DTU

DTU Health Technology
Bioinformatics

Introduction to NGS technology

Gabriel Renaud
Associate Professor
Section of Bioinformatics
Technical University of Denmark
gabriel.reno@gmail.com

Outline

- 2nd generation NGS
- Illumina movietime!
- Your turn to basecall
- 3rd generation NGS

2 main types of approaches

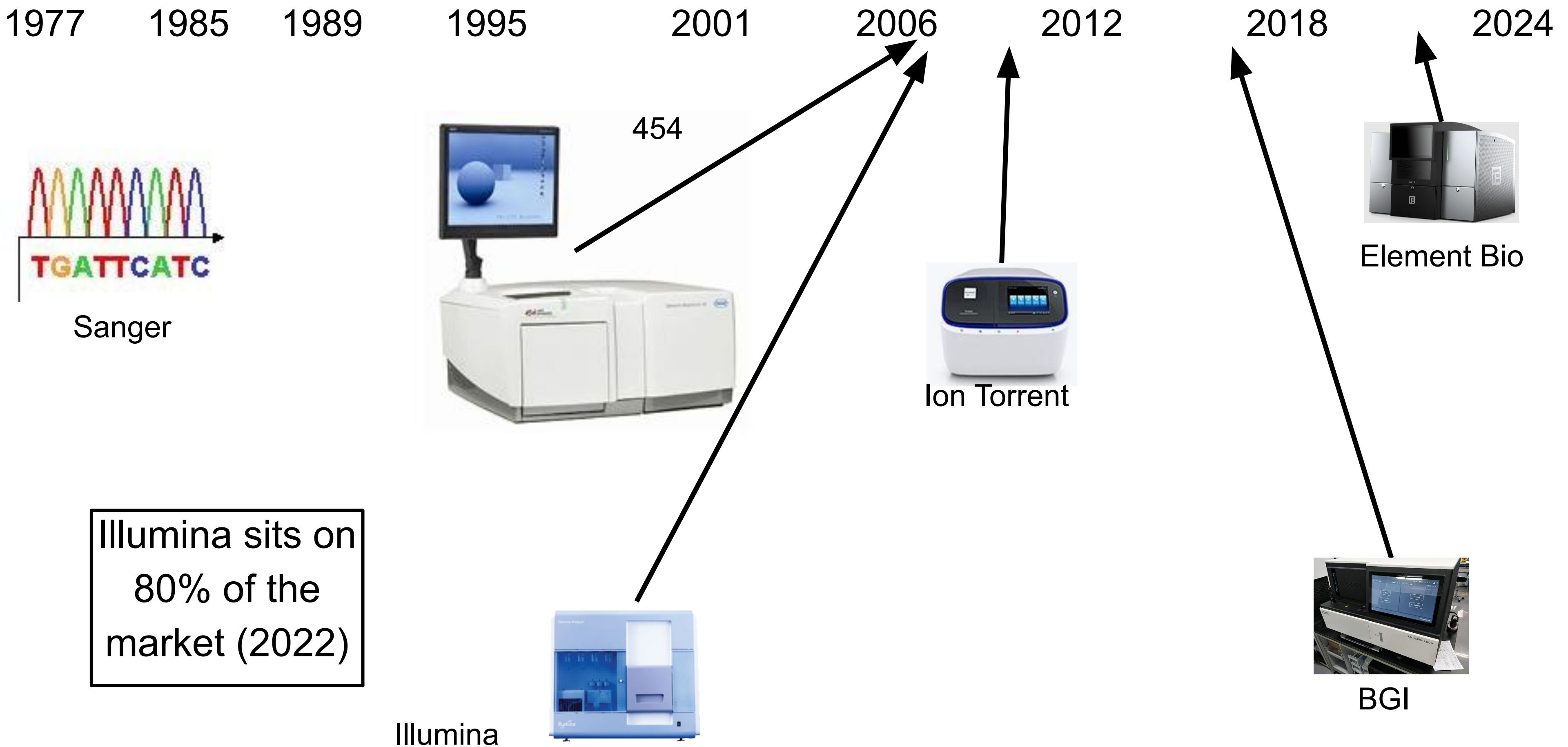
1) Amplify and sequence one base at a time

1:A 2:G 3:G 4:T = AGGT

2) Amplify and count how many of the same base

1:1A 2:2G 3:1T = AGGT

2nd generation

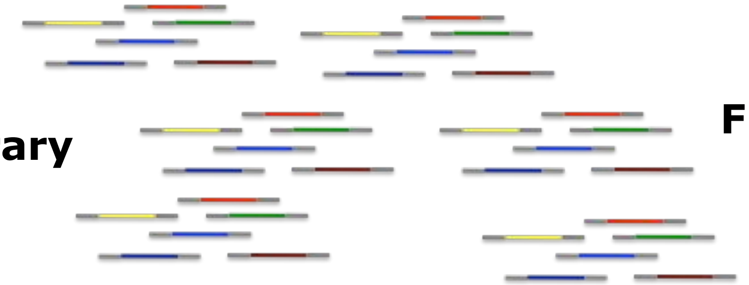


Illumina sits on 80% of the market (2022)

General library preparation steps

1. Create library molecules
2. Amplification (PCR)
3. Massive parallel sequencing (strength over Sanger)

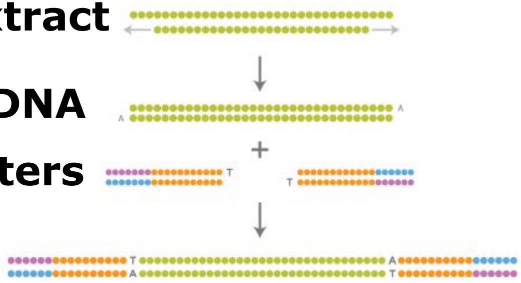
Library



DNA from extract

Fragment & polish DNA

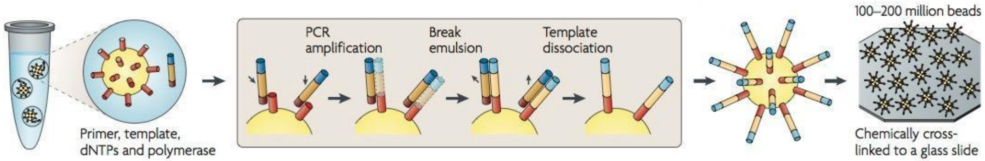
Adapters



Library molecule

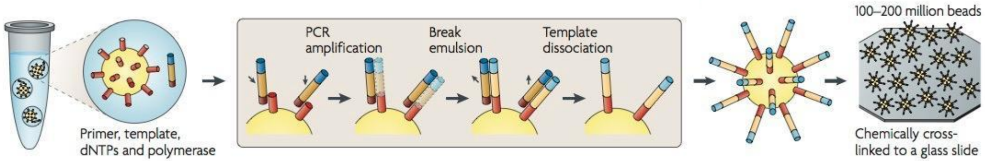
What is common: Amplification and immobilization

- Emulsion PCR (454, SOLiD, IonTorrent): Water, oil, beads, one DNA template/droplet

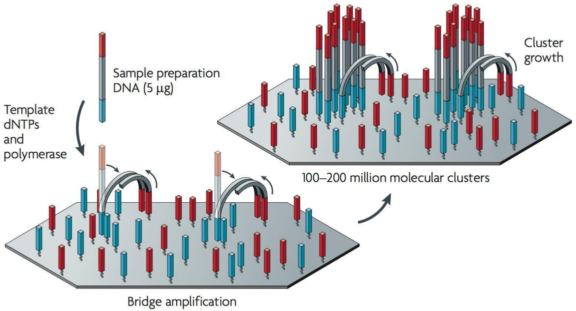


What is common: Amplification and immobilization

- Emulsion PCR (454, SOLiD, IonTorrent): Water, oil, beads, one DNA template/droplet



Bridge PCR (Illumina): One DNA template/cluster, primers on surface, grow by bridging primers



2 main types of approaches

1) Amplify and sequence one base at a time

1:A 2:G 3:G 4:T = AGGT

2) Amplify and count how many of the same base

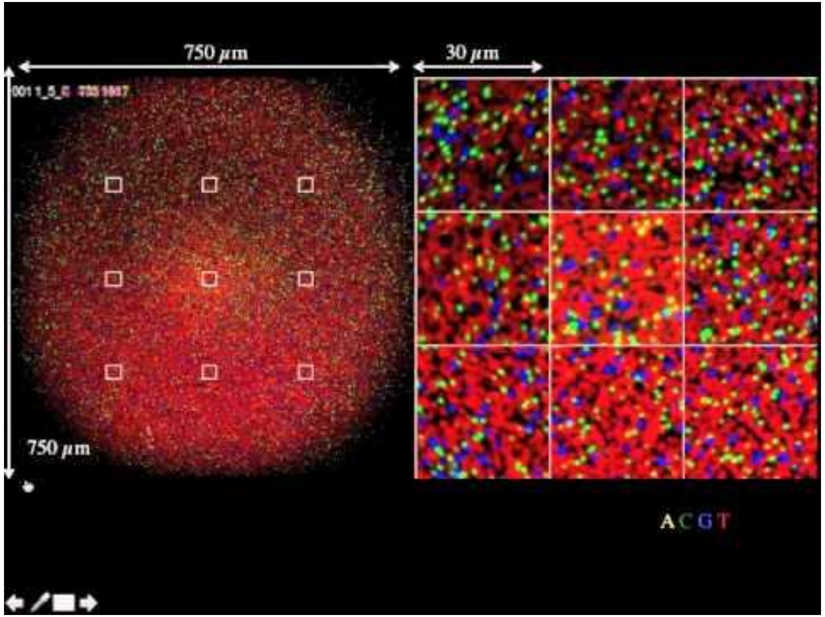
1:1A 2:2G 3:1T = AGGT

Illumina sequencing

corporate propaganda:
https://www.youtube.com/watch?v=HMyCqWhwB8E

Amplicon sequencing on Illumina

- Why can’t you just fill your Illumina flow cell with amplicon libraries (i.e. the same sequence over and over)?



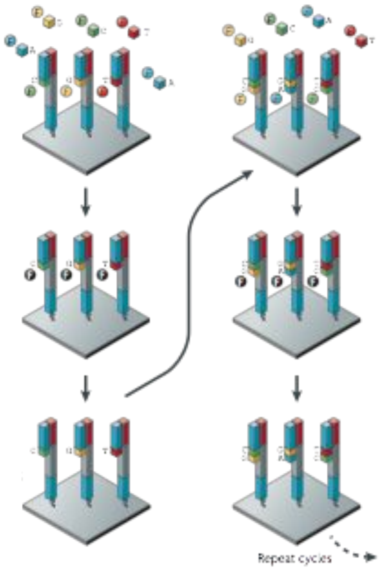
Fluorescence detection

Illumina - Cyclic reversible termination

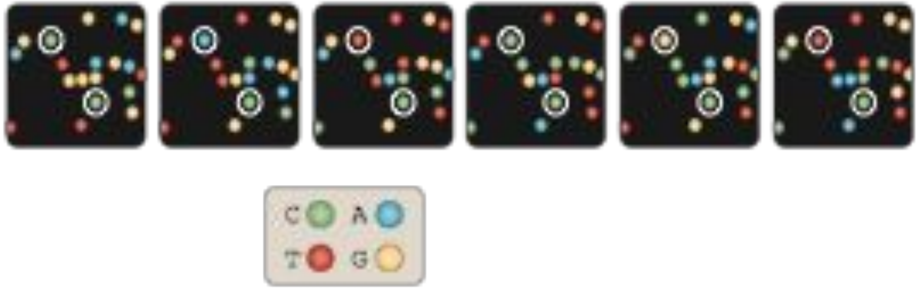
Add all dNTPs labelled w. diff dye

Create four-color image

Cleave dye and repeat next cycle



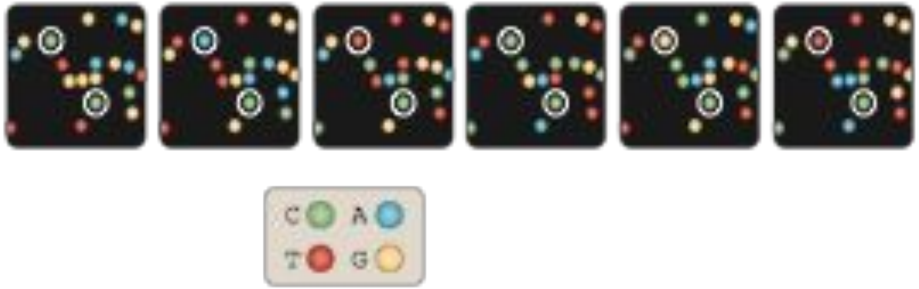
2G: Imaging



Illumina 1:_________

Illumina 2:_________

2G: Imaging Answers!



