
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
```

Second generation sequencing

- Illumina sits on 75% of the market

Illumina

Element Bio

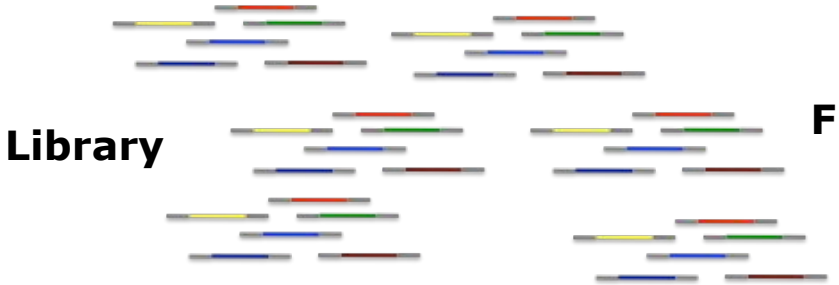
454

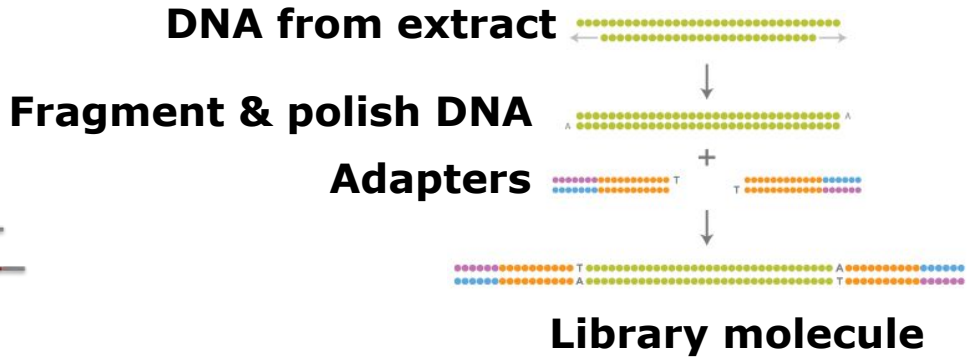
BGISEQ

Ion Torrent

General library preparation steps

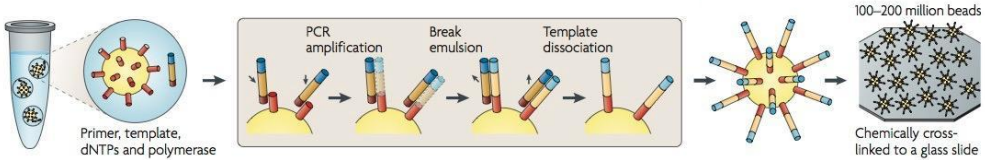
1. Create library molecules
2. Amplification (PCR)
3. Massive parallel sequencing (strength over Sanger)





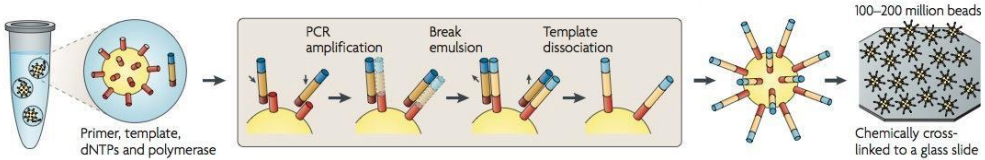
Amplification and immobilization

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

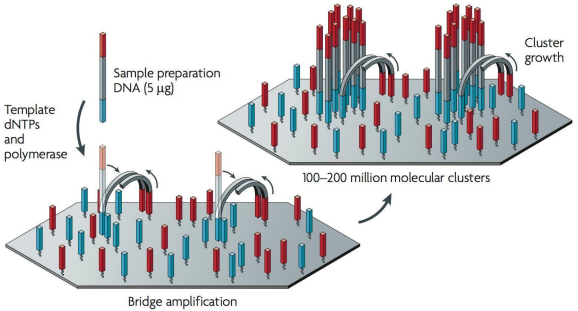


Amplification and immobilization

- 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

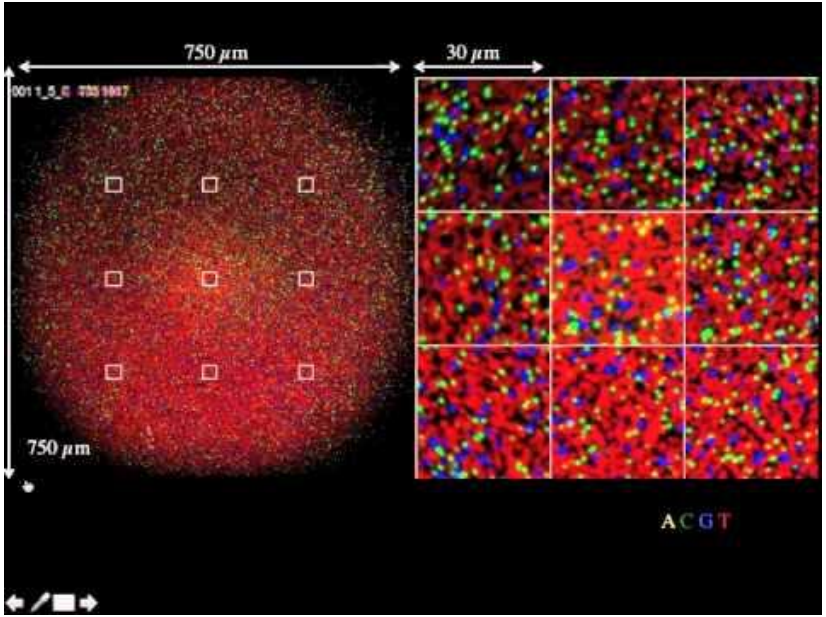
```
1:1A  2:2G  3:1T        =   AGGT
```

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)?



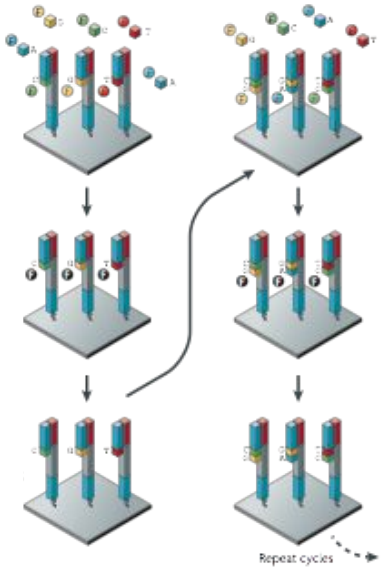
Fluorescence detection

Illumina - Cyclic reversible termination

Add all dNTPs labelled w. diff dye

Create four-color image

Cleave dye and repeat next cycle



2G: Imaging handout

Illumina 1:_____

Illumina 2:_____

2G: Imaging handout Answers!

