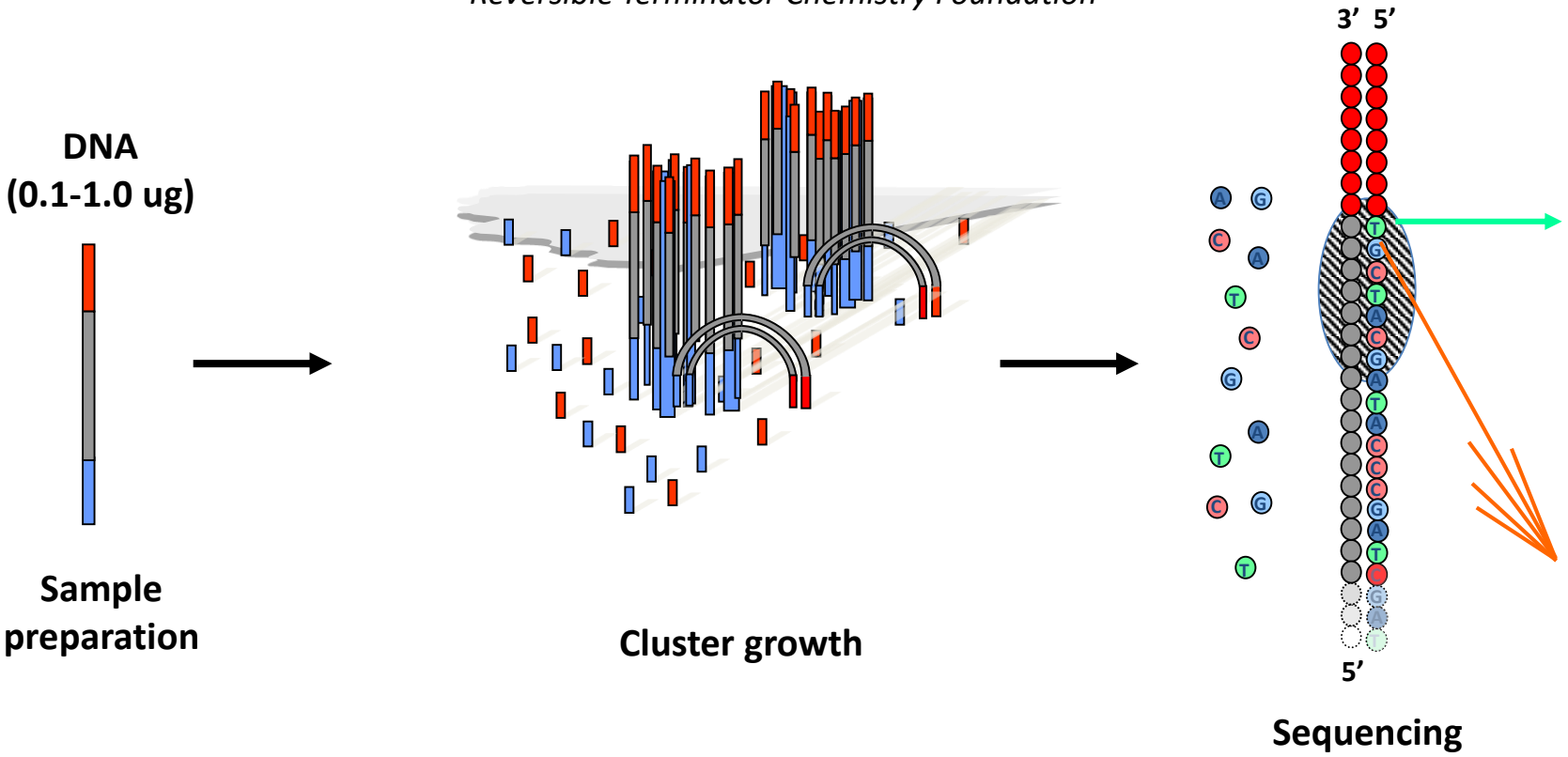
Consideration of each platform

2. Reversible Terminator Bases are Essential Technology Used in Many Chemistries

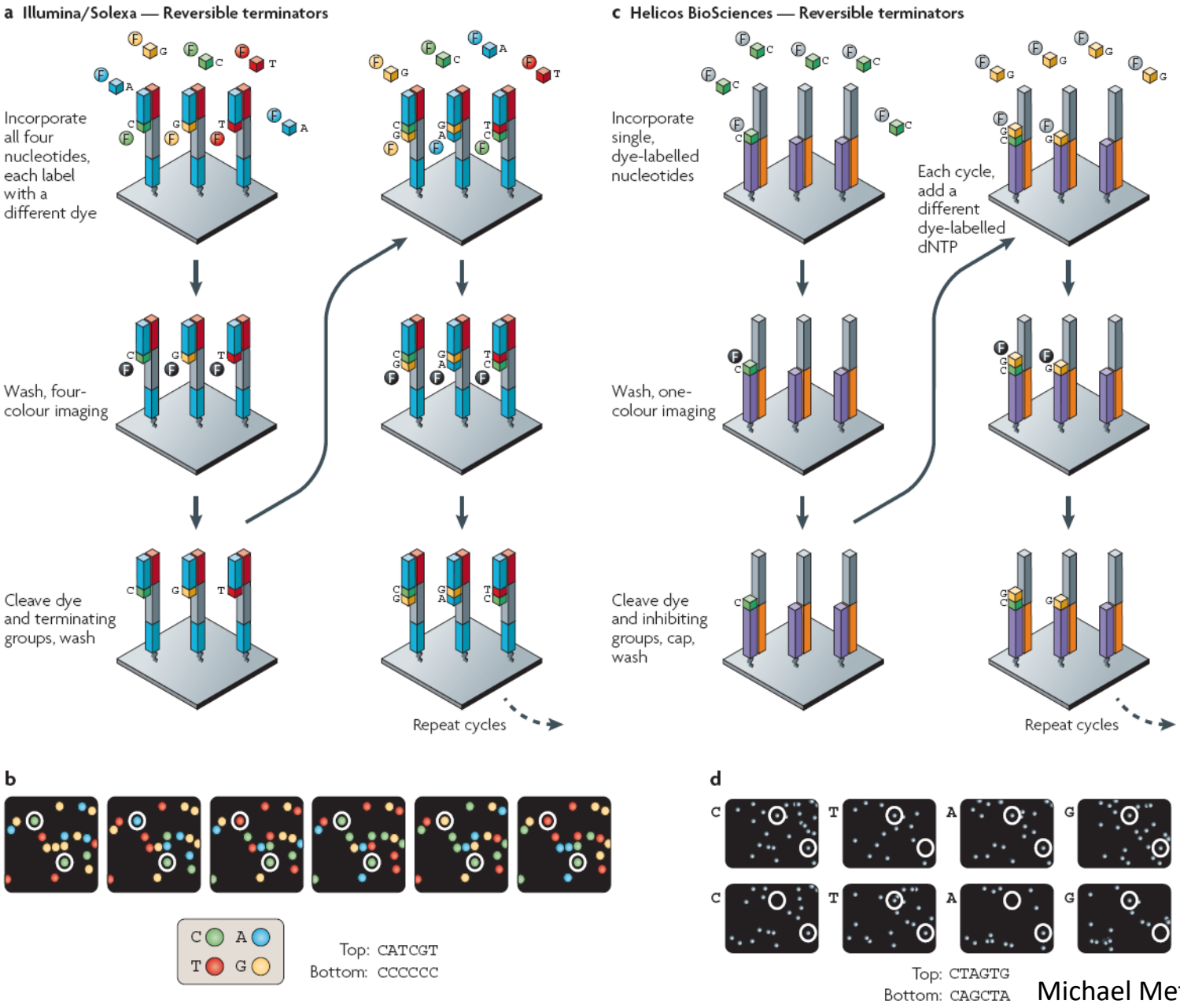


Illumina SBS Technology

Reversible Terminator Chemistry Foundation



Sequencing by Synthesis (SBS)



Michael Metzker, 2010

Now three kinds of chemistry

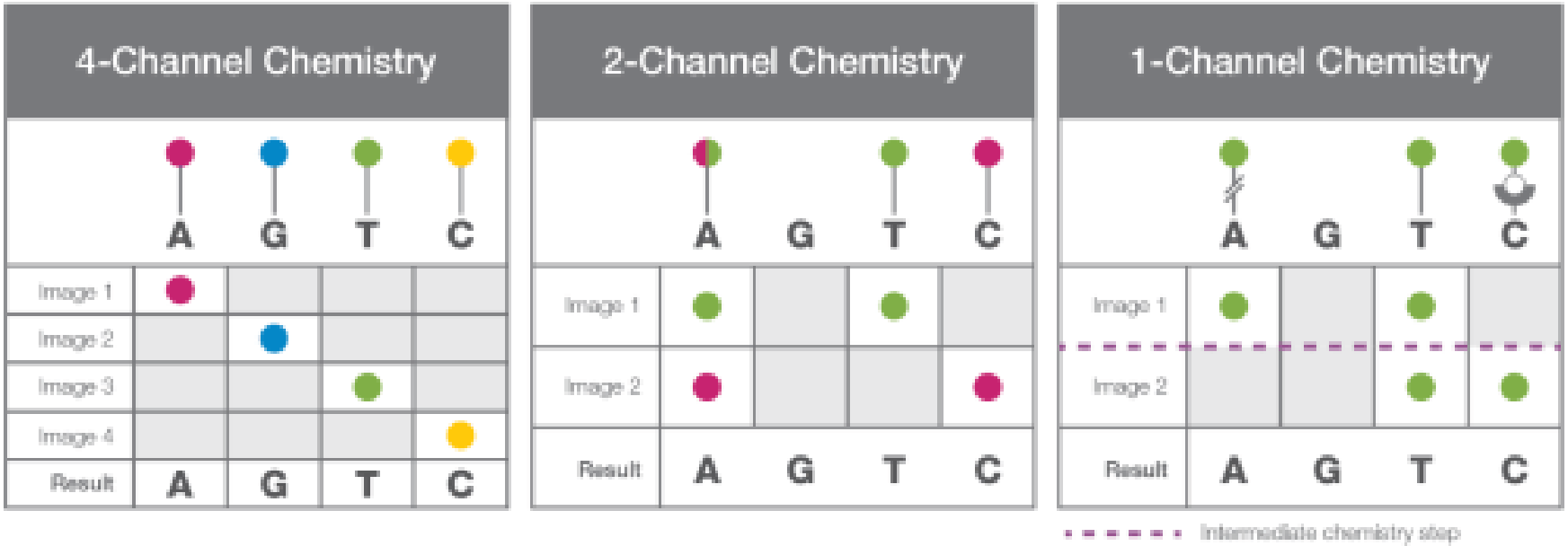
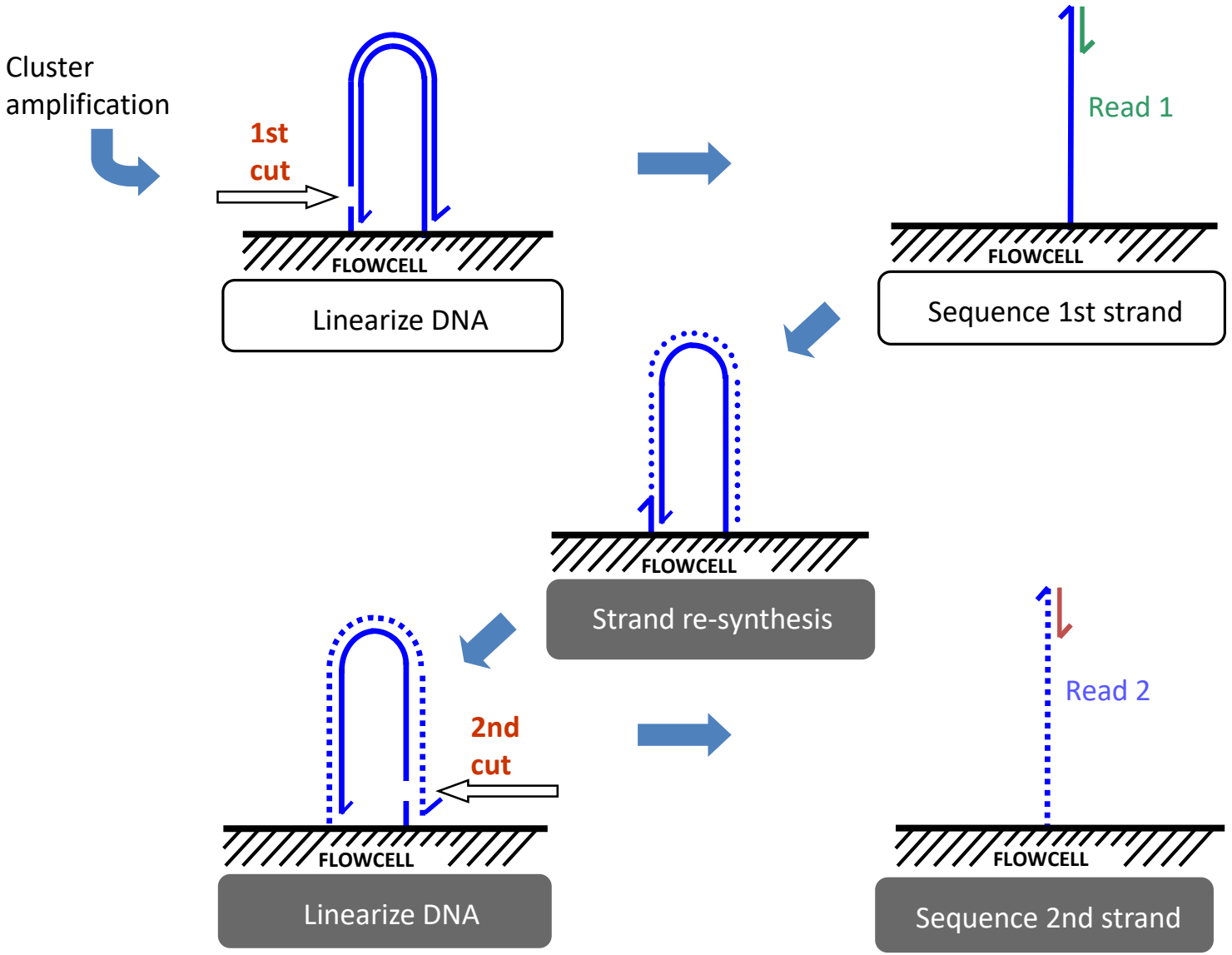
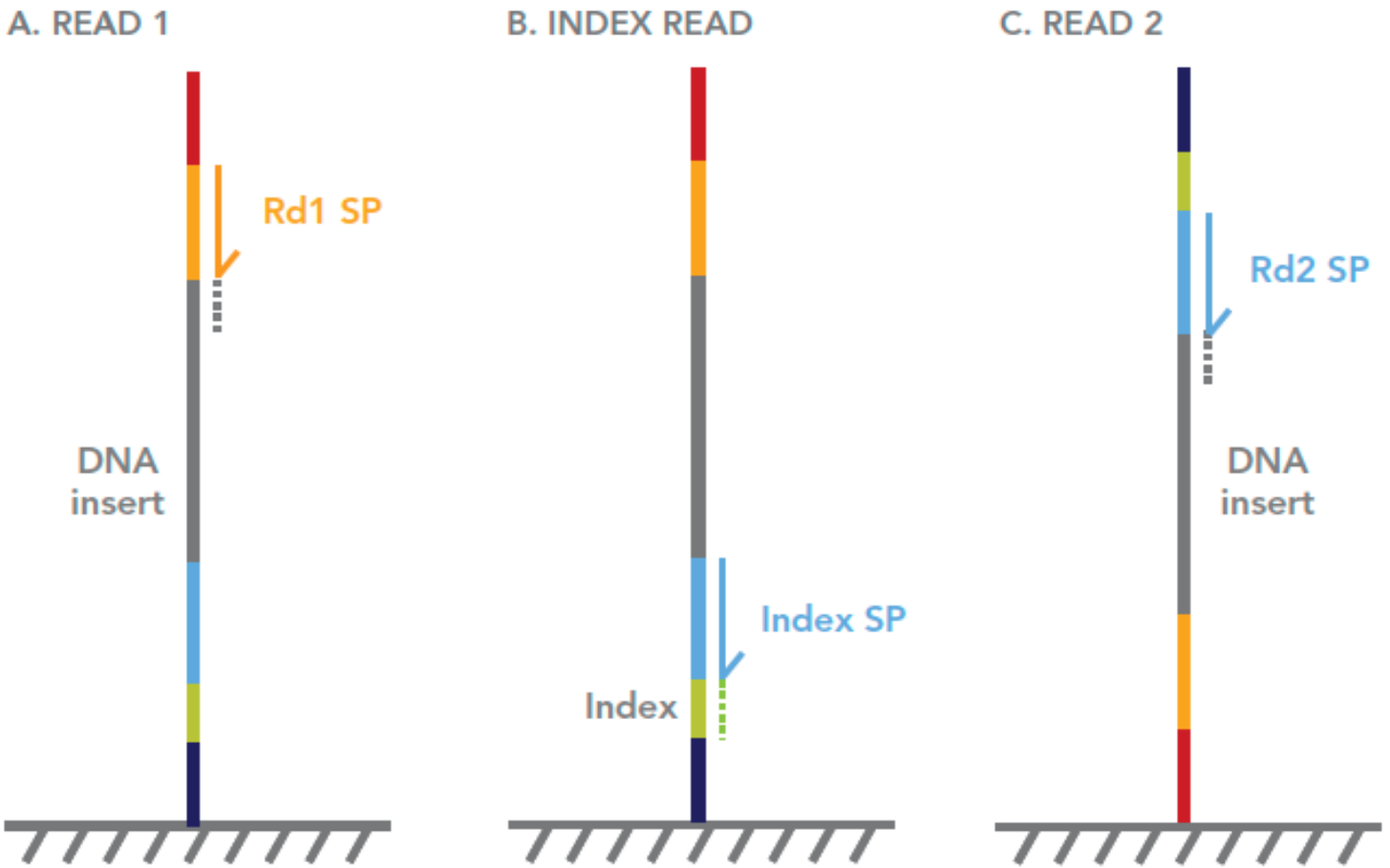


Figure 2: Four-, Two-, and One-Channel Chemistry—Four-channel chemistry uses a mixture of nucleotides labeled with four different fluorescent dyes. Two-channel chemistry uses two different fluorescent dyes, and one-channel chemistry uses only one dye. The images are processed by image analysis software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

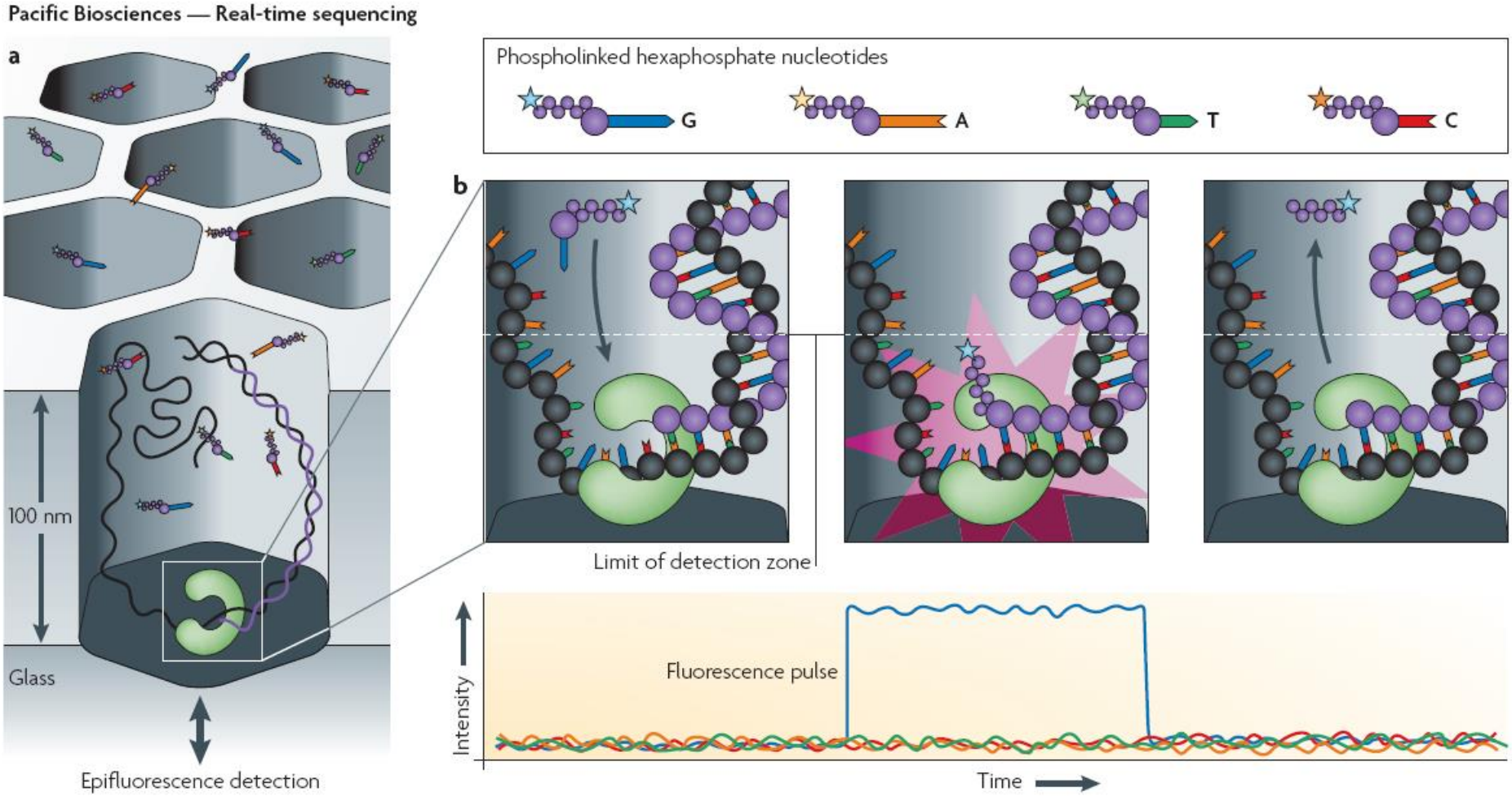
Paired-End Sequencing allows for two looks at a sequence



