CENTER OF
PLANT SCIENCES

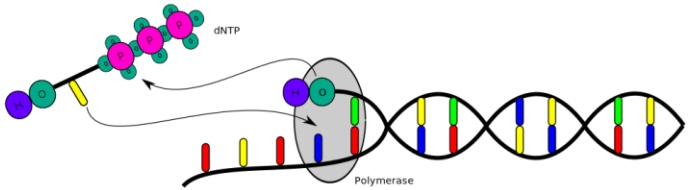
Sant'Anna
Scuola Universitaria Superiore Pisa

Advanced Genomics

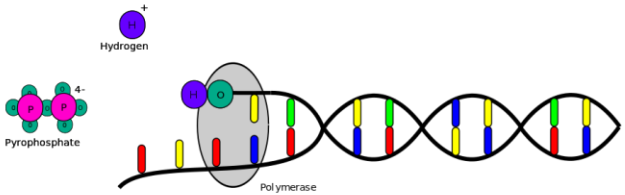
From MPS to III gen sequencing

Ion Torrent

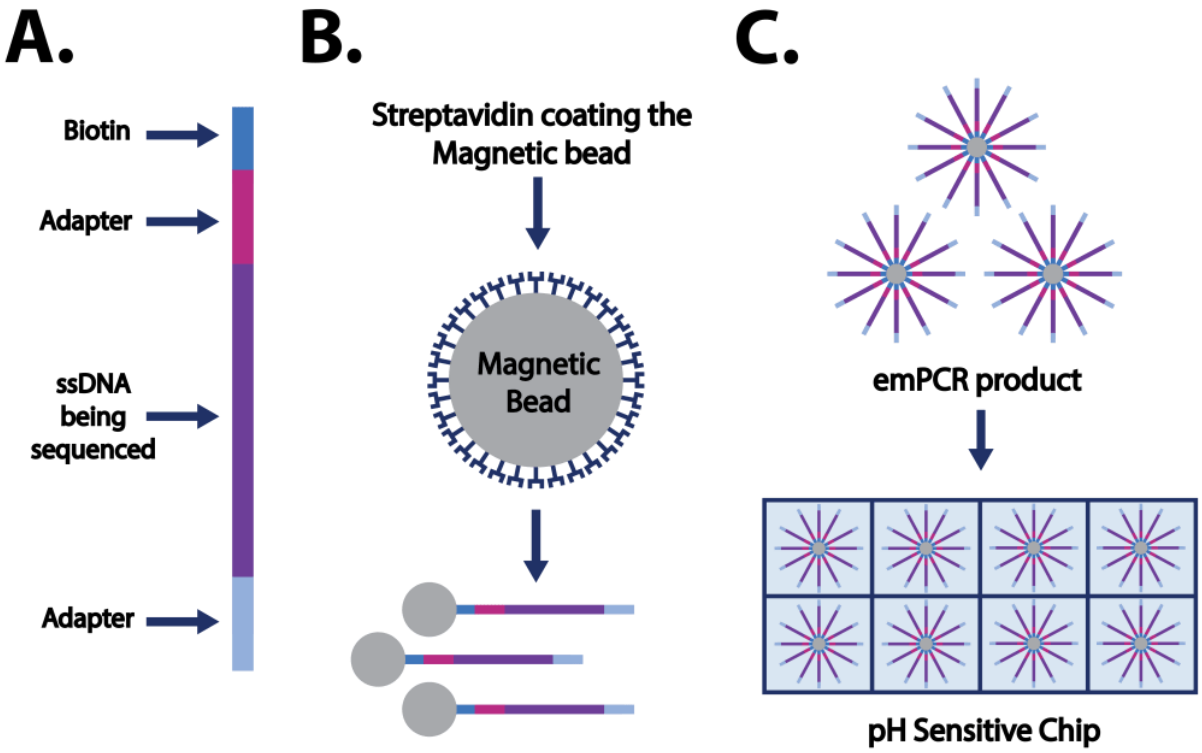
- Sequencing by synthesis
- Detection is based on release of H+ ion in solution
- Easy measurement, no complex chemistry



