3rd generation sequencing technologies

Doc. dr. sc. Krešimir Križanović

University of Zagreb

Faculty of Electrical Engineering and Computing

Laboratory for Bioinformatics and Computational Biology (LBCB)

Contents

- History of sequencing
  - First generation technologies
  - Second generation technologies - NGS
  - New technologies – third generation?
- Pacific Biosciences
- Oxford Nanopore Technologies
- What is on the horizon?

History of sequencing

- First generation sequencing
  - 1973 – Walter Gilbert and Allan Maxam: "DNA sequencing by chemical degradation"
  - 1977 – Frederic Sanger: "DNA sequencing with chain-terminating inhibitors"
- From the computer scientists point of view
  - Read length: from 500-600 to 800-1000 base pairs
  - Error rate: 0,1% - accepted standard
  - Higher error rate at the beginning of each read
- From the applicability point of view
  - Speed: slow!
  - Cost: $2.400 for 1M nucleotide

Human Genome Project

- Officially started in 1990.
- Planned duration – 15 years
- Announced complete in 2003. – 2 years ahead of schedule
  - In May 2006. the sequence of the last chromosome was published in Nature
- Advances in sequencing technology enabled earlier project completion
  - Applied Biosystems – ABI PRISM, technology based on Sanger sequencing, parallel sequencing of a large number of samples
- Estimated project cost – $3 billion ($5 billion adjusting for inflation)
- Combined sequence of several individuals
  - NOT A PERSONAL GENOME

Next Generation Sequencing - NGS

- *Pyrosequencing*
  - Pyrosequencing AB (1999) -> Biotage (2003) -> Qiagen (2008)
  - 454 Life Sciences -> Roche (2007) -> closed 2013
- *Ion semiconductor sequencing*
  - Ion Torrent Systems (2010)
- *Illumina dye sequencing*
  - Solexa (1998) -> Illumina (2007)
  - **The largest fish in the pond!**