Illumina 1:_________

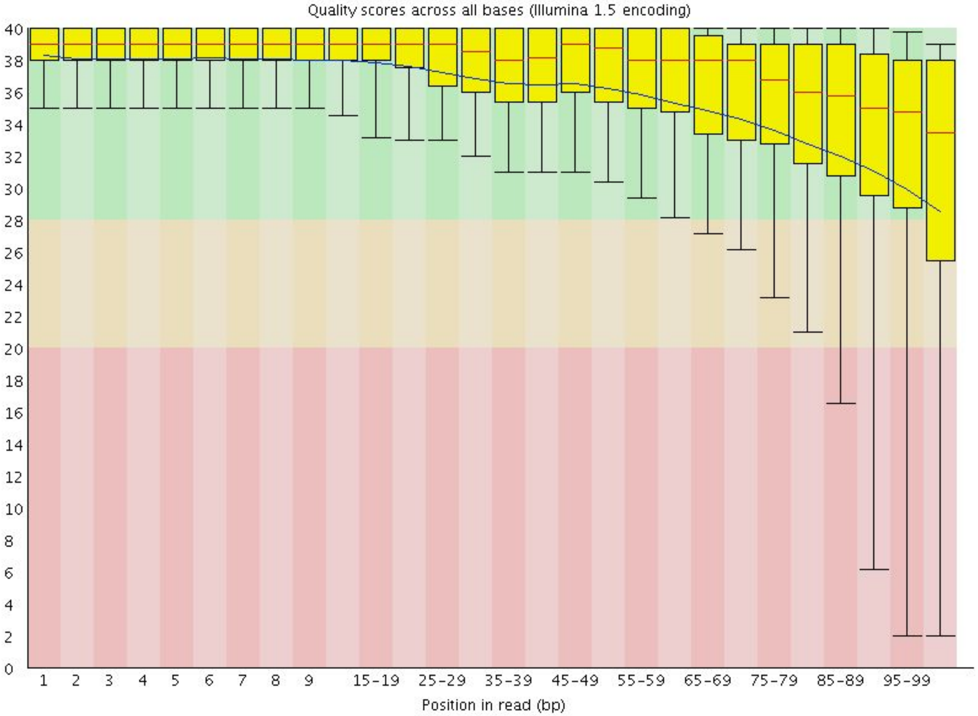
Illumina 2:_________

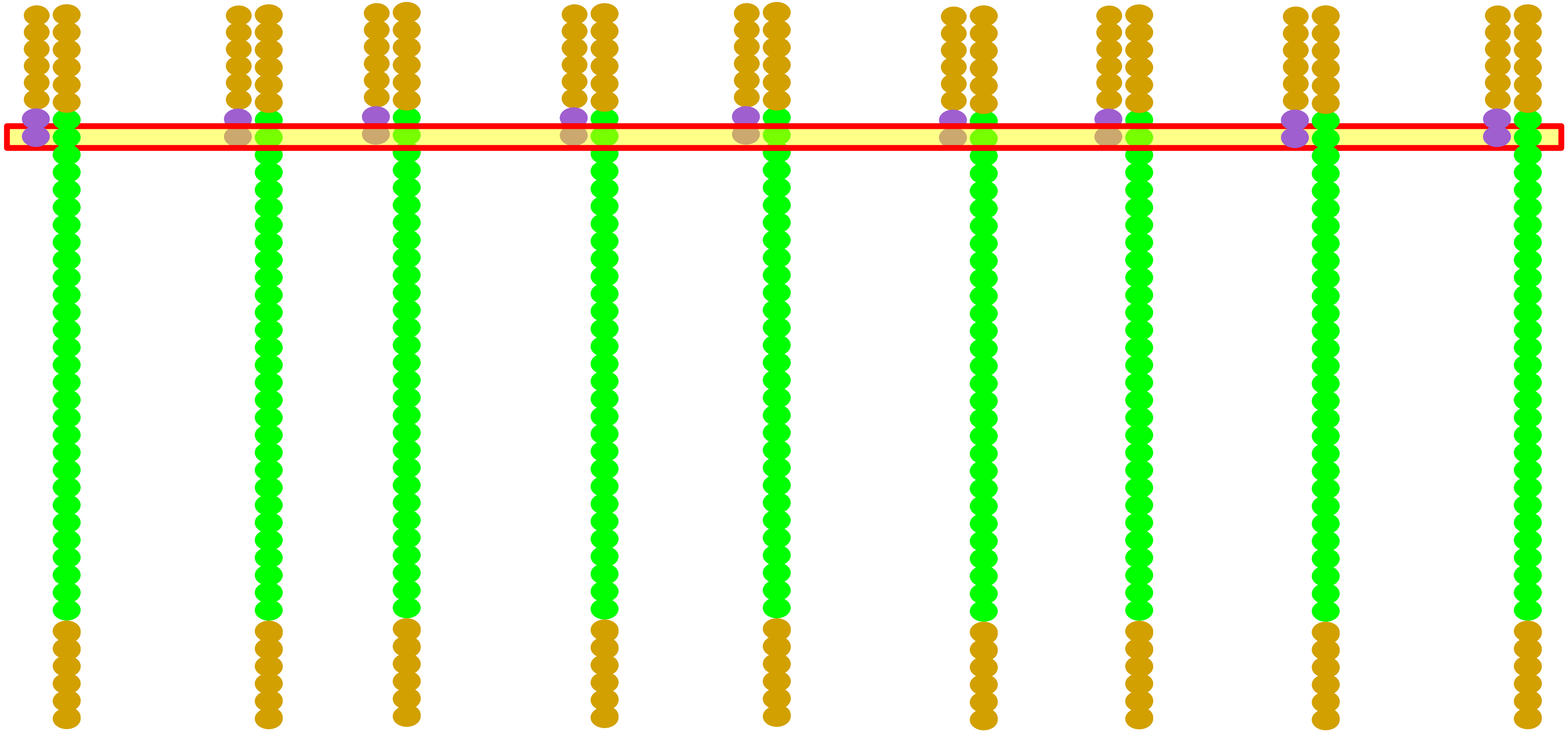
TOP: **CATCGT**

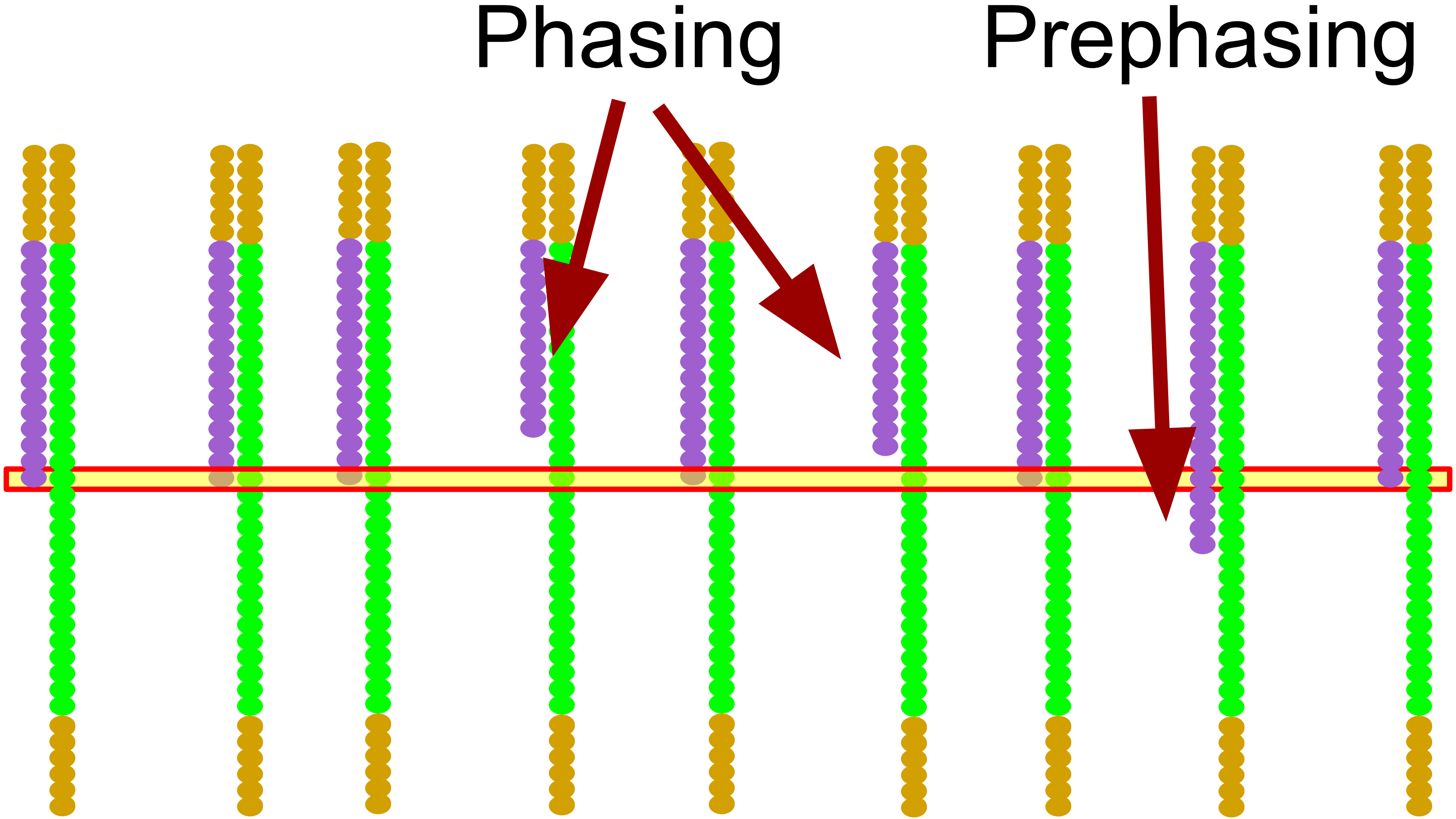
BOTTOM: **CCCCCC**

Illumina: Quality deterioration

- Quality goes down
- Especially 2nd read
- Can you think of why?

