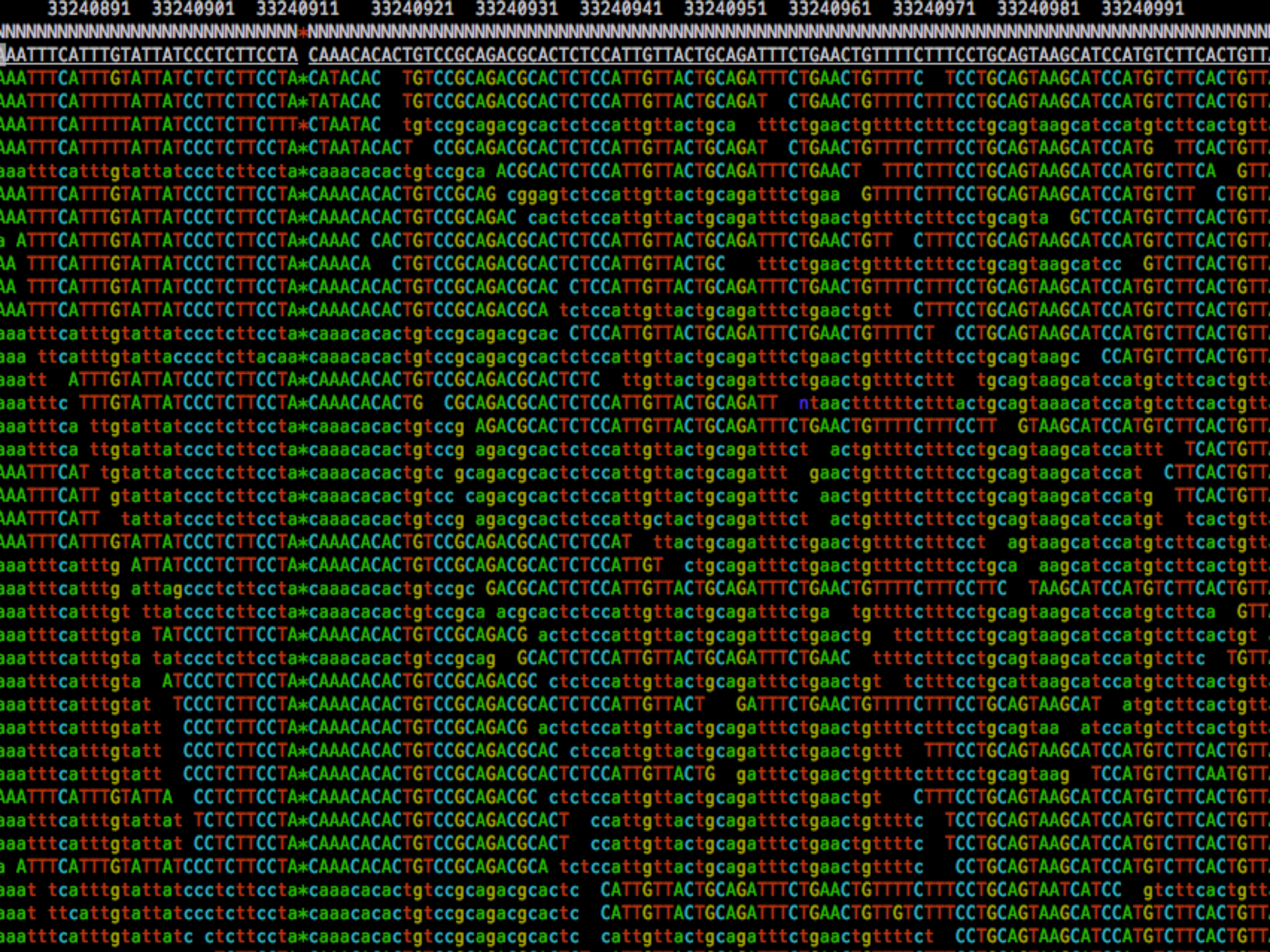


```
    33240891  33240901  33240911   33240921  33240931  33240941  33240951  33240961  33240971  33240981  33240991
NNNNNNNNNNNNNNNNNNNNNNNNNNNNN*NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
AAATTTCATTTGTATTATCCCTCTTCCTA CAAACACACTGTCCGCAGACGCACTCTCCATTGTTACTGCAGATTTCTGAACTGTTTTCTTTCCTGCAGTAAGCATCCATGTCTTCACTGTT
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  ntaactttttctttactgcagtaaacatccatgtcttcactgtt
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 2018
Lectures 01-02-03

Professor:

Christopher E. Mason, Ph.D.

TAs:

Ebrahim Afshinnekoo

Alexa McIntyre

Course Over Eight Sessions:

I. **Sequencing Methods, Single-Cell Dynamics, and Molecular Detection Techniques (March 14th)**

II. **RNA Sequencing, Epitranscriptomes, and Gene Fusions (March 21st)**

III. **Epigenomes, DNA Modifications, and Chromatin Dynamics (March 28th)**

IV. **Microbiome and Metagenome Characterizations and Cross-Species Analysis (April 4th)**

V. **Complex Genome Re-arrangements, Transposons, and Tools for Genetic Variant Calling (April 11th)**

VI. **Cancer Genomics, Non-coding Regulation and Variation, and Statistical Power (April 25th)**

VII. **Systems Biology, Big Data, and Disease Classification (May 2nd)**

VIII. **Big Health, Sculpting Evolution, Synthetic Biology, & Genome Engineering (May 9th)**

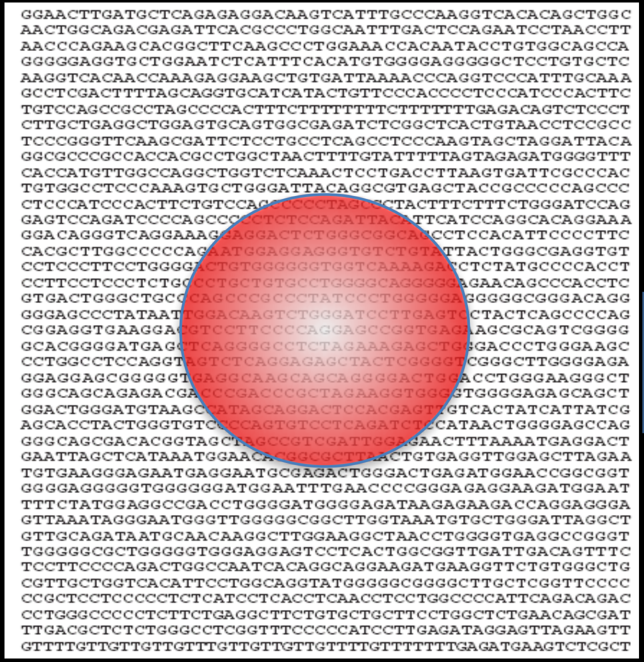
All classes on Wednesday, 10:00-11:30
1305 York Avenue, 13th floor, Y13-01

Stay updated with the course webpage:

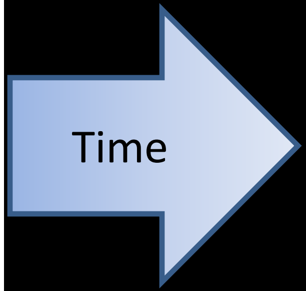
http://physiology.med.cornell.edu/faculty/mason/lab/clinicalgenomics/schedule.html

Start

Finish

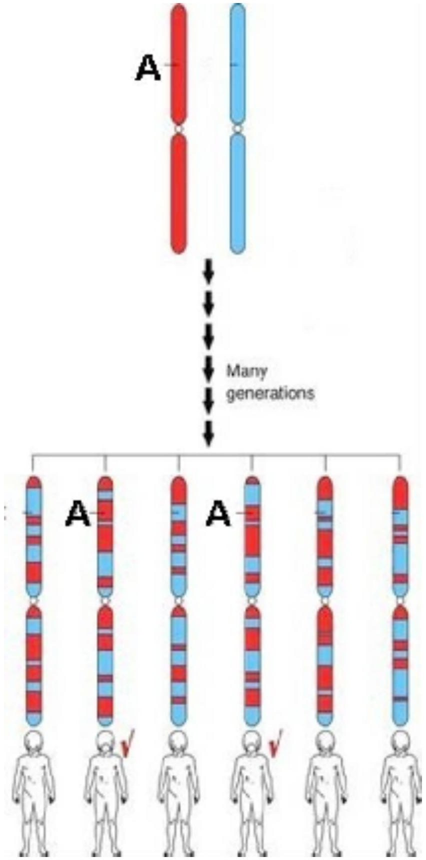


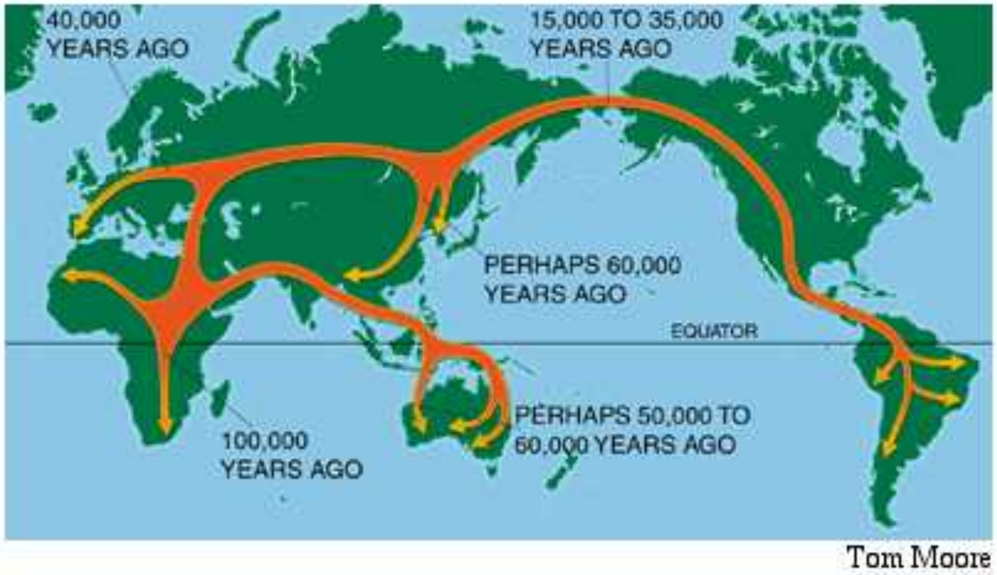
```
GGAACTTGATGCTCAGAGAGGACAAGTCATTTGCCCAAGGTCACACAGCTGGC
AACTGGCAGACGAGATTCACGCCCTGGCAATTTGACTCCAGAATCCTAACCTT
AACCCAGAAGCACGGCTTCAAGCCCTGGAAACCACAATACCTGTGGCAGCCA
GGGGGAGGTGCTGGAATCTCATTTCACATGTGGGGAGGGGGCTCCTGTGCTC
AAGGTCACAACCAAAGAGGAAGCTGTGATTAAAACCCAGGTCCCATTTGCAAA
GCCTCGACTTTTAGCAGGTGCATCATACTGTTCCCACCCCTCCCATCCCACTTC
TGTCCAGCCGCCTAGCCCCACTTTCTTTTTTTTCTTTTTTTGAGACAGTCTCCCT
CTTGCTGAGGCTGGAGTGCAGTGGCGAGATCTCGGCTCACTGTAACCTCCGCC
TCCCGGGTTCAAGCGATTCTCCTGCCTCAGCCTCCCAAGTAGCTAGGATTACA
GGCGCCCGCCACCACGCCTGGCTAACTTTTGTATTTTTAGTAGAGATGGGGTTT
CACCATGTTGGCCAGGCTGGTCTCAAACTCCTGACCTTAAGTGATTCGCCCAC
TGTGGCCTCCCAAAGTGCTGGGATTACAGGCGTGAGCTACCGCCCCCAGCCC
CTCCCATCCCACTTCTGTCCAGCCCCCTAGCCCTACTTTCTTTCTGGGATCCAG
GAGTCCAGATCCCCAGCCCCCTCTCCAGATTACATTCATCCAGGCACAGGAAA
GGACAGGGTCAGGAAAGGAGGACTCTGGGCGGCAGCCTCCACATTCCCCTTC
CACGCTTGGCCCCCAGAATGGAGGAGGGTGTCTGTATTACTGGGCGAGGTGT
CCTCCCTTCCTGGGGACTGTGGGGGGTGGTCAAAAGACCTCTATGCCCCACCT
CCTTCCTCCCTCTGCCCTGCTGTGCCTGGGGCAGGGGGAGAACAGCCCACCTC
GTGACTGGGCTGCCCAGCCCGCCCTATCCCTGGGGGAGGGGGCGGGACAGG
GGGAGCCCTATAATTGGACAAGTCTGGGATCCTTGAGTCCTACTCAGCCCCAG
CGGAGGTGAAGGACGTCCTTCCCCAGGAGCCGGTGAGAAGCGCAGTCGGGG
GCACGGGGATGAGCTCAGGGGCCTCTAGAAAGAGCTGGGACCCTGGGAAGC
CCTGGCCTCCAGGTAGTCTCAGGAGAGCTACTCGGGGTCGGGCTTGGGGAGA
GGAGGAGCGGGGGTGAGGCAAGCAGCAGGGGACTGGACCTGGGAAGGGCT
GGGCAGCAGAGACGACCCGACCCGCTAGAAGGTGGGGTGGGGAGAGCAGCT
GGACTGGGATGTAAGCCATAGCAGGACTCCACGAGTTGTCACTATCATTATCG
AGCACCTACTGGGTGTCCCCAGTGTCCTCAGATCTCCATAACTGGGGAGCCAG
GGGCAGCGACACGGTAGCTAGCCGTCGATTGGAGAACTTTAAAATGAGGACT
GAATTAGCTCATAAATGGAACACGGCGCTTAACTGTGAGGTTGGAGCTTAGAA
TGTGAAGGGAGAATGAGGAATGCGAGACTGGGACTGAGATGGAACCGGCGGT
GGGGAGGGGGTGGGGGGATGGAATTTGAACCCCGGGAGAGGAAGATGGAAT
TTTCTATGGAGGCCGACCTGGGGATGGGGAGATAAGAGAAGACCAGGAGGGA
GTTAAATAGGGAATGGGTTGGGGGCGGCTTGGTAAATGTGCTGGGATTAGGCT
GTTGCAGATAATGCAACAAGGCTTGGAAGGCTAACCTGGGGTGAGGCCGGGT
TGGGGGCGCTGGGGGTGGGAGGAGTCCTCACTGGCGGTTGATTGACAGTTTC
TCCTTCCCCAGACTGGCCAATCACAGGCAGGAAGATGAAGGTTCTGTGGGCTG
CGTTGCTGGTCACATTCCTGGCAGGTATGGGGGCGGGGCTTGCTCGGTTCCCC
CCGCTCCTCCCCCTCTCATCCTCACCTCAACCTCCTGGCCCCATTCAGACAGAC
CCTGGGCCCCCTCTTCTGAGGCTTCTGTGCTGCTTCCTGGCTCTGAACAGCGAT
TTGACGCTCTCTGGGCCTCGGTTTCCCCCATCCTTGAGATAGGAGTTAGAAGTT
GTTTTGTTGTTGTTGTTTGTTGTTGTTTTGTTTTTTTGAGATGAAGTCTCGCT
```



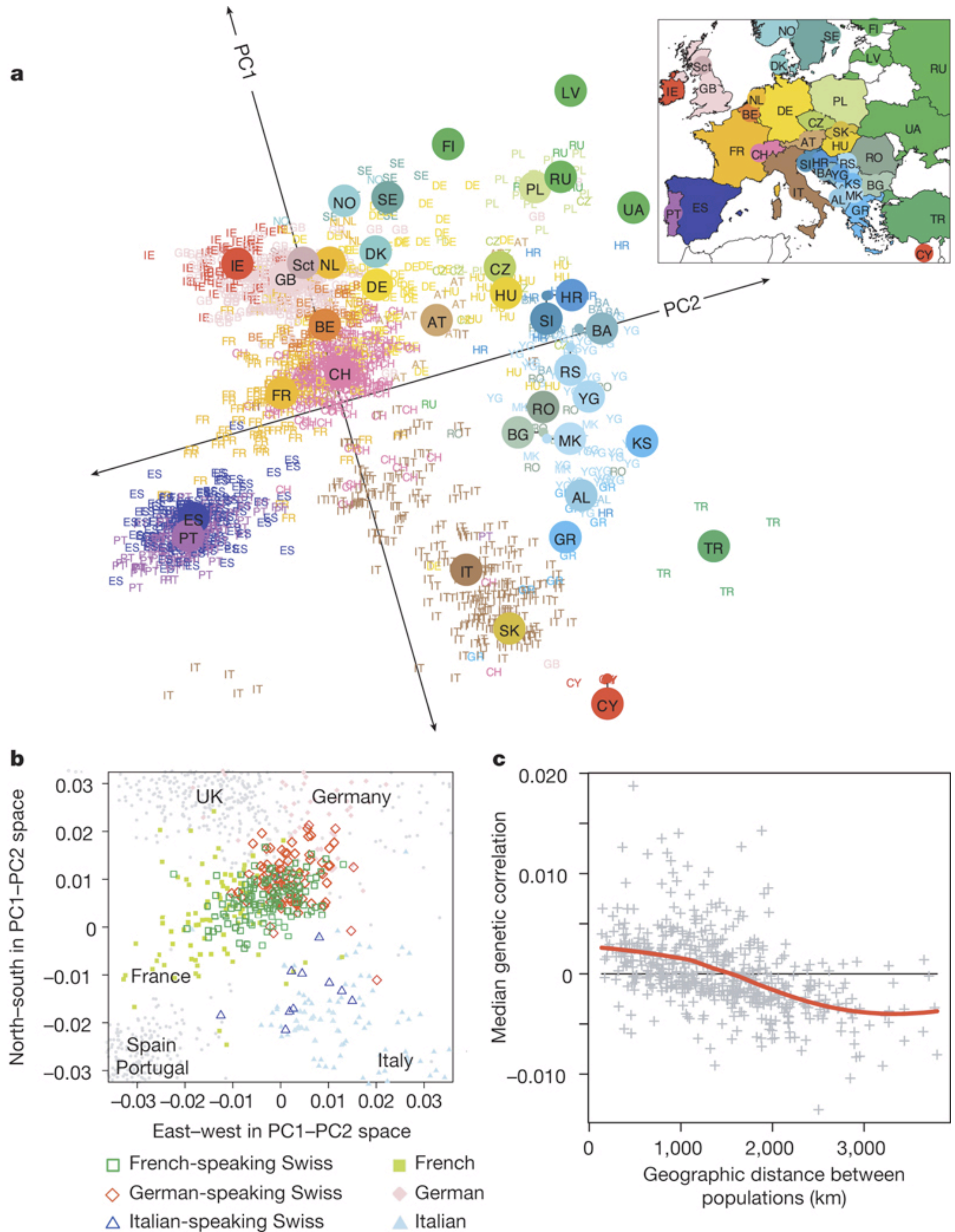
Time

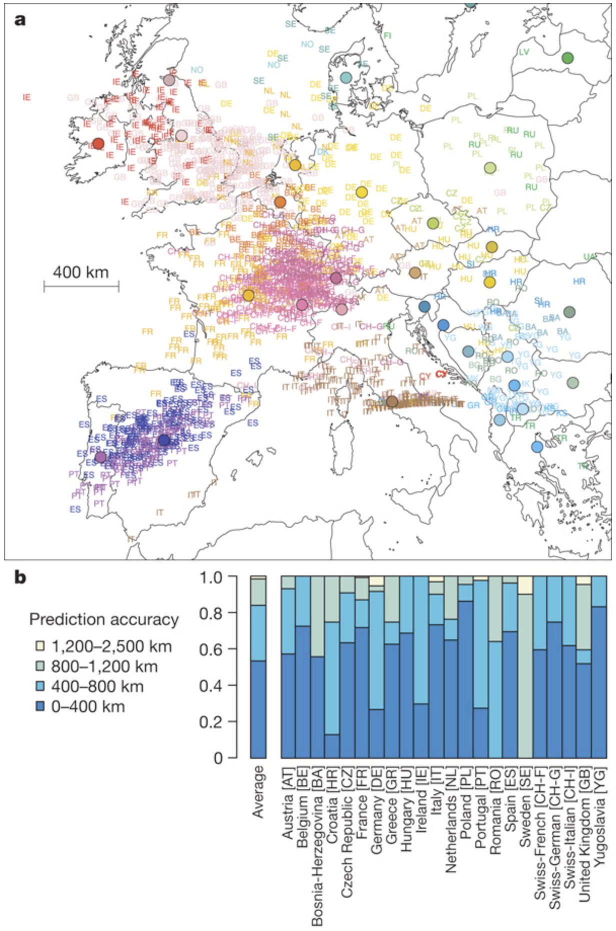
Our genes come from the migration patterns of haplotypes throughout human history ("Population Stratification")





Genotype data can even predict your birthplace





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

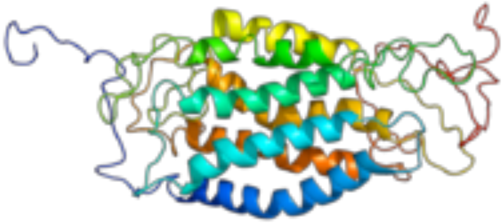
Specific genes can have significant impact

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.

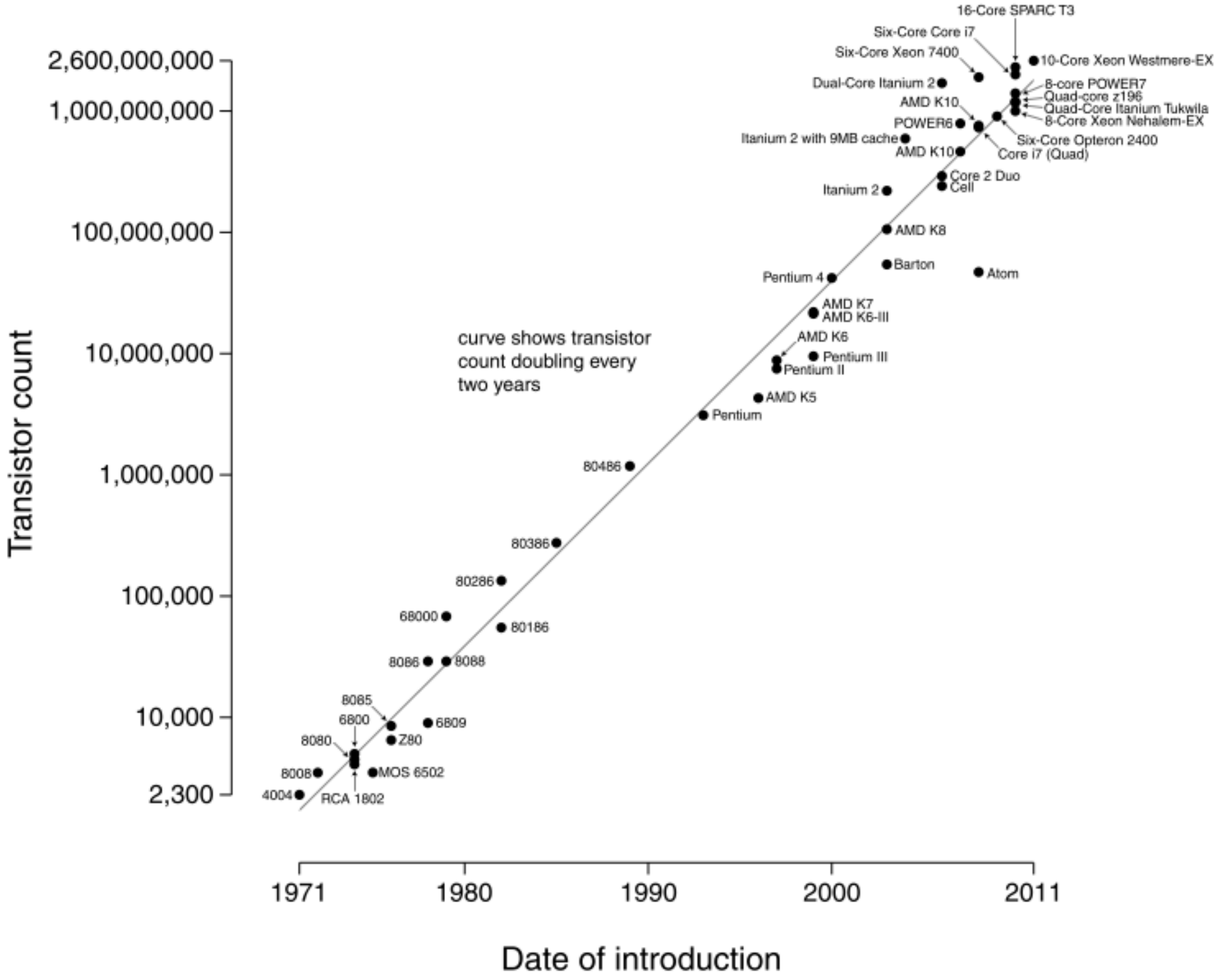
The first control group of 1896 healthy adults, including 1079 women (age range 15–91, mean 58.3) and 817 men (age range 23–90, mean 59.4). These controls were selected at random from the computerized patient lists of five large family practices located in the region of Szczecin. These healthy adults 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 effects from Moore’s Law ushered in a whole new era of technology

Microprocessor Transistor Counts 1971-2011 & Moore’s Law



By Wgsimon

Initially we expected a $1K Genome in 2040

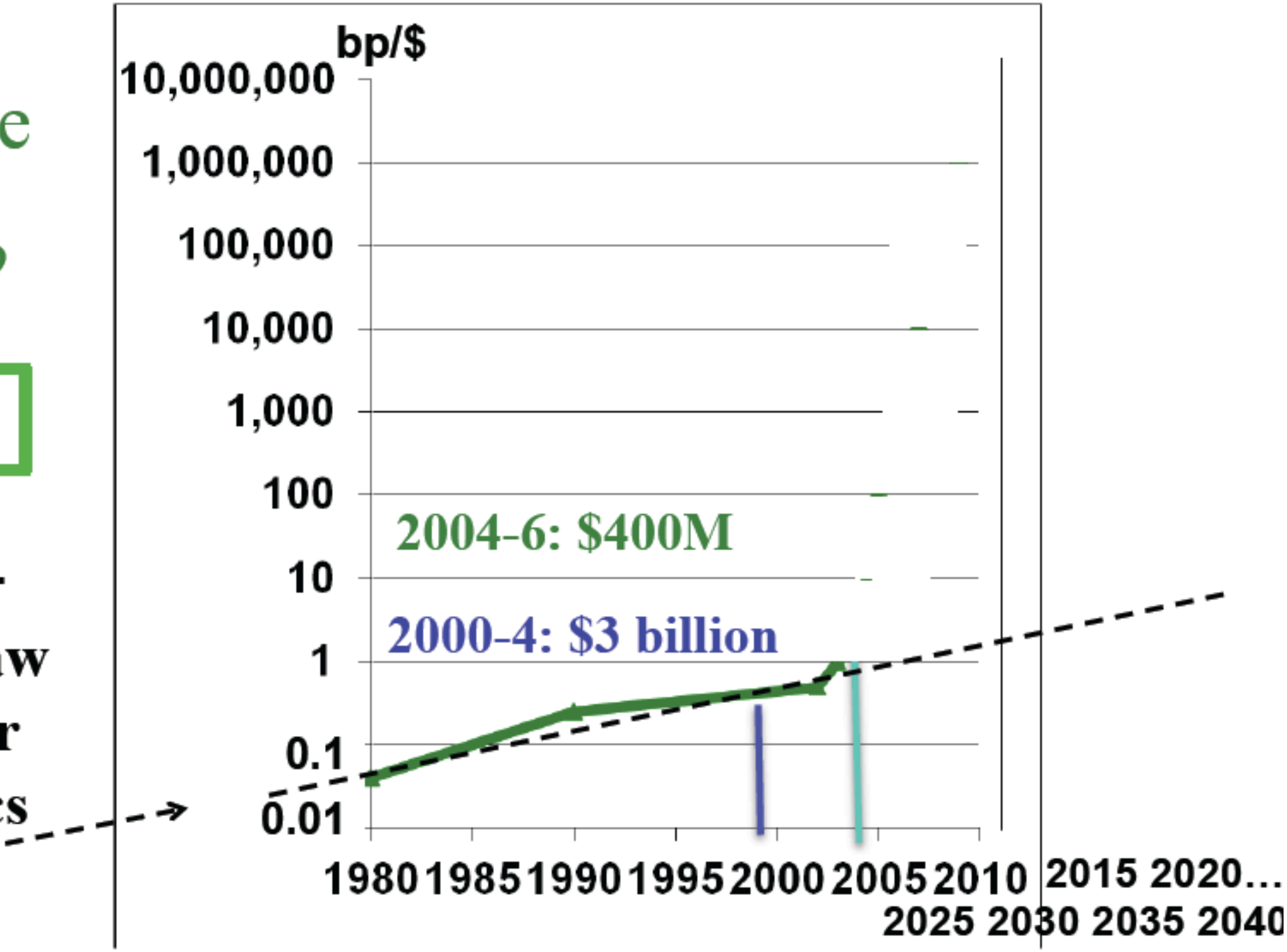
$1000 Genome

When?

2040

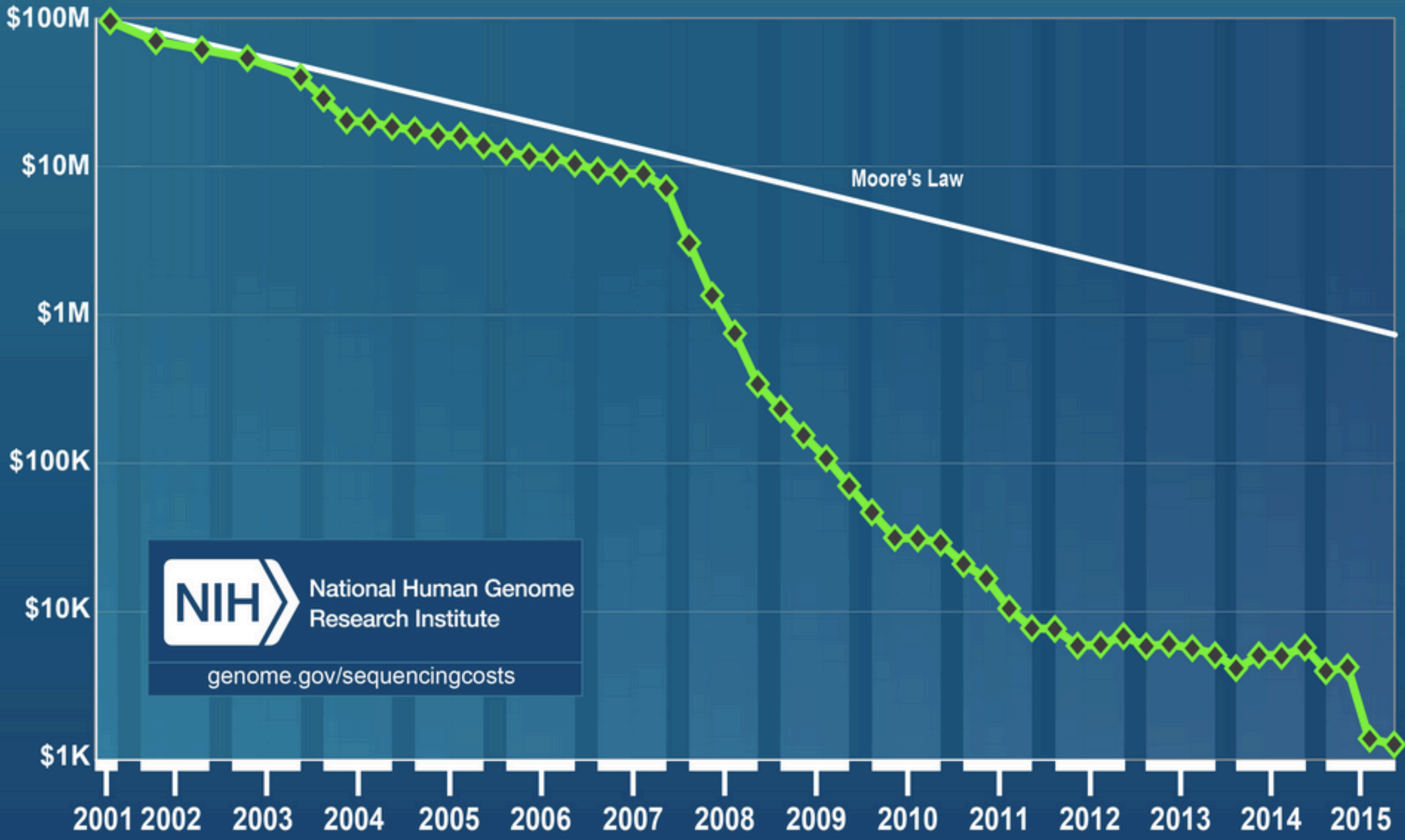
- - - - - -

Moore’ s law 1.5x/yr for electronics



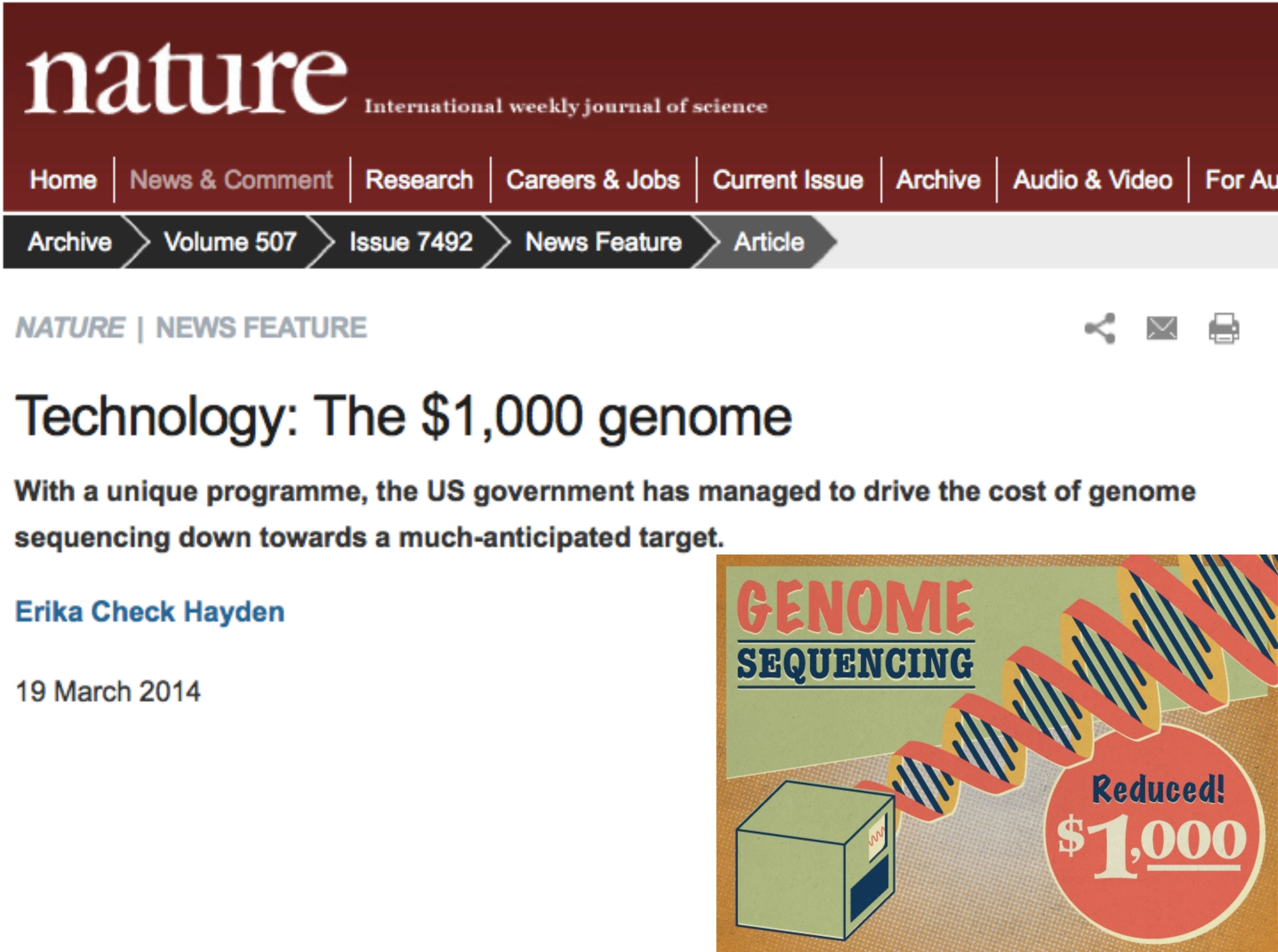
George Church

Cost per Genome



National Human Genome Research Institute

genome.gov/sequencingcosts



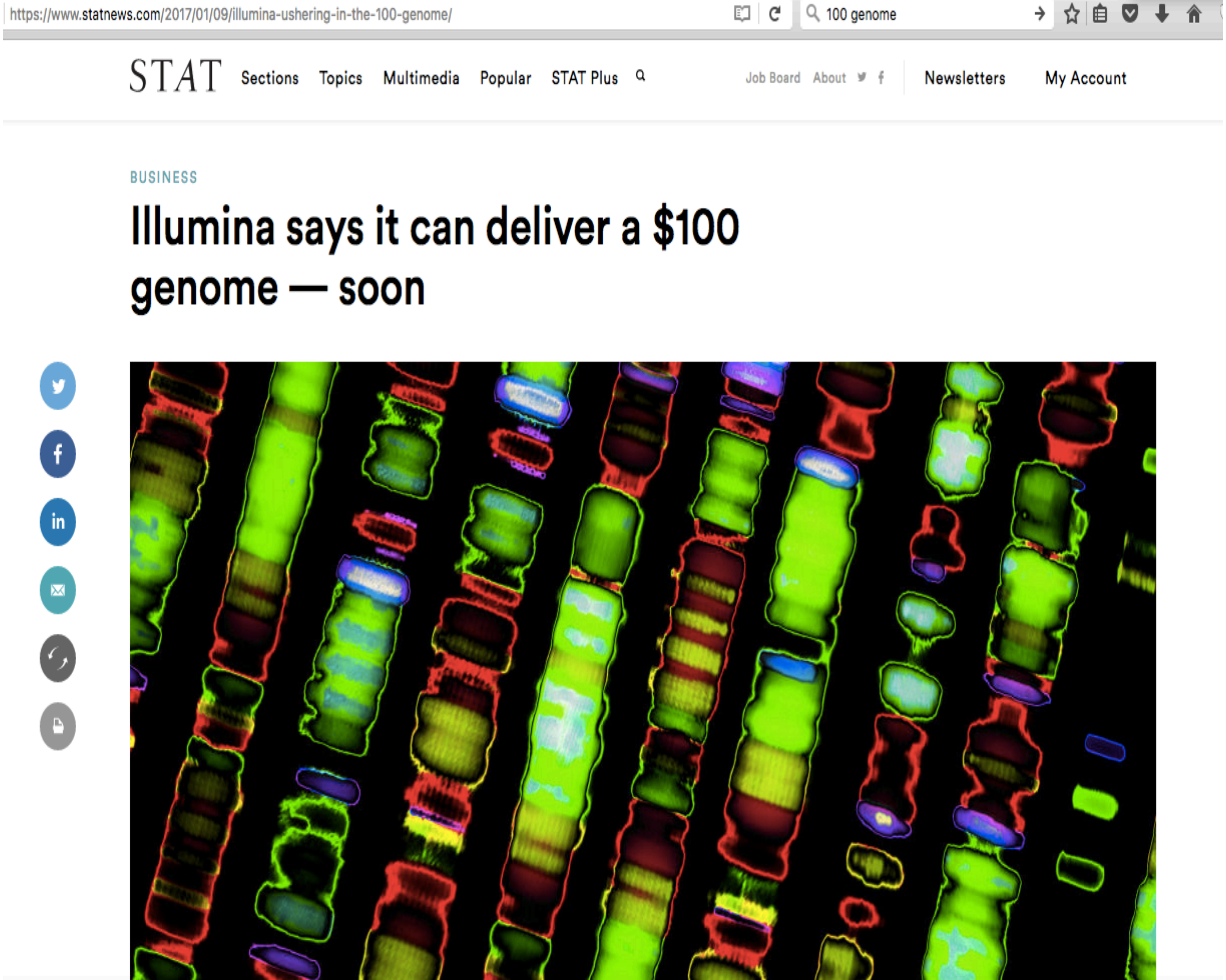
NATURE | NEWS FEATURE

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.

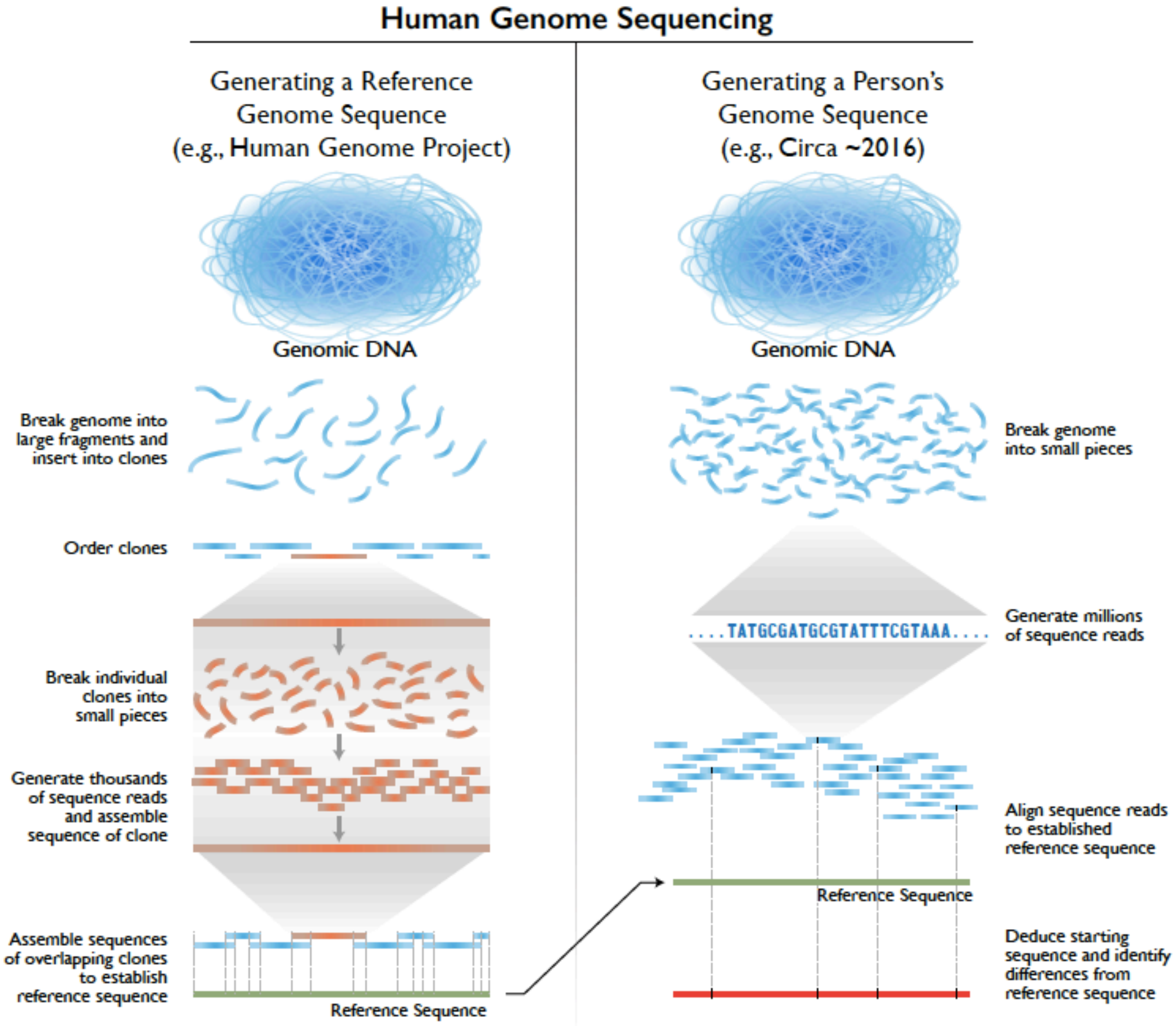
Erika Check Hayden

19 March 2014



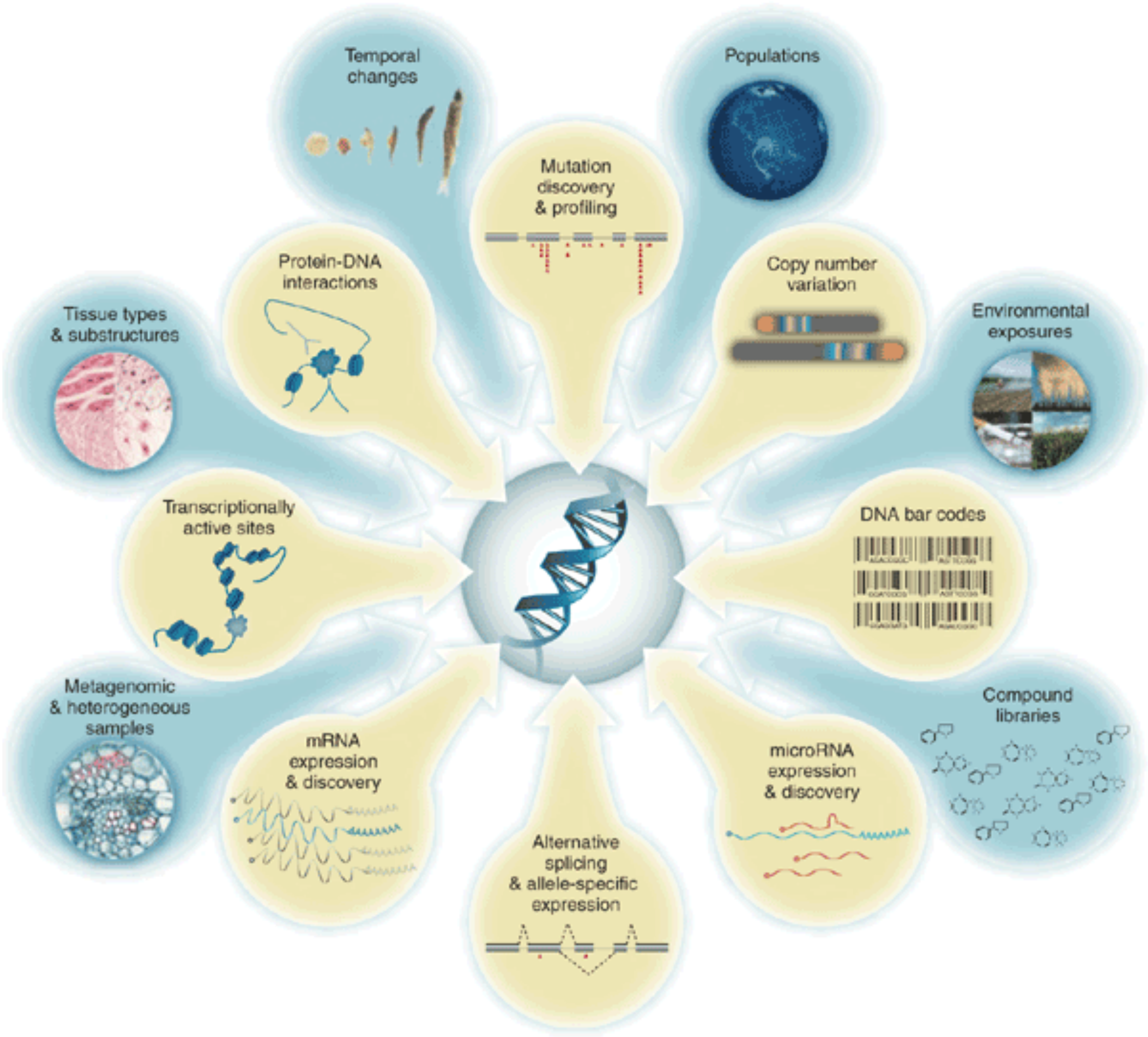
BUSINESS

Illumina says it can deliver a $100 genome — soon


Human Genome Sequencing
Generating a Reference Genome Sequence (e.g., Human Genome Project)
Generating a Person's Genome Sequence (e.g., Circa ~2016)
Genomic DNA
Genomic DNA
Break genome into large fragments and insert into clones
Break genome into small pieces
Order clones
Generate millions of sequence reads
....TATGCGATGCGTATTTCGTAAA....
Break individual clones into small pieces
Generate thousands of sequence reads and assemble sequence of clone
Align sequence reads to established reference sequence
Reference Sequence
Assemble sequences of overlapping clones to establish reference sequence
Deduce starting sequence and identify differences from reference sequence
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

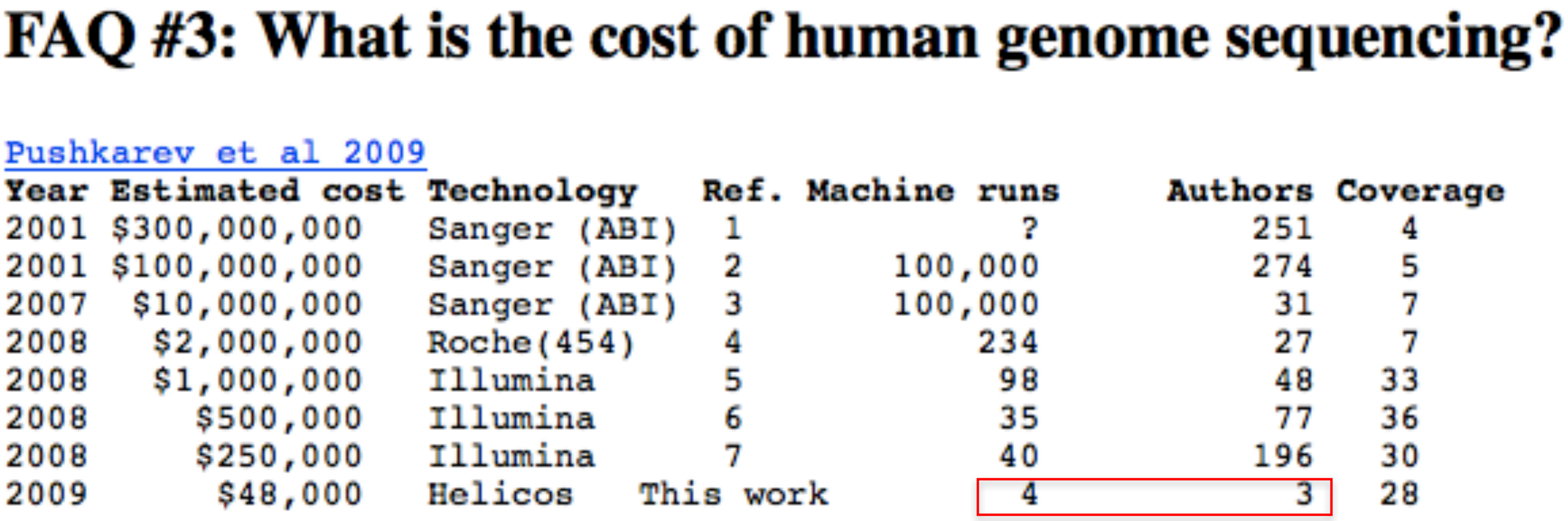
Kahvejian, 2008

The future is already here;

 it's just not evenly distributed.

—William Gibson

NGS has also enabled a democratization of the genomes by 2009, making it personal and ubiquitous



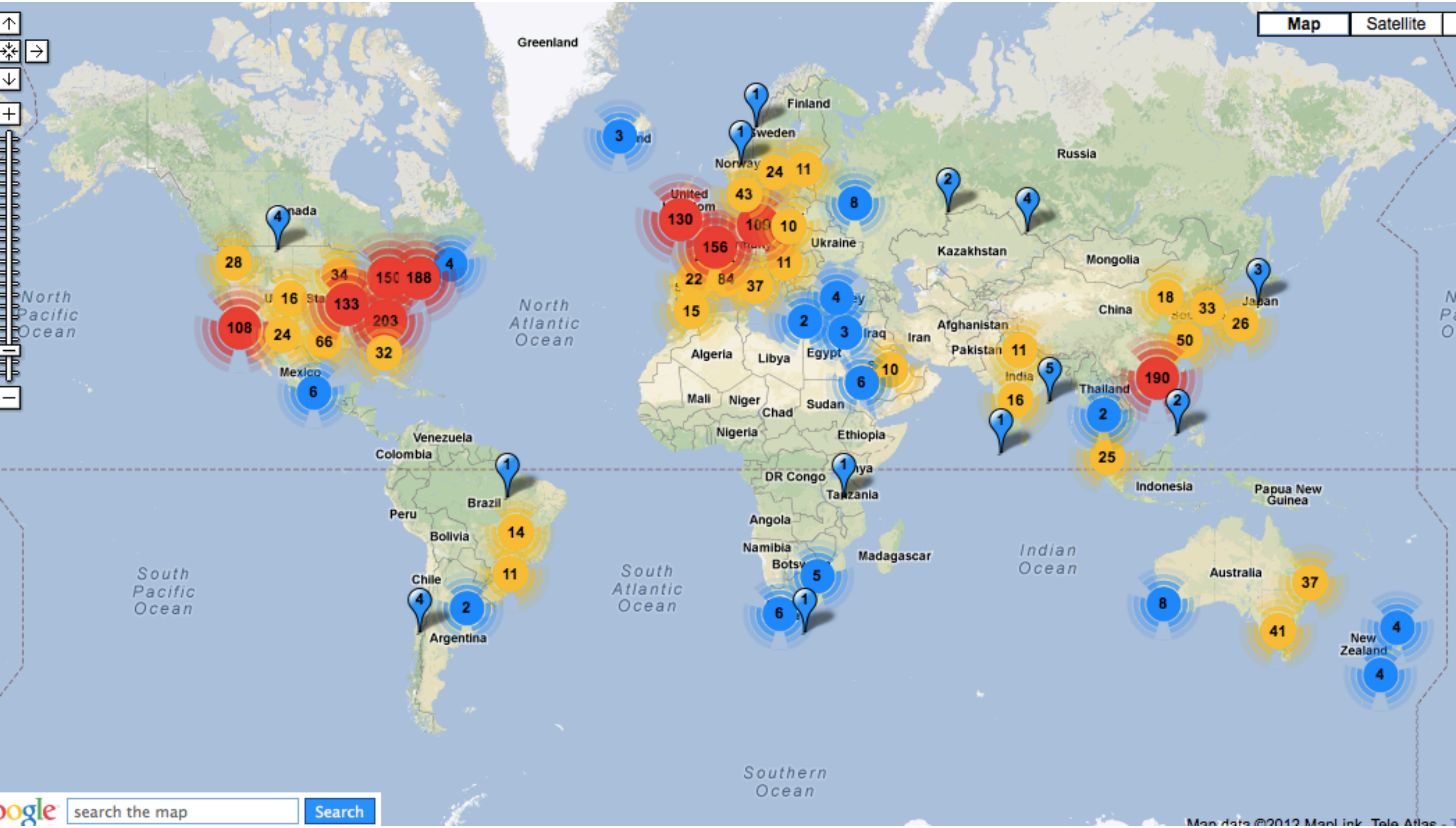
FAQ #3: What is the cost of human genome sequencing?

Pushkarev et al 2009

Year	Estimated cost	Technology	Ref.	Machine runs	Authors	Coverage
2001	$300,000,000	Sanger (ABI)	1	?	251	4
2001	$100,000,000	Sanger (ABI)	2	100,000	274	5
2007	$10,000,000	Sanger (ABI)	3	100,000	31	7
2008	$2,000,000	Roche(454)	4	234	27	7
2008	$1,000,000	Illumina	5	98	48	33
2008	$500,000	Illumina	6	35	77	36
2008	$250,000	Illumina	7	40	196	30
2009	$48,000	Helicos	This work	4	3	28

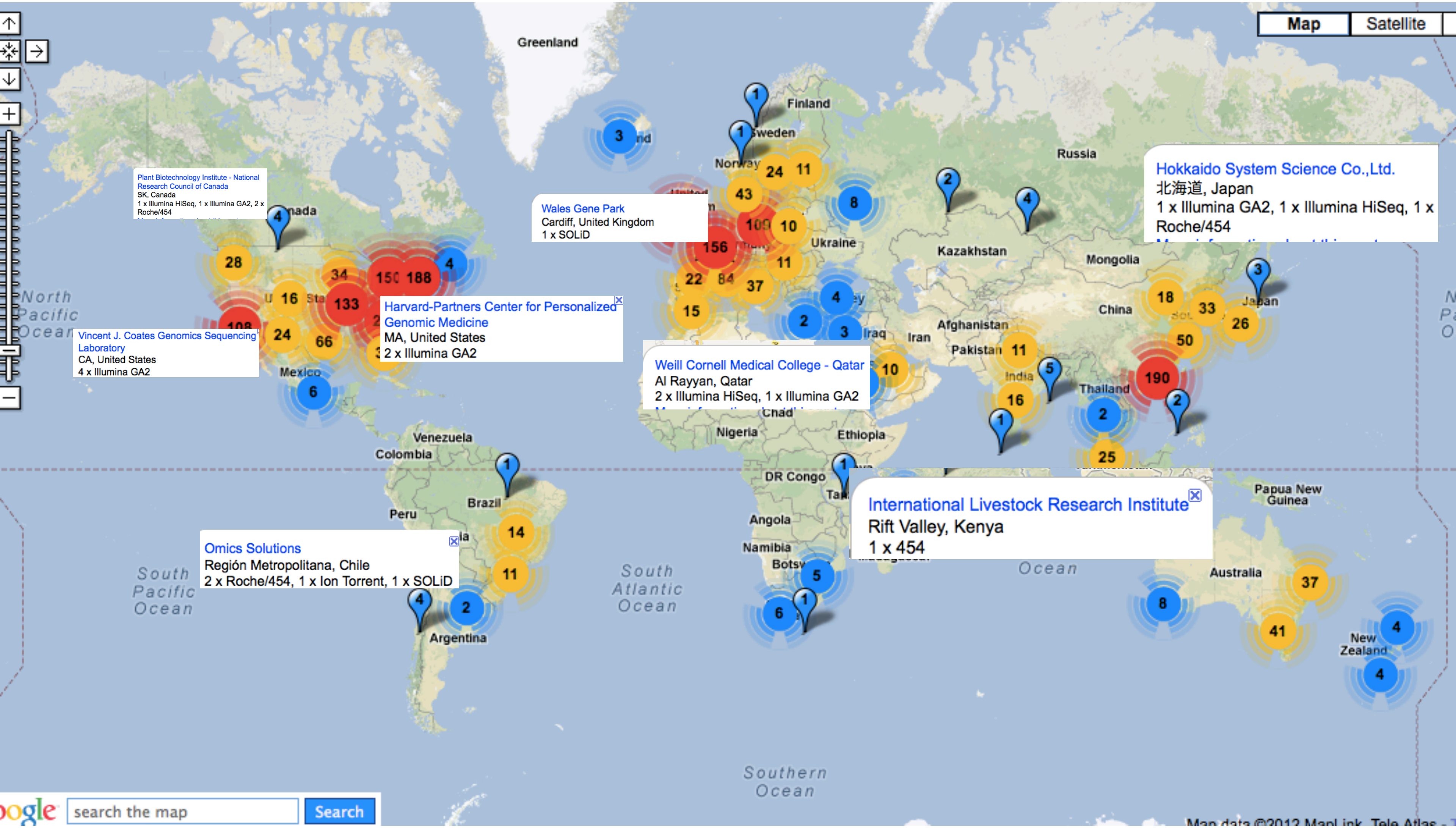
http://arep.med.harvard.edu/gmc/nexgen.html

NGS sites are globally distributed



http://omicsmaps.com/

And cover a wide range of applications in academia, government, and industry



Plant Biotechnology Institute - National Research Council of Canada
SK, Canada
1 x Illumina HiSeq, 1 x Illumina GA2, 2 x Roche/454

Wales Gene Park
Cardiff, United Kingdom
1 x SOLiD

Hokkaido System Science Co.,Ltd.