Illumina 1:_____

Illumina 2:_____

TOP: **CATCGT**

BOTTOM: **CCCCCC**

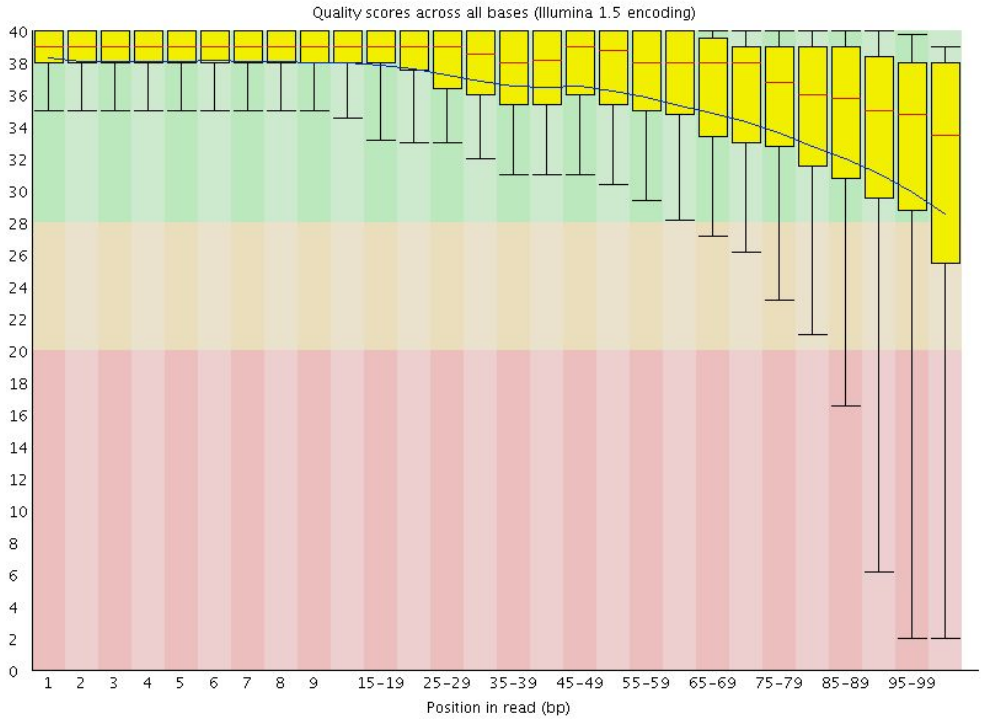
___

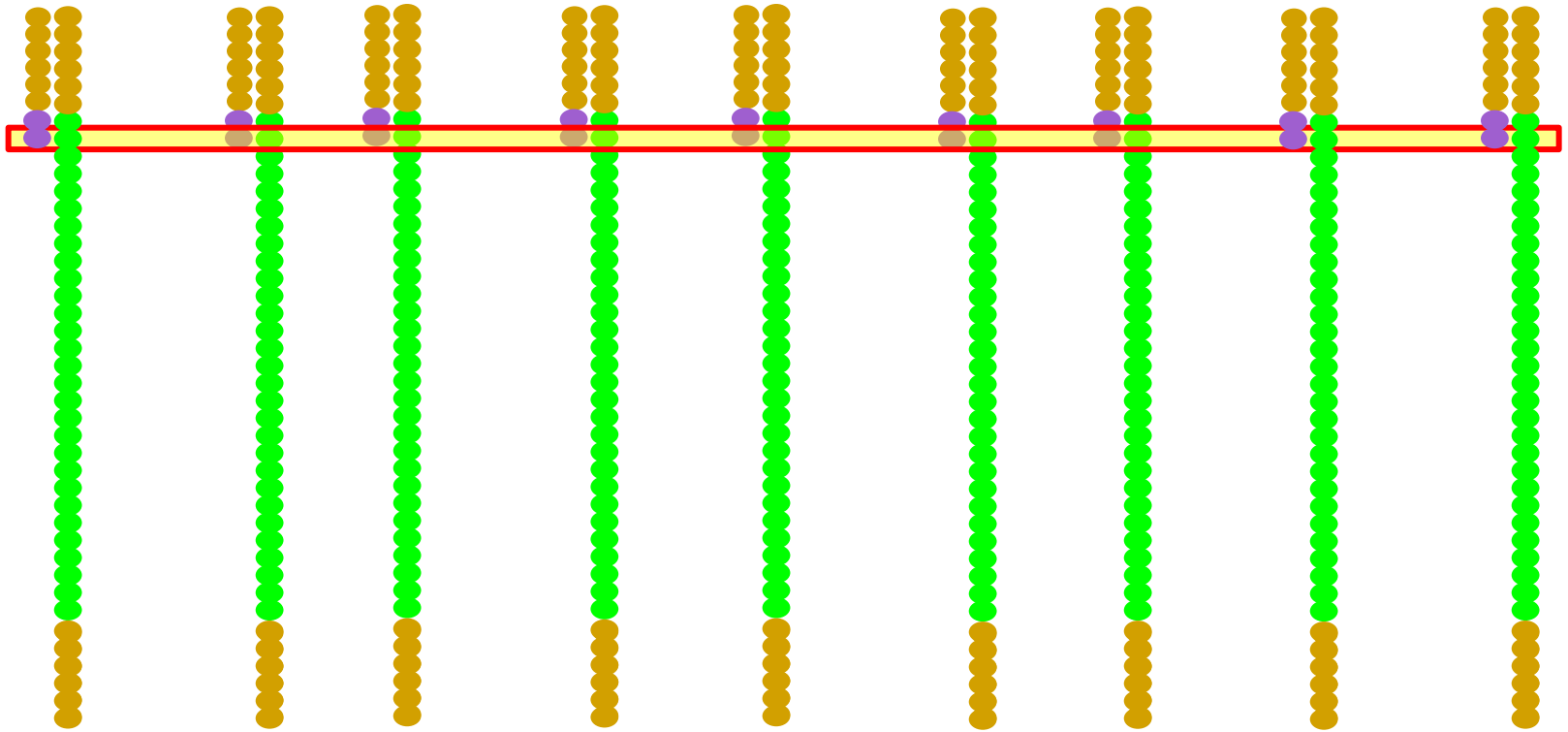
__

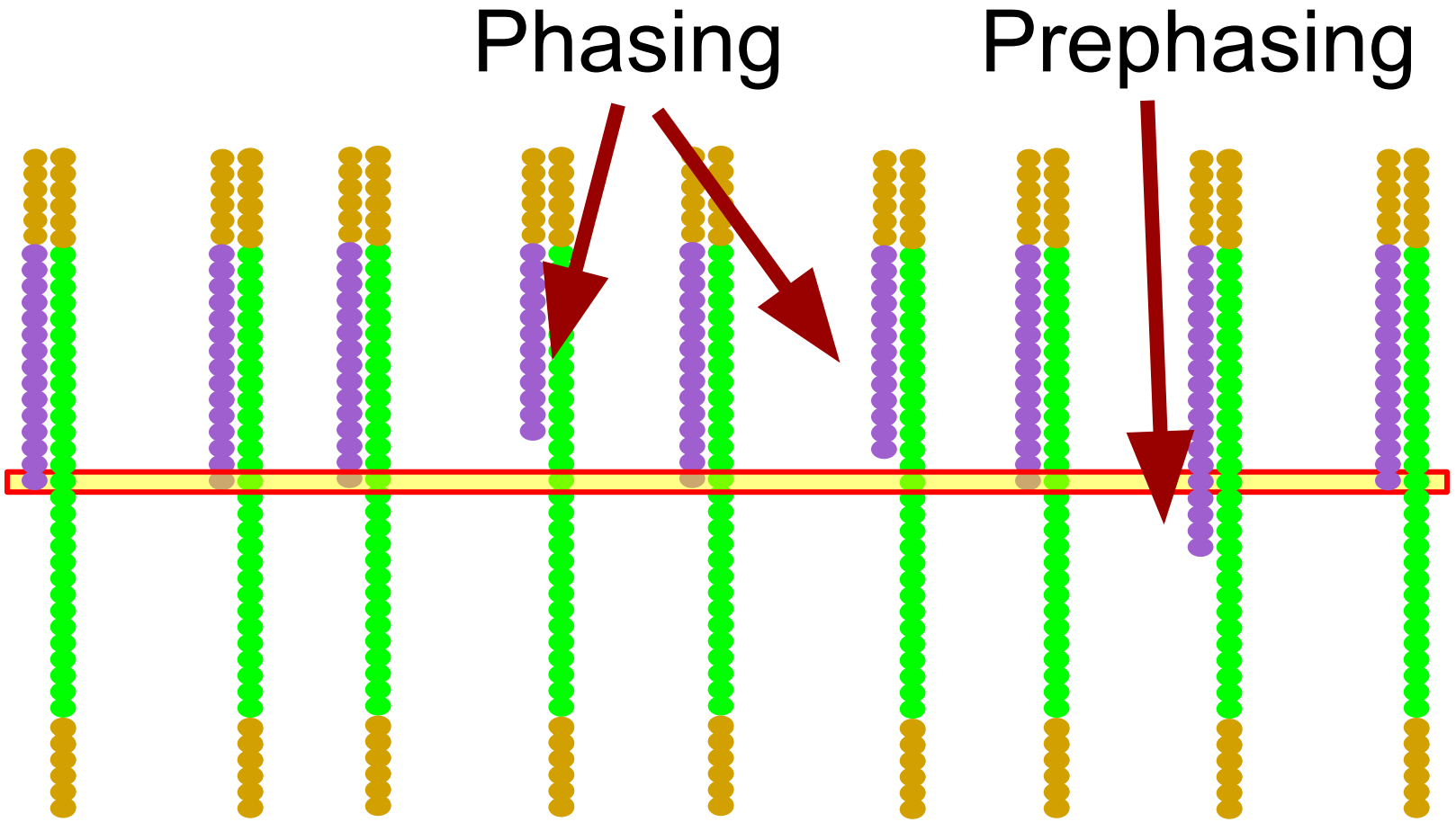
__

Illumina: Quality deterioration

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

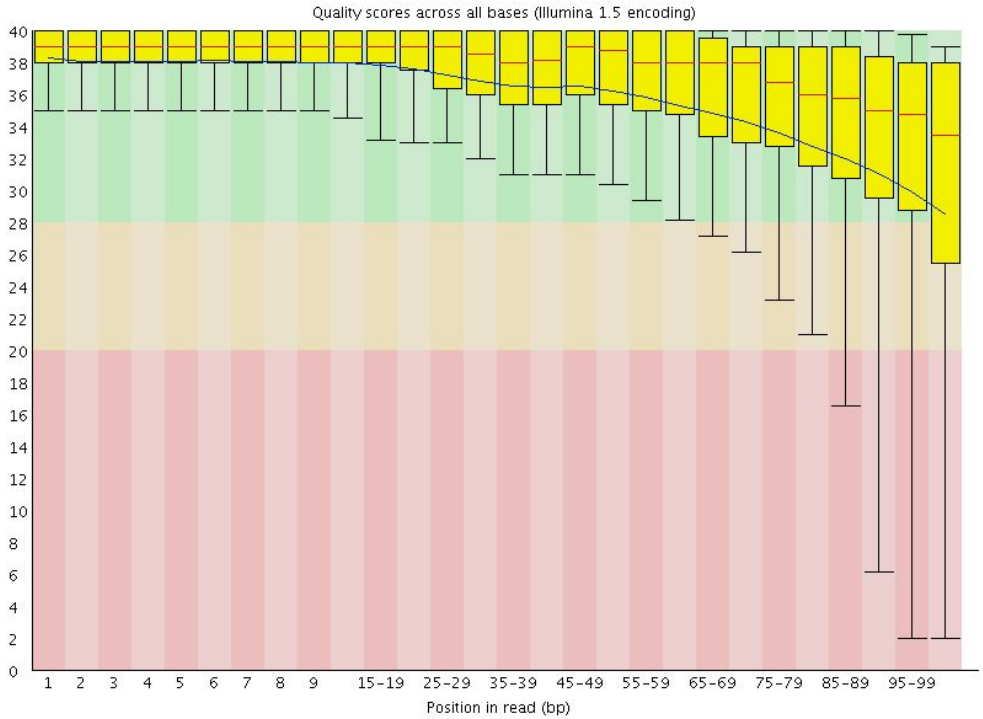





Phasing
Prephasing

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

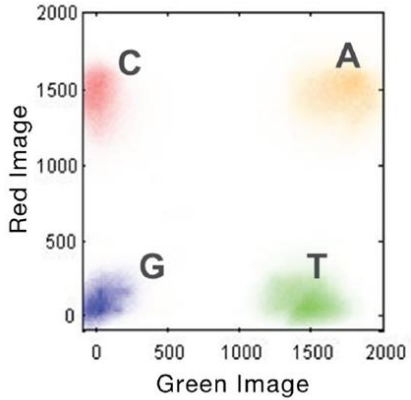