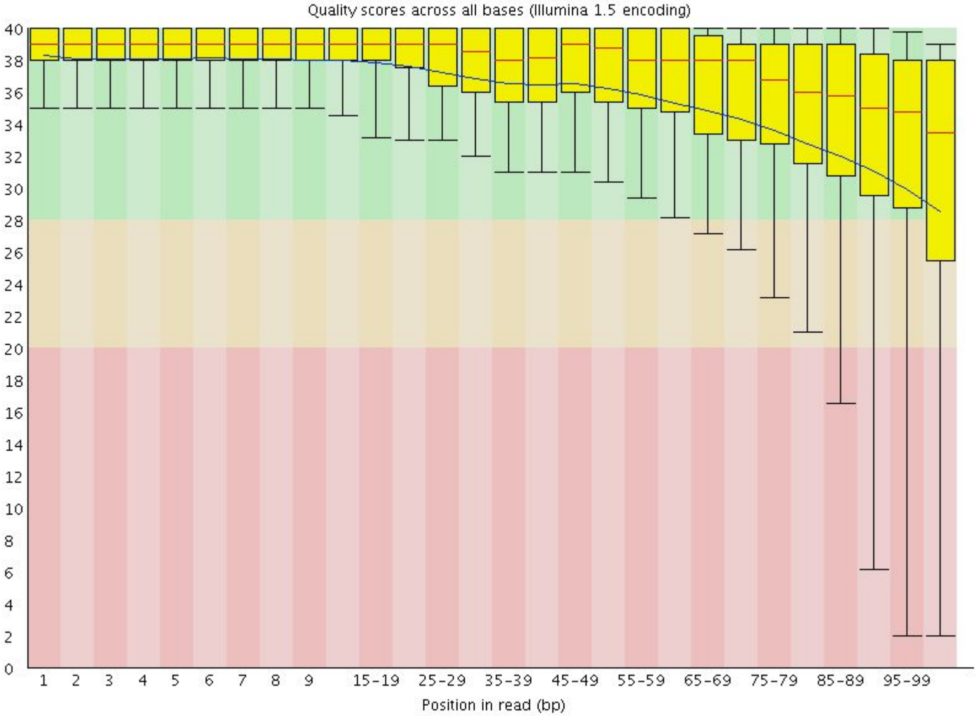





Phasing
Prephasing

Illumina: Quality deterioration

- Quality goes down
- Especially 2nd read
- Can you think of why?



- Efficiency of incorporation
- Phasing
- Prephasing

Brief side note about multiplexing/demultiplexing

- If we sequence a small virus (ex: bacteriophage Phi-X174 with a genome size of 5386 nucleotides), do we need 1B reads?

- Idea to save costs: pool multiple samples together on the same run

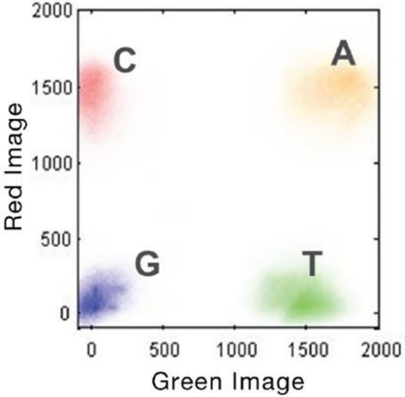
Brief side note about spike-in

- How to know if the sequencing run was successful (low error rate)?

- Idea: Let's spike-in a small virus (ex: bacteriophage Phi-X174 with a genome size of 5386 nucleotides)

NextSeq/NovaSeq (2015-)

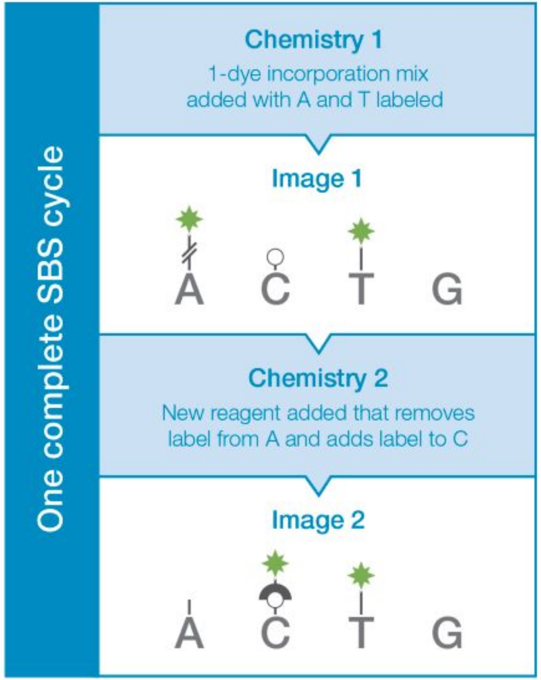
- Chemistry is not based 4 dyes (as before) but 2 dyes
  - – T (green), C (red), A (both) and G (none = “dark”)
  - – Faster processing rate and cheaper reagents
  - – Slightly increases error rate
  - – Problem with G stretches because G is not dyed



source: Illumina

1 dye, 2 images

A.



One complete SBS cycle

Chemistry 1

1-dye incorporation mix added with A and T labeled

Image 1

Chemistry 2

New reagent added that removes label from A and adds label to C

Image 2

B.

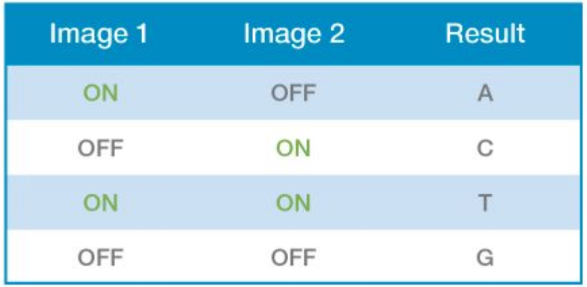
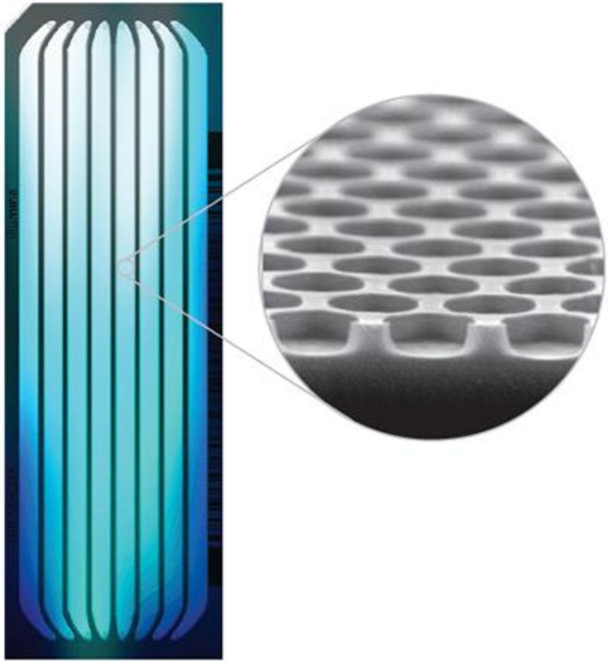


Image 1	Image 2	Result
ON	OFF	A
OFF	ON	C
ON	ON	T
OFF	OFF	G

source: Illumina

Patterned flowcell

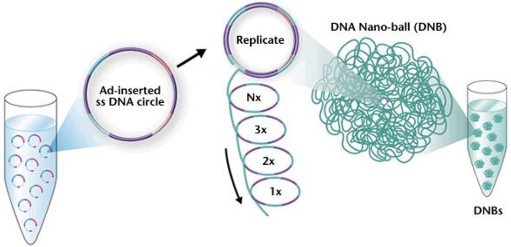
- Patterned wells
- Solves overloading flowcell
- More duplicates



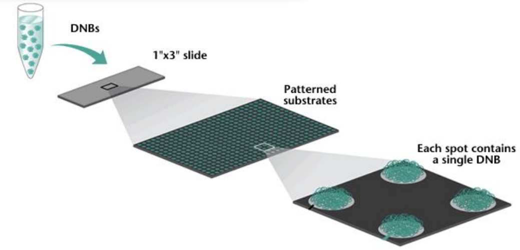
source: Illumina

BGI-Seq

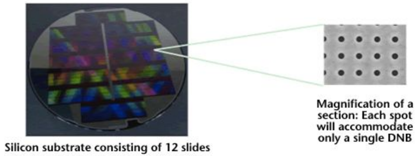
ssDNA -> DNA nanoballs



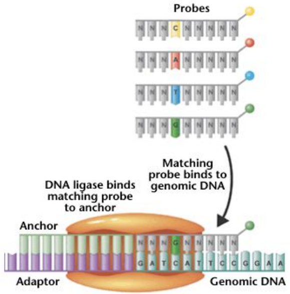
Place DNBs into each spot



Use silicon chips with sticky spots

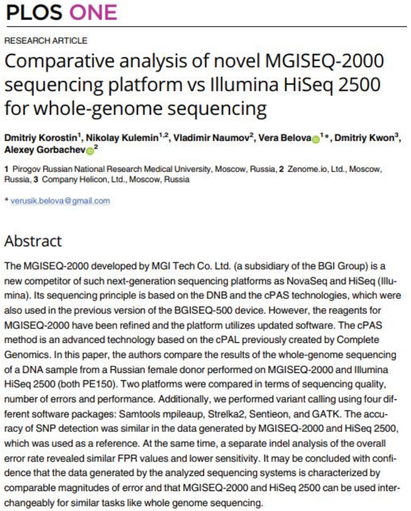


Sequence using ligase and fluorescent labeled probes



BGI-Seq

2020



PLOS ONE

RESEARCH ARTICLE

Comparative analysis of novel MGISEQ-2000 sequencing platform vs Illumina HiSeq 2500 for whole-genome sequencing

Dmitriy Korostin[1], Nikolay Kulemin[1,2], Vladimir Naumov[2], Vera Belova[1]*, Dmitriy Kwon[3], Alexey Gorbachev[2]

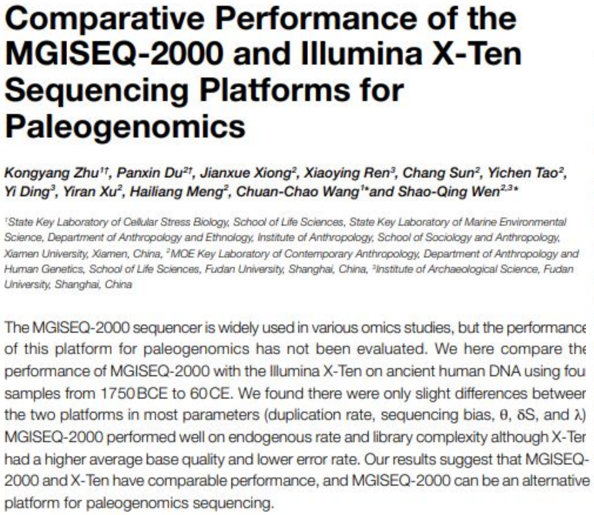
1 Pirogov Russian National Research Medical University, Moscow, Russia, 2 Zenome.io, Ltd., Moscow, Russia, 3 Company Helicon, Ltd., Moscow, Russia

* verusik.belova@gmail.com

Abstract

The MGISEQ-2000 developed by MGI Tech Co. Ltd. (a subsidiary of the BGI Group) is a new competitor of such next-generation sequencing platforms as NovaSeq and HiSeq (Illumina). Its sequencing principle is based on the DNB and the cPAS technologies, which were also used in the previous version of the BGISEQ-500 device. However, the reagents for MGISEQ-2000 have been refined and the platform utilizes updated software. The cPAS method is an advanced technology based on the cPAL previously created by Complete Genomics. In this paper, the authors compare the results of the whole-genome sequencing of a DNA sample from a Russian female donor performed on MGISEQ-2000 and Illumina HiSeq 2500 (both PE150). Two platforms were compared in terms of sequencing quality, number of errors and performance. Additionally, we performed variant calling using four different software packages: Samtools mpileaup, Strelka2, Sentieon, and GATK. The accuracy of SNP detection was similar in the data generated by MGISEQ-2000 and HiSeq 2500, which was used as a reference. At the same time, a separate indel analysis of the overall error rate revealed similar FPR values and lower sensitivity. It may be concluded with confidence that the data generated by the analyzed sequencing systems is characterized by comparable magnitudes of error and that MGISEQ-2000 and HiSeq 2500 can be used interchangeably for similar tasks like whole genome sequencing.