北海道, Japan
1 x Illumina GA2, 1 x Illumina HiSeq, 1 x Roche/454

Harvard-Partners Center for Personalized Genomic Medicine
MA, United States
2 x Illumina GA2

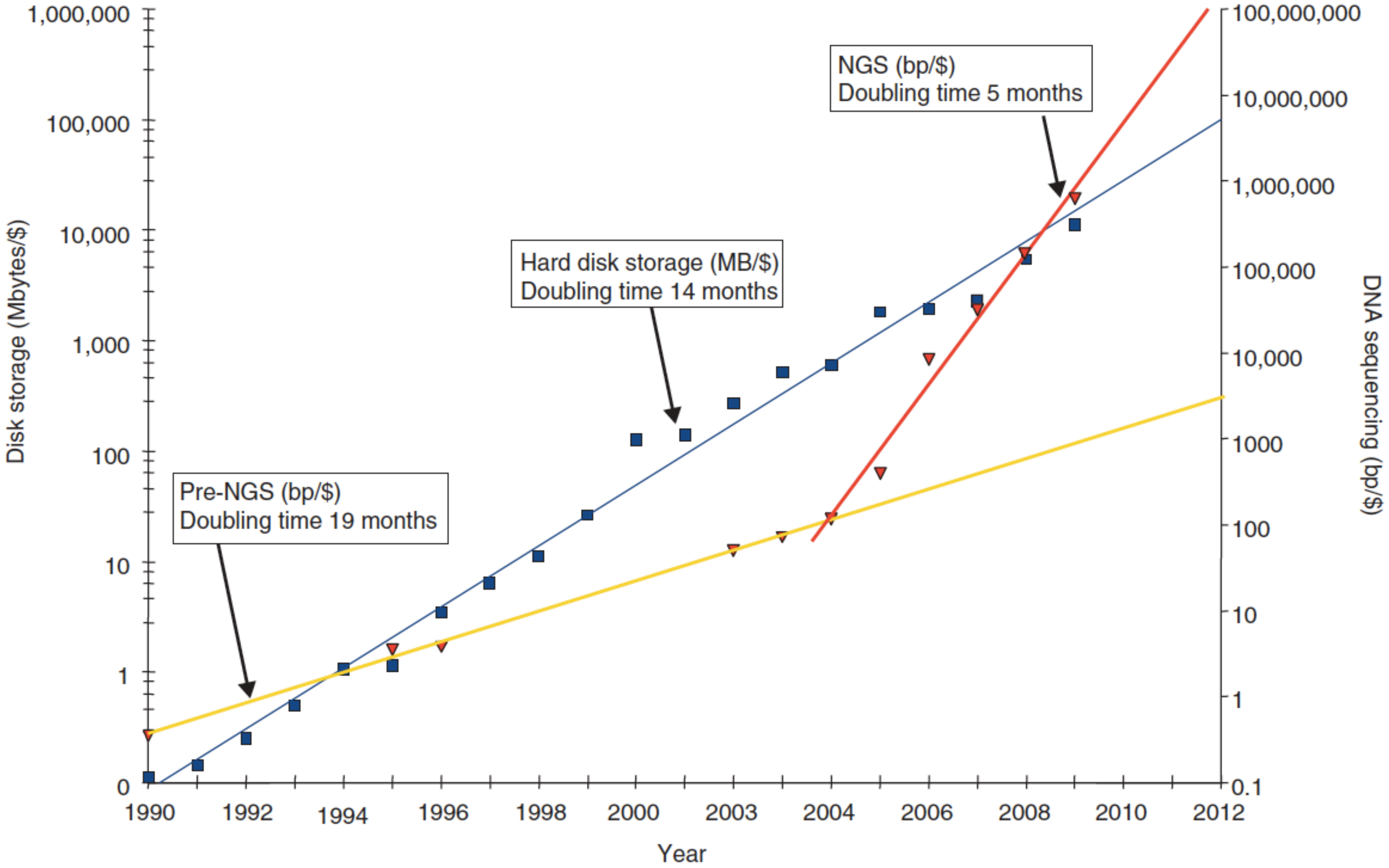
Vincent J. Coates Genomics Sequencing Laboratory
CA, United States
4 x Illumina GA2

Weill Cornell Medical College - Qatar
Al Rayyan, Qatar
2 x Illumina HiSeq, 1 x Illumina GA2

International Livestock Research Institute
Rift Valley, Kenya
1 x 454

Omics Solutions
Región Metropolitana, Chile
2 x Roche/454, 1 x Ion Torrent, 1 x SOLiD

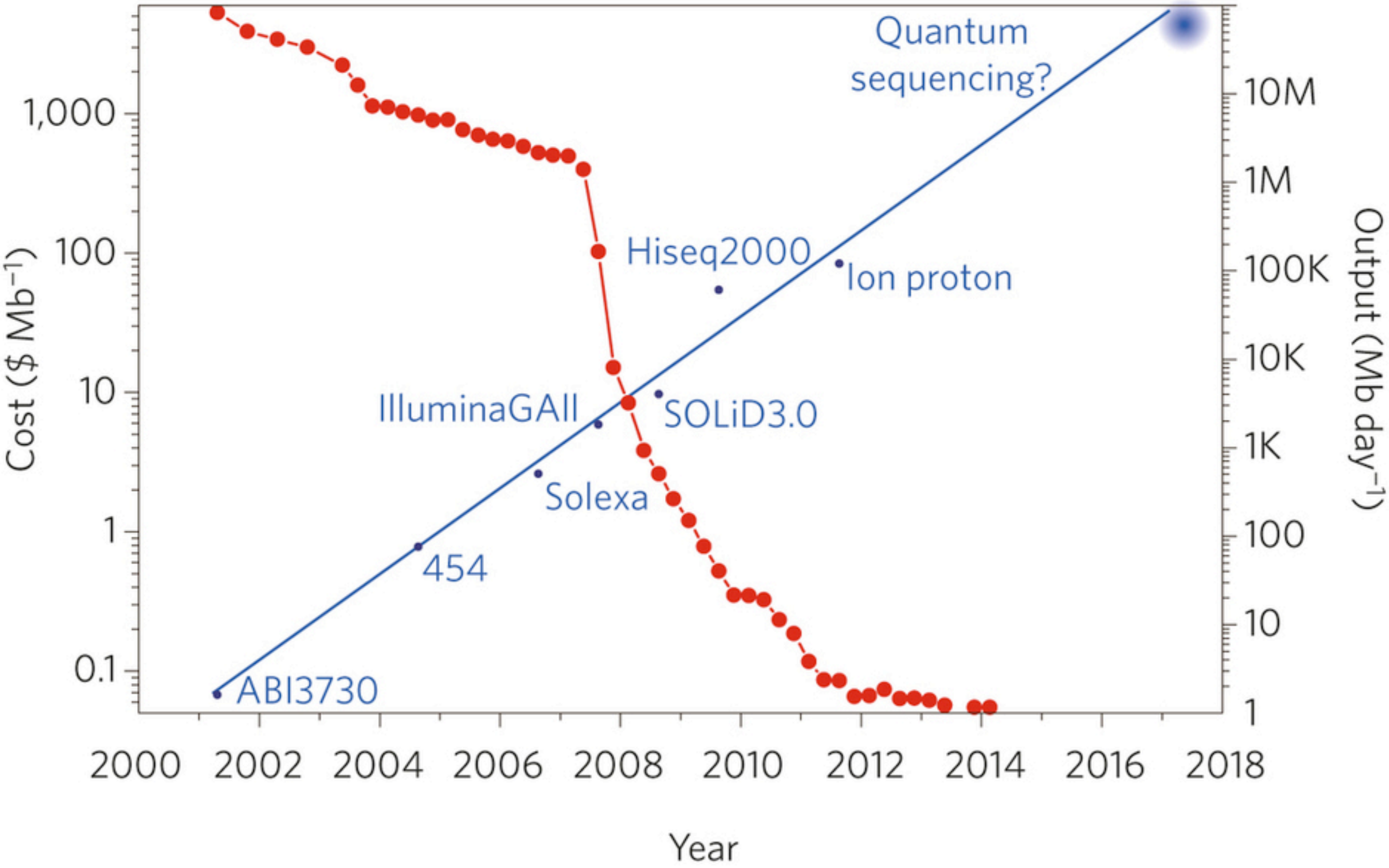
But, hard drive space is not keeping pace, creating a phalanx of companies aimed at the cloud



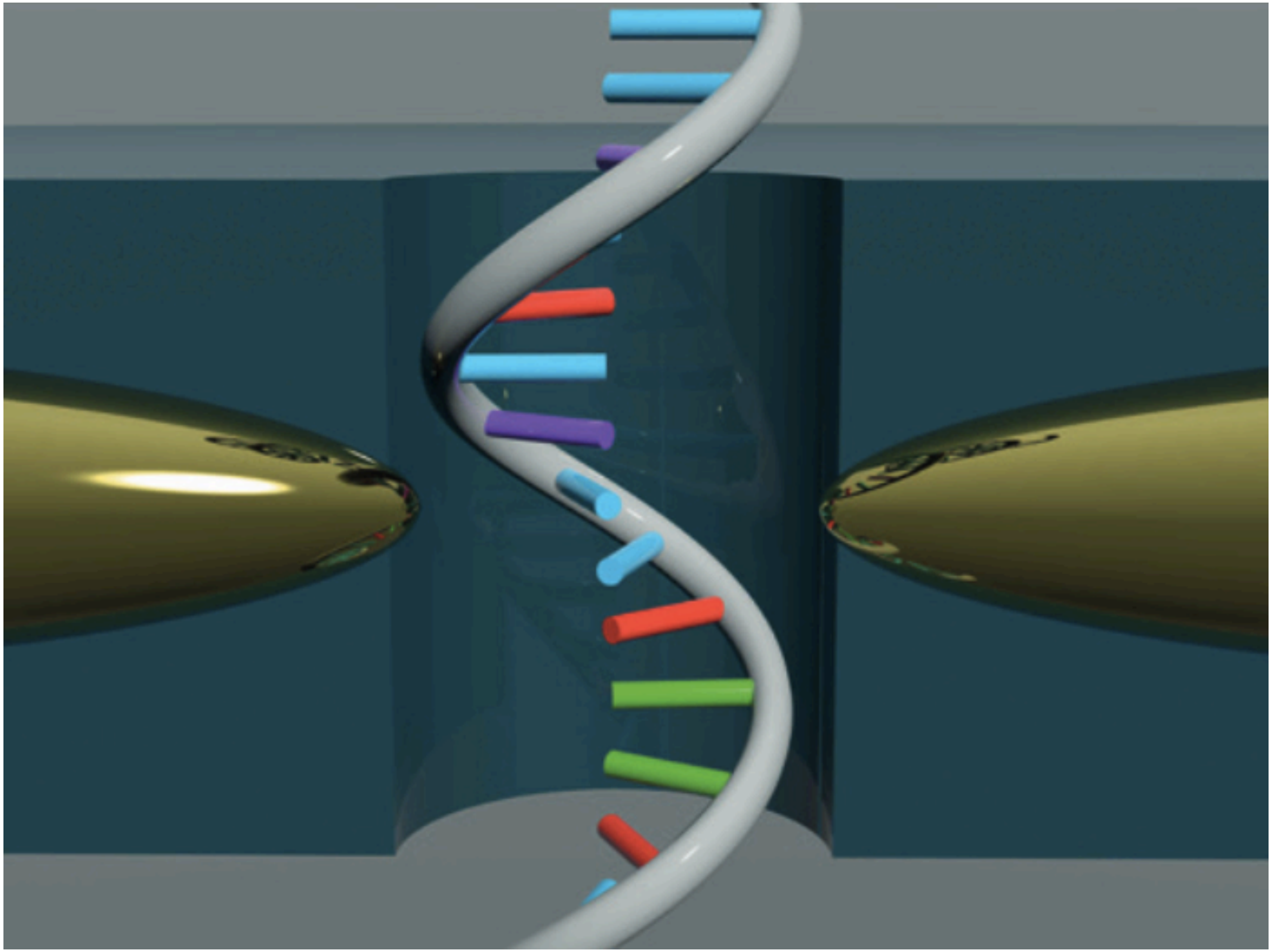
Does a $1,000 genome need a $100,000 interpretation? At least a big phone bill.

Genome, NGS, and Clinical Standards Groups	Genome in a Bottle Consortium GeT-RM Genetic Testing Reference Materials PROGRAM ABRF-NGS Nex-StoCT II Next Generation Sequencing-Standardization of Clinical Testing 2 DAYS of INFORMATICS HARMONY Oct. 11-12, 2012 Atlanta, GA THE HUMAN VARIOME PROJECT FDA-SEQC ARCHON GENOMICS X PRIZE PRESENTED BY EXPRESS SCRIPTS® CMS.gov Centers for Medicare & Medicaid Services cap
Personalized Medicine using Variant Annotation and Contextualization	GenomeQuest Counsyl Omicia Cypher Genomics GEDI Gene Discovery Genetic Disorder Personalis™ Knome® FOUNDATION MEDICINE INGENUITY® SYSTEMS geospiza® 23andMe Ve!P Oncotator Variant Analysis™ ancestry.com®
Cloud-based Approaches to Informatics and Sequencing	globus online Gene Pattern GENOMESPACE Galaxy BlueSEQ BGI Cloud DNAnexus CAGI aspera Sage BIONETWORKS Yabi @YabiProject Yabi makes High P workflow design. U atmospheric resear Perth, Western Austr WINGS Taverna
Patient and Data Sharing Initiatives for Treatment	genomera BE THE MATCH® QS Quantified Self self knowledge through numbers Personal Genome Project patientslikeme® DIYgenomics MatchingDonors Pulse OF THE PATIENT PMC Personalized Medicine Coalition genomes unzipped public personal genomics CureTogether AltheaHealth WhatNext™

Quantum sequencing?



Tunneling to measure base changes



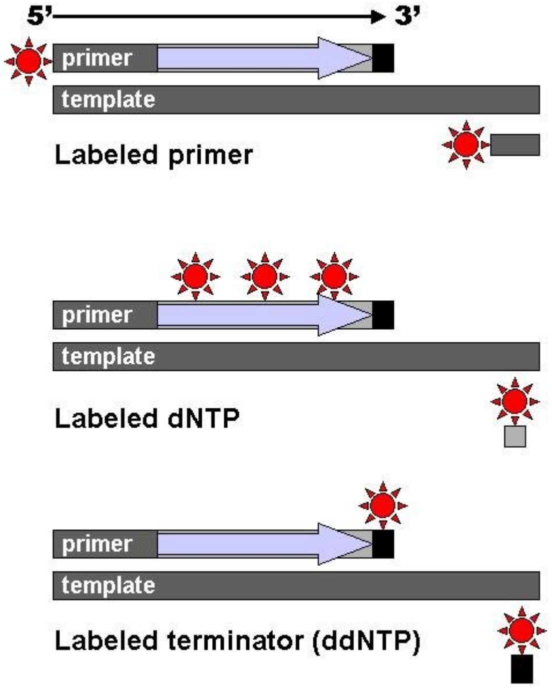
Sequencing Technologies

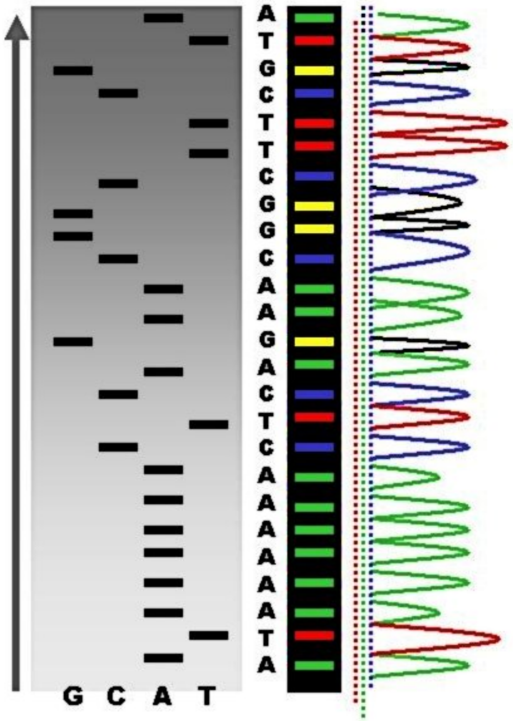
1. “Old School” dye-terminator sequencing (Sanger). 300-1000bp

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

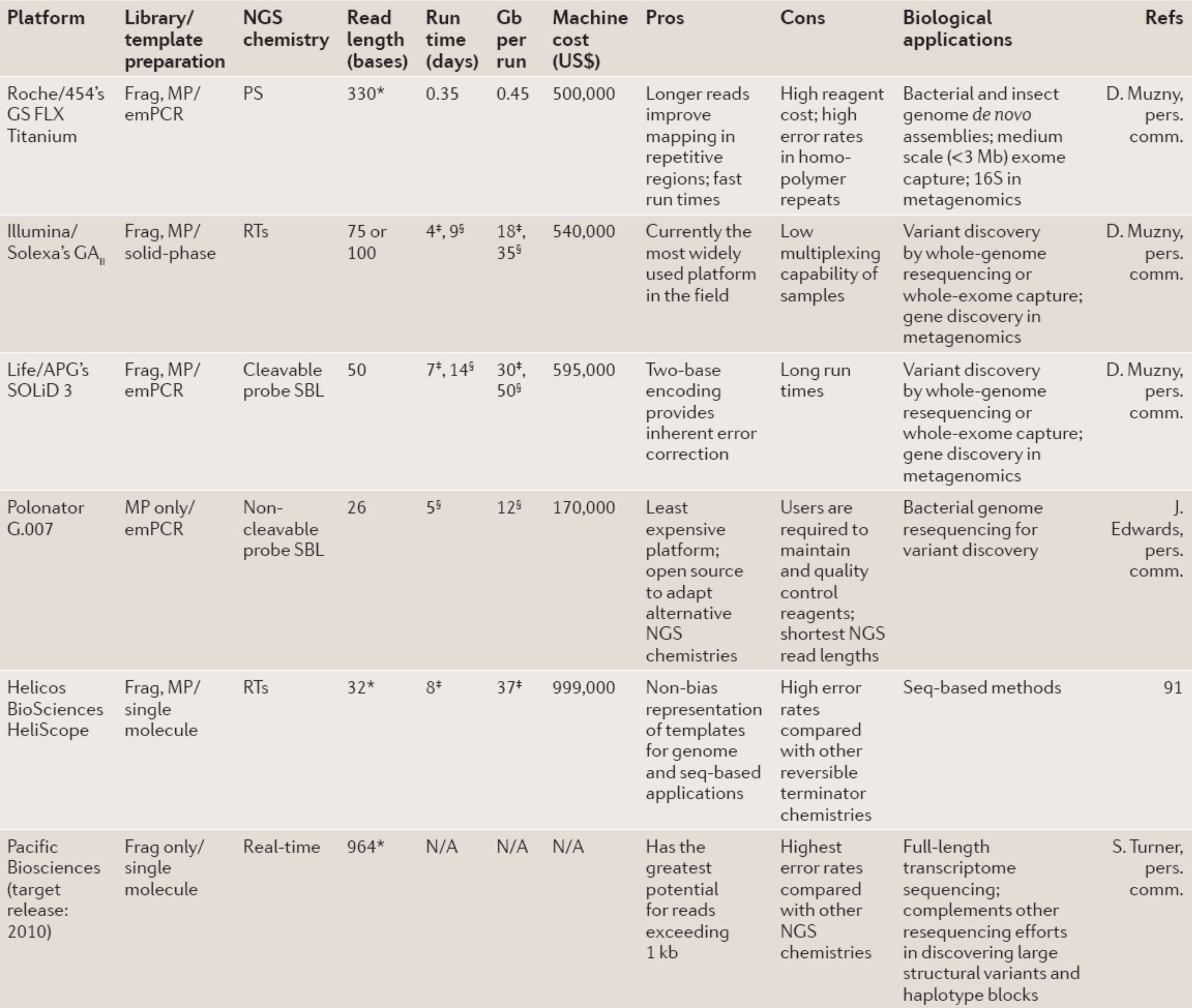
Sequencing Technologies

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





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

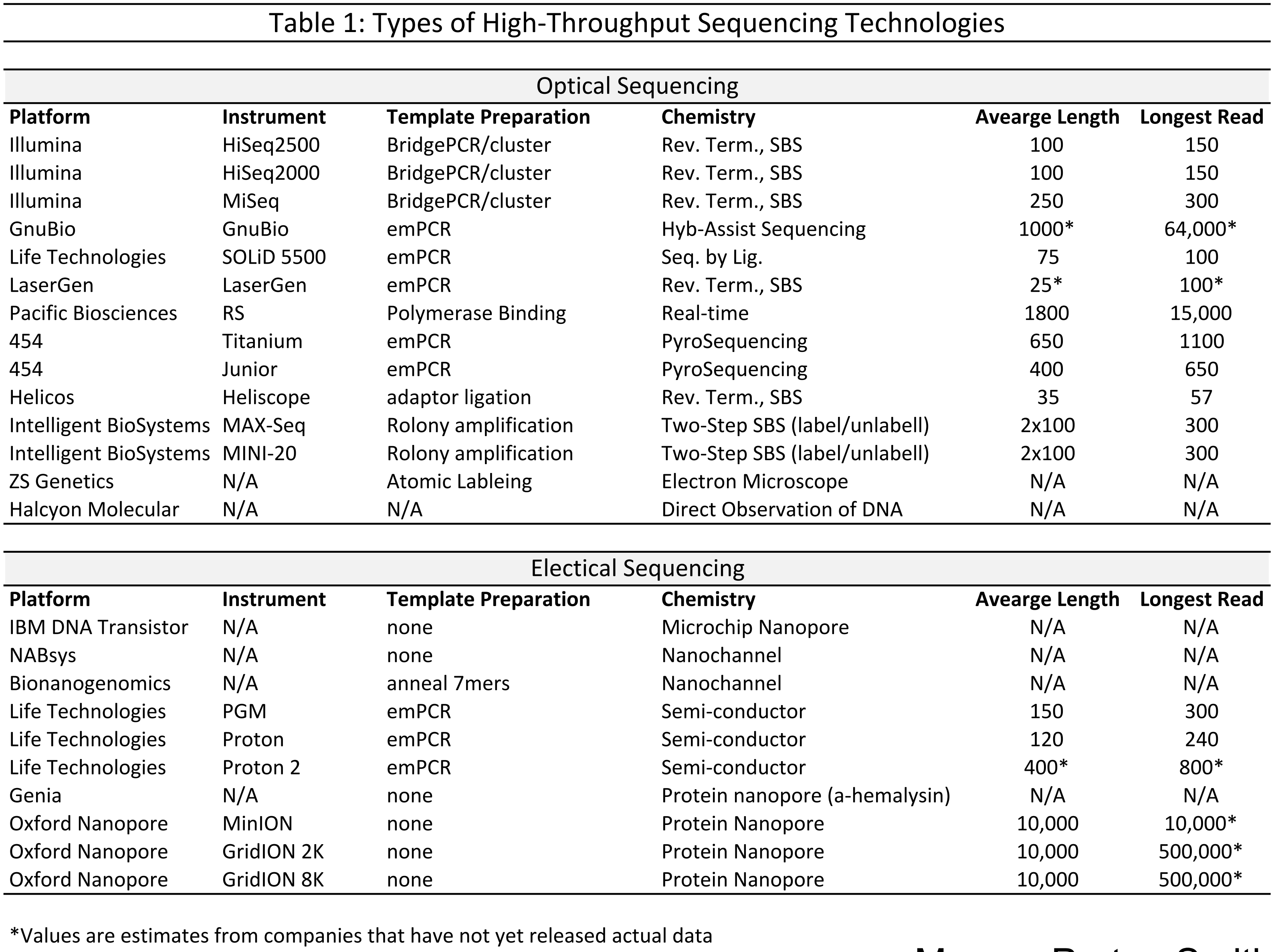


Table 1: Types of High-Throughput Sequencing Technologies

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.

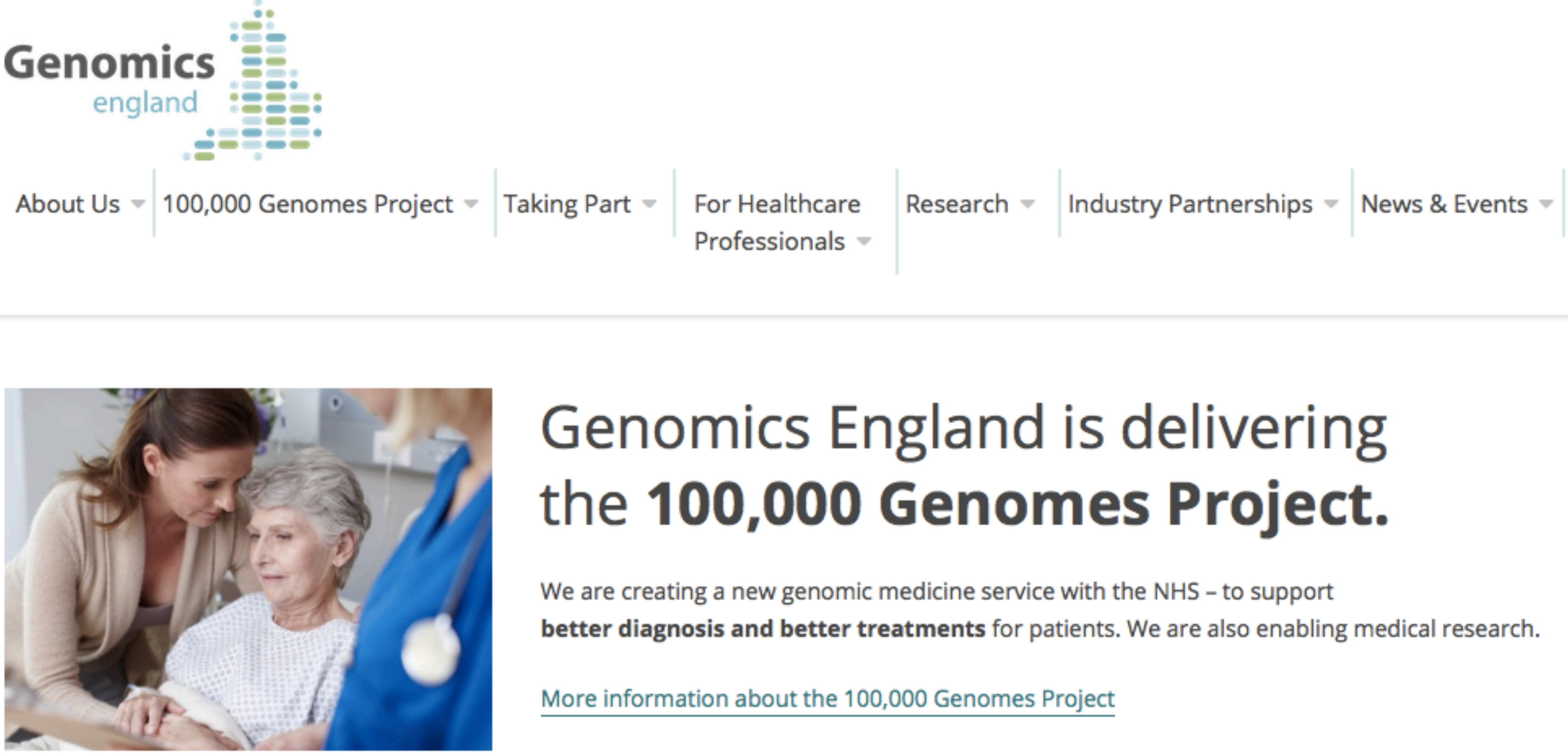
Costs vary widely, some unknown

2018 Sequencing Costs Per Platform

Chemistry	Company	Release	Instrument	Notes	Instrument	Run Time (h)	wells / pores / clusters / channels	active wells / pores / cluster	PassFilter Reads ----------- ------ Active Pores	Output / Sequence Site or Pore	Mean Read Length	Mb / Run	Gb / Run	Raw Cost / Run ($)	Reagent Cost /Gb ($)	Cost / 30X Human Genome ($)
ExAmp	Illumina	Q1 2017	NovaSeq6000	6Tb run (dual FC S4)	$950,000	48	20,000,000,000	100%	############	1.00	300	6,000,000	6,000	#########	$ 10.17	$ 915
ExAmp	Illumina	Q1 2017	NovaSeq5000	2Tb run (dual FC S2)	$850,000	60	6,800,000,000	95%	6,460,000,000	1.00	300	1,938,000	1,938	#########	$ 15.43	$ 1,389
ExAmp	Illumina	Q1 2014	X10	1Tb run	$1,000,000	72	6,200,000,000	95%	5,890,000,000	1.00	302	1,778,780	1,779	#########	$ 7.17	$ 645
ExAmp	Illumina	Q1 2015	X5	1Tb run	$1,000,000	72	6,200,000,000	95%	5,890,000,000	1.00	302	1,778,780	1,779	#########	$ 10.79	$ 971
ExAmp	Illumina	Q1 2015	HiSeq4000	Regular (v4, 1TB)	$900,000	144	5,200,000,000	97%	5,044,000,000	1.00	300	1,513,200	1,513	#########	$ 19.76	$ 1,778
TruSBS	Illumina	Q1 2012	HiSeq2500	Regular	$740,000	__	4,000,000,000	95%	3,800,000,000	1.00	250	950,000	950	#########	$ 31.47	$ 2,833
TruSBS	Illumina	Q1 2012	HiSeq2500	RapidGenome	$740,000	__	600,000,000	95%	570,000,000	1.00	300	171,000	171	$ 6,972.00	$ 40.77	$ 3,669
TruSBS	Illumina	Q1 2015	NextSeq	2x150bp run	$225,000	30	520,000,000	95%	494,000,000	1.00	300	148,200	148	$ 4,000.00	$ 26.99	$ 2,429
TruSBS	Illumina	Q1 2015	NextSeq	2x75bp run	$225,000	30	520,000,000	95%	494,000,000	1.00	150	74,100	74	$ 2,500.00	$ 33.74	$ 3,036
TruSBS	Illumina	Q1 2015	NextSeq	1x75bp run	$225,000	30	520,000,000	95%	494,000,000	1.00	75	37,050	37	$ 1,300.00	$ 35.09	$ 3,158
TruSBS	Illumina	Q1 2013	MiSeq	v2	$125,000	24	25,000,000	95%	23,750,000	1.00	500	11,875	12	$ 1,000.00	$ 84.21	$ 7,579
TruSBS	Illumina	Q1 2016	MiniSeq	v1	$49,500	24	26,000,000	95%	24,700,000	1.00	300	7,410	7	$ 1,000.00	$ 134.95	$ 12,146
Solid-state	Illumina	Q3 2017	Firefly	v1	$19,900	4	5,000,000	95%	4,750,000	1.00	300	1,425	1	$ 400.00	$ 280.70	$ 25,263
Nanopore	Genia	2019?	UNK	v1	unk	48	8,000	50%	4,000	500	5,000	10,000	7	$ 1,000.00	$ 100.00	$ 9,000
cPAS-DNB	BGI	Q1 2018	MGISEQ-2000	2x100		48					200	600,000.00	600	$ 5,000.00	$ 8.33	
cPAS-DNB	BGI	Q1 2018	MGISEQ-200	2x100	$150,000	48					200		60			
cPAS-DNB	BGI	Q1 2018	MGIFLP	2x100?												
Sanger	LifeTech	Q1 1995	3730xl	capillary/Sanger	$300,000	2	96	100%	96	1.00	750	0.07	#######	$ 90.00	$ 1,250.00	$ 112,500
IonTorrent	LifeTech	Q1 2010	PGM	318chip	$75,000	2	11,000,000	50%	5,500,000	1.00	400	2,200	2	$ 1,100.00	$ 500.00	$ 45,000
IonTorrent	LifeTech	Q3 2012	Proton	Proton 1	$225,000	2	100,000,000	65%	65,000,000	1.00	120	7,800	8	$ 1,525.00	$ 195.51	$ 17,596
IonTorrent	LifeTech	Q3 2015	S5 / S5XL	520 chip	$50,000	2	5,000,000	95%	4,750,000	1.00	400	1,900	2	$ 300.00	$ 157.89	$ 14,211
IonTorrent	LifeTech	Q3 2015	S5 / S5XL	530 chip	$50,000	2	20,000,000	95%	19,000,000	1.00	400	7,600	8	$ 300.00	$ 39.47	$ 3,553
IonTorrent	LifeTech	Q3 2015	S5 / S5XL	540 chip	$50,000	2	80,000,000	95%	76,000,000	1.00	200	15,200	15	$ 300.00	$ 19.74	$ 1,776
IonTorrent	LifeTech	Q1 2015	Proton	Proton 2	$225,000	6	300,000,000	80%	240,000,000	1.00	120	28,800	29	$ 1,000.00	$ 34.72	$ 3,125
CsgG	Oxford Nanopore	Q2 2015	MinION	Min500	$500	6	512	75%	384	778	15000	4,479	4.48	$ 500.00	$ 111.63	$ 10,047
CsgG	Oxford Nanopore	Q2 2017	GridIONx5	5 pores	$125,000		2,560	75%	1,920	3,888	15,000	111,974	111.97	2,500	$ 22.33	$ 2,009
CsgG	Oxford Nanopore	Q2 2017	PrOmethION	100,000 pores	$75,000	6	98,304	75%	73,728	6,221	15000	6,879,707	6,879.71	#########	$ 4.33	$ 389
DNA Pol	PacBio	Q1 2014	RSII	C2XL (120 min)	$700,000	6	150,000	45%	67,500	1.00	11000	743	0.74	$ 150.00	$ 202.02	$ 18,182
DNA Pol	PacBio	Q1 2016	Sequel	C2XL (360 min)	$350,000	8	1,000,000	60%	600,000	1.00	11000	6,600	6.60	$ 700.00	$ 106.06	$ 9,545
DNA Pol	PacBio	Q4 2018	Sequel	v3 (P6-c4)	$350,000	8	8,000,000	35%	2,800,000	1.00	11000	30,800	30.80	$ 350.00	$ 11.36	$ 1,023
SBS	QIAGEN	Q1 2015	GeneReader	150 bp run	$225,000	33	16,000,000	95%	15,200,000	1.00	150	2,280	2.28	$ 500.00	$ 219.30	$ 19,737
Pyroseq	Roche	Q1 2007	454	FLX	$100,000	8	1,600,000	65%	1,040,000	1.00	500	520	0.52	$ 1,200.00	$ 2,307.69	$ 207,692

The $1000 genome is here!

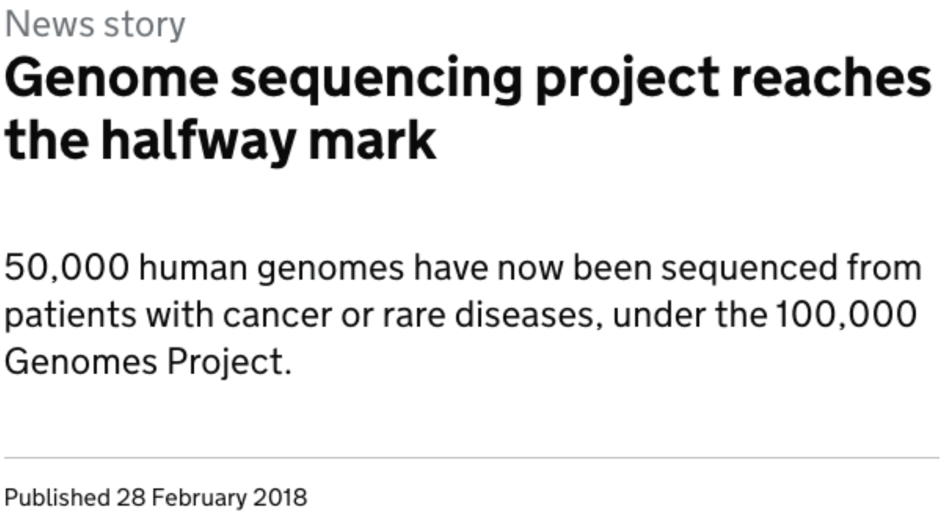
- More often ~$1100 per genome. Coming down.
- Exome sequencing costs also are dropping
- Certain platforms are better suited for certain tasks:
  - Counting applications (ChIP-Seq, RNA-Seq) need more reads
  - *De novo* assembly work needs longer reads
  - Whole genome re-sequencing requires lower errors rate and high processivity



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



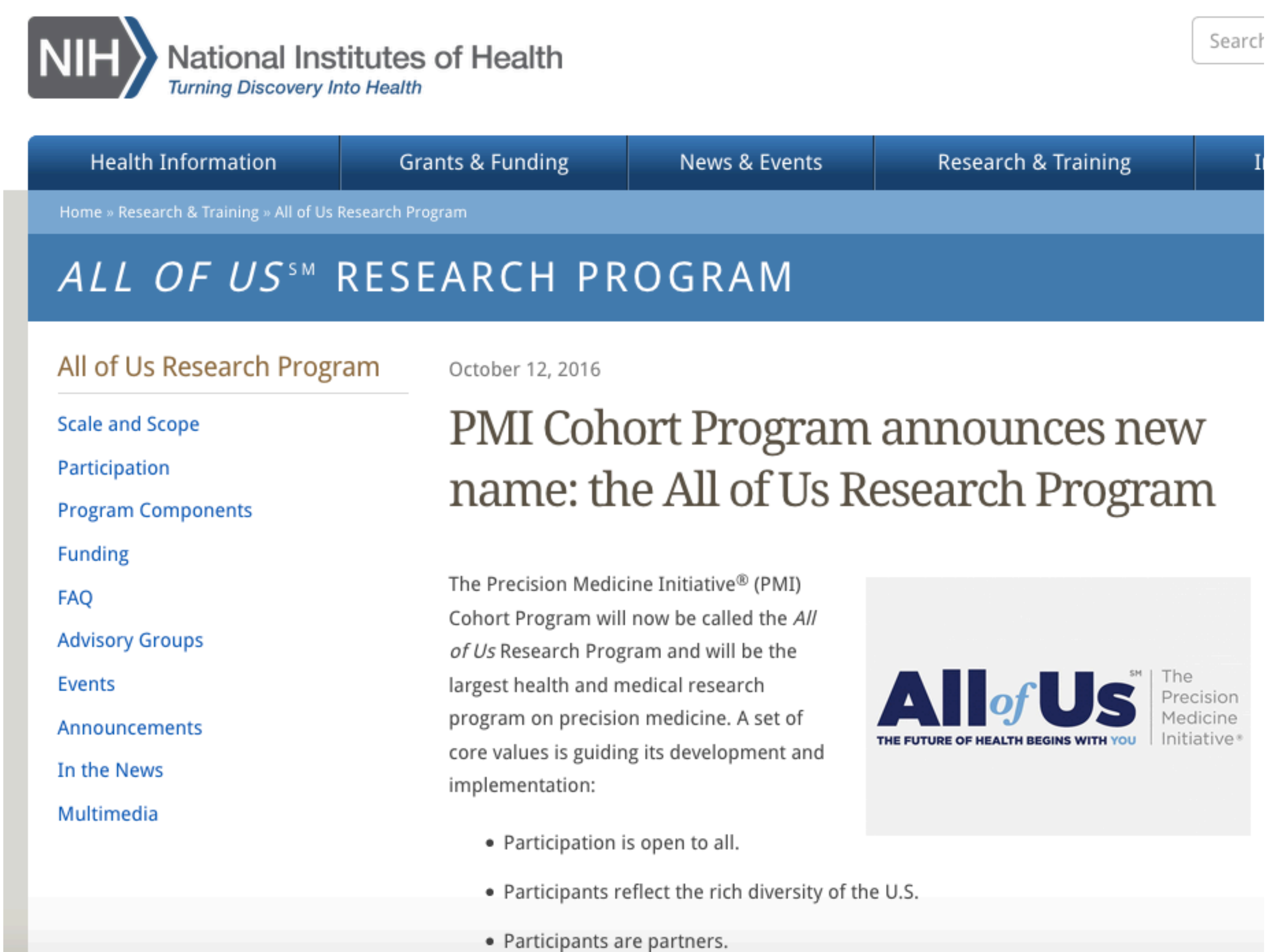
News story

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

*ALL OF US*SM RESEARCH PROGRAM

All of Us Research Program

October 12, 2016

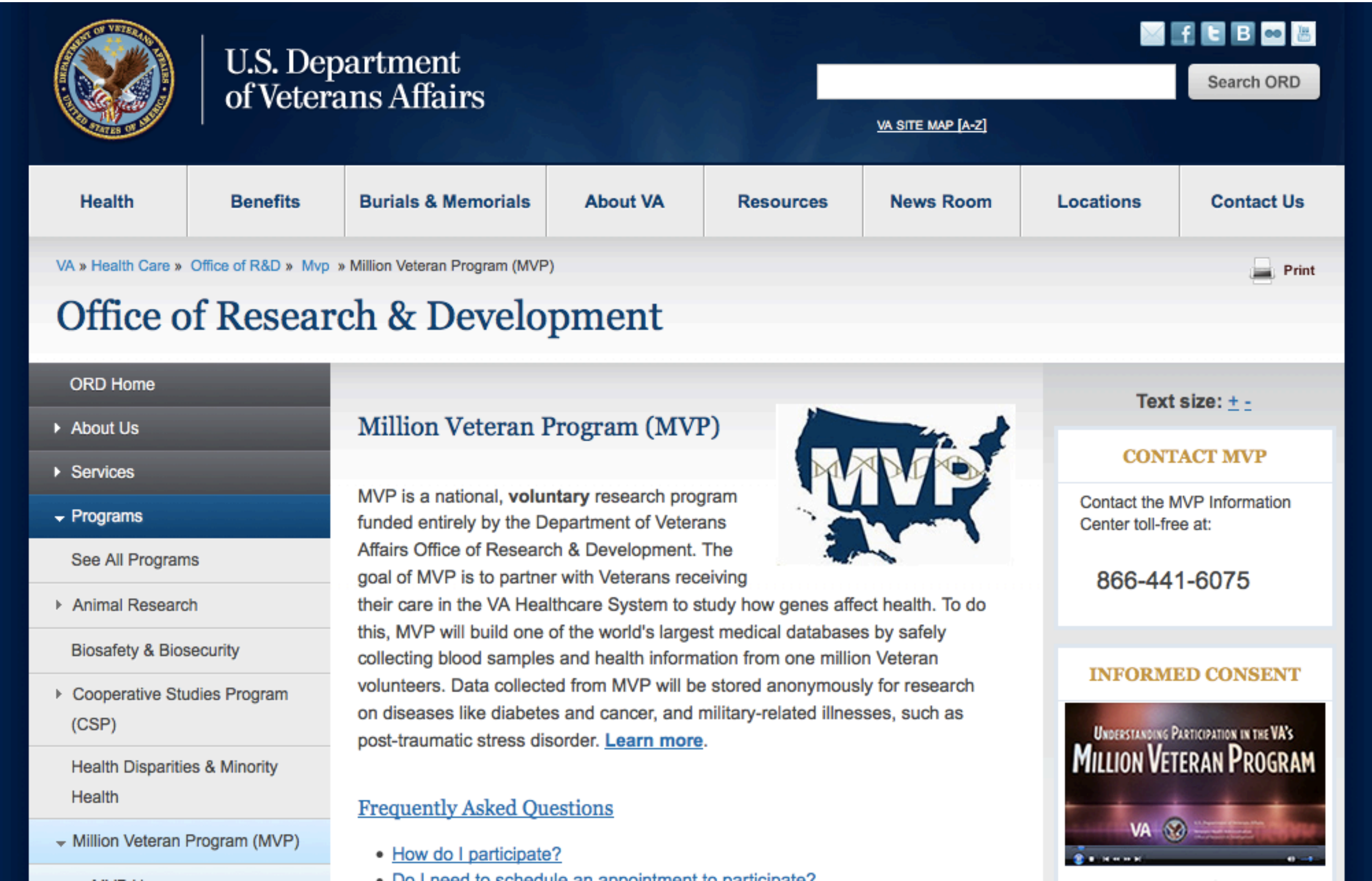
PMI Cohort Program announces new name: the All of Us Research Program

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 SM
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 too!


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
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)
Text size: + -
MVP
Understanding Participation in the VA's Million Veteran Program
VA

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.

CONTACT MVP

Contact the MVP Information Center toll-free at:

866-441-6075

INFORMED CONSENT

Frequently Asked Questions

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

A lot of genomic and medical data coming

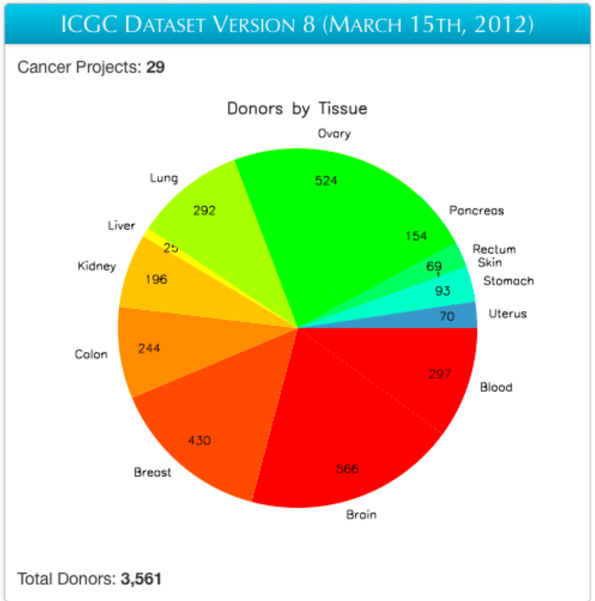
Announcements of Large Genome Consortia:

- AllOfUs – 1M U.S. Patients with medical data
- Netherlands GoNL– 250trios – preclinical (http://www.nlgenome.nl/)
- Faroer islands 100k –pre-clinical
- Qatar 300k – pre-clinical
- Iceland 2.5k – pre-clinical
- UK 100k – clinical
- Genomics Medicine Ireland (GMI) with AbbVie
- Finland, number unknown – clinical (https://www.fimm.fi/en/research/grand-challenge-programs/finnish-genome-sequencing-and-preventive-health-care)
- Poland 100K
- Swiss Genome 100K
- Geisinger Health 100K (with Regeneron)
- Astrozenica (2M with HLI)
- 1 million U.S. Veterans Project
- Newfoundland 100K

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



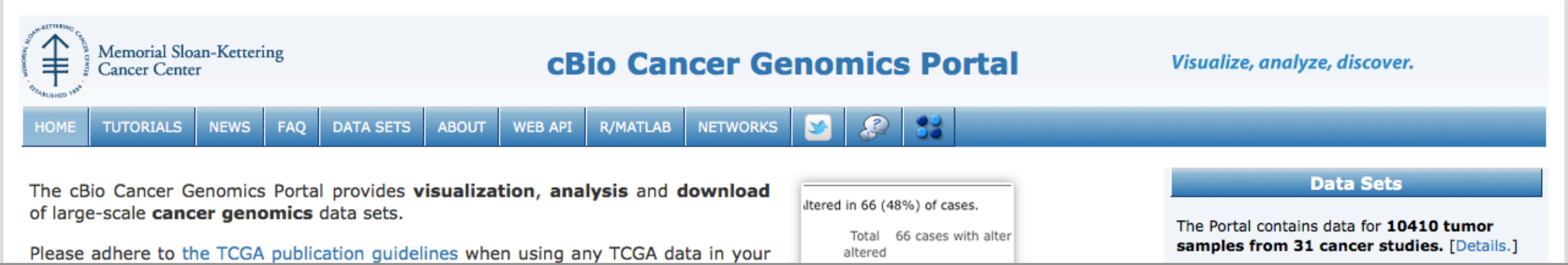
ICGC Dataset Version 8 (March 15th, 2012)

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.



cBio Cancer Genomics Portal

Visualize, analyze, discover.

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

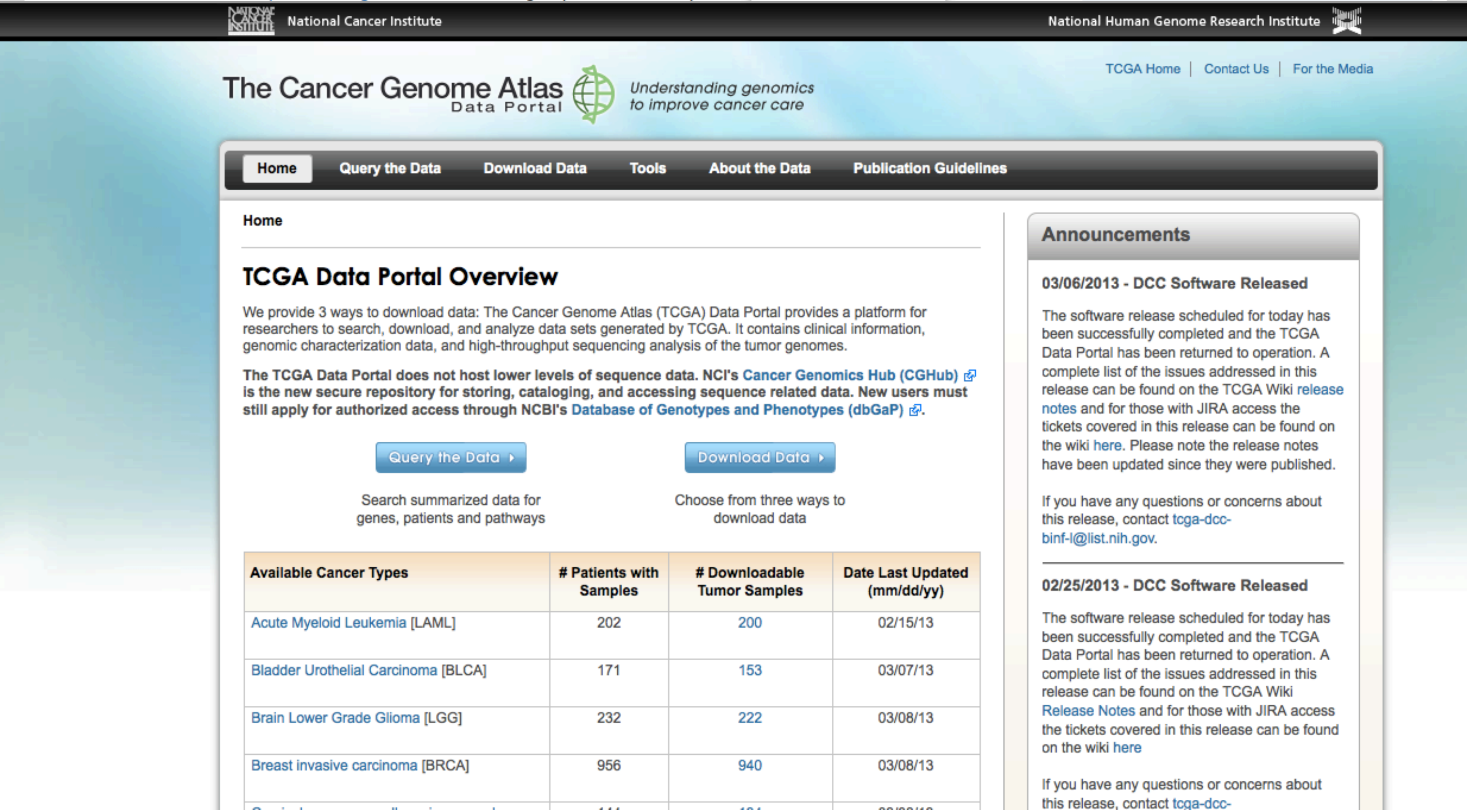
Please adhere to the TCGA publication guidelines when using any TCGA data in your

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).

Search summarized data for genes, patients and pathways

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.

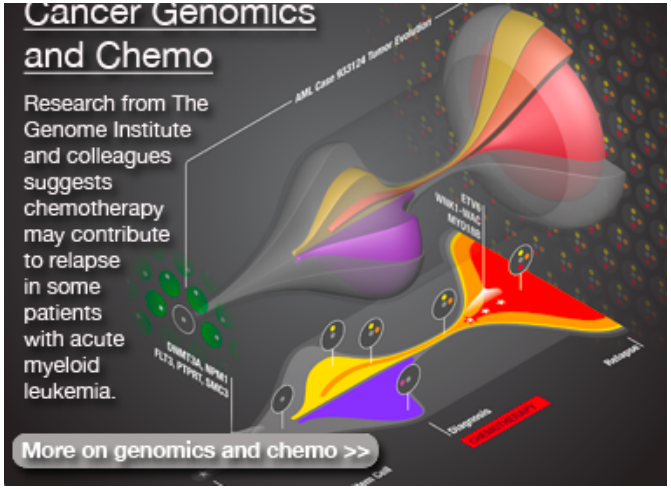
If you have any questions or concerns about this release, contact tcga-dcc-binf-l@list.nih.gov.

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-

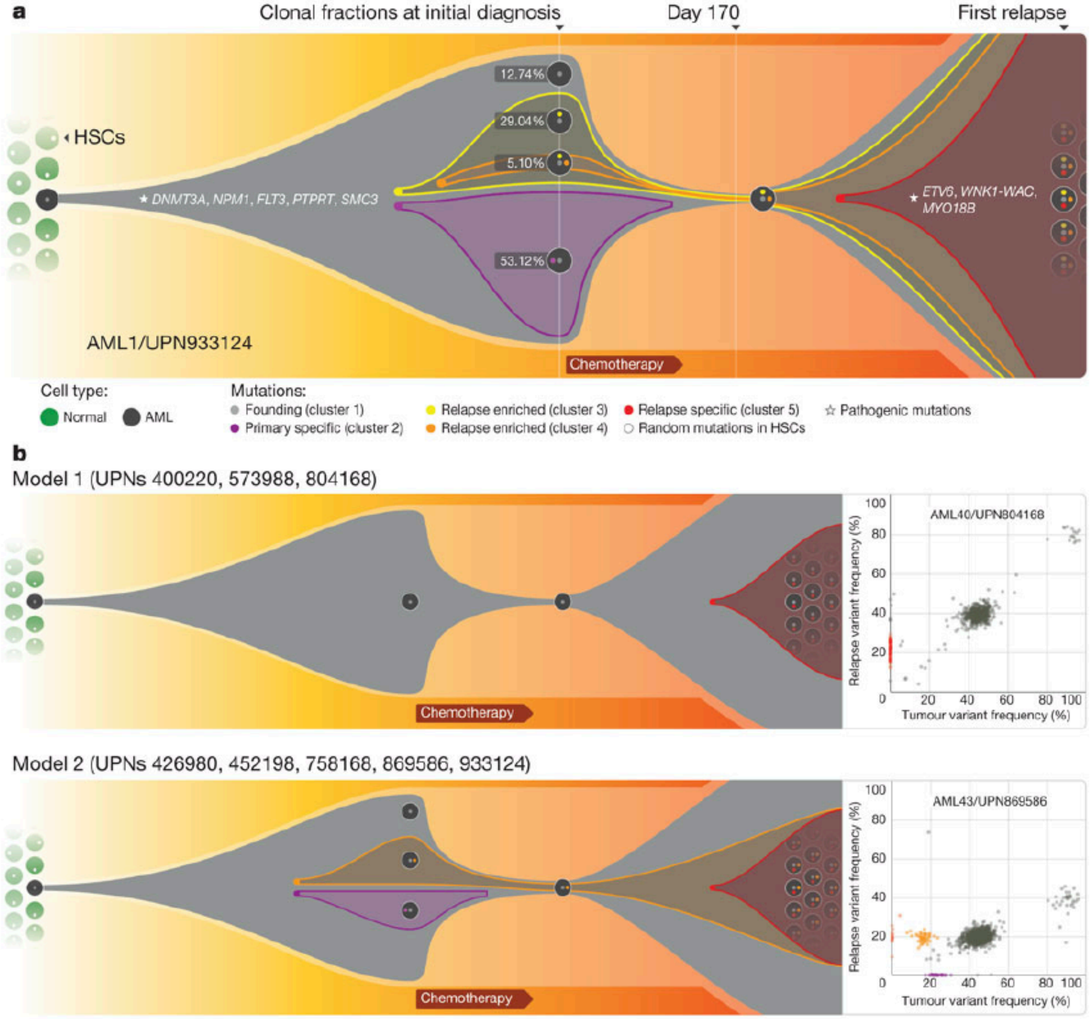
We can also observe the dynamics and evolution of cancers



Cancer Genomics and Chemo

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

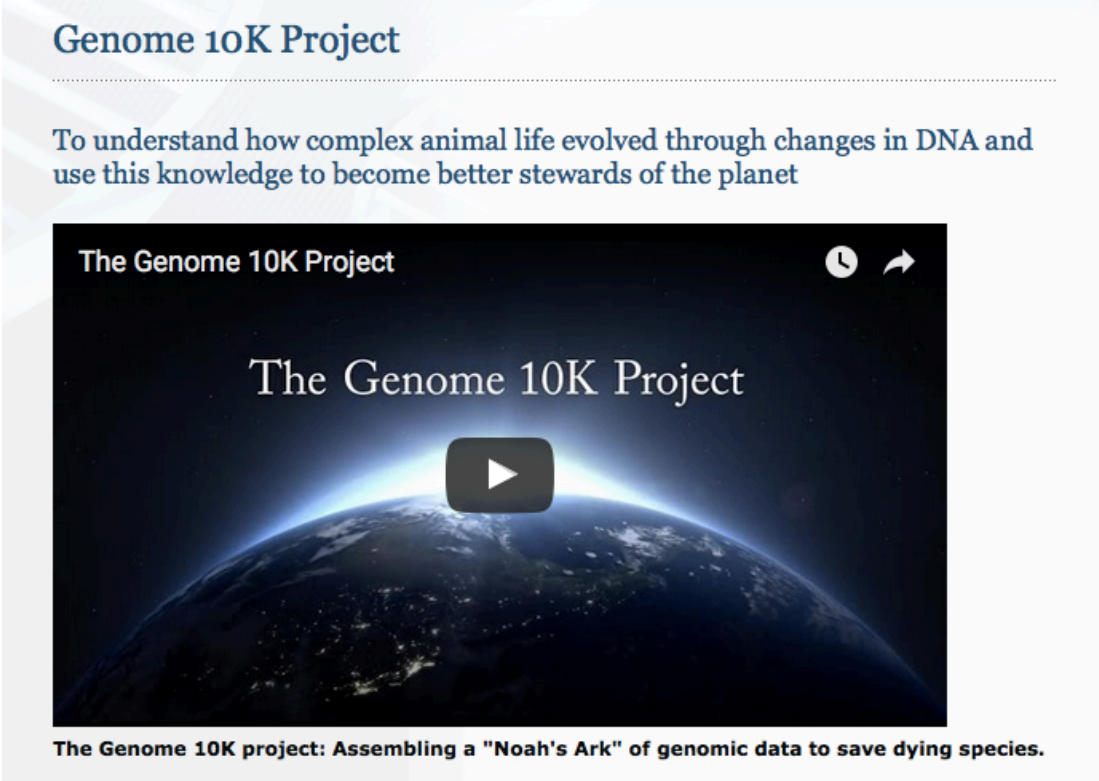
More on genomics and chemo >>



a
Clonal fractions at initial diagnosis
Day 170
First relapse
HSCs
12.74%
29.04%
5.10%
53.12%
★ DNMT3A, NPM1, FLT3, PTPRT, SMC3
★ ETV6, WNK1-WAC, MYO18B
AML1/UPN933124
Chemotherapy
Cell type:
● Normal
● AML
Mutations:
● Founding (cluster 1)
● Primary specific (cluster 2)
● Relapse enriched (cluster 3)
● Relapse enriched (cluster 4)
● Relapse specific (cluster 5)
○ Random mutations in HSCs
☆ Pathogenic mutations
b
Model 1 (UPNs 400220, 573988, 804168)
Chemotherapy
AML40/UPN804168
Relapse variant frequency (%)
Tumour variant frequency (%)
0
20
40
60
80
100
Model 2 (UPNs 426980, 452198, 758168, 869586, 933124)
Chemotherapy
AML43/UPN869586
Relapse variant frequency (%)
Tumour variant frequency (%)
0
20
40
60
80
100

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

The Genome 10K Project

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!

