THIRD GENERATION SEQUENCING

- Overview of technologies, data properties and impact on genomic

SUMMARY

- Introduction
- Main technologies
- TGS data properties
- Impact of TGS technologies

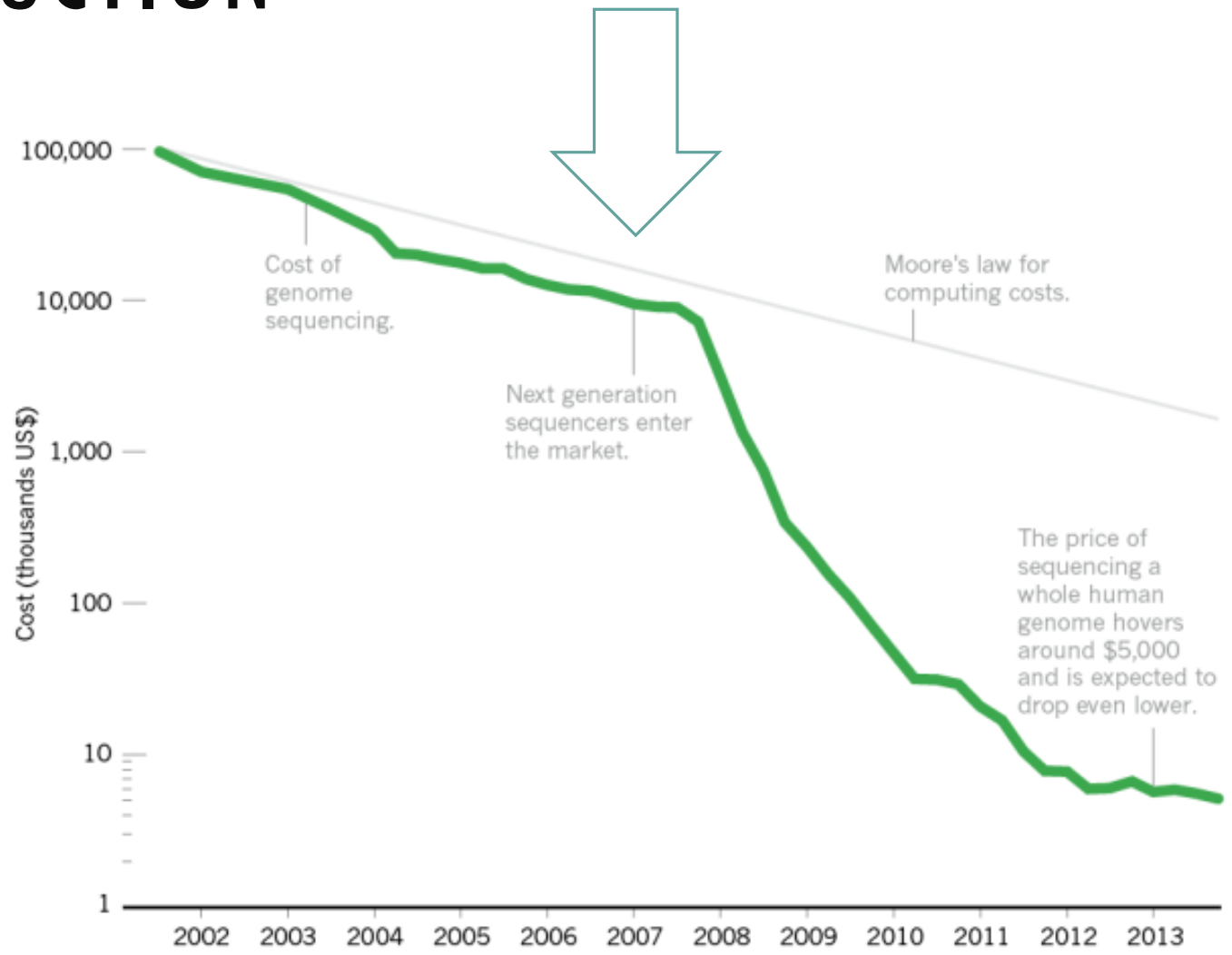
SUMMARY

- **Introduction**
- Main technologies
- TGS data properties
- Impact of TGS technologies