Indexed sequencing method is now standard for single and paired reads



Pacific Biosciences
Single Molecule Real-Time (SMRT) Sequencing



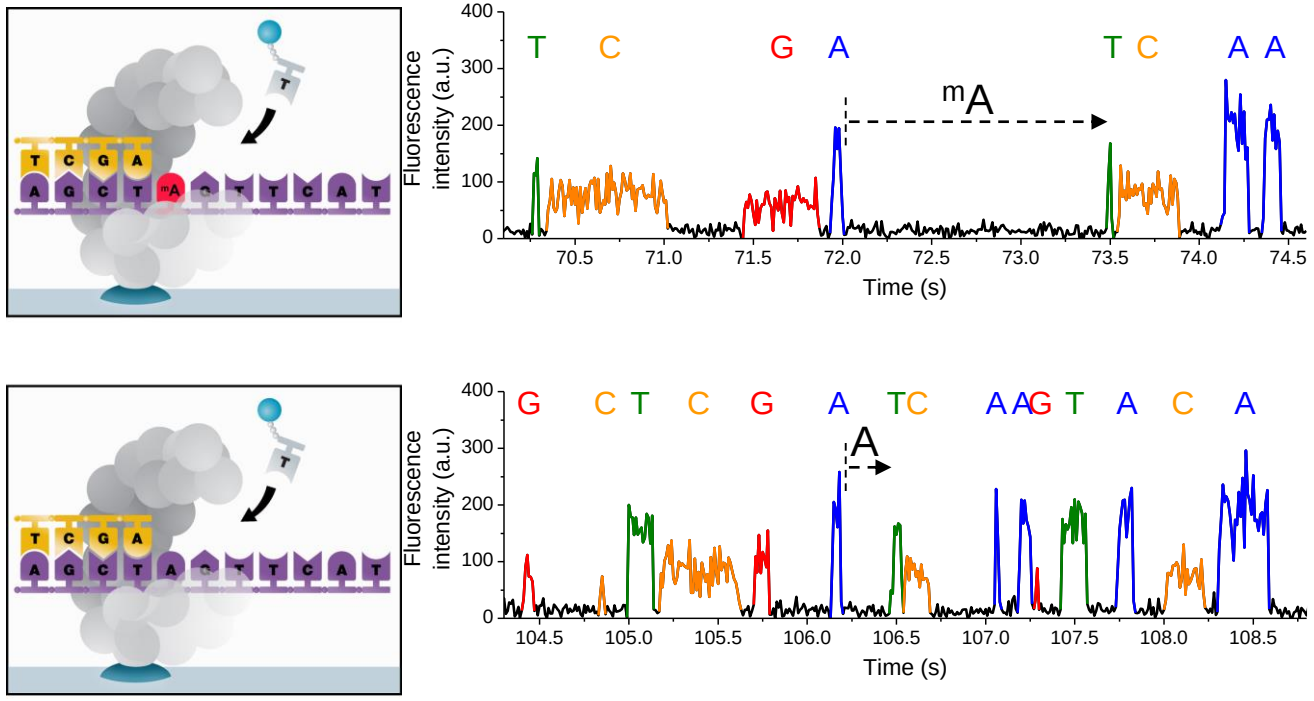
https://www.pacb.com/videos/video-overview-of-smrt-technology/

Metzker, 2010

Single Molecule Kinetics Allow for the Direct Detection of Methylation

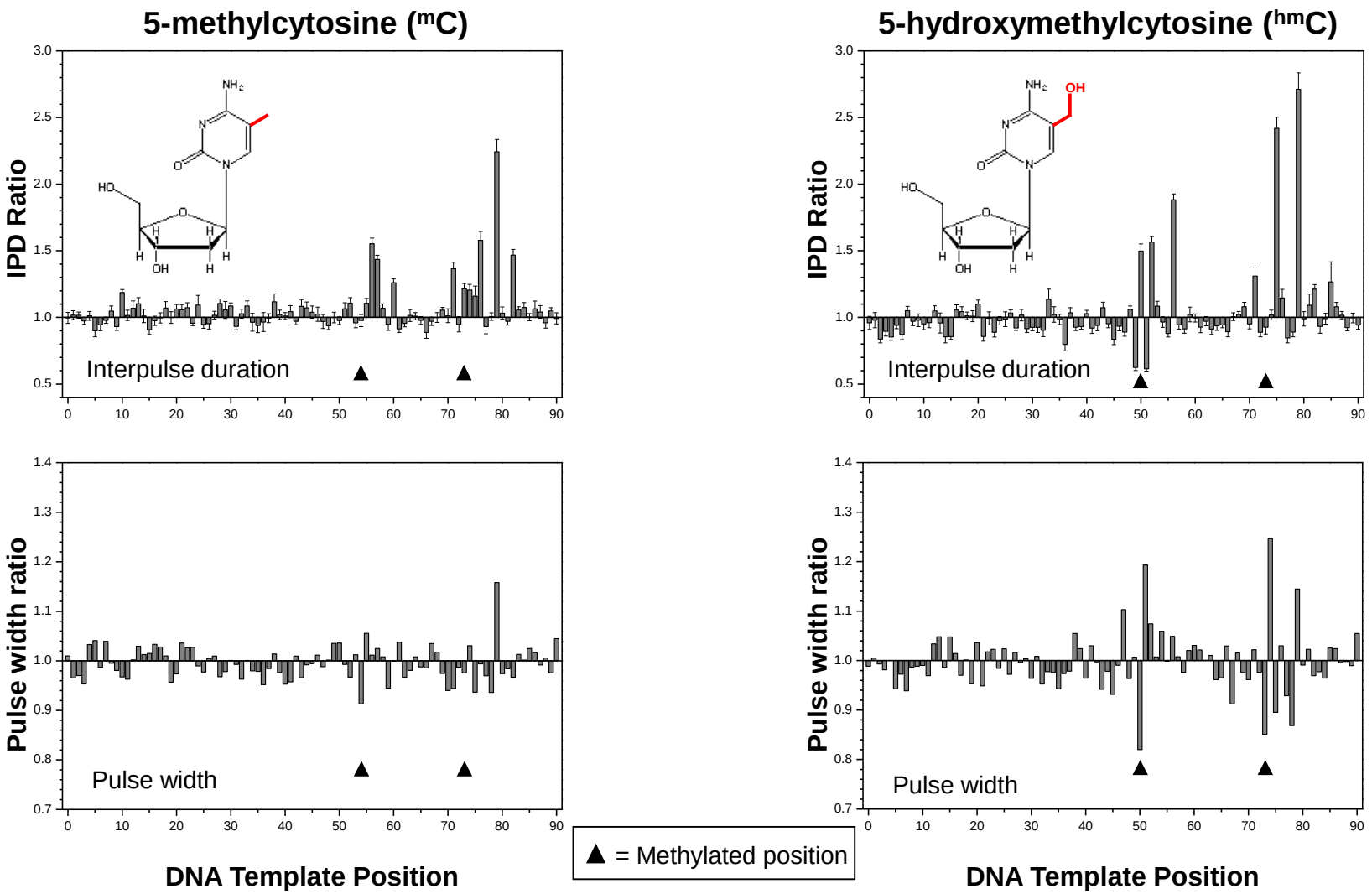
Approach: Kinetic detection of methylated bases during SMRT DNA sequencing

Example: N^6-methyladenosine (^mA)

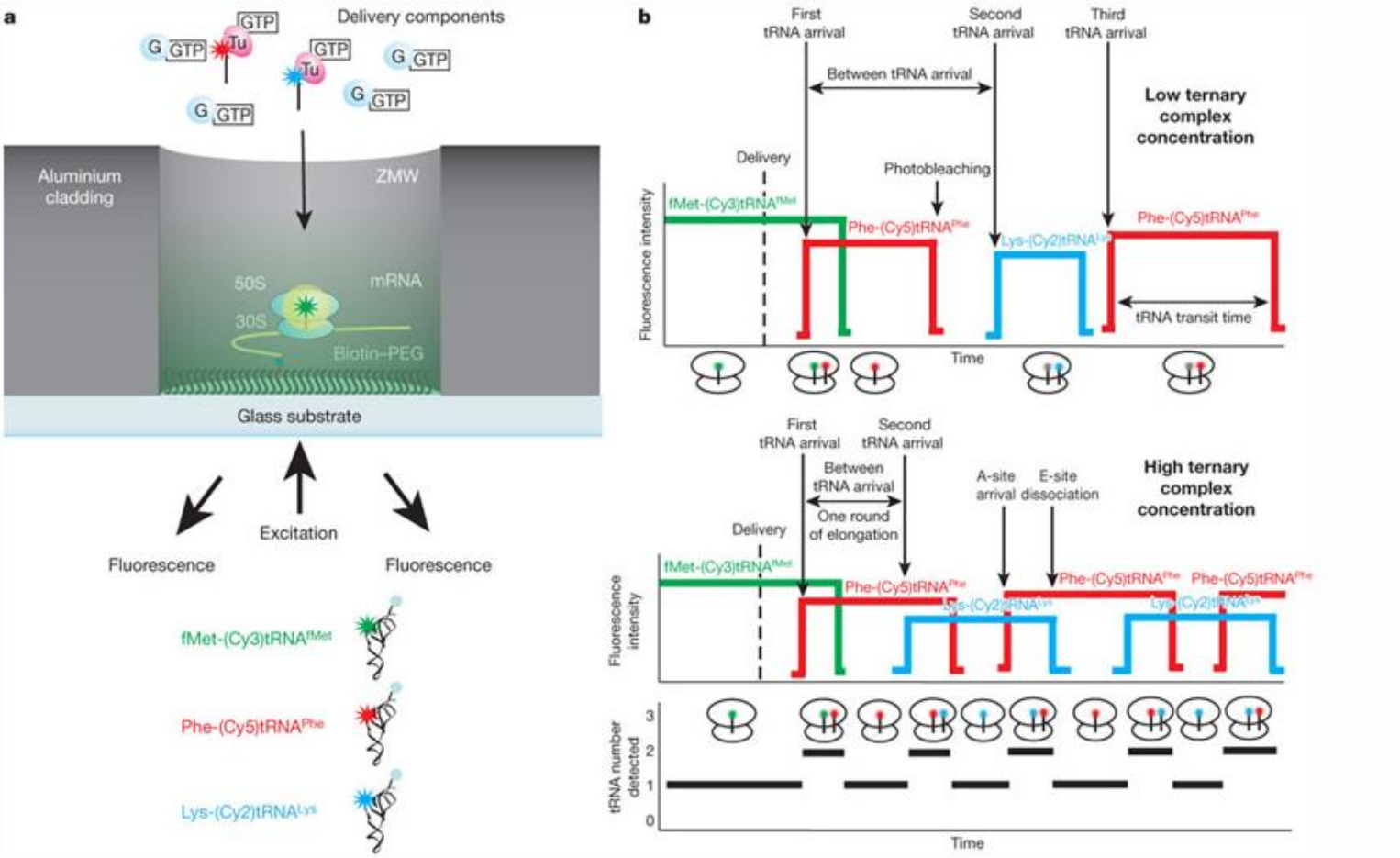


Flusberg et al., 2010.

Kinetics can detect other base modifications



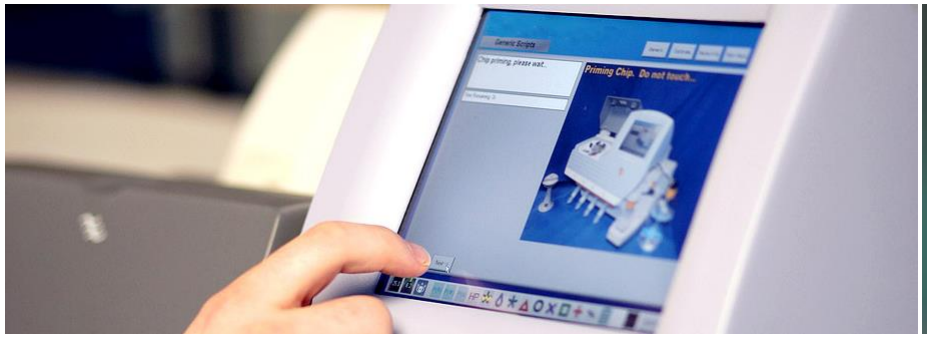
Kinetics allow one to watch protein translation as it occurs

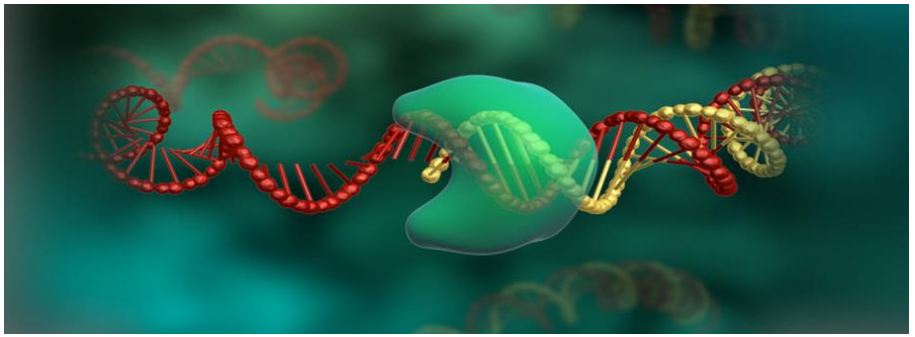


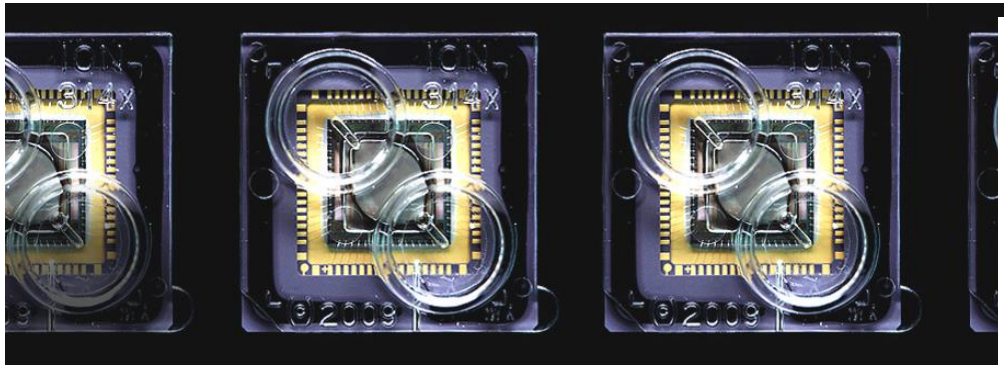
Uemura et al., 2010

“Post-Light,”
Semi-Conductor Sequencing:

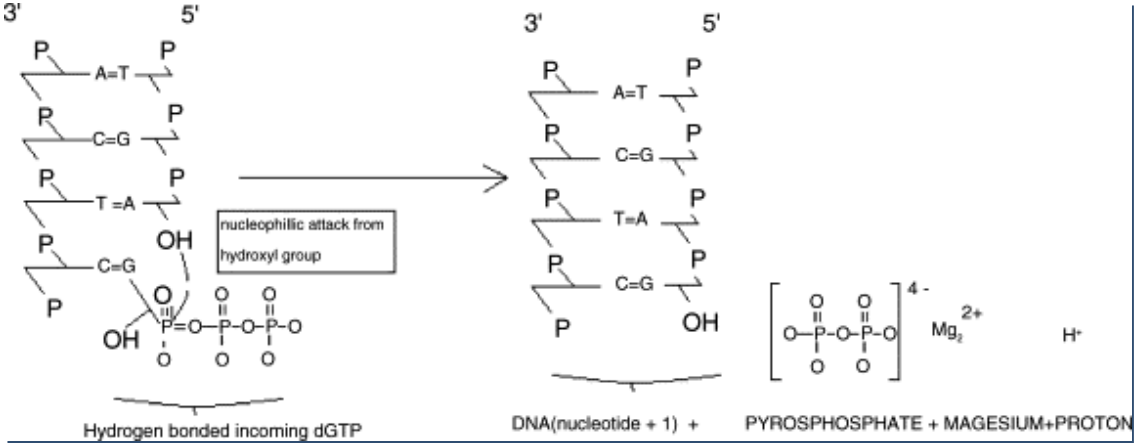
Thermo Fisher’s Personal Genome Machine (PGM), the Proton I and Proton II, and S5







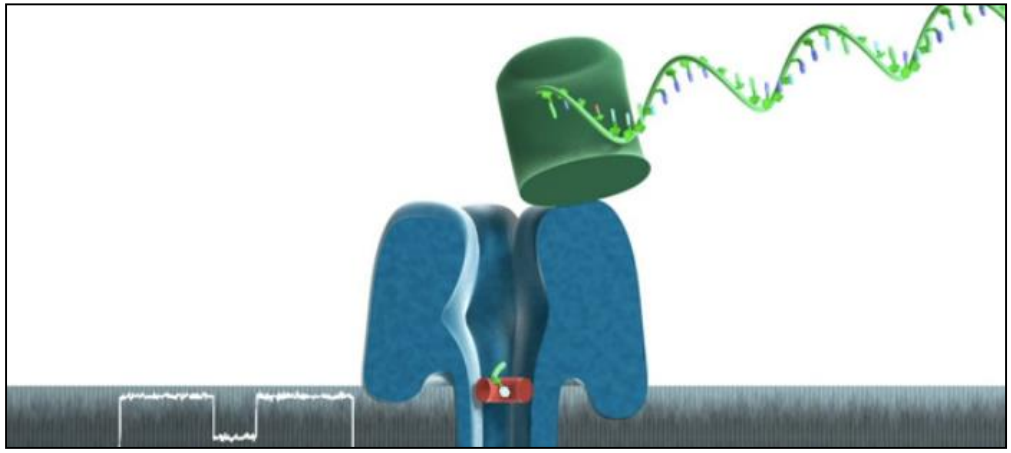
Essentially, Millions of very small pH meters



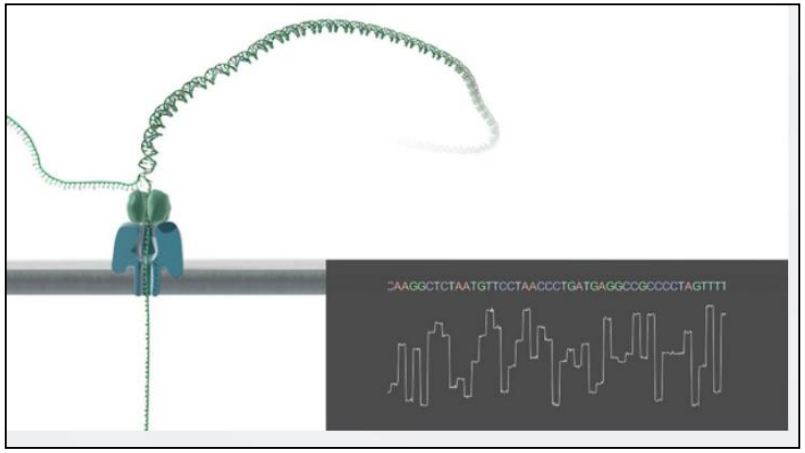
Purushothaman *et al*, 2005
IonTorrent, Inc.