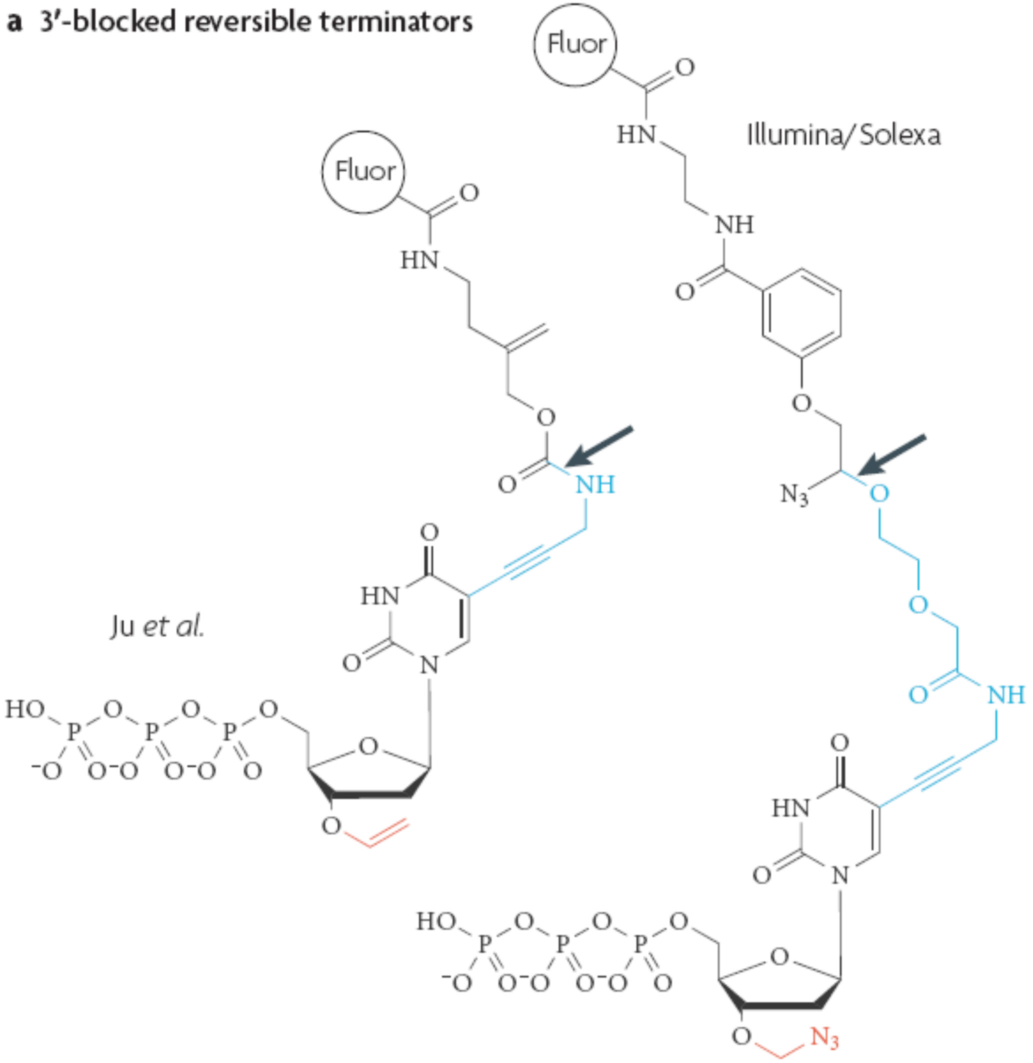
华大基因
BGI

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

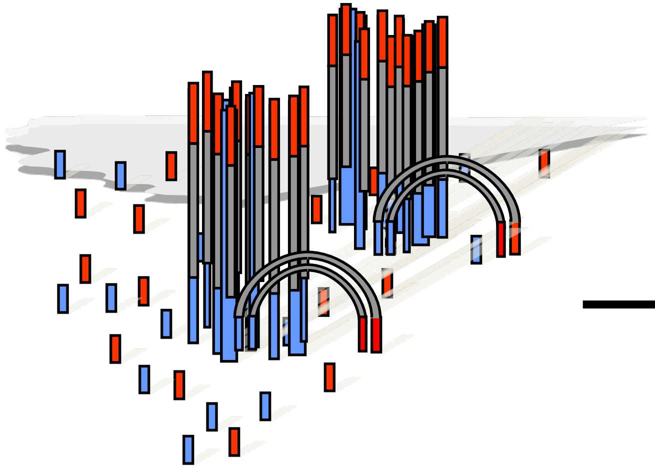


Illumina SBS Technology

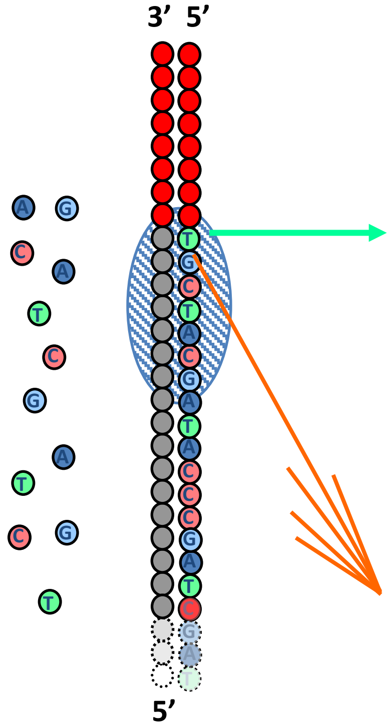
Reversible Terminator Chemistry Foundation

DNA
(0.1-1.0 ug)

Sample preparation



Cluster growth



Sequencing

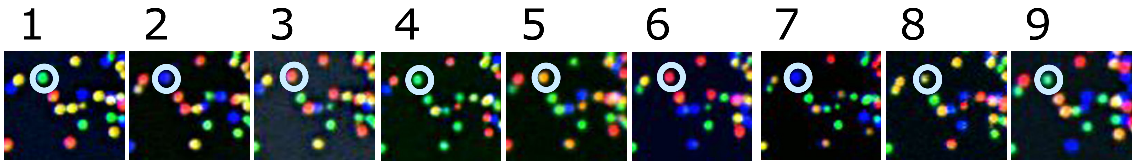
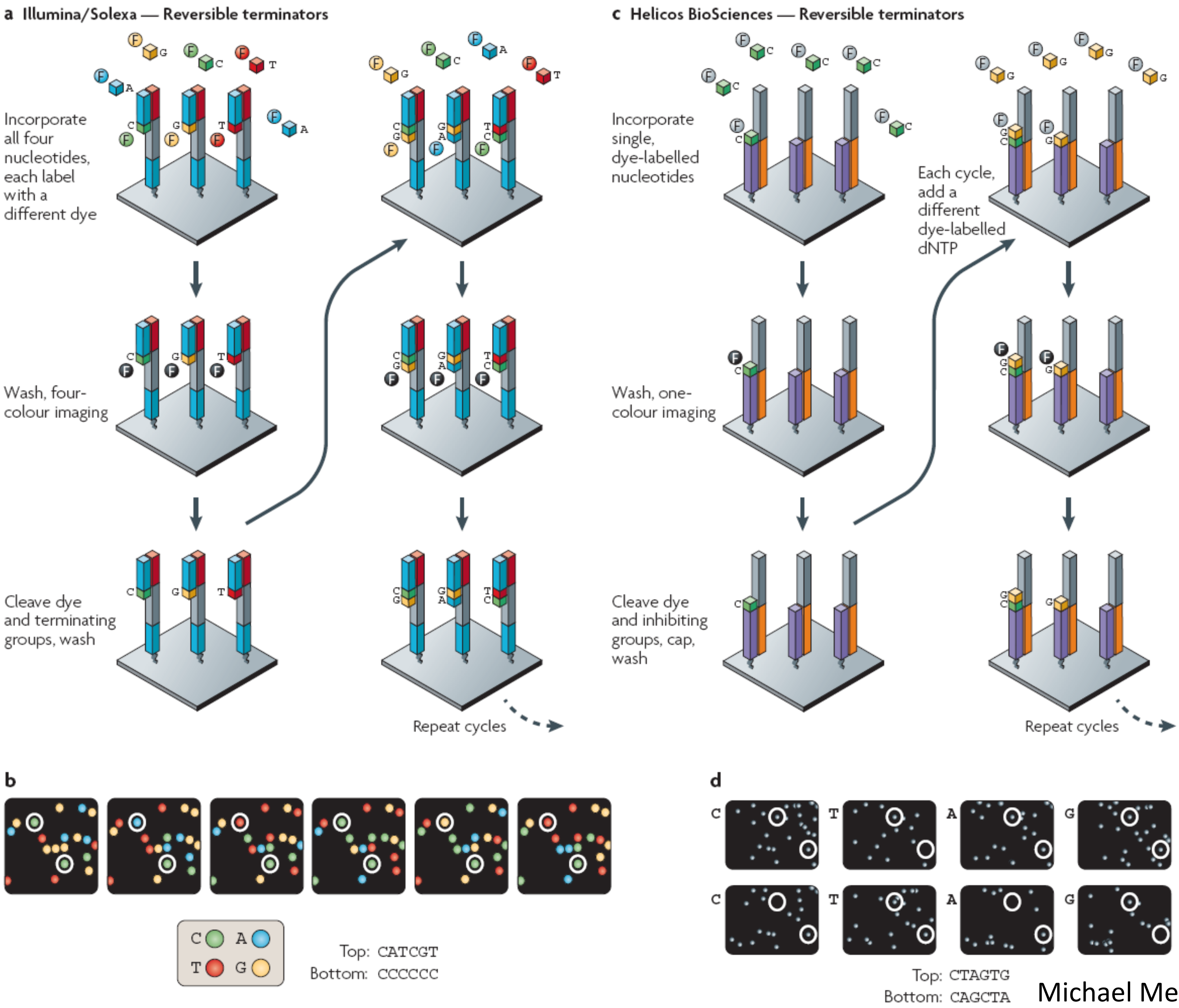


Image acquisition

Base calling

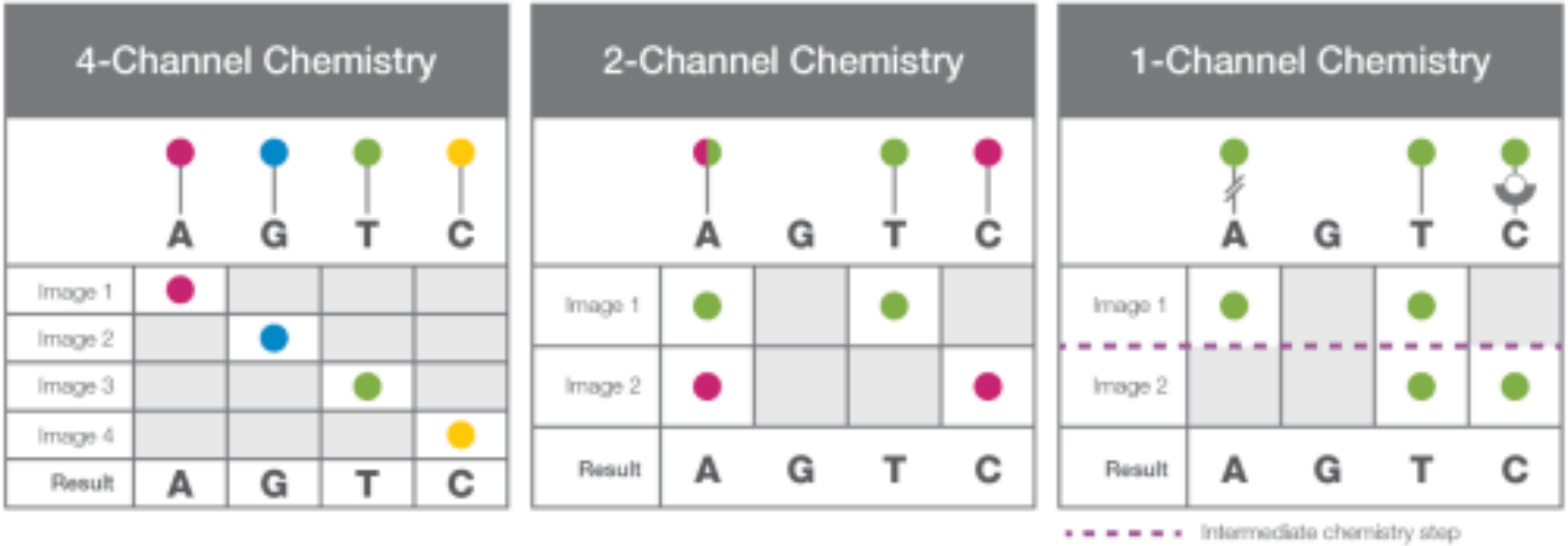
http://www.illumina.com/technology/sequencing_technology.ilmn

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
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

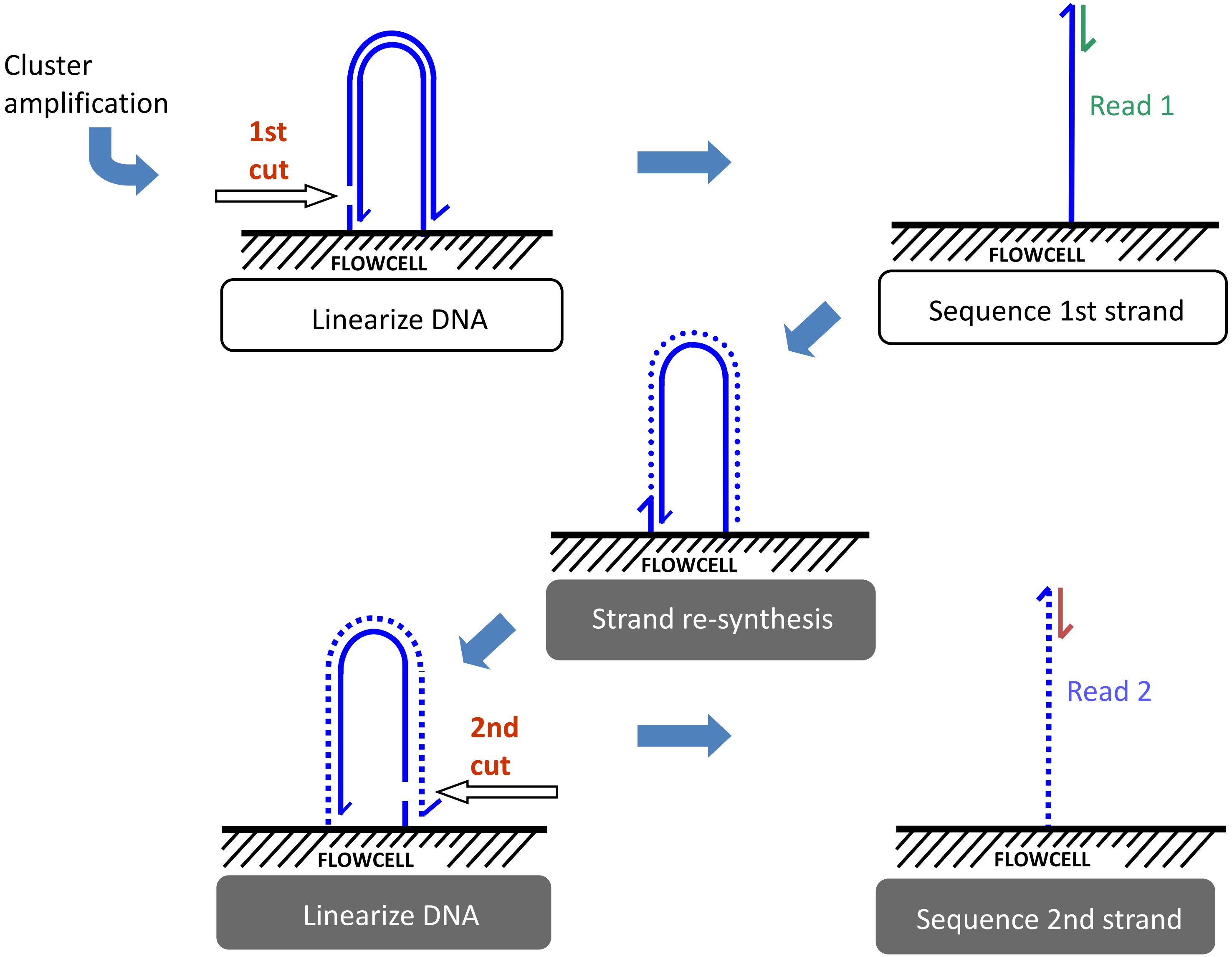
1-Channel Chemistry

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

- - - - - - Intermediate chemistry step

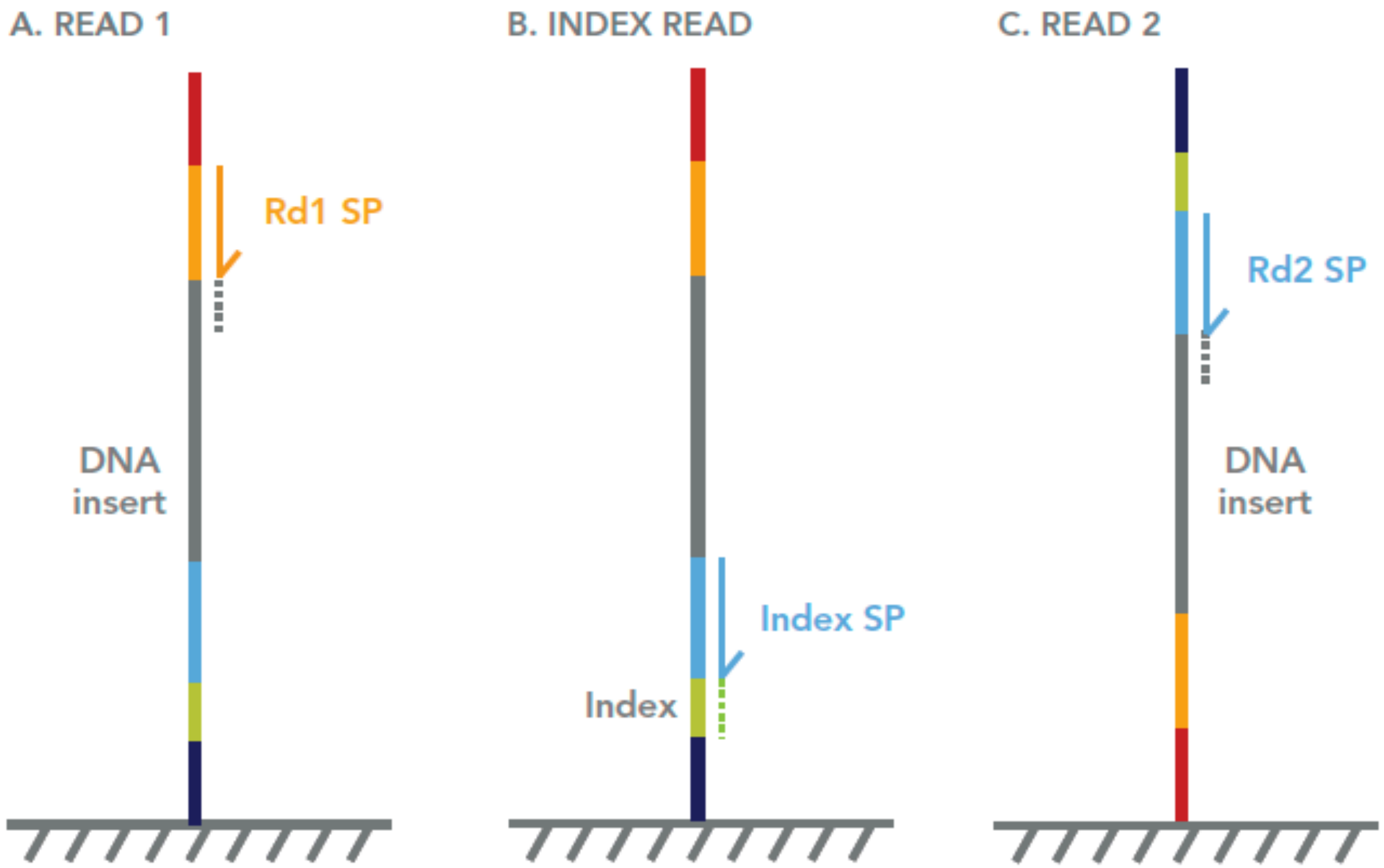
Figure 2: Four-, Two-, and One-Channel Chemistry—Four-channel chemistry uses a mixture of nucleotides labeled with four different fluorescent dyes. Two-channel chemistry uses two different fluorescent dyes, and one-channel chemistry uses only one dye. The images are processed by image analysis software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

Paired-End Sequencing allows for two looks at a sequence



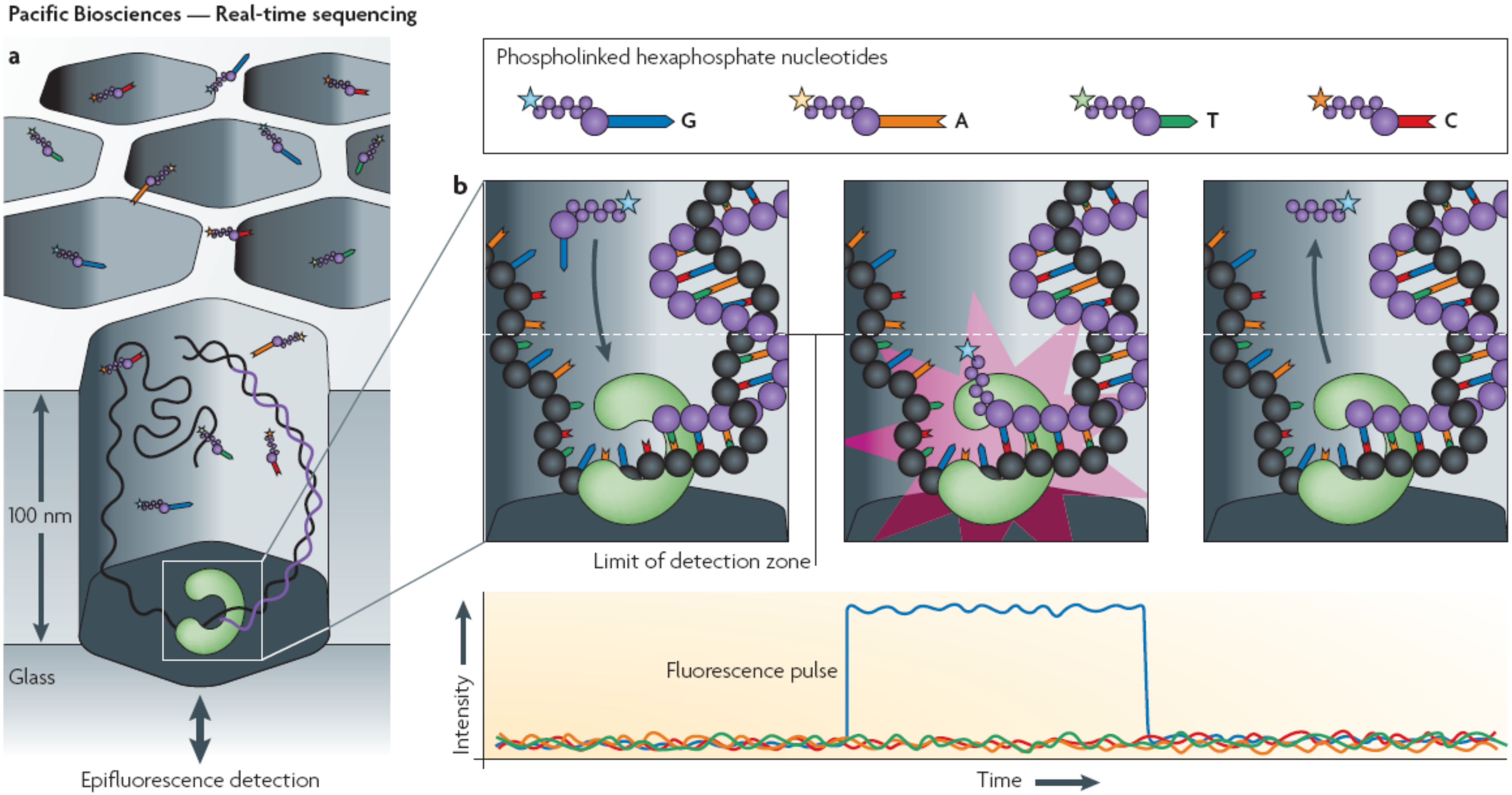
© Illumina, Inc.

Indexed sequencing method is now standard for single and paired reads



© Illumina, Inc.

Pacific Biosciences
Single Molecule Real-Time (SMRT) Sequencing

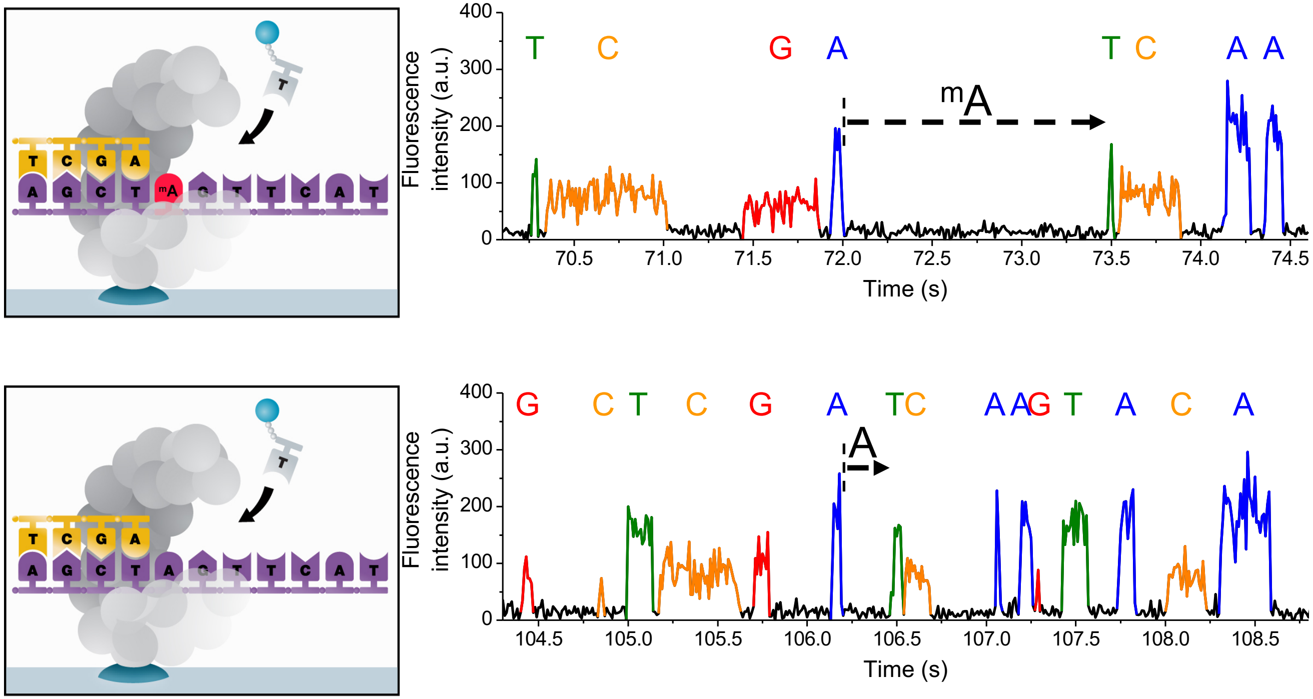

Pacific Biosciences — Real-time sequencing
a
100 nm
Glass
Epifluorescence detection
Phospholinked hexaphosphate nucleotides
G
A
T
C
b
Limit of detection zone
Fluorescence pulse
Intensity
Time

Metzker, 2010

Single Molecule Kinetics Allow for the Direct Detection of Methylation

Approach: Kinetic detection of methylated bases during SMRT DNA sequencing

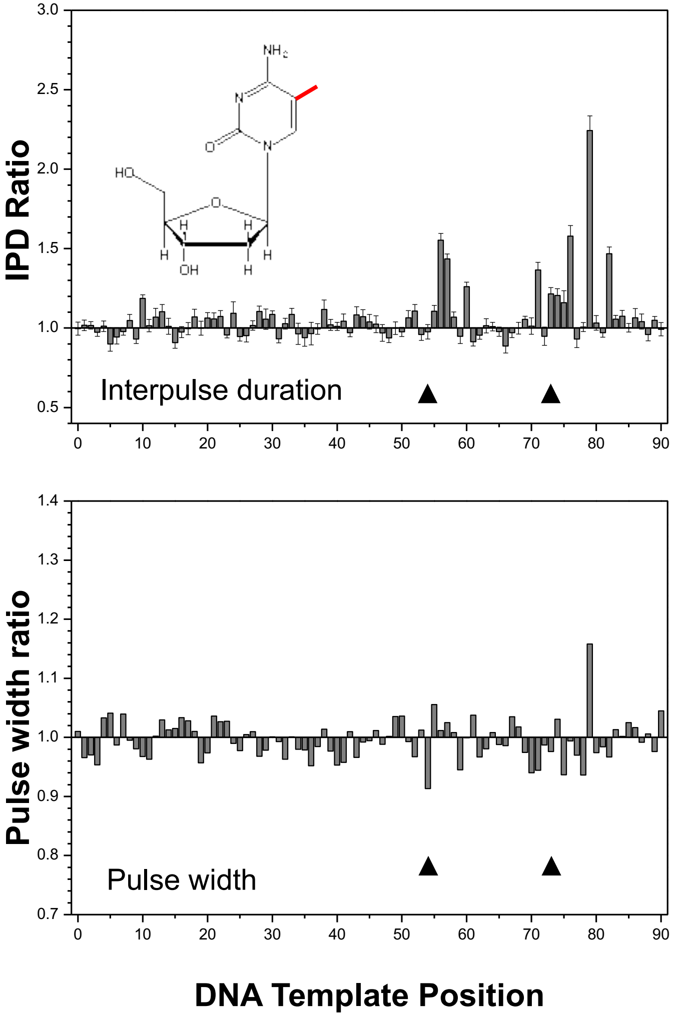
Example: N^6-methyladenosine (^{m}A)



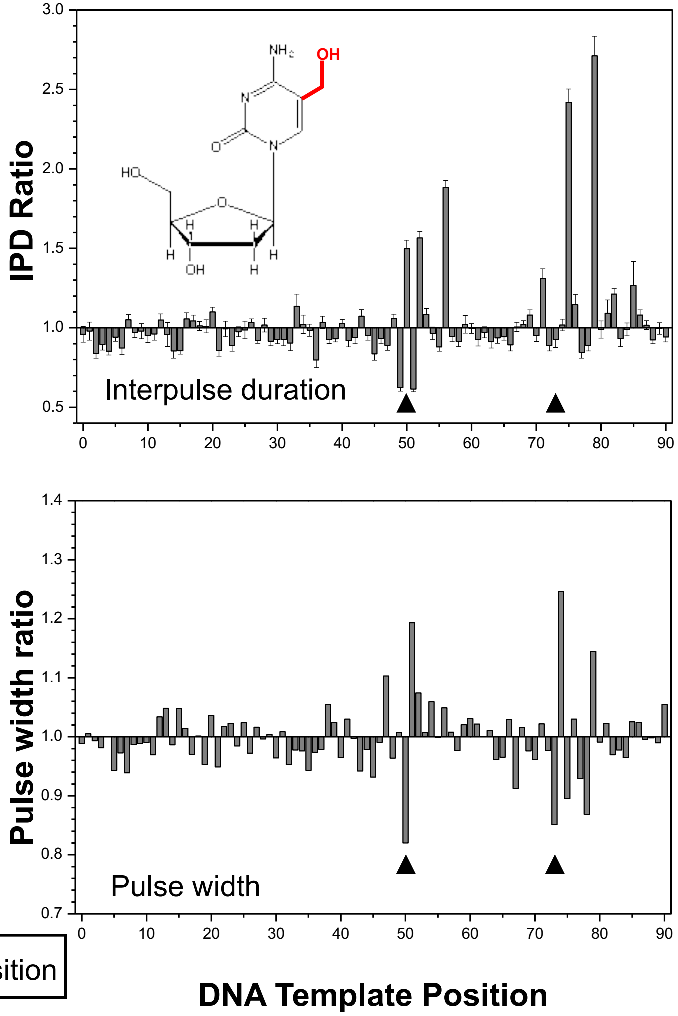
Flusberg et al., 2010.

Kinetics can detect other base modifications

5-methylcytosine (^{m}C)

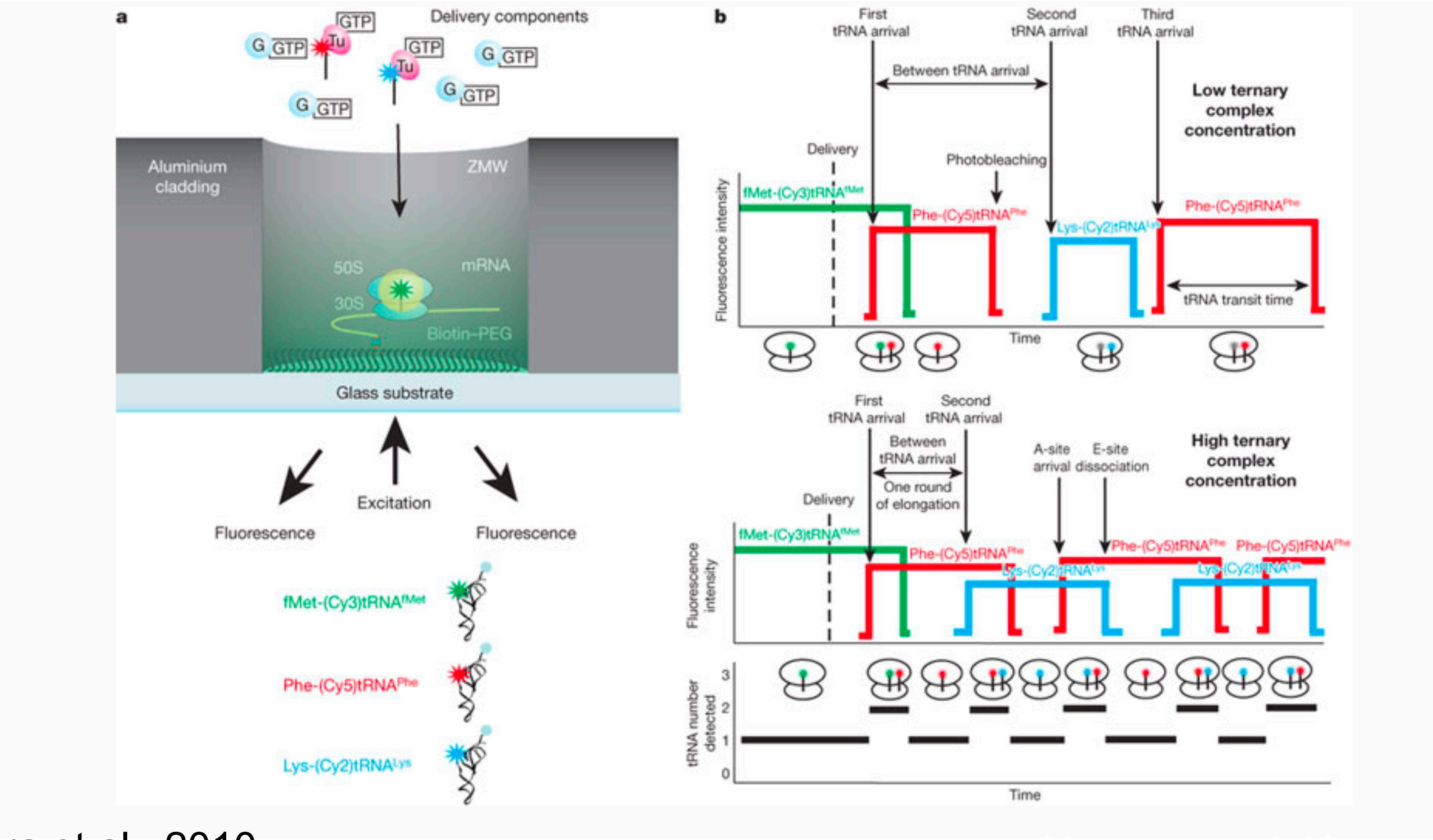


5-hydroxymethylcytosine (^{hm}C)



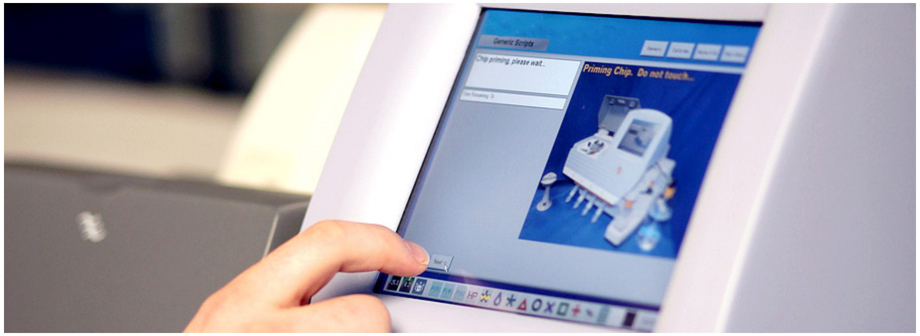
▲ = Methylated position

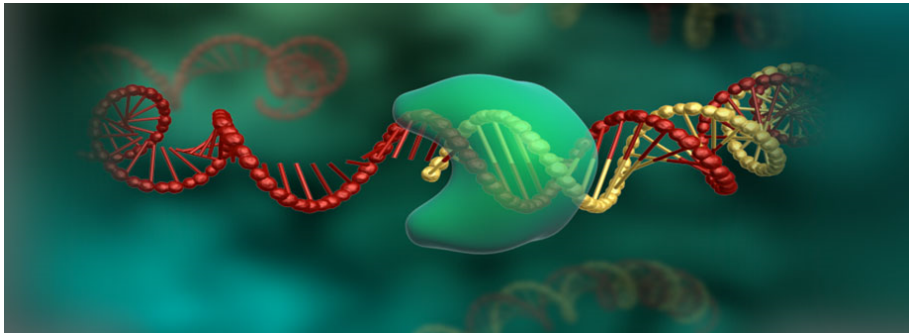
Kinetics allow one to watch protein translation as it occurs

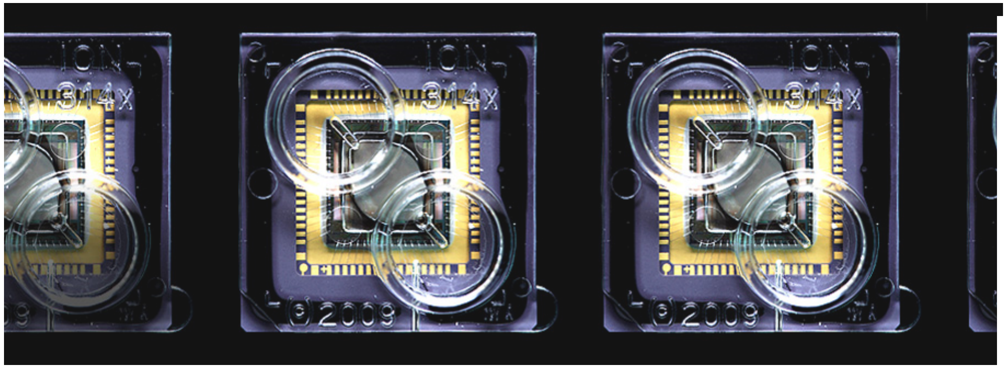

a
Delivery components
G
GTP
Tu
Aluminium cladding
ZMW
50S
30S
mRNA
Biotin–PEG
Glass substrate
Excitation
Fluorescence
Fluorescence
fMet-(Cy3)tRNA$^{\text{fMet}}$
Phe-(Cy5)tRNA$^{\text{Phe}}$
Lys-(Cy2)tRNA$^{\text{Lys}}$
b
First tRNA arrival
Second tRNA arrival
Third tRNA arrival
Between tRNA arrival
Low ternary complex concentration
Delivery
Photobleaching
Fluorescence intensity
fMet-(Cy3)tRNA$^{\text{fMet}}$
Phe-(Cy5)tRNA$^{\text{Phe}}$
Lys-(Cy2)tRNA$^{\text{Lys}}$
Phe-(Cy5)tRNA$^{\text{Phe}}$
tRNA transit time
Time
First tRNA arrival
Second tRNA arrival
Between tRNA arrival
One round of elongation
A-site arrival
E-site dissociation
High ternary complex concentration
Delivery
Fluorescence intensity
fMet-(Cy3)tRNA$^{\text{fMet}}$
Phe-(Cy5)tRNA$^{\text{Phe}}$
Lys-(Cy2)tRNA$^{\text{Lys}}$
Phe-(Cy5)tRNA$^{\text{Phe}}$
Phe-(Cy5)tRNA$^{\text{Phe}}$
Lys-(Cy2)tRNA$^{\text{Lys}}$
tRNA number detected
3
2
1
0
Time

Uemura et al., 2010

"Post-Light," Semi-Conductor Sequencing: Life Technologies Personal Genome Machine (PGM) and the Proton I and Proton II





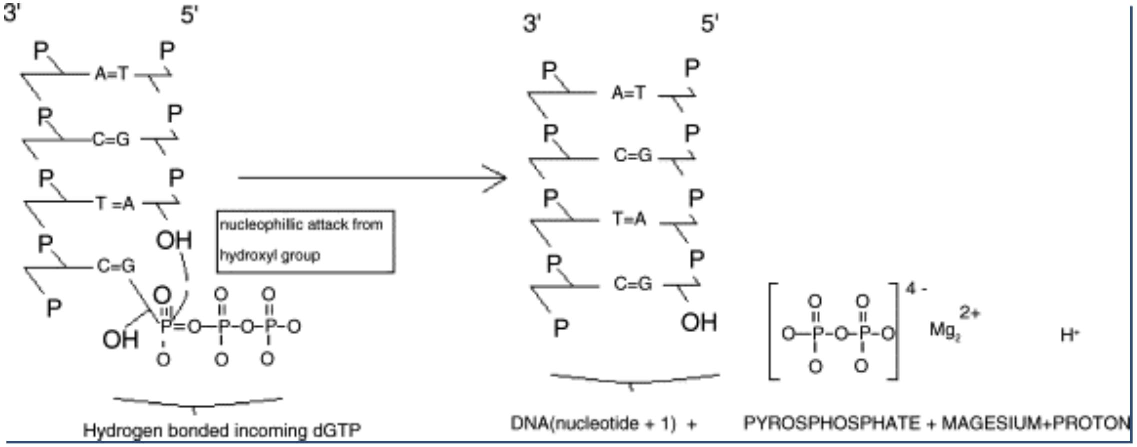


Simple chip loading

Intuitive graphical user interface

State-of-the-art electronics

Integrated reagent delivery


3'
5'
A=T
C=G
T=A
C=G
P
OH
nucleophilic attack from hydroxyl group
Hydrogen bonded incoming dGTP
DNA(nucleotide + 1) +
PYROSPHOSPHATE + MAGESIUM+PROTON
4 -
Mg
2+
H+

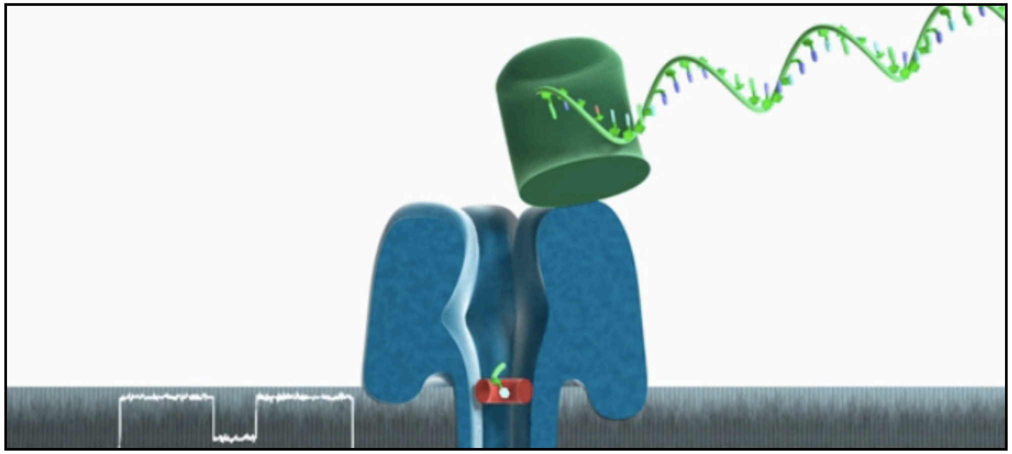
Essentially, 11 million very small pH meters

Purushothaman *et al*, 2005
IonTorrent, Inc.

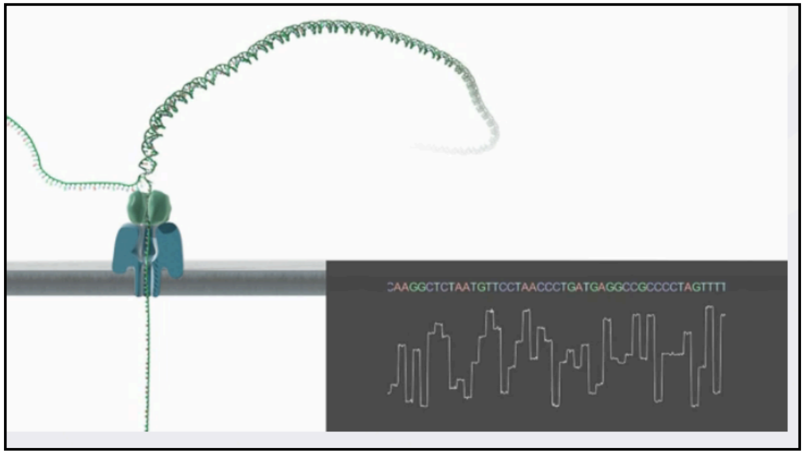
Latest Ion Platforms
Thermo Fisher's Ion S5 & S5 XL

Oxford NANOPORE Technologies®

DNA Sequencing with a protein nanopore

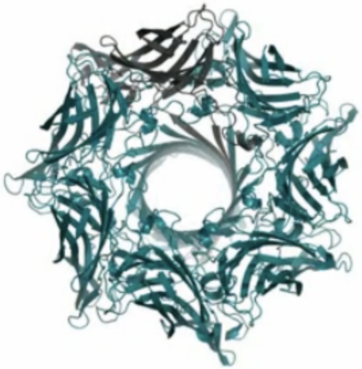


Exonuclease-Seq



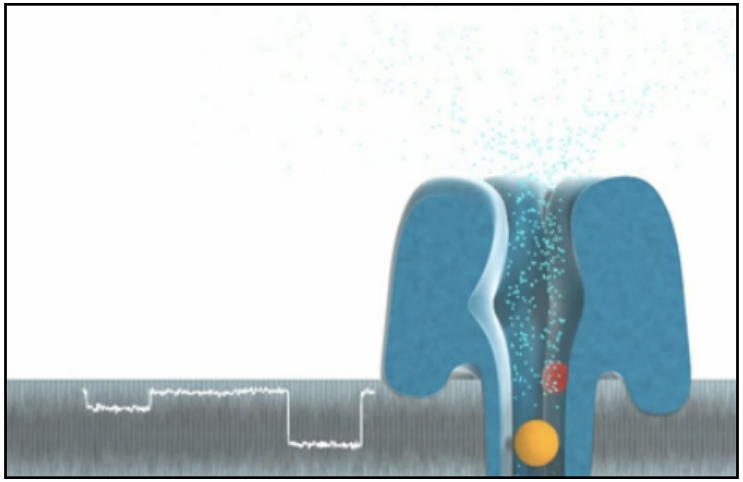
Strand-Seq

MinION

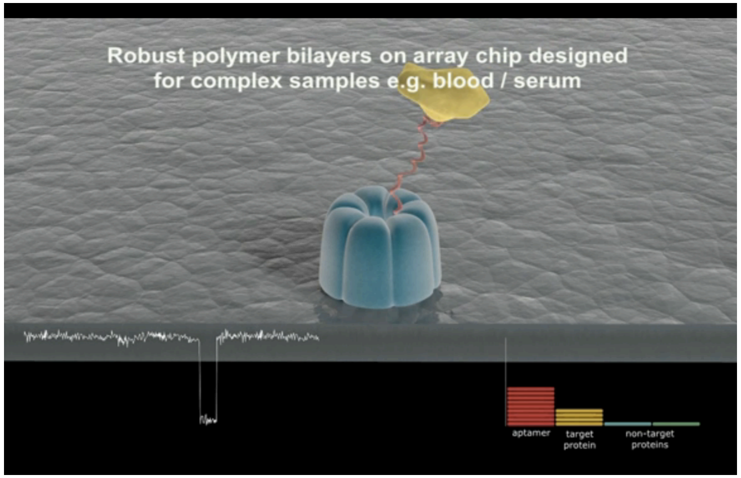


PromethION

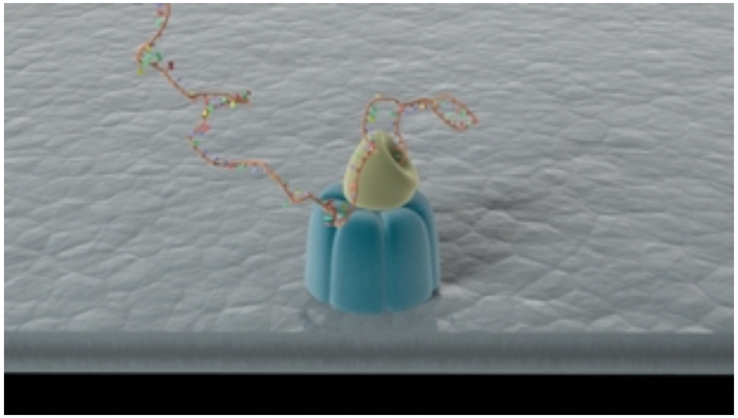
Other (Maybe Killer) Apps



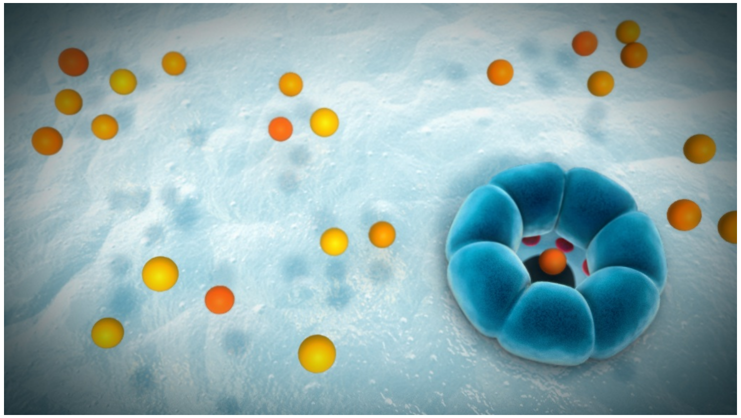
Analyte


Robust polymer bilayers on array chip designed for complex samples e.g. blood / serum
aptamer
target protein
non-target proteins

Protein Aptamer

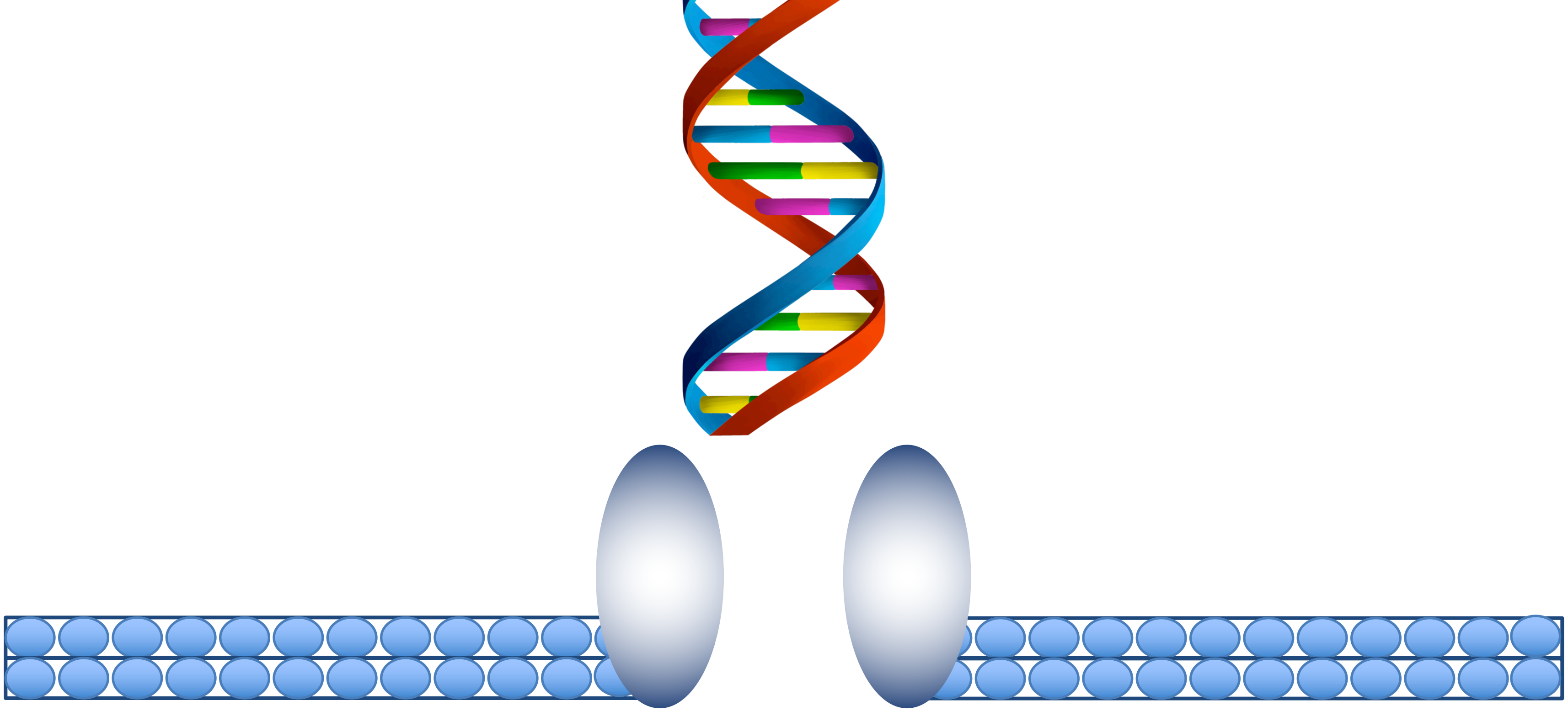


Direct RNA Sequencing



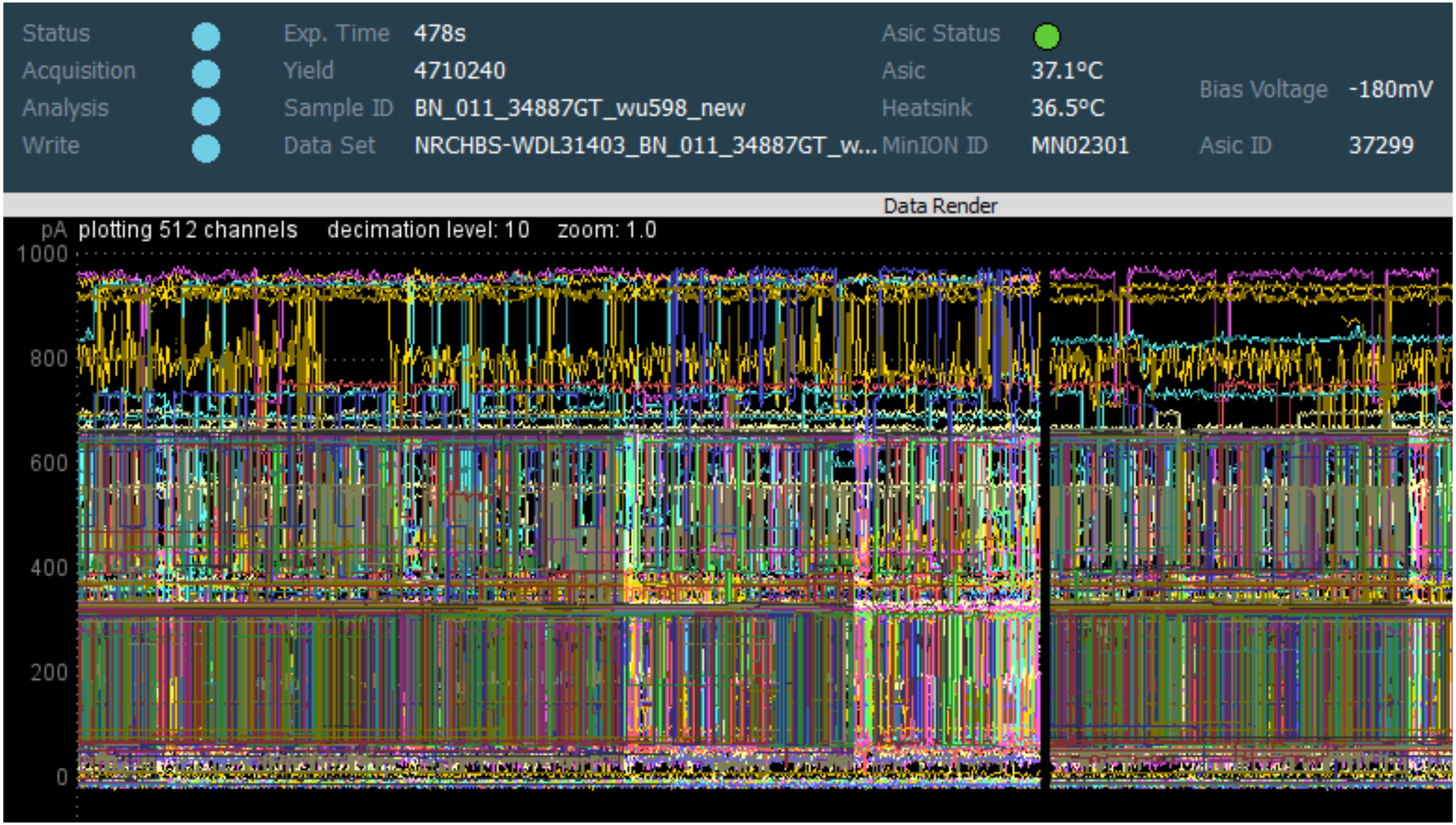
Small molecule

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

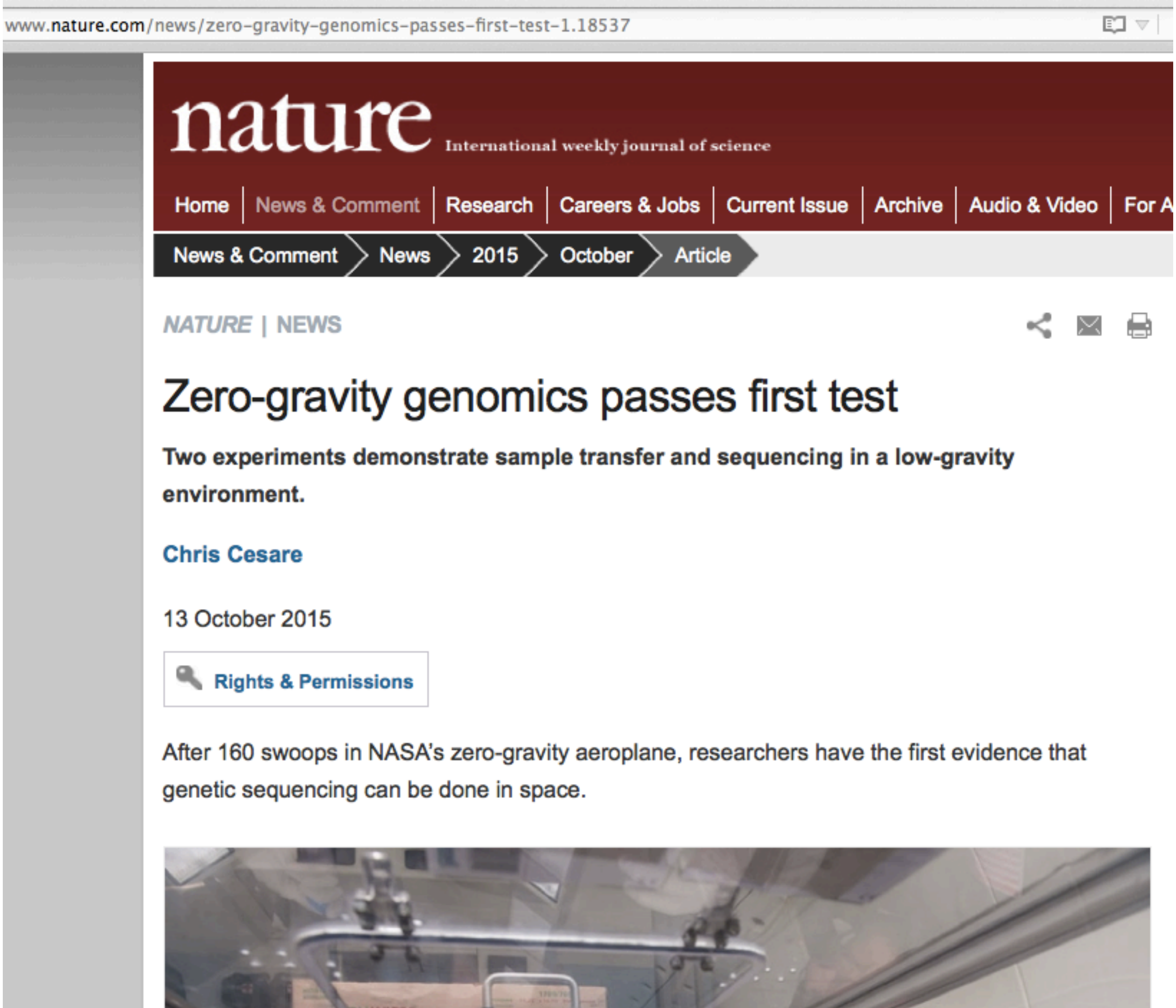
Bias Voltage -180mV
Asic ID 37299

Data Render

plotting 512 channels decimation level: 10 zoom: 1.0

Zero-G Pipetting: Hardest Lab Job Ever

Dr. Andrew Feinberg



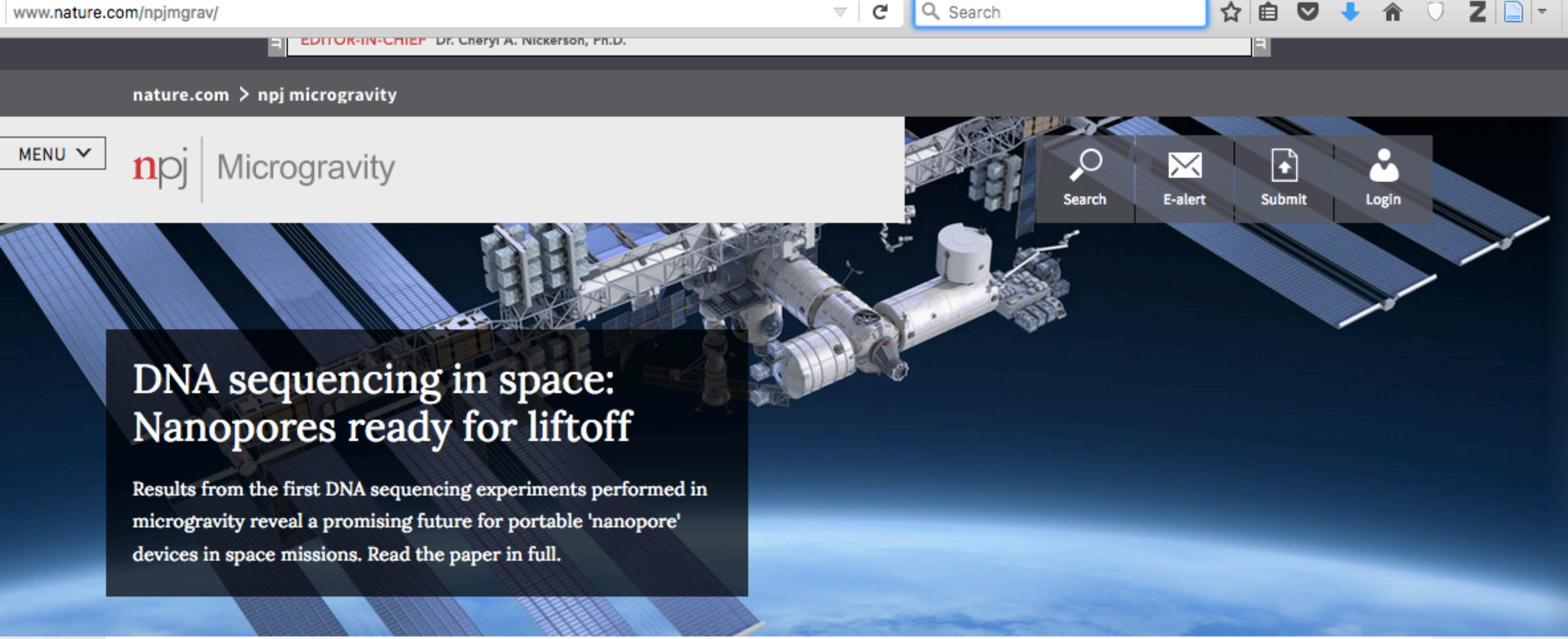
Zero-gravity genomics passes first test

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

Chris Cesare

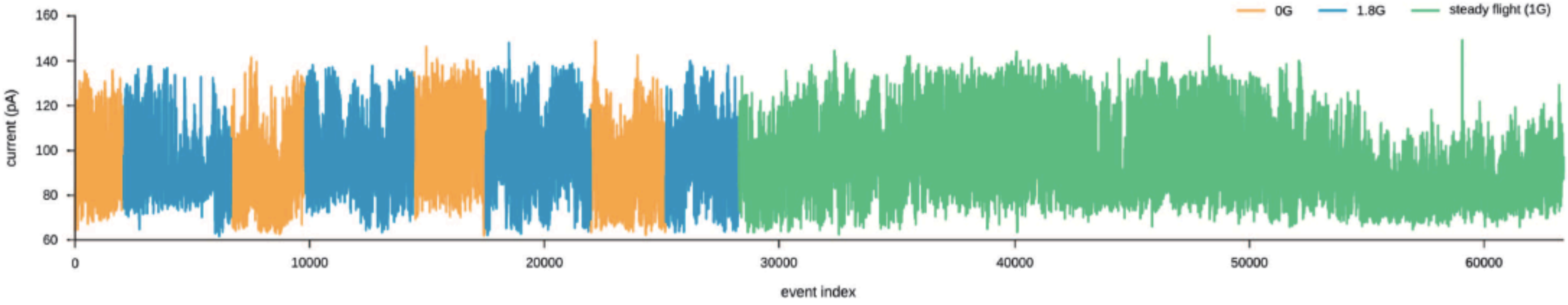
13 October 2015

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

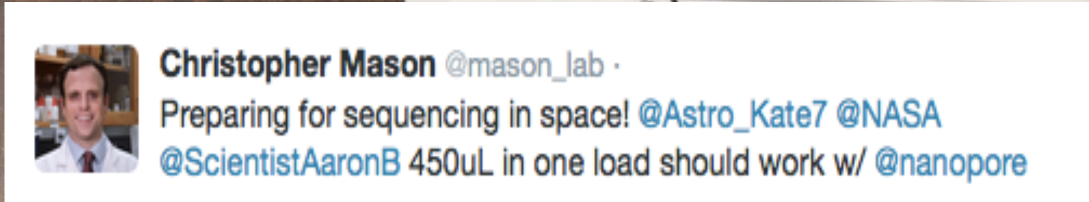

www.nature.com/npjmgrav/
Search
EDITOR-IN-CHIEF Dr. Cheryl A. Nickerson, Ph.D.