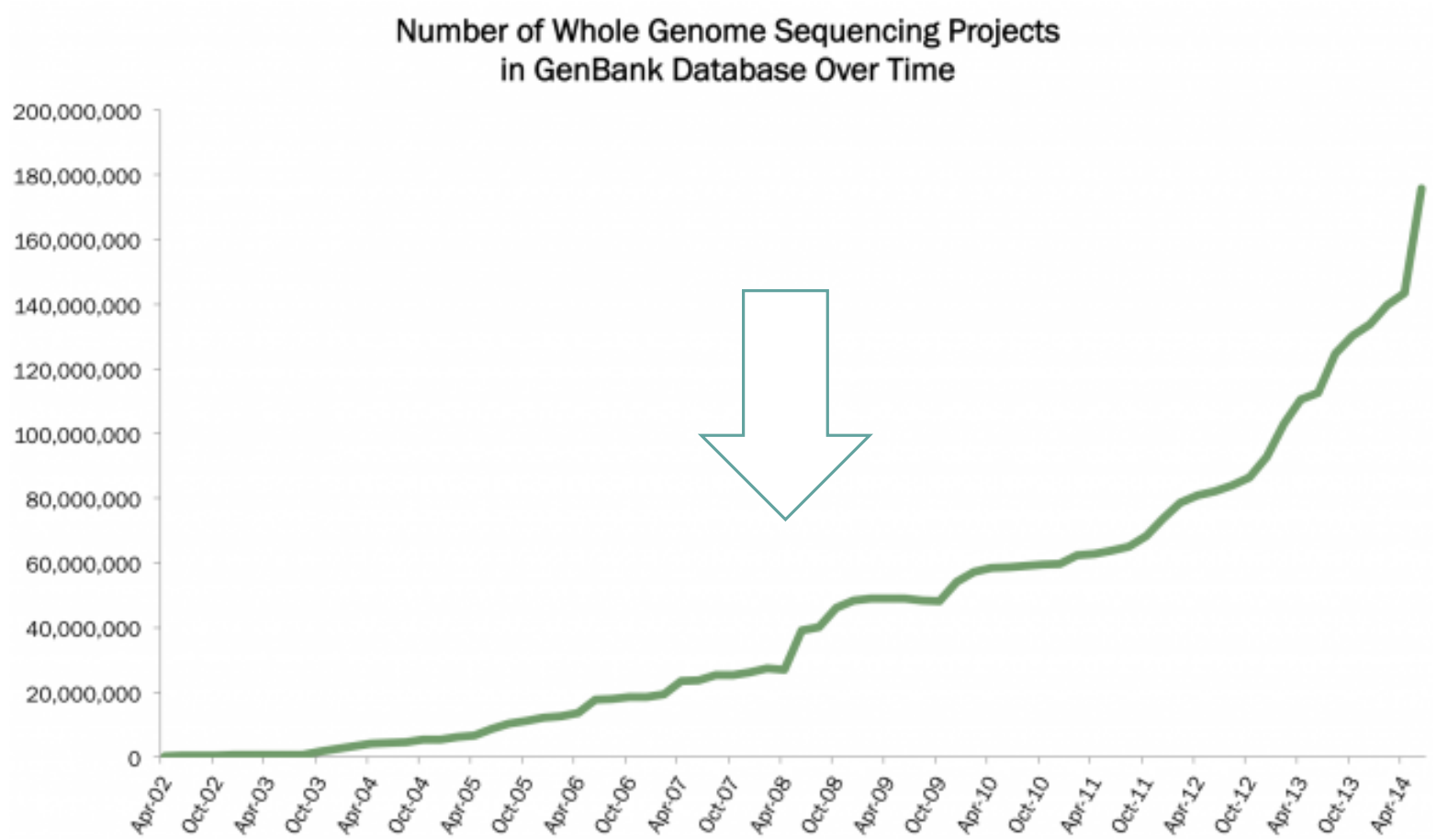
INTRODUCTION

- NGS had a crucial impact of genomics
  - Its massive parallelism reduced the sequencing cost
  - Enabled the sequencing of thousand of new organisms
  - Enabled population-scale sequencing projects
  - Allowed the characterization of some diseases at the genetic level