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

2014:Sequencing with a protein nanopore

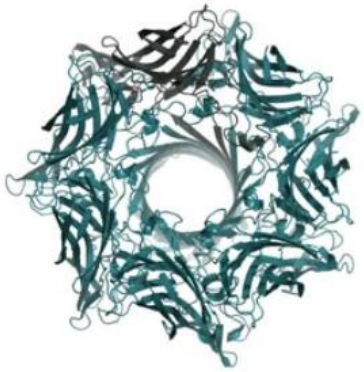


Exonuclease-Seq

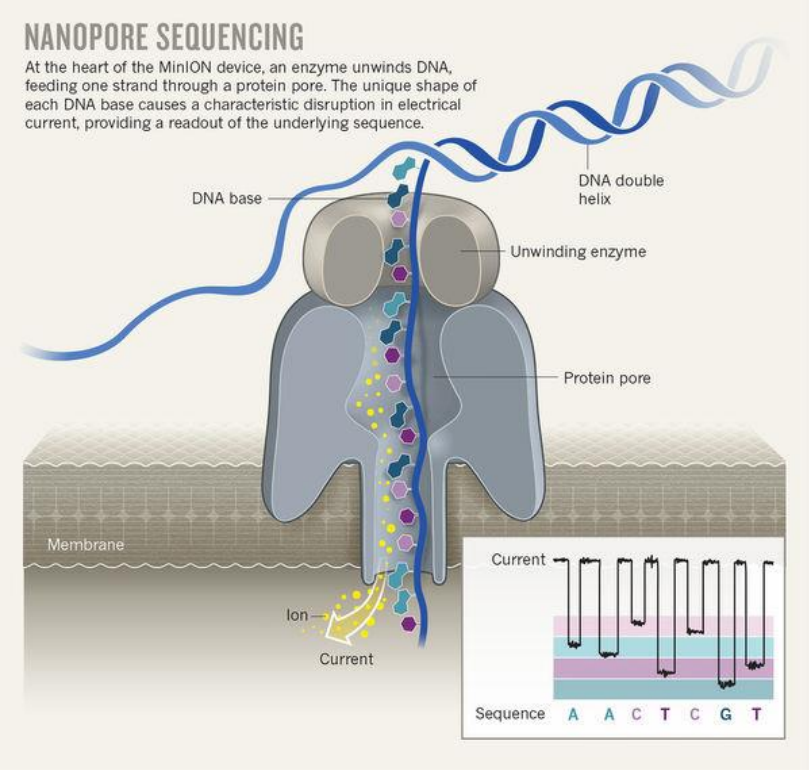


Strand-Seq

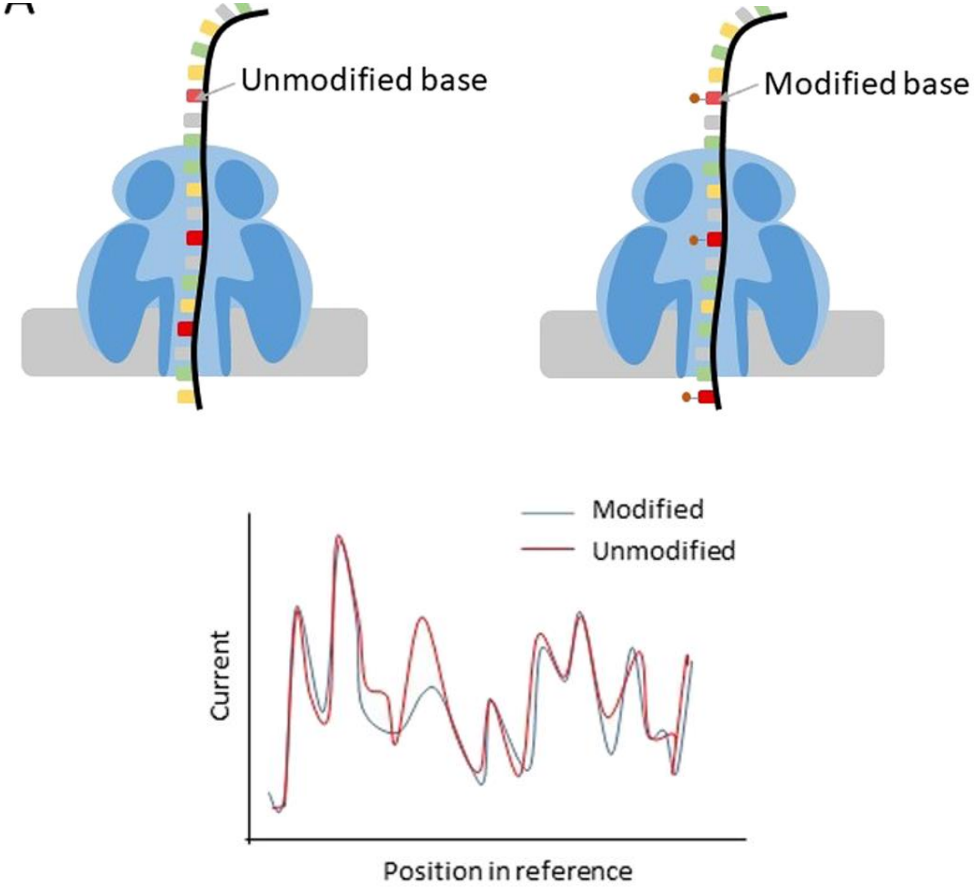
MinION



PromethION



http://blogs.nature.com/naturejobs/2017/10/16/techblog-the-nanopore-toolbox/



2022

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.

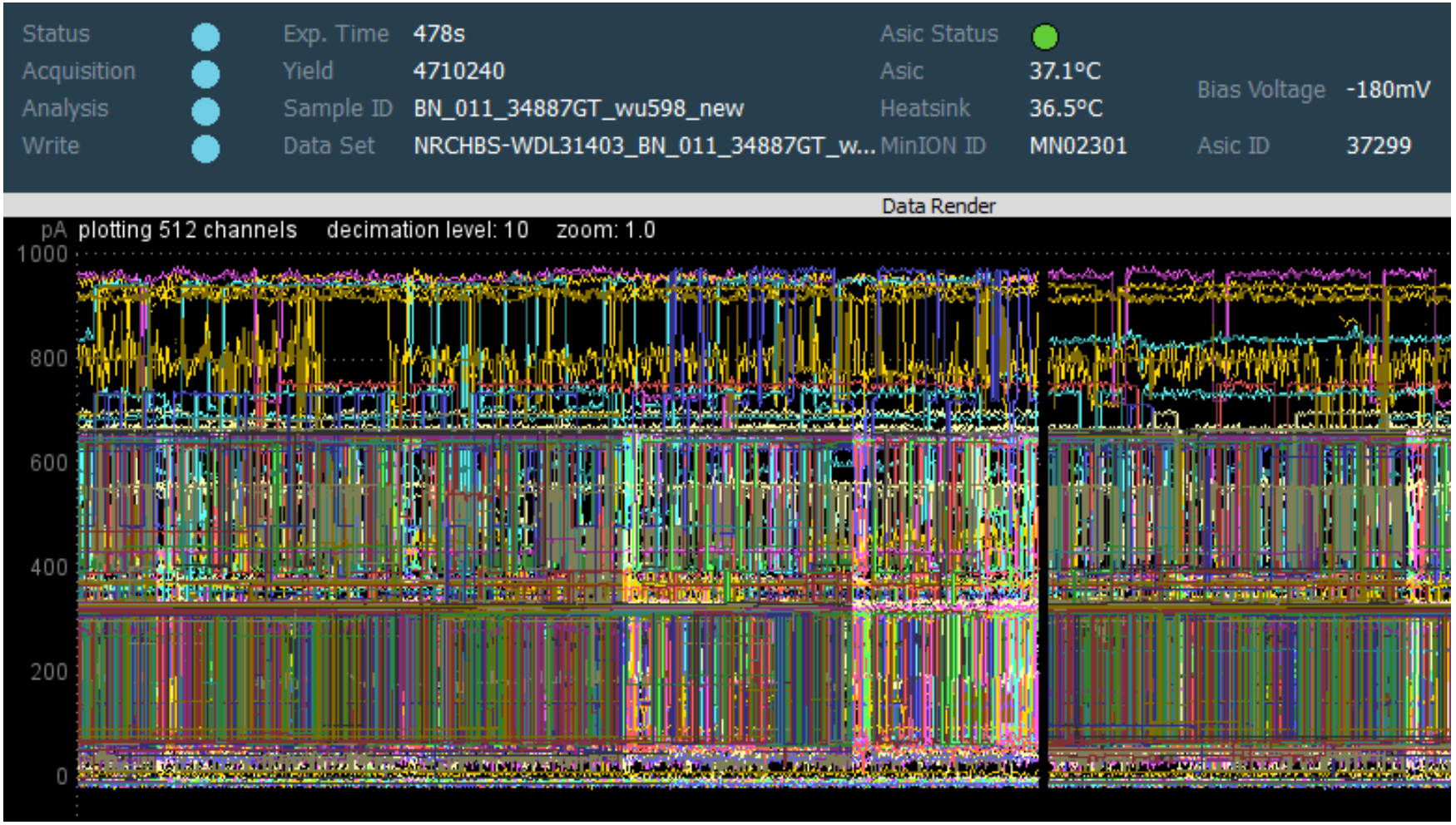
Q-Line ›

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

https://nanoporetech.com/

They are small

Base space is now “squiggle space”



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

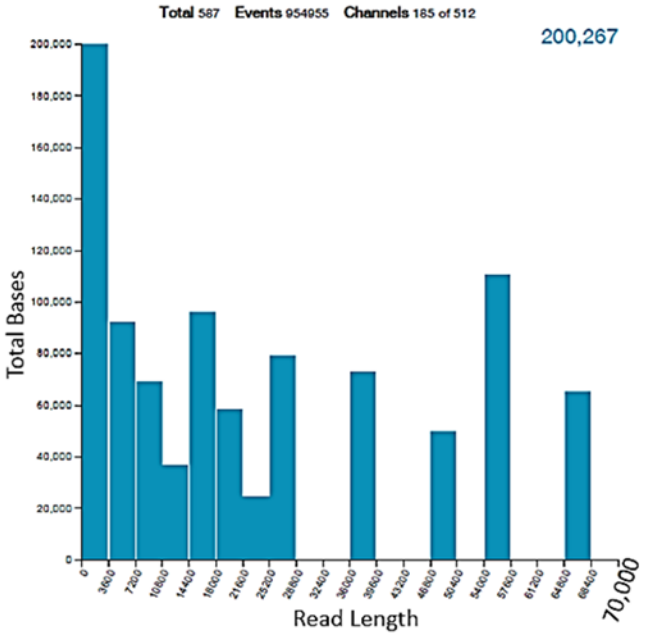
https://www.nature.com/articles/nature16996

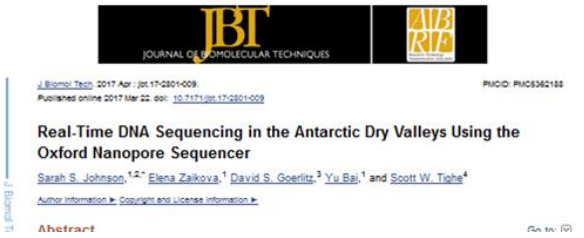

N403IL
NATIONAL SCIENCE FOUNDATION


Scott Tighe

Lake Fryxell, Antarctica
Scott Tighe

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





JBT JOURNAL OF BIOMOLECULAR TECHNIQUES

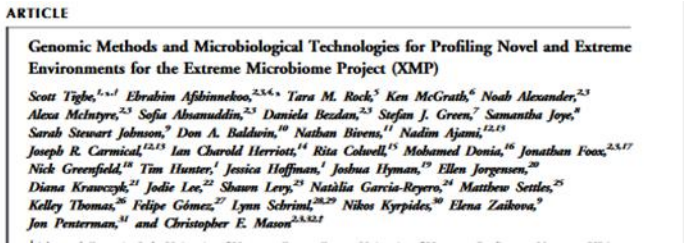
J Biomol Tech. 2017 Apr; jbt.17-2801-009. PMCID: PMC5382188
Published online 2017 Mar 22. doi: 10.7171/jbt.17-2801-009

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



ARTICLE

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

Scott Tighe,[1,2,†] 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

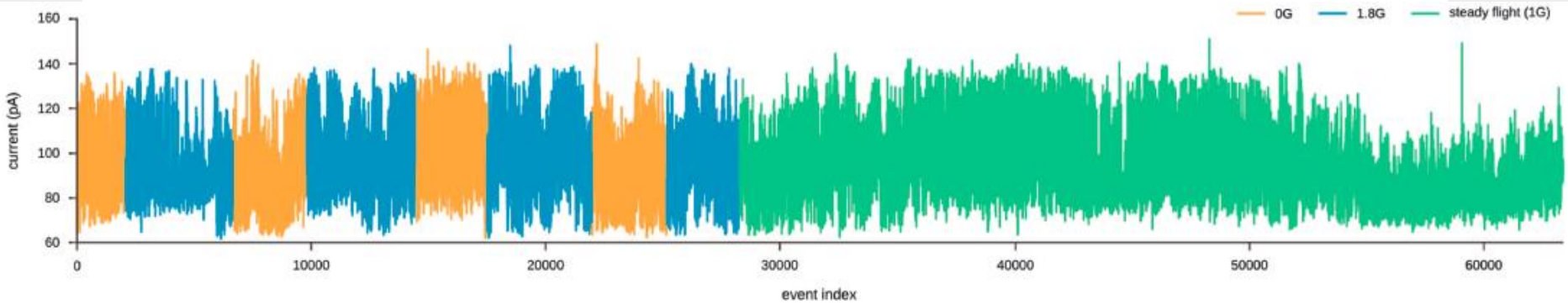
MENU

npj | Microgravity

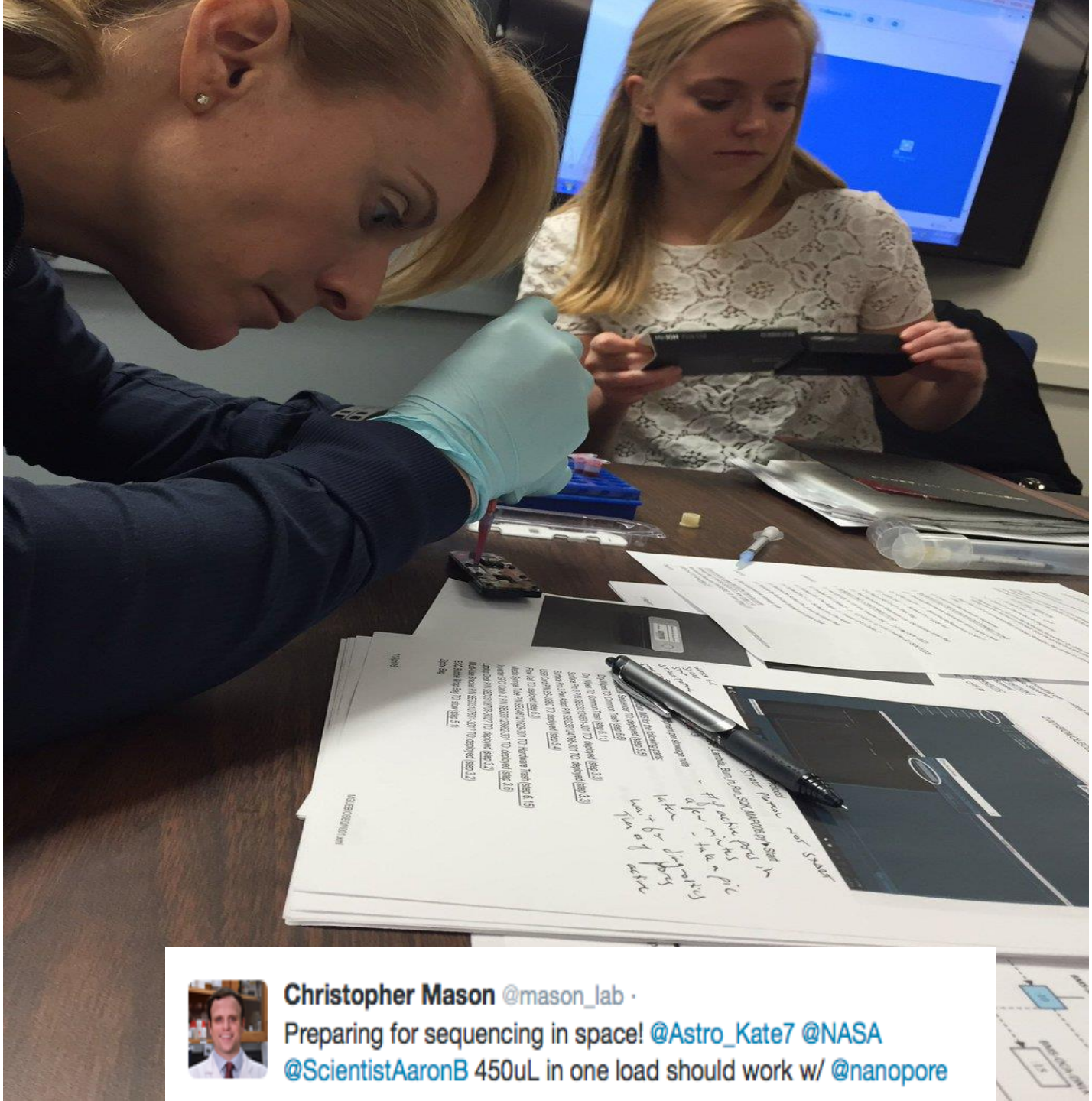
Search E-alert Submit Login

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:

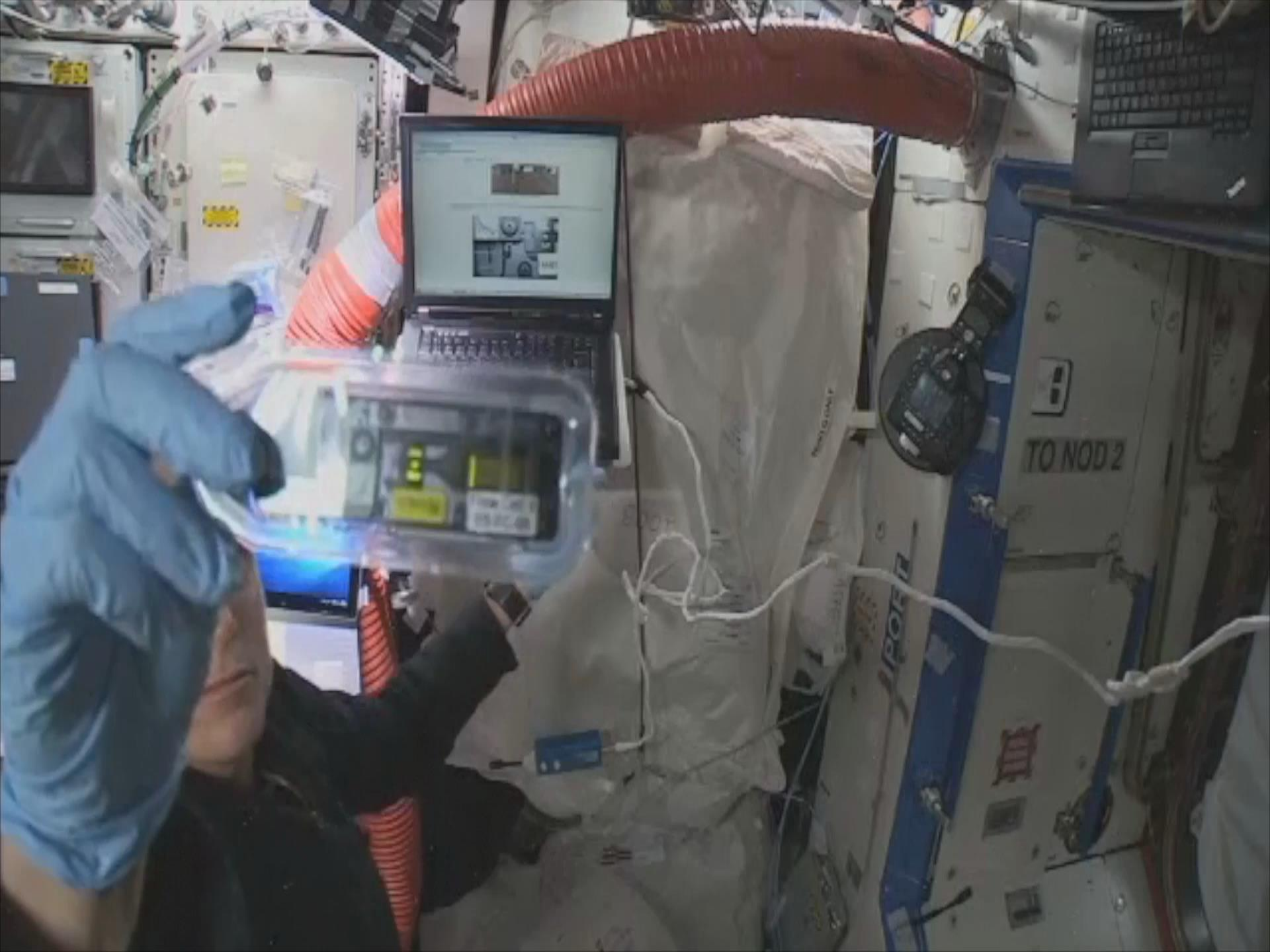
As NASA astronaut Scott Kelley tweeted on Sunday, June 28, 2015, "space is hard."