Brief side note about spike-in

- 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-)

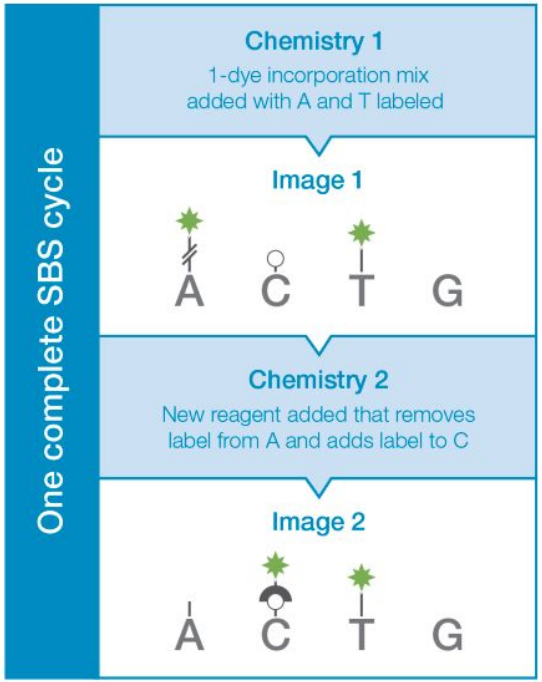
- Chemistry is not based 4 dyes (as before) but 2 dyes
  - T (red), C (green), A (both) and G (none = “dark”)
  - Faster processing rate and cheaper reagents
  - Slightly increases error rate
  - Problem with G stretches because G is not dyed



source: Illumina

1 dye, 2 images

A.



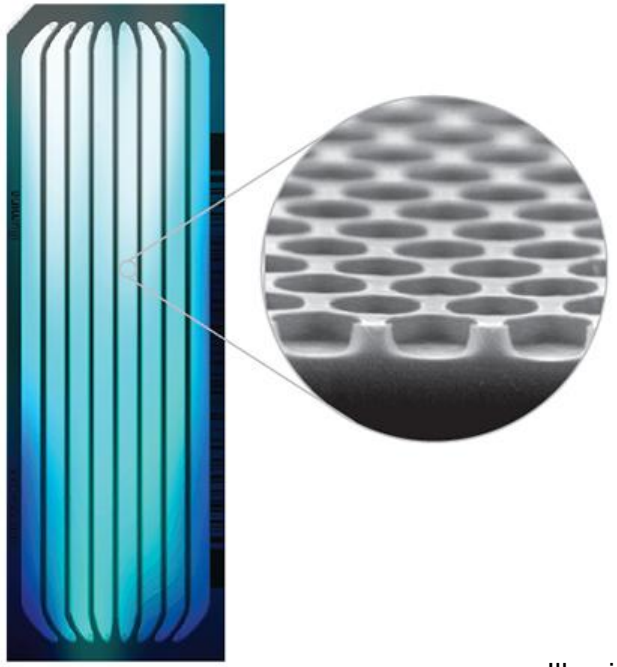
B.