INTRODUCTION



INTRODUCTION



INTRODUCTION

- However, reads NGS technologies produce are short
  - In general, 50 to 500 base pairs long
- Such reads cause issues in computational analyses
  - Multiple alignments -> problems in the identification of the right location in the genome of the reads
  - Detecting InDels longer than approximately 50 base-pairs long
  - Assembling repetitive genome portions
  - Resolving structural variants longer than the average read length
- TGS technologies were designed for avoiding such limitations

SUMMARY

- Introduction
- **Main technologies**
- TGS data properties
- Impact of TGS technologies

MAIN TECHNOLOGIES

- TGS technologies start appearing in 2012
- Third Generation Sequencing technologies properties:
  - No wash-and-scan protocols, the sequencing reaction is not interrupted
  - Completely different kind of data generated
    - Multi kilobase-pairs reads
    - Error rate higher than 25%-30% in some cases

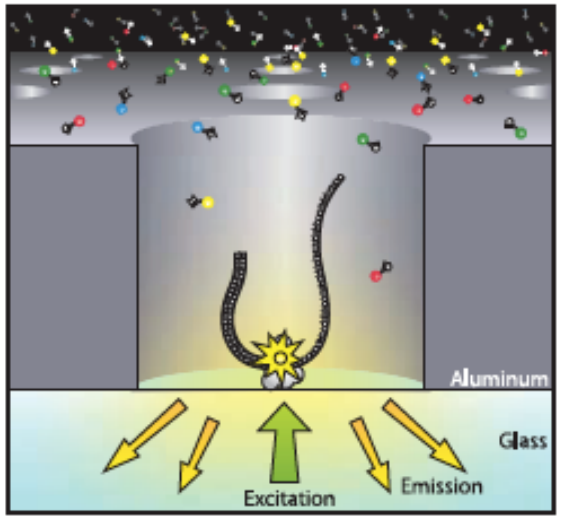
MAIN TECHNOLOGIES

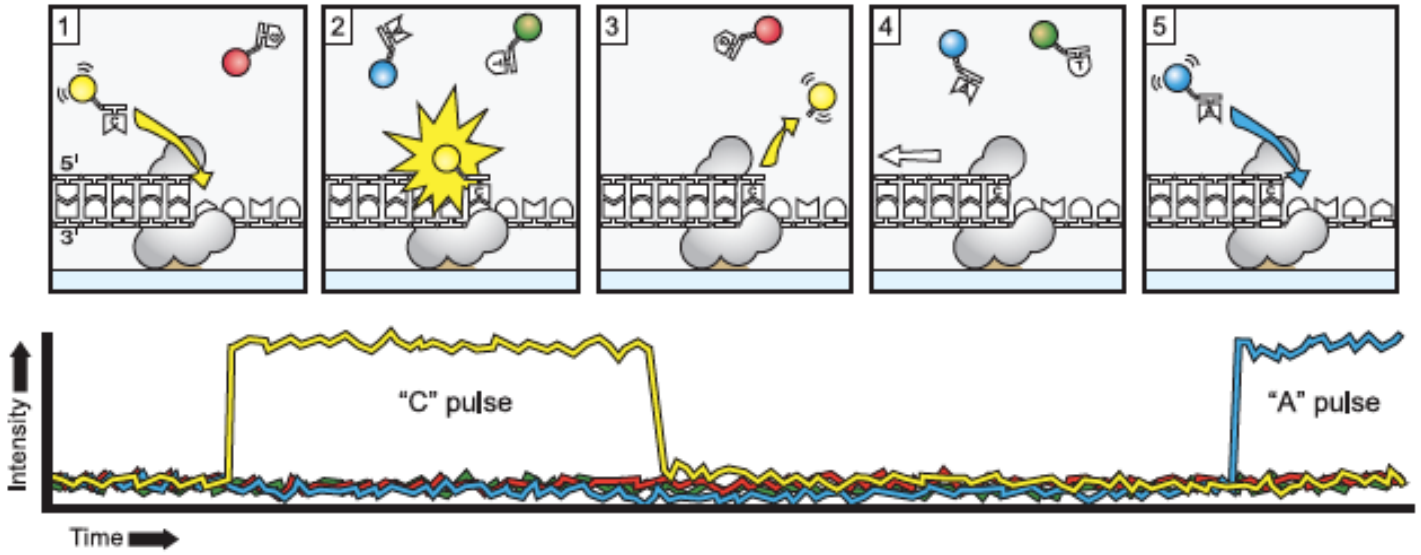
- Up to date, two main technologies on the market:
  - Single Molecule Real Time sequencing, by PacificBiosciences
  - Nanopore sequencing, by Oxford Nanopore Technologies
- Illumina uses proprietary long-read protocol, called *Moleculo*
  - Not a real TGS technology
  - ~10 kilobase-pairs reads are assembled from short NGS reads
  - Short reads coming from similar genomic regions are recognized by looking at a special tag attached during library preparation

MAIN TECHNOLOGIES – SMRT SEQUENCING

- Sequencing-by-synthesis technology
- Direct observation of DNA polymerase at work
  - Sequencing happens at the bottom of a particular nanophotonic visualization chamber called Zero Mode Waveguide (ZMW)
  - DNA polymerase is tightly attached at the bottom of ZMW
  - Each base incorporation releases a different colored fluorescent label
  - A sensor detects different light pulse and performs base-calling