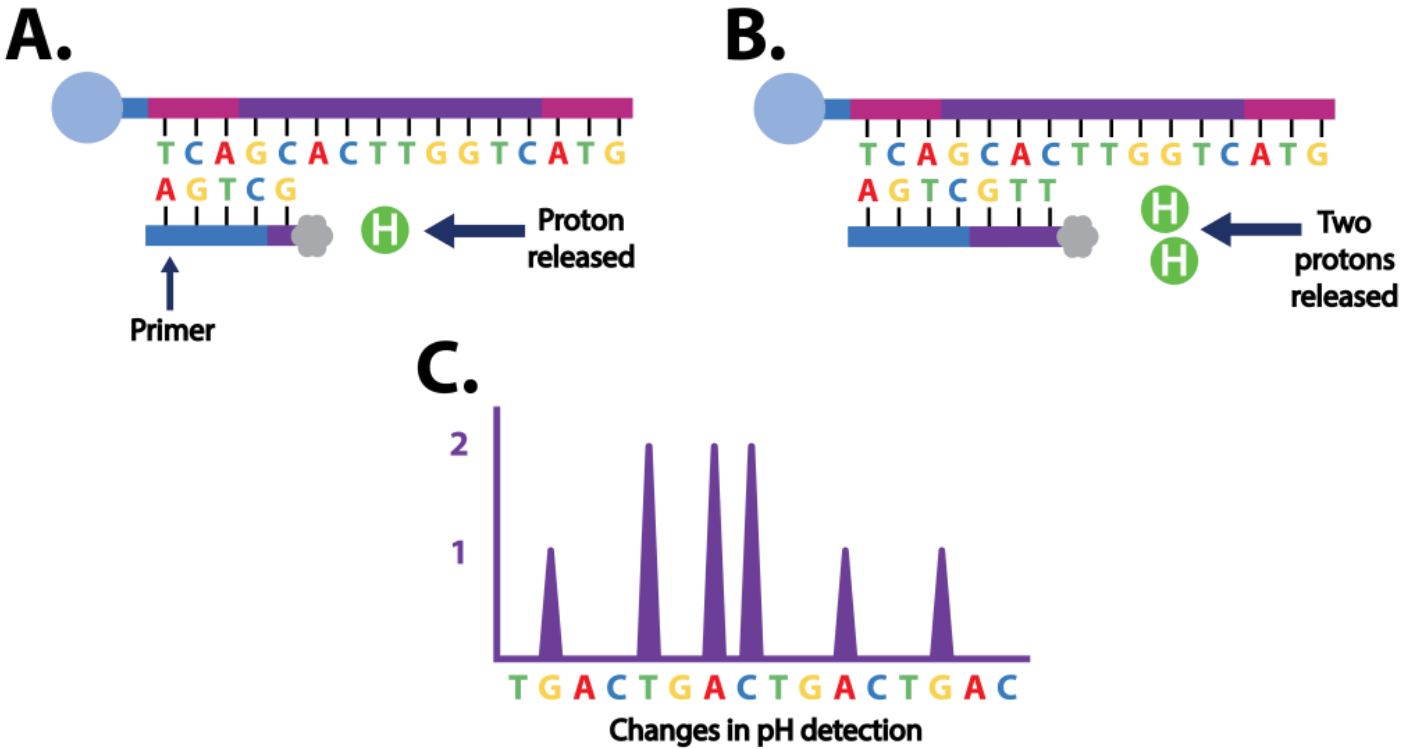
Polymerase integrates a nucleotide.