nature.com > npj microgravity
MENU
npj Microgravity
Search
E-alert
Submit
Login

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

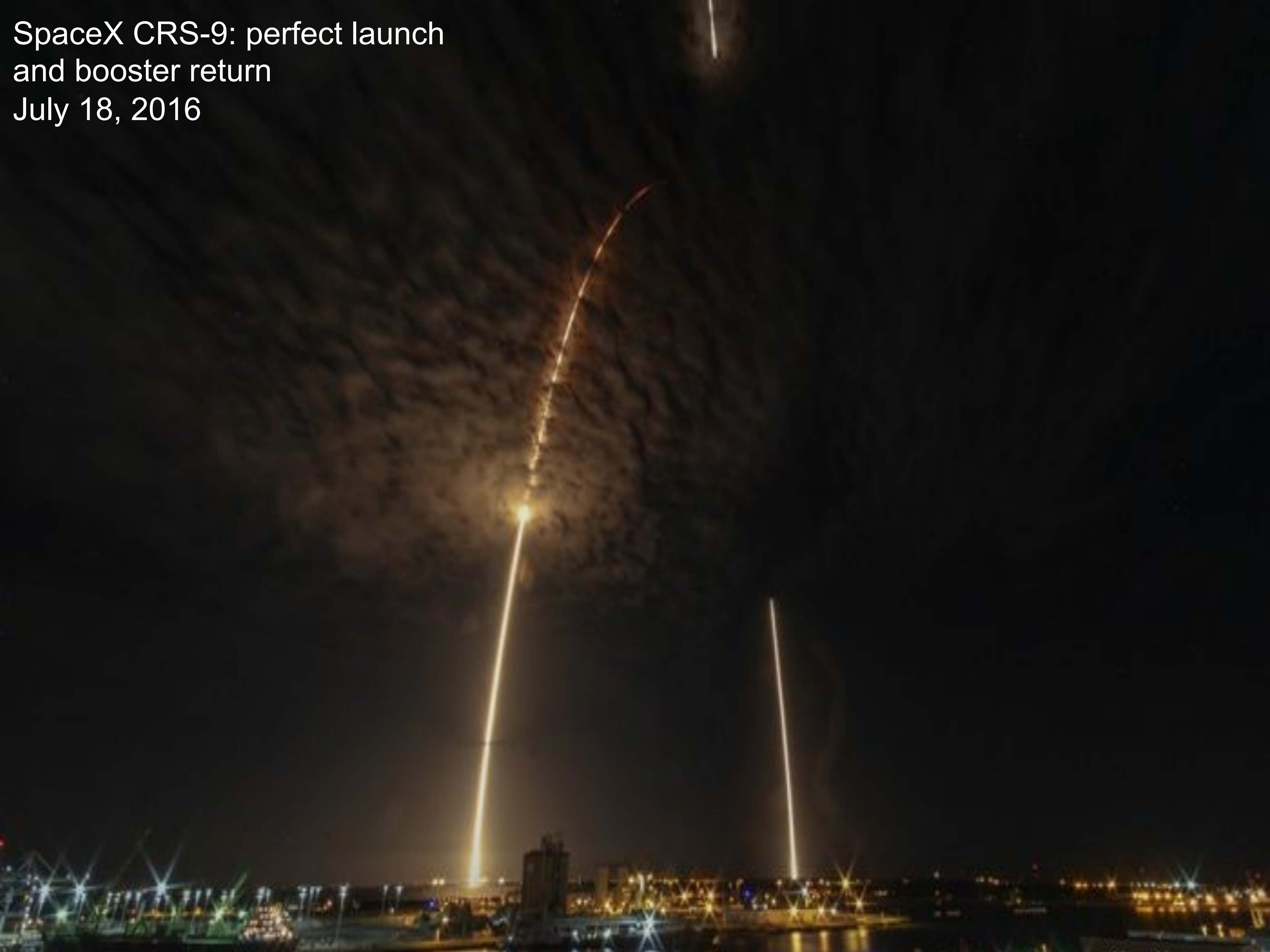
Dear Dr. Mason:

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.

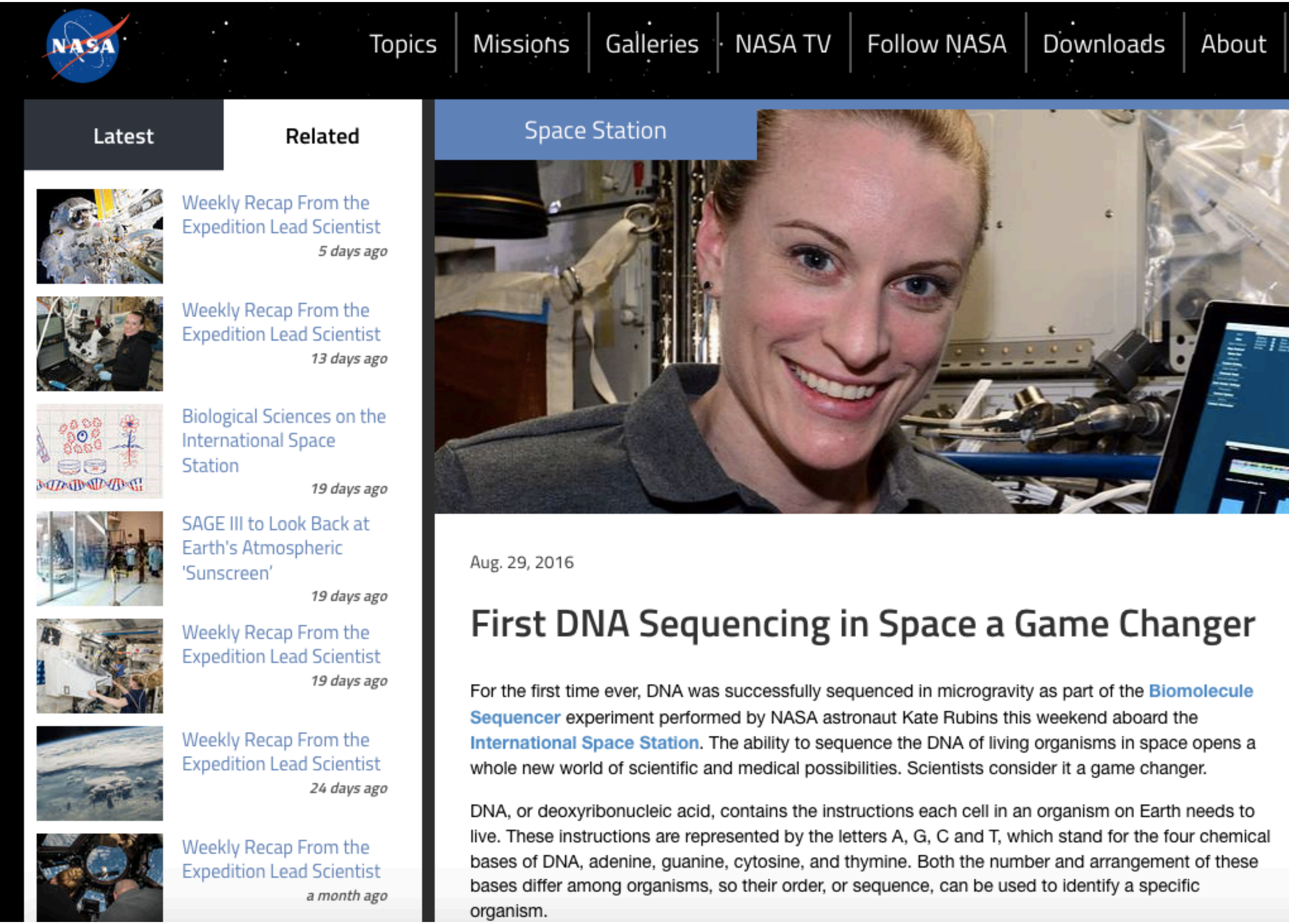
I am writing to encourage you – and in fact, to urge you – to continue your inquiry. The story of space exploration is the story of people just like you who meet adversity, head on, with determination and scientific and technological advancement. If you think about it, virtually every major innovation 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.

ВНИМАН
ЗОНА ПЕРЕМЕ
НЕ ДОПУС



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

TO NOD 2
PORT



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

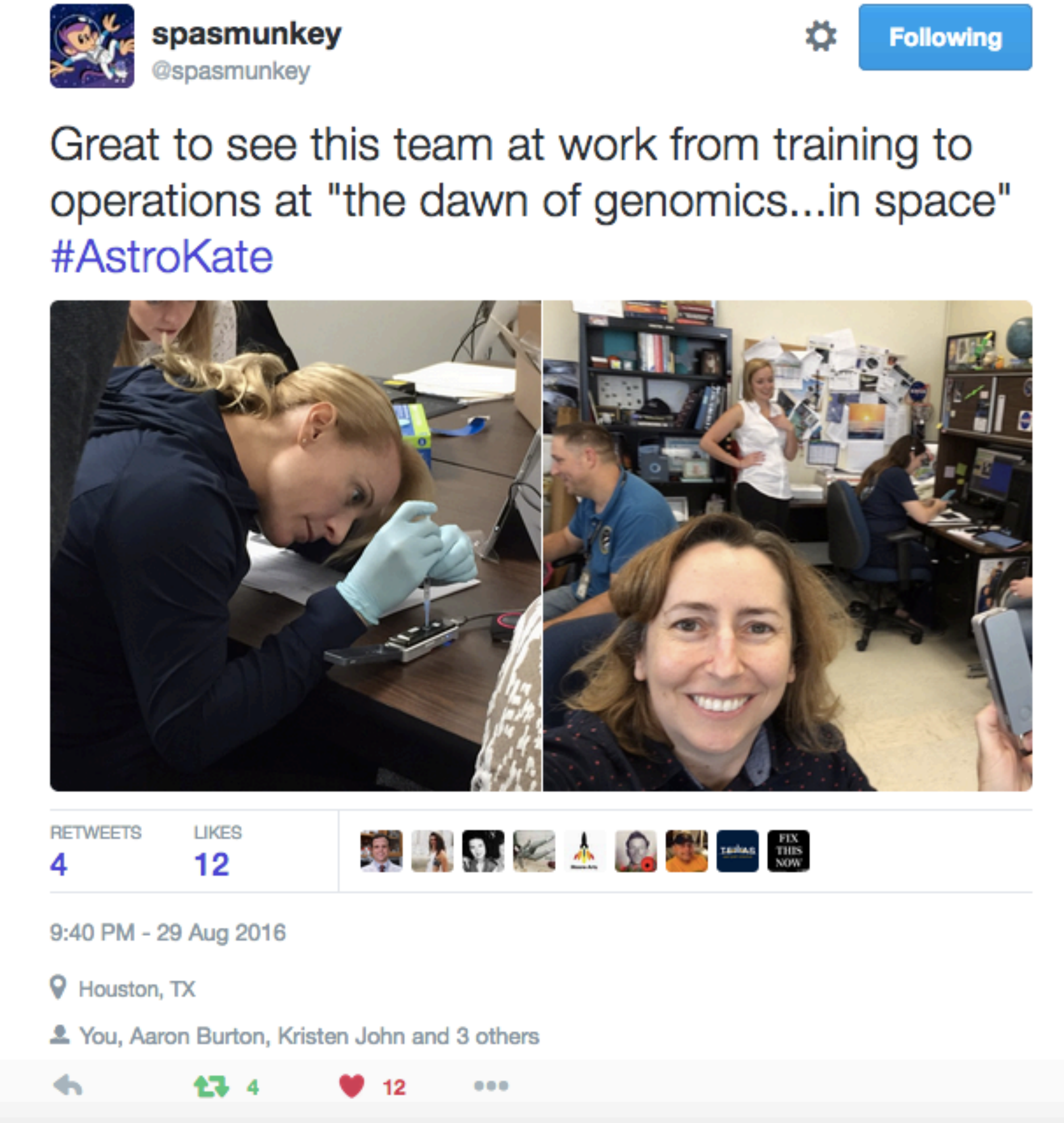
Weekly Recap From the Expedition Lead Scientist
a month ago

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

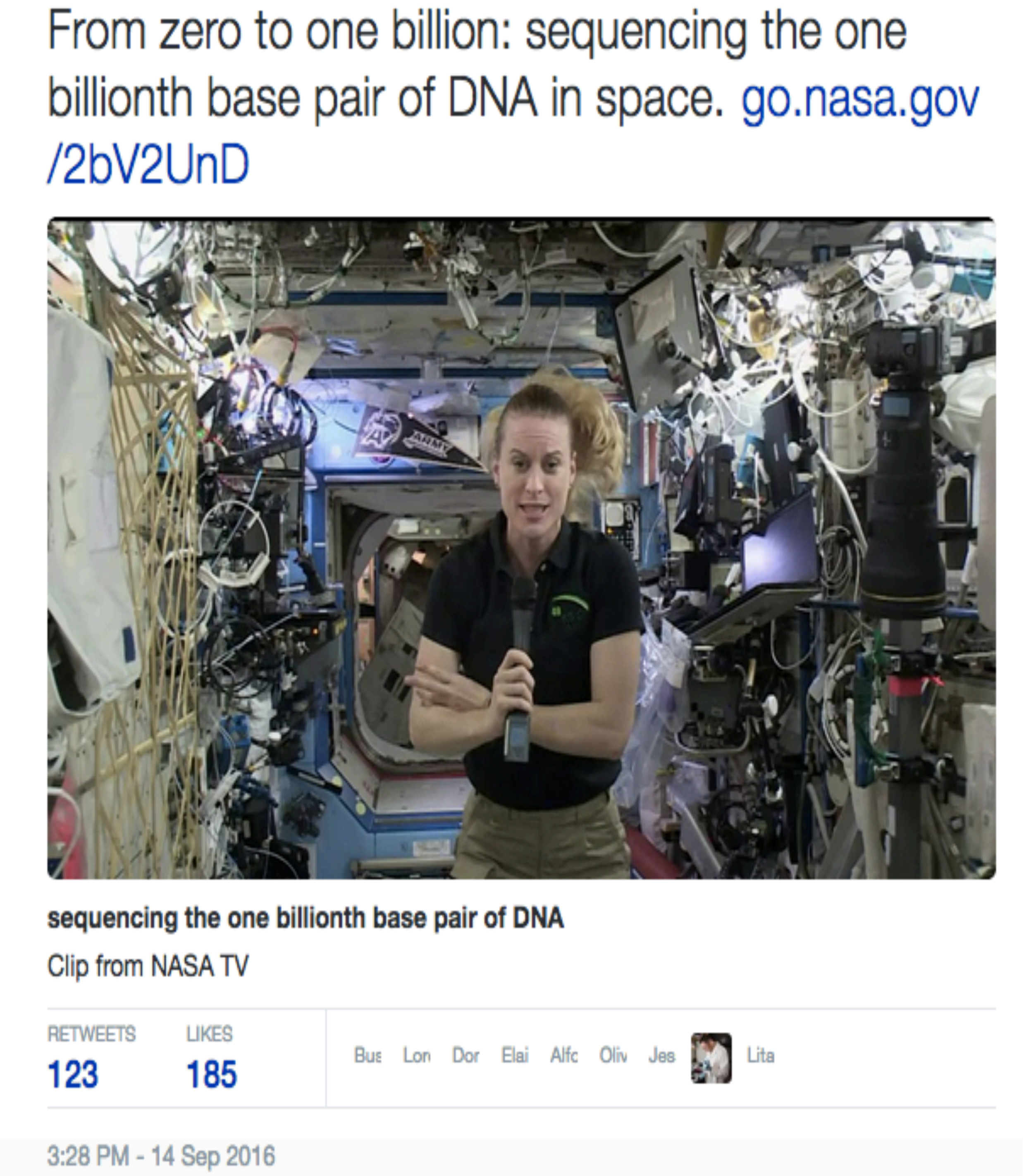
Great to see this team at work from training to operations at "the dawn of genomics...in space" #AstroKate

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

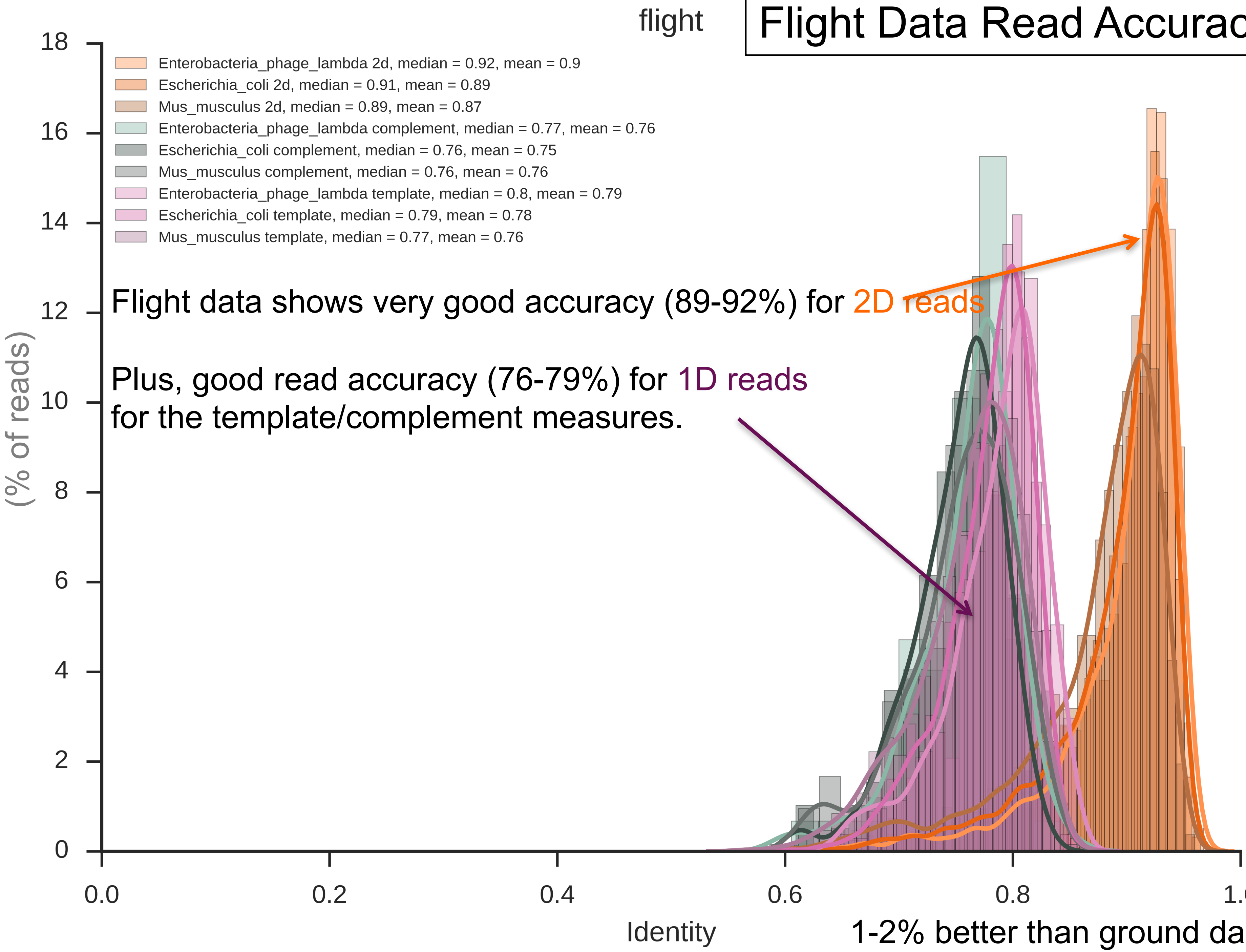
RETWEETS 123

LIKES 185

Bus Lon Dor Elai Alfc Oliv Jes Lita

3:28 PM - 14 Sep 2016

Flight Data Read Accuracy

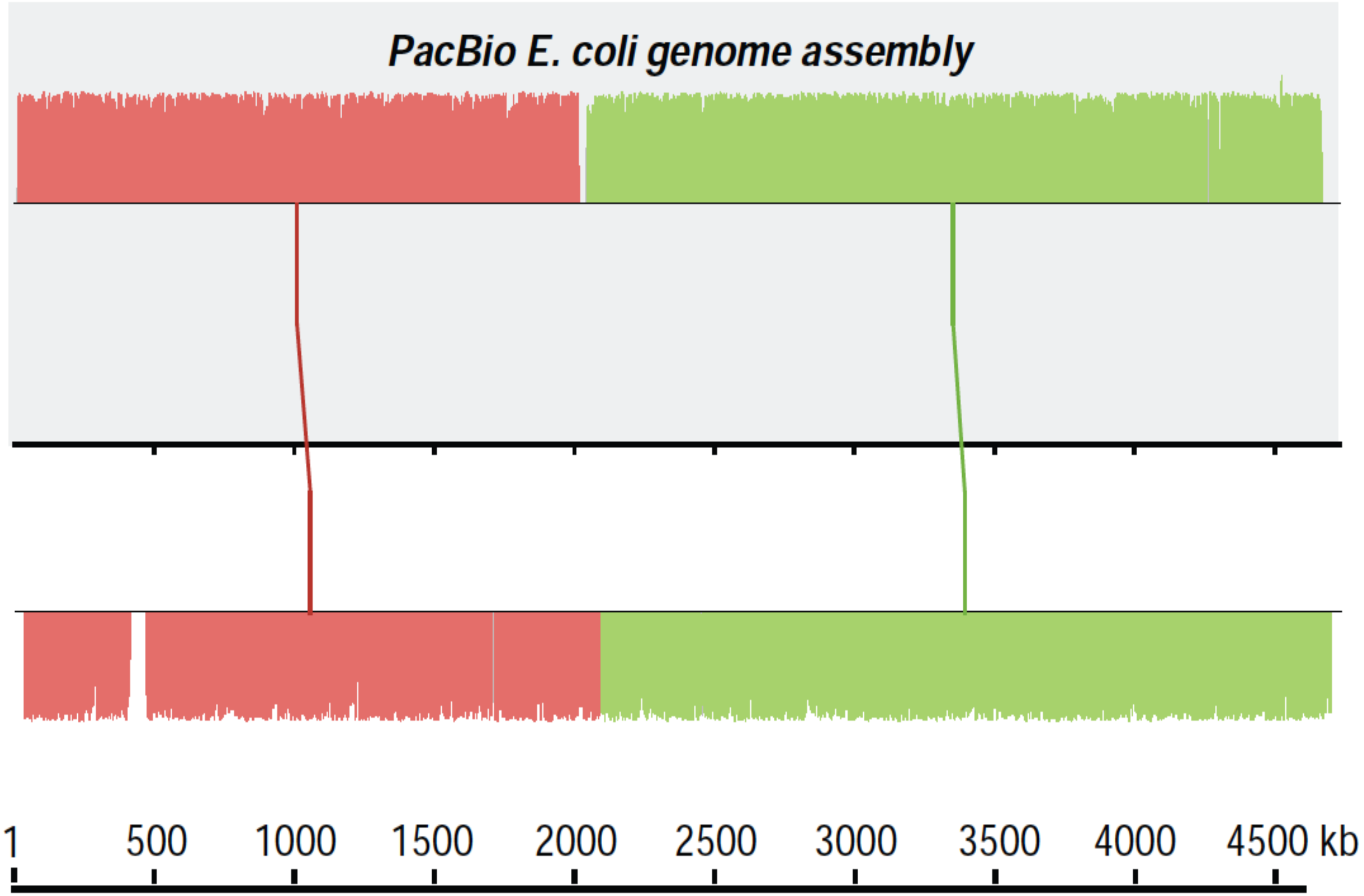


Flight data shows very good accuracy (89-92%) for 2D reads

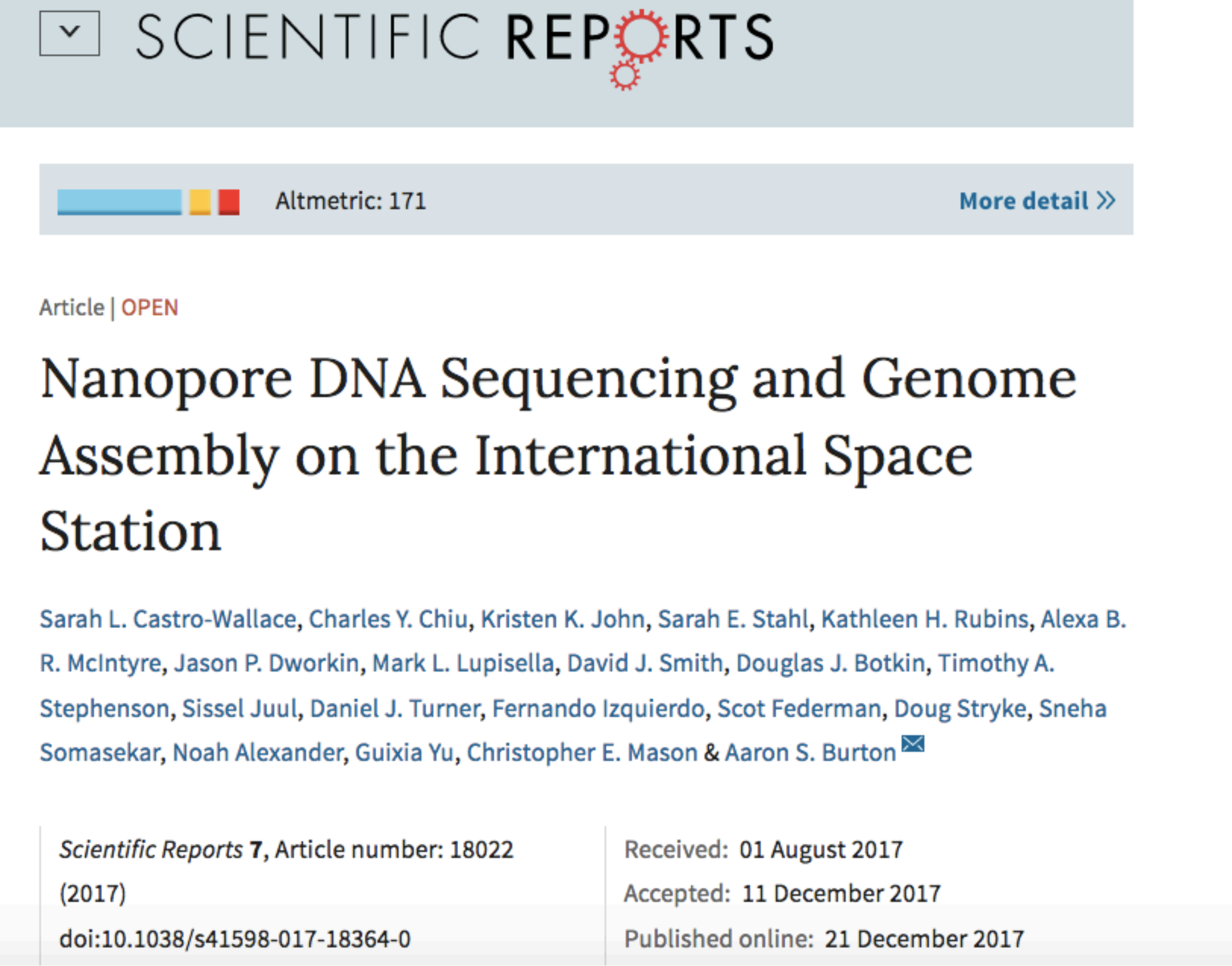
Plus, good read accuracy (76-79%) for 1D reads
for the template/complement measures.

1-2% better than ground data

Almost perfect when compared to PacBio



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



Article | OPEN

Nanopore DNA Sequencing and Genome Assembly on the International Space Station

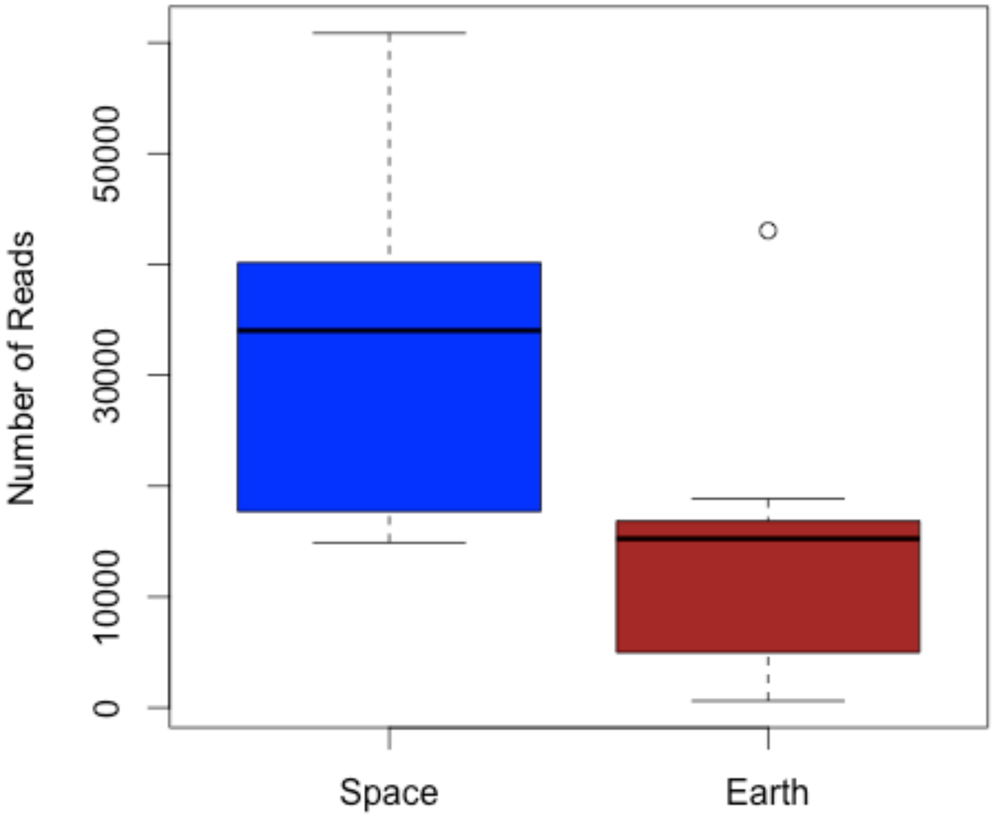
Sarah L. Castro-Wallace, Charles Y. Chiu, Kristen K. John, Sarah E. Stahl, Kathleen H. Rubins, Alexa B. R. McIntyre, Jason P. Dworkin, Mark L. Lupisella, David J. Smith, Douglas J. Botkin, Timothy A. Stephenson, Sissel Juul, Daniel J. Turner, Fernando Izquierdo, Scot Federman, Doug Stryke, Sneha Somasekar,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

Scientific Reports **7**, Article number: 18022 (2017)
doi:10.1038/s41598-017-18364-0

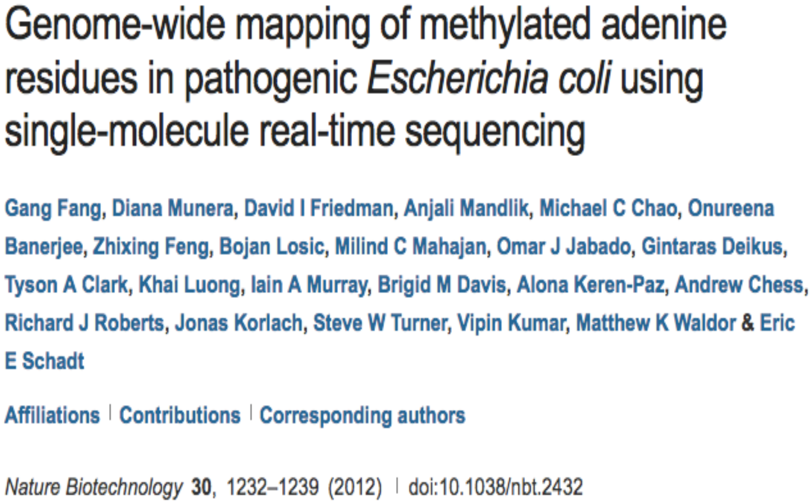
Received: 01 August 2017
Accepted: 11 December 2017
Published online: 21 December 2017

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

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



Bacteria are splattered with epigenetic marks

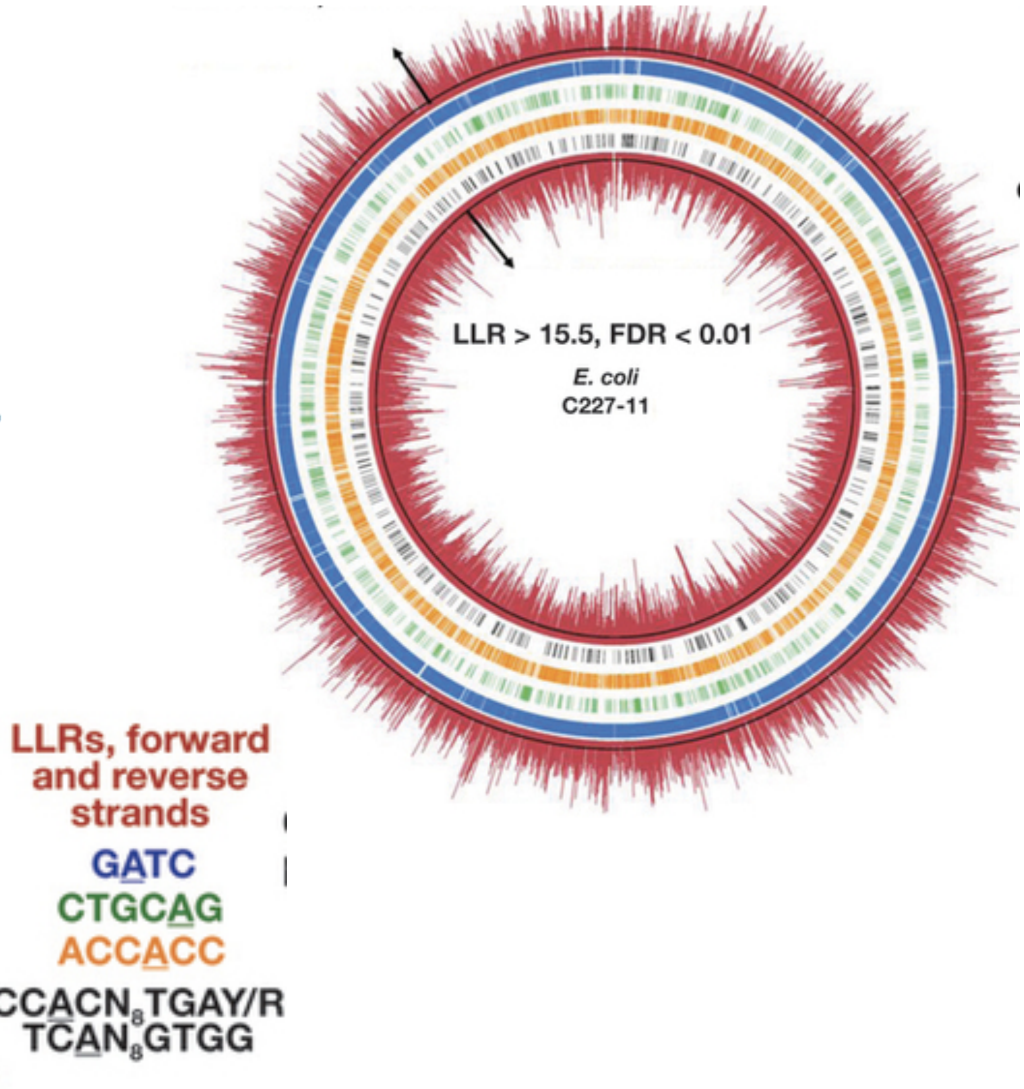


Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing

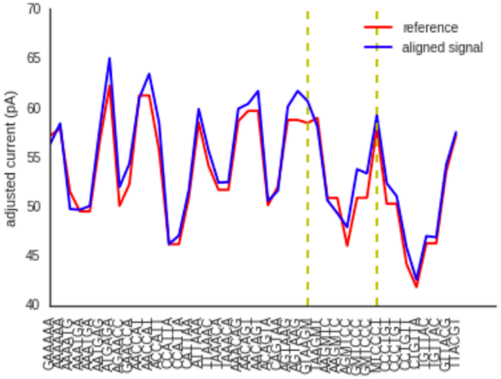
Gang Fang, Diana Munera, David I Friedman, Anjali Mandlik, Michael C Chao, Onureena Banerjee, Zhixing Feng, Bojan Losic, Milind C Mahajan, Omar J Jabado, Gintaras Deikus, Tyson A Clark, Khai Luong, Iain A Murray, Brigid M Davis, Alona Keren-Paz, Andrew Chess, Richard J Roberts, Jonas Korlach, Steve W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

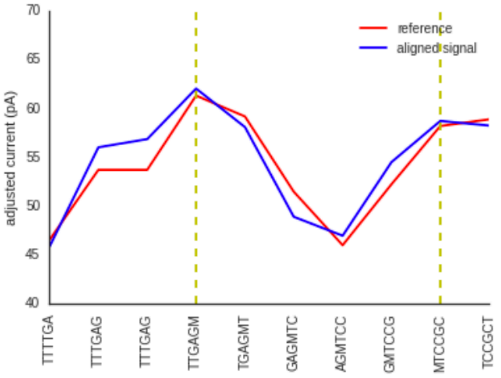
Affiliations | Contributions | Corresponding authors

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

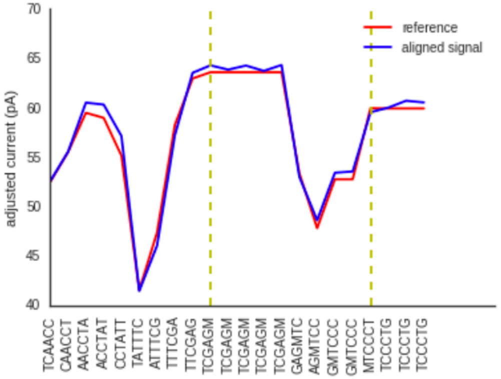


Calling current (pA) differences, similar to PacBio

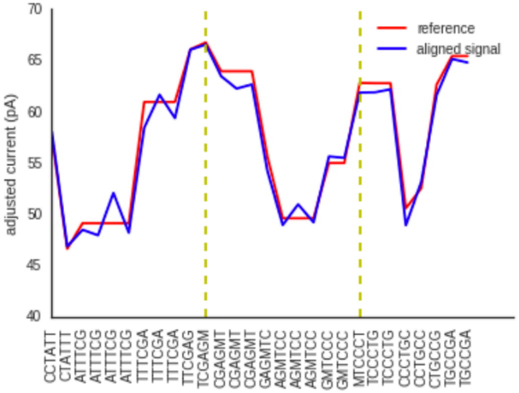


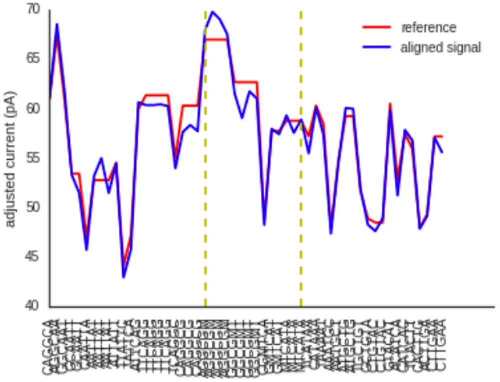


Reads aligned to same positions

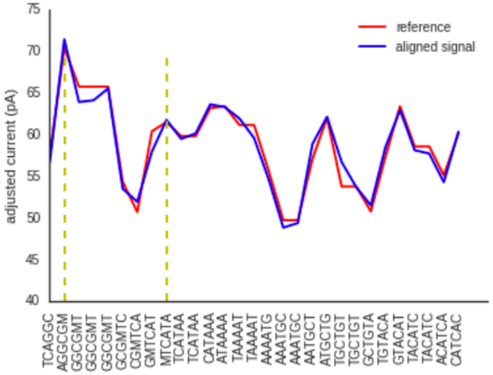


=

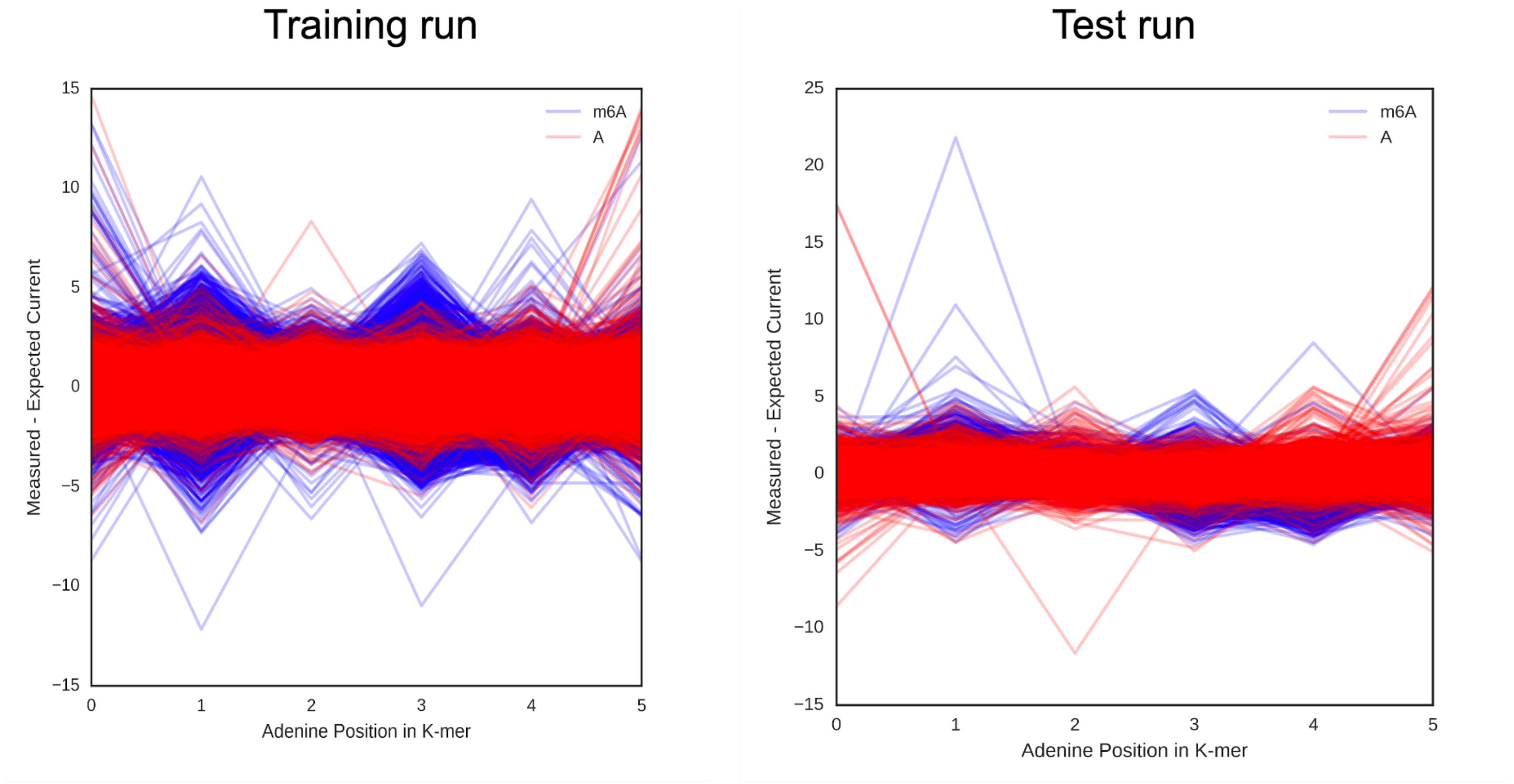


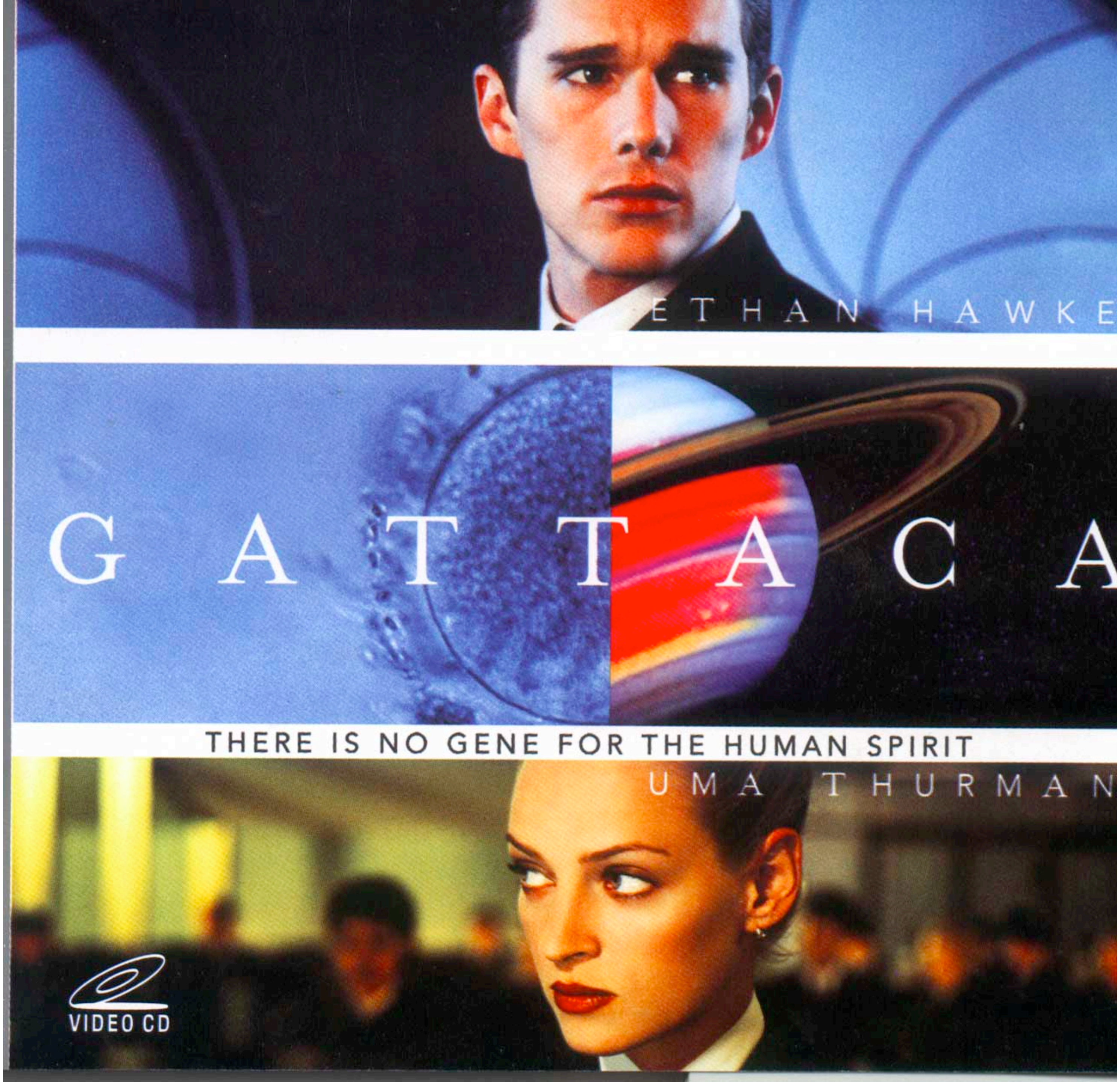


=



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

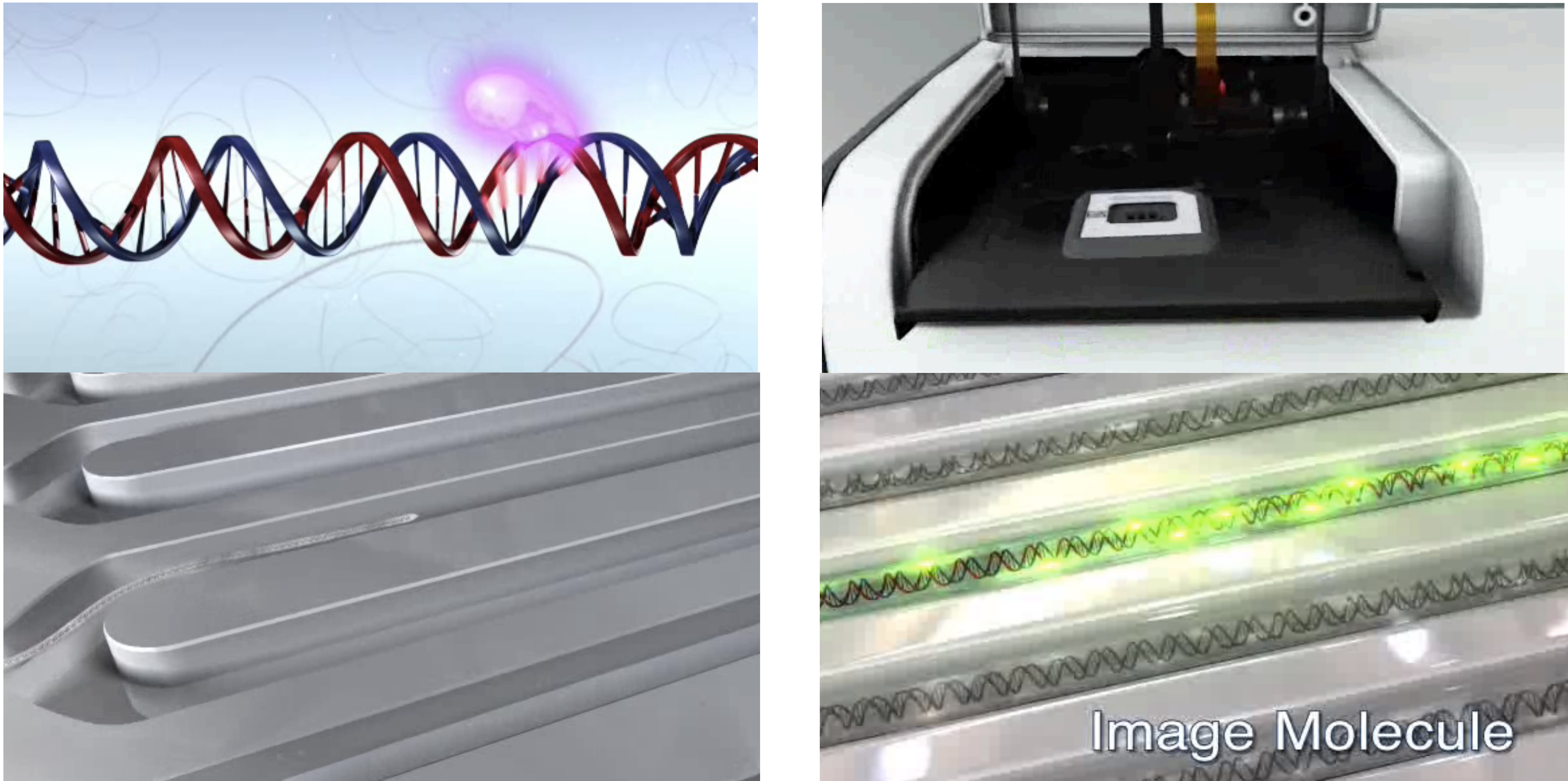
Is a 2.6 minute genome possible?