MAIN TECHNOLOGIES – SMRT SEQUENCING

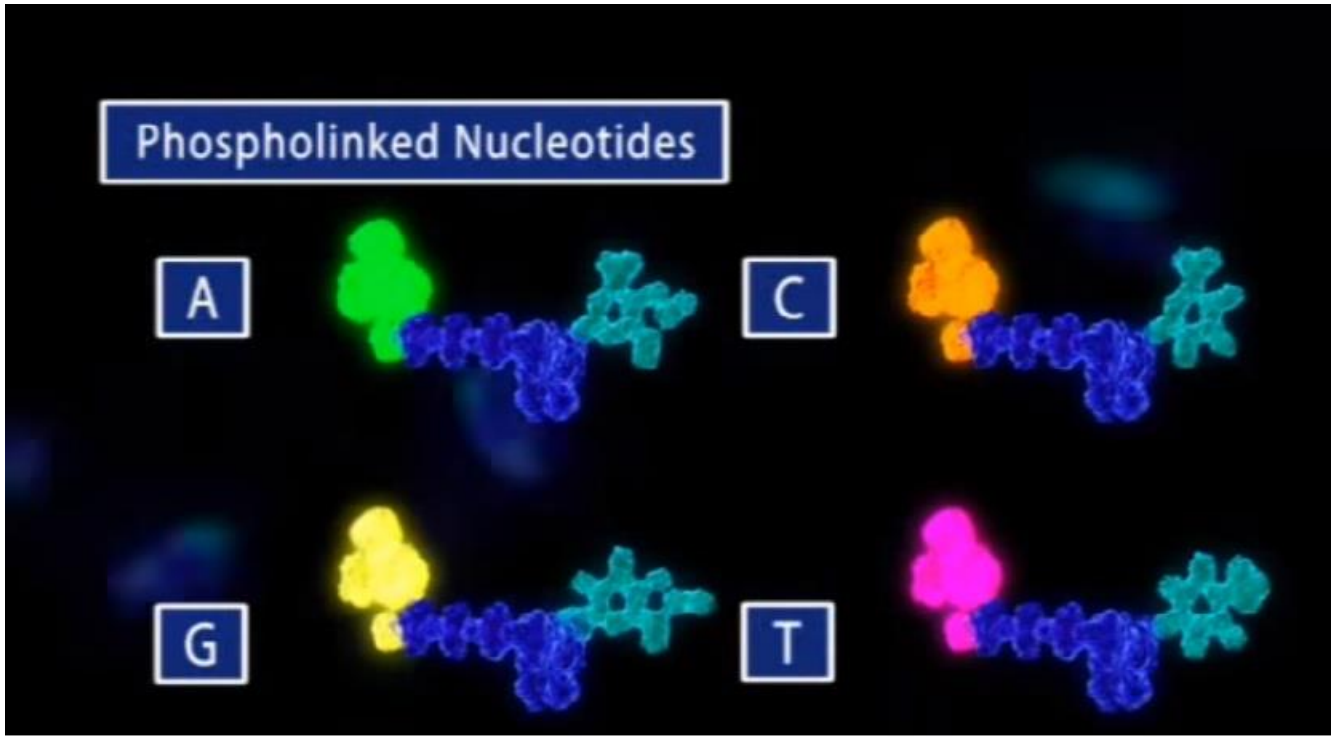




YouTube video : https://www.youtube.com/watch?v=v8p4ph2MAvI

MAIN TECHNOLOGIES – SMRT SEQUENCING

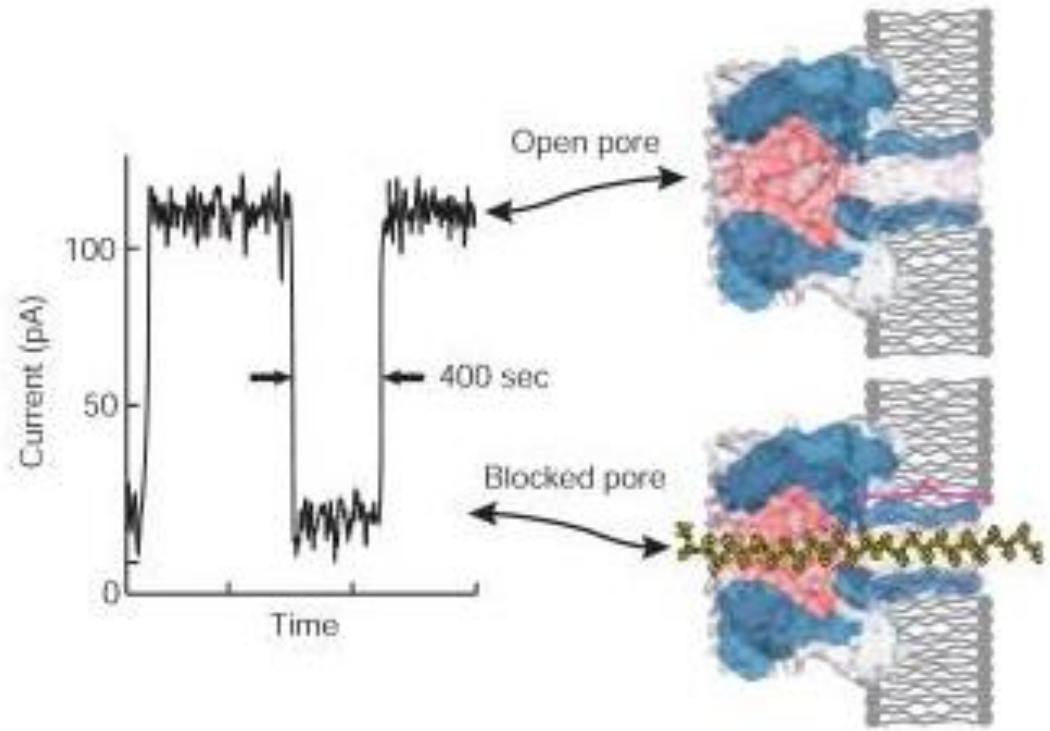
- The colored fluorescent label is incorporated to each base at the terminal phosphate rather than the base
- DNA-pol releases the fluorescent label as part of the incorporation process leaving behind a natural DNA strand