- Short reads
- Massively parallel sequencing

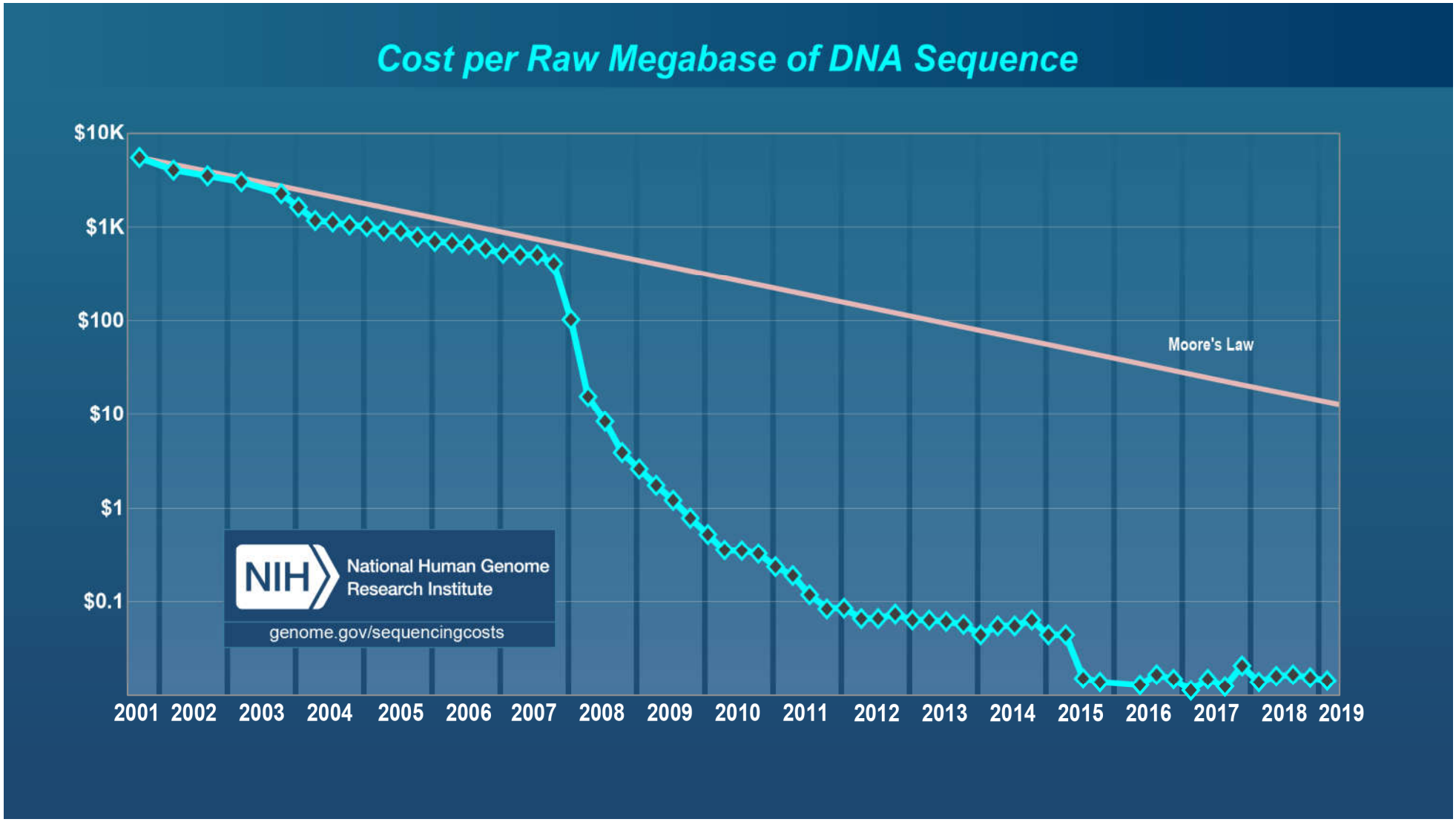
Next Generation Sequencing - NGS

- From the computer science and applicability point of view:

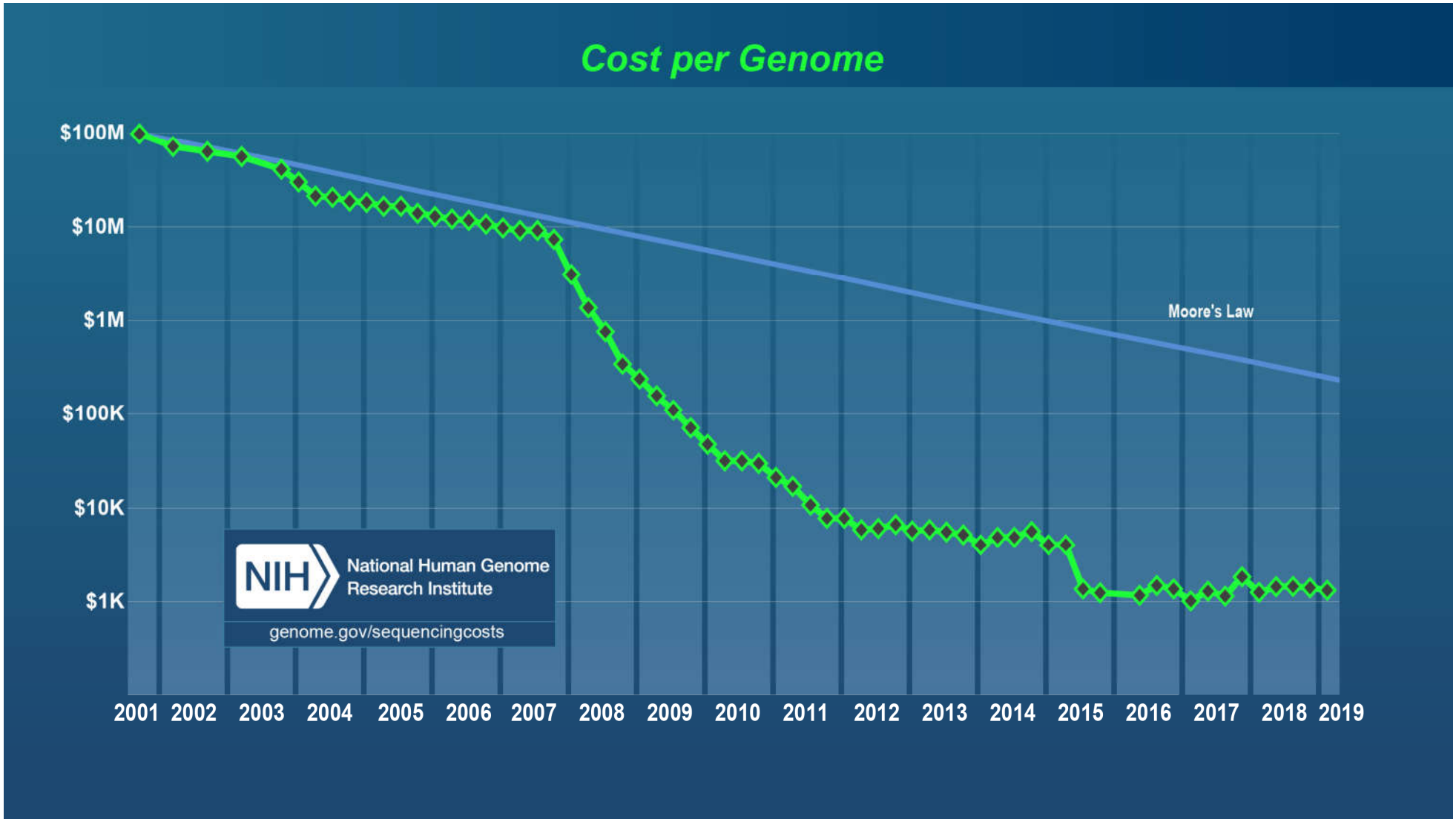
Technology	Read length	Accuracy	Data throughput	Cost for 1M nucleotides	Error types
Ion Torrent	to 600 bp	99,6%	up to 50 Gbp in 2h	$1	Homopolymer
Pyrosequencing	700 bp	99,9%	up to 1 Gbp in 24h	$10	Homopolymer
Illumina (various devices)	50 – 600 bp	99,9%	up to several Tbp in 24h	$0,05 - $0,15	More errors toward read ends

Source: en.wikipedia.org/wiki/DNA_sequencing

- **Fast, accurate, cheap !**
- **Very short reads !**



Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata. 20. listopada 2019.



Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata. 20. listopada 2019.

What is wrong with short reads?

- Short reads are insufficient for assembling large genomes and genomes with repetitive parts
  - Repetitive Elements May Comprise Over Two-Thirds of the Human Genome (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3228813/)
- Short reads are often insufficient to uniquely identify RNA transcripts

- Solution: longer reads, especially reads that completely cover entire repetitive regions and cover entire RNA transcripts
  - Combining long and short reads
  - Long reads for generating draft genome assembly
  - Short reads for higher per-base accuracy

New technologies

- Third generation sequencing technologies or long read sequencing technologies

- Pacific Biosciences (PacBio)
  - Single Molecule Real Time Sequencing – SMRT Sequencing
  - 2011 – PacBio RS – first commercially available device
    - ”Sequencing mainframe”
- Oxford Nanopore Technologies (ONT)
  - Nanopore sequencing
  - 2015 – MinION – first commercially available device
    - ”Personal sequencing device”

New technologies

- From the computer science and applicability point of view:

Technology	Read length	Accuracy	Data throughput	Cost for 1M nucleotides	Error types
PacBio	30.000 bp (N50) Longest read 100.000 bp	87%	around 30 Gbp in 24h	$0,05 - $0,08	Random error
ONT	Depends on library preparation Longest read over 2 Mbp	92 - 97%	MinION – 30 Gbp per cell PromethION – 160 Gbp per cell	$0,1 - $0,5 (MinION)	Homopolymer