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

source: Illumina

Patterned flowcell

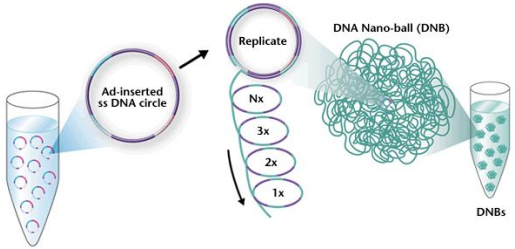
- Patterned wells
- Solves overloading flowcell
- More duplicates



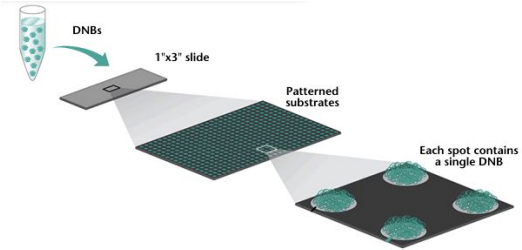
source: Illumina

BGI-Seq

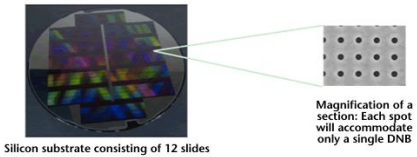
ssDNA -> DNA nanoballs



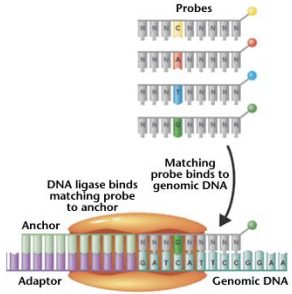
Place DNBs into each spot



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 four 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

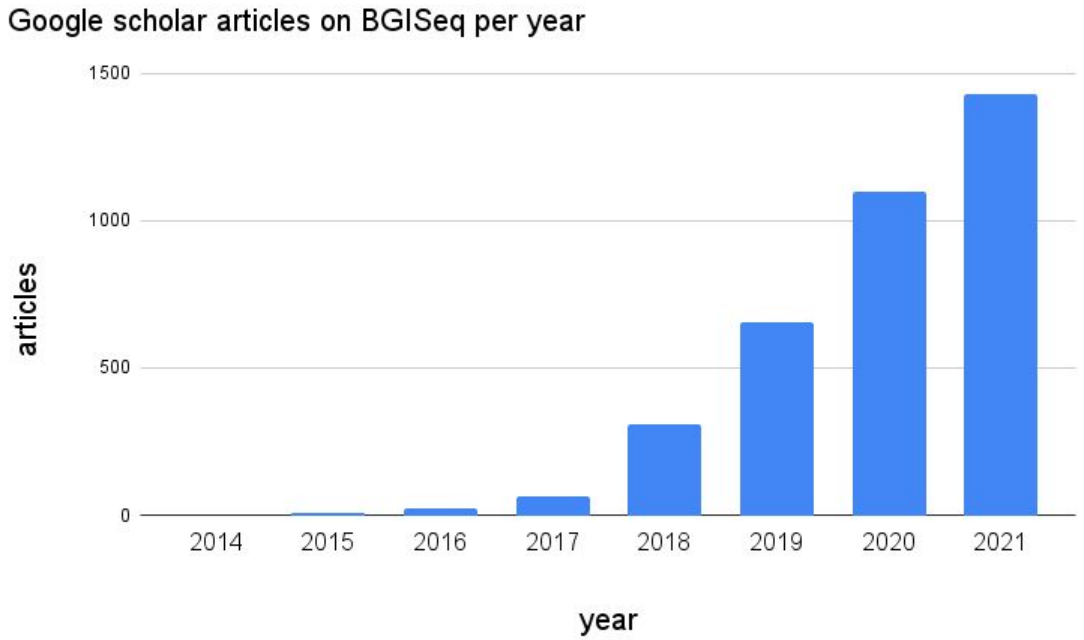
Comparative analysis of 7 short-read sequencing platforms using the Korean Reference Genome: MGI and Illumina sequencing benchmark for whole-genome sequencing

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 per year
1500
1000
500
0
articles
2014
2015
2016
2017
2018
2019
2020
2021
year

Avidity sequencing

Element Biosciences

New Results

Follow this preprint

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

doi: https://doi.org/10.1101/2022.12.07.519512

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

2 main types of approaches

1) Amplify and sequence one base at a time

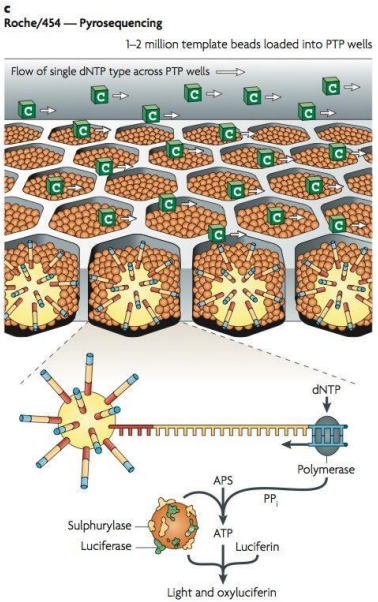
```
1:A   2:G   3:G   4:T   =   AGGT
```

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



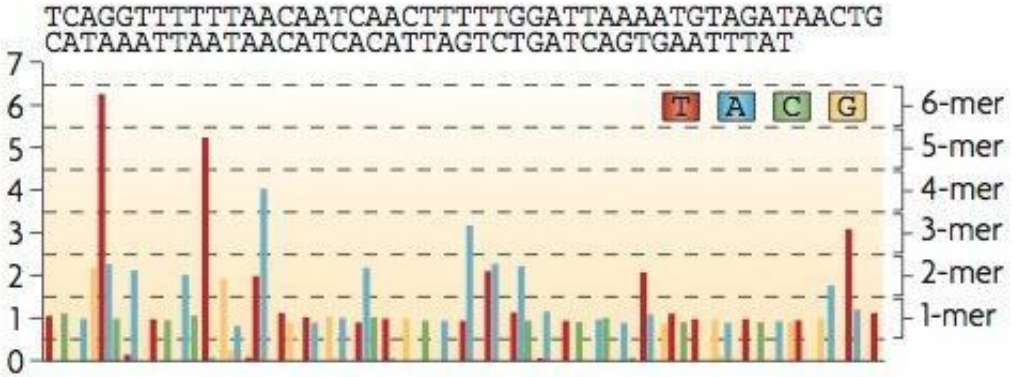
2G: Imaging handout

454: _____

_____

_____

2G: Imaging handout Answers!