2021

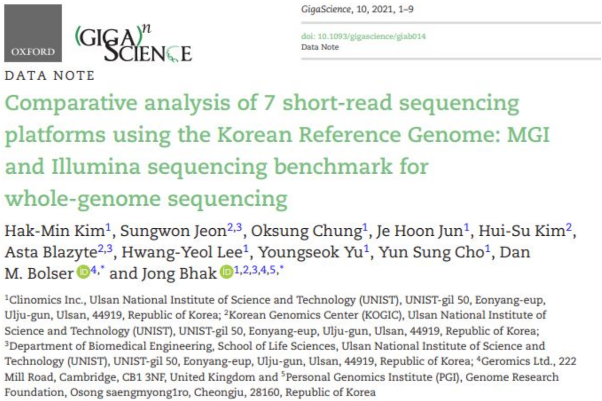


Comparative Performance of the MGISEQ-2000 and Illumina X-Ten Sequencing Platforms for Paleogenomics

Kongyang Zhu[1†], Panxin Du[2†], Jianxue Xiong[2], Xiaoying Ren[3], Chang Sun[2], Yichen Tao[2], Yi Ding[3], Yiran Xu[2], Hailiang Meng[2], Chuan-Chao Wang[1]and Shao-Qing Wen[2,3*]*

[1]State Key Laboratory of Cellular Stress Biology, School of Life Sciences, State Key Laboratory of Marine Environmental Science, Department of Anthropology and Ethnology, Institute of Anthropology, School of Sociology and Anthropology, Xiamen University, Xiamen, China, [2]MOE Key Laboratory of Contemporary Anthropology, Department of Anthropology and Human Genetics, School of Life Sciences, Fudan University, Shanghai, China, [3]Institute of Archaeological Science, Fudan University, Shanghai, China

The MGISEQ-2000 sequencer is widely used in various omics studies, but the performance of this platform for paleogenomics has not been evaluated. We here compare the performance of MGISEQ-2000 with the Illumina X-Ten on ancient human DNA using fou samples from 1750 BCE to 60 CE. We found there were only slight differences between the two platforms in most parameters (duplication rate, sequencing bias, θ, δS, and λ) MGISEQ-2000 performed well on endogenous rate and library complexity although X-Ten had a higher average base quality and lower error rate. Our results suggest that MGISEQ-2000 and X-Ten have comparable performance, and MGISEQ-2000 can be an alternative platform for paleogenomics sequencing.



OXFORD (GIGA)^n SCIENCE

GigaScience, 10, 2021, 1–9

doi: 10.1093/gigascience/giab014
Data Note

DATA NOTE

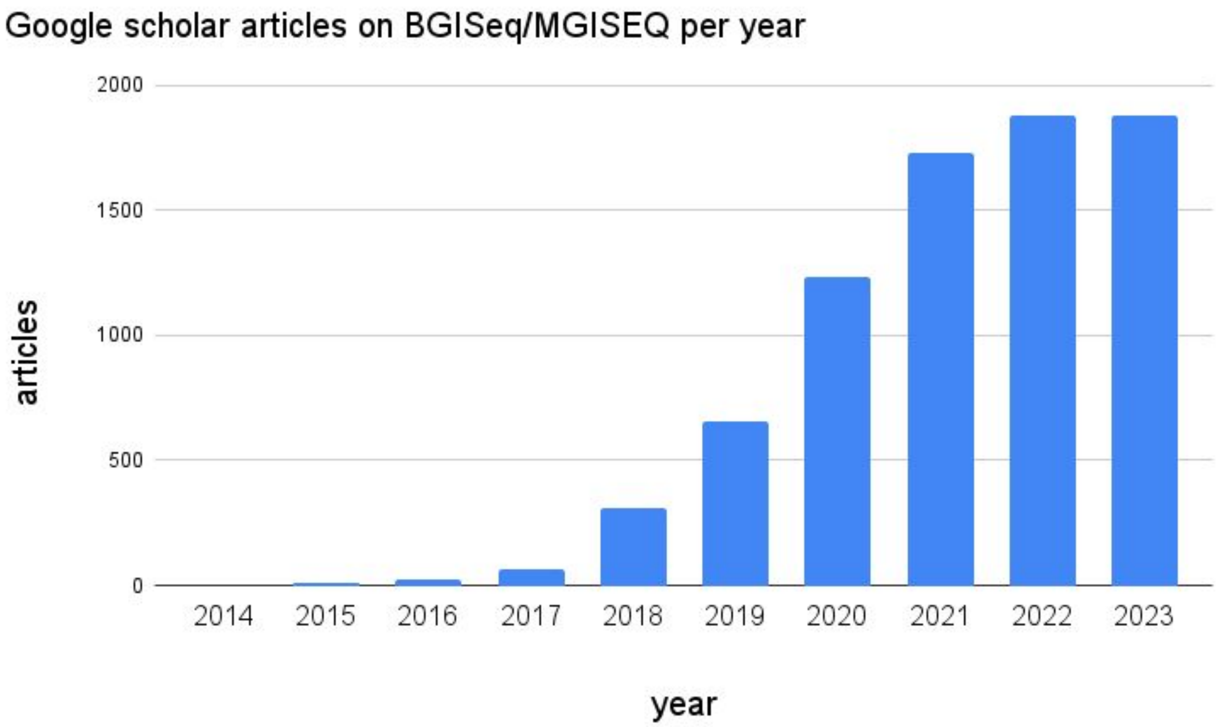
Comparative analysis of 7 short-read sequencing platforms using the Korean Reference Genome: MGI and Illumina sequencing benchmark for whole-genome sequencing

Hak-Min Kim[1], Sungwon Jeon[2,3], Oksung Chung[1], Je Hoon Jun[1], Hui-Su Kim[2], Asta Blazyte[2,3], Hwang-Yeol Lee[1], Youngseok Yu[1], Yun Sung Cho[1], Dan M. Bolser[4,*] and Jong Bhak[1,2,3,4,5,*]

[1]Clinomics Inc., Ulsan National Institute of Science and Technology (UNIST), UNIST-gil 50, Eonyang-eup, Ulju-gun, Ulsan, 44919, Republic of Korea; [2]Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology (UNIST), UNIST-gil 50, Eonyang-eup, Ulju-gun, Ulsan, 44919, Republic of Korea; [3]Department of Biomedical Engineering, School of Life Sciences, Ulsan National Institute of Science and Technology (UNIST), UNIST-gil 50, Eonyang-eup, Ulju-gun, Ulsan, 44919, Republic of Korea; [4]Geromics Ltd., 222 Mill Road, Cambridge, CB1 3NF, United Kingdom and [5]Personal Genomics Institute (PGI), Genome Research Foundation, Osong saengmyong1ro, Cheongju, 28160, Republic of Korea

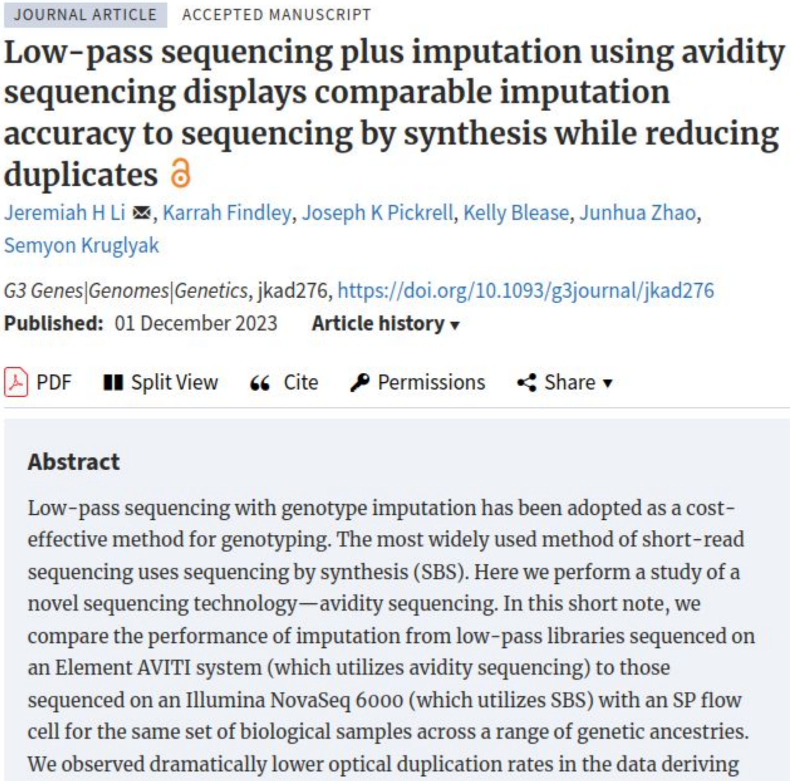
Conclusion: BGI = Illumina in terms of errors but cheaper

BGI-Seq


Google scholar articles on BGISeq/MGISEQ per year
2000
1500
1000
500
0
articles
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
year

Avidity sequencing (new 2022/2023)

Element Biosciences



JOURNAL ARTICLE ACCEPTED MANUSCRIPT

Low-pass sequencing plus imputation using avidity sequencing displays comparable imputation accuracy to sequencing by synthesis while reducing duplicates

Jeremiah H Li, Karrah Findley, Joseph K Pickrell, Kelly Blease, Junhua Zhao, Semyon Kruglyak

G3 Genes|Genomes|Genetics, jkad276, https://doi.org/10.1093/g3journal/jkad276

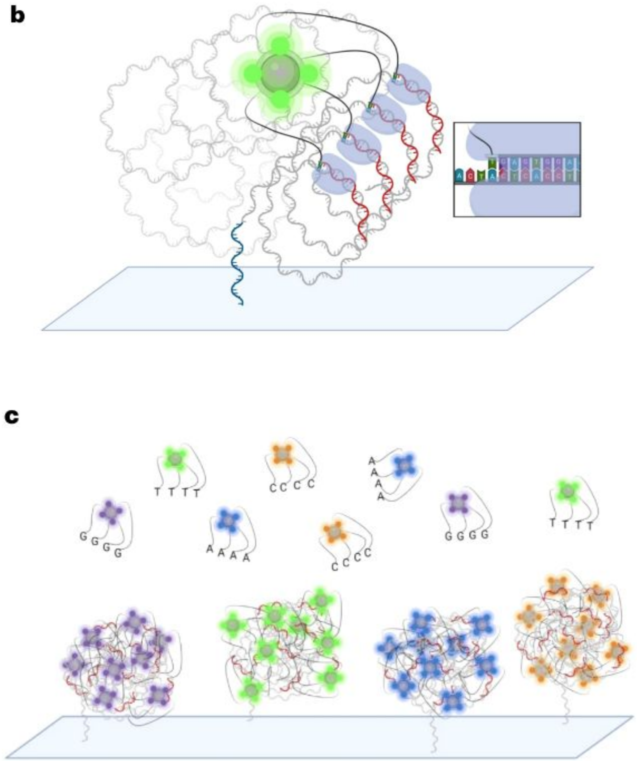
Published: 01 December 2023 **Article history** ▾

Abstract

Low-pass sequencing with genotype imputation has been adopted as a cost-effective method for genotyping. The most widely used method of short-read sequencing uses sequencing by synthesis (SBS). Here we perform a study of a novel sequencing technology—avidity sequencing. In this short note, we compare the performance of imputation from low-pass libraries sequenced on an Element AVITI system (which utilizes avidity sequencing) to those sequenced on an Illumina NovaSeq 6000 (which utilizes SBS) with an SP flow cell for the same set of biological samples across a range of genetic ancestries. We observed dramatically lower optical duplication rates in the data deriving

Avidity sequencing (new 2022/2023)

Element Biosciences



Avidity sequencing (new 2022/2023)

Element
Biosciences

Claims to:

- Less errors than Illumina
- Cheaper than Illumina

2 main types of approaches

1) Amplify and sequence one base at a time

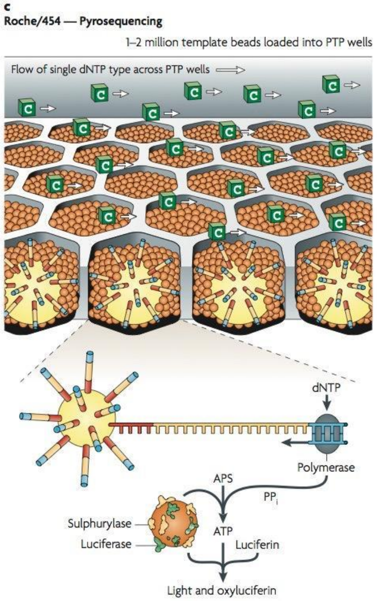
1:A 2:G 3:G 4:T = AGGT

2) Amplify and count how many of the same base

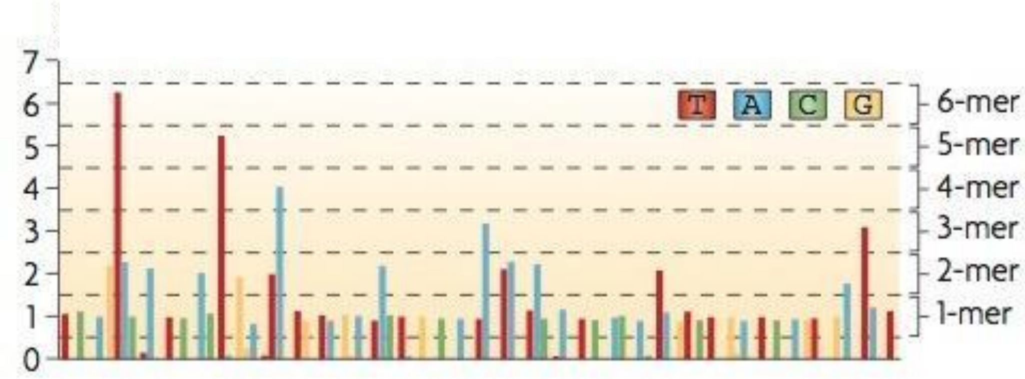
1:1A 2:2G 3:1T = AGGT

454: Pyrosequencing

1. Load template beads into wells
2. Flow one dNTP across wells
3. Polymerase incorporates nucleotide
4. Release of PPi leads to light
5. Light intensity= # of bases
6. Imaging, next dNTP


c
Roche/454 — Pyrosequencing
1–2 million template beads loaded into PTP wells
Flow of single dNTP type across PTP wells
C
dNTP
Polymerase
APS
PP$_i$
ATP
Sulphurylase
Luciferase
Luciferin
Light and oxyluciferin

2G: Imaging handout

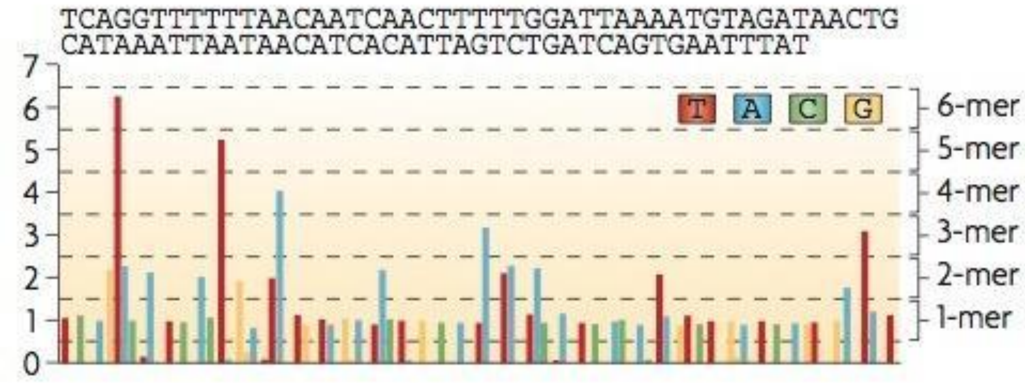


454: ____________________

____

____

2G: Imaging handout Answers!



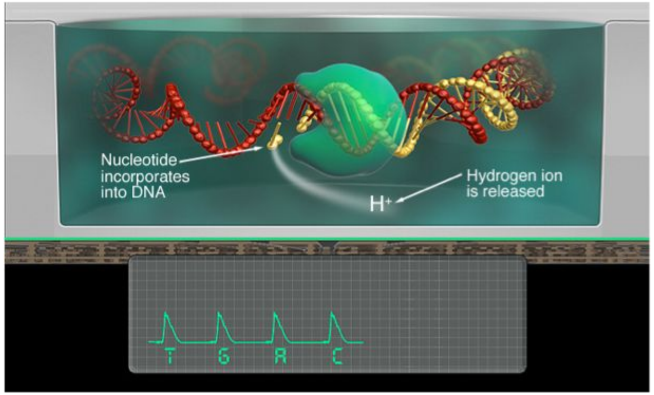
454: ____________________

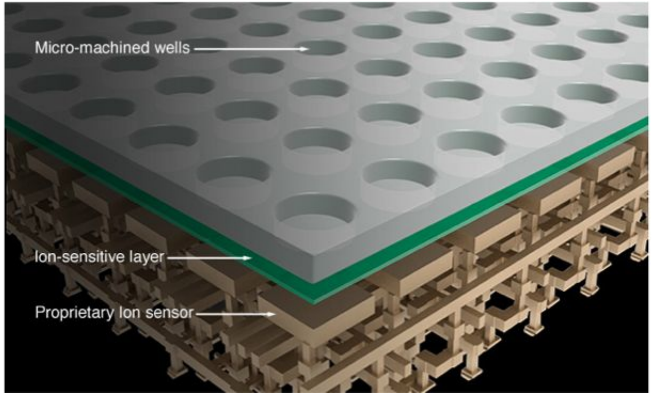
____

____

Ion Torrent

- Similar principle to 454
- Library: Emulsion PCR
- Based on semiconductors
- Detection is based on H ions (pH) changes





Let’s remember the types of errors

mismatch

```
AGCAATCTCAATTACAAATATACACCAACAAA

AGCAATCTCAATTACAGATATACACCAACAAA
```

insert

```
AGCAATCTCAATTACA-AATATACACCAACAA

AGCAATCTCAATTACACAATATACACCAACAA
```

deletion

```
AGCAATCTCAATTACAAATATACACCAACAA

AGCAATCTCAATTACA-ATATACACCAACAA
```

Quiz!

Which of the the 2 main types of approaches would be more prone to indels?

1) Amplify and sequence one base at a time

1:A 2:G 3:G 4:T = AGGT

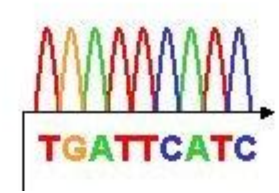
2) Amplify and count how many of the same base

1:1A 2:2G 3:1T = AGGT

Technology	read length	# of reads	errors?
Sanger	400 to 900 bp	96	mm 0.01%
Illumina MiSeq	2x 200-300bp	20-30 M per flow cell	mm 0.1-0.2%
Illumina NextSeq	2x 100-150bp	~400M-1G per flow cell	mm 0.1-0.2%
Illumina NovaSeq	2x 100-250bp	~20G per flow cell	mm 0.1%?
MGI-DNBSEQ-T7	2x 100-200bp	~20G per flow cell	<mm 0.1%
AVITI	2x150 bp	1G reads?	<mm 0.1%?

3rd generation

1977 1985 1989 1995 2001 2006 2012 2018 2024



Sanger

PacBio

Oxford Nanopore

PacBio's revio

3rd generation

- Single-molecule sequencing
- No amplification -> less bias -> observations are more independent

Helicos

PacBio

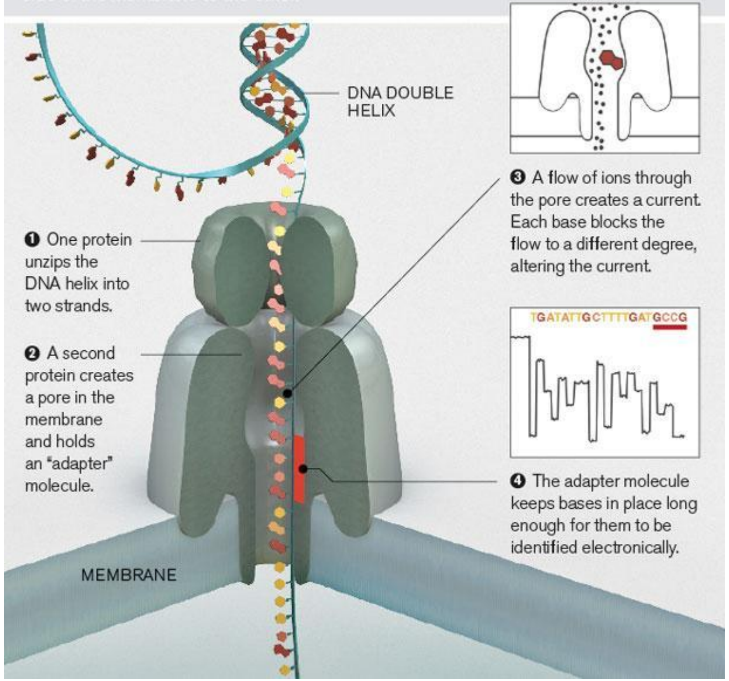
Oxford Nanopore

Oxford Nanopore

- Literal nanopores
- Current per base
- Non-random errors
- https://www.youtube.com/watch?v=RcP85JHLmnI
- Very high error rate

“If a nanopore was the size of a fist, a 1MB strand of DNA passing through that nanopore would be 2 miles (3.2 km) long”
-Adam Philippy, NHGRI

DNA can be sequenced by threading it through a microscopic pore in a membrane. Bases are identified by the way they affect ions flowing through the pore from one side of the membrane to the other.



❶ One protein unzips the DNA helix into two strands.

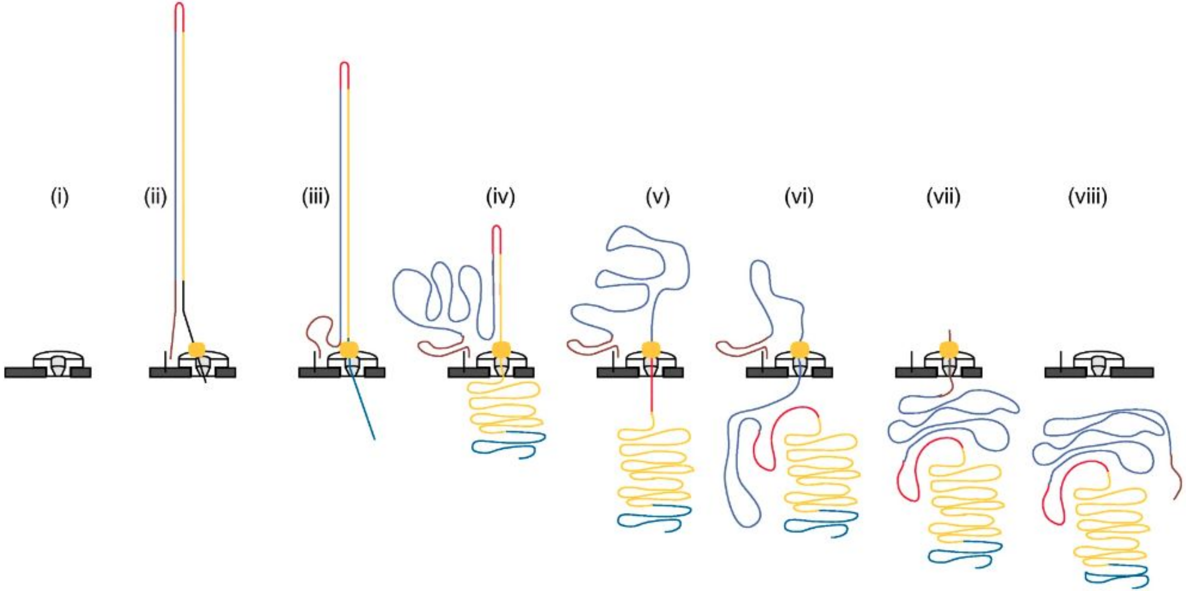
❷ A second protein creates a pore in the membrane and holds an “adapter” molecule.