No today, but if the physics holds up…

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

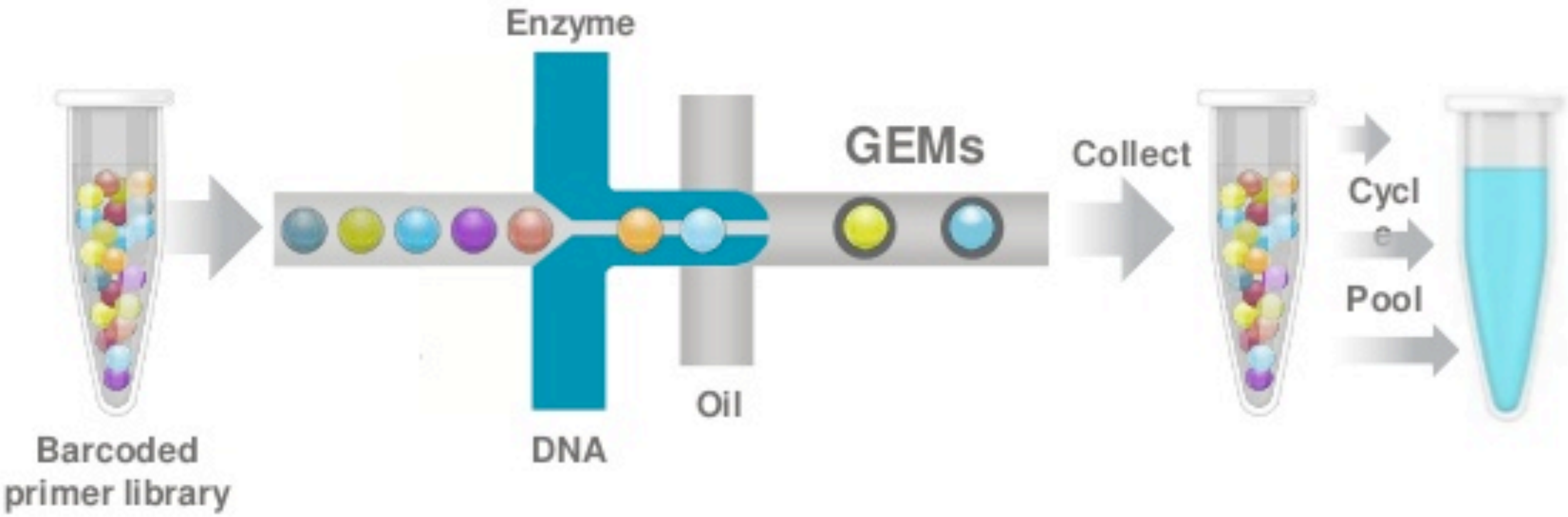
Bionanogenomics - Irys System

BIONANO
GENOMICS



QIAGEN GeneReader

>100,000 Reactions Assembled in < 5 min



Solid phase reagent delivery

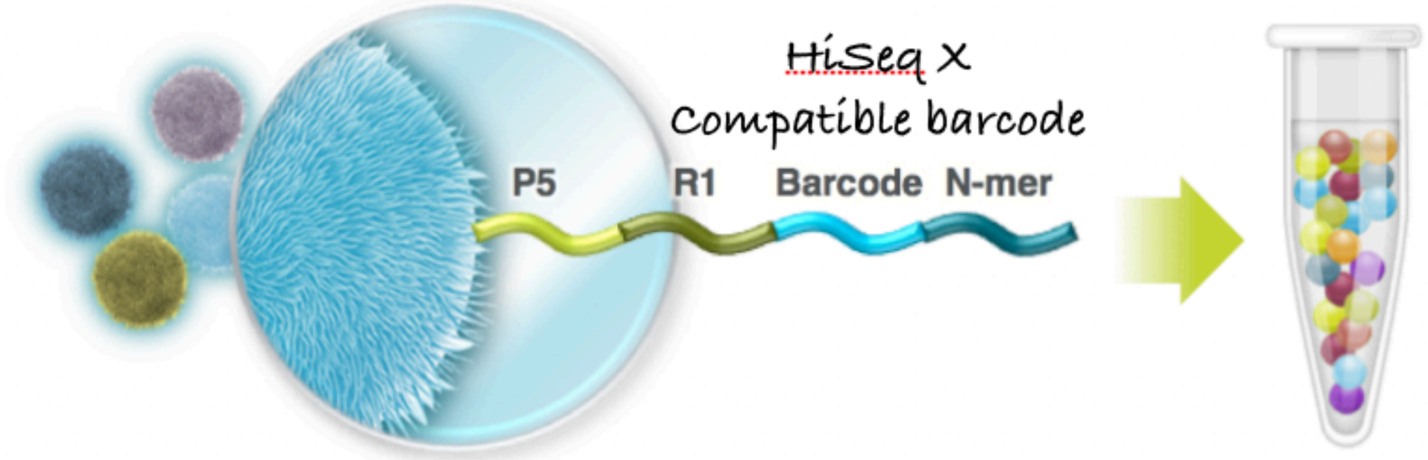
Fluid partitioning

Liquid phase biochemistry

4,000,000

~~750,000~~ Barcodes in One Tube

Gel bead scaffold → Functional oligo with barcode → High-diversity library



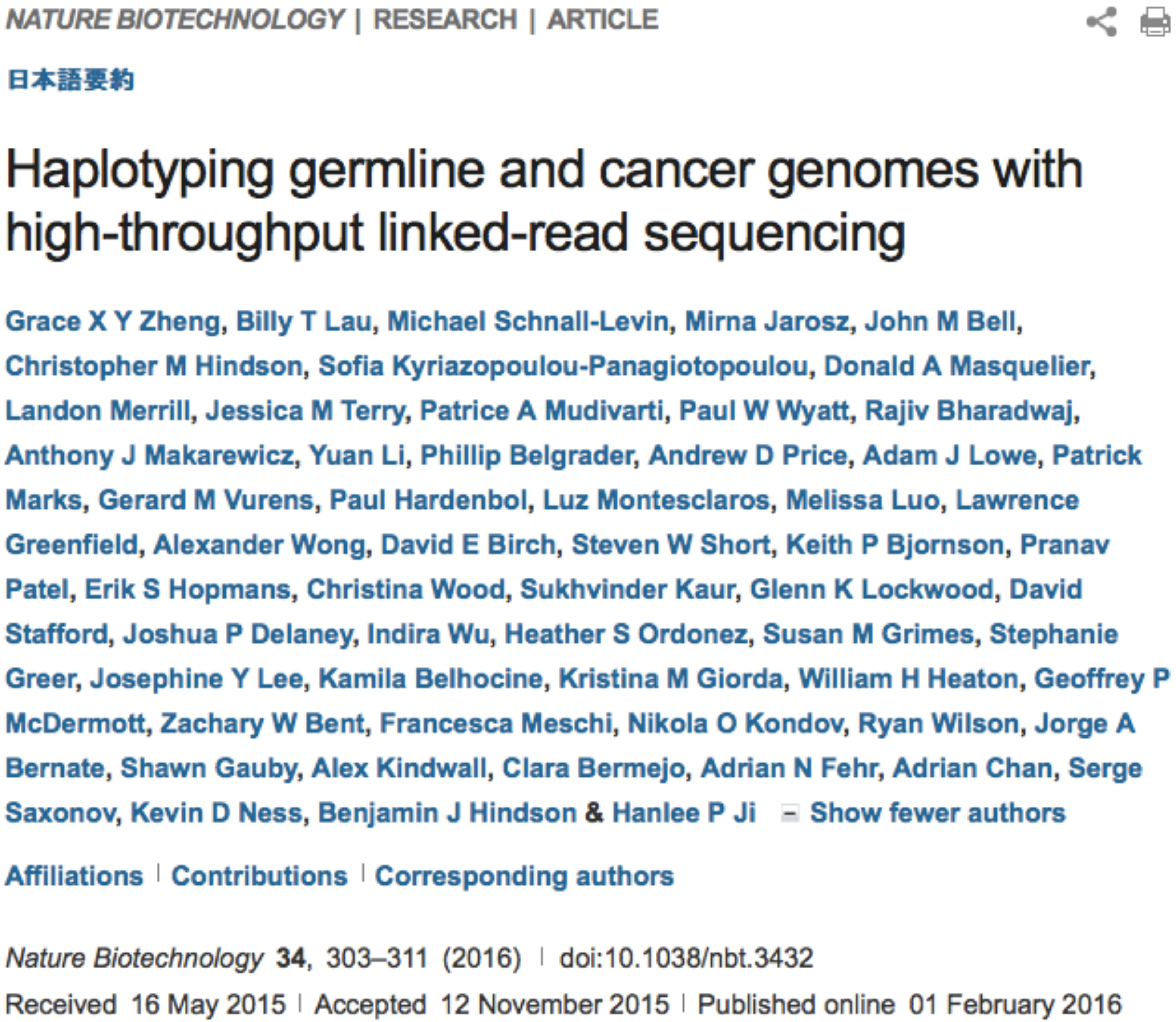
HiSeq X compatible barcode

Assembled in ~15 mins

- ~~14~~ 16 bp barcode
- Defined sequence
- Highly uniform size and representation
- Built-in sequencing adapter and primer content

Chromium: 1M Partitions from 4M Barcode Pool

	Conventional Approaches	GemCode	Chromium
Partitions	384	>100,000	>1,000,000
Barcode pool	384	750,000	4,000,000
Input DNA	100ng+	1ng	1ng



NATURE BIOTECHNOLOGY | RESEARCH | ARTICLE

日本語要約

Haplotyping germline and cancer genomes with high-throughput linked-read sequencing

Grace X Y Zheng, Billy T Lau, Michael Schnall-Levin, Mirna Jarosz, John M Bell, Christopher M Hindson, Sofia Kyriazopoulou-Panagiotopoulou, Donald A Masquelier, Landon Merrill, Jessica M Terry, Patrice A Mudivarti, Paul W Wyatt, Rajiv Bharadwaj, Anthony J Makarewicz, Yuan Li, Phillip Belgrader, Andrew D Price, Adam J Lowe, Patrick Marks, Gerard M Vurens, Paul Hardenbol, Luz Montesclaros, Melissa Luo, Lawrence Greenfield, Alexander Wong, David E Birch, Steven W Short, Keith P Bjornson, Pranav Patel, Erik S Hopmans, Christina Wood, Sukhvinder Kaur, Glenn K Lockwood, David Stafford, Joshua P Delaney, Indira Wu, Heather S Ordonez, Susan M Grimes, Stephanie Greer, Josephine Y Lee, Kamila Belhocine, Kristina M Giorda, William H Heaton, Geoffrey P McDermott, Zachary W Bent, Francesca Meschi, Nikola O Kondov, Ryan Wilson, Jorge A Bernate, Shawn Gauby, Alex Kindwall, Clara Bermejo, Adrian N Fehr, Adrian Chan, Serge Saxonov, Kevin D Ness, Benjamin J Hindson & Hanlee P Ji ⊟ Show fewer authors

Affiliations | Contributions | Corresponding authors

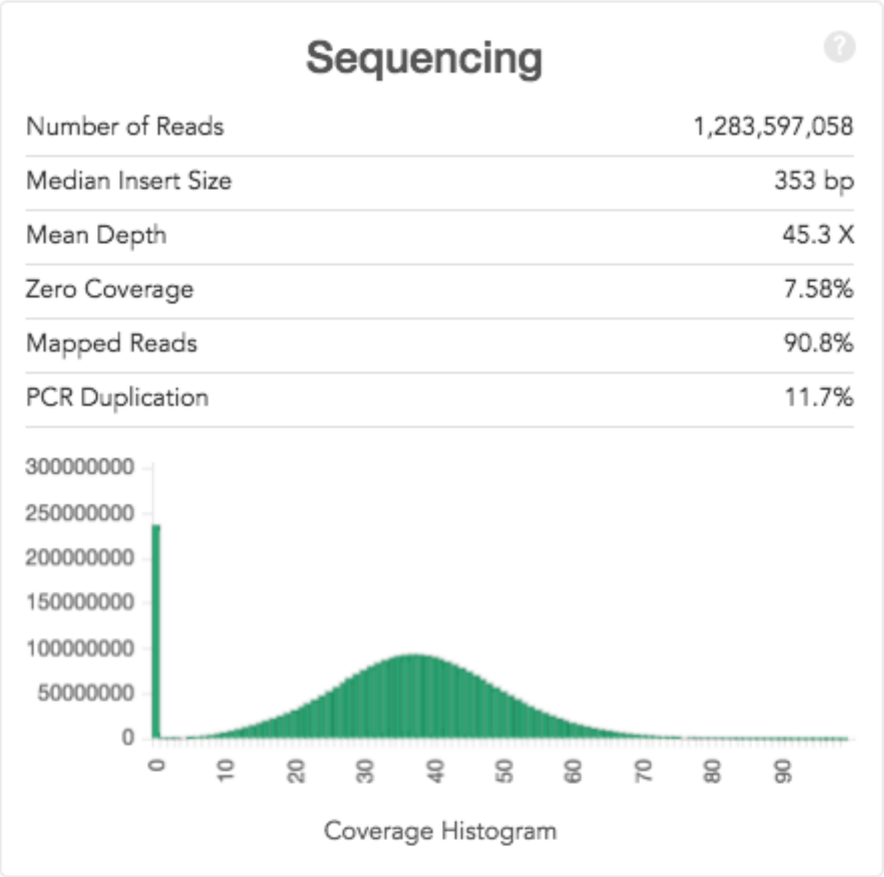
Nature Biotechnology **34**, 303–311 (2016) | doi:10.1038/nbt.3432

Received 16 May 2015 | Accepted 12 November 2015 | Published online 01 February 2016

Summary: Subject 2

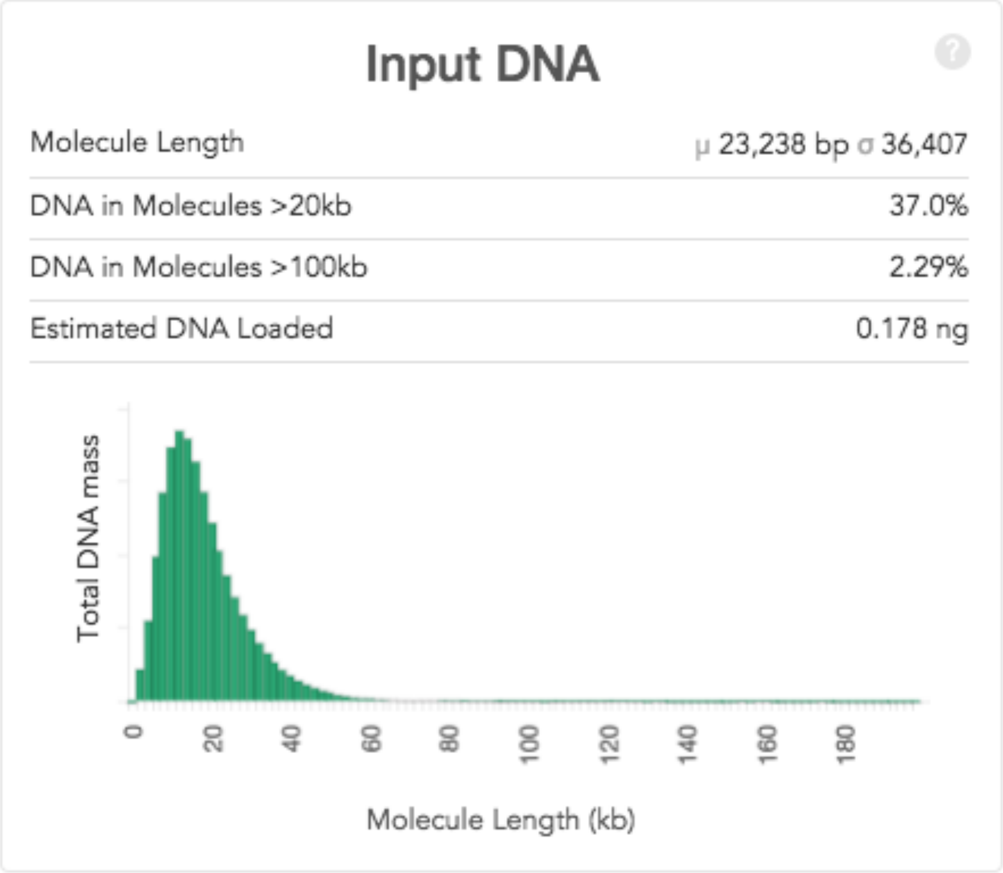
- 45.3X Sequencing depth
  – 2 lanes, 2x150 HiSeq 4k

Sample	Peak Size
Subject 1	26,169 BP
Subject 2	20,976 BP



Sequencing

Number of Reads	1,283,597,058
Median Insert Size	353 bp
Mean Depth	45.3 X
Zero Coverage	7.58%
Mapped Reads	90.8%
PCR Duplication	11.7%



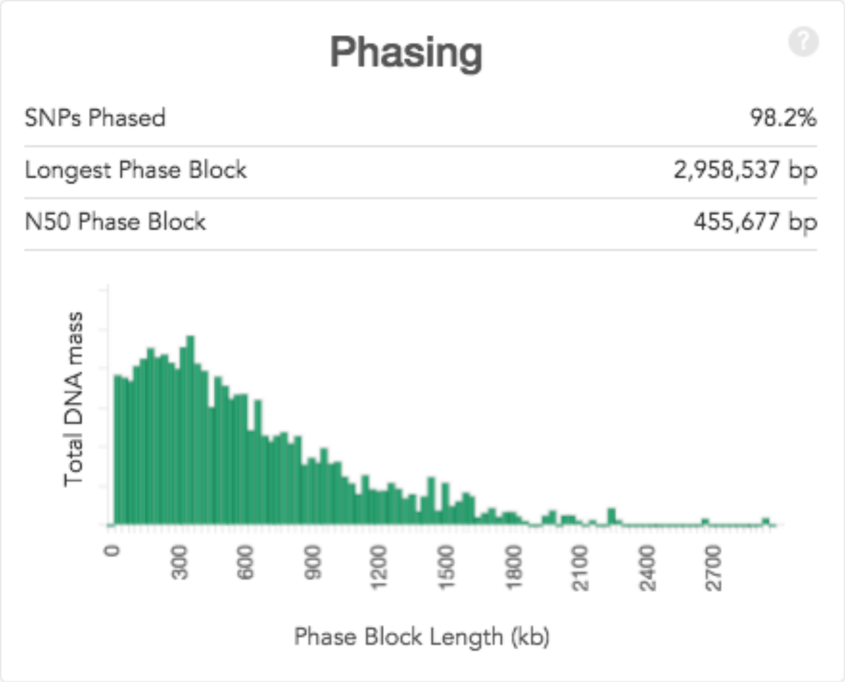
Input DNA

Molecule Length	μ 23,238 bp σ 36,407
DNA in Molecules >20kb	37.0%
DNA in Molecules >100kb	2.29%
Estimated DNA Loaded	0.178 ng

Summary: Subject 2

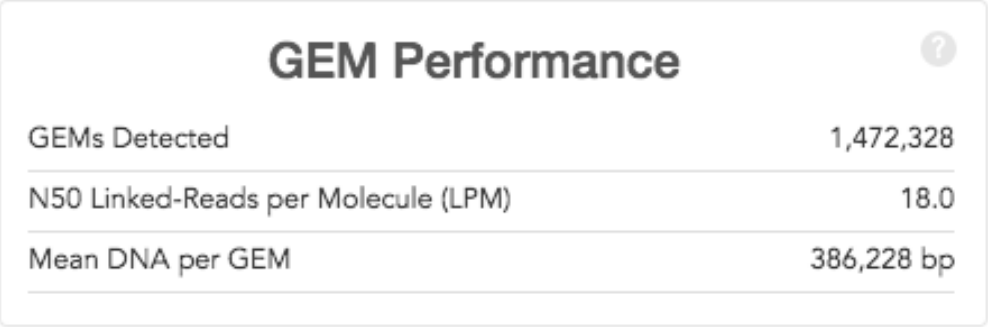
- 1.5 M GEMs detected
- 18 N50 LPM

- 456 k N50 phase block



Phasing

SNPs Phased	98.2%
Longest Phase Block	2,958,537 bp
N50 Phase Block	455,677 bp



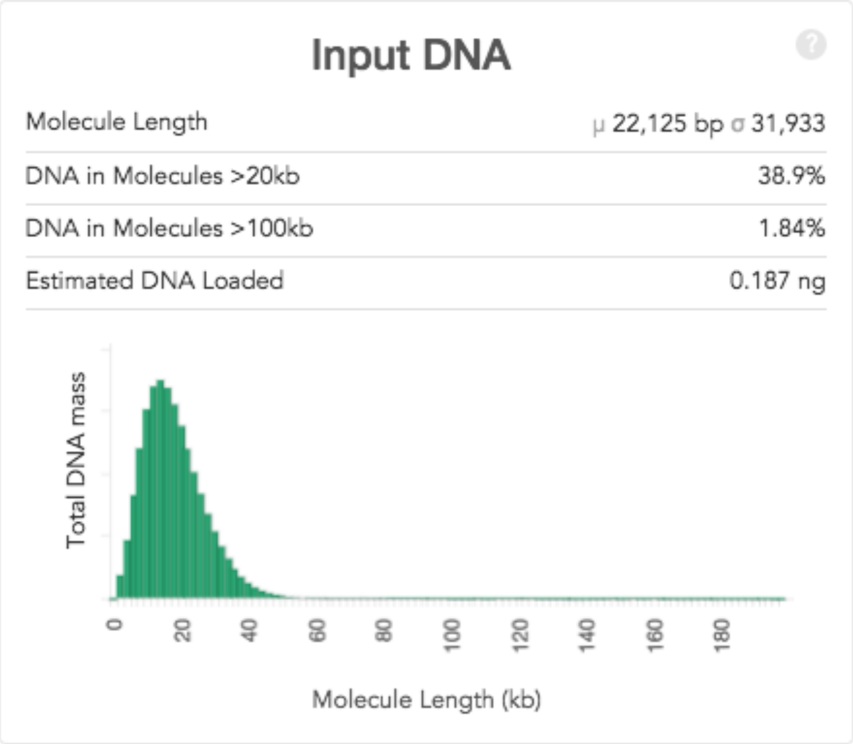
GEM Performance

GEMs Detected	1,472,328
N50 Linked-Reads per Molecule (LPM)	18.0
Mean DNA per GEM	386,228 bp

Comparison to NA12878 HMW control

- EA Qiagen MagAttract protocol and chemistry
  - ~95 kb mean DNA molecule length

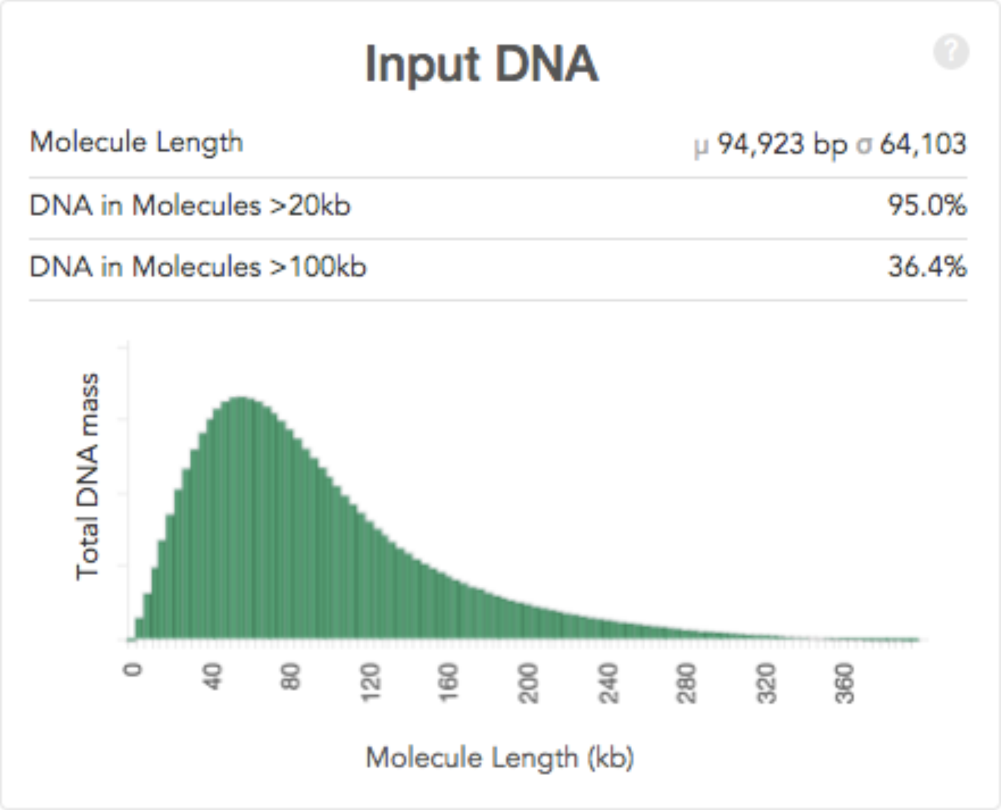
Subject 1



Input DNA

Molecule Length	μ 22,125 bp σ 31,933
DNA in Molecules >20kb	38.9%
DNA in Molecules >100kb	1.84%
Estimated DNA Loaded	0.187 ng

NA12878



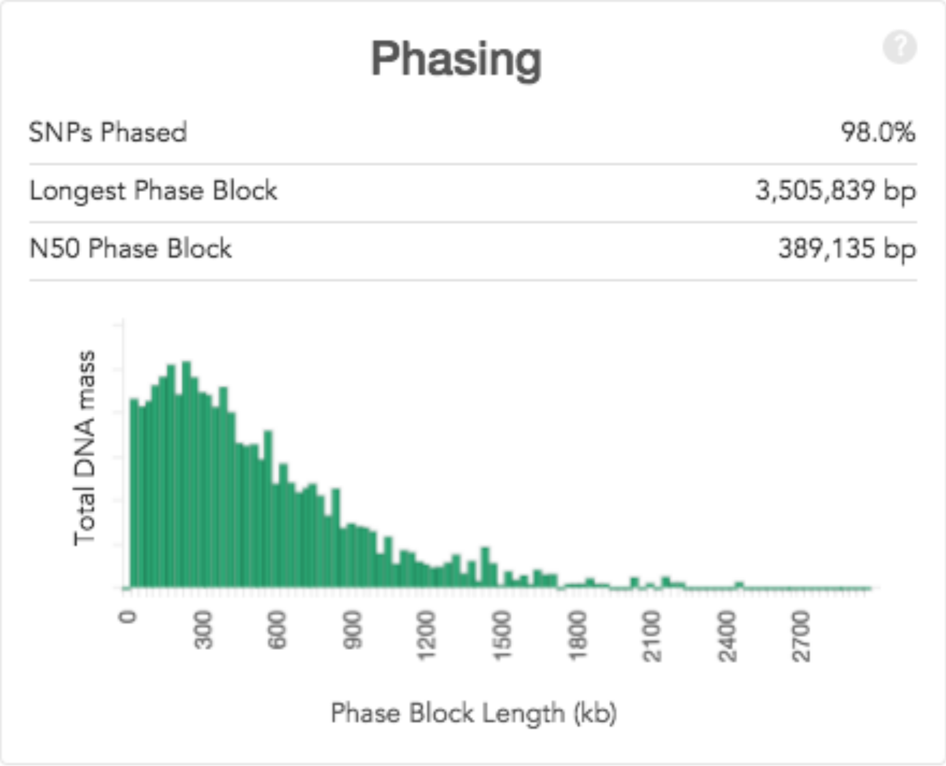
Input DNA

Molecule Length	μ 94,923 bp σ 64,103
DNA in Molecules >20kb	95.0%
DNA in Molecules >100kb	36.4%

Comparison to NA12878 HMW control

- 24X increase in N50 phase block length

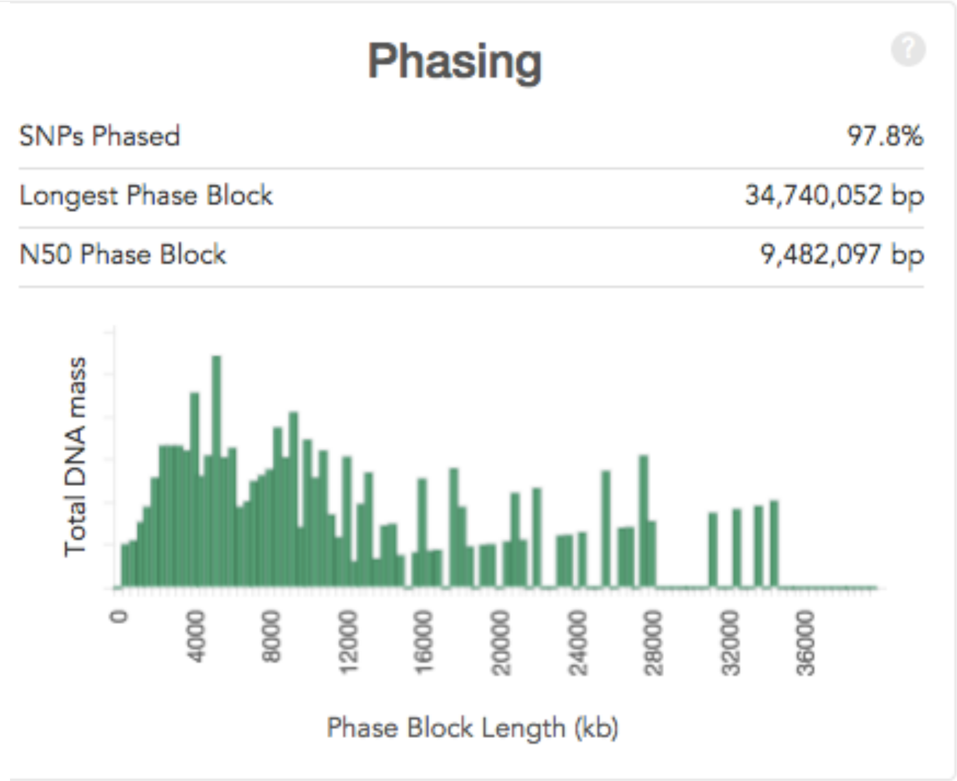
Subject 1



Phasing

SNPs Phased	98.0%
Longest Phase Block	3,505,839 bp
N50 Phase Block	389,135 bp

NA12878

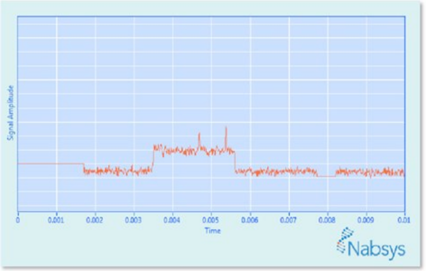


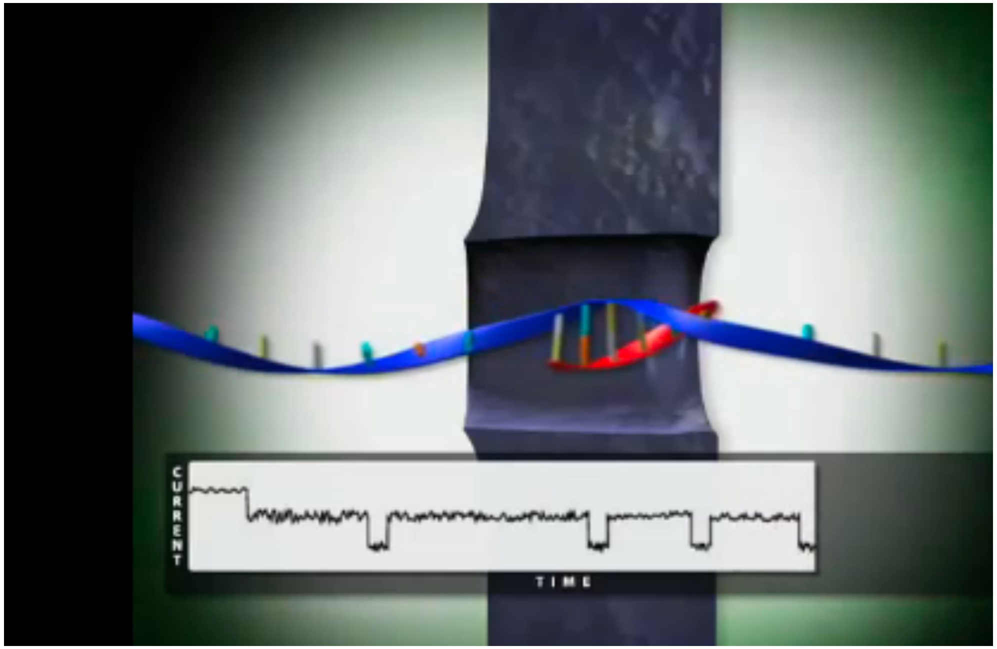
Phasing

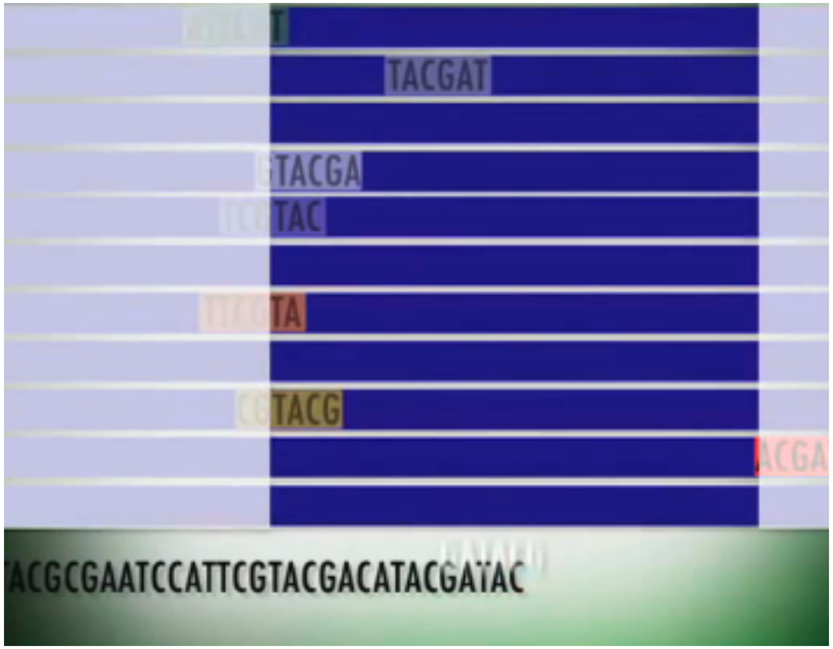
SNPs Phased	97.8%
Longest Phase Block	34,740,052 bp
N50 Phase Block	9,482,097 bp

Emerging Technologies

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.

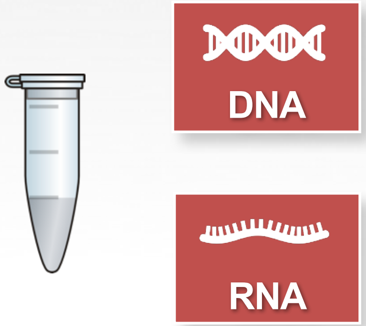
Nanostring's Hyb & Seq

Simple Workflow

No library preparation or amplification required

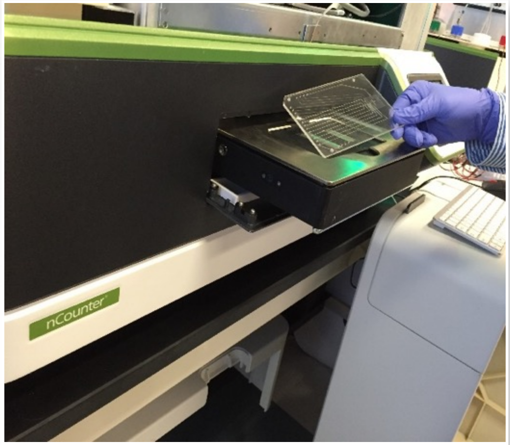
<30 minutes of hands on time Flexible input type (tissue, swabs, cells, etc.)

Single Tube Assay



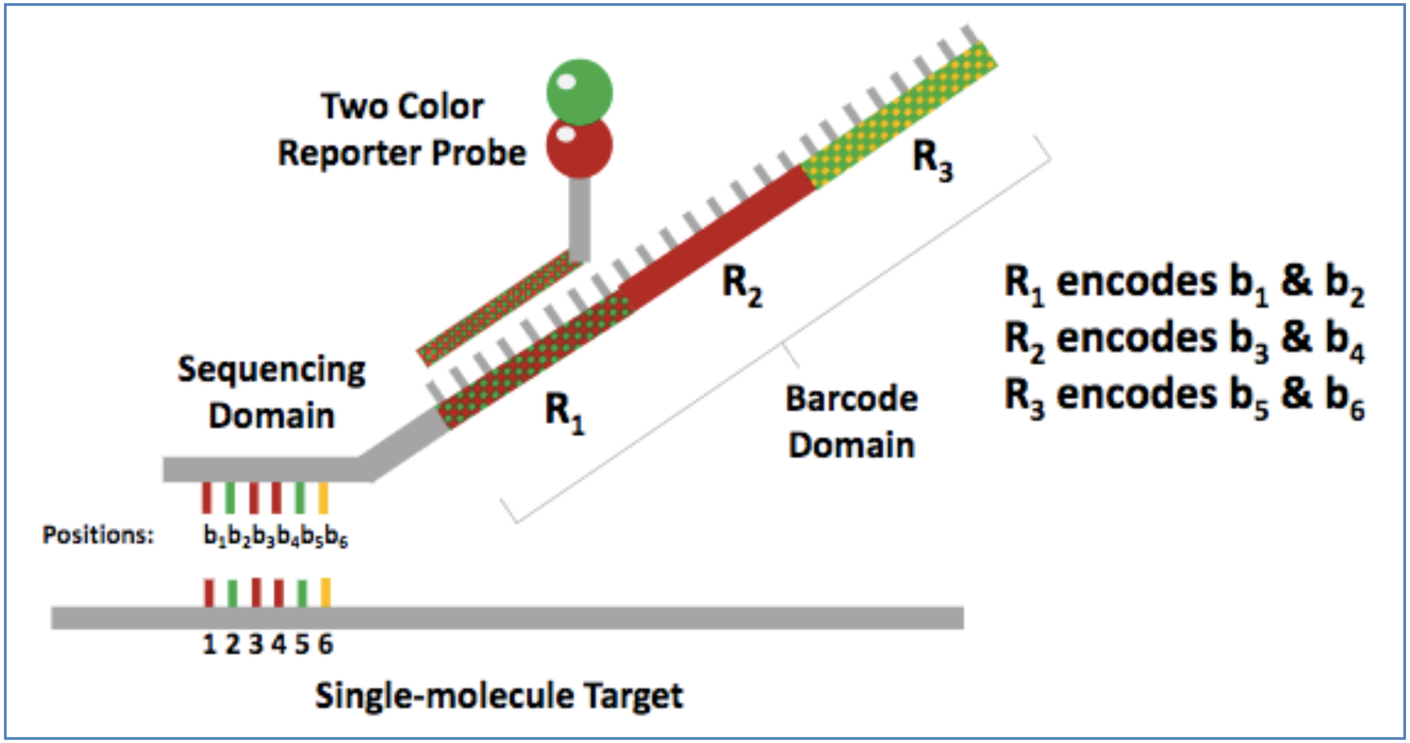
Enabling simultaneous and direct DNA and DNA sequencing

Clinically-relevant Timeframe



Sample-to-results in 4 hrs

Hyb & Seq



R_1 encodes b_1 & b_2
R_2 encodes b_3 & b_4
R_3 encodes b_5 & b_6

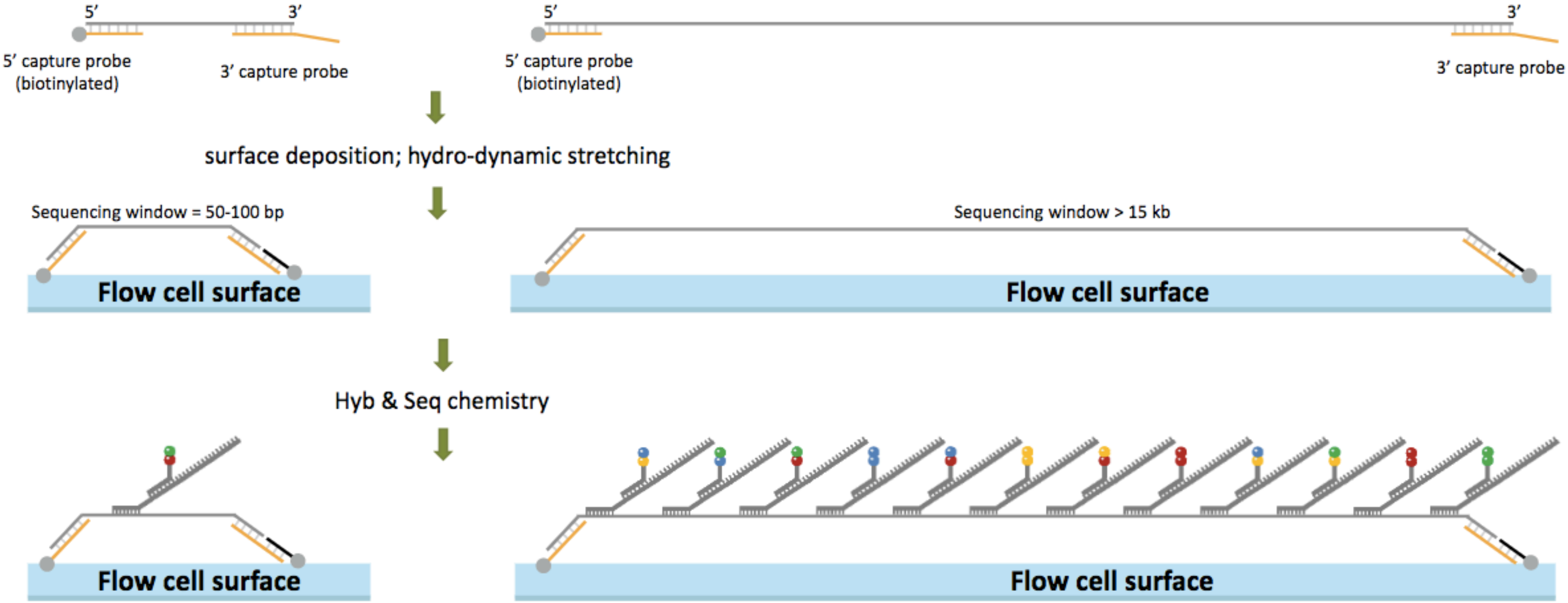
Sequencing Probes:

- Sequencing domain base-pairs with single-molecule target
- Barcode domain has three regions (R_1, R_2, R_3) encoding hexamer sequence
- Set of 4096 sequencing probes enables sequencing of any target sequence

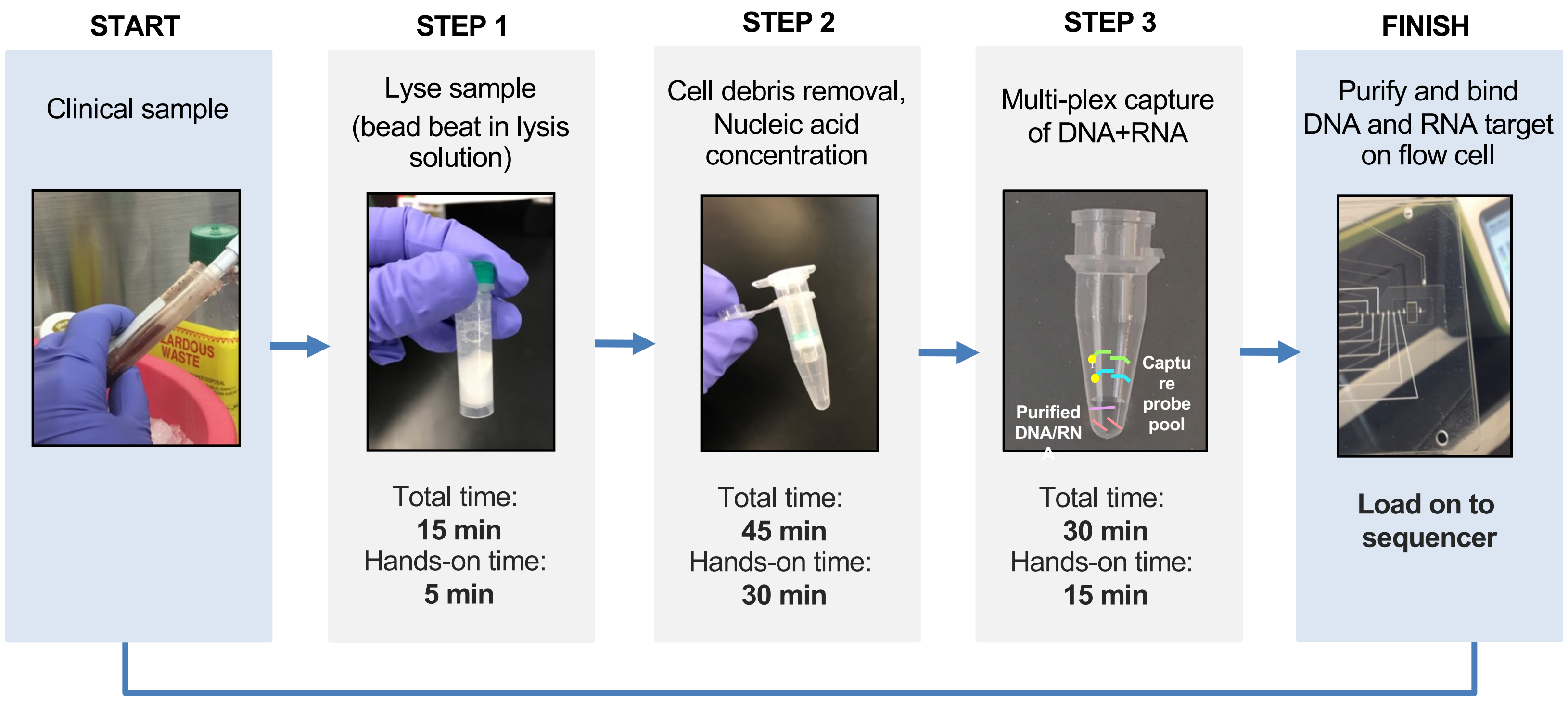
Two-color Reporter Probes:

- Three reporter probes bind sequentially to barcode domain (R_1, R_2 and R_3)
- Each reporter probe represents a dinucleotide sequence

Long and Short reads possible (up to 33kb)


5′
3′
5′ capture probe (biotinylated)
3′ capture probe
5′
3′
5′ capture probe (biotinylated)
3′ capture probe
surface deposition; hydro-dynamic stretching
Sequencing window = 50-100 bp
Sequencing window > 15 kb
Flow cell surface
Flow cell surface
Hyb & Seq chemistry
Flow cell surface
Flow cell surface

Clinical Sample Processing



START

Clinical sample

STEP 1

Lyse sample (bead beat in lysis solution)

Total time:
15 min
Hands-on time:
5 min

STEP 2

Cell debris removal, Nucleic acid concentration

Total time:
45 min
Hands-on time:
30 min

STEP 3

Multi-plex capture of DNA+RNA

Total time:
30 min
Hands-on time:
15 min

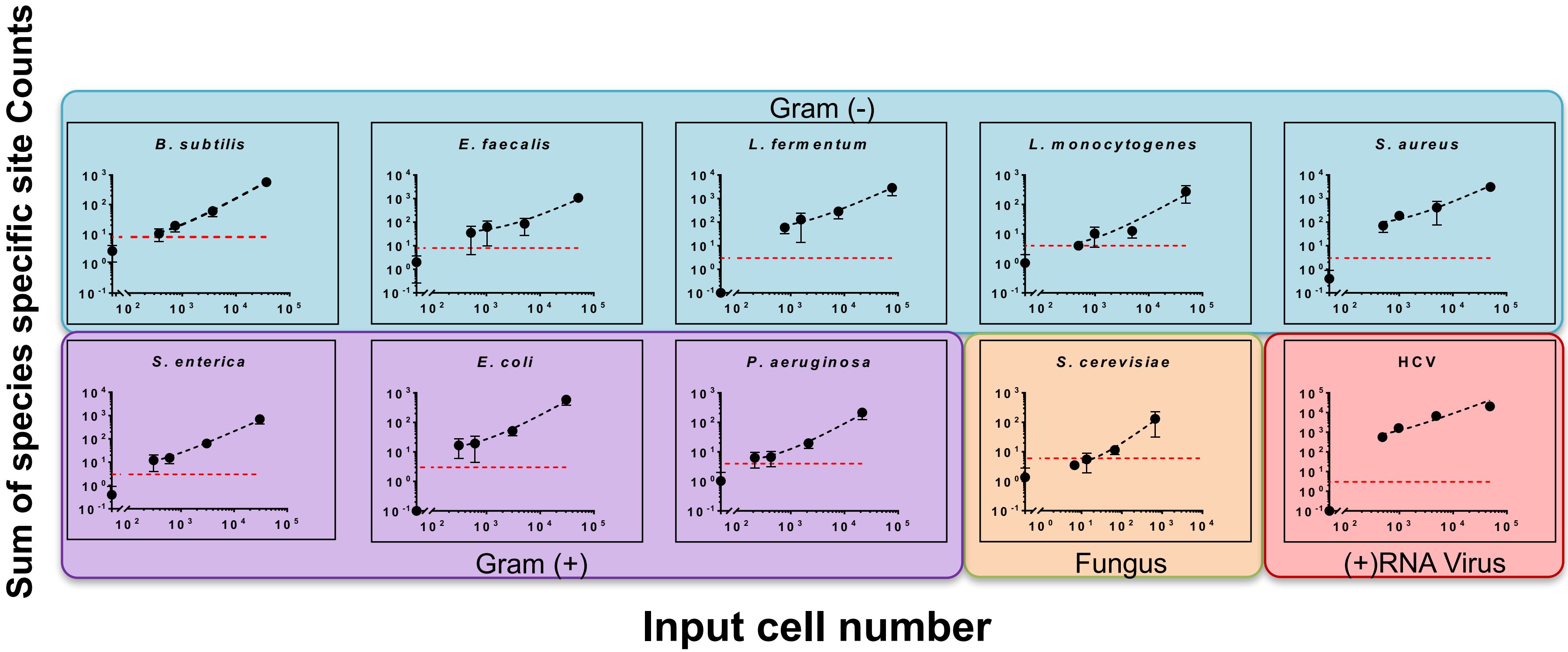
FINISH

Purify and bind DNA and RNA target on flow cell

Load on to sequencer

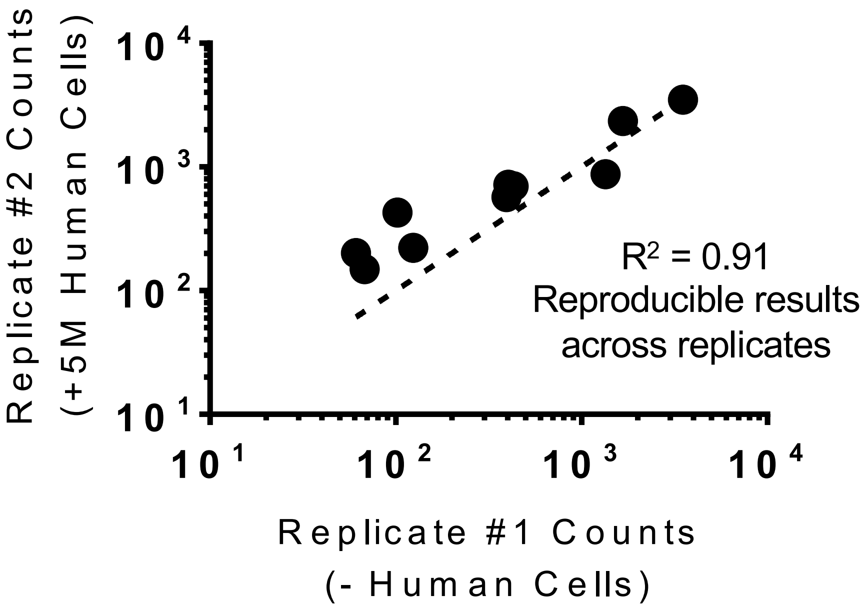
Completed in 90 min
No amplification, No library preparation

Assay Validation: Limit of Detection



Hyb & Seq simultaneously detected 10 pathogens at ≤ 1000 cells/ml from a same sample using a single tube assay

Assay Validation: No cross reactions with human DNA



$R^2 = 0.91$
Reproducible results across replicates

- Amplification-free sequencing of pathogens even in the presence of human cell background (5 million cells, cell line GM19240/NA12878)
- High concordance of sequencing results with or without excess of human cells background
- Same workflow regardless of sample background (swab, cells, tissue, etc)
- Eliminates reads waste due to carrier human DNA/RNA

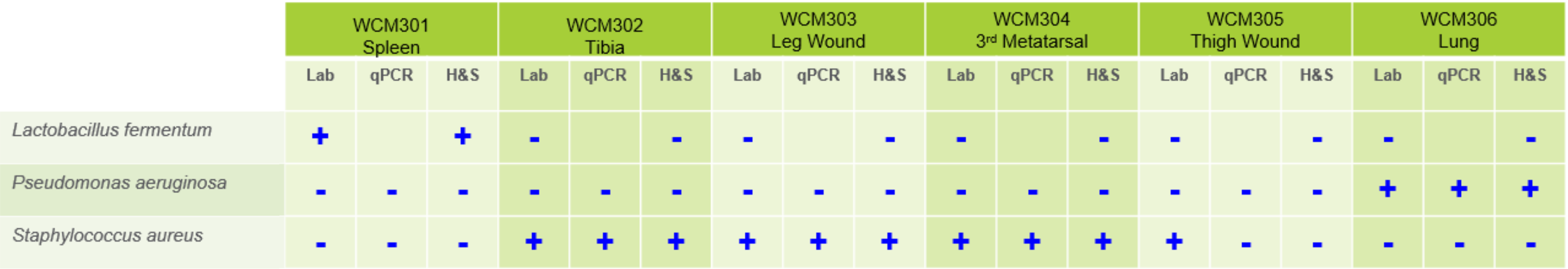
Clinical samples from WCM

Sample Name	Site	Final microbiology report
WCM300	Head Epidural Fluid	Sparse P. aeruginosa, Sparse Enterococcus faecalis
WCM301	Spleen	Sparse E. coli, Sparse Proteus mirabilis, Sparse Lactobacillus sp. (no final speciation)*
WCM302	R tibia	Sparse MRSA
WCM303	R leg wound	MSSA
WCM304	R 3rdmetatarsal	Sparse Proteus mirabilis, Few Staphylococcus agalactiae, Sparse MSSA
WCM305	L thigh wound	MSSA
WCM306	Lung	Many Pseudomonas aeruginosa

With Lars Westblade

Precision Clinical Metagenomics
IRB#: 1606017347

Hyb & Seq Sequencing Results



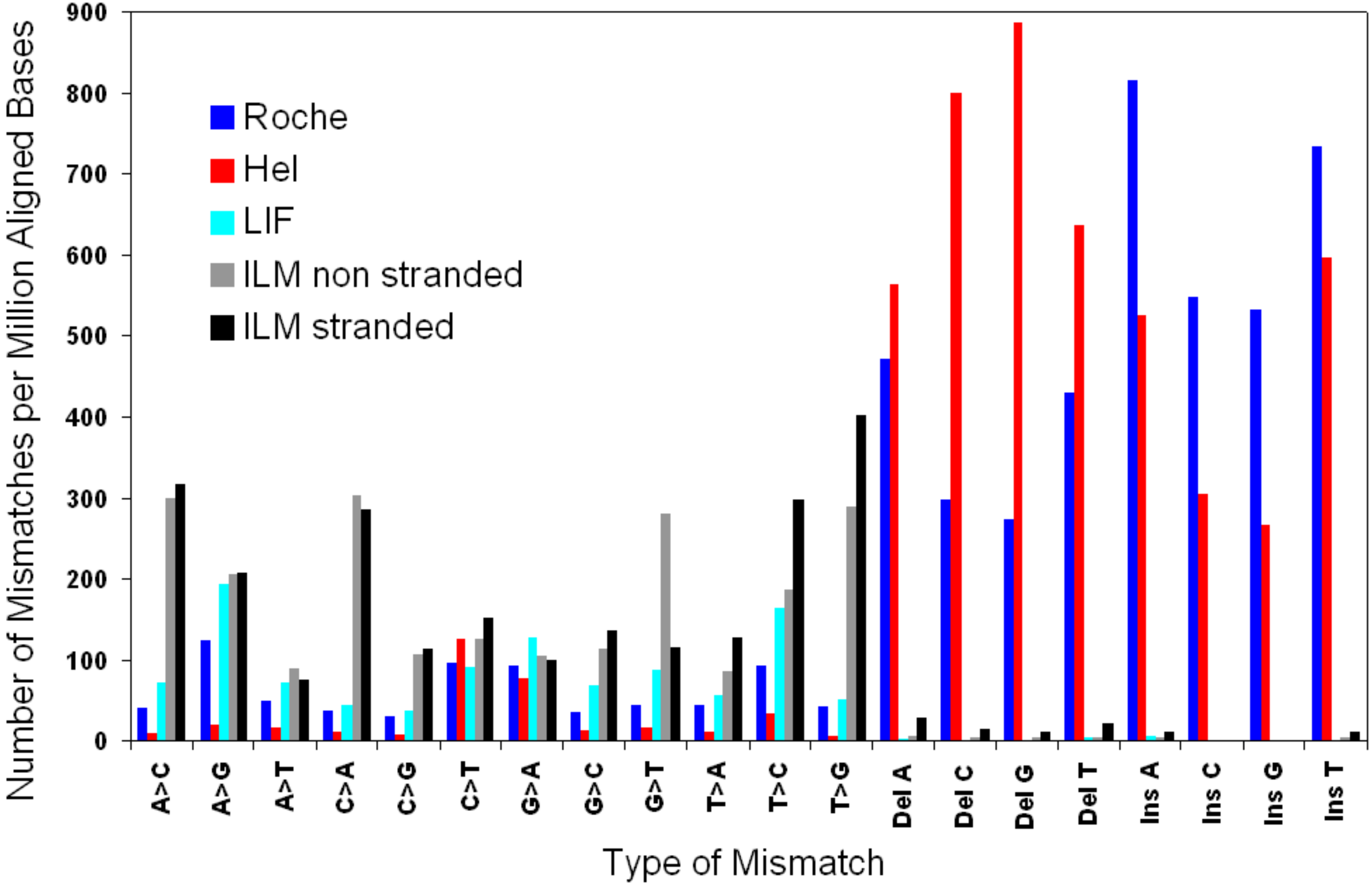
	WCM301 Spleen			WCM302 Tibia			WCM303 Leg Wound			WCM304 3rd Metatarsal			WCM305 Thigh Wound			WCM306 Lung		
	Lab	qPCR	H&S	Lab	qPCR	H&S	Lab	qPCR	H&S	Lab	qPCR	H&S	Lab	qPCR	H&S	Lab	qPCR	H&S
Lactobacillus fermentum	+		+	-		-	-		-	-		-	-		-	-		-
Pseudomonas aeruginosa	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+
Staphylococcus aureus	-	-	-	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-

- Six different clinical samples were analyzed
- Five positive calls across three kingdom of organisms
- High concordance with **pathology lab analysis (98%; 65/66)** and **100% concordance with PCR** analysis
- Simultaneously detected intra- and inter-species DNA and RNA
- *One discordant same was only found in the broth and flagged as ambiguous

Each Platform has various sources of noise, and thus Error

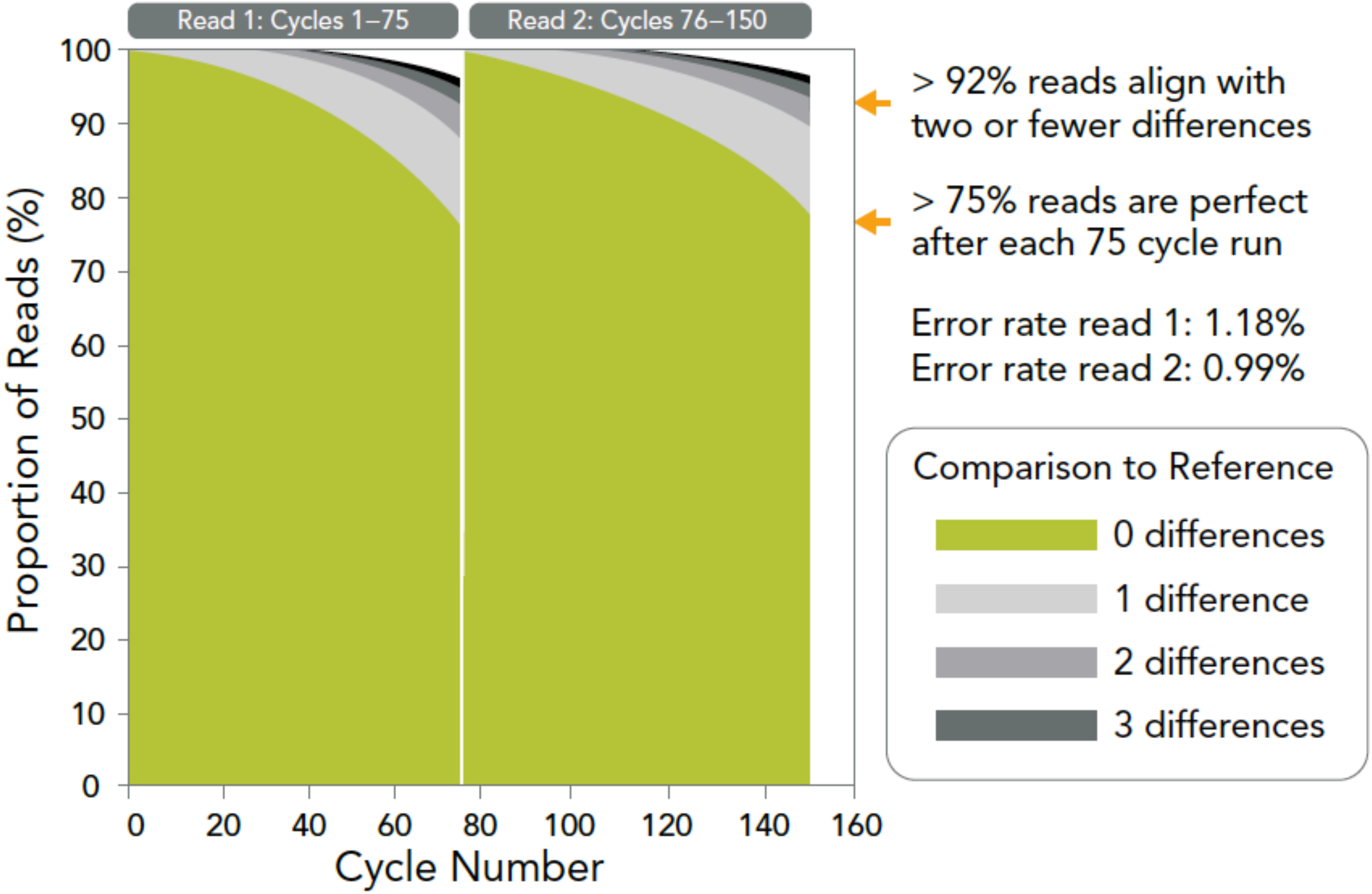
- De-Phasing
  - Lagging strand dephasing from incomplete extension
  - Leading strand dephasing from over-extension
- Dark Nucleotides
- Polymerase errors (10^{-5} to 10^{-7})
- Single molecule challenges
  - High noise
  - Polymerase "wiggling" from tail
- Platform-specific errors
  - Illumina more 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

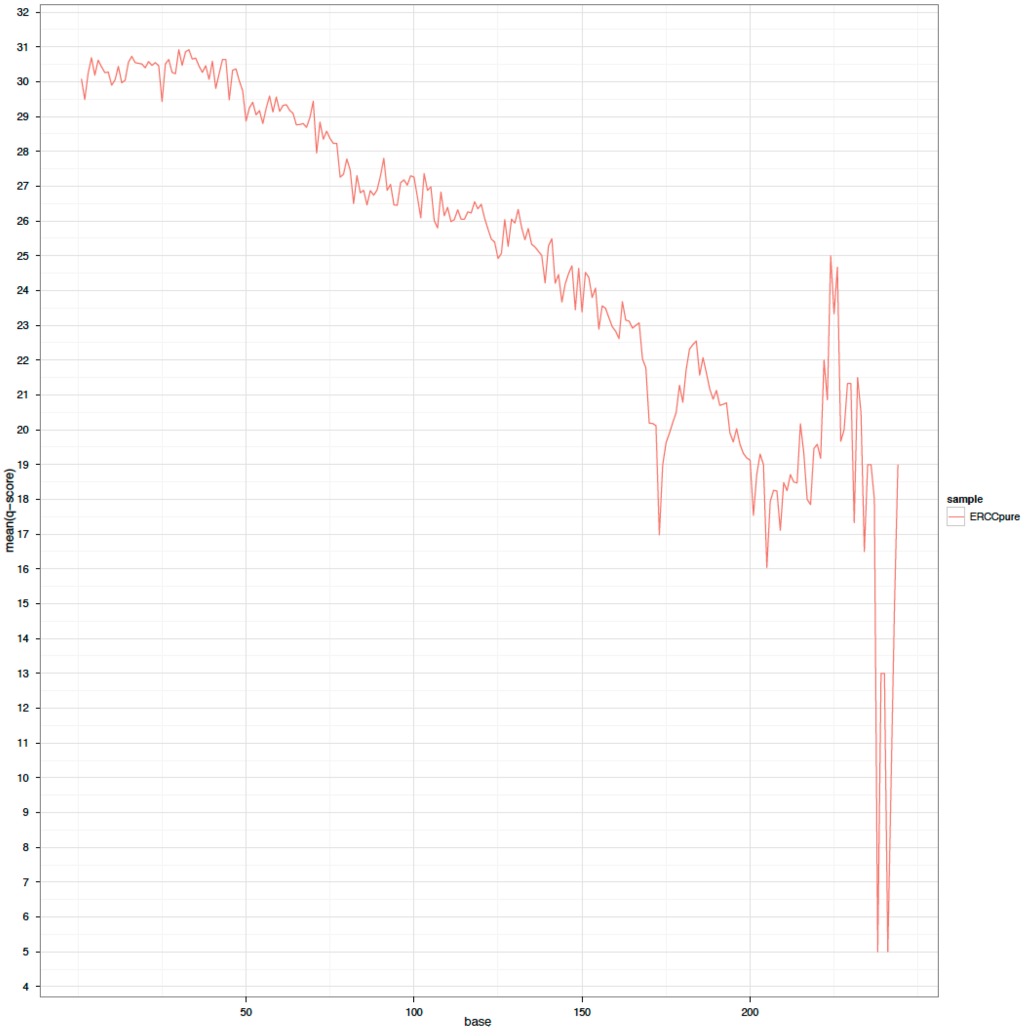


> 92% reads align with two or fewer differences