Source: en.wikipedia.org/wiki/DNA_sequencing

- **Less fast, less accurate i less cheap ?**
- **(Very) long reads!**

New technologies

- Third generation sequencing technologies or long read sequencing technologies
- Synthetic long reads - combining short read technology with barcoding (or similar procedure) to obtain highly accurate long sequences
  - 10x Genomics - Linked-reads technology
  - Illumina – Synthetic long reads
  - Complete Genomics - Long Fragment Read
- Alternative methods used in combination with sequencing to improve genome assembly
  - Optical mapping
  - Genome wide chromosome conformation capture methods (use NGS)

Pacific Biosciences

- Two types of reads
  - Subread – ”regular read”
    - Error rate roughly 10%
    - Read length up to 100 kbp
  - Read of insert – special read obtained using Circular Consensus technology, reading a regular reads multiple times dramatically increasing accuracy
    - Error rate below 1% (as the manufacturer claims)
    - Read length 10 – 20 kbp

Pacific Biosciences

- The latest sequencing machine introduced in April 2019.
  - Sequel II
- High accuracy, high throughput, but also a very high initial cost
- Under controlled conditions reads of insert of length 13.5 kbp and consensus accuracy of 99.9% were produced

Pacific Biosciences

- In 2013 it was estimated that most of the bacteria and arachaea genomes and be sequenced and completely assembled using solely PacBio long read technology
  - Genome Biology (Sep 13, 2013) "Reducing assembly complexity of microbial genomes with single-molecule sequencing"
- A paper was published in 2013. that demonstrates the use of PacBio devices for transcriptome analysis, completely capturing all isoforms
  - Nature Biotechnology (Oct 13, 2013) "A single-molecule long-read survey of the human transcriptome"

Pacific Biosciences

- In November 2018. Illumina agreed to purchase Pacific Biosciences for $1.2 billion! The deal is expected to complete by the end of 2019.

Oxford Nanopore Technologies

- Initially, two different read types, chemistry version R6
  - 1D reads – regular reads
    - **Error rate up to 30% !**
  - 2D reads – increased accuracy is obtained by reading both DNA strands connecting them with a hairpin construct
    - **Error rate up to 10-15%**
- 2016. – Chemistry version R9
  - $1D^2$ reads, reading both strands without physically connecting them
- With improvements in chemistry, current error rate is 5 - 15%, depending on the sample preparation protocol
- Chemistry version R10 was announced in 2019.
  - Better homopolymer handling
  - Consensus accuracy 99,99%