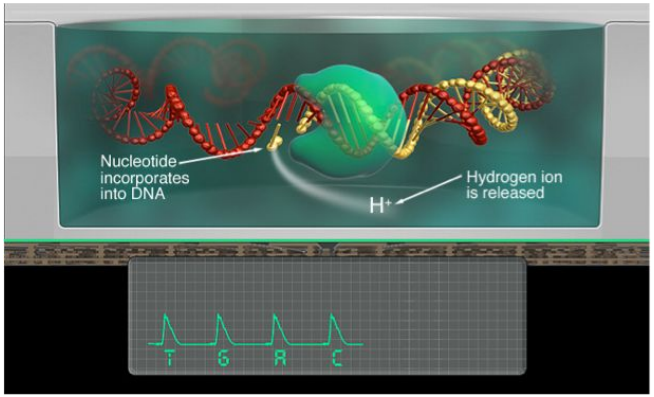
454: _____

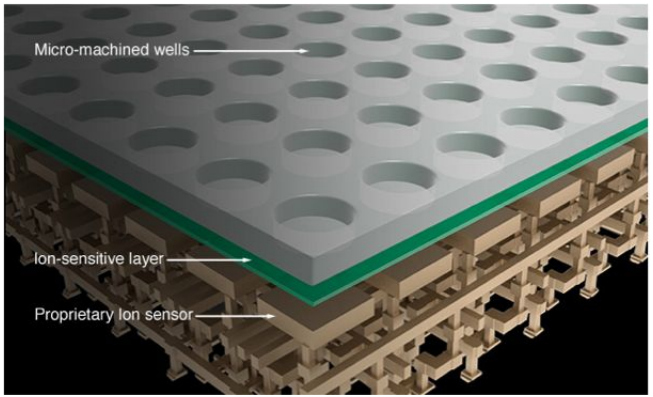
_____

_____

Ion Torrent

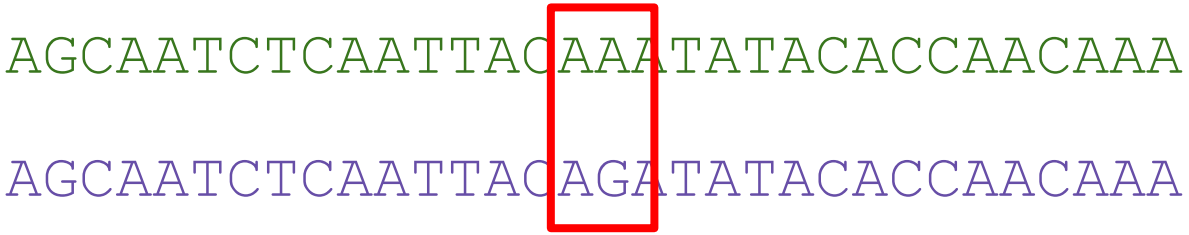
- Corporate propaganda: https://www.youtube.com/watch?v=zBPKj0mMcDg
- Similar principle to 454
- Library: Emulsion PCR
- Based on semiconductors
- Detection is based on H ions (pH) changes



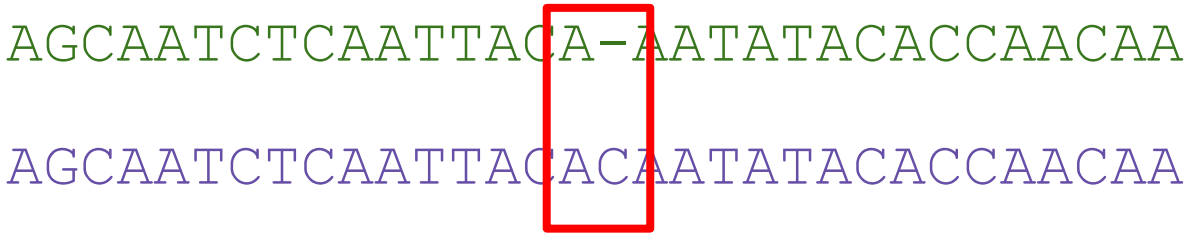


Let’s remember the types of errors

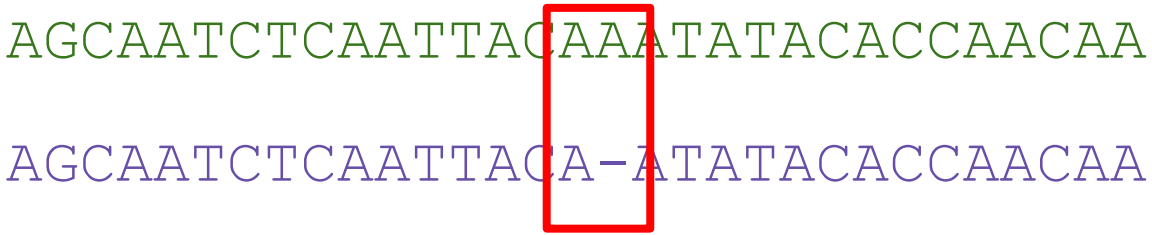
mismatch



insert



deletion



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
```

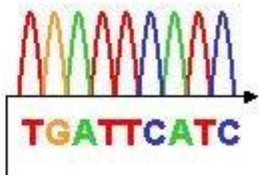
2) Amplify and count how many of the same base

```
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%
Ion Torrent	~200-400 bp	5-150M reads	indel 0.46 to 2.4%

3rd generation

1977 1980 1983 1986 1989 1992 1995 1998 2001 2004 2007 2010 2013 2016 2019



Sanger

454

PacBio

Ion Torrent

SOLiD

BGI

Oxford Nanopore

Illumina

3rd generation

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

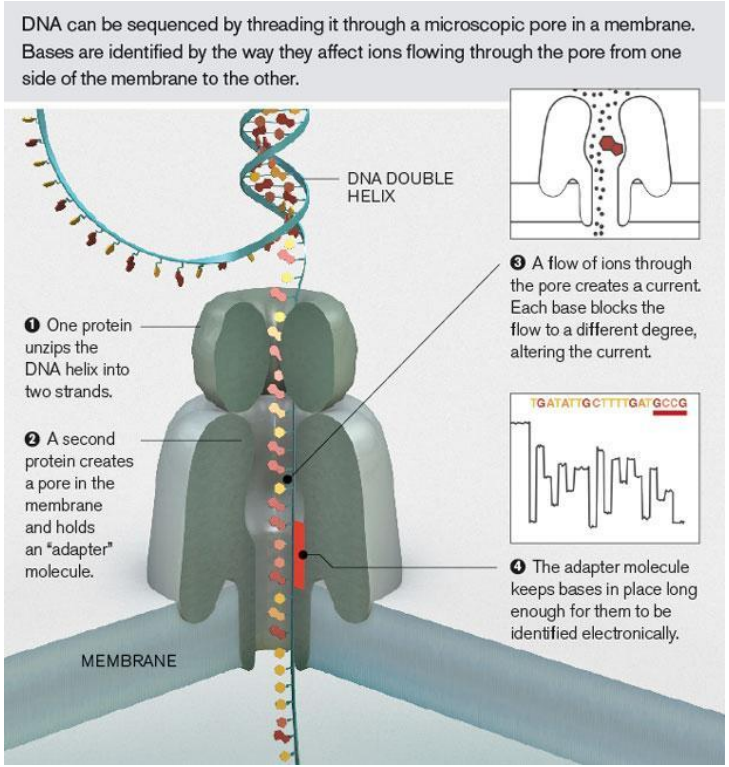
Helicos

PacBio

Oxford Nanopore

Oxford Nanopore

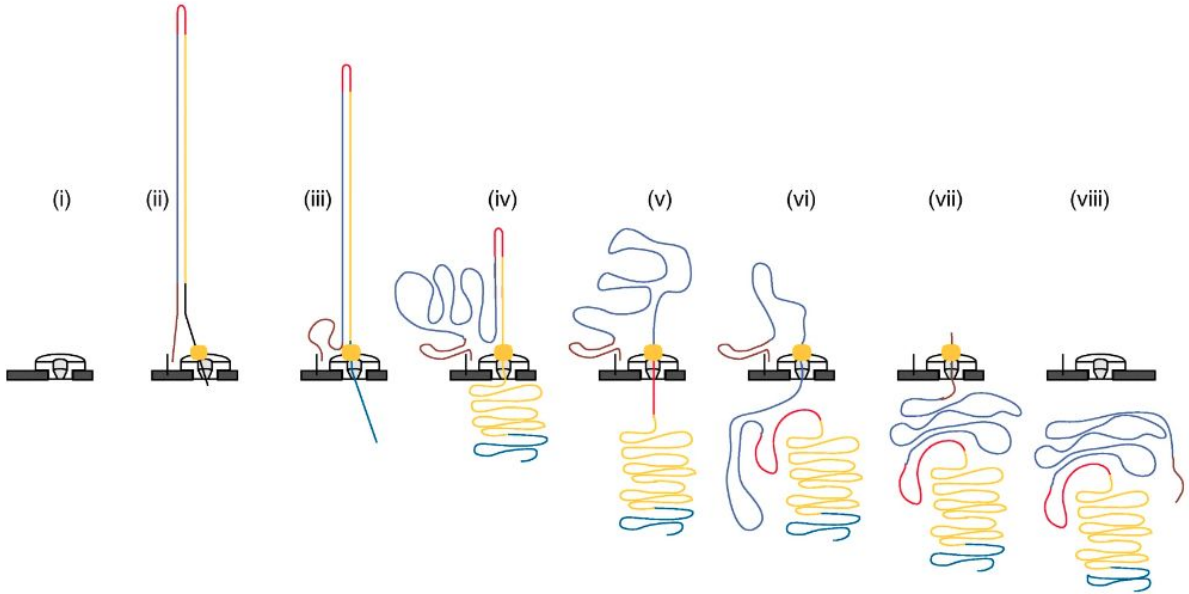
- Literal nanopores
- Current per base
- Non-random errors
- https://www.youtube.com/watch?v=RcP85JHLmnI
- Very high error rate



"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