> 75% reads are perfect after each 75 cycle run

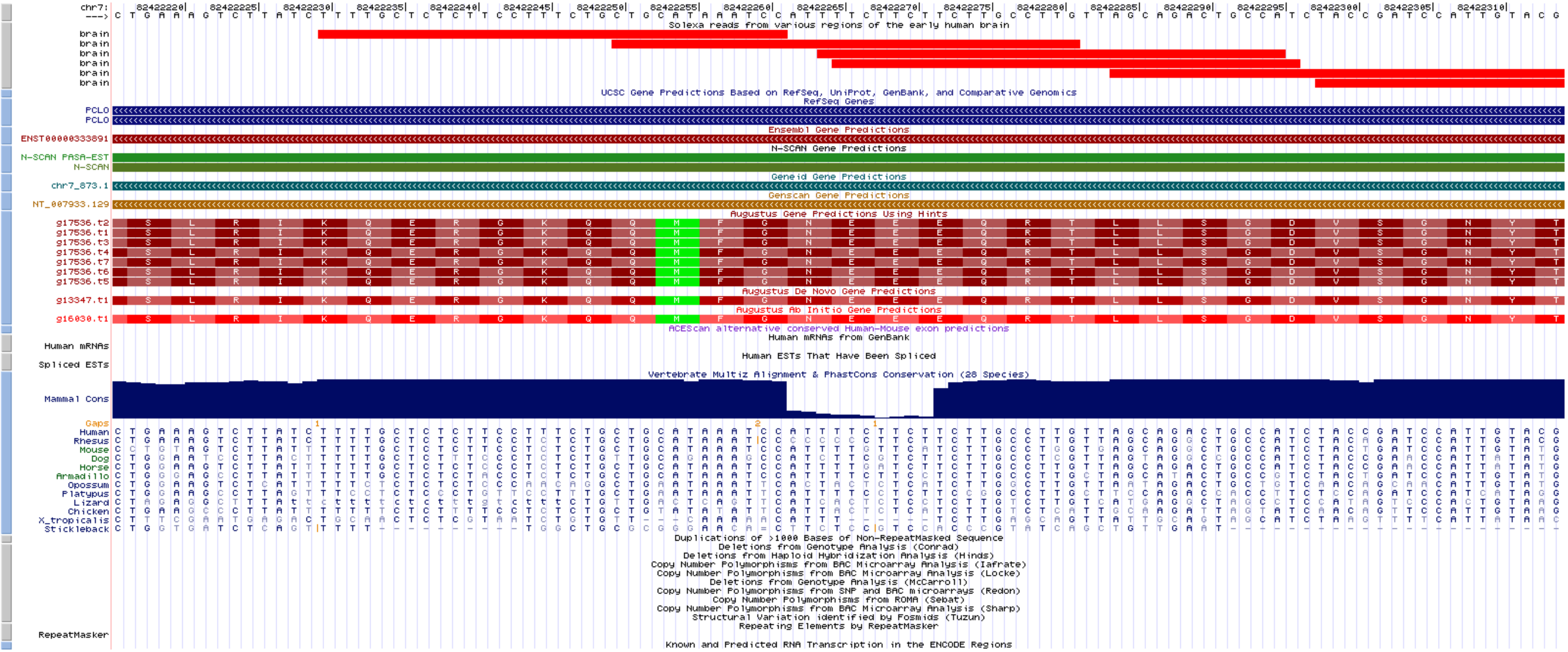
Error rate read 1: 1.18%
Error rate read 2: 0.99%

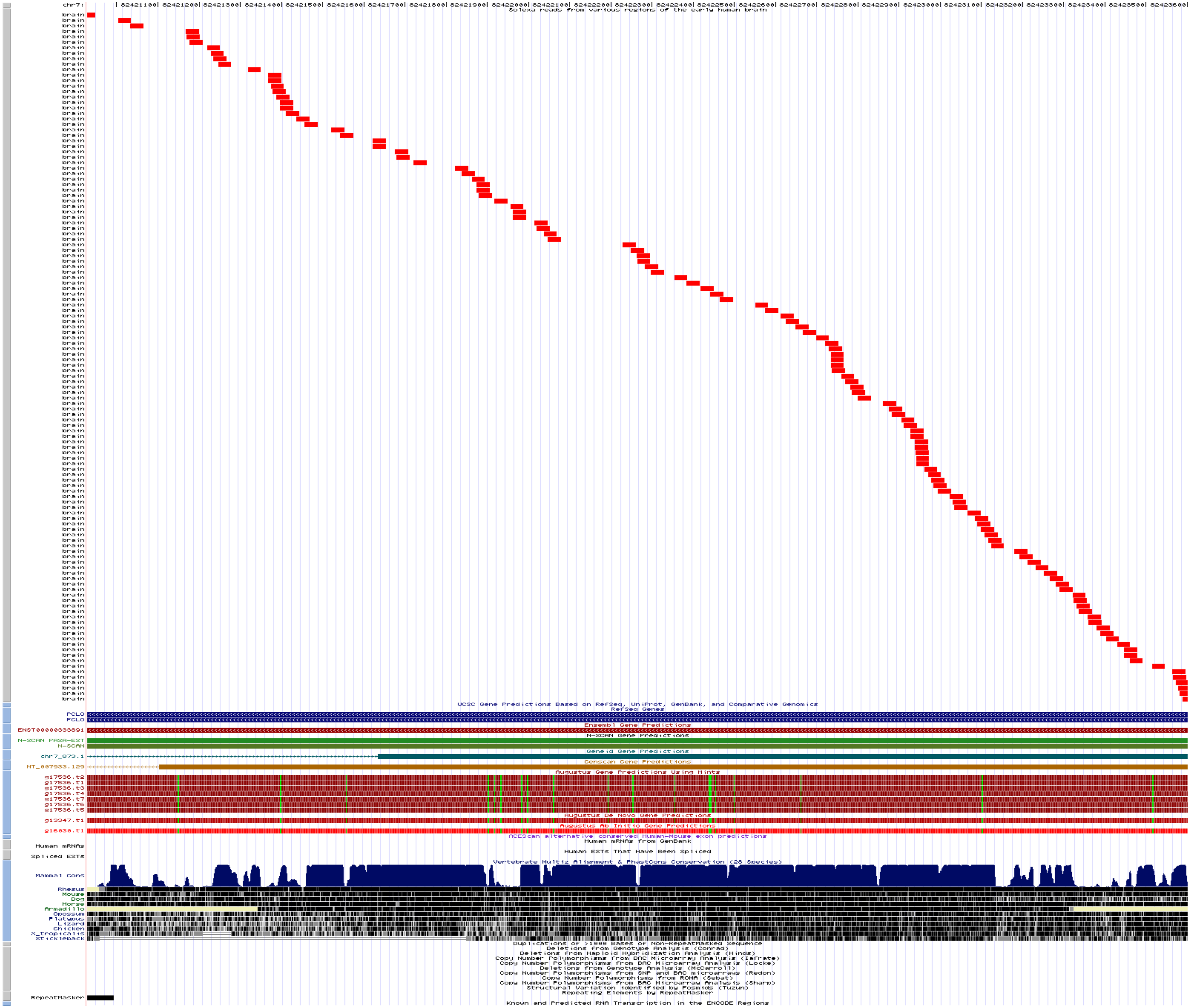
Many platforms are cycle-dependent on error rate - ION



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
ENST00000333891
Ensembl Gene Predictions
N-SCAN PASA-EST
N-SCAN
N-SCAN Gene Predictions
chr7_873.1
Geneid Gene Predictions
NT_007933.129
Genscan Gene Predictions
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
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
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
UCSC Gene Predictions Based on RefSeq, UniProt, GenBank, and Comparative Genomics
PCLO
PCLO
RefSeq Genes
ENST00000333891
Ensembl Gene Predictions
N-SCAN PASA-EST
N-SCAN
N-SCAN Gene Predictions
chr7_873.1
Geneid Gene Predictions
NT_007933.129
Genscan Gene Predictions
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 $Q_{value} = (-10\log_{10}(0.01))$

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

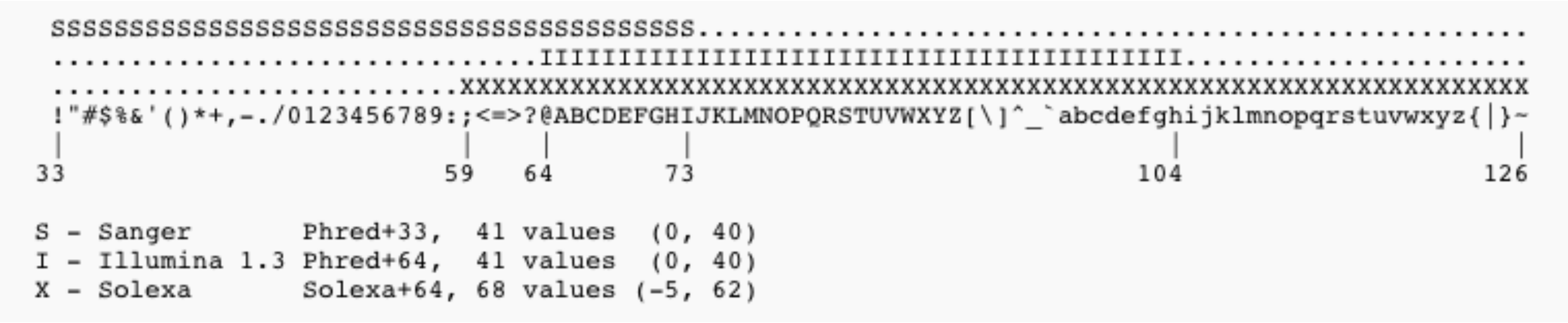
If p=0.001, then $Q_{value} = (-10\log_{10}(0.001))$

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

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.

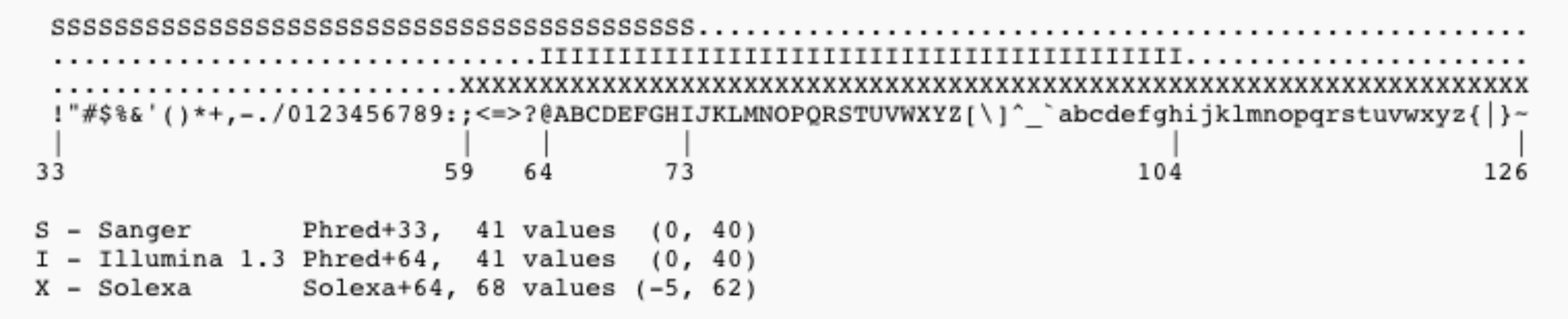
The first 32 symbols in ASCII are control characters, so we start at 33.



```
 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.

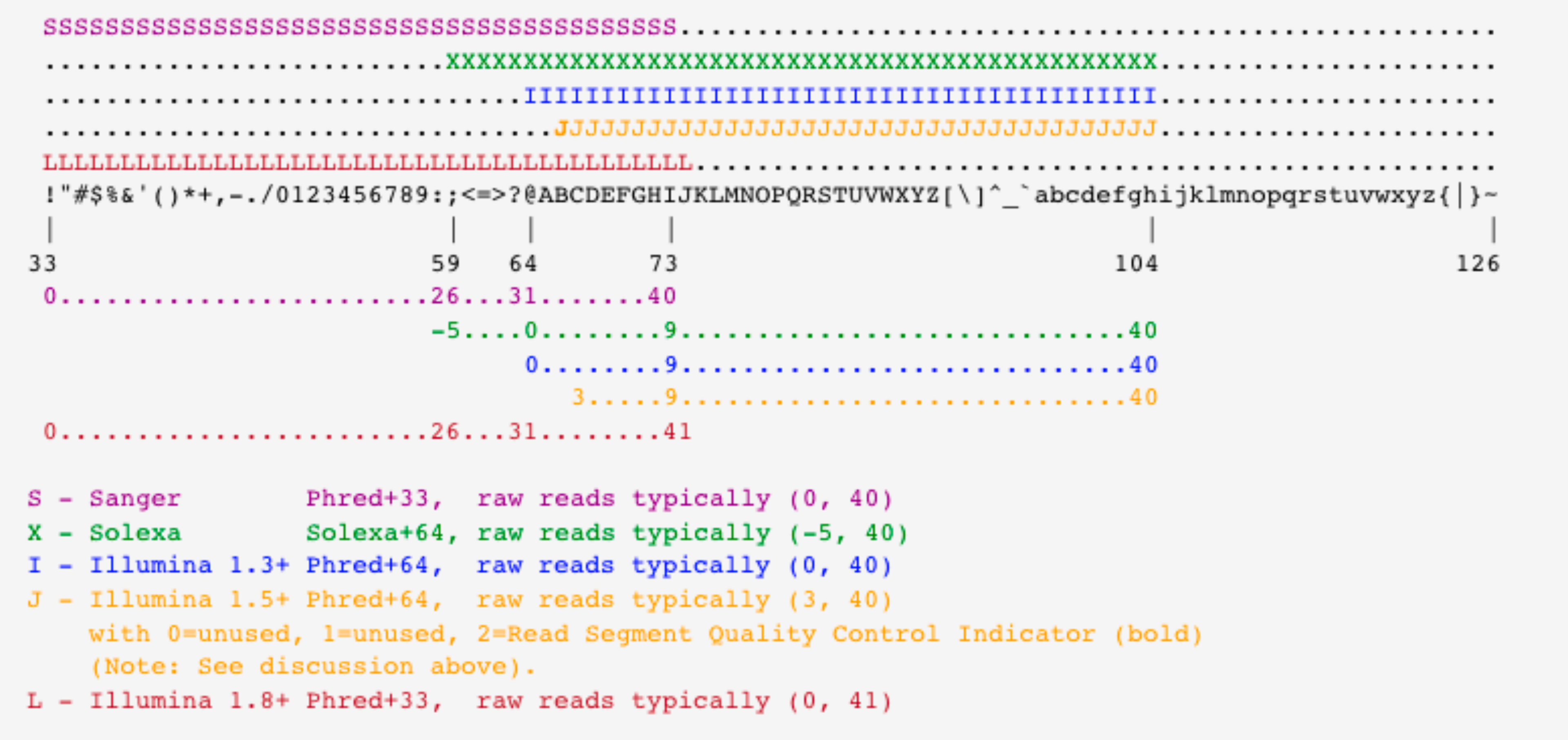
If your ASCII character is ‘h’, then 104-64=40, so

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

-4.0 = $\log_{10}p$

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

Phred-Based Base Quality Today

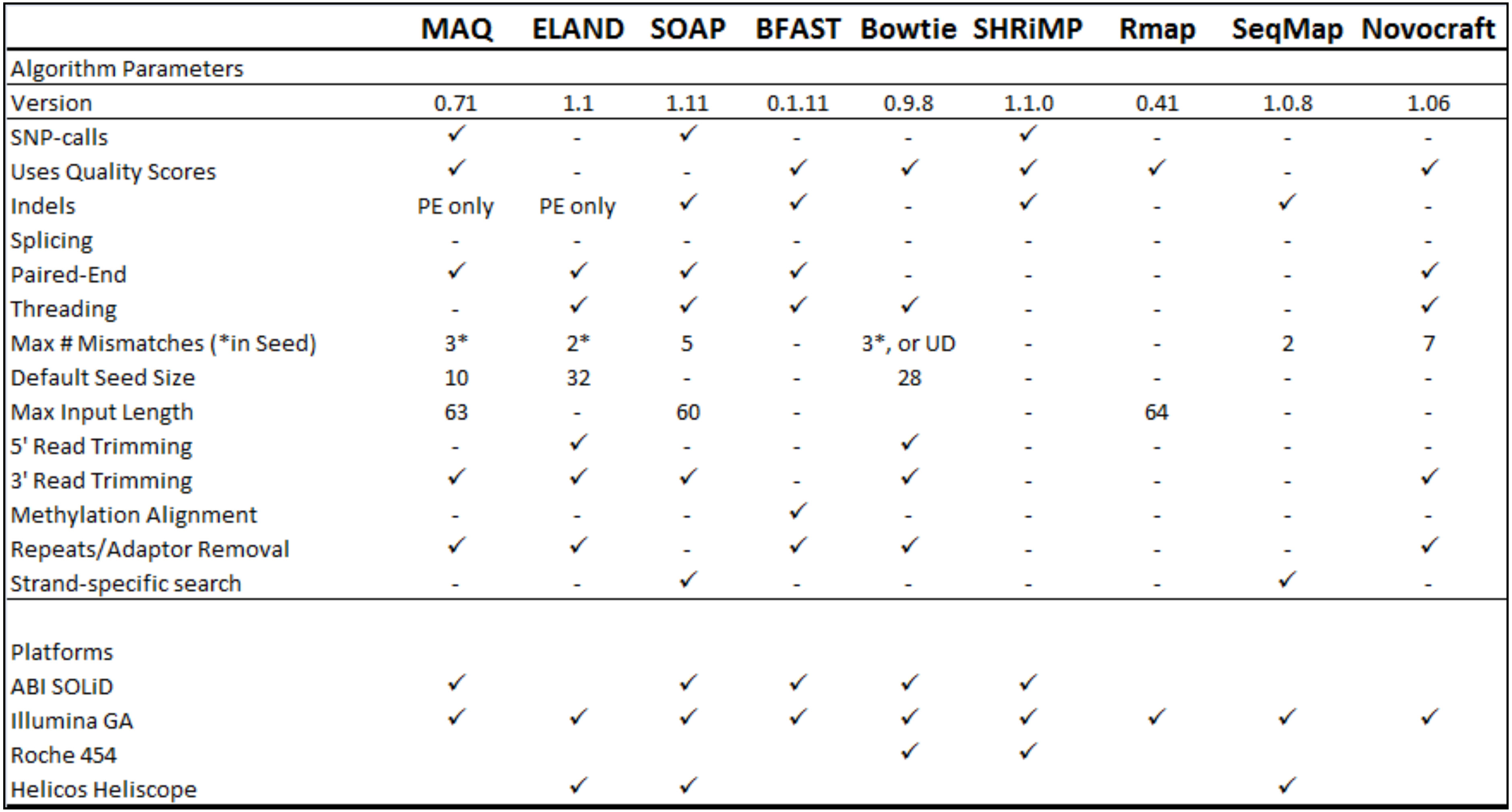


```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
  ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
  .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
  LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
 33                        59   64       73                            104                   126
  0........................26...31.......40
                           -5....0........9.............................40
                                 0........9.............................40
                                    3.....9.............................40
  0........................26...31........41

 S - Sanger        Phred+33,  raw reads typically (0, 40)
 X - Solexa        Solexa+64, raw reads typically (-5, 40)
 I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
 J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
     with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
     (Note: See discussion above).
 L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

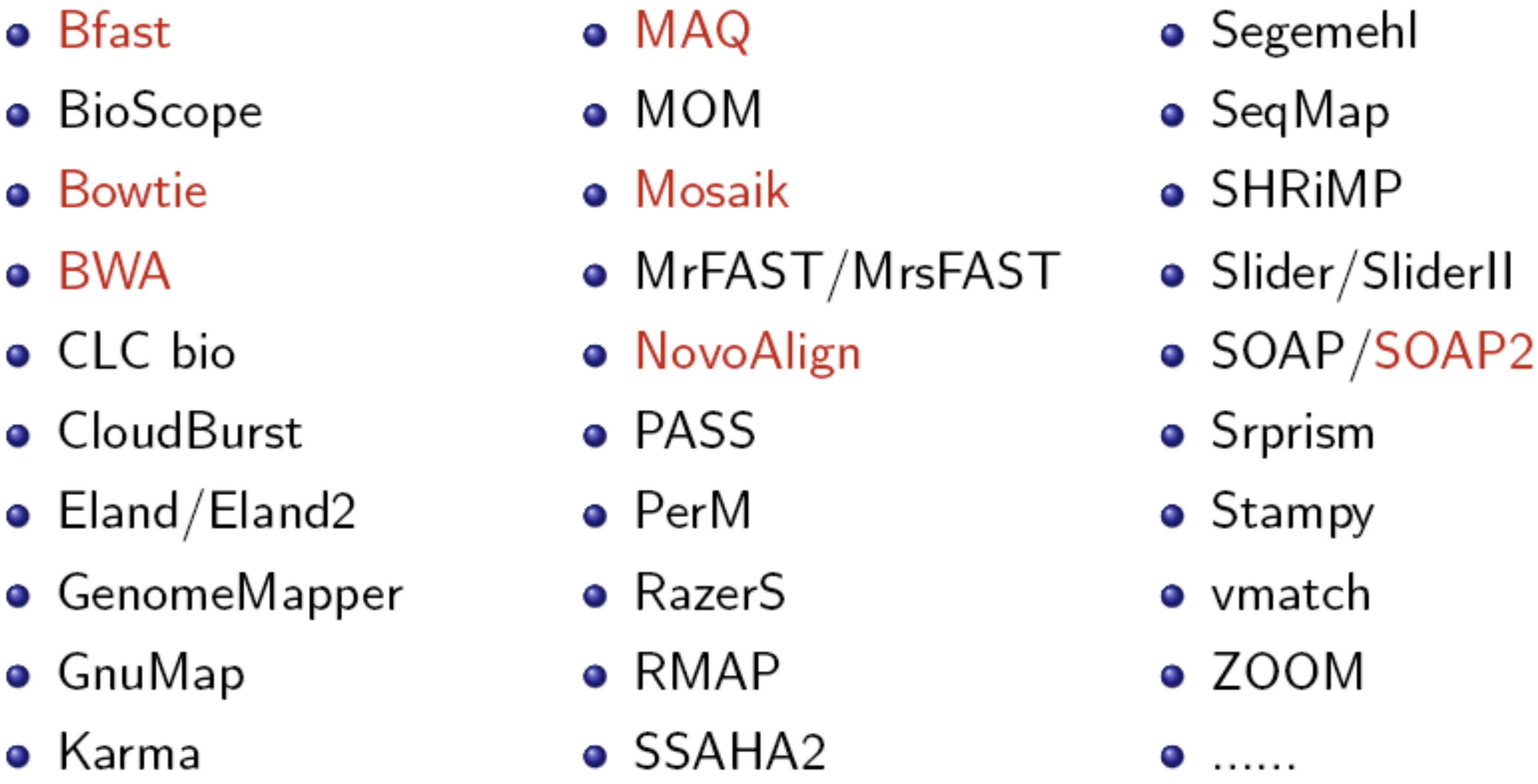
Cock et al (2009) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research,

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		✓	✓					✓	

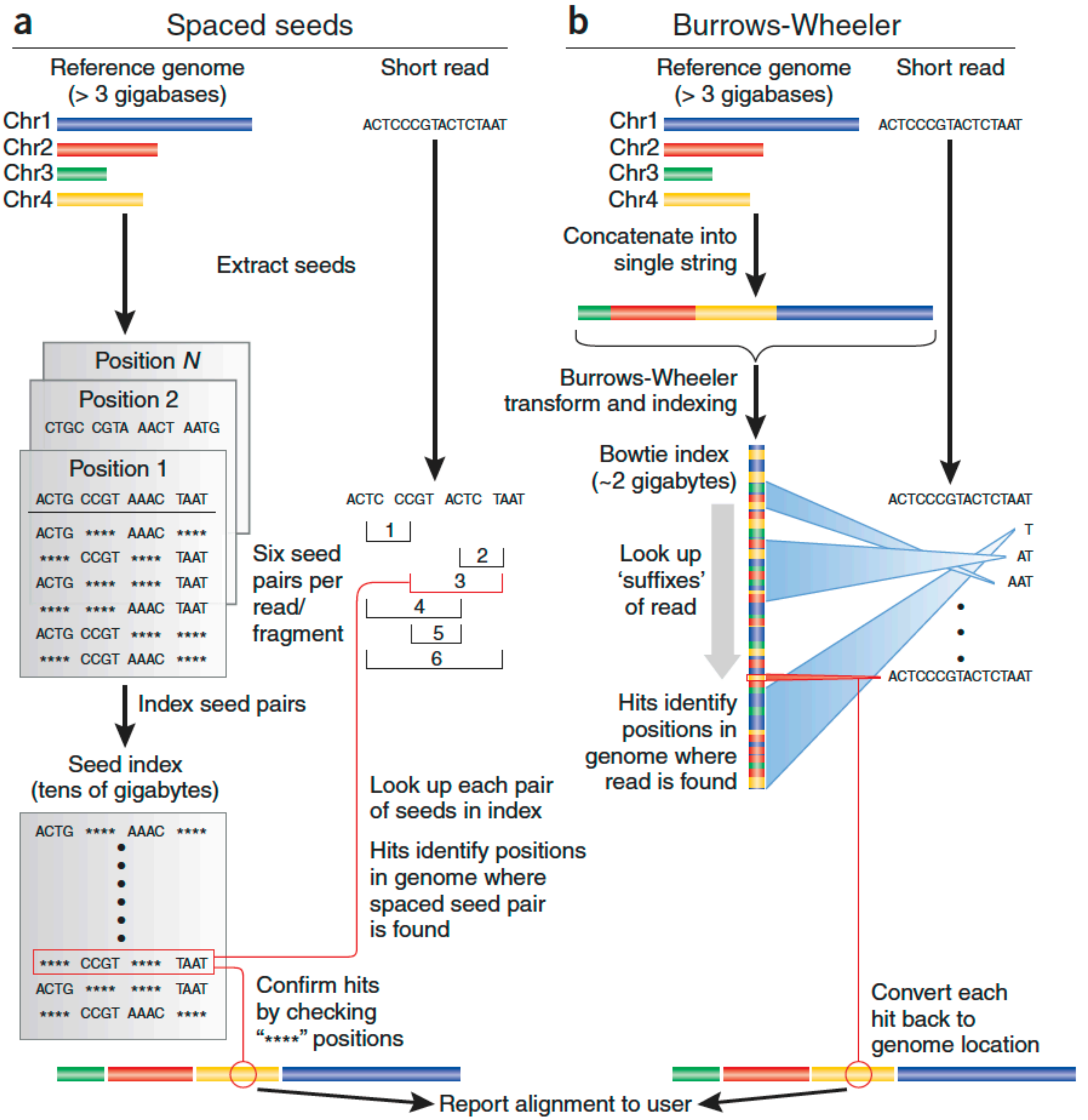
Many Options for Alignment - 2018



- 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
- ......

Li et al, 2010

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/

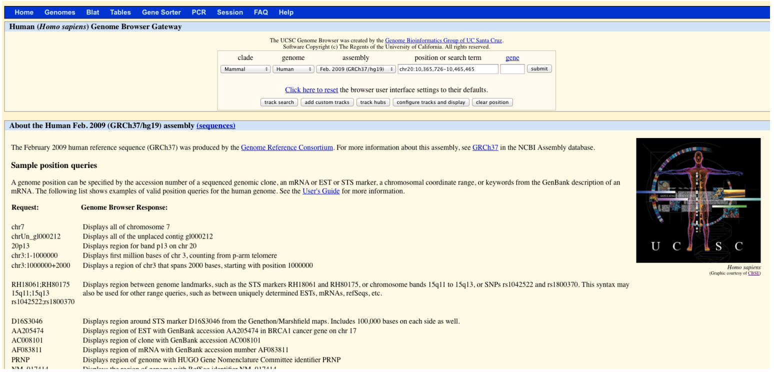
Li H. and Durbin R. “Fast and accurate short read alignment with Burrows-Wheeler transform.” (2009) *Bioinformatics*, 25, 1754-60.

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

Plan ahead for all genomes to be sequenced and available

ETHAN HAWKE UMA THURMAN JUDE LAW
FROM ANDREW NICCOL, WRITER OF "THE TRUMAN SHOW"
GATTACA
THERE IS NO GENE FOR THE HUMAN SPIRIT
DVD

However, your internet browser home page will likely change:



Human (*Homo sapiens*) Genome Browser Gateway

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.
Software Copyright (c) The Regents of the University of California. All rights reserved.

clade	genome	assembly	position or search term	gene
Mammal	Human	Feb. 2009 (GRCh37/hg19)	chr20:10,365,726-10,465,465	

Click here to reset the browser user interface settings to their defaults.

About the Human Feb. 2009 (GRCh37/hg19) assembly (sequences)

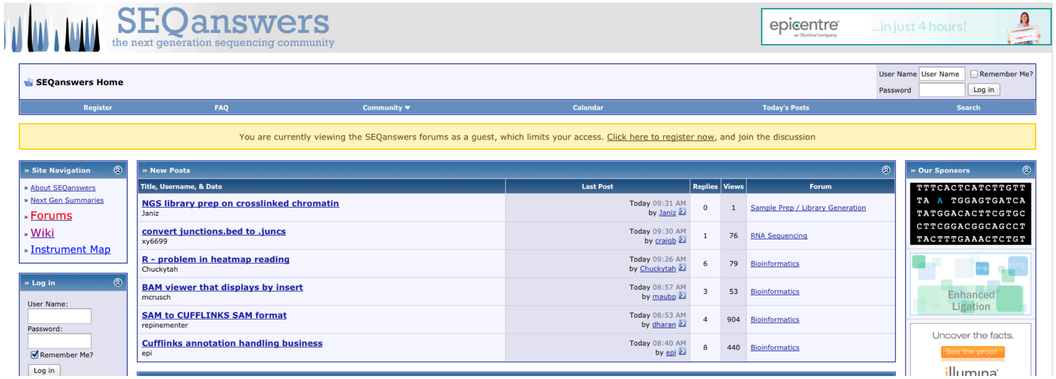
The February 2009 human reference sequence (GRCh37) was produced by the Genome Reference Consortium. For more information about this assembly, see GRCh37 in the NCBI Assembly database.

Sample position queries

A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the User's Guide for more information.

Request:	Genome Browser Response:
chr7	Displays all of chromosome 7
chrUn_gl000212	Displays all of the unplaced contig gl000212
20p13	Displays region for band p13 on chr 20
chr3:1-1000000	Displays first million bases of chr 3, counting from p-arm telomere
chr3:1000000+2000	Displays a region of chr3 that spans 2000 bases, starting with position 1000000
RH18061;RH80175 15q11;15q13 rs1042522;rs1800370	Displays region between genome landmarks, such as the STS markers RH18061 and RH80175, or chromosome bands 15q11 to 15q13, or SNPs rs1042522 and rs1800370. This syntax may also be used for other range queries, such as between uniquely determined ESTs, mRNAs, refSeqs, etc.
D16S3046	Displays region around STS marker D16S3046 from the Genethon/Marshfield maps. Includes 100,000 bases on each side as well.
AA205474	Displays region of EST with GenBank accession AA205474 in BRCA1 cancer gene on chr 17
AC008101	Displays region of clone with GenBank accession AC008101
AF083811	Displays region of mRNA with GenBank accession number AF083811
PRNP	Displays region of genome with HUGO Gene Nomenclature Committee identifier PRNP
NM_017414	Displays the region of genome with RefSeq identifier NM_017414

Homo sapiens
(Graphic courtesy of CBSE)



SEQanswers Home

User Name: User Name ☐ Remember Me?
Password:

You are currently viewing the SEQanswers forums as a guest, which limits your access. Click here to register now, and join the discussion

» Site Navigation

- About SEQanswers
- Next Gen Summaries
- Forums
- Wiki
- Instrument Map

» Log in

User Name:
Password:
☑ Remember Me?

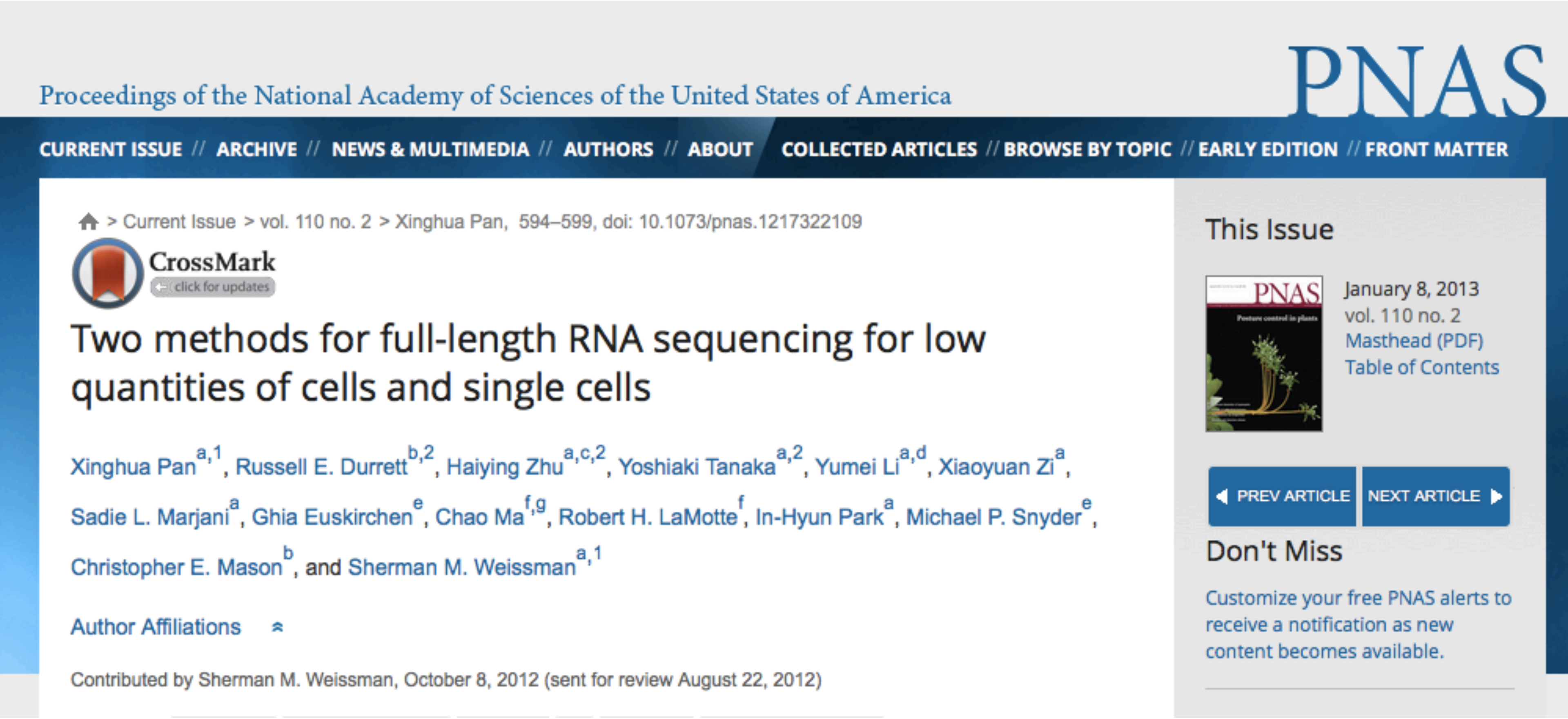
» New Posts

Title, Username, & Date	Last Post	Replies	Views	Forum
NGS library prep on crosslinked chromatin Janiz	Today 09:31 AM by Janiz	0	1	Sample Prep / Library Generation
convert junctions.bed to .juncs xy6699	Today 09:30 AM by craigb	1	76	RNA Sequencing
R - problem in heatmap reading Chuckytah	Today 09:26 AM by Chuckytah	6	79	Bioinformatics
BAM viewer that displays by insert mcrusch	Today 08:57 AM by maubp	3	53	Bioinformatics
SAM to CUFFLINKS SAM format repinementer	Today 08:53 AM by dharan	4	904	Bioinformatics
Cufflinks annotation handling business epi	Today 08:40 AM by epi	8	440	Bioinformatics

» Our Sponsors

Single cells

Used to be very hard to look at individual cells


Proceedings of the National Academy of Sciences of the United States of America
PNAS
CURRENT ISSUE // ARCHIVE // NEWS & MULTIMEDIA // AUTHORS // ABOUT
COLLECTED ARTICLES // BROWSE BY TOPIC // EARLY EDITION // FRONT MATTER
CrossMark
click for updates
PREV ARTICLE
NEXT ARTICLE

> Current Issue > vol. 110 no. 2 > Xinghua Pan, 594–599, doi: 10.1073/pnas.1217322109

Two methods for full-length RNA sequencing for low quantities of cells and single cells

Xinghua Pan[a,1], Russell E. Durrett[b,2], Haiying Zhu[a,c,2], Yoshiaki Tanaka[a,2], Yumei Li[a,d], Xiaoyuan Zi[a], Sadie L. Marjani[a], Ghia Euskirchen[e], Chao Ma[f,g], Robert H. LaMotte[f], In-Hyun Park[a], Michael P. Snyder[e], Christopher E. Mason[b], and Sherman M. Weissman[a,1]

Author Affiliations

Contributed by Sherman M. Weissman, October 8, 2012 (sent for review August 22, 2012)

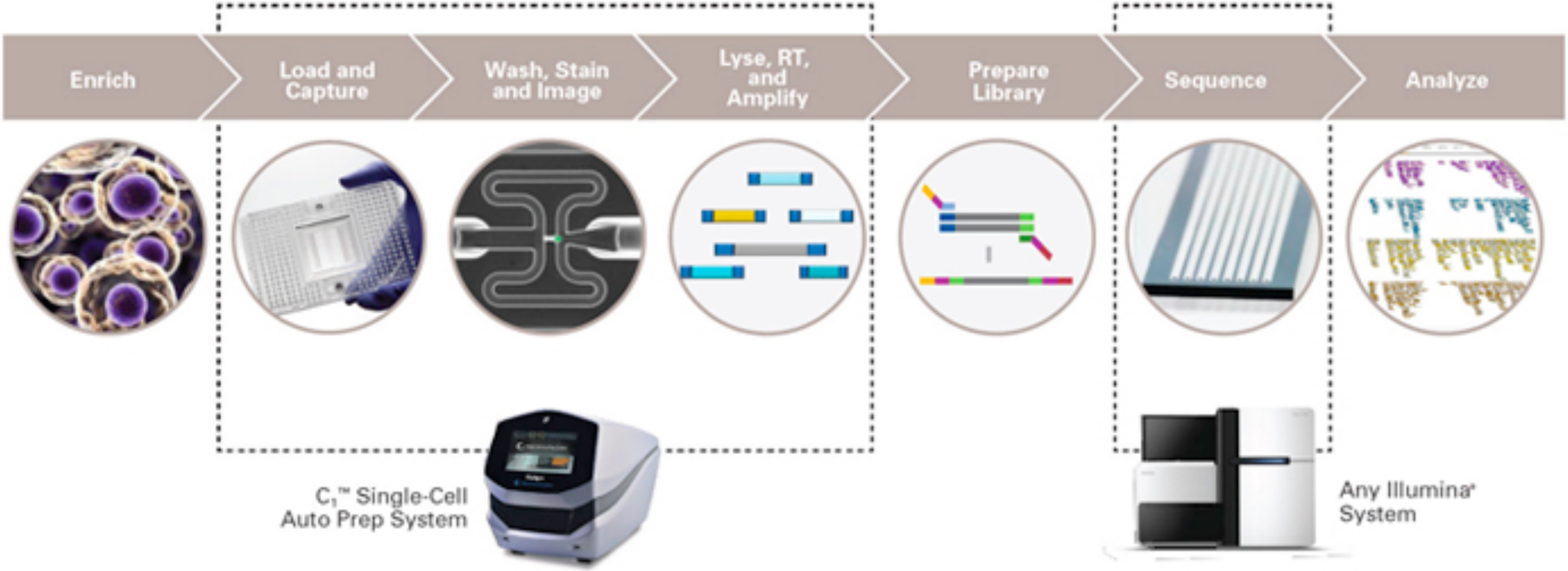
This Issue

January 8, 2013
vol. 110 no. 2
Masthead (PDF)
Table of Contents

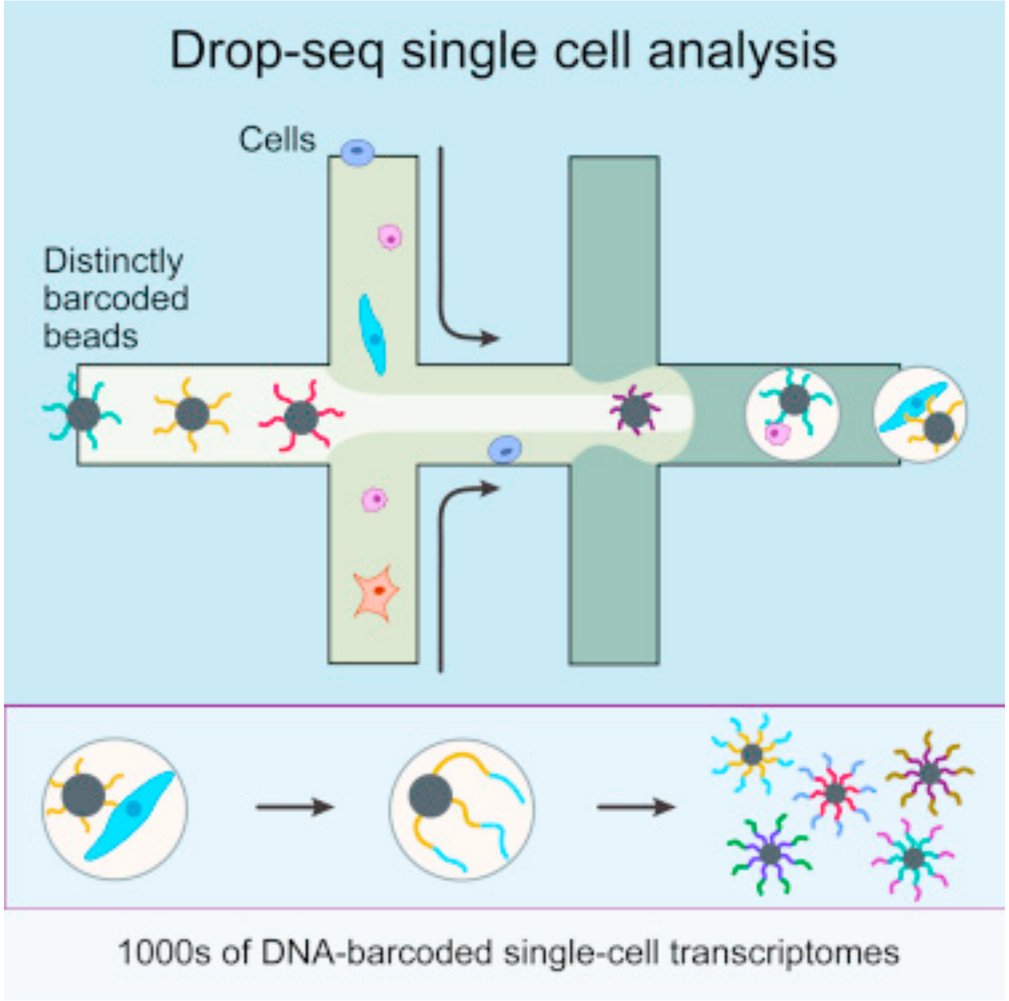
Don't Miss

Customize your free PNAS alerts to receive a notification as new content becomes available.

But now it’s very easy – Fluidigm C1



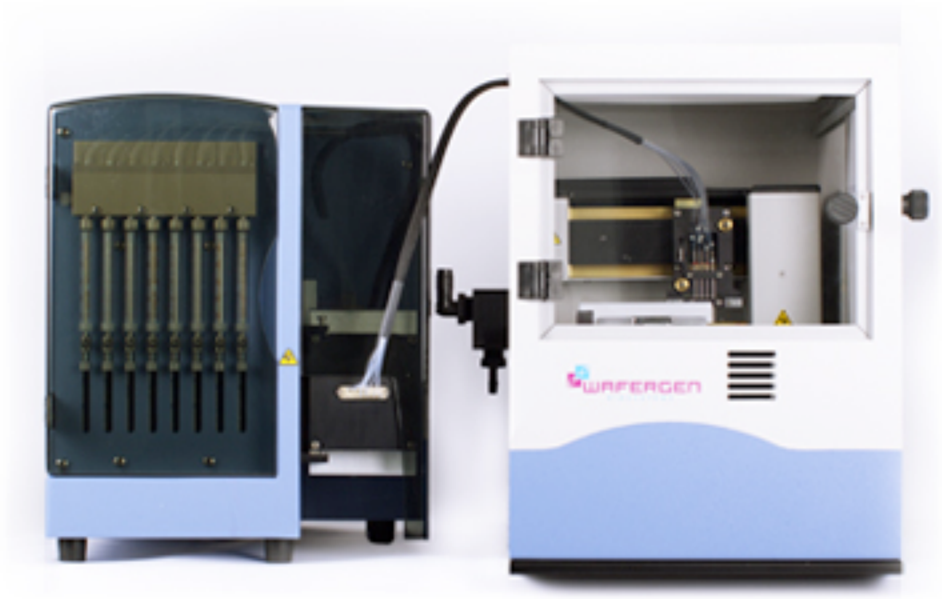
Drop-Seq


Drop-seq single cell analysis
Cells
Distinctly barcoded beads
1000s of DNA-barcoded single-cell transcriptomes

http://mccarrolllab.com/dropseq/

http://www.cell.com/abstract/S0092-8674%2815%2900549-8

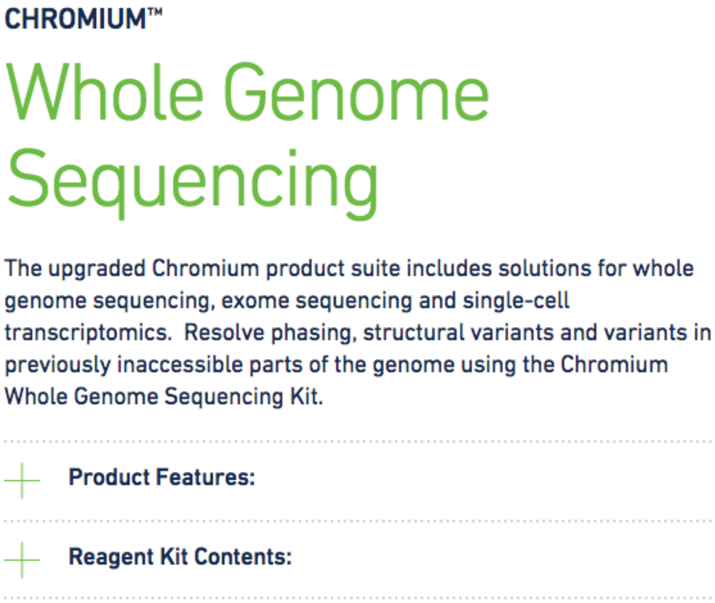
WaferGen iCell8



BioRad QX200 & ILMN system

QX200™ Droplet Digital™ PCR System

Chromium NGS



CHROMIUM™

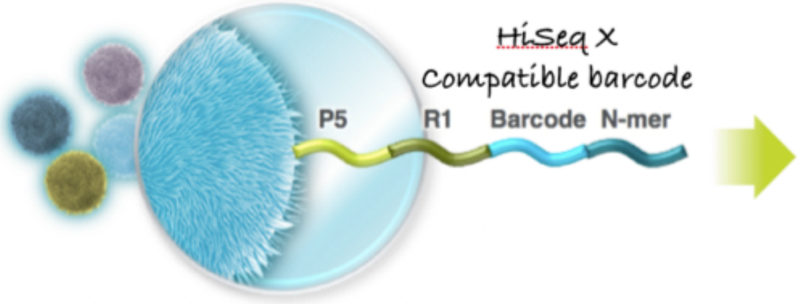
Whole Genome Sequencing

The upgraded Chromium product suite includes solutions for whole genome sequencing, exome sequencing and single-cell transcriptomics. Resolve phasing, structural variants and variants in previously inaccessible parts of the genome using the Chromium Whole Genome Sequencing Kit.