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.

ВНИМАНИЕ


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


TO NOD 2

Latest | Related


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

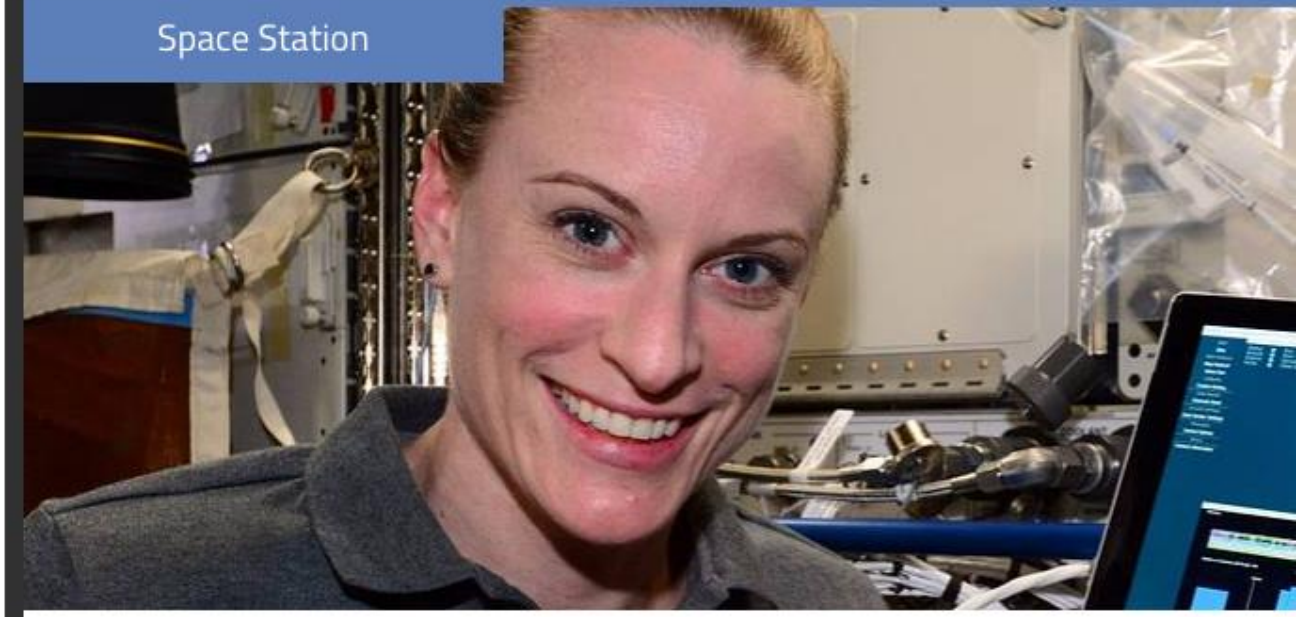

SAGE III to Look Back at Earth's Atmospheric 'Sunscreen'
19 days ago


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

Space Station



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

Following

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

4 12

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

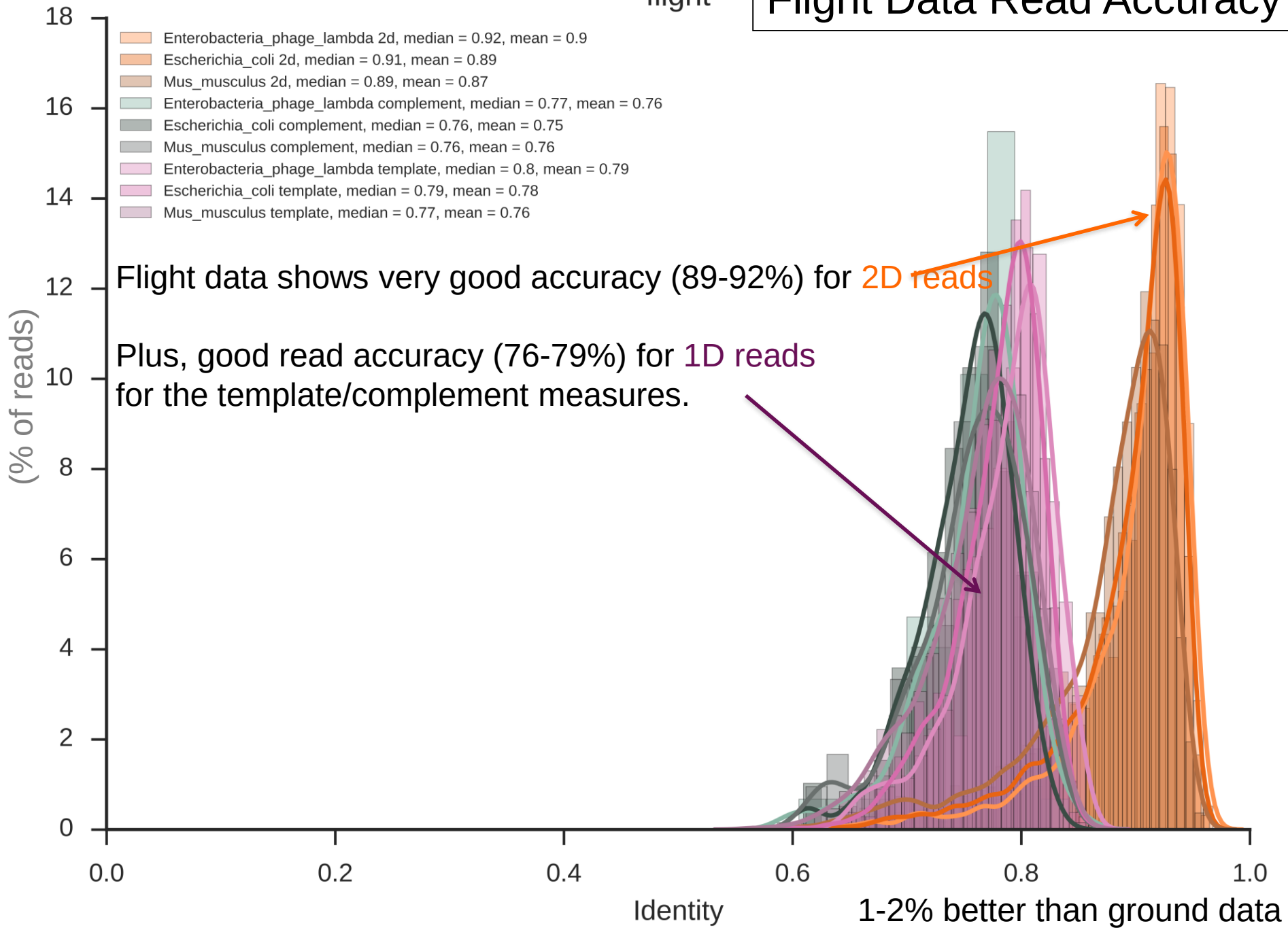
Clip from NASA TV

RETWEETS	LIKES
123	185

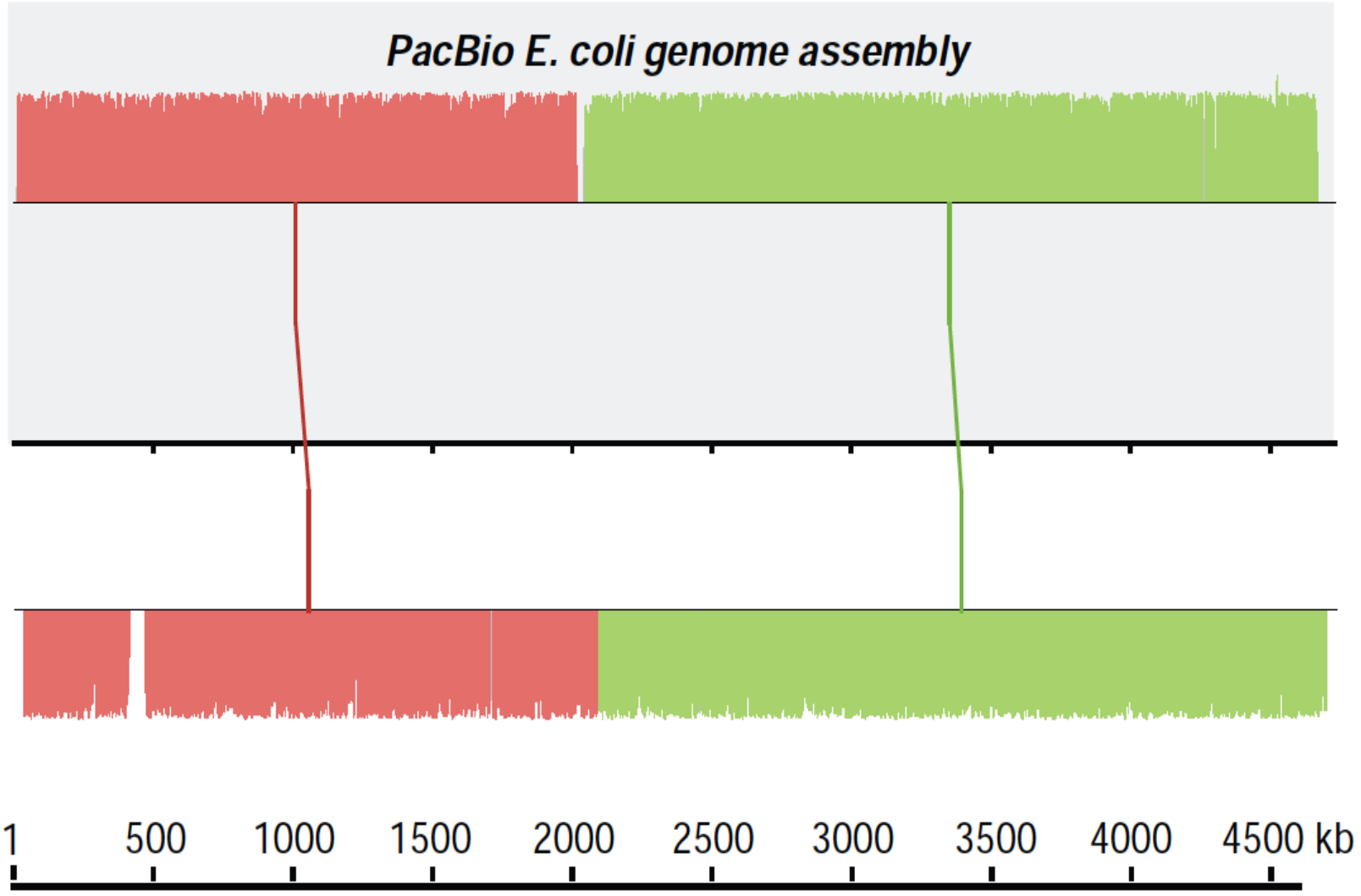
Bus Lon Dor Elai Alfc Oliv Jes Lita

3:28 PM - 14 Sep 2016

Flight Data Read Accuracy



Almost perfect when compared to PacBio



The first genome sequence and assembly 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

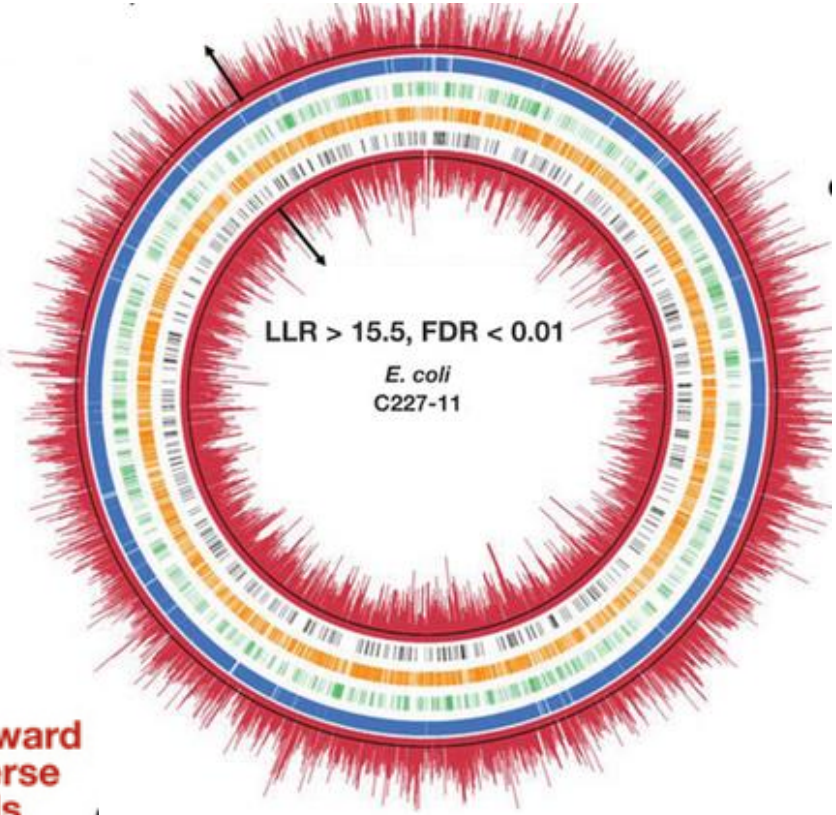
Bacteria are splattered with epigenetic marks

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



LLRs, forward and reverse strands

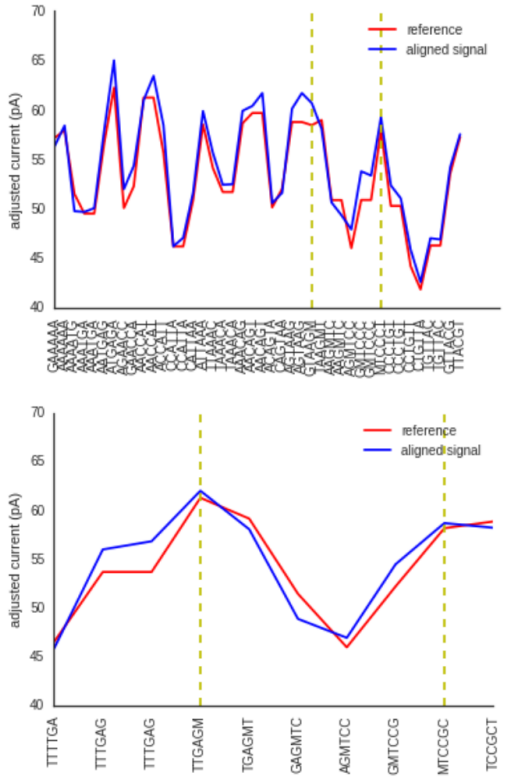
GATC

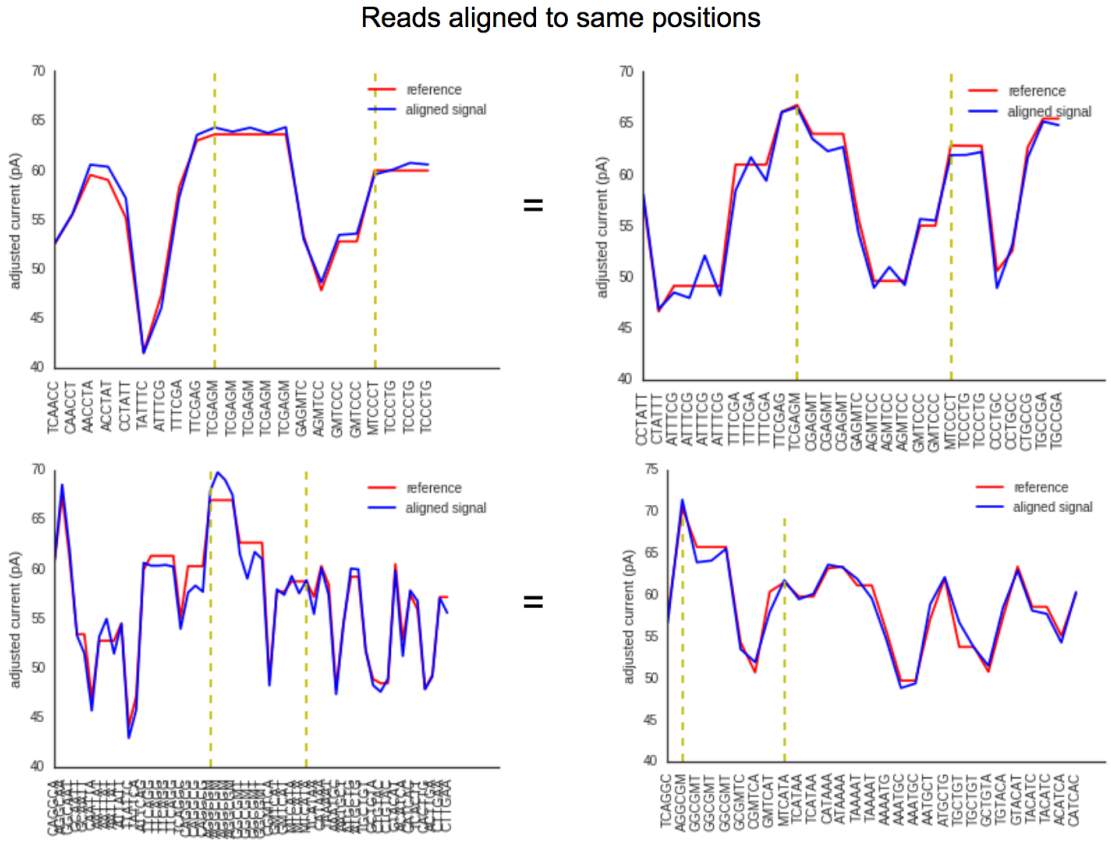
CTGCAG

ACCACC

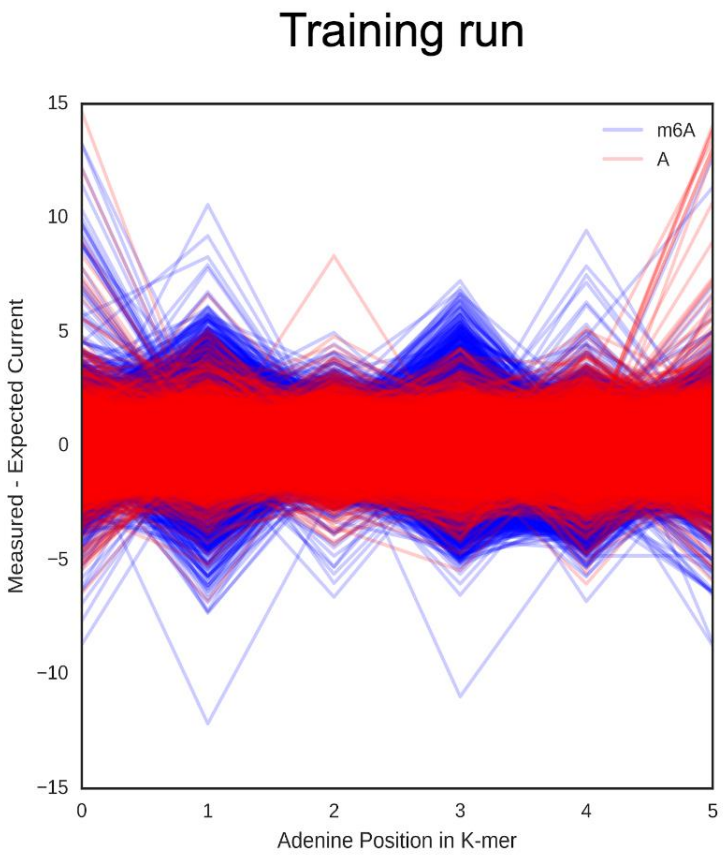
CCACN$_8$TGAY/R
TCAN$_8$GTGG

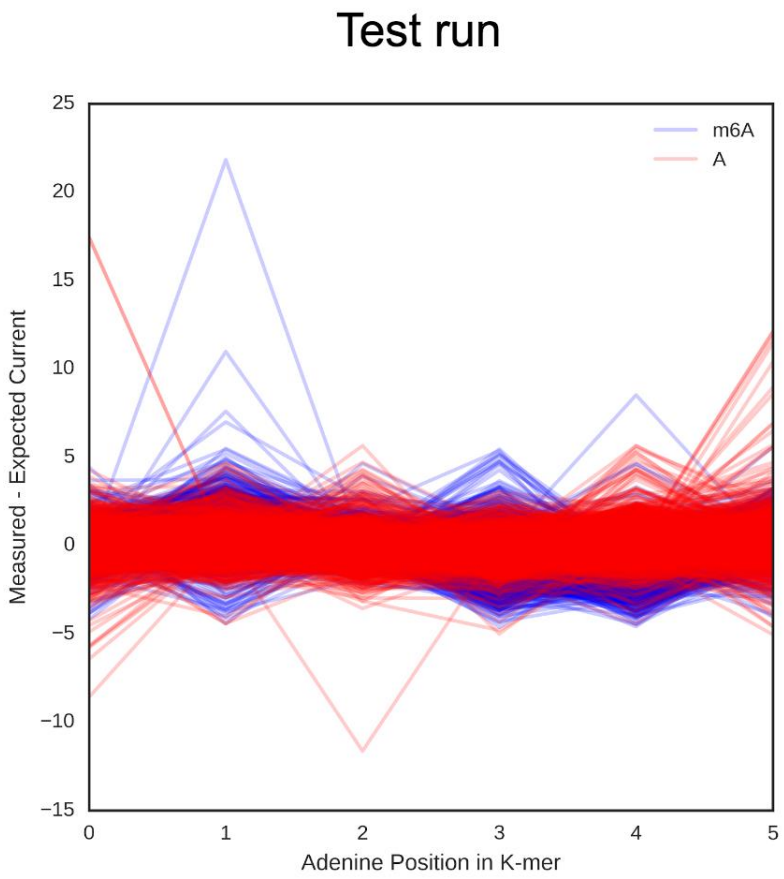
Calling current (pA) differences, similar to PacBio