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

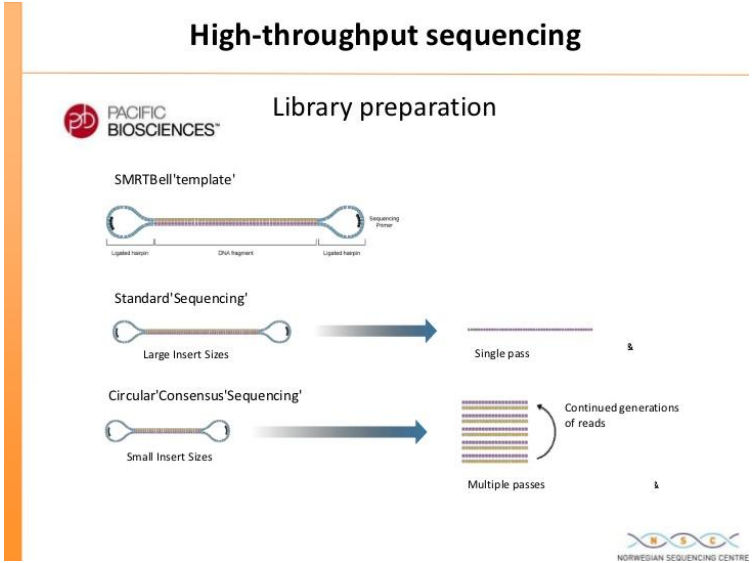
- Long reads, low quality
- Low establishment and maintenance costs
- Portability

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

- Expensive machinery
- Not very portable

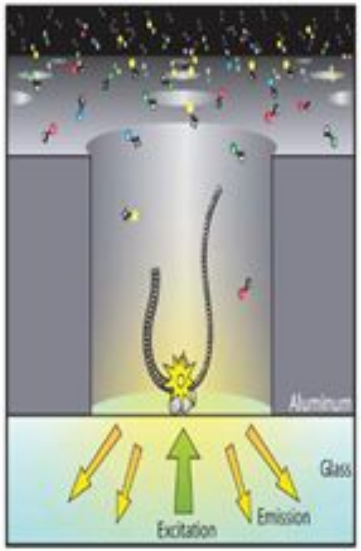
PacBio

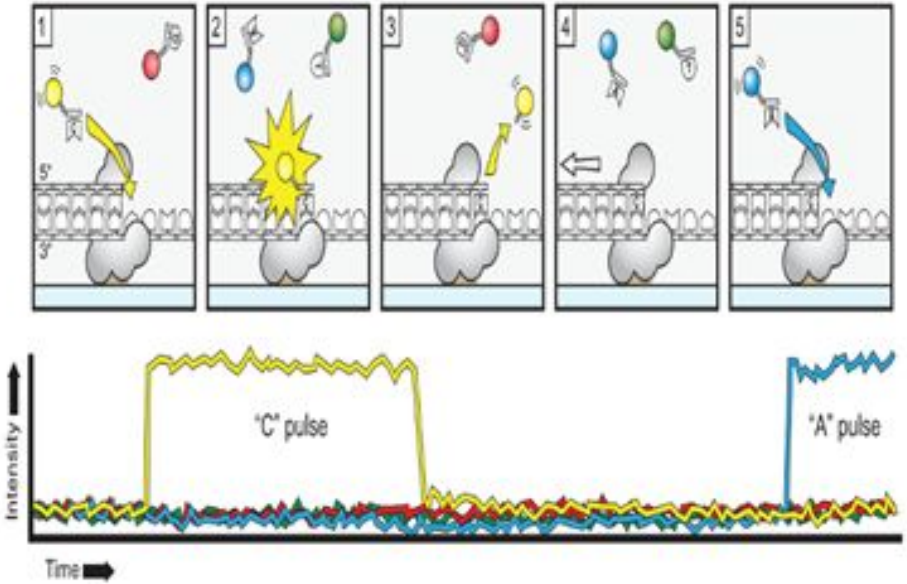
- Flexibility
  - Long but low quality or shorter but better reads
  - Robust
  - https://www.youtube.com/watch?v=_ID8JyAbwEo
  - New 2019: HiFi read same fragment multiple times



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: ~0.7% 1D: indel+mm 20% 2D: indel+mm 7%
PacBio	10-20 kbp	500k-4M	indel+mm 13-15% HiFi: 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

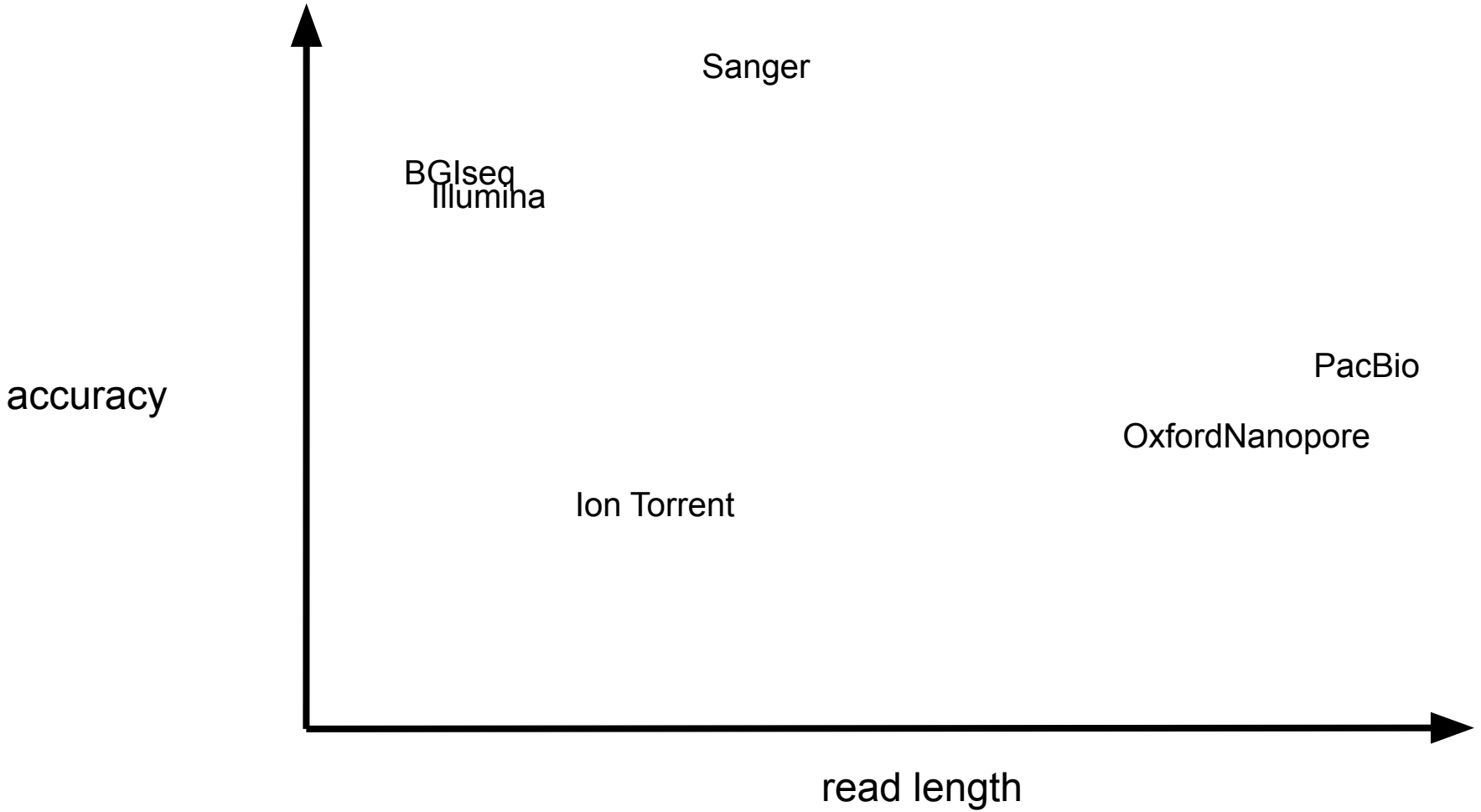
Takeaways:

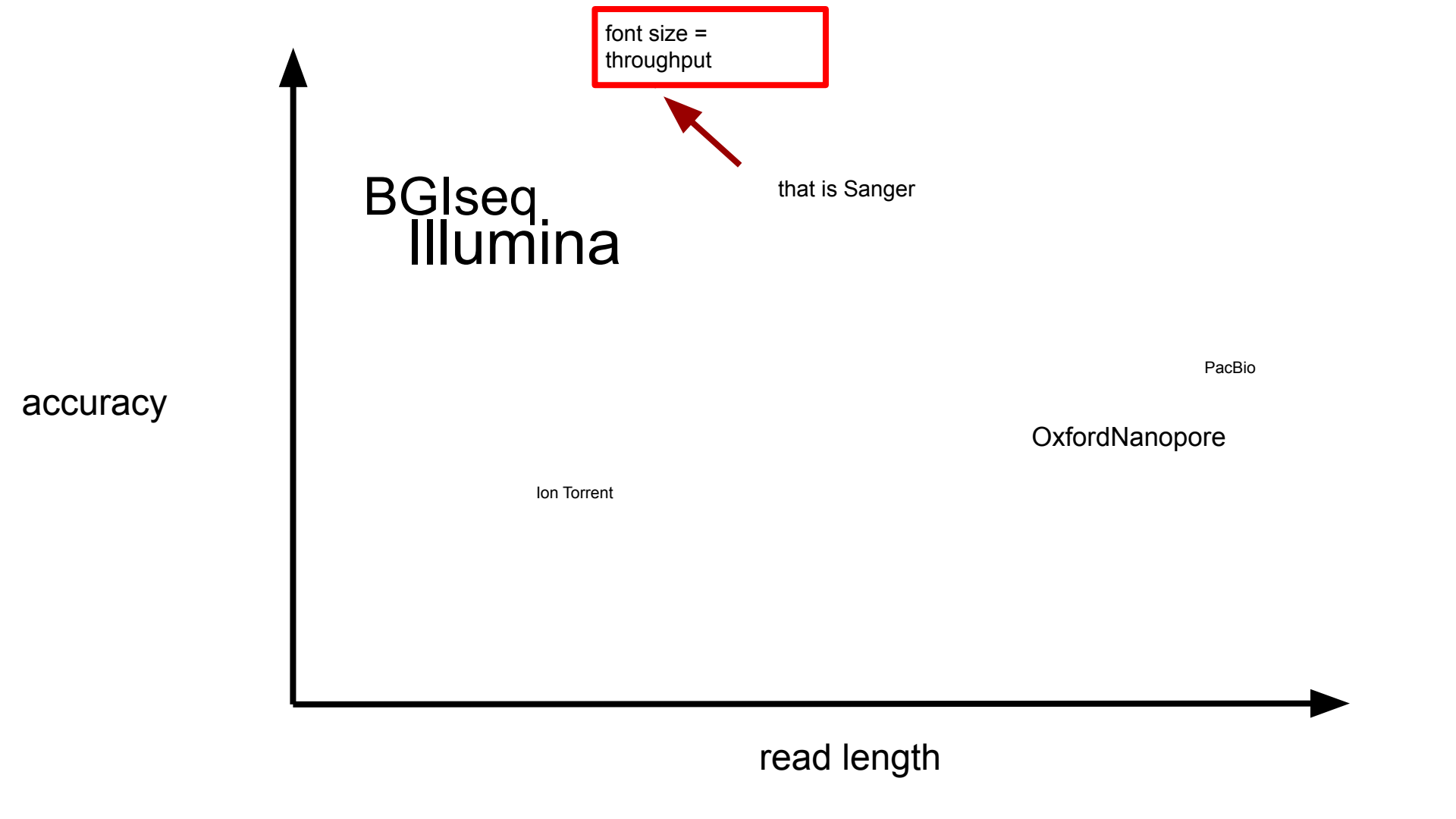
Short reads

- Illumina cheapest
- BGI most accurate

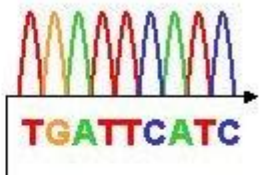
Long reads:

- 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
PacBio
accuracy
OxfordNanopore
Ion Torrent
read length

1977 1980 1983 1986 1989 1992 1995 1998 2001 2004 2007 2010 2013 2016 2022



Sanger

454

PacBio

Element Bio

Ultima Genomics

Ion Torrent

SOLiD

Oxford Nanopore

BGI

Illumina

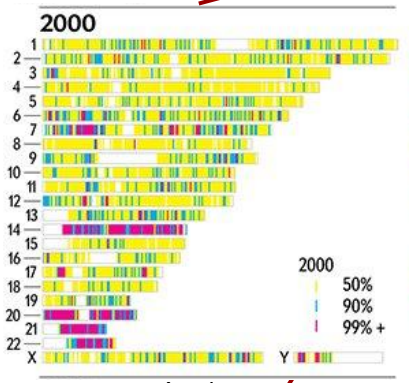