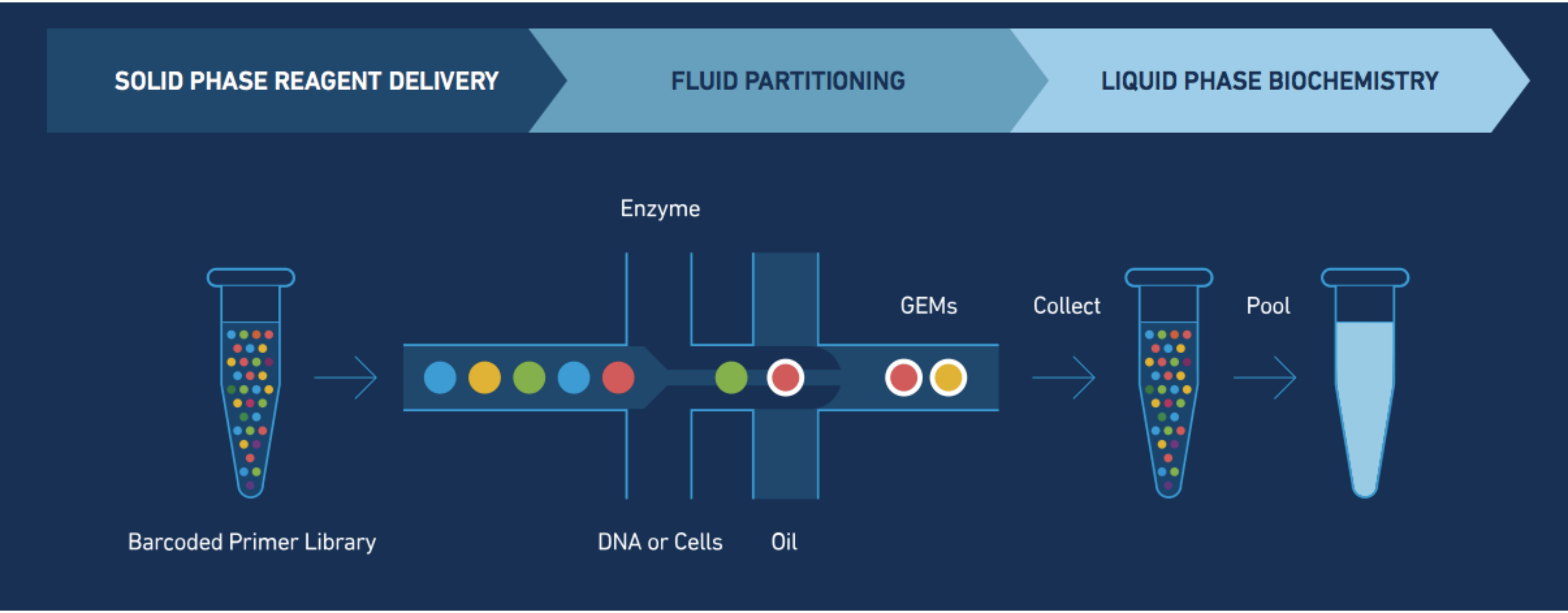
＋ **Product Features:**

＋ **Reagent Kit Contents:**

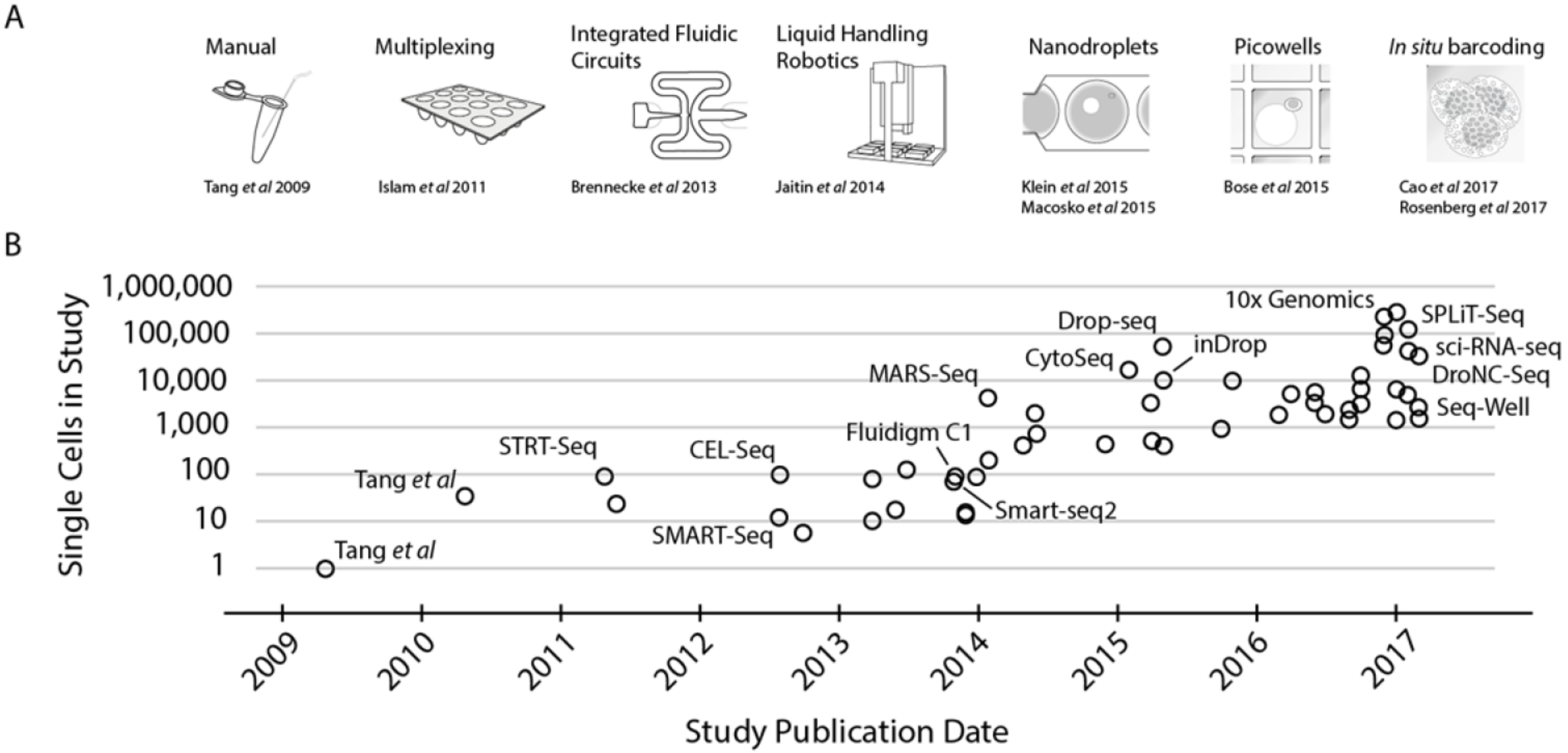
Chromium
Single Cell 3' Reagents Module 1
Chromium
Single Cell 3' Reagents Module 2
Chromium
Single Cell 3' Sample Index Plate
10x



10X Genomics Single-Cell

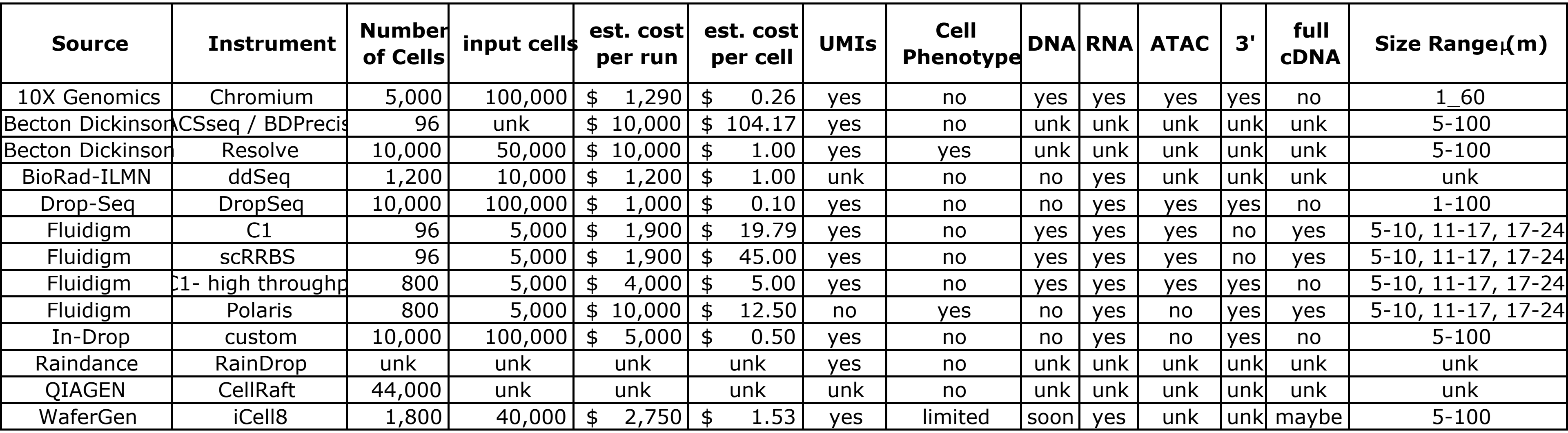


The explosion of scRNA-seq experiments



Svennson *et al*., arXiv 2017

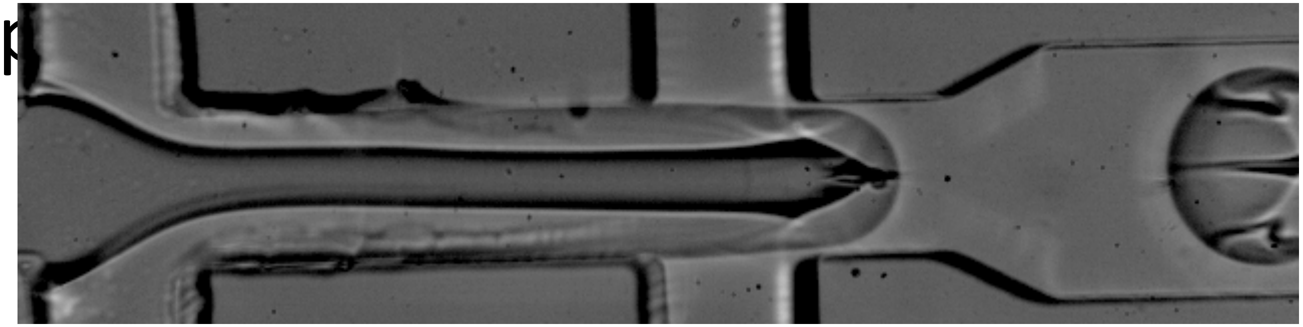
Many options today for single-cell sequencing



Source	Instrument	Number of Cells	input cells	est. cost per run	est. cost per cell	UMIs	Cell Phenotype	DNA	RNA	ATAC	3'	full cDNA	Size Rangeμ(m)
10X Genomics	Chromium	5,000	100,000	$ 1,290	$ 0.26	yes	no	yes	yes	yes	yes	no	1_60
Becton Dickinson	ACSseq / BDPrecis	96	unk	$ 10,000	$ 104.17	yes	no	unk	unk	unk	unk	unk	5-100
Becton Dickinson	Resolve	10,000	50,000	$ 10,000	$ 1.00	yes	yes	unk	unk	unk	unk	unk	5-100
BioRad-ILMN	ddSeq	1,200	10,000	$ 1,200	$ 1.00	unk	no	no	yes	unk	unk	unk	unk
Drop-Seq	DropSeq	10,000	100,000	$ 1,000	$ 0.10	yes	no	no	yes	yes	yes	no	1-100
Fluidigm	C1	96	5,000	$ 1,900	$ 19.79	yes	no	yes	yes	yes	no	yes	5-10, 11-17, 17-24
Fluidigm	scRRBS	96	5,000	$ 1,900	$ 45.00	yes	no	yes	yes	yes	no	yes	5-10, 11-17, 17-24
Fluidigm	C1- high throughp	800	5,000	$ 4,000	$ 5.00	yes	no	yes	yes	yes	yes	no	5-10, 11-17, 17-24
Fluidigm	Polaris	800	5,000	$ 10,000	$ 12.50	no	yes	no	yes	no	yes	yes	5-10, 11-17, 17-24
In-Drop	custom	10,000	100,000	$ 5,000	$ 0.50	yes	no	no	yes	no	yes	no	5-100
Raindance	RainDrop	unk	unk	unk	unk	yes	no	unk	unk	unk	unk	unk	unk
QIAGEN	CellRaft	44,000	unk	unk	unk	unk	no	unk	unk	unk	unk	unk	unk
WaferGen	iCell8	1,800	40,000	$ 2,750	$ 1.53	yes	limited	soon	yes	unk	unk	maybe	5-100

Single cell capture and RNA chemistry using nanodroplets

- Drop

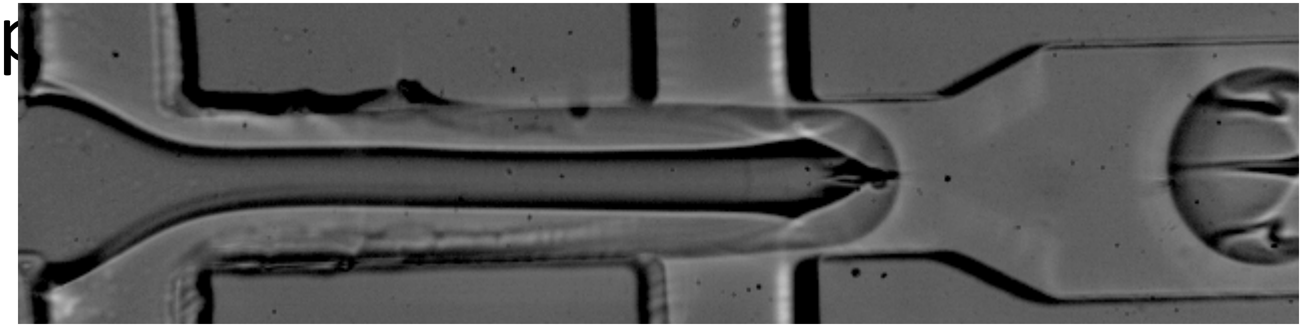


http://mccarrolllab.com/dropseq/
Macosko *et al.*, Cell 2015

Single cell capture and RNA chemistry using nanodroplets

- Drop

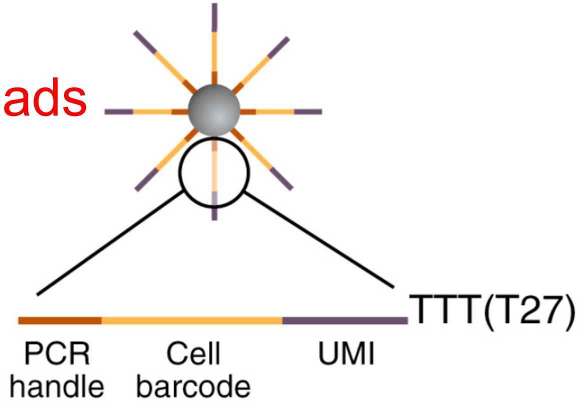
Beads



Cells + Enzymes

Oil

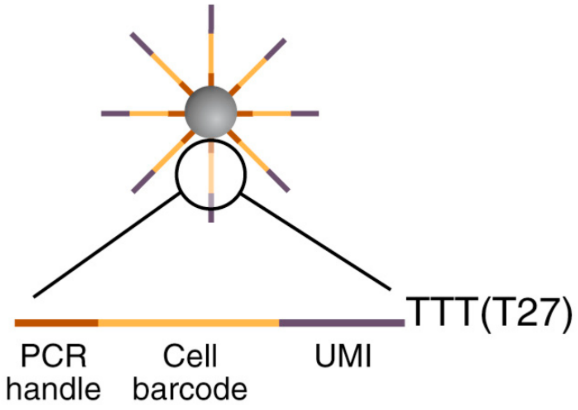
Barcoded beads

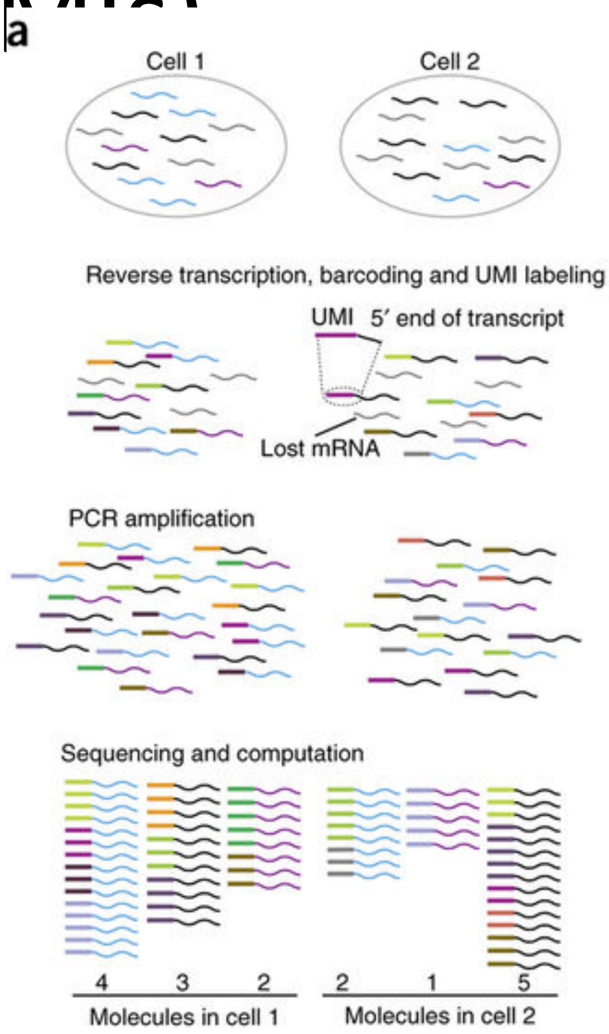


http://mccarrolllab.com/dropseq/
Macosko *et al.*, Cell 2015

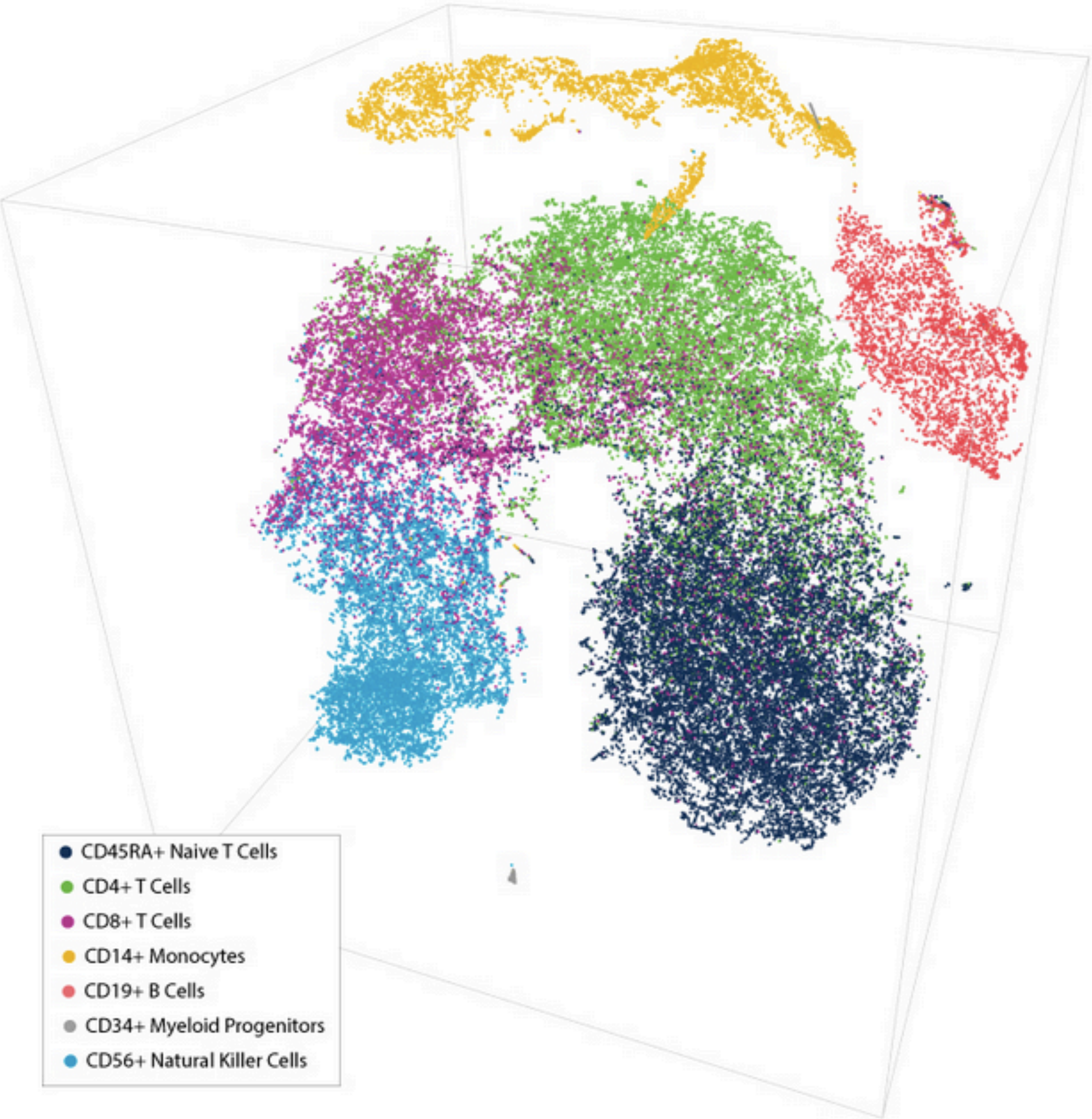
Unique Molecular Identifiers (UMIs)

Barcoded beads

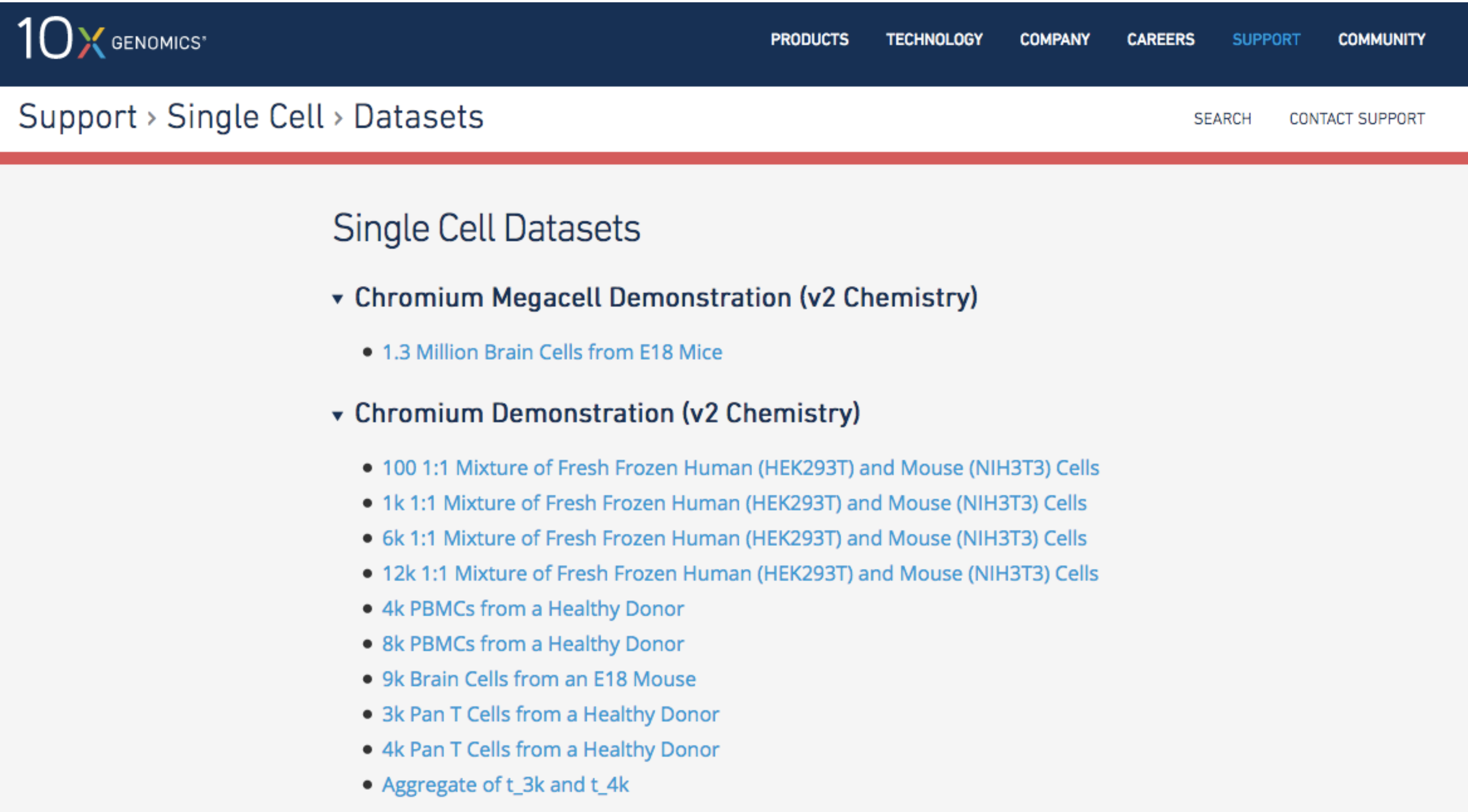




Islam *et al.*, Nature Methods 2014


● CD45RA+ Naive T Cells
● CD4+ T Cells
● CD8+ T Cells
● CD14+ Monocytes
● CD19+ B Cells
● CD34+ Myeloid Progenitors
● CD56+ Natural Killer Cells

1.3 million neurons catalogued



Support › Single Cell › Datasets

Single Cell Datasets

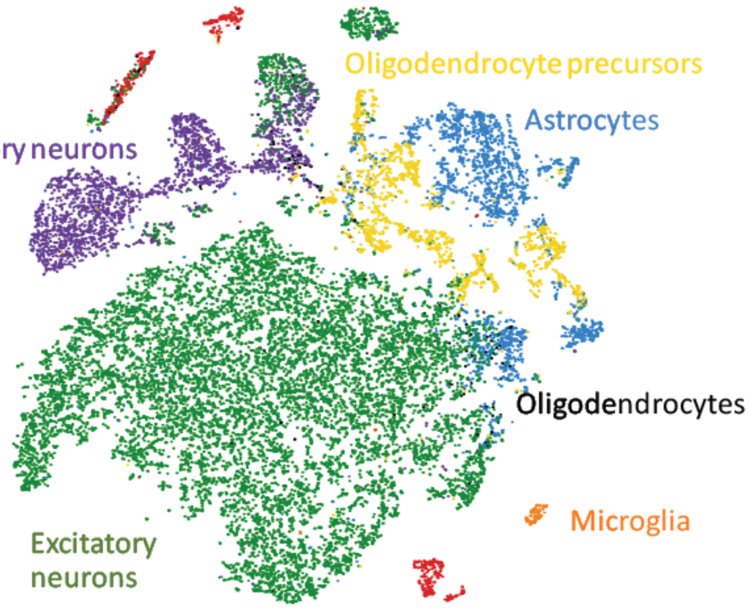
▾ Chromium Megacell Demonstration (v2 Chemistry)

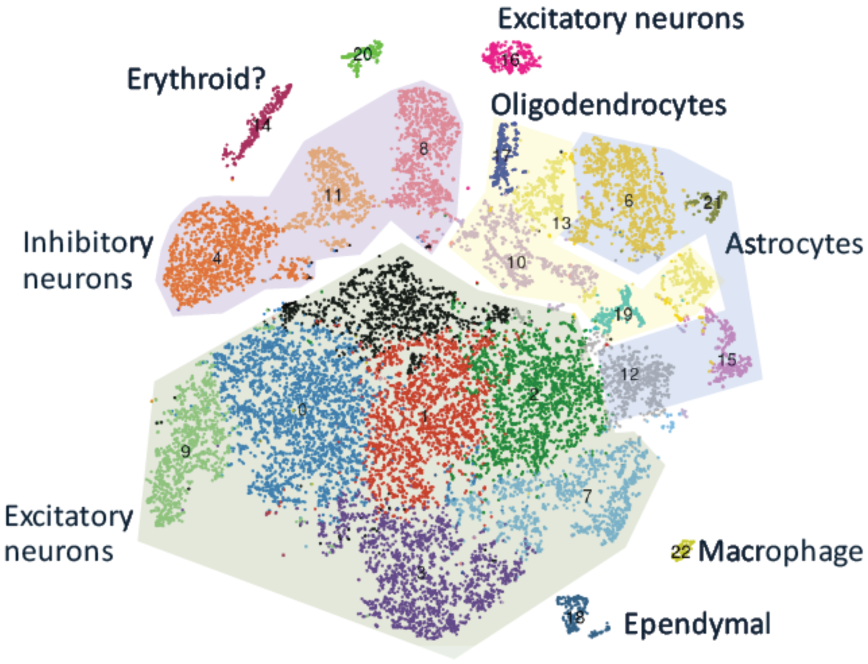
- 1.3 Million Brain Cells from E18 Mice

▾ Chromium Demonstration (v2 Chemistry)

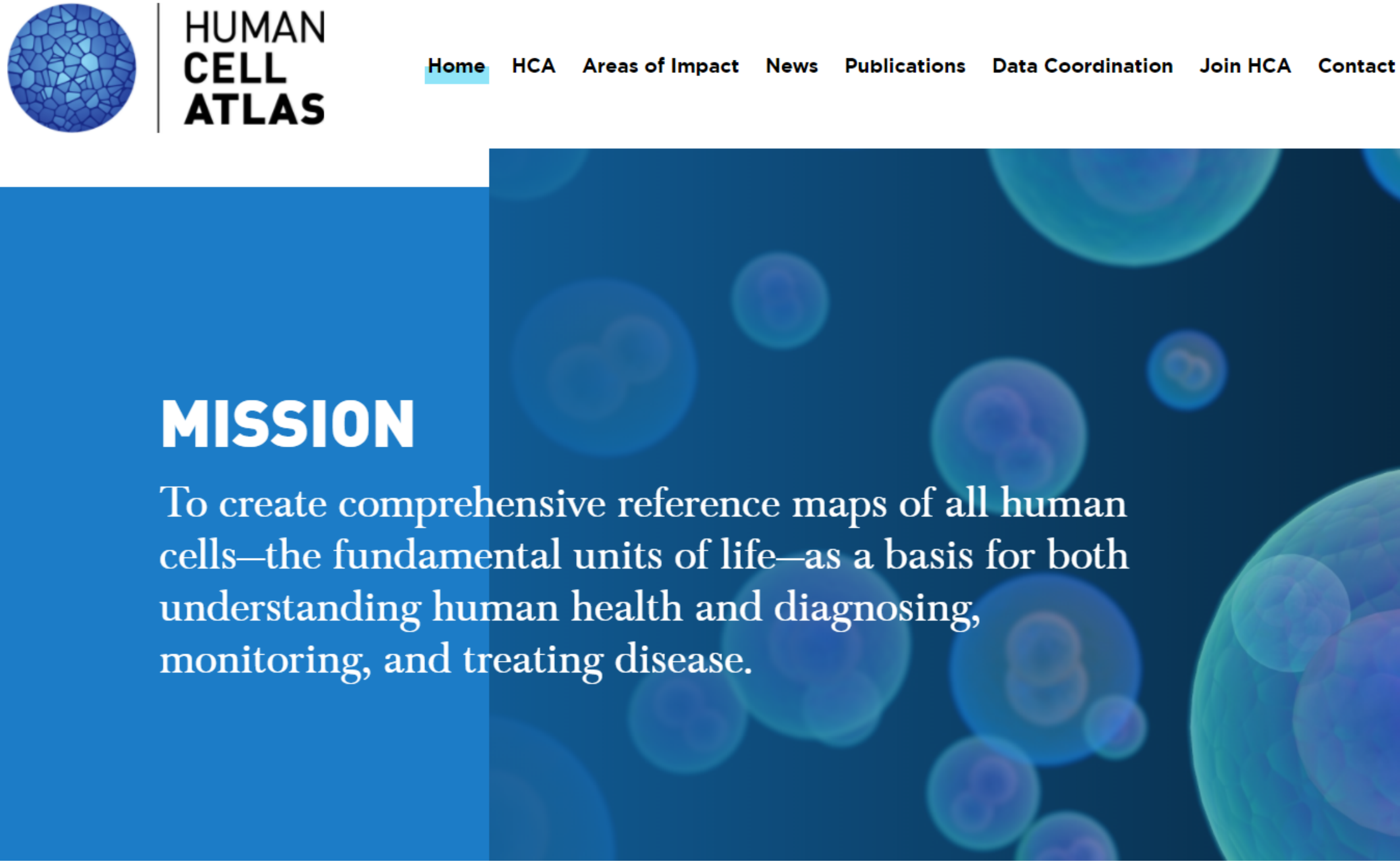
- 100 1:1 Mixture of Fresh Frozen Human (HEK293T) and Mouse (NIH3T3) Cells
- 1k 1:1 Mixture of Fresh Frozen Human (HEK293T) and Mouse (NIH3T3) Cells
- 6k 1:1 Mixture of Fresh Frozen Human (HEK293T) and Mouse (NIH3T3) Cells
- 12k 1:1 Mixture of Fresh Frozen Human (HEK293T) and Mouse (NIH3T3) Cells
- 4k PBMCs from a Healthy Donor
- 8k PBMCs from a Healthy Donor
- 9k Brain Cells from an E18 Mouse
- 3k Pan T Cells from a Healthy Donor
- 4k Pan T Cells from a Healthy Donor
- Aggregate of t_3k and t_4k

1.3 million mouse embryonic brain cells, 10X Chromium





10x Genomics | LIT000015 Chromium™ Million Brain Cells Application Note



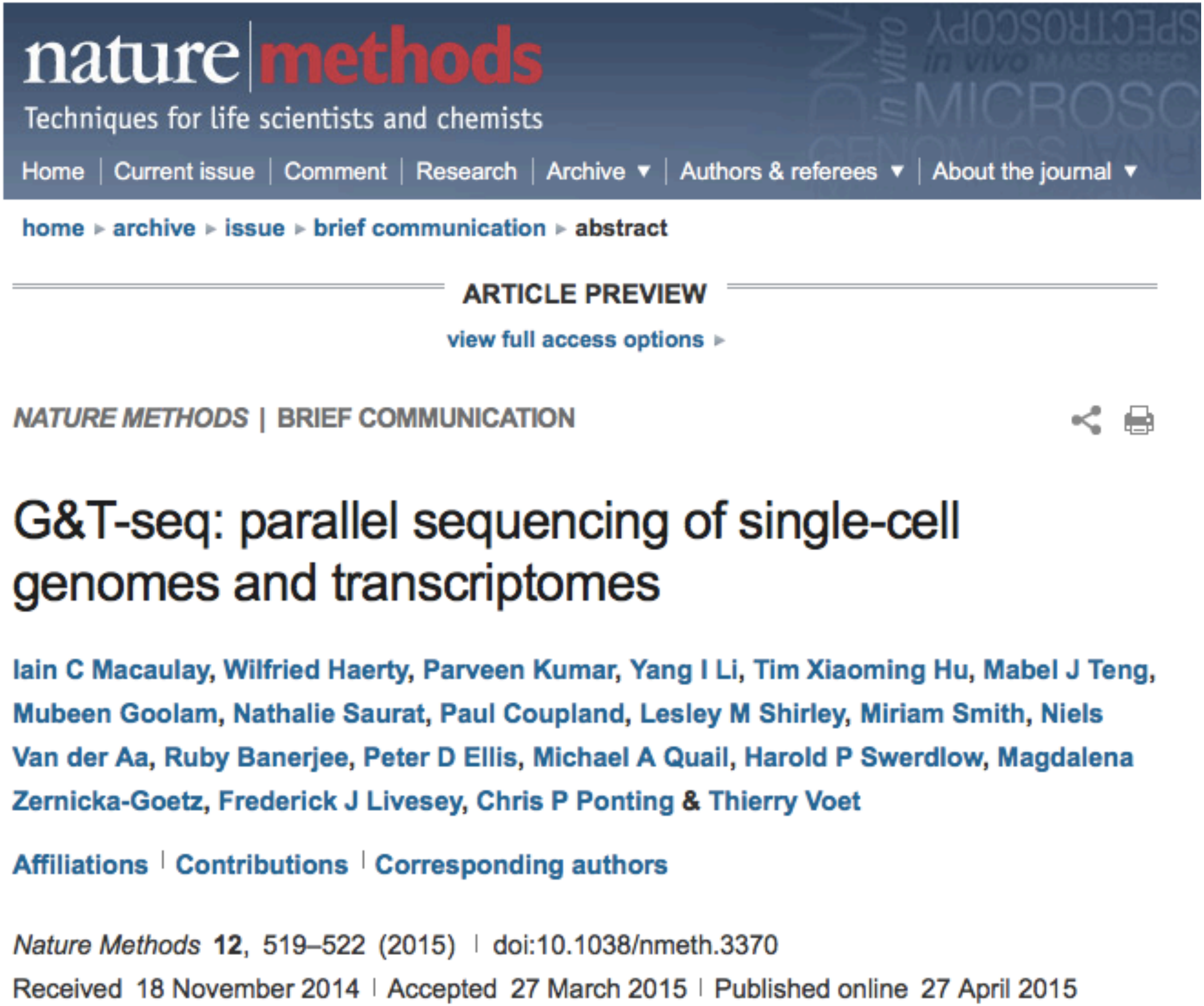
MISSION

To create comprehensive reference maps of all human cells—the fundamental units of life—as a basis for both understanding human health and diagnosing, monitoring, and treating disease.

www.humancellatlas.org

Beyond single cell RNA-seq

Single nuclei sequencing	scNuc-seq
Epigenomics	scBS-seq, scRRBS-seq, scCHIP-seq, scATAC-seq, scDNase-seq
Genomics	Whole genome, exome
Multiple simultaneous measurements	
RNA + DNA	DR-seq, G&T-seq
RNA + methylation	scM&T-seq, scMT-seq
RNA + DNA + methylation	scTrio-seq
RNA + protein	index sorting, CITE-seq
RNA + genome editing	Perturb-seq, CRISP-seq, CROP-seq



nature methods
Techniques for life scientists and chemists

Home | Current issue | Comment | Research | Archive ▼ | Authors & referees ▼ | About the journal ▼

home ▸ archive ▸ issue ▸ brief communication ▸ abstract

ARTICLE PREVIEW
view full access options ▸

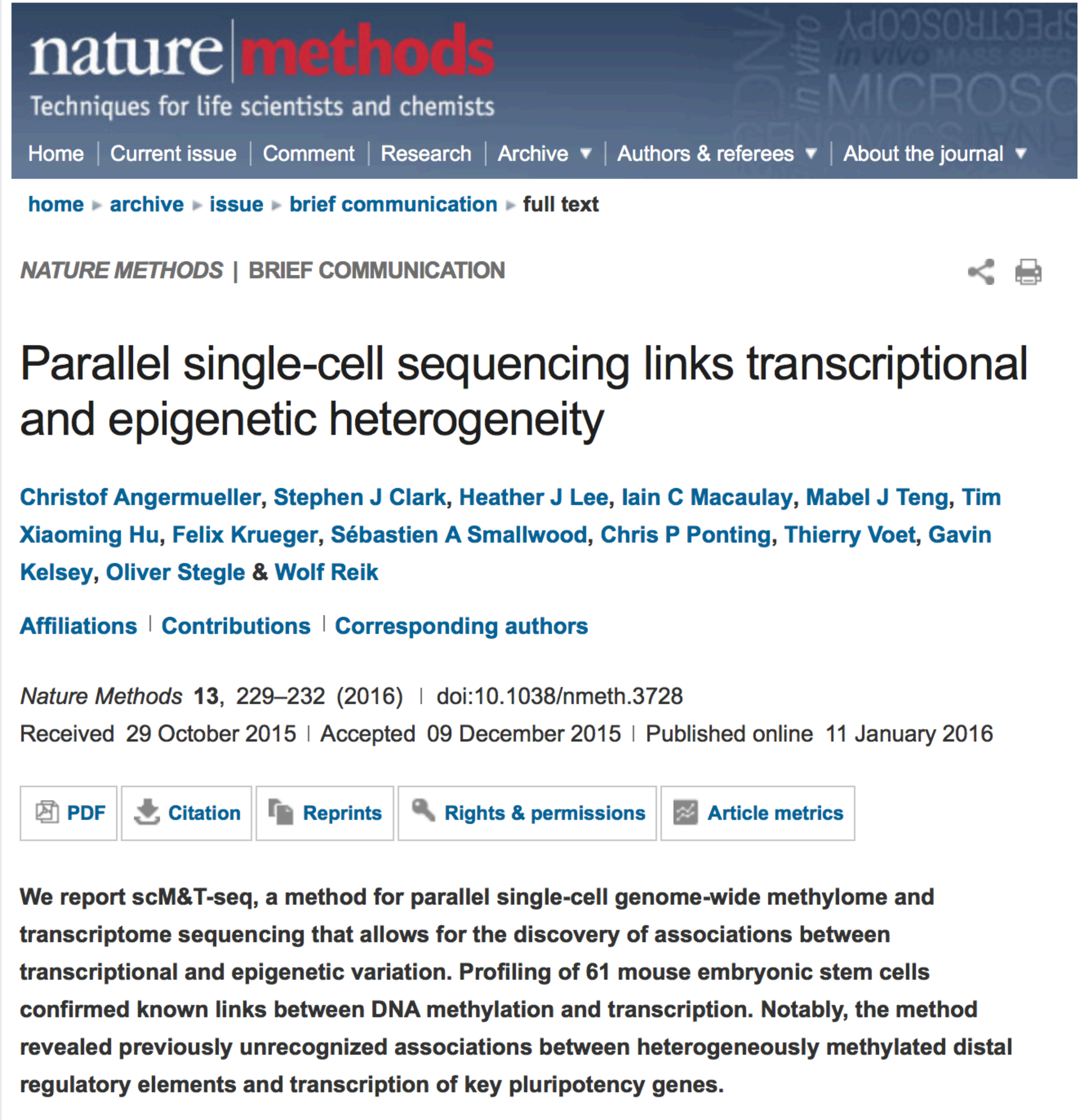
NATURE METHODS | BRIEF COMMUNICATION

G&T-seq: parallel sequencing of single-cell genomes and transcriptomes

Iain C Macaulay, Wilfried Haerty, Parveen Kumar, Yang I Li, Tim Xiaoming Hu, Mabel J Teng, Mubeen Goolam, Nathalie Saurat, Paul Coupland, Lesley M Shirley, Miriam Smith, Niels Van der Aa, Ruby Banerjee, Peter D Ellis, Michael A Quail, Harold P Swerdlow, Magdalena Zernicka-Goetz, Frederick J Livesey, Chris P Ponting & Thierry Voet

Affiliations | Contributions | Corresponding authors

Nature Methods **12**, 519–522 (2015) | doi:10.1038/nmeth.3370
Received 18 November 2014 | Accepted 27 March 2015 | Published online 27 April 2015


nature methods
Techniques for life scientists and chemists
Home
Current issue
Comment
Research
Archive ▼
Authors & referees ▼
About the journal ▼
home ▸ archive ▸ issue ▸ brief communication ▸ full text

NATURE METHODS | BRIEF COMMUNICATION

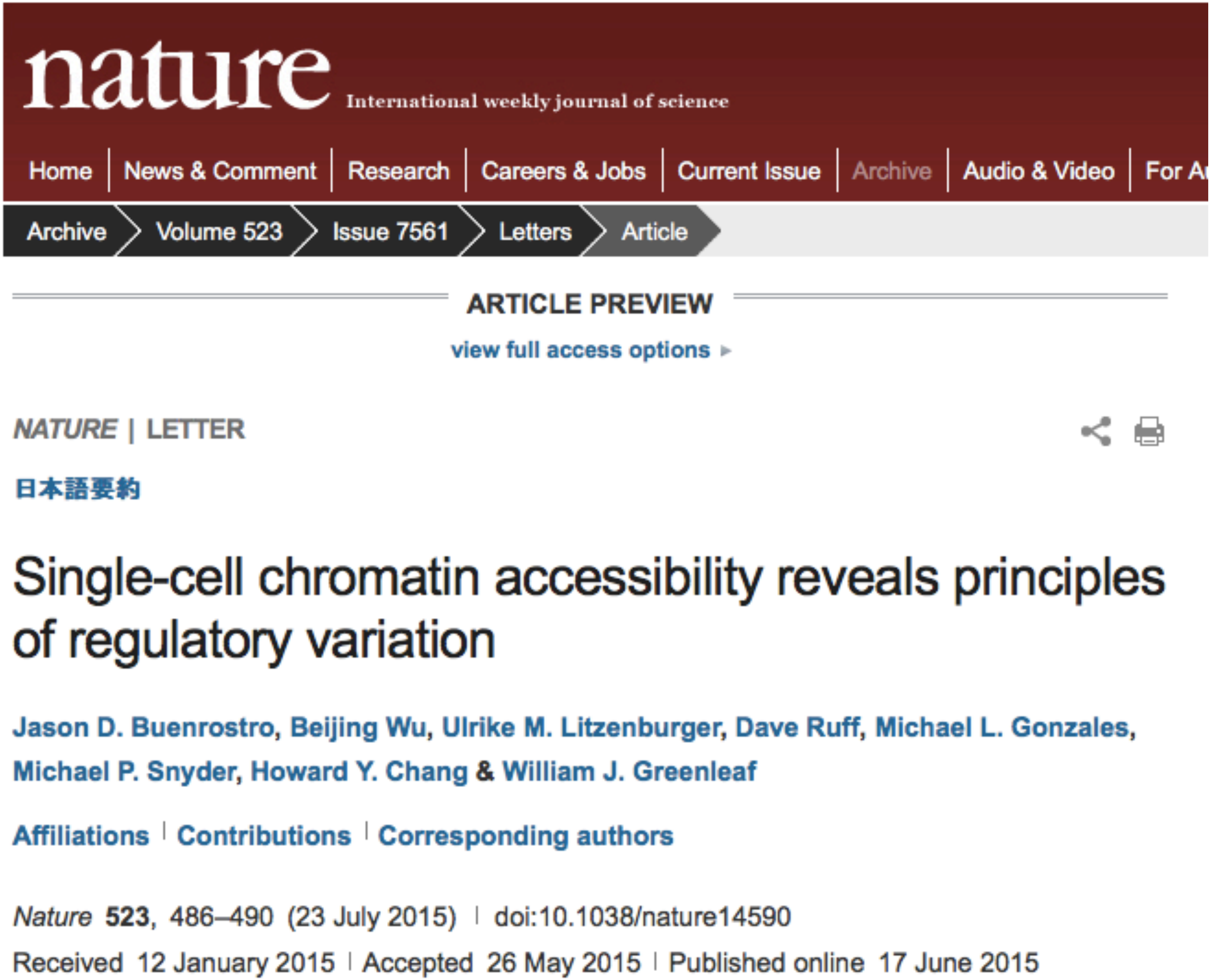
Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity

Christof Angermueller, Stephen J Clark, Heather J Lee, Iain C Macaulay, Mabel J Teng, Tim Xiaoming Hu, Felix Krueger, Sébastien A Smallwood, Chris P Ponting, Thierry Voet, Gavin Kelsey, Oliver Stegle & Wolf Reik

Affiliations | Contributions | Corresponding authors

Nature Methods **13**, 229–232 (2016) | doi:10.1038/nmeth.3728
Received 29 October 2015 | Accepted 09 December 2015 | Published online 11 January 2016

We report scM&T-seq, a method for parallel single-cell genome-wide methylome and transcriptome sequencing that allows for the discovery of associations between transcriptional and epigenetic variation. Profiling of 61 mouse embryonic stem cells confirmed known links between DNA methylation and transcription. Notably, the method revealed previously unrecognized associations between heterogeneously methylated distal regulatory elements and transcription of key pluripotency genes.



NATURE | LETTER

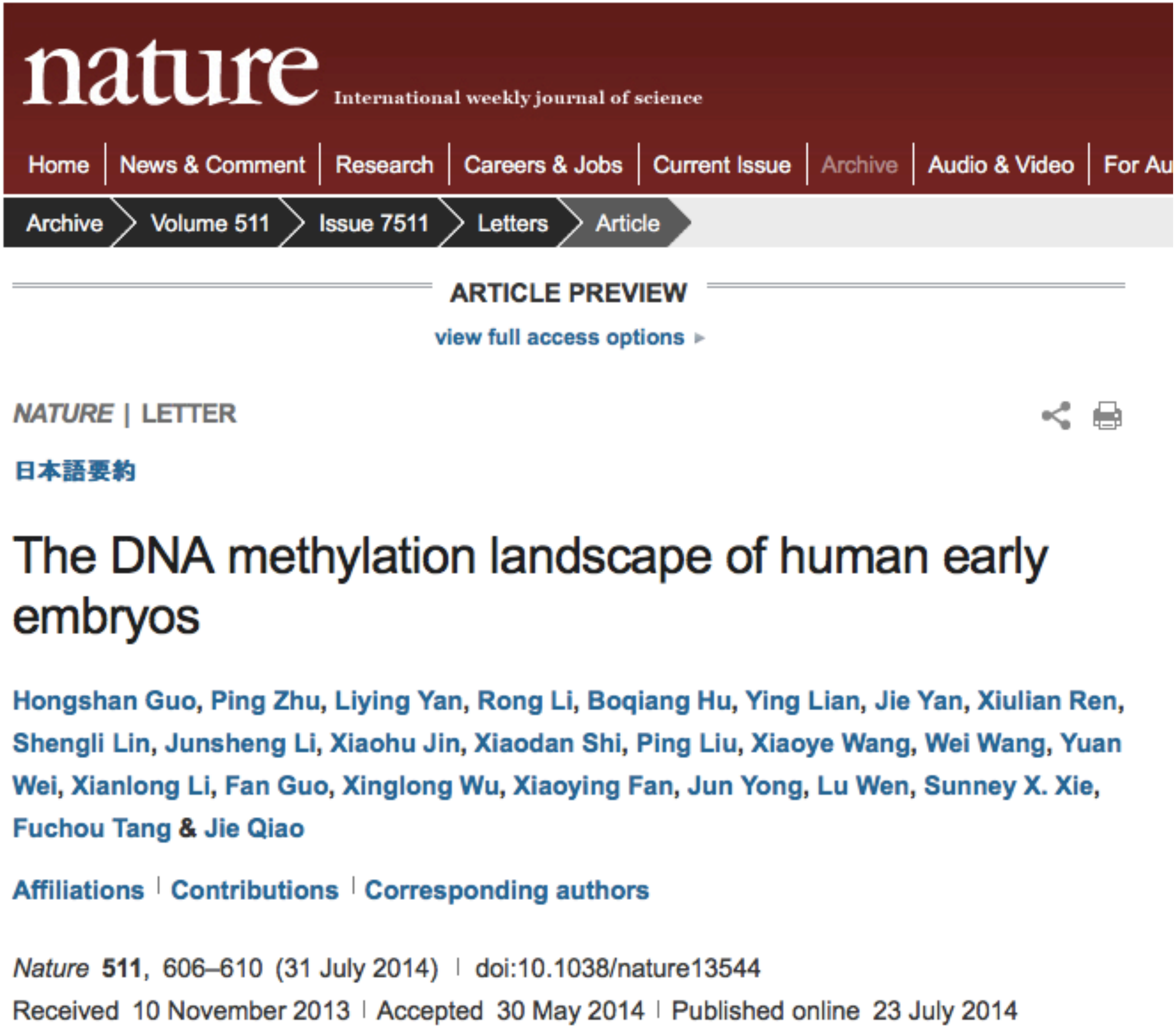
日本語要約

Single-cell chromatin accessibility reveals principles of regulatory variation

Jason D. Buenrostro, Beijing Wu, Ulrike M. Litzenburger, Dave Ruff, Michael L. Gonzales, Michael P. Snyder, Howard Y. Chang & William J. Greenleaf

Affiliations | Contributions | Corresponding authors

Nature **523**, 486–490 (23 July 2015) | doi:10.1038/nature14590
Received 12 January 2015 | Accepted 26 May 2015 | Published online 17 June 2015



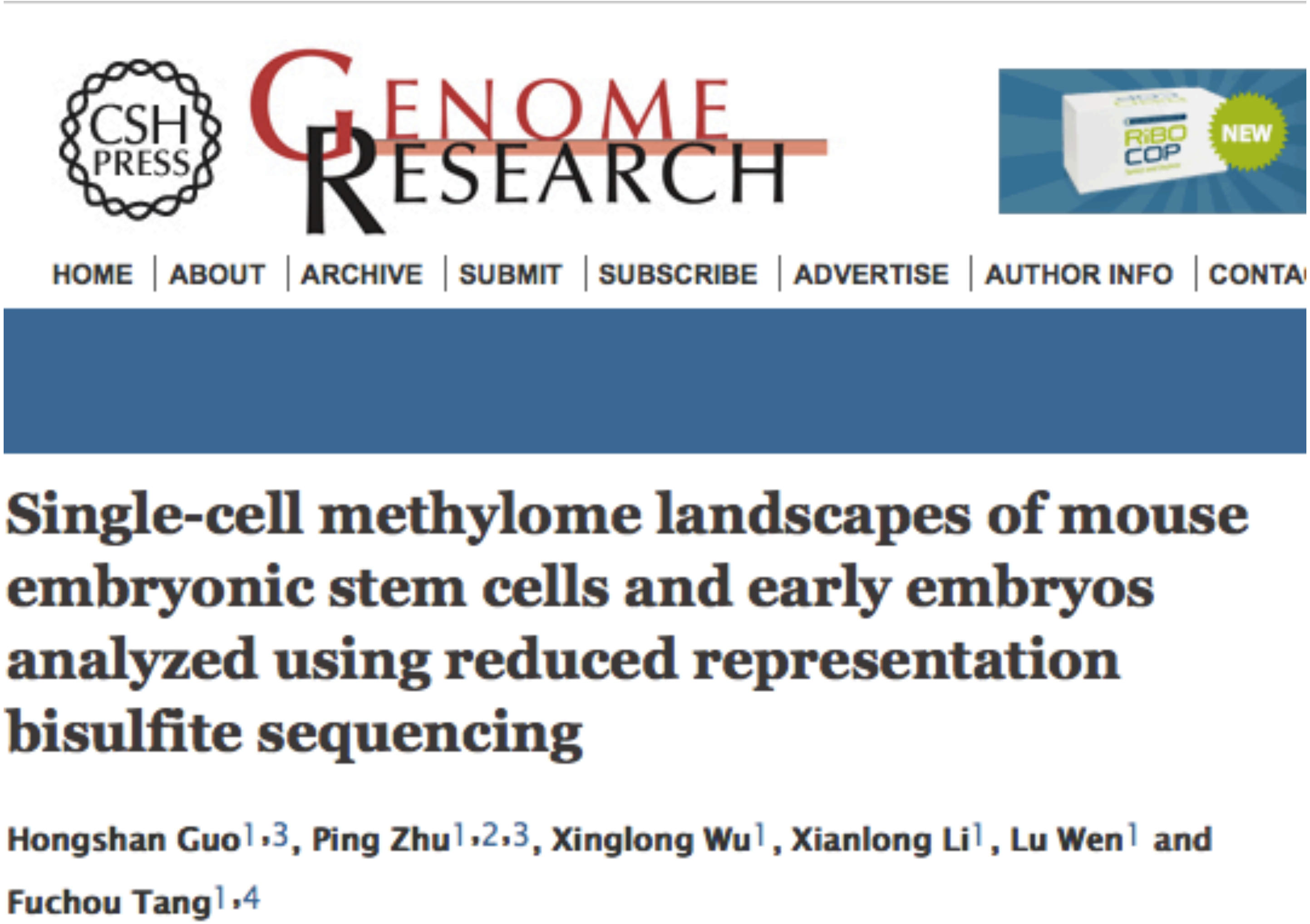
NATURE | LETTER

日本語要約

The DNA methylation landscape of human early embryos

Hongshan Guo, Ping Zhu, Liying Yan, Rong Li, Boqiang Hu, Ying Lian, Jie Yan, Xiulian Ren, Shengli Lin, Junsheng Li, Xiaohu Jin, Xiaodan Shi, Ping Liu, Xiaoye Wang, Wei Wang, Yuan Wei, Xianlong Li, Fan Guo, Xinglong Wu, Xiaoying Fan, Jun Yong, Lu Wen, Sunney X. Xie, Fuchou Tang & Jie Qiao

Nature **511**, 606–610 (31 July 2014) | doi:10.1038/nature13544
Received 10 November 2013 | Accepted 30 May 2014 | Published online 23 July 2014



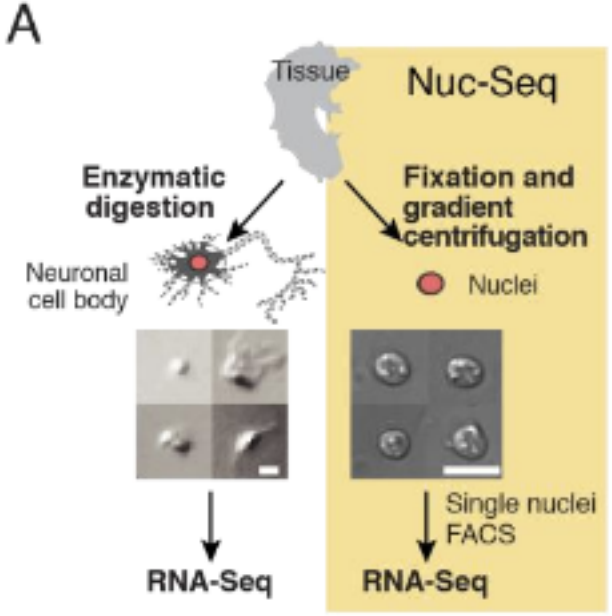
CSH PRESS

GENOME RESEARCH

Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing

Hongshan Guo[1,3], Ping Zhu[1,2,3], Xinglong Wu[1], Xianlong Li[1], Lu Wen[1] and Fuchou Tang[1,4]

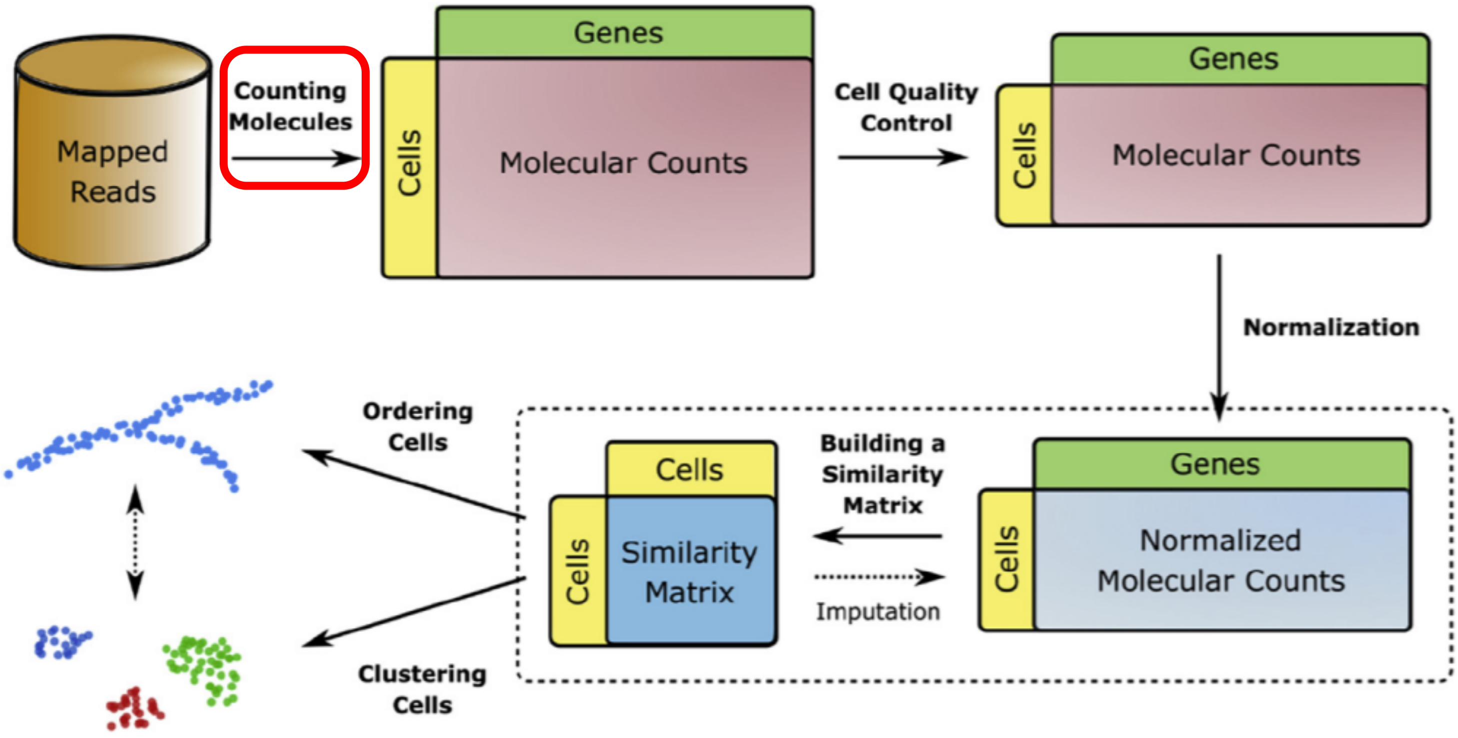
Other methods also emerging



Div-Seq: A single nucleus RNA-Seq method reveals dynamics of rare adult newborn neurons in the CNS

http://biorxiv.org/content/biorxiv/early/2016/03/27/045989.full.pdf

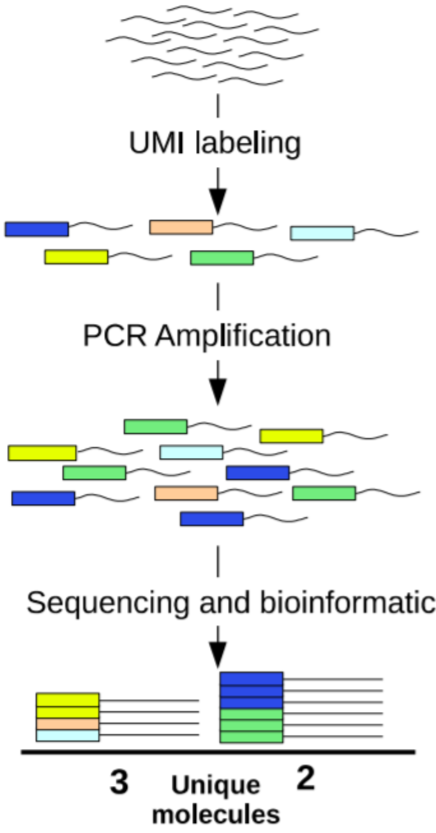
Analysis:
Structure of a generic pipeline



Camara, Current Opinions in Systems Biology, 2018

Counting Molecules

- Counting reads
  - featureCounts, etc.