❸ A flow of ions through the pore creates a current. Each base blocks the flow to a different degree, altering the current.

❹ The adapter molecule keeps bases in place long enough for them to be identified electronically.

Oxford Nanopore

- Hairpin allows double sequencing (2D)



Jain, M., Olsen, H.E., Paten, B. et al. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol 17, 239 (2016). https://doi.org/10.1186/s13059-016-1103-0

Cheap & mobile

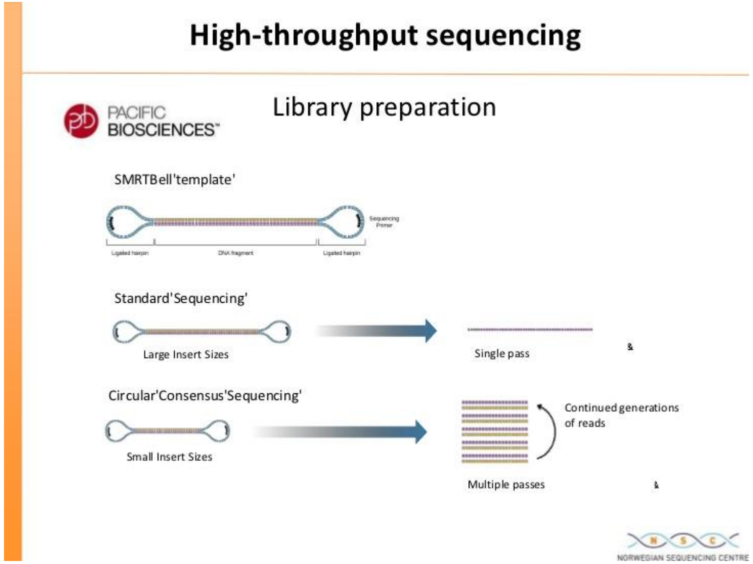
- Long reads, low quality
- Low establishment and maintenance costs
- Portability

PacBio: Single-molecule real-time (SMRT) sequencing

- Expensive machinery
- Not very portable

PacBio

- Flexibility
  - Long but low quality or shorter but better reads
  - Robust
  - https://www.youtube.com/watch?v=_lD8JyAbwEo
  - 2019: HiFi read same fragment multiple times
  - New 2022: Revio
    “Revio is designed to provide customers with the ability to sequence up to 1,300 human whole genomes per year at 30-fold coverage for less than $1,000 per genome. “

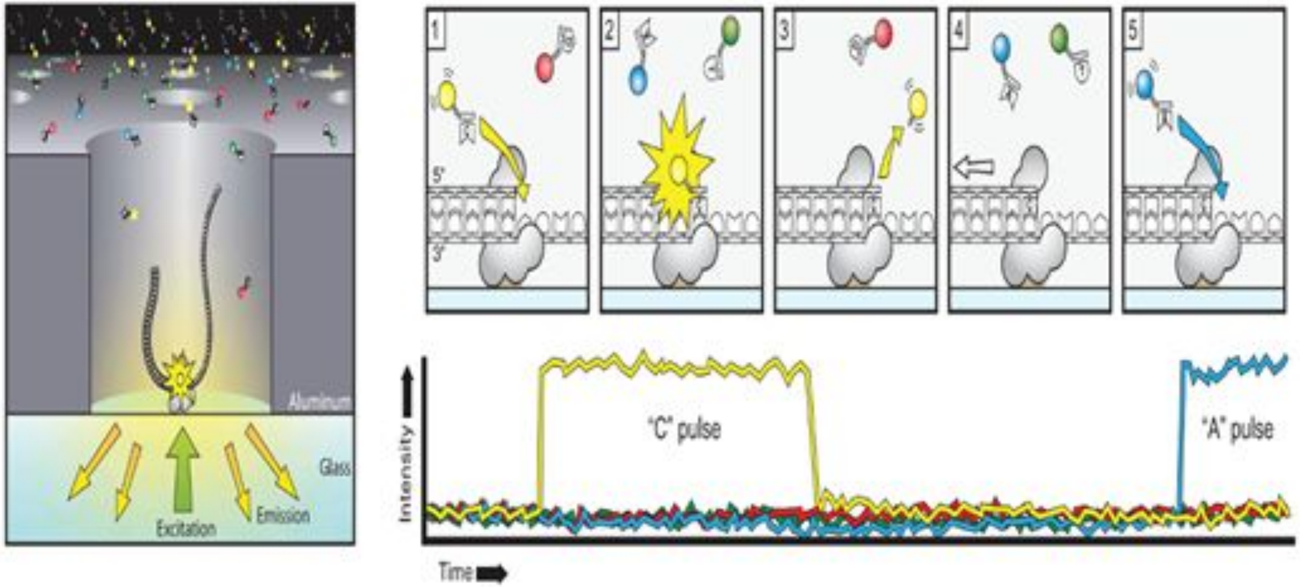


High-throughput sequencing

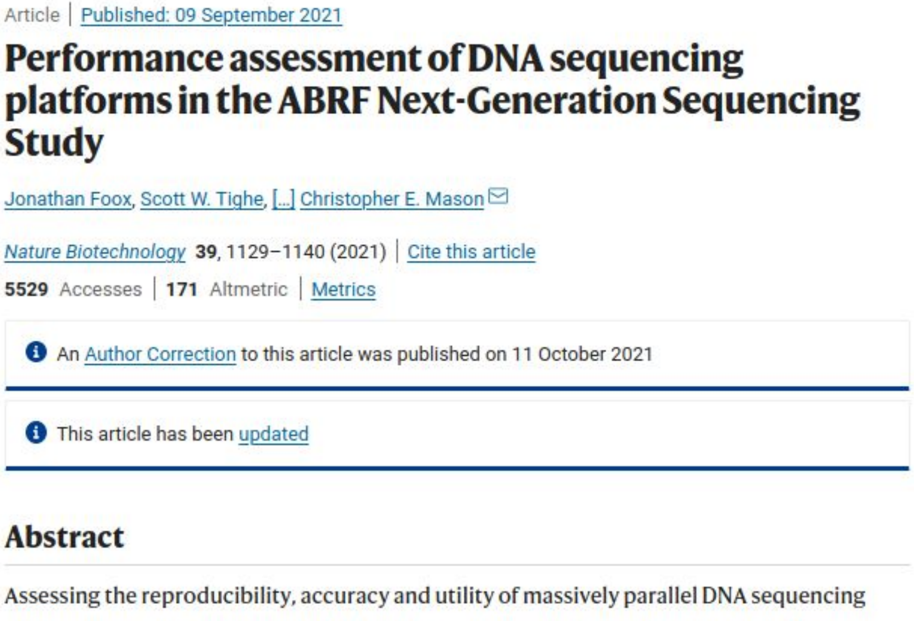
Library preparation

Tiny wells

- 1 million wells per cell
- Hit the lights



Technology	read length	# of reads	errors?
Oxford Nanopore	avg. 2 kbp-20 kbp	2M-6G	2022 update: ~1-3% 1D: indel+mm 20% 2D: indel+mm 7%
PacBio	10-20 kbp	500k-4M	indel+mm 13-15% HiFi: indel 1%+mm 0.1%
PacBio’s REVIO	10-20 kbp	6M+	indel 1%+mm 0.1%



Article | Published: 09 September 2021

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

Jonathan Foox, Scott W. Tighe, […] Christopher E. Mason

Nature Biotechnology **39**, 1129–1140 (2021) | Cite this article

5529 Accesses | **171** Altmetric | Metrics

An Author Correction to this article was published on 11 October 2021

This article has been updated

Abstract

Assessing the reproducibility, accuracy and utility of massively parallel DNA sequencing

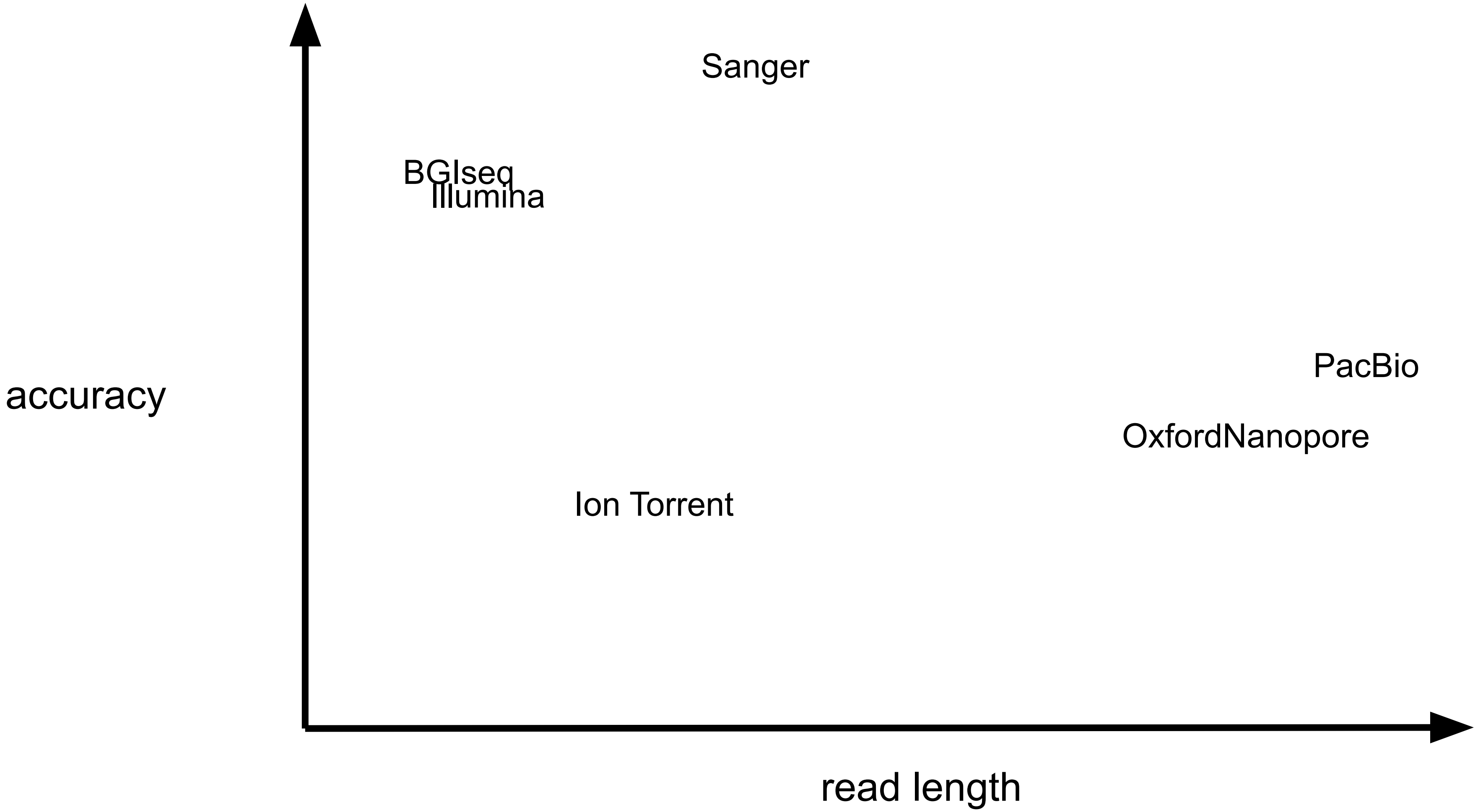
Takeaways:

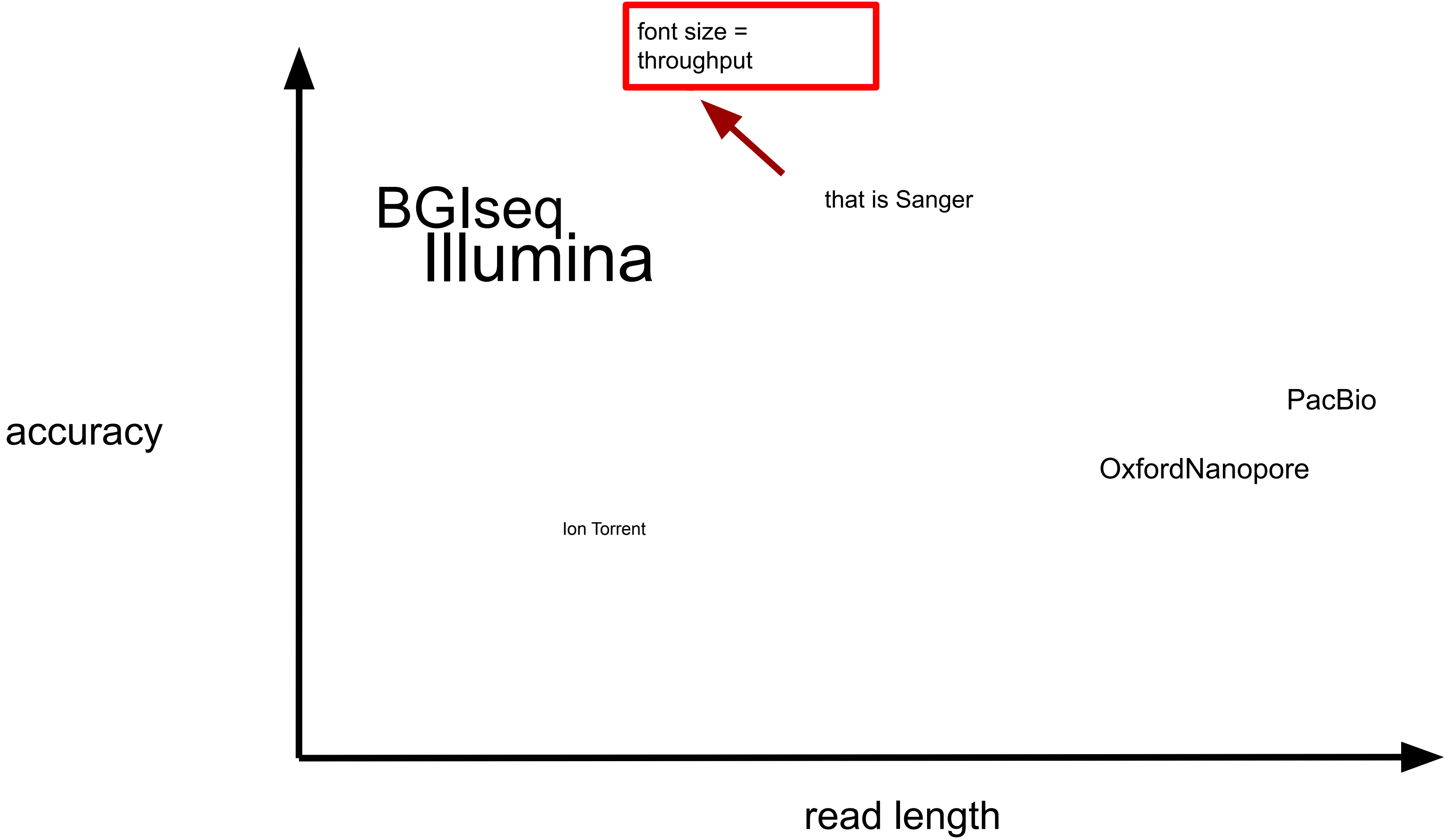
Short reads

- Illumina cheapest
- BGI most accurate

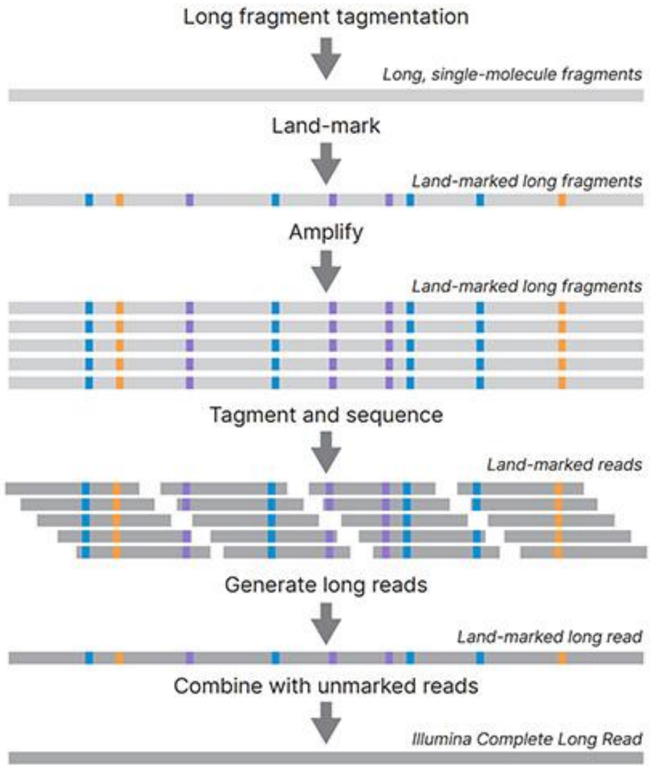
Long reads:

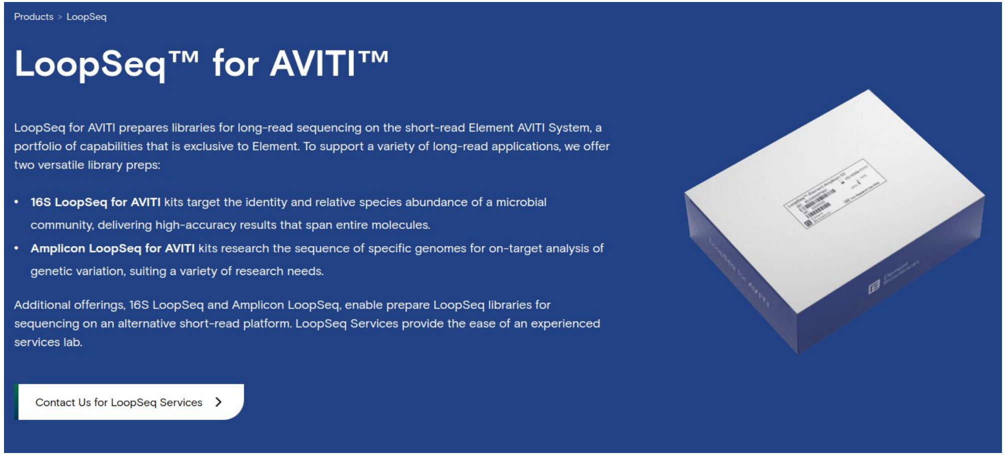
- Most mapping with PacBio
- Oxford/Pacbio good with repeats


Sanger
BGIseq
Illumina
PacBio
accuracy
OxfordNanopore
Ion Torrent
read length


font size =
throughput
that is Sanger
BGIseq
Illumina
accuracy
PacBio
OxfordNanopore
Ion Torrent
read length

New in 2023: Long read and accurate?


Long fragment tagmentation
Long, single-molecule fragments
Land-mark
Land-marked long fragments
Amplify
Land-marked long fragments
Tagment and sequence
Land-marked reads
Generate long reads
Land-marked long read
Combine with unmarked reads
Illumina Complete Long Read

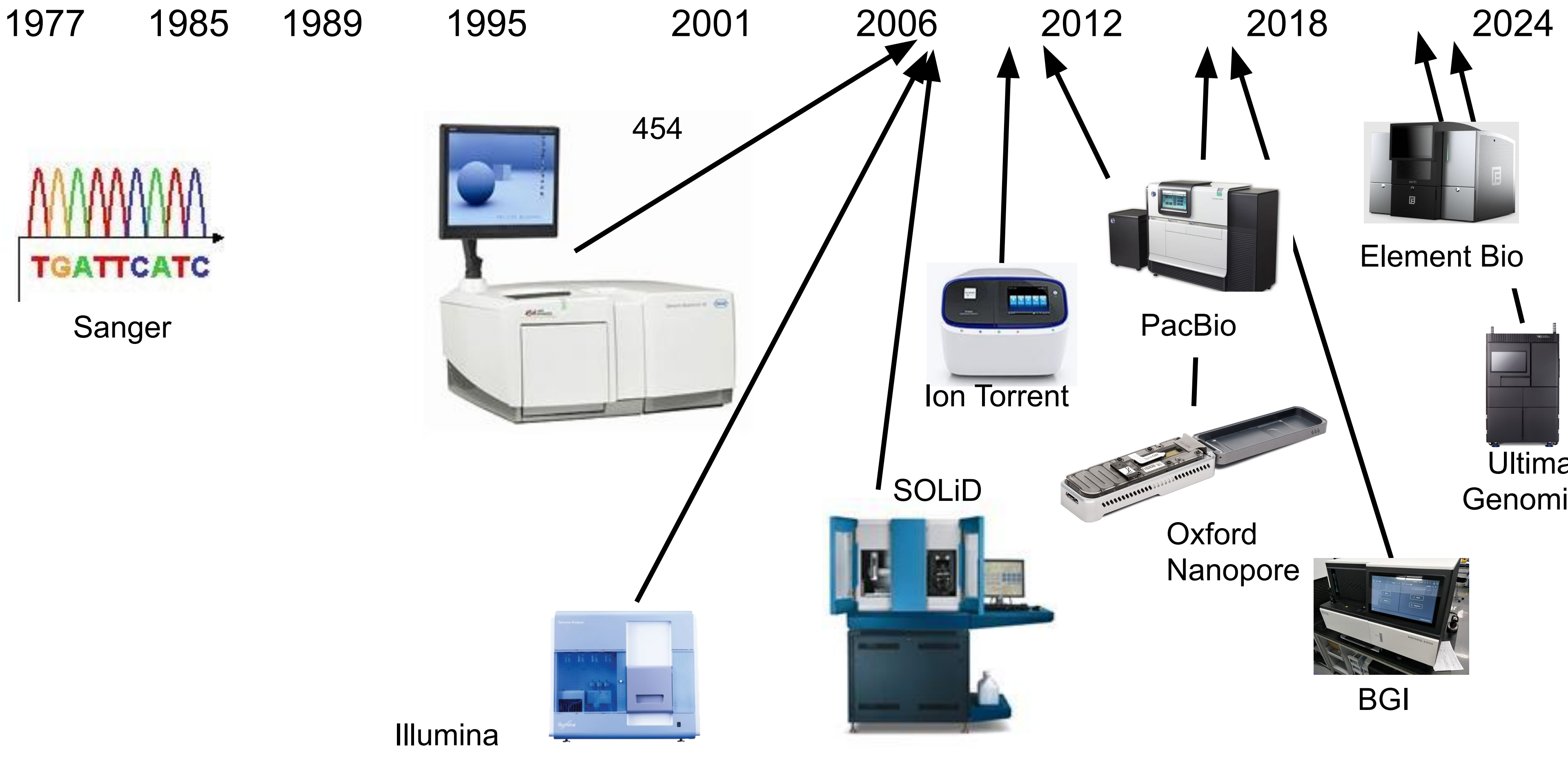

Products > LoopSeq

LoopSeq™ for AVITI™

LoopSeq for AVITI prepares libraries for long-read sequencing on the short-read Element AVITI System, a portfolio of capabilities that is exclusive to Element. To support a variety of long-read applications, we offer two versatile library preps:

- **16S LoopSeq for AVITI** kits target the identity and relative species abundance of a microbial community, delivering high-accuracy results that span entire molecules.
- **Amplicon LoopSeq for AVITI** kits research the sequence of specific genomes for on-target analysis of genetic variation, suiting a variety of research needs.

Additional offerings, 16S LoopSeq and Amplicon LoopSeq, enable prepare LoopSeq libraries for sequencing on an alternative short-read platform. LoopSeq Services provide the ease of an experienced services lab.