MAIN TECHNOLOGIES – NANOPORE SEQUENCING

- Relies on current measurements over a nanopore inserted in a polymer membrane with very high electrical resistence
  - A voltage bias is imposed across the membrane
  - Whenever ions in solution flow through a nanopore a current is measured
  - When a DNA strand flow through the pore, the ions flow is perturbed
  - The current varies differently depending on the nucleotide in the pore
- Observing the current evolution base-call is possible
- No imaging techniques required

MAIN TECHNOLOGIES – NANOPORE SEQUENCING



YouTube video : https://www.youtube.com/watch?v=CGWZvHli3i0
https://www.youtube.com/watch?v=GUb1TZvMWsw

SUMMARY

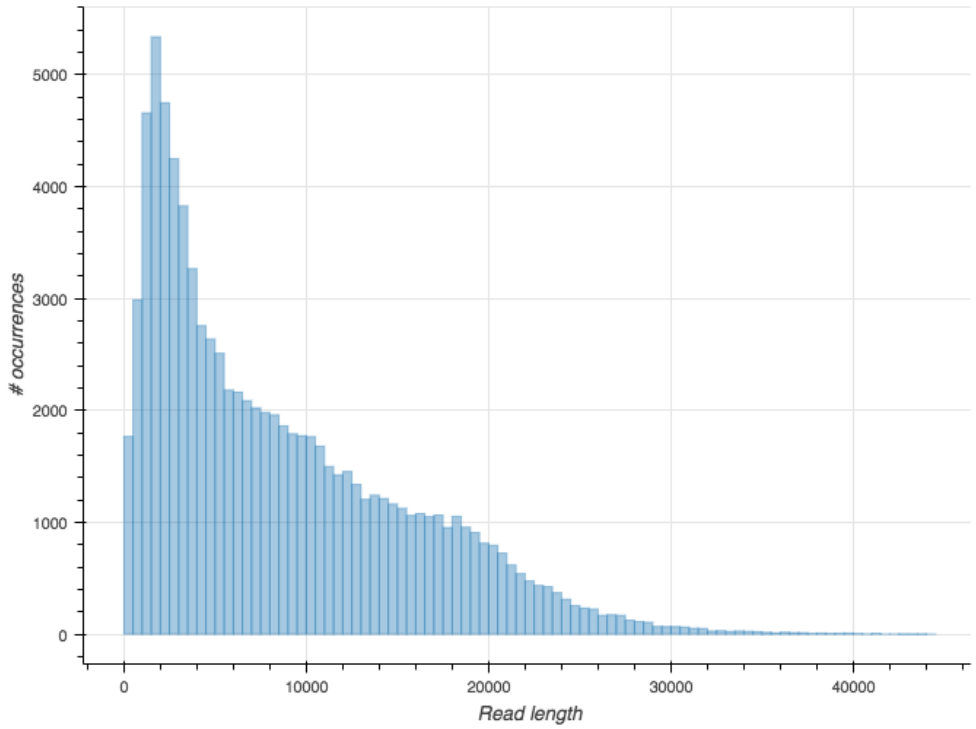
- Introduction
- Main technologies
- **TGS data properties**
- Impact of TGS technologies