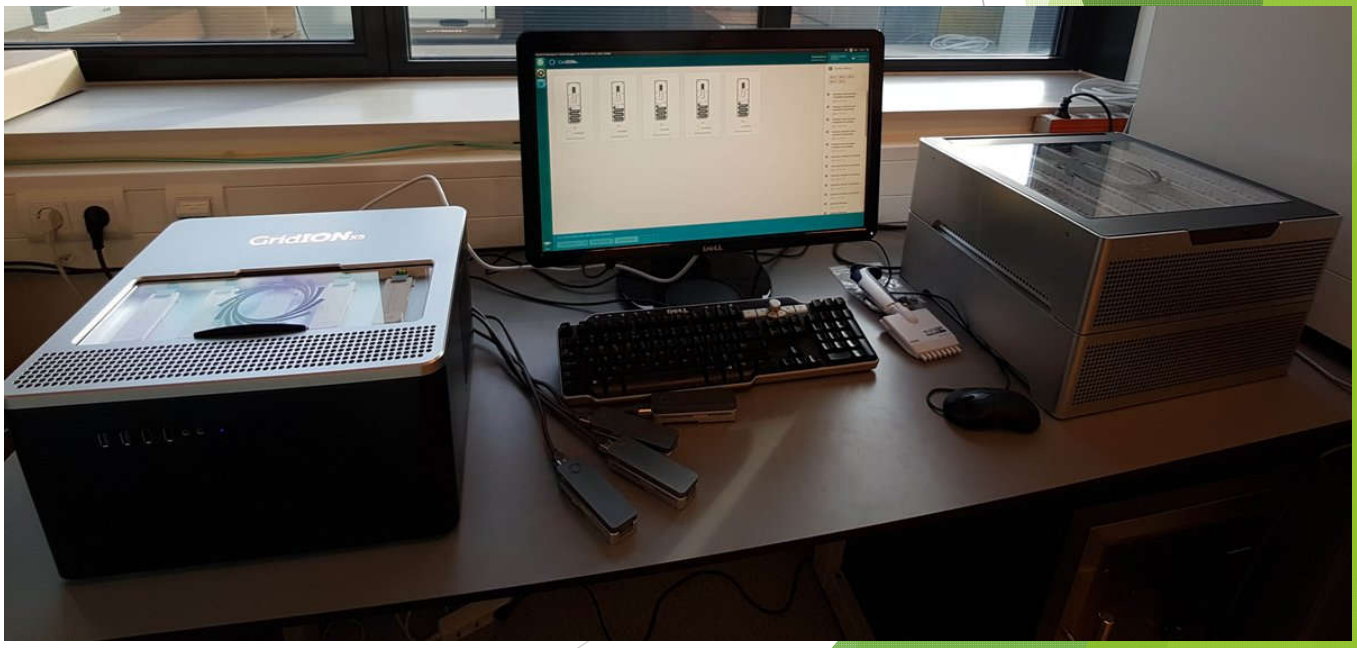
Oxford Nanopore Technologies

- MinION
  - Small portable device, connected to a PC through a USB port
  - Up to 30 Gbp in up to 48h
  - **Can sequence for shorter time!**
  - **Constantly produces results!**

- Starting price:
  - Mk1B $1000
  - Mk1C $4900
    - Includes an online data processing system

Oxford Nanopore Technologies

- GridION
  - 5 connected MinIONs
  - Up to 150 Gbp in 48h



Oxford Nanopore Technologies

- PromethION
  - New flowcell
  - Yield 100 – 180 Gbp
  - 48 parallel cells
  - Sequencing can last up to 72h
  - Theoretical maximum yield around 10Tbp

Oxford Nanopore Technologies

- SmidgeION
  - Sequencing on a mobile phone
- VolTRAX
  - Automatic sample preparation
- Fongle
  - adapter for MinION or GridION X5 that enables direct, real-time DNA or RNA sequencing on smaller, single-use flow cells & delivering up to 1.8 Gb of data.
  - No pipettes!
  - Basecalling on local PC/laptop

Oxford Nanopore Technologies

- April 2015, MinION was used for real time genomic surveillance of the ongoing Ebola epidemic (http://nature.com/articles/nature16996)
- In July 2016, a MiniON nanopore sequencer was included on the ninth NASA/SpaceX commercial cargo resupply services mission to the International Space Station.
  - Samples prepared on Earth were sequenced
- April 2019. - the highest throughput yet: PromethION breaks the 7 Terabase mark
  - 7.3 Tbp / 81h

What is on the horizon?

- *Sequencing the entire range of Earth's biodiversity is not a pipe dream anymore. In fact, it might become tangible reality within the lifetime of the current generation of scientists*, Eugene Koonin, NCBI

- Sequencing and assembling complex individual genomes
- Sequencing and assembling diploid and polyploid genomes
- Sequencing and assembling metagenomes (microbiomes)

- **Long read technologies will be essential!**

What is on the horizon?

- *Everybody talks about the $1,000 genome, but they don't talk about the $2,000 mapping problem behind the $1,000 genome*," Peter Tonellato, University of Wisconsin

- Sequencing produces extremely large amounts of data that needs to be processed, stored and efficiently searched at later time
  - Already a problem and will only get worse
  - New data storage models
  - New indexing structures
  - New algorithms for fast searching

Questions?

Thank you for your attention!