1977 1980 1983 1986 1989 1992 1995 1998 2001 2004 2007 2010 2013 2016 2022

DOE Holds First Human Genome Contractor/Grantee Workshop

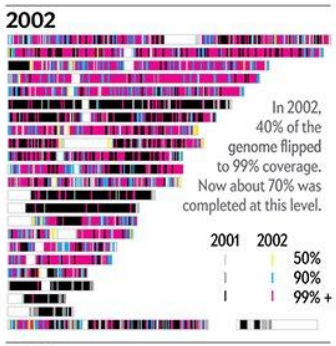
Genome Data To Spark Expansion in Biological Research

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



hg12

Images: Martin Krzywinski
Scientific American August 2022

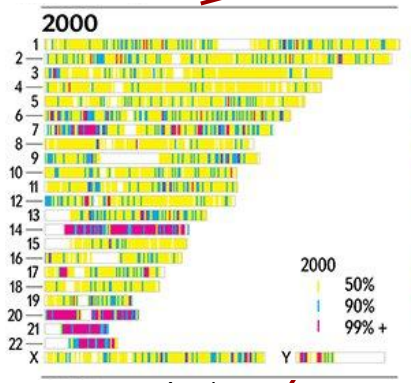
1977 1980 1983 1986 1989 1992 1995 1998 2001 2004 2007 2010 2013 2016 2022

DOE Holds First Human Genome Contractor/Grantee Workshop

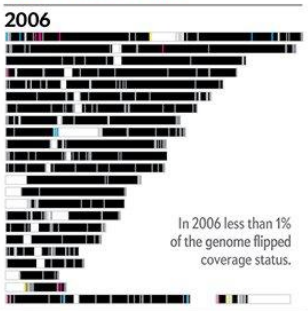
Genome Data To Spark Expansion in Biological Research

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

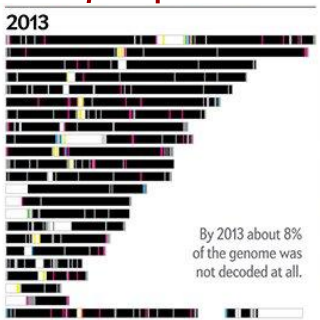
1990: Human genome project launched



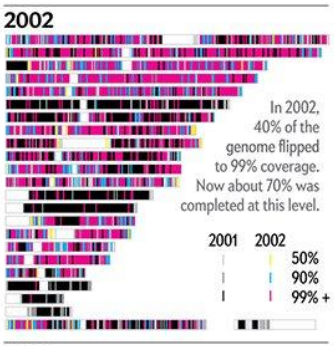
hg1



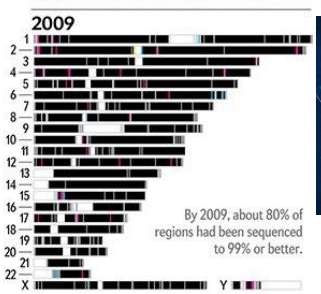
hg18



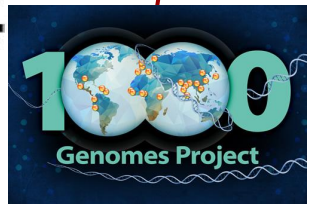
hg38



hg12



hg19



2012

Images: Martin Krzywinski
Scientific American August 2022

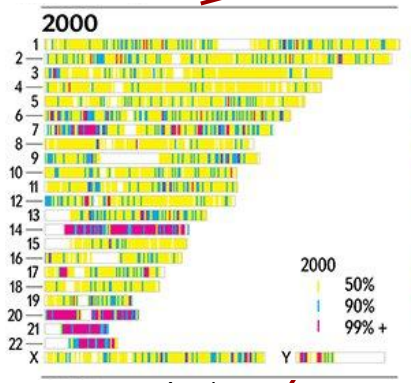
1977 1980 1983 1986 1989 1992 1995 1998 2001 2004 2007 2010 2013 2016 2022

DOE Holds First Human Genome Contractor/Grantee Workshop

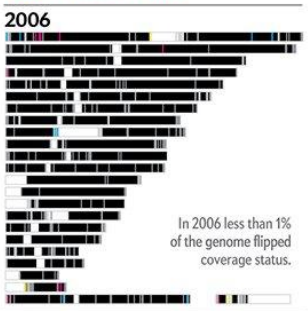
Genome Data To Spark Expansion in Biological Research

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