TGS DATA PROPERTIES

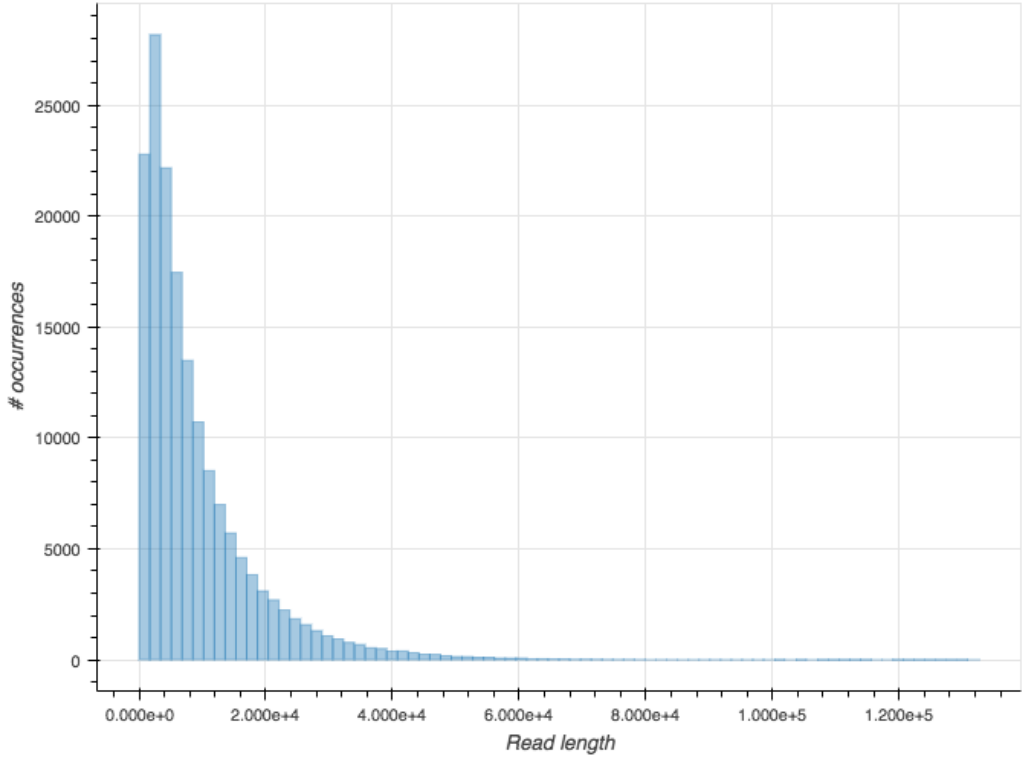
- TGS data have radically different properties w.r.t. previous data
  - Much longer reads, usually longer than 5 kilo base-pairs in the average
  - Great variety of read lengths, from 500 to more than 100000 bases
  - High error rate, in general higher than 10%, sometime over 30%
- Such data request for new approaches in designing pipeline for genetic analyses

TGS DATA PROPERTIES

SMRT read length profile



Oxford nanopore read length profile



TGS DATA PROPERTIES

		Error rates [%]			
Dataset	Tool	Substitution	Insertion	Deletion	Total
SMRT	BWA-MEM	1.9	7.2	2.6	11.7
	Minimap2	1.7	8.0	2.7	12.4
Nanopore	BWA-MEM	7.4	2.7	7.7	17.8
	Minimap2	6.2	3.3	8.3	17.8

SUMMARY

- Introduction
- Main technologies
- TGS data properties
- **Impact of TGS technologies**

IMPACT OF TGS TECHNOLOGIES

- TGS long reads promise to increase quality of assemblies
  - Spanning repetitive regions in complex genomes
  - Detecting different structural variants
  - Increase size of contigues (created by overlapping reads)

IMPACT OF TGS DATA

- High error rate do not affect assembly per-base quality
  - TGS errors are randomly distributed and independent of genome content
- Explicit error-correction steps embedded in assembly pipeline
  - Self-correction algorithms
    - Long reads overlapped one against the other and polished running consensus algorithms
  - Hybrid correction algorithms
    - Accurate NGS reads aligned over long noisy reads which are then corrected resorting to consensus again