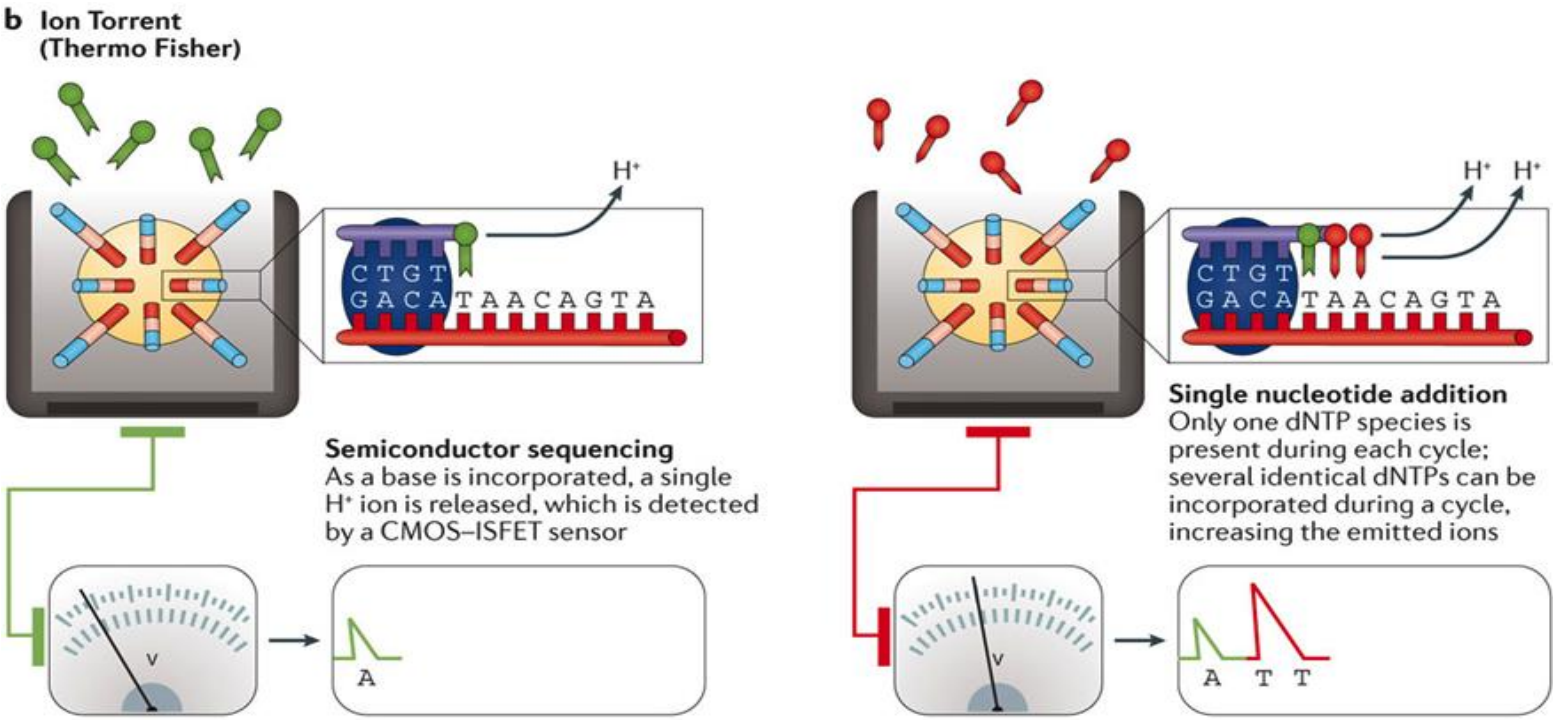
Hydrogen and pyrophosphate are released.