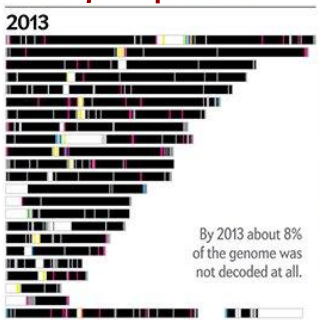
1990: Human genome project launched



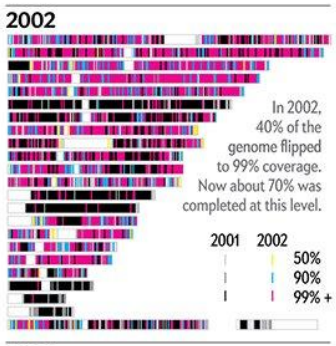
hg1



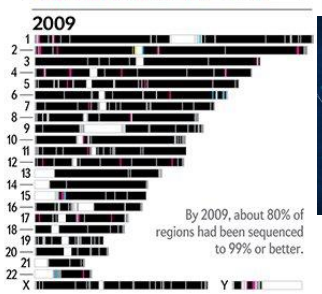
hg18



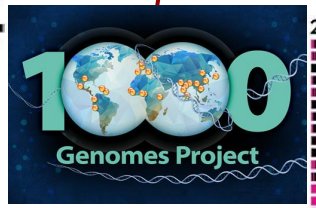
hg38



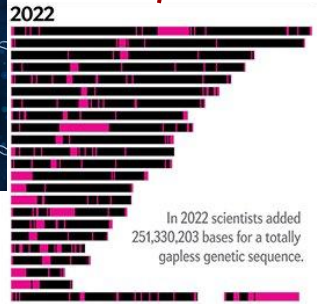
hg12



hg19



2012



CHM13v2

Images: Martin Krzywinski
Scientific American August 2022

Summary

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

Article | Open Access | Published: 14 July 2020

Telomere-to-telomere assembly of a complete human X chromosome

Karen H. Miga, Sergey Koren, Arang Rhie, Mitchell R. Vollger, Ariel Gershman, Andrey Bzikadze, Shelise Brooks, Edmund Howe, David Porubsky, Glennis A. Logsdon, Valerie A. Schneider, Tamara Potapova, Jonathan Wood, William Chow, Joel Armstrong, Jeanne Fredrickson, Evgenia Pak, Kristof Tigyi, Milinn Kremitzki, Christopher Markovic, Valerie Maduro, Amalia Dutra, Gerard G. Bouffard, Alexander M. Chang, Nancy F. Hansen, Amy B. Wilfert, Françoise Thibaud-Nissen, Anthony D. Schmitt, Jon-Matthew Belton, Siddarth Selvaraj, Megan Y. Dennis, Daniela C. Soto, Ruta Sahasrabudhe, Gulhan Kaya, Josh Quick, Nicholas J. Loman, Nadine Holmes, Matthew Loose, Urvashi Surti, Rosa ana Risques, Tina A. Graves
Lindsay, Robert Fulton, Ira Hall, Benedict Paten, Kerstin Howe, Winston Timp, Alice Young, James C.