- Counting UMIs
  - Unique
    - does not account for PCR and sequencing errors
  - Directional adjacency graph (UMI-tools)
  - Bayesian (dropEst)
  - Proprietary (SevenBridges for BD Precise)

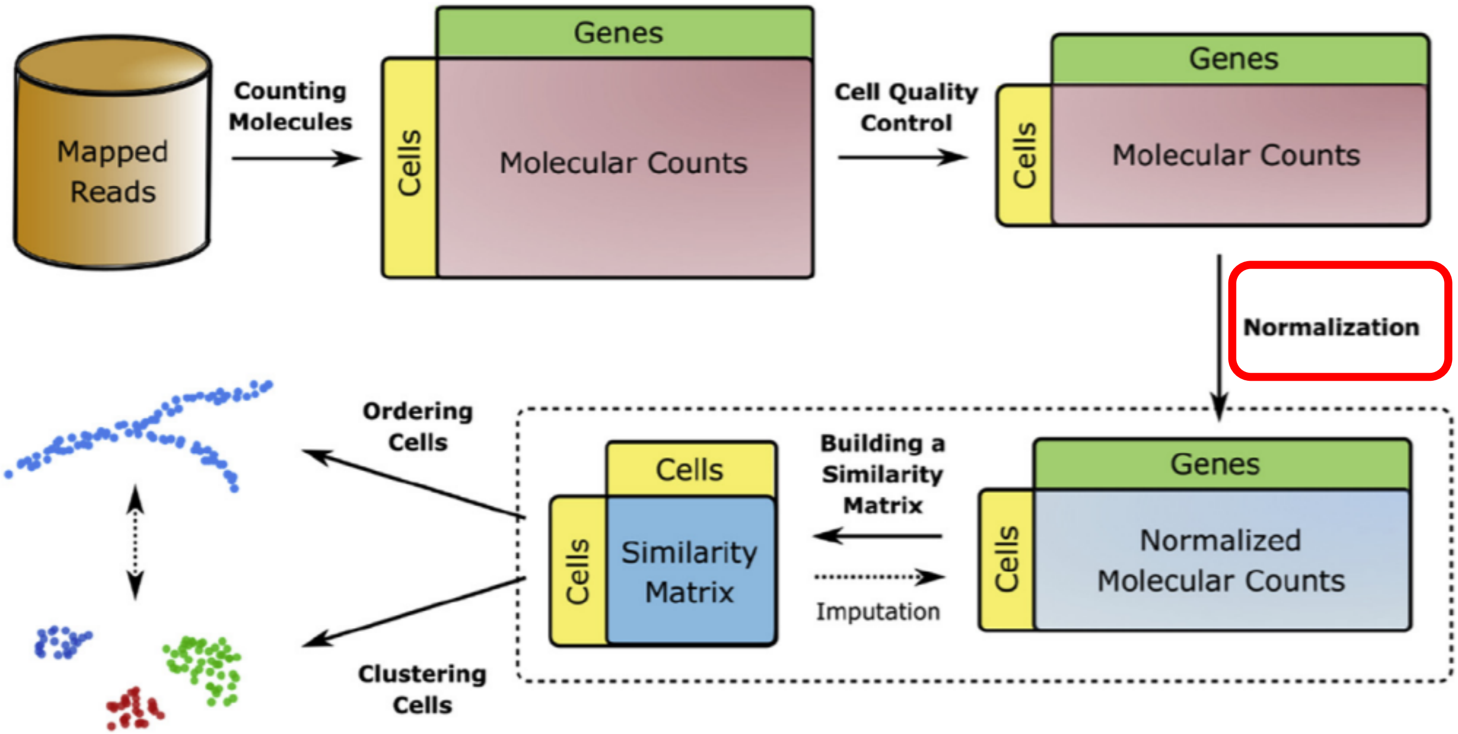


Smith *et al.*, Genome Research 2017

Commonly used open-source tools

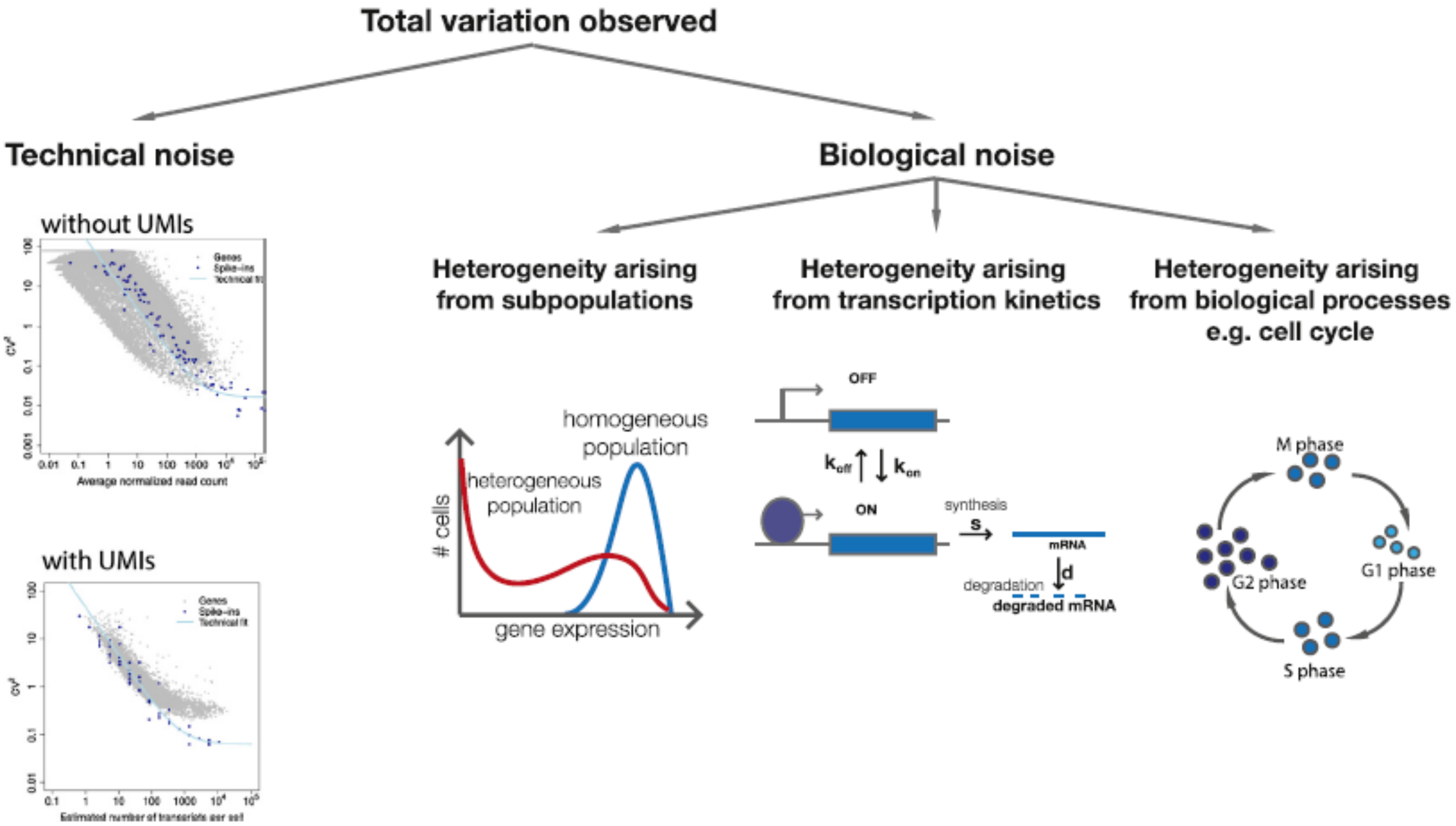
1. Infer which barcodes come from valid cells – **UMI-tools**
2. Extract cell barcodes and UMIs from R1 and add to R2 – **UMI-tools**
3. Align to reference genome (GRCh38) – **STAR**
4. Assign reads to genes (Ensembl) – **featureCounts**
5. Count unique UMIs per gene – **UMI-tools**
6. QC – **fastqc, picard, multiqc, custom scripts**

Structure of a generic pipeline



Camara, Current Opinions in Systems Biology, 2018

Normalization challenges



Total variation observed

Technical noise

Biological noise

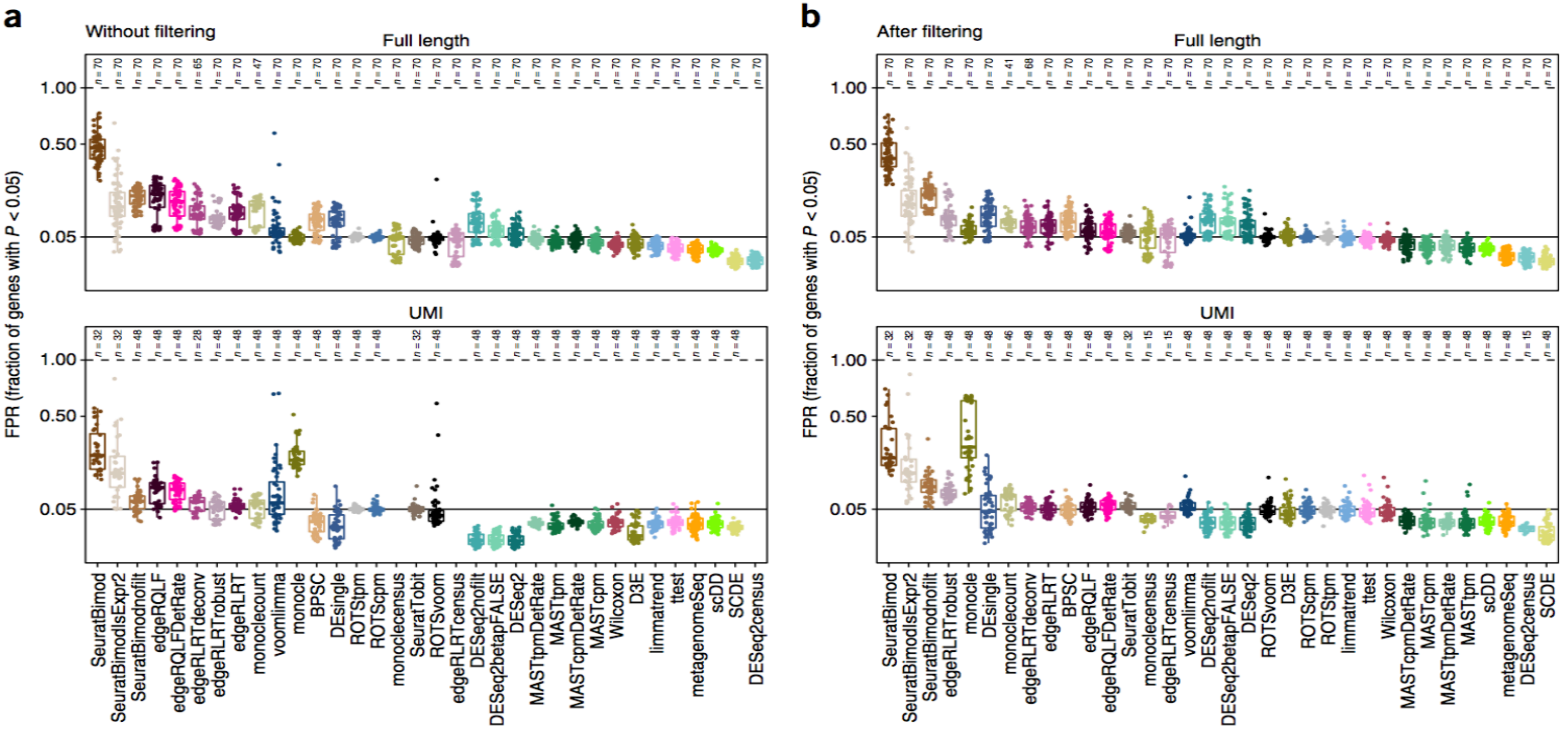
Heterogeneity arising from subpopulations

Heterogeneity arising from transcription kinetics

Heterogeneity arising from biological processes e.g. cell cycle

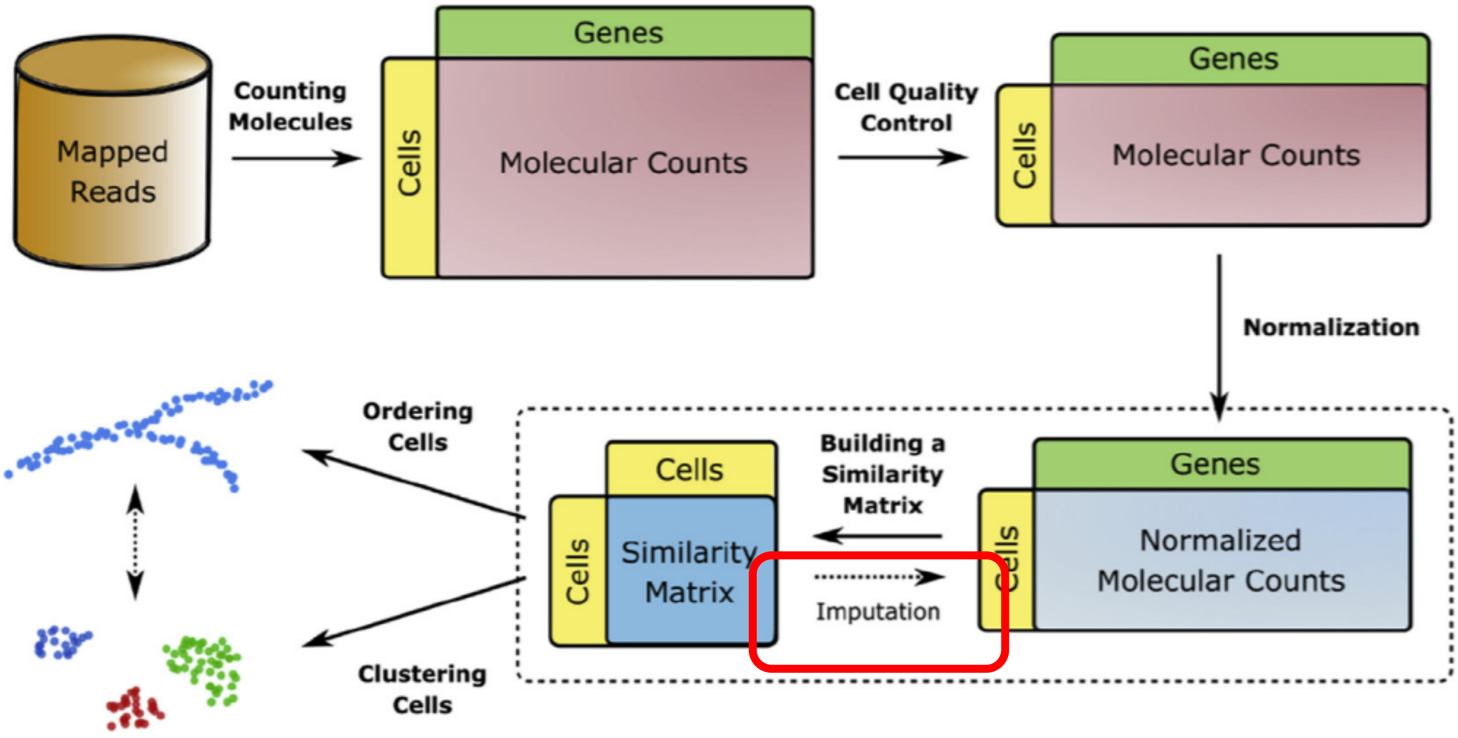
Kolodziejczyk *et al*., Briefings in Functional Genomics 2017

Normalization + Differential Expression Analysis



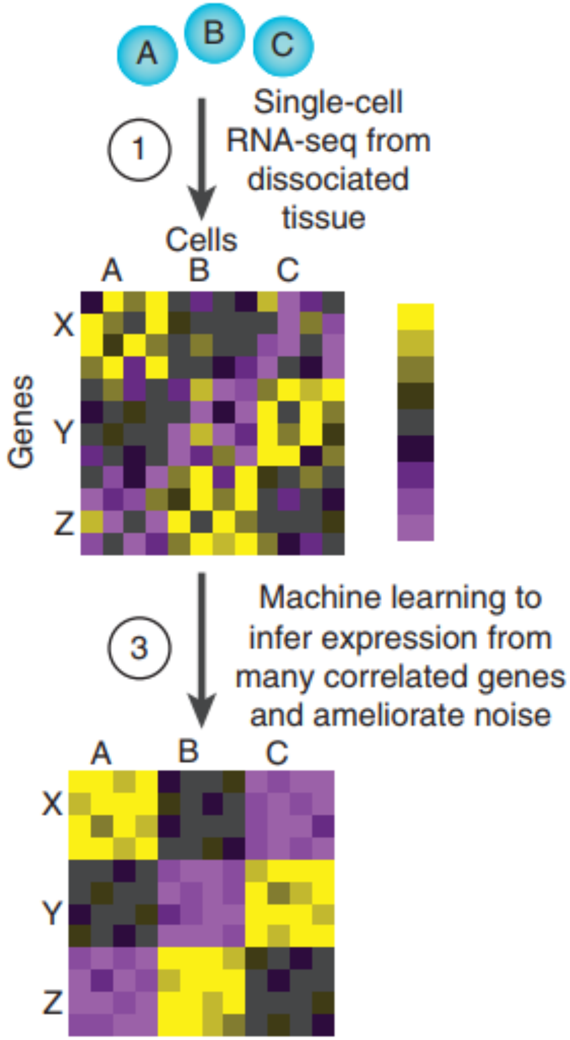
Soneson and Robinson, Nature Methods 2018

Structure of a generic pipeline



Camara, Current Opinions in Systems Biology, 2018

Gene Expression Imputation


A
B
C
1
Single-cell RNA-seq from dissociated tissue
Cells
A
B
C
Genes
X
Y
Z
3
Machine learning to infer expression from many correlated genes and ameliorate noise
A
B
C
X
Y
Z

Satija *et al.*, Nature Biotechnology 2015

Gene Expression Imputation

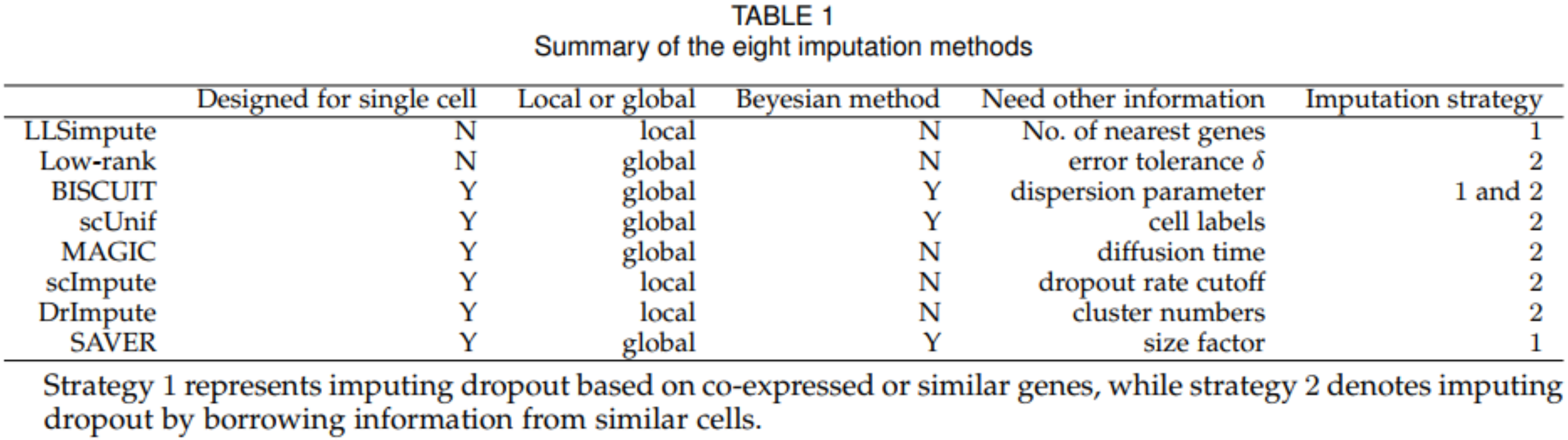


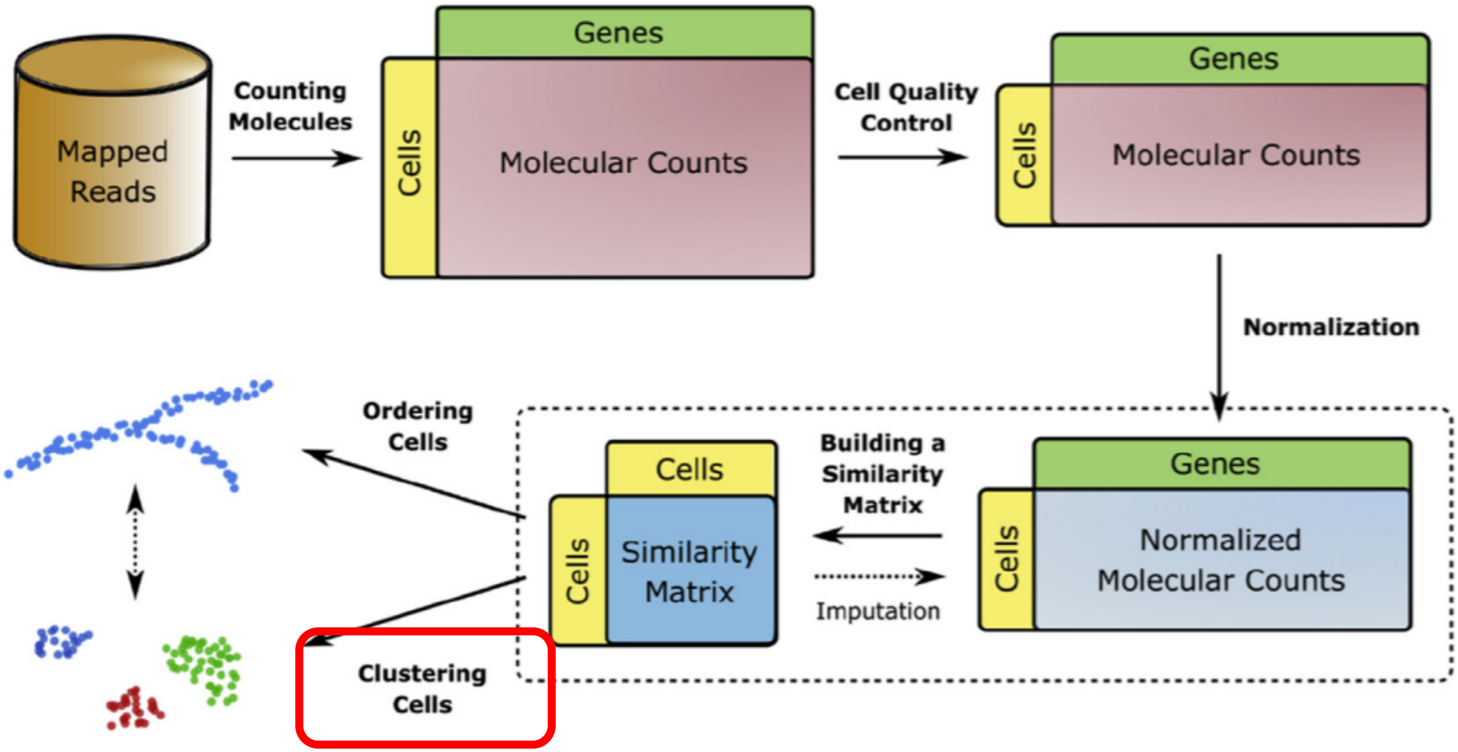
TABLE 1
Summary of the eight imputation methods

	Designed for single cell	Local or global	Beyesian method	Need other information	Imputation strategy
LLSimpute	N	local	N	No. of nearest genes	1
Low-rank	N	global	N	error tolerance δ	2
BISCUIT	Y	global	Y	dispersion parameter	1 and 2
scUnif	Y	global	Y	cell labels	2
MAGIC	Y	global	N	diffusion time	2
scImpute	Y	local	N	dropout rate cutoff	2
DrImpute	Y	local	N	cluster numbers	2
SAVER	Y	global	Y	size factor	1

Strategy 1 represents imputing dropout based on co-expressed or similar genes, while strategy 2 denotes imputing dropout by borrowing information from similar cells.

Zhang and Zhang, Biorxiv 2017

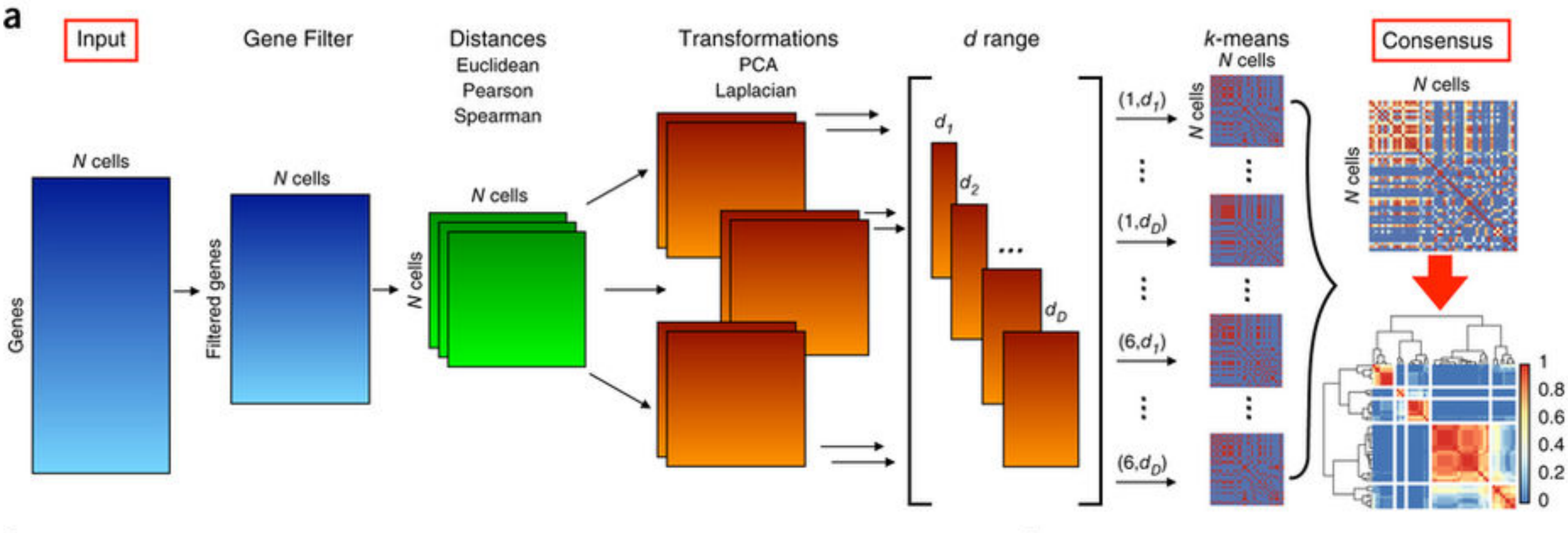
Structure of a generic pipeline



Camara, Current Opinions in Systems Biology, 2018

Clustering Cells

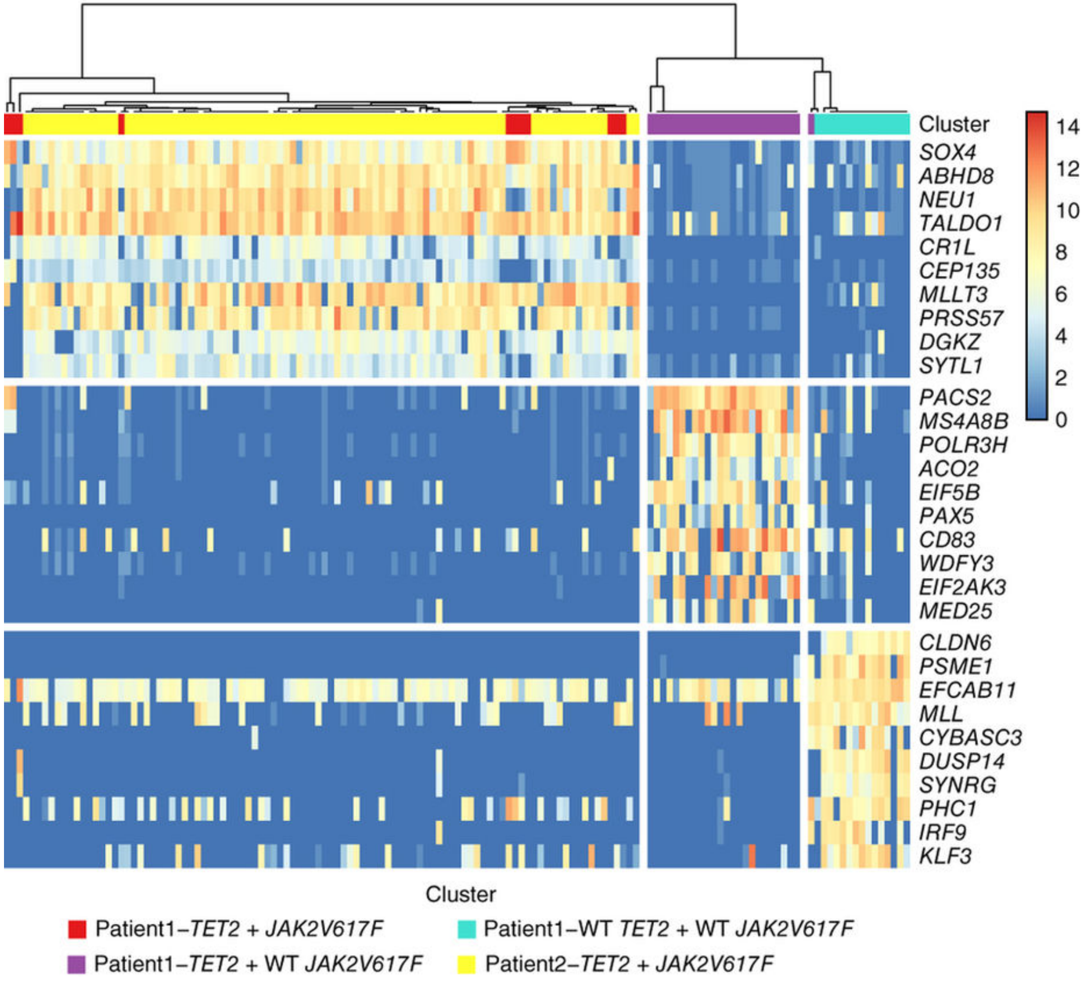
SC3: consensus clustering of single-cell RNA-seq data



Kiselev *et al.*, Nature Methods 2017

Differential Expression Analysis

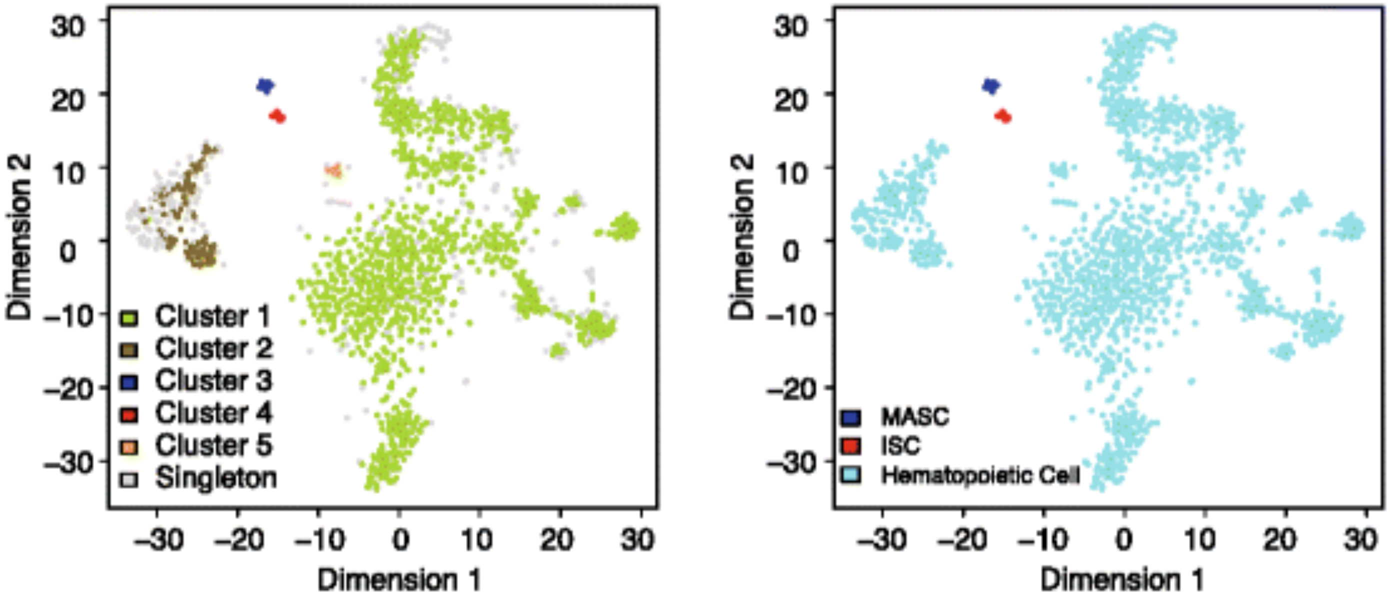
SC3: consensus clustering of single-cell RNA-seq data



Kiselev *et al.*, Nature Methods 2017

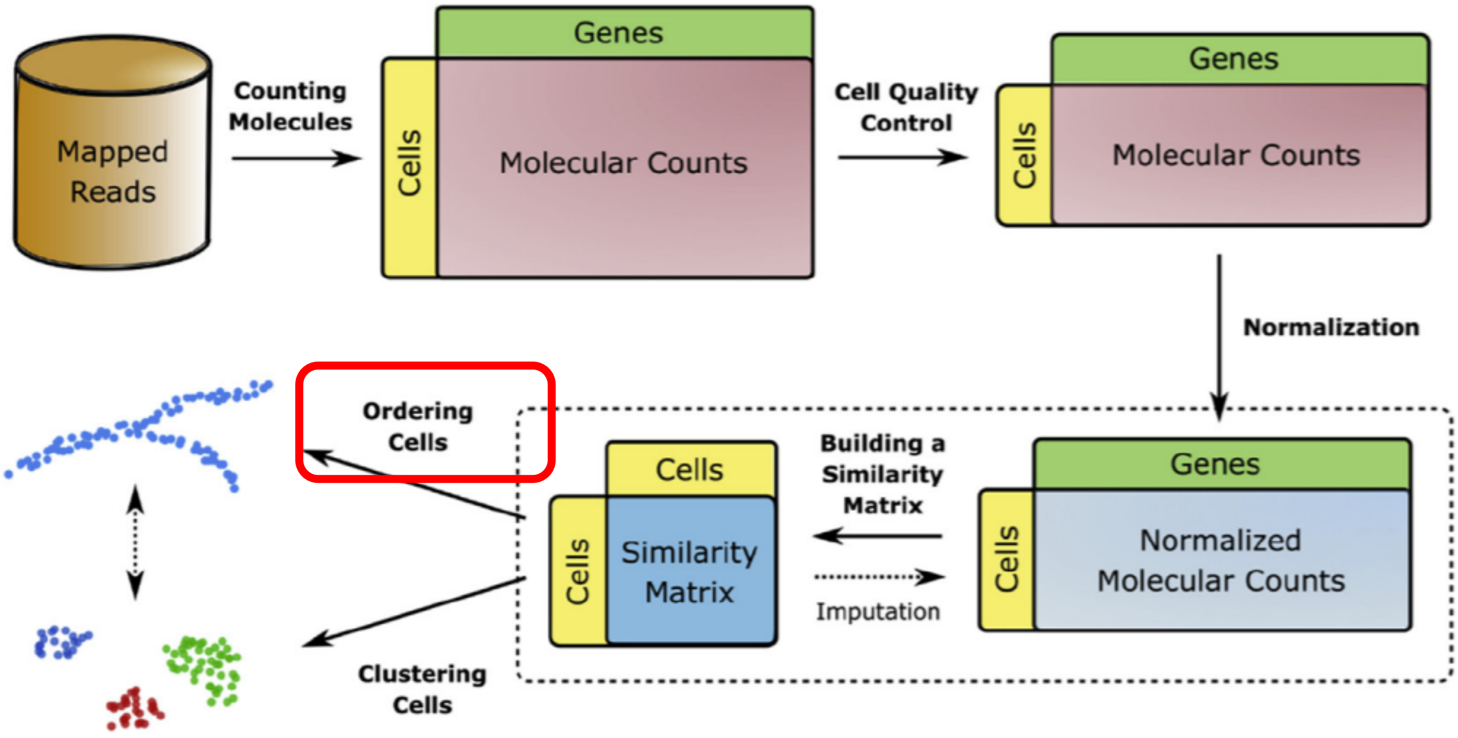
Clustering Cells

GiniClust: detecting rare cell types from single-cell gene expression data with Gini index



Jiang *et al*., Genome Biology 2016

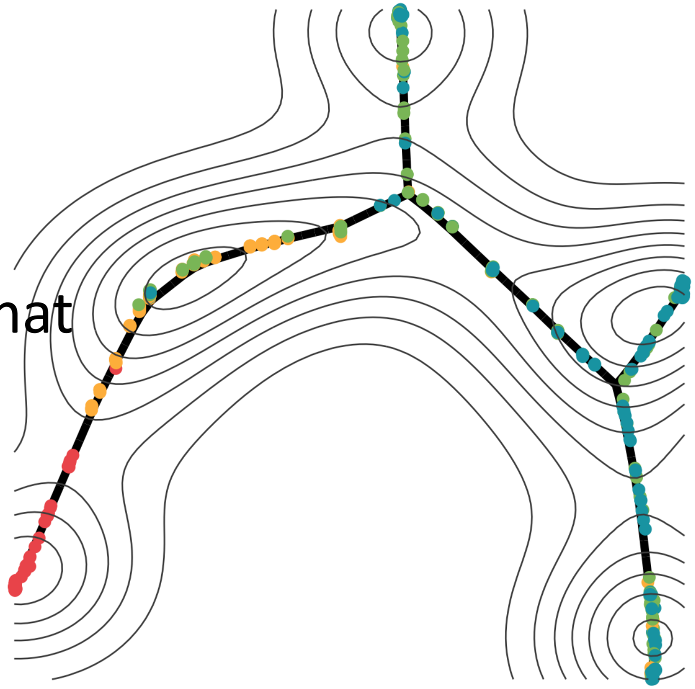
Structure of a generic pipeline



Camara, Current Opinions in Systems Biology, 2018

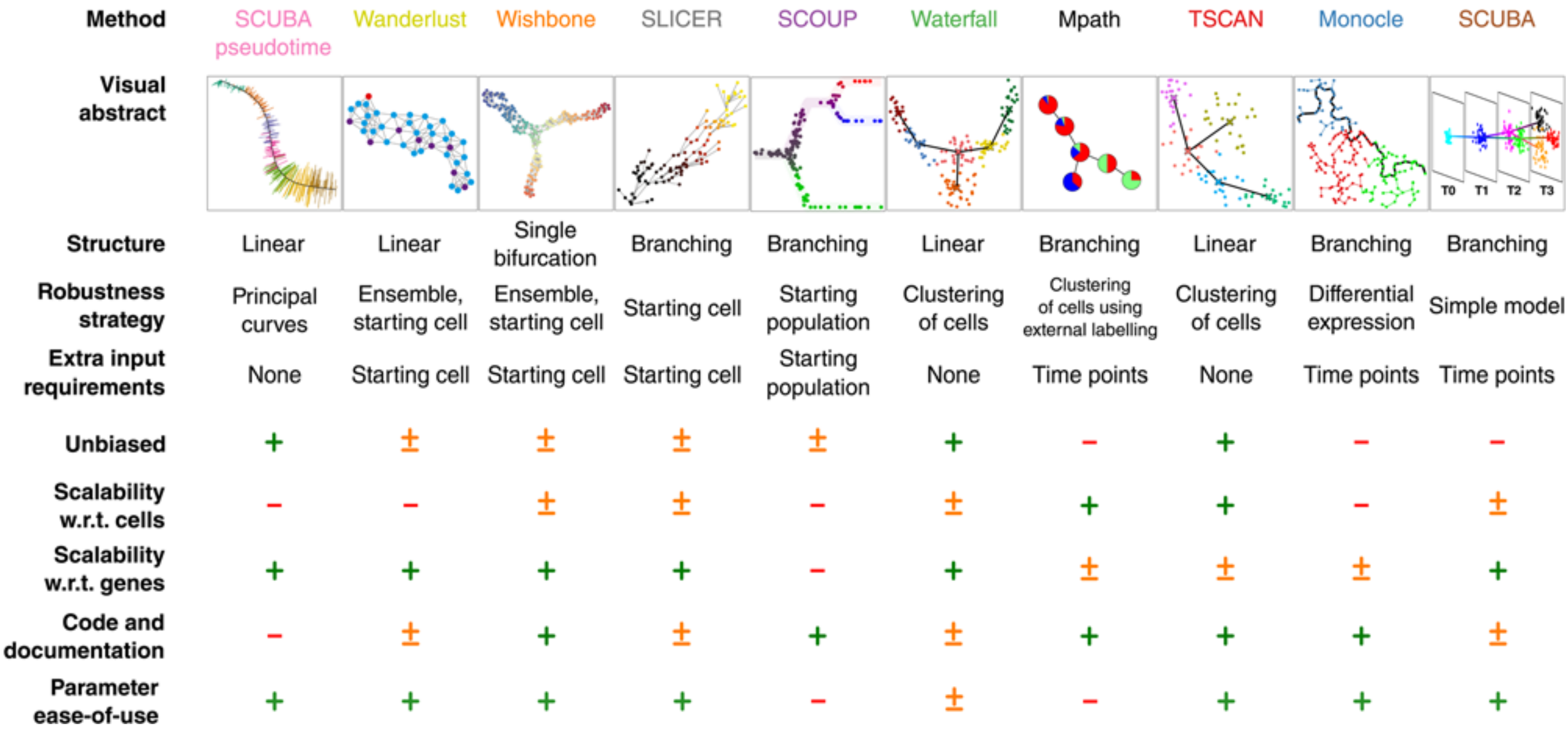
Single Cell Trajectory Inference

- "Pseudotime" introduced in Trapnell *et al.*, Nature Biotechnology 2014 (Monocle)
- Steps:
  1. (Optional) Choose genes that define a biological process
  2. Reduce dimensionality
  3. Order cells



http://cole-trapnell-lab.github.io/monocle-release/

Single Cell Trajectory Inference



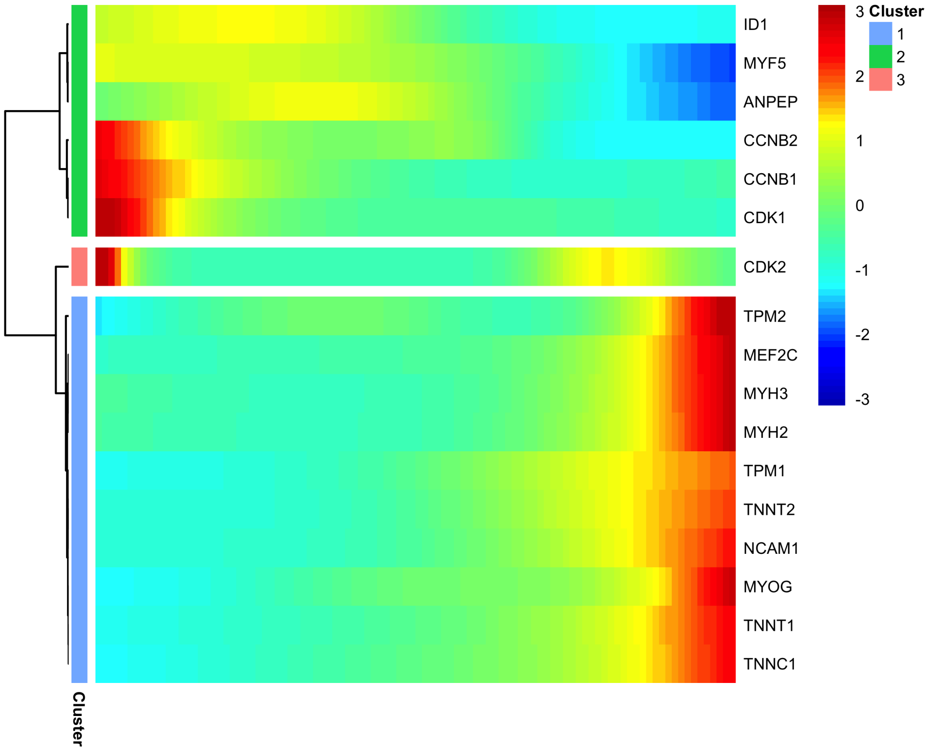
Method	SCUBA pseudotime	Wanderlust	Wishbone	SLICER	SCOUP	Waterfall	Mpath	TSCAN	Monocle	SCUBA
Visual abstract										
Structure	Linear	Linear	Single bifurcation	Branching	Branching	Linear	Branching	Linear	Branching	Branching
Robustness strategy	Principal curves	Ensemble, starting cell	Ensemble, starting cell	Starting cell	Starting population	Clustering of cells	Clustering of cells using external labelling	Clustering of cells	Differential expression	Simple model
Extra input requirements	None	Starting cell	Starting cell	Starting cell	Starting population	None	Time points	None	Time points	Time points
Unbiased	+	±	±	±	±	+	−	+	−	−
Scalability w.r.t. cells	−	−	±	±	−	±	+	+	−	±
Scalability w.r.t. genes	+	+	+	+	−	+	±	±	±	+
Code and documentation	−	±	+	±	+	±	+	+	+	±
Parameter ease-of-use	+	+	+	+	−	±	−	+	+	+

Cannoodt *et al.*, 2016

Single Cell Trajectory Inference

- “Pseudotime” introduced in Trapnell *et al.*, Nature Biotechnology 2014 (Monocle)
- Steps:
    1. (Optional) Choose genes that define a biological process
    2. Reduce dimensionality

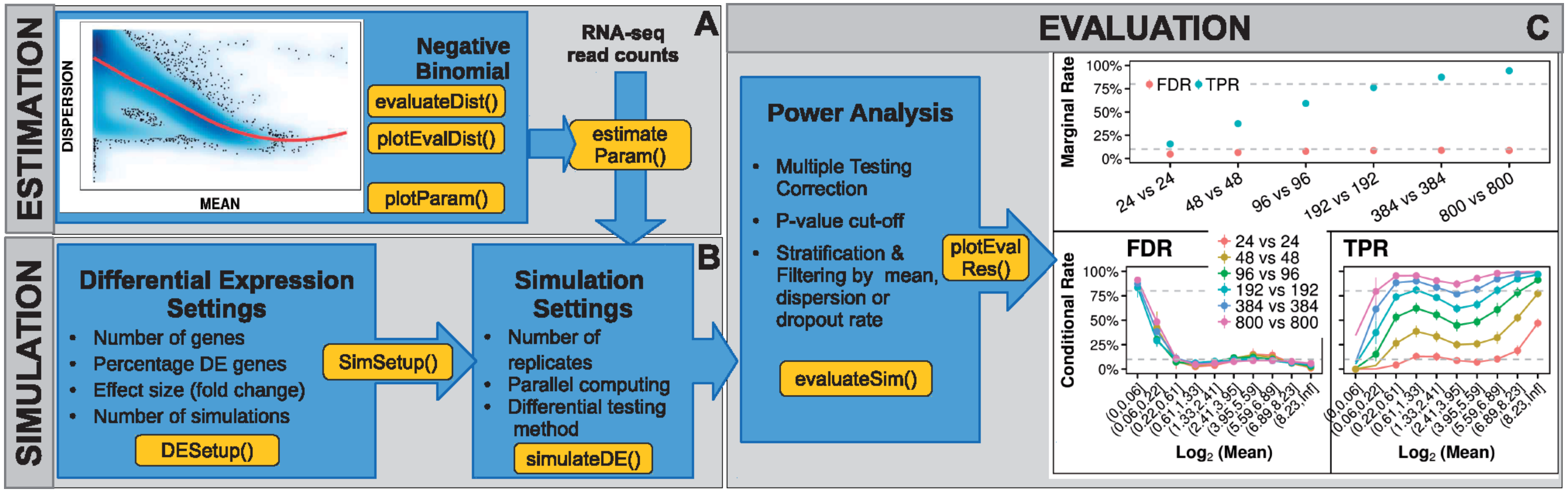
Differential Expression Analysis using Monocle



Qiu *et al.*, Nature Methods 2017

Simulating scRNA-seq data

PowSimR



ESTIMATION

Negative Binomial

evaluateDist()

plotEvalDist()

plotParam()

RNA-seq read counts

estimate Param()

SIMULATION

Differential Expression Settings

- Number of genes
- Percentage DE genes
- Effect size (fold change)
- Number of simulations

DESetup()

SimSetup()

Simulation Settings

- Number of replicates
- Parallel computing
- Differential testing method

simulateDE()

EVALUATION

Power Analysis

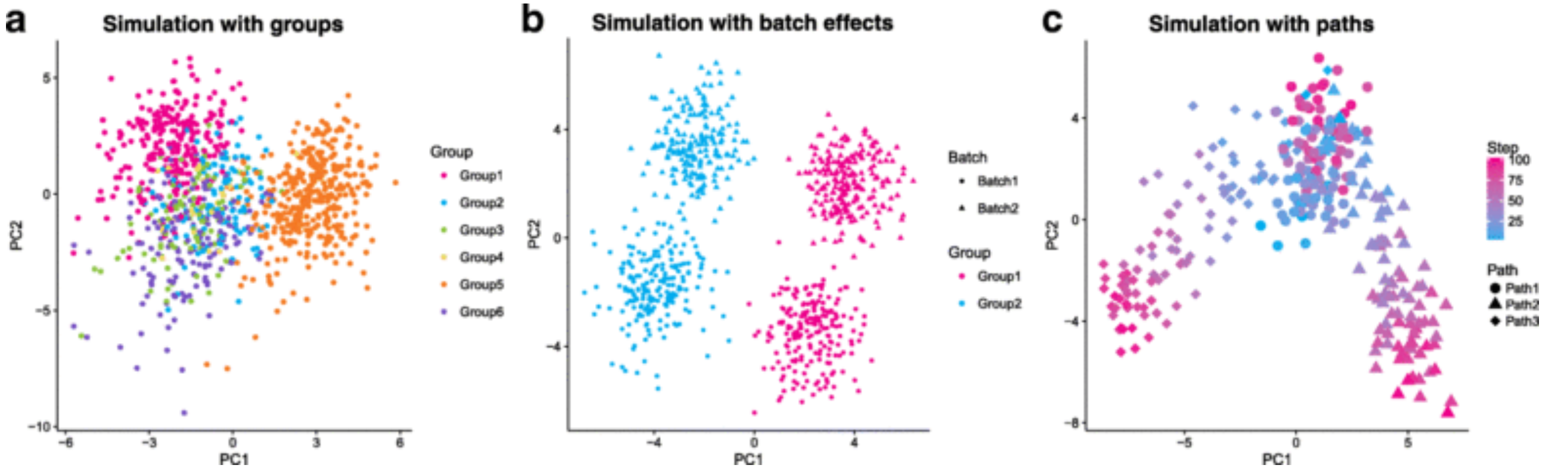
- Multiple Testing Correction
- P-value cut-off
- Stratification & Filtering by mean, dispersion or dropout rate

plotEvalRes()

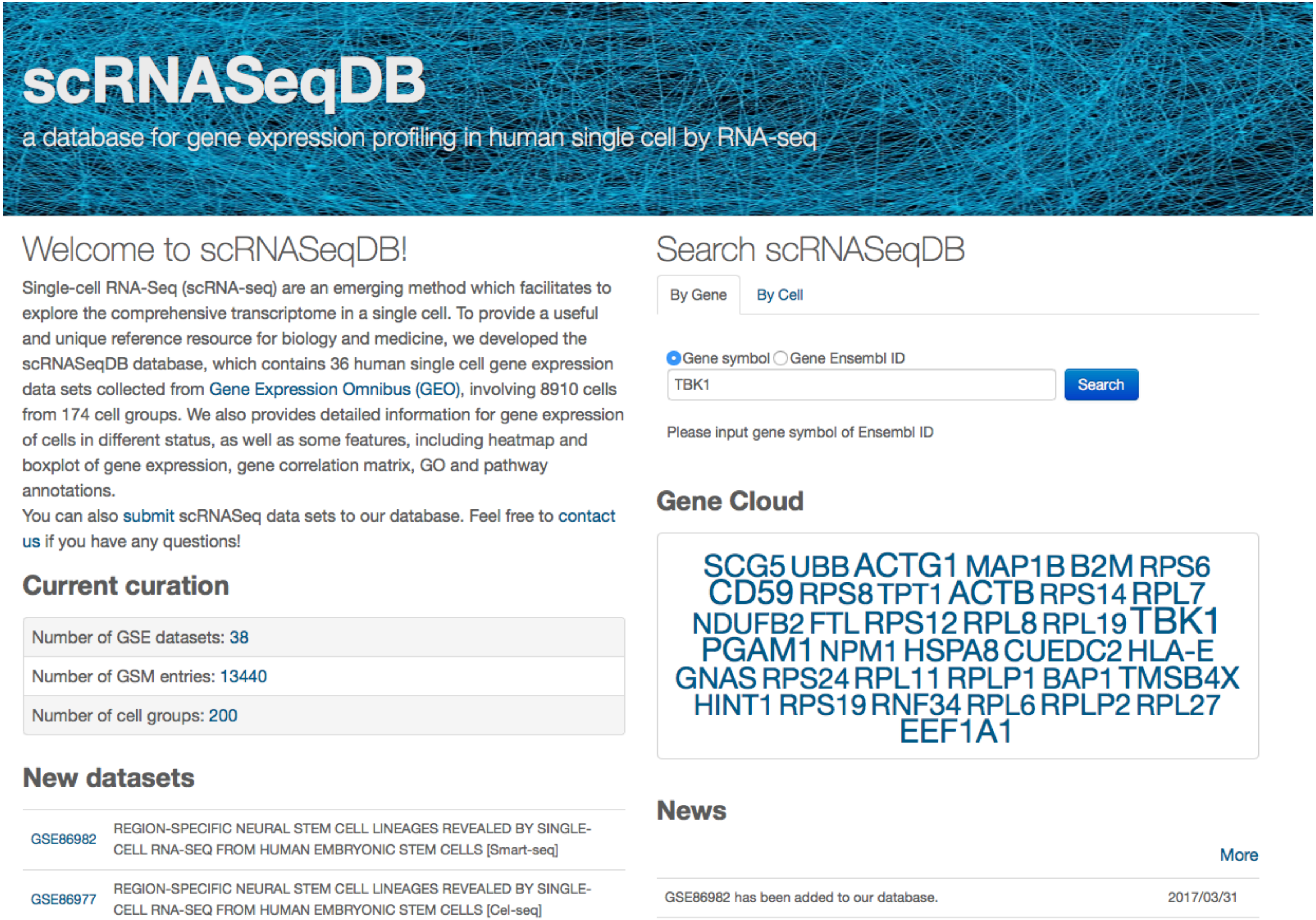
evaluateSim()

Vieth *et al.*, Bioinformatics 2017

Splatter



Zappia *et al.*, Genome Biology 2017



scRNASeqDB

a database for gene expression profiling in human single cell by RNA-seq

Welcome to scRNASeqDB!

Single-cell RNA-Seq (scRNA-seq) are an emerging method which facilitates to explore the comprehensive transcriptome in a single cell. To provide a useful and unique reference resource for biology and medicine, we developed the scRNASeqDB database, which contains 36 human single cell gene expression data sets collected from Gene Expression Omnibus (GEO), involving 8910 cells from 174 cell groups. We also provides detailed information for gene expression of cells in different status, as well as some features, including heatmap and boxplot of gene expression, gene correlation matrix, GO and pathway annotations.

You can also submit scRNASeq data sets to our database. Feel free to contact us if you have any questions!

Current curation

Number of GSE datasets: 38	
Number of GSM entries: 13440	
Number of cell groups: 200	

New datasets

GSE86982	REGION-SPECIFIC NEURAL STEM CELL LINEAGES REVEALED BY SINGLE-CELL RNA-SEQ FROM HUMAN EMBRYONIC STEM CELLS [Smart-seq]
GSE86977	REGION-SPECIFIC NEURAL STEM CELL LINEAGES REVEALED BY SINGLE-CELL RNA-SEQ FROM HUMAN EMBRYONIC STEM CELLS [Cel-seq]

Search scRNASeqDB

By Gene | By Cell

● Gene symbol ○ Gene Ensembl ID

TBK1 | Search

Please input gene symbol of Ensembl ID

Gene Cloud

SCG5 UBB ACTG1 MAP1B B2M RPS6
CD59 RPS8 TPT1 ACTB RPS14 RPL7
NDUFB2 FTL RPS12 RPL8 RPL19 TBK1
PGAM1 NPM1 HSPA8 CUEDC2 HLA-E
GNAS RPS24 RPL11 RPLP1 BAP1 TMSB4X
HINT1 RPS19 RNF34 RPL6 RPLP2 RPL27
EEF1A1

News

More

GSE86982 has been added to our database.	2017/03/31

https://bioinfo.uth.edu/scrnaseqdb/index.php?r=site/index

Questions?

Thanks also to Dr. Priyanka Vijay!