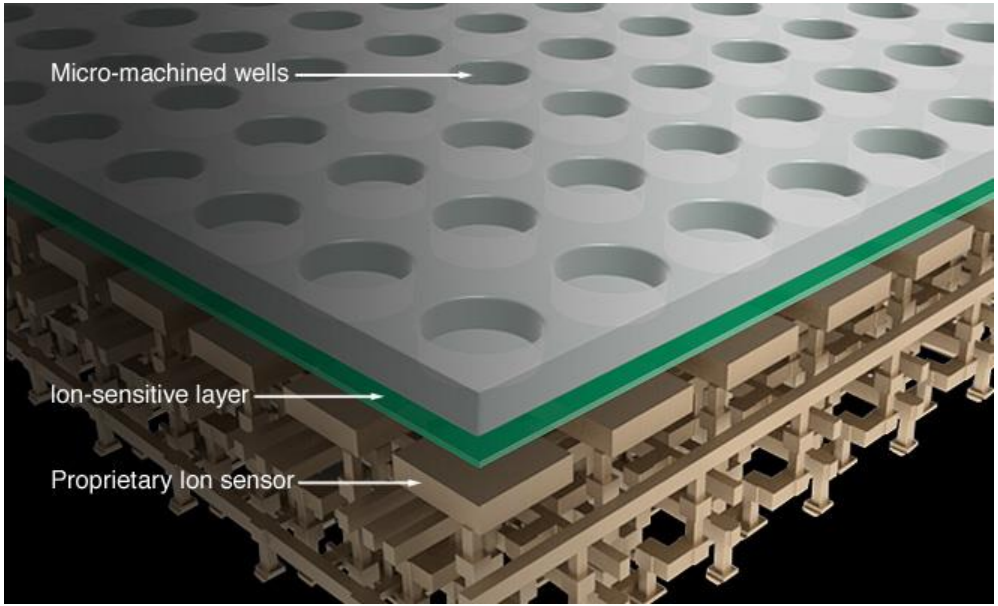
Library prep – DNA fragments are linked to adapted and to beads. Fragments are amplified (one per bead) and put in pH sensitive chip



Sequencing – Nts are put one after the other; whenever a Nt gets incorporated, an H+ ion is released and can be detected by a perturbation of voltage


b Ion Torrent
(Thermo Fisher)
H+
C T G T
G A C A T A A C A G T A
Semiconductor sequencing
As a base is incorporated, a single H+ ion is released, which is detected by a CMOS–ISFET sensor
V
A
H+ H+
C T G T
G A C A T A A C A G T A
Single nucleotide addition
Only one dNTP species is present during each cycle; several identical dNTPs can be incorporated during a cycle, increasing the emitted ions
V
A T T