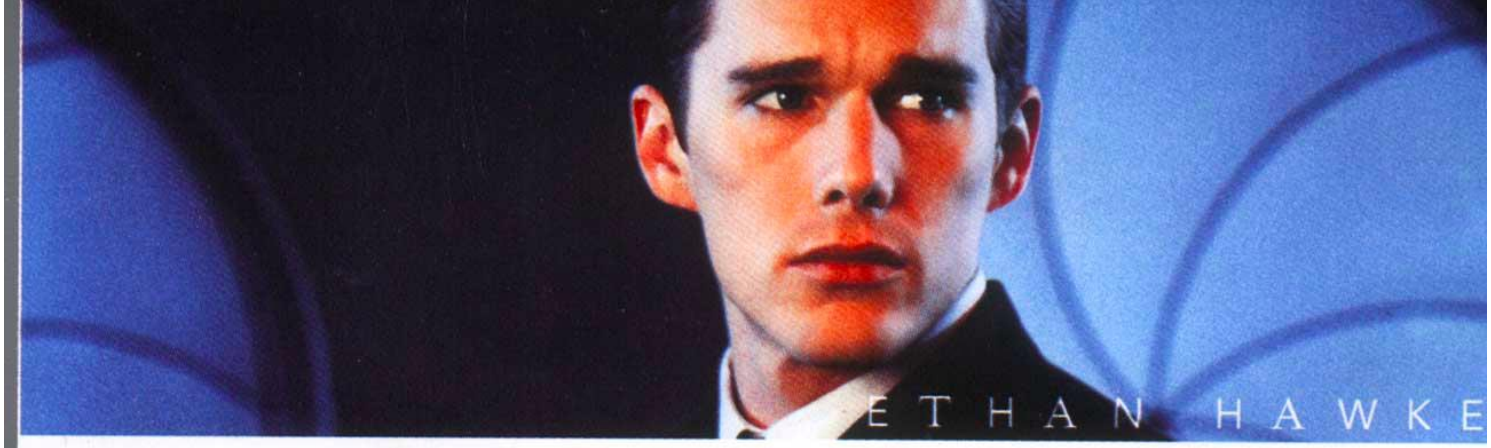




Certain positions of the pore and more informative then others







GATTACA

THERE IS NO GENE FOR THE HUMAN SPIRIT

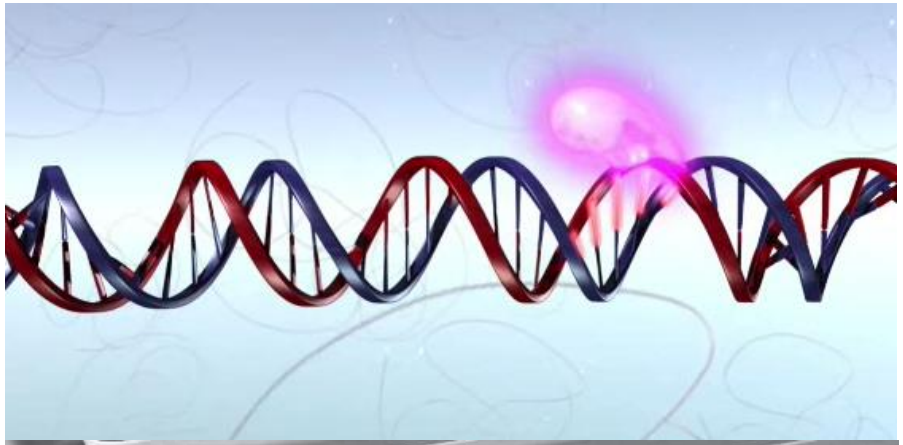
Is a 2.6 minute genome possible?
No today, but if the physics holds up…

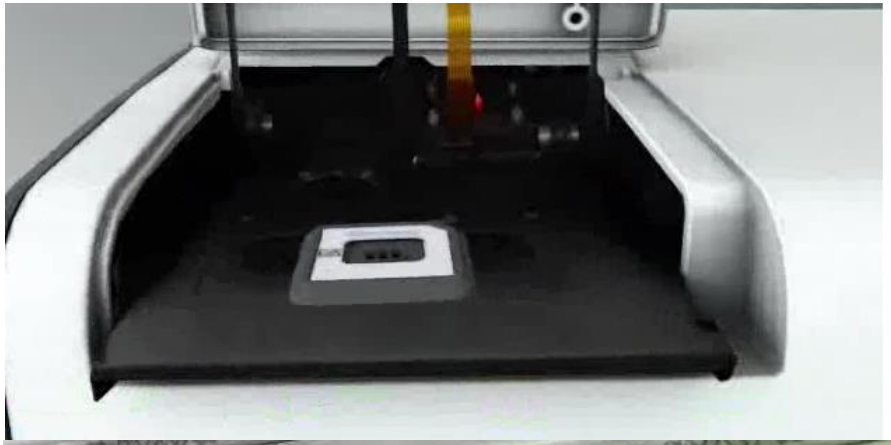
Table 2: Nanopore and Nanochannel Sequencing Considerations

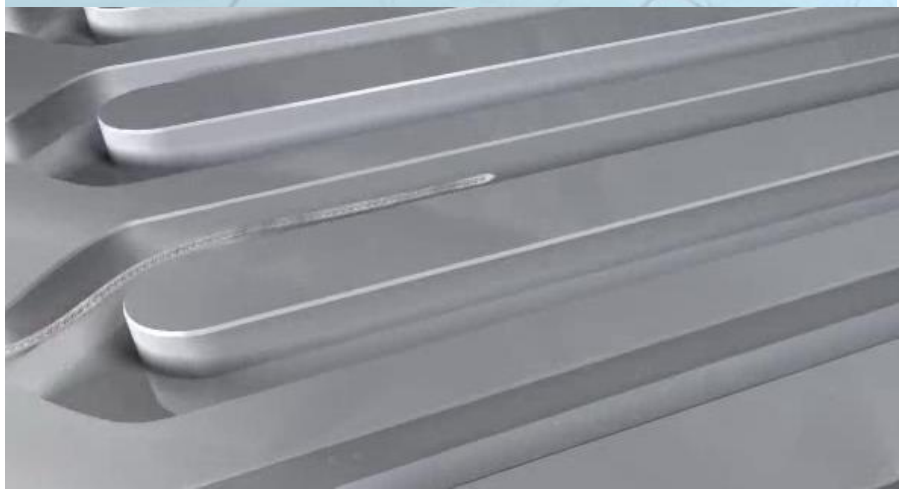
Parameter	DNA fragment (aveargebp)	Pore Speed (bp/s)	# nanopores	% of Pores Functional	transit time (seconds)	transit time (minutes)	run time (hours)	max # molecules / pore / run	% of time pores have DNA	actual # molecules/ pore/run	# of bases sequenced per device	Run Cost ($)	$ / Mb	$ / Gb	Hours for 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

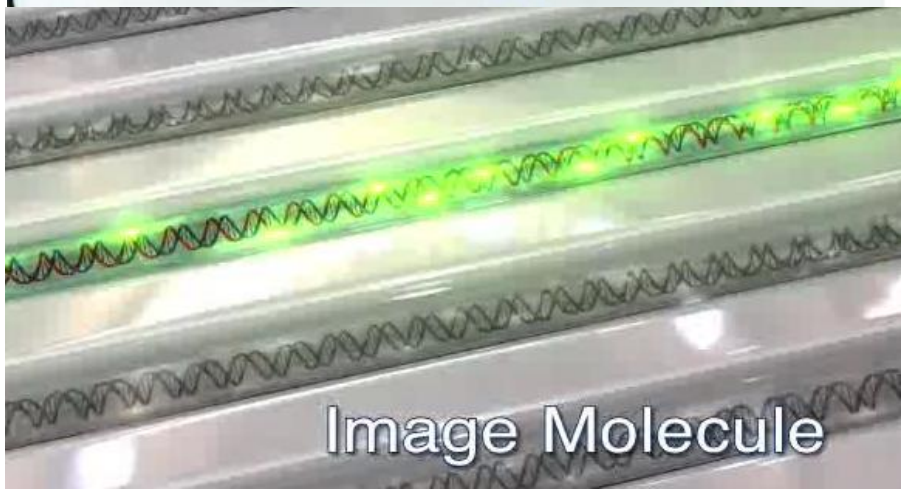
Bionanogenomics - Irys System

BIONANO GENOMICS





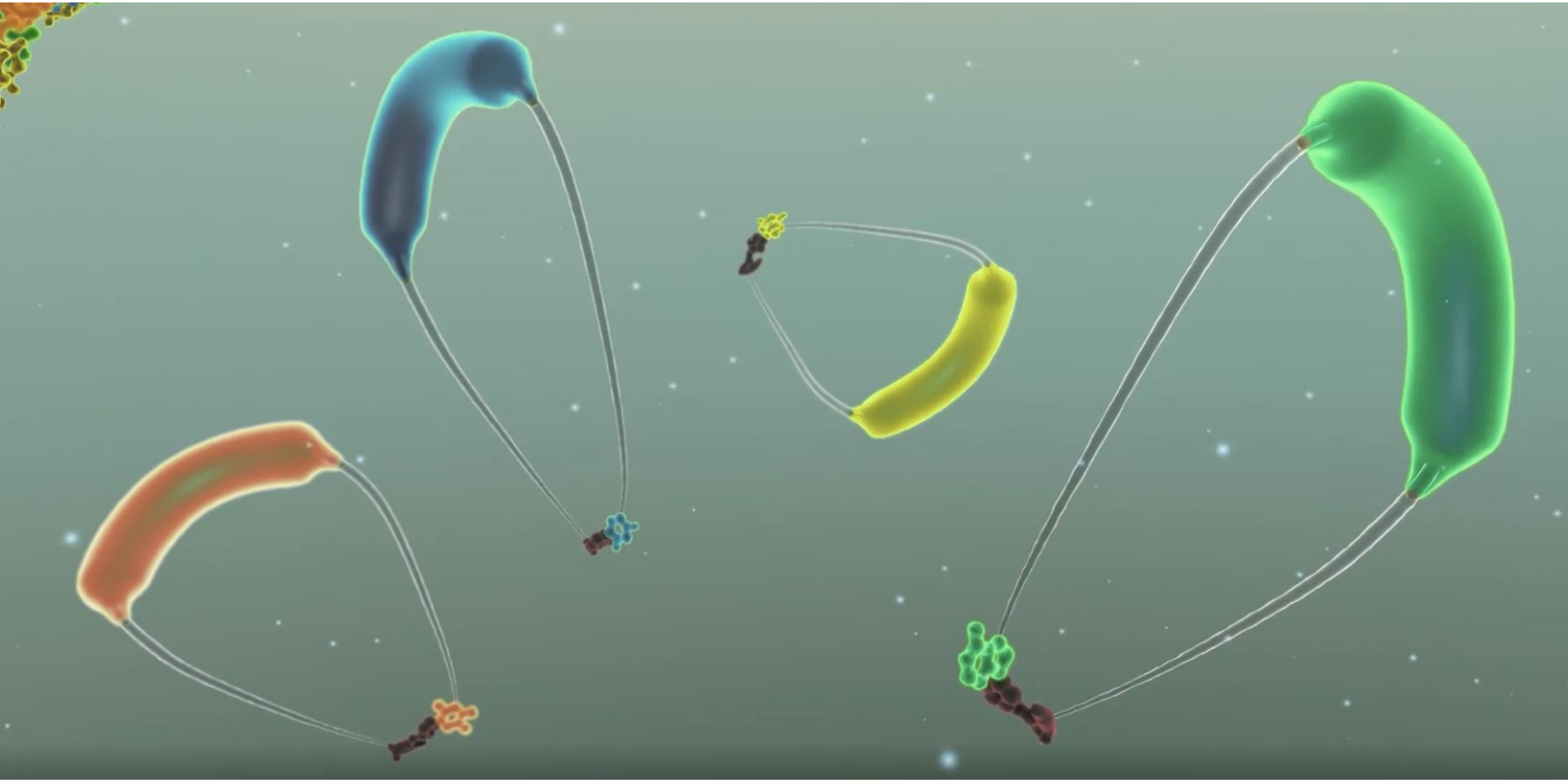




The new Illumina iSeq100 can sequence in <6h.

Emerging Technologies

Roche’s nanopore tech



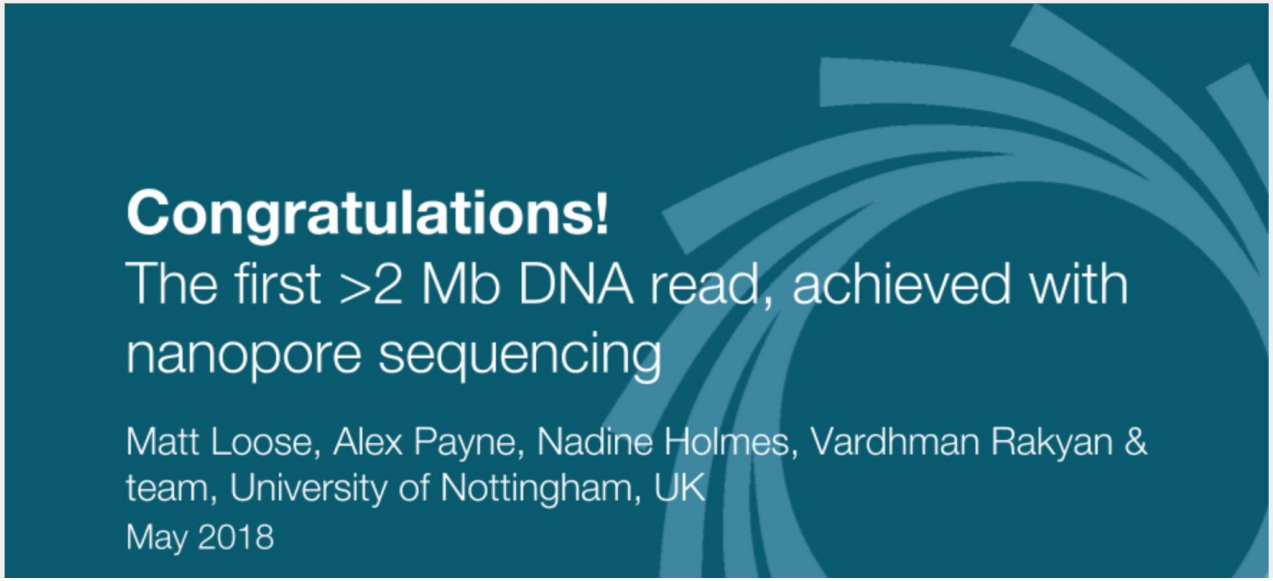
Sequencing by eXpansion (SBX)

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

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



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

News

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.

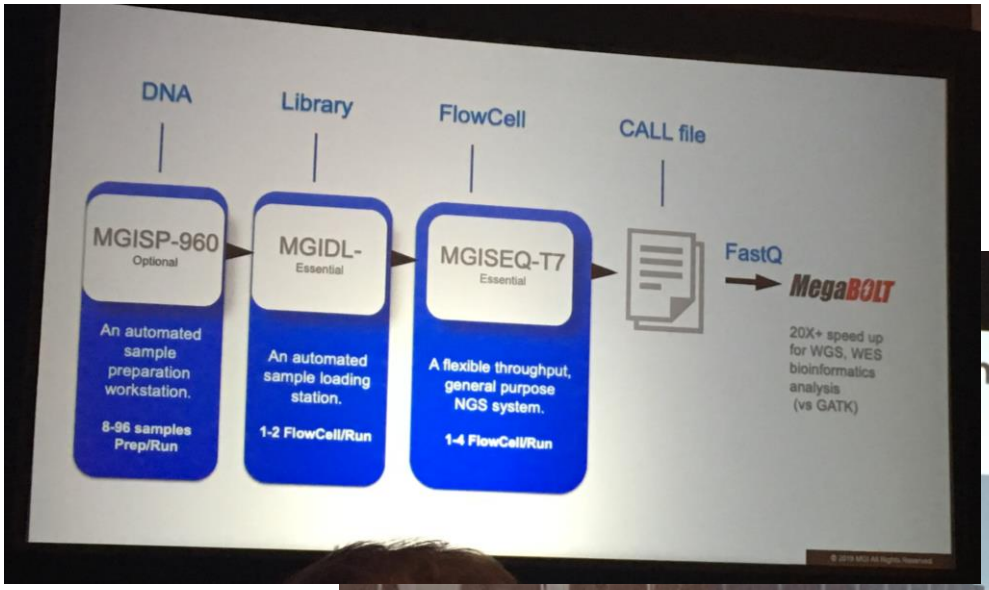
The newly unveiled MGISEQ-T7 is the most powerful sequencing system from BGI's MGI subsidiary, with a daily output capability of 6Tb of data.