1977 1985 1989 1995 2001 2006 2012 2018 2024

Sanger

454

Illumina

SOLiD

Ion Torrent

PacBio

Oxford Nanopore

BGI

Element Bio

Ultima Genomi

1977 1985 1989 1995 2001 2006 2012 2018 2024

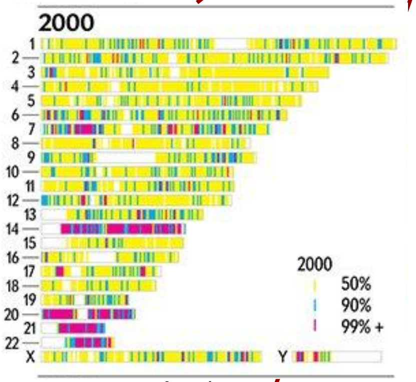
DOE Holds First Human Genome Contractor/Grantee Workshop

Genome Data To Spark Expansion in Biological Research

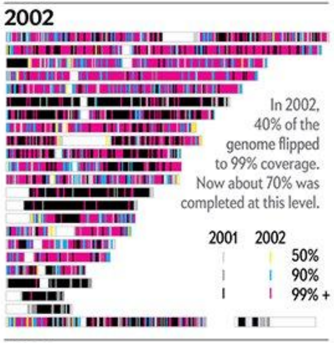
At the first Contractor/Grantee Workshop for the DOE Human Genome Program, Benjamin J. Barnhart, Program Manager, told participants that data produced by the inter-

critically necessar pletion of the genc workshop has led work, including in

1990: Human genome project launched



hg1



In 2002, 40% of the genome flipped to 99% coverage. Now about 70% was completed at this level.

hg12

Images: Martin Krzywinski
Scientific American August 2022

1977 1985 1989 1995 2001 2006 2012 2018 2024

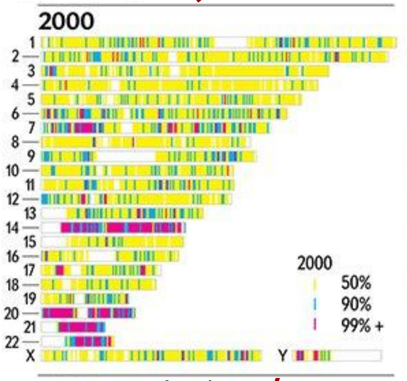
DOE Holds First Human Genome Contractor/Grantee Workshop

Genome Data To Spark Expansion in Biological Research

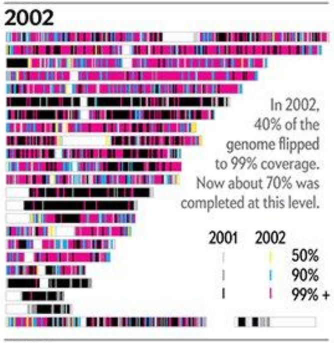
At the first Contractor/Grantee Workshop for the DOE Human Genome Program, Benjamin J. Barnhart, Program Manager, told participants that data produced by the inter-

critically necessar
pletion of the genc
workshop has led
work, including in

1990: Human genome project launched

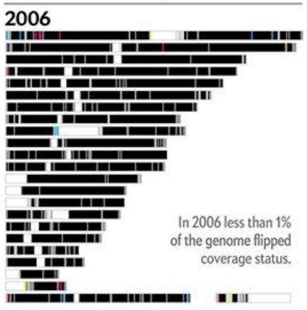


hg1



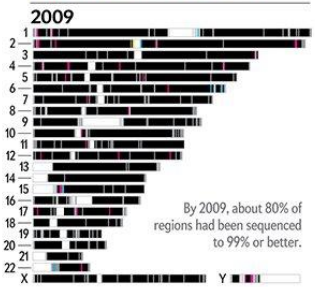
In 2002, 40% of the genome flipped to 99% coverage. Now about 70% was completed at this level.

hg12



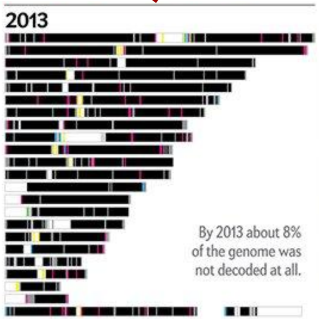
In 2006 less than 1% of the genome flipped coverage status.

hg18



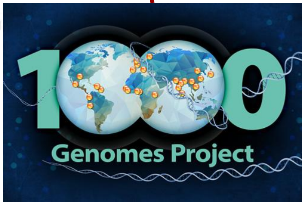
By 2009, about 80% of regions had been sequenced to 99% or better.

hg19



By 2013 about 8% of the genome was not decoded at all.

hg38



2012

Images: Martin Krzywinski
Scientific American August 2022

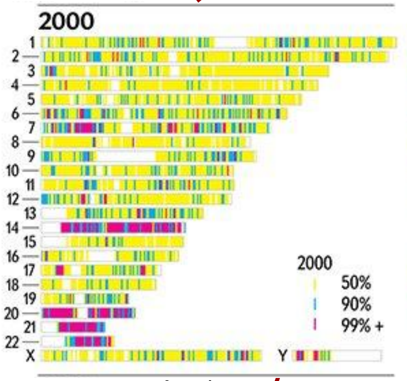
1977 1985 1989 1995 2001 2006 2012 2018 2024

DOE Holds First Human Genome Contractor/Grantee Workshop

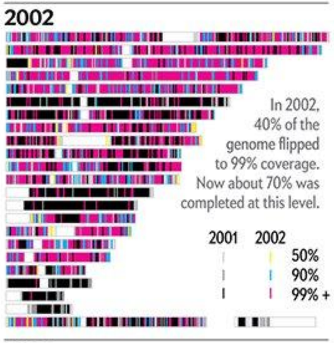
Genome Data To Spark Expansion in Biological Research

At the first Contractor/Grantee Workshop for the DOE Human Genome Program, Benjamin J. Barnhart, Program Manager, told participants that data produced by the inter- critically necessar pletion of the genc workshop has led work, including in

1990: Human genome project launched

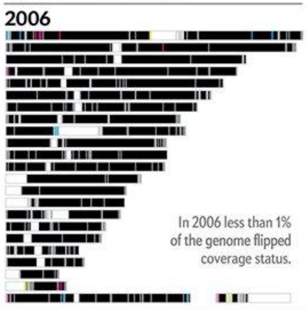


hg1



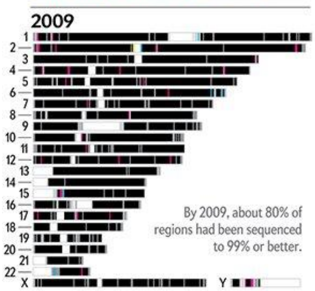
In 2002, 40% of the genome flipped to 99% coverage. Now about 70% was completed at this level.

hg12



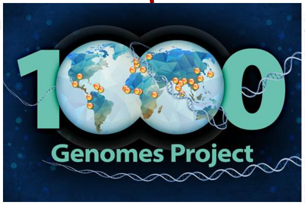
In 2006 less than 1% of the genome flipped coverage status.

hg18

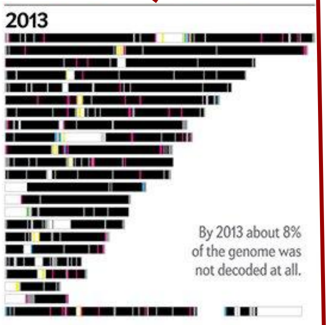


By 2009, about 80% of regions had been sequenced to 99% or better.

hg19

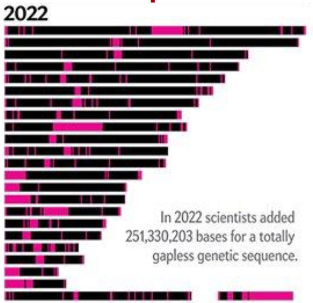


2012



By 2013 about 8% of the genome was not decoded at all.

hg38



In 2022 scientists added 251,330,203 bases for a totally gapless genetic sequence.

CHM13v2

Images: Martin Krzywinski
Scientific American August 2022

Summary

- I did not mention a very important factor: **cost**
- I did not mention another important factor: **runtime**

https://twitter.com/AlbertVilella

Albert Vilella
@AlbertVilella

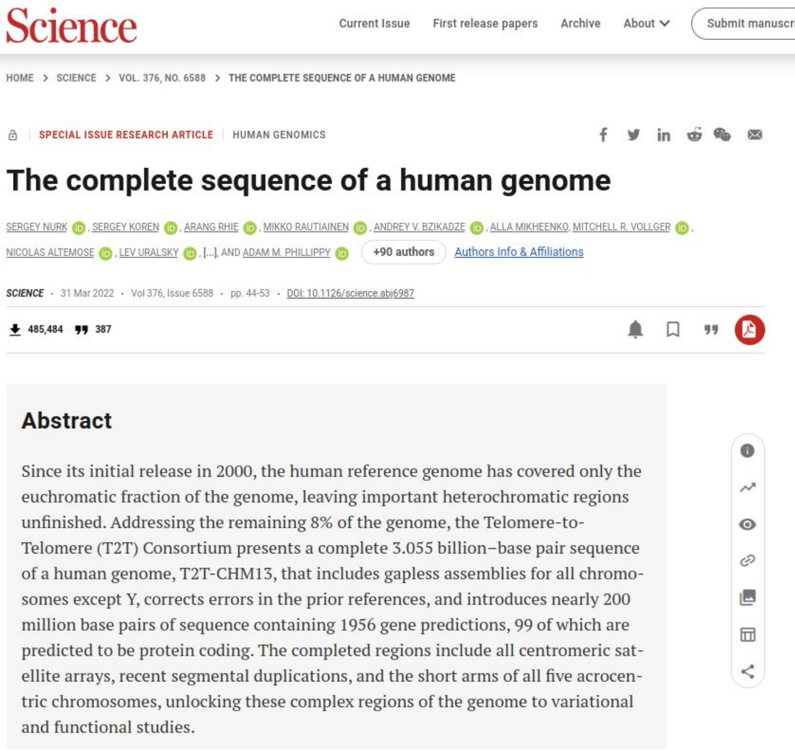
Experienced Bioinformatics Scientist, Next-Generation Sequencing, Single Cell, Spatial Biology, Liquid Biopsy, Epigenomics, Synthetic Biology.

Science & Technology · Cambridge, England · linktr.ee/albertvilella
Joined July 2012

26 Following **19.4K** Followers

Summary

- Each tech has advantages, pick the most appropriate for your question
- Illumina is the current workhorse
  - Great for many applications
- Long read technology
  - Adding information
  - Resolves difficult regions during genome assembly



HOME > SCIENCE > VOL. 376, NO. 6588 > THE COMPLETE SEQUENCE OF A HUMAN GENOME

SPECIAL ISSUE RESEARCH ARTICLE | HUMAN GENOMICS

The complete sequence of a human genome

SERGEY NURK, SERGEY KOREN, ARANG RHIE, MIKKO RAUTIAINEN, ANDREY V. BZIKADZE, ALLA MIKHEENKO, MITCHELL R. VOLLGER, NICOLAS ALTEMOSE, LEV URALSKY, [...], AND ADAM M. PHILLIPPY +90 authors Authors Info & Affiliations

SCIENCE • 31 Mar 2022 • Vol 376, Issue 6588 • pp. 44-53 • DOI: 10.1126/science.abj6987

485,484 387

Abstract

Since its initial release in 2000, the human reference genome has covered only the euchromatic fraction of the genome, leaving important heterochromatic regions unfinished. Addressing the remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium presents a complete 3.055 billion–base pair sequence of a human genome, T2T-CHM13, that includes gapless assemblies for all chromosomes except Y, corrects errors in the prior references, and introduces nearly 200 million base pairs of sequence containing 1956 gene predictions, 99 of which are predicted to be protein coding. The completed regions include all centromeric satellite arrays, recent segmental duplications, and the short arms of all five acrocentric chromosomes, unlocking these complex regions of the genome to variational and functional studies.