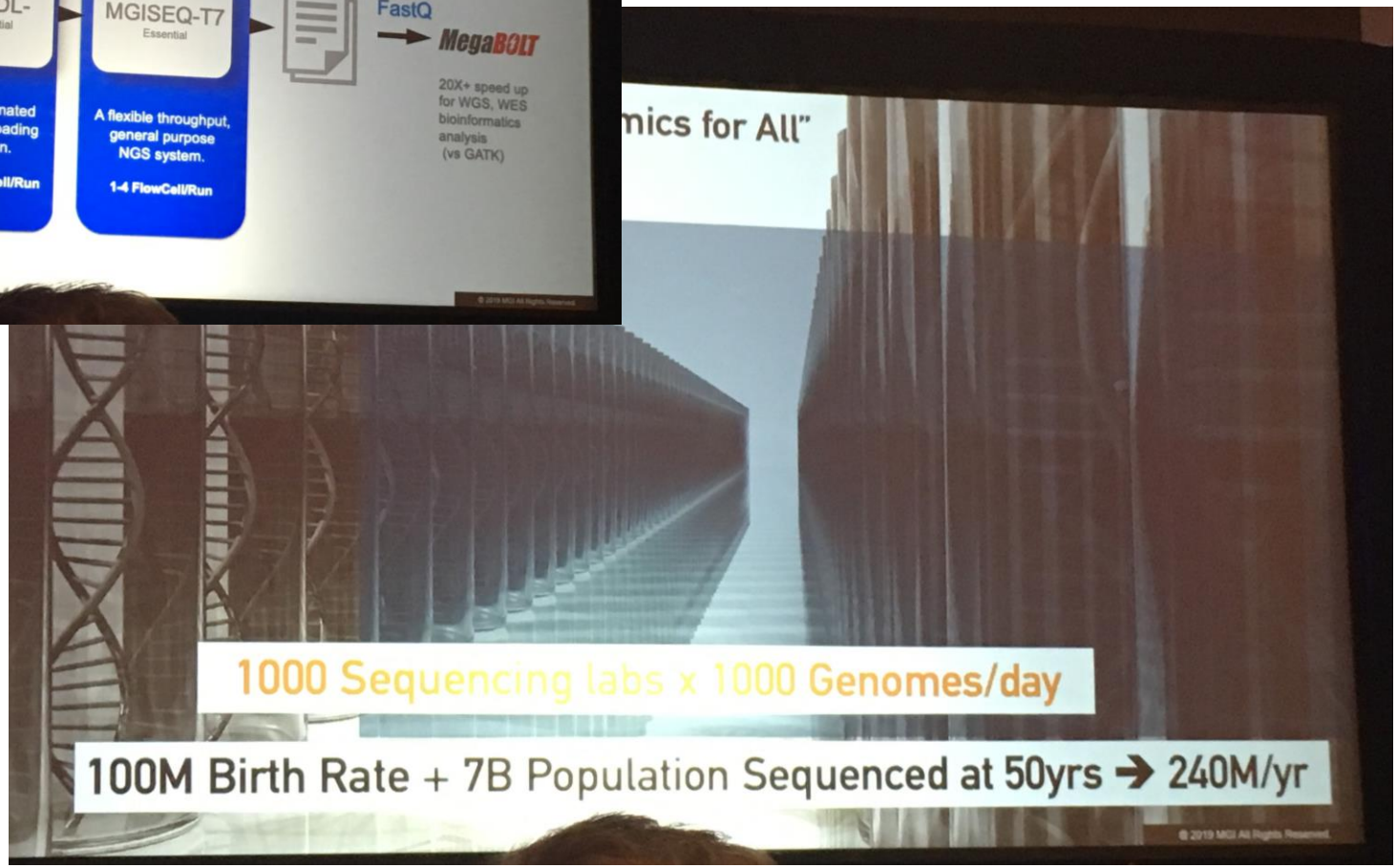
The MGISEQ-T7 is able to complete 60 human genomes in a single day, with essentially error-free sequencing from BGI's DNBseq sequencing technology.

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

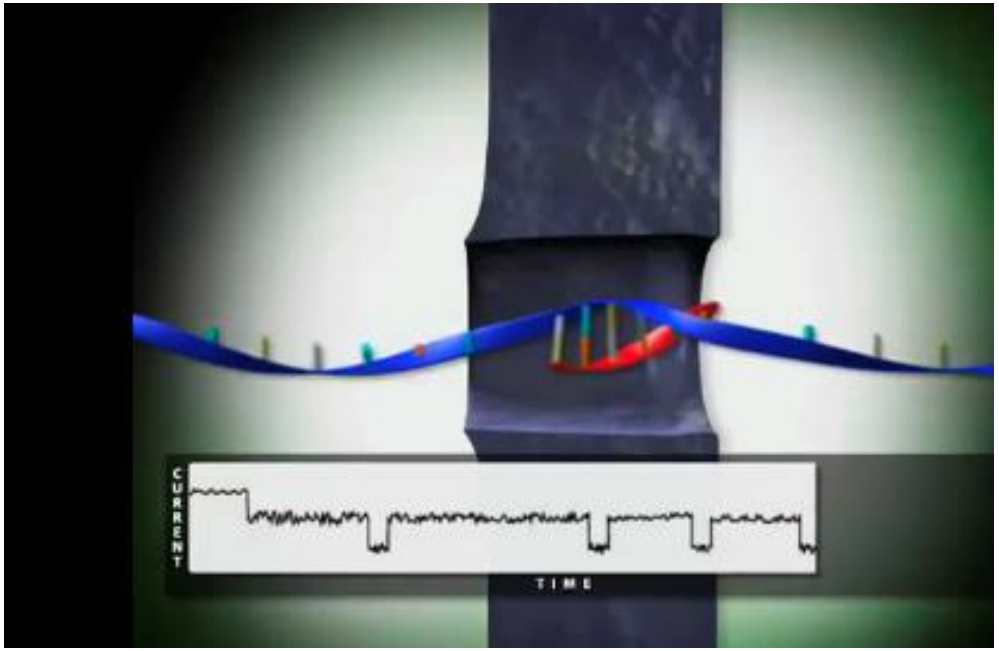
T-1000?

BGI – NGS streets





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

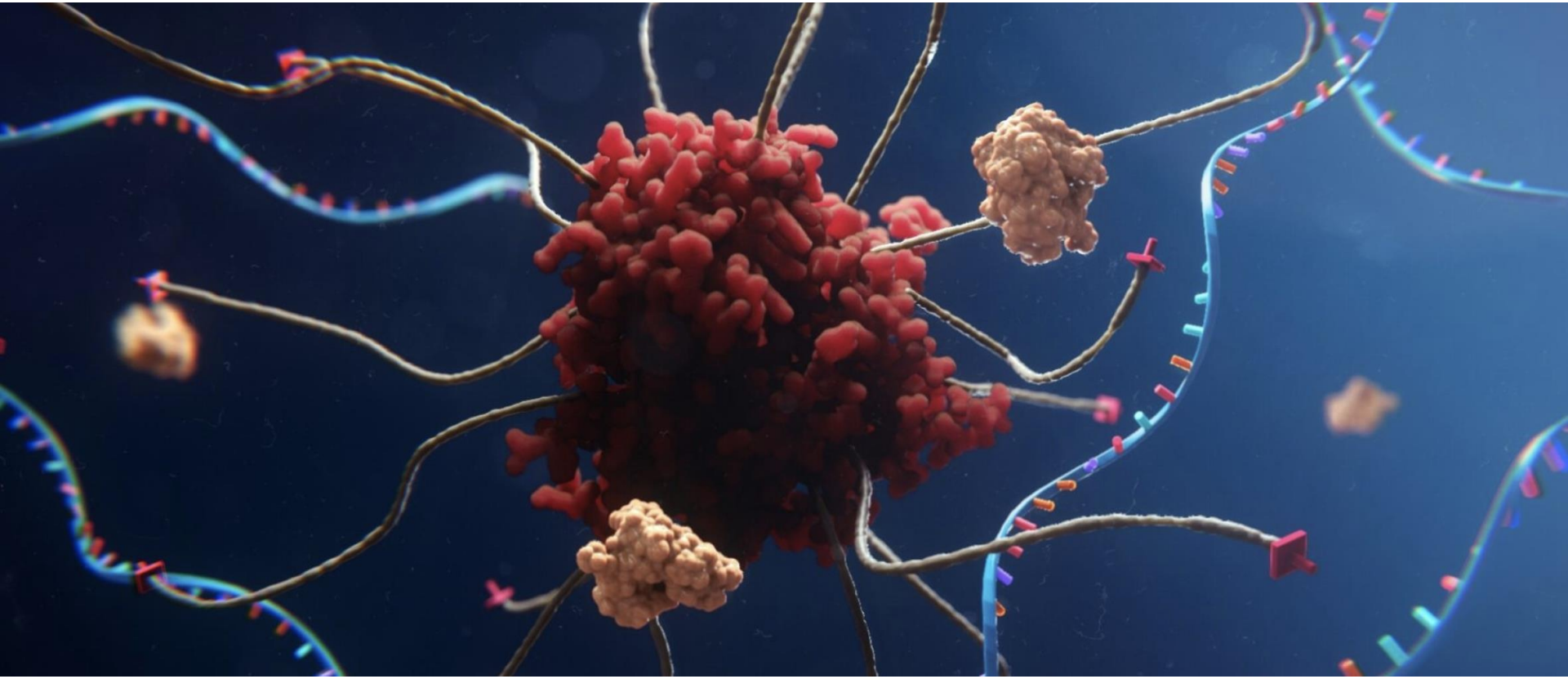
-Long Reads (20kb)
-Single molecule

GenapSys

(1M, 16M or 144M)

GenapSys

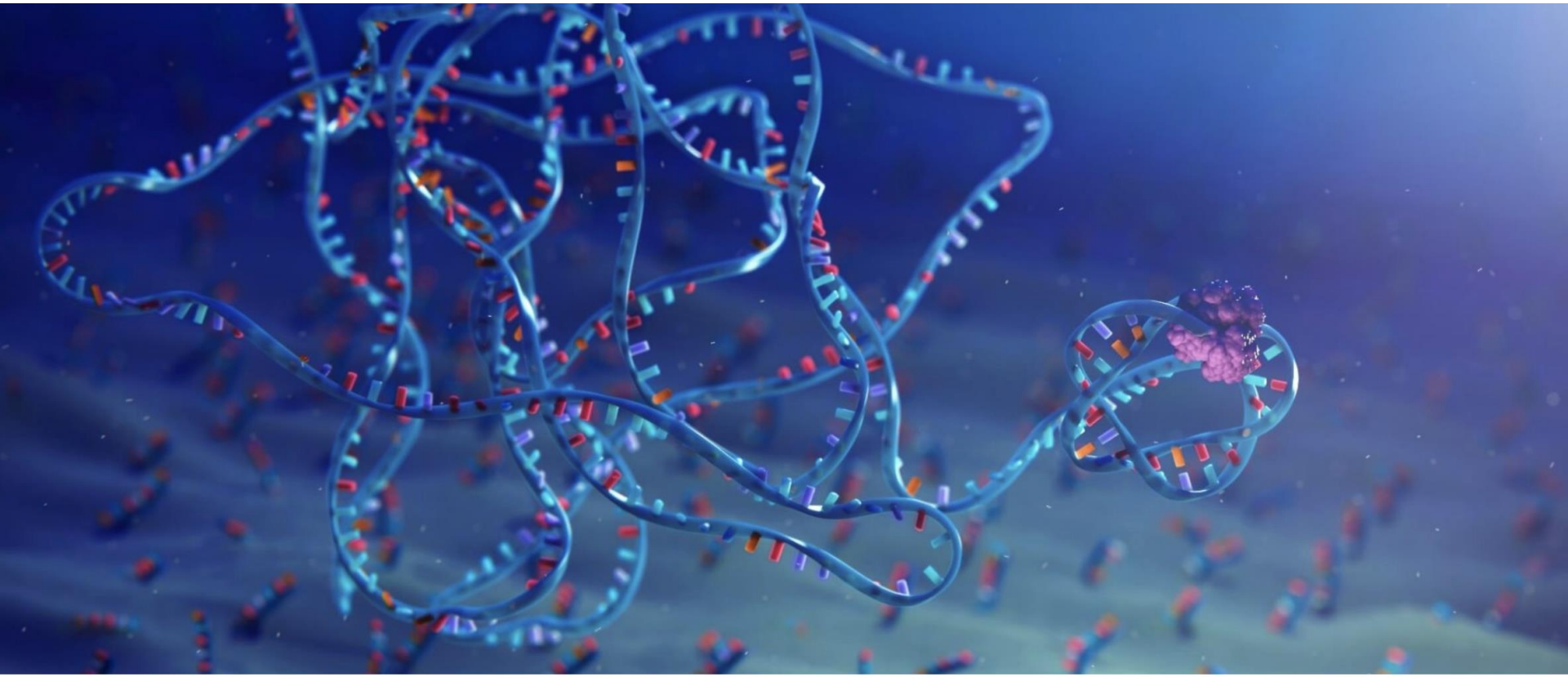
New Kid on the Block: Element Biosciences



Element Biosciences

https://www.elementbiosciences.com/technology

Surface-anchored amplification sites



Element Biosciences

https://www.elementbiosciences.com/technology

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

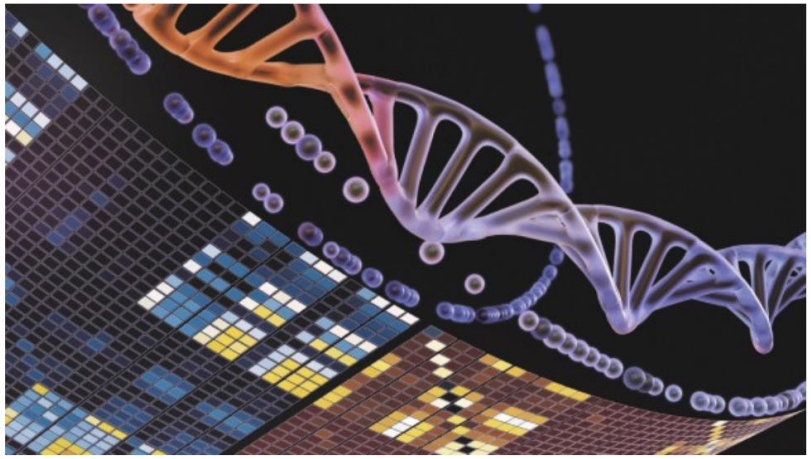
How can we be sure we sequenced correctly?

nature > nature biotechnology > collection

COLLECTION | 09 SEPTEMBER 2021

Sequencing Quality Control 2

This Web Collection presents the results of the Sequencing Quality Control 2 (SEQC2) project that sought to evaluate quality-control metrics and human, bacterial and metagenomic reference materials and datasets for next-generation sequencing (NGS) in both regulatory... show more



https://www.nature.com/collections/seqc2

Test them!

nature biotechnology

Article | Published: 09 September 2021

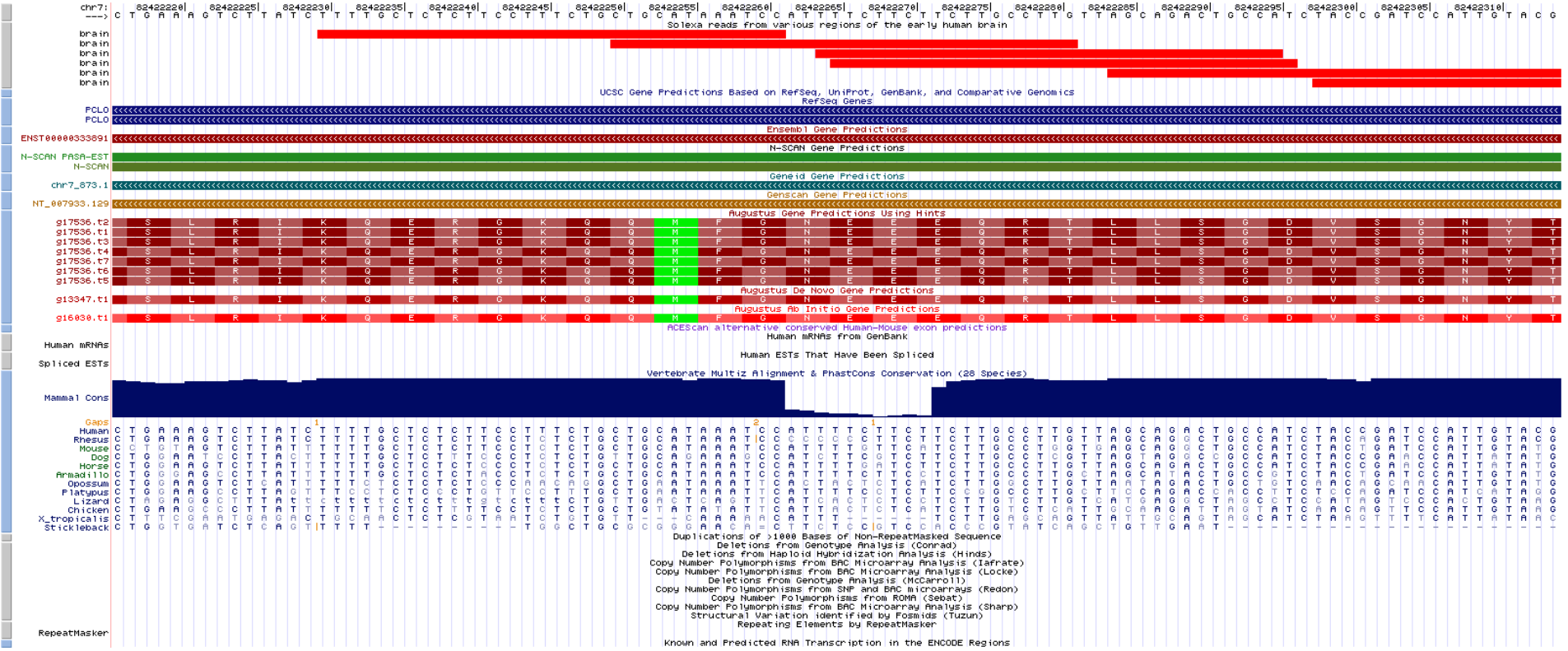
Performance assessment of DNA sequencing platforms in the ABRF Next-Generation Sequencing Study

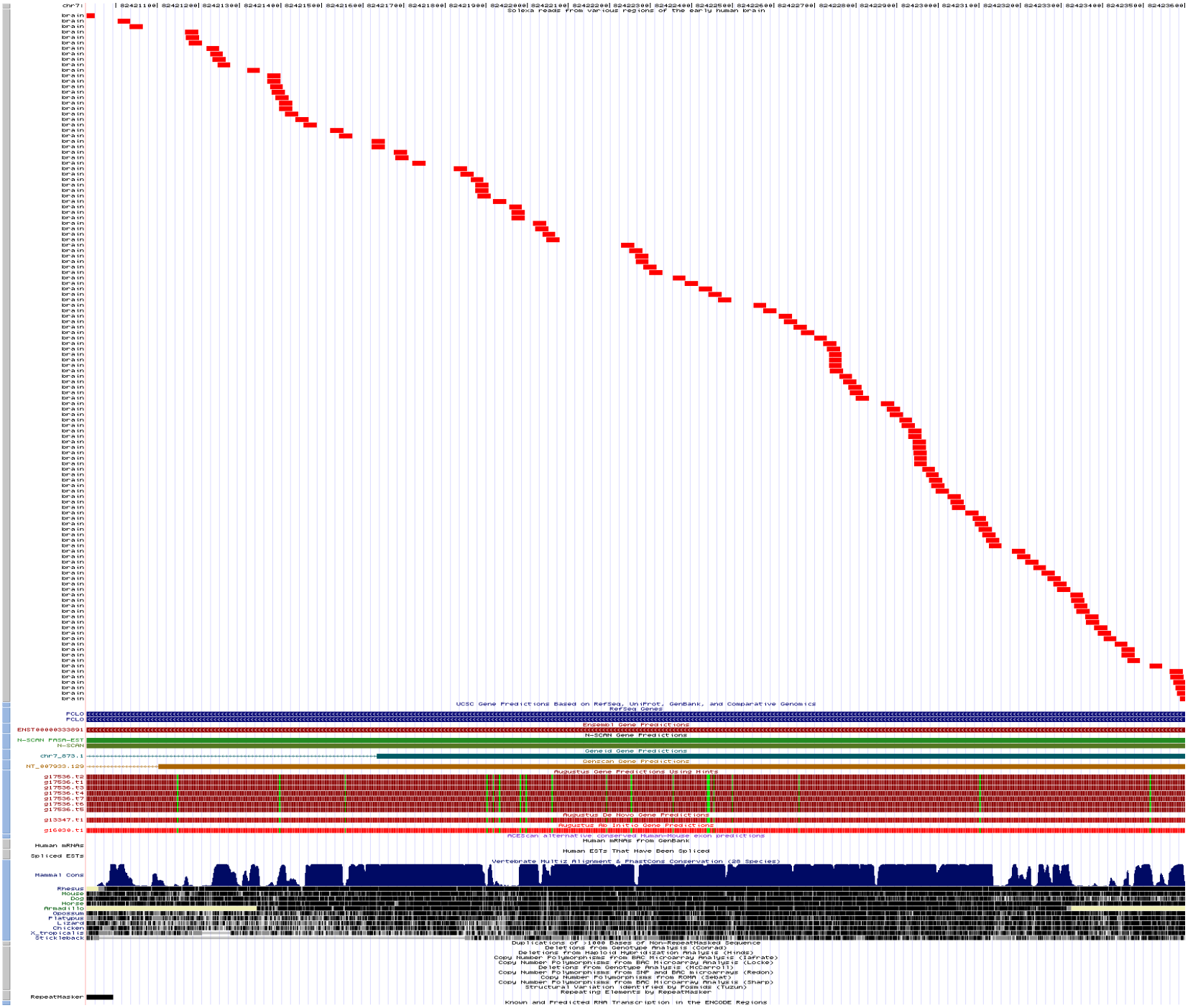
Jonathan Foox, Scott W. Tighe, ... Christopher E. Mason + Show authors

https://www.nature.com/articles/s41587-021-01049-5

What do you do with the reads?

Alignment to the genome




chr7:
Solexa reads from various regions of the early human brain
brain
UCSC Gene Predictions Based on RefSeq, UniProt, GenBank, and Comparative Genomics
RefSeq Genes
FCLO
Ensembl Gene Predictions
ENST00000333691
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 alternatively 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.

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
 !"#$%&'()*+,-./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

```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
  ............................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
  !"#$%&'()*+,-./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

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

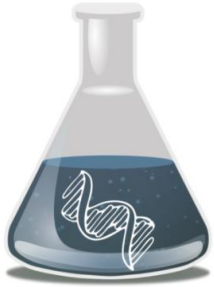
$-4.0 = \log_{10}p$

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

https://jimb.stanford.edu/giab/

GENOME IN A BOTTLE

Our mission is to provide the authoritative characterization of human genomes.



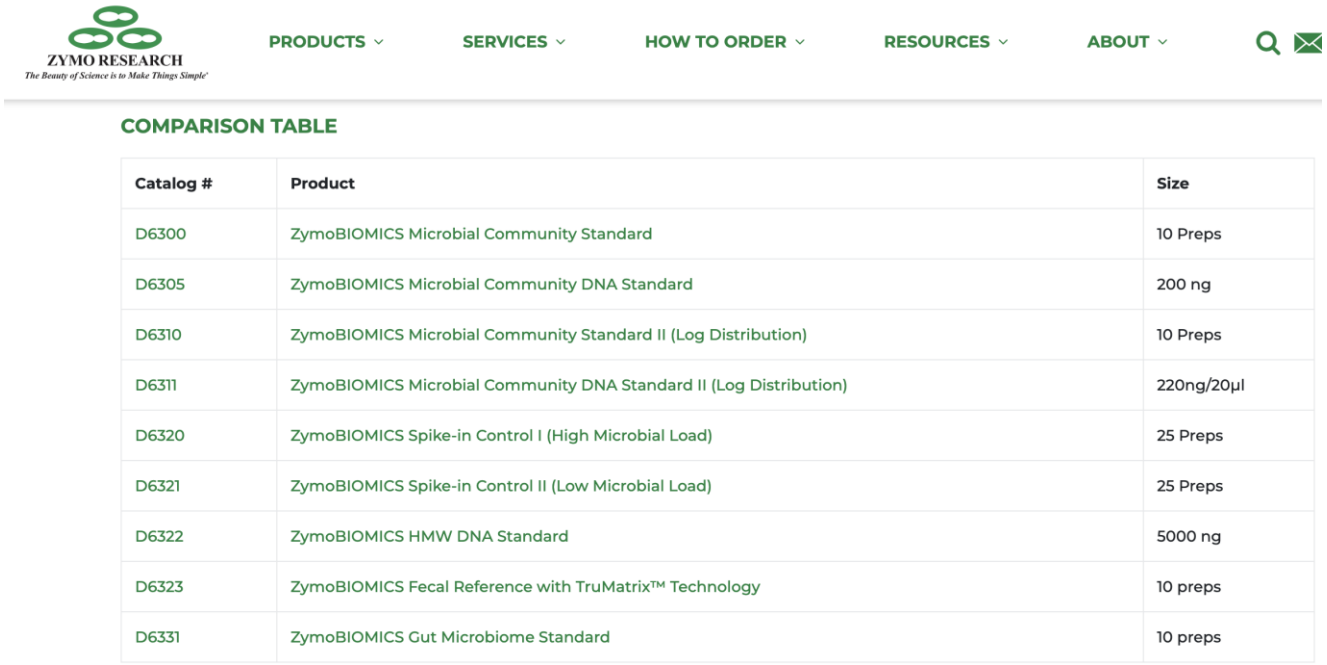
Reference Materials and Data

The Genome in a Bottle Consortium has selected several genomes to produce and characterize as reference materials. The National Institute of Standards and Technology (NIST) is developing NIST Reference Materials from these genomes, which are DNA extracted from a large homogenized growth of B lymphoblastoid cell lines from the Coriell Institute for Medical Research. Note that there may be small differences between the NIST DNA and the Coriell DNA since they come from different growths of cells, though we do not expect these differences to be large for most applications.

The NIST Reference Materials available and planned are listed below, along with links to their data.

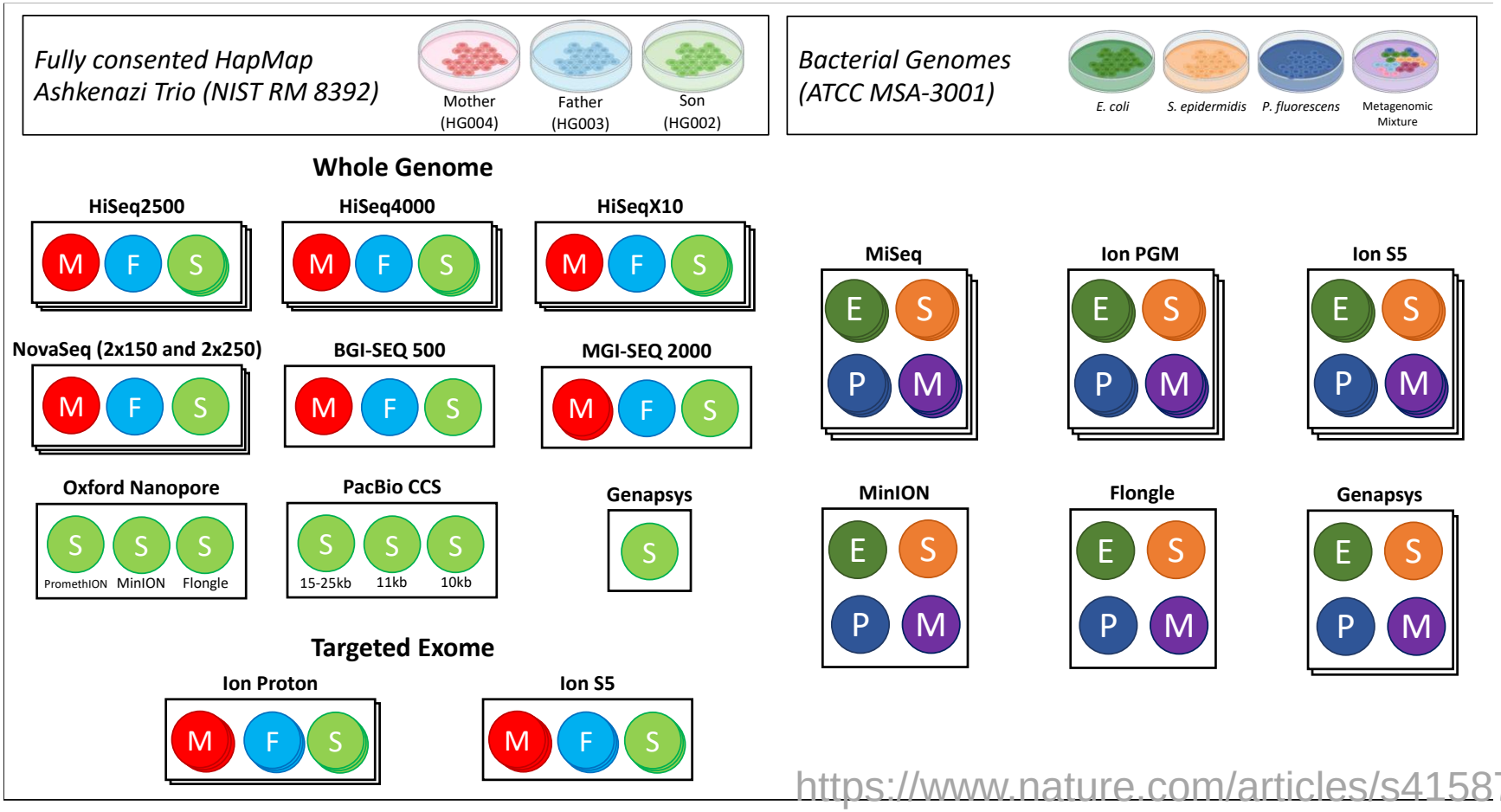
A description of data generated by GIAB for all the genomes below is published here, and characterization of small variants is published here. Ongoing work to characterize more difficult variants and regions is announced in the GIAB Analysis Team google group.

Metagenome in a bottle



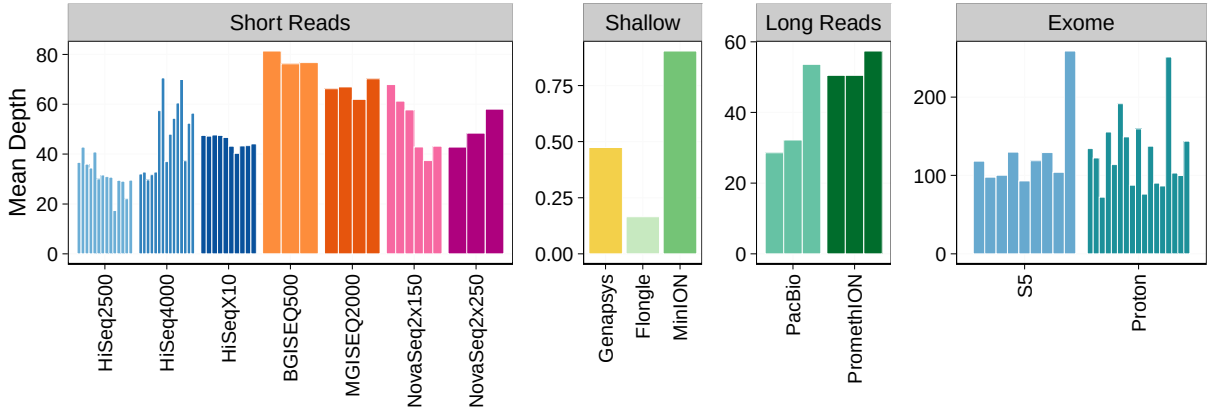
https://www.zymoresearch.com/collections/zymobiomics-microbial-community-standards

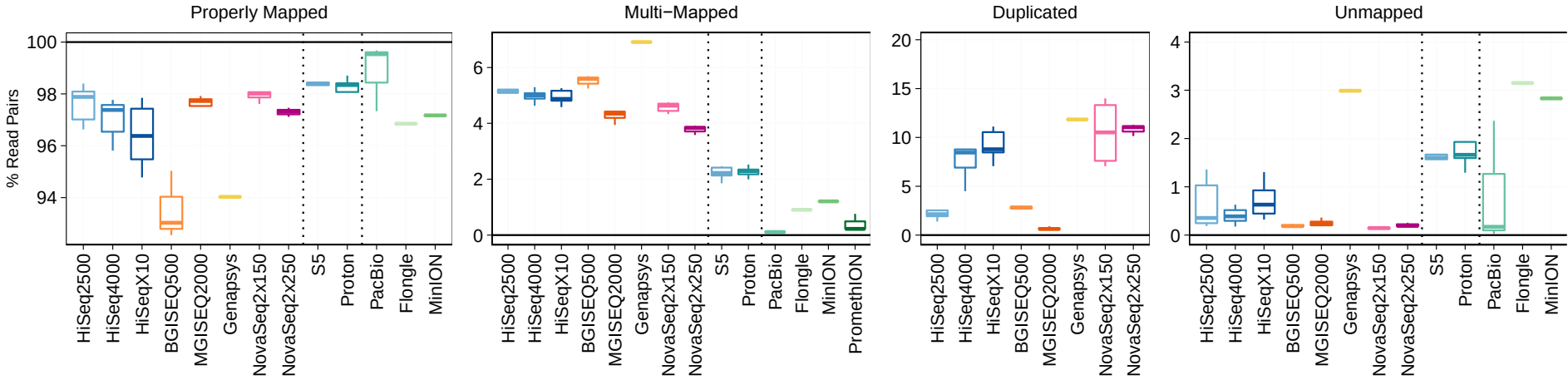
ABRF-NGS Study Overview



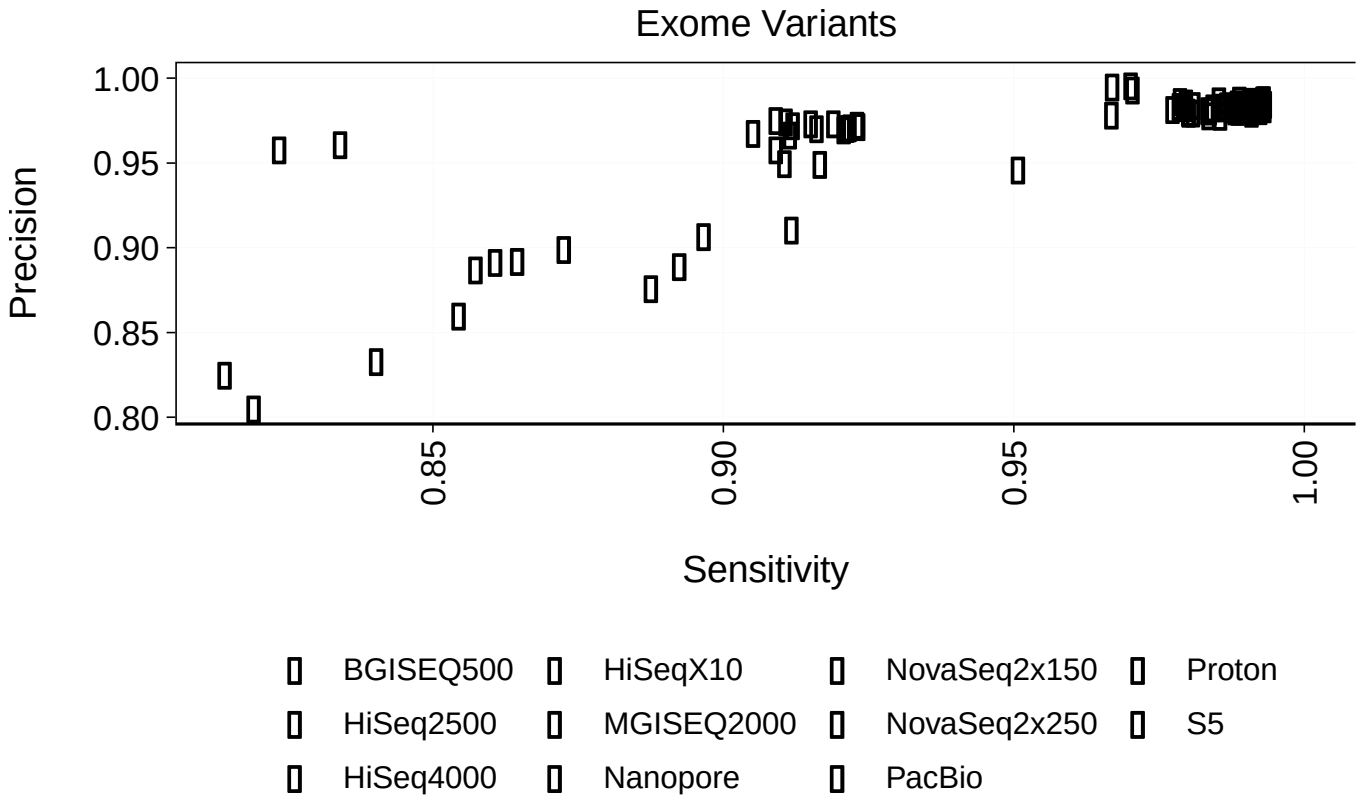
https://www.nature.com/articles/s41587-021-01049-5

Sequencing Depth and Mapping Efficiencies (Human Samples)

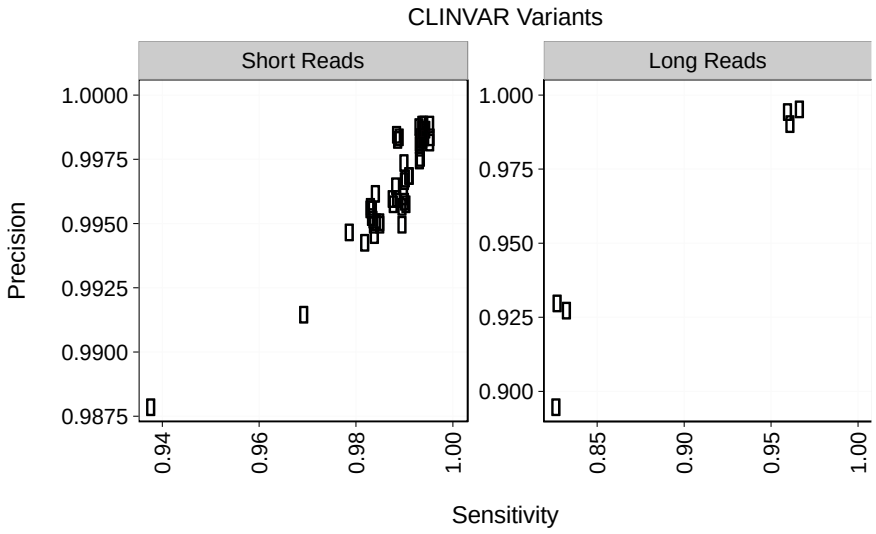


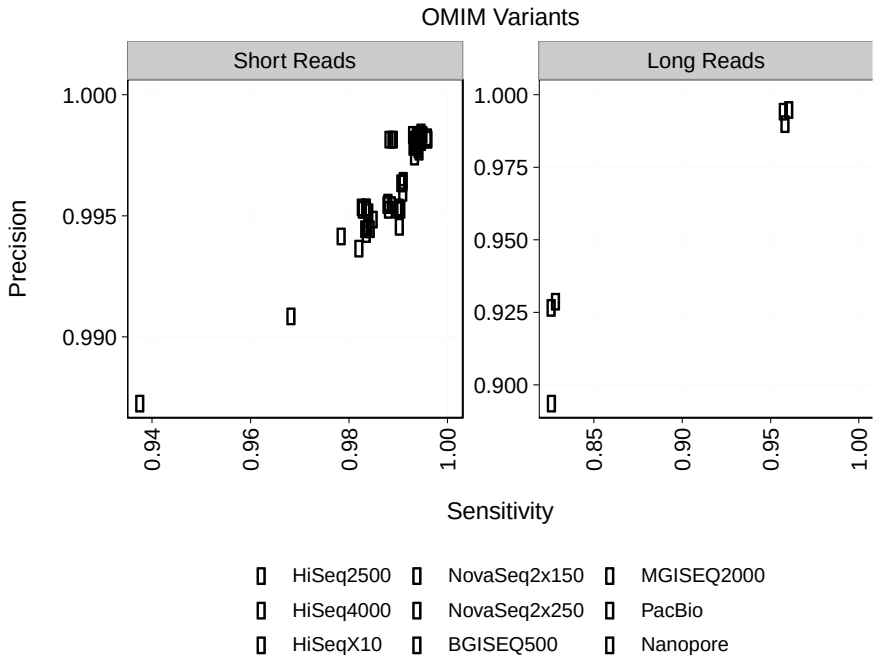


Variant Calling in…

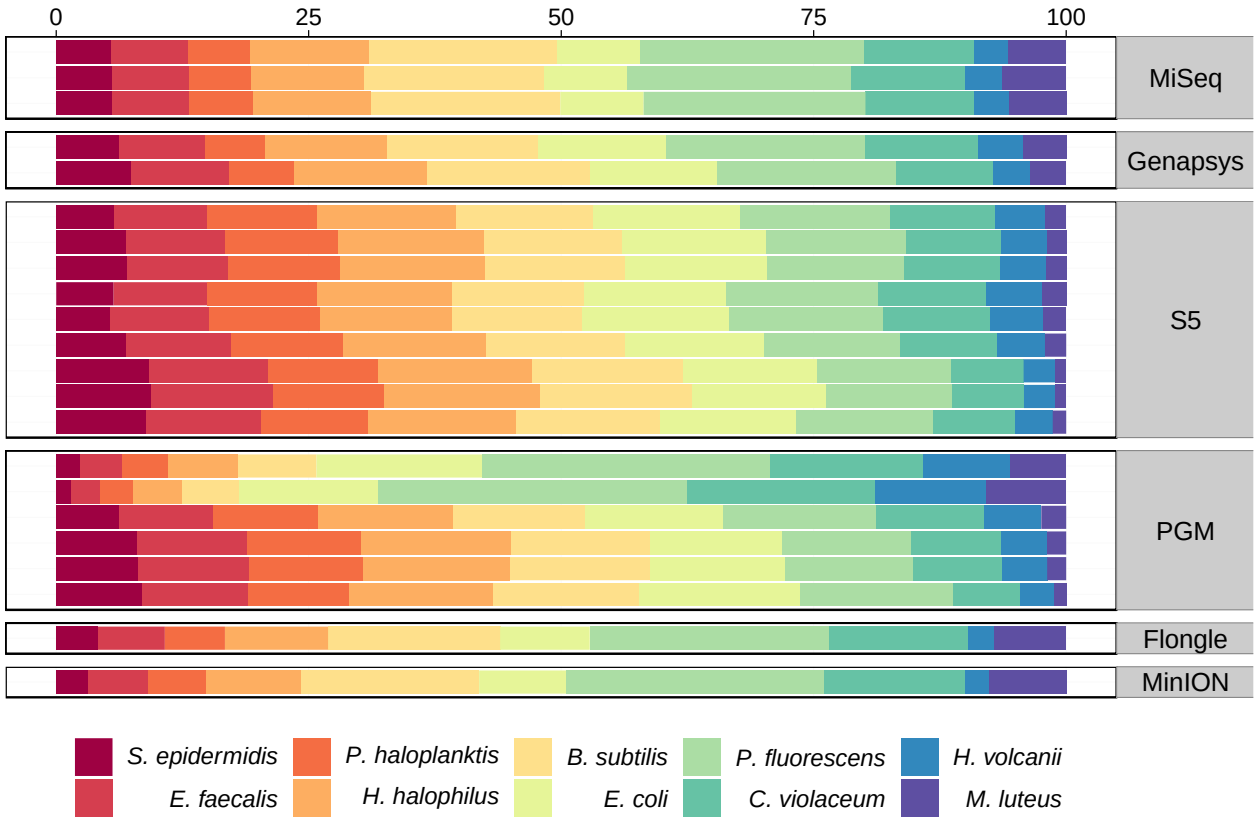


Variant Calling in…
Medically Relevant Genes



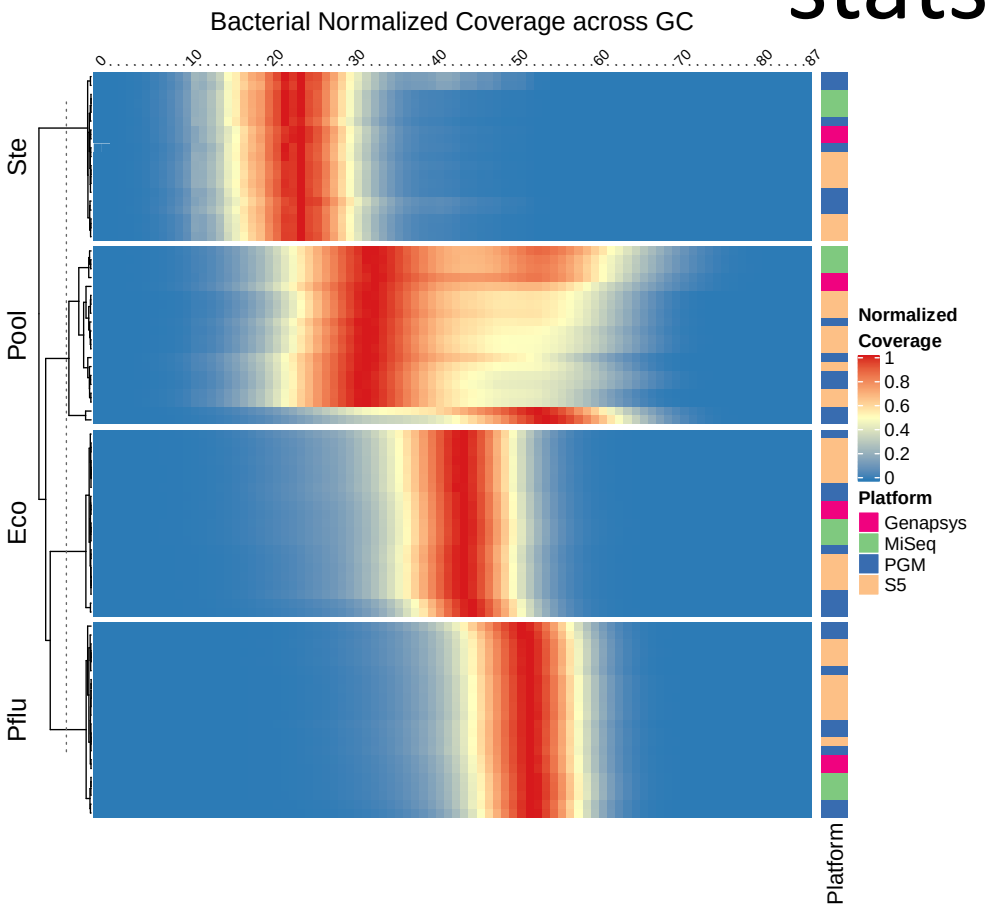


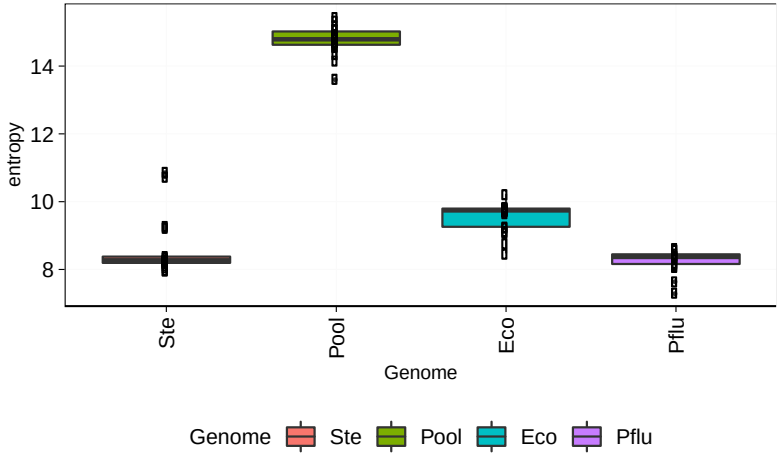
Metagenomic Samples are relatively similar



https://www.nature.com/articles/s41587-021-01049-5

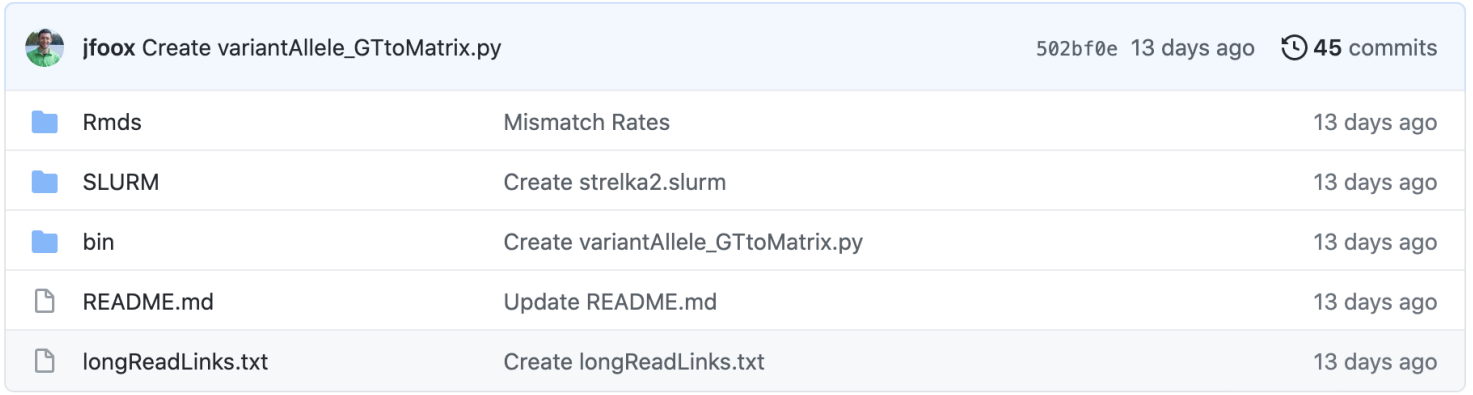
Mixed samples mirror the pooled GC stats





https://www.nature.com/articles/s41587-021-01049-5

All Code and Data Available!



jfoox Create variantAllele_GTtoMatrix.py — `502bf0e` 13 days ago — **45** commits

Rmds	Mismatch Rates	13 days ago
SLURM	Create strelka2.slurm	13 days ago
bin	Create variantAllele_GTtoMatrix.py	13 days ago
README.md	Update README.md	13 days ago
longReadLinks.txt	Create longReadLinks.txt	13 days ago

README.md

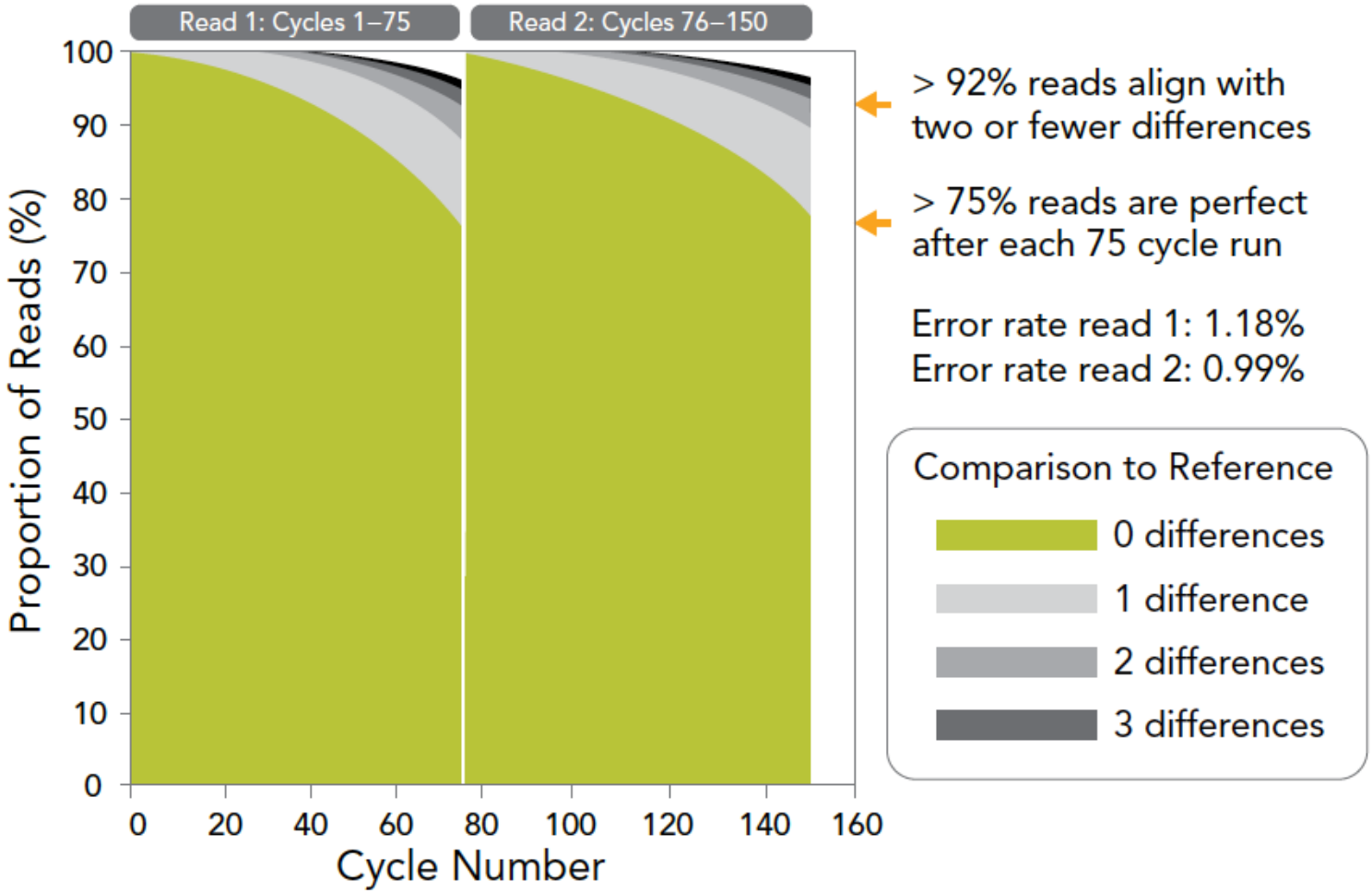
ABRF NGS Phase II

Analysis and figure generation code for the ABRF NGS Phase II Study on DNA-seq reproducibility. This repository includes scripts to run heavy lifting such as alignment and variant calling (SLURM), shell scripts to do post-processing calculations (bin), and R scripts used to create figures (Rmds).

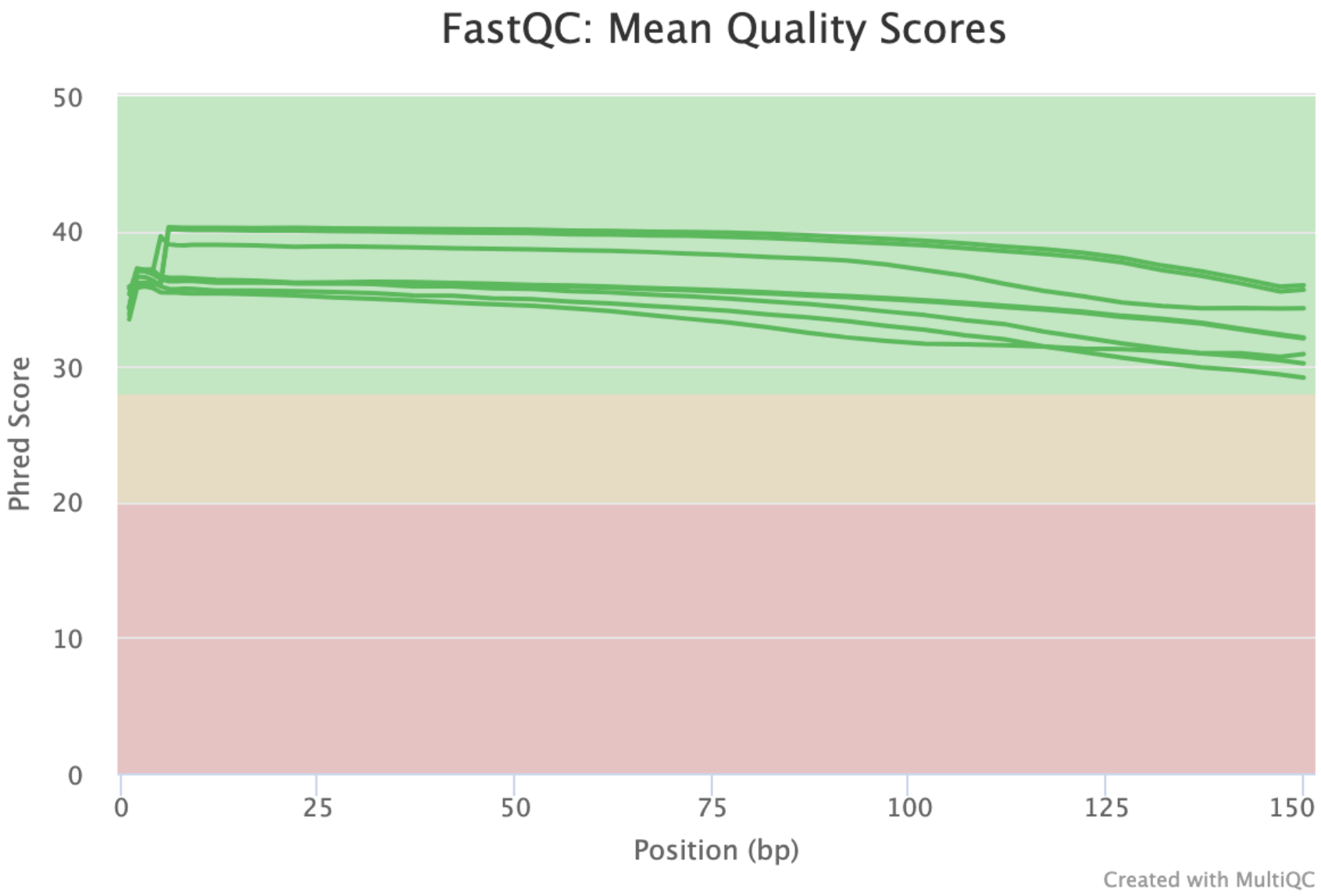
https://github.com/jfoox/abrfngs2

https://www.nature.com/articles/s41587-021-01049-5

Many platforms are cycle-dependent on error rate - ILMN



Element looks better so far



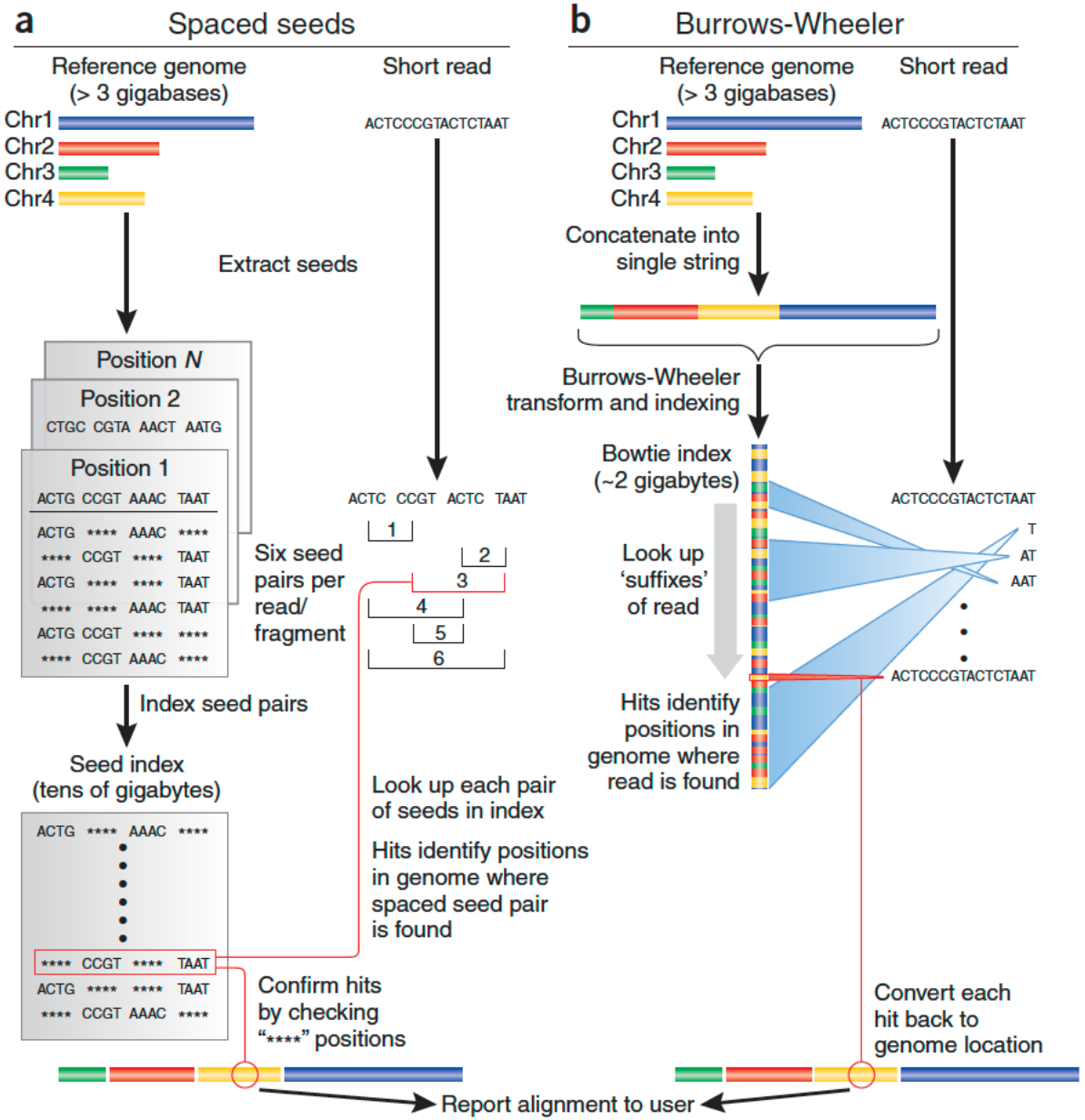
Many Options for Alignment - 2009

	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 - 2022

- 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?