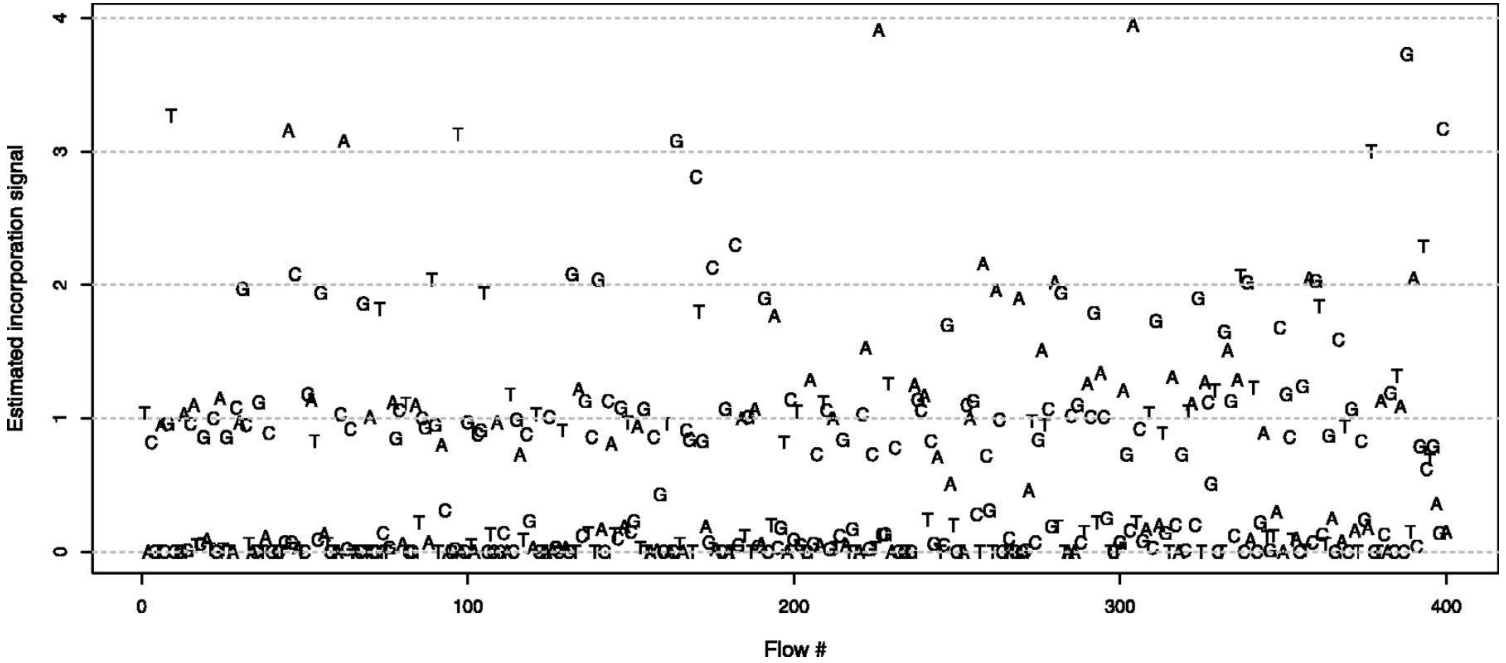

Ion Proton™ I Chip
2 human exomes
165 million wells
$1,000 per run

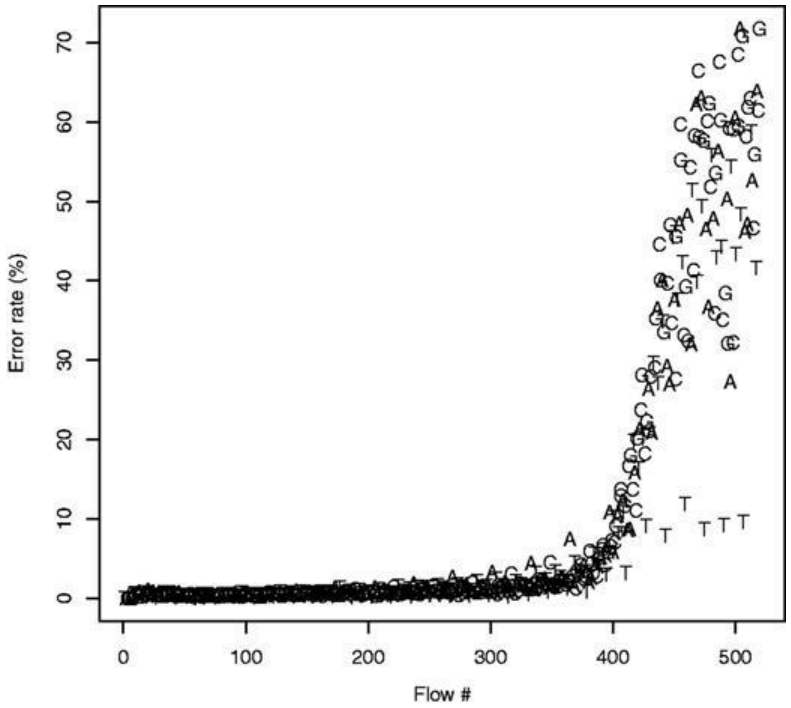

Micro-machined wells
Ion-sensitive layer
Proprietary Ion sensor

https://www.youtube.com/watch?v=zBPKj0mMcDg&ab_channel=ThermoFisherScientific

Machine readouts

Ideal signals are expected to be integers, indicating exactly how many nucleotides were incorporated during each flow. However, the actual signal at each flow is noisy, and the noise increases as the sequencing process advances





- As the sequencing cycles advance, error rate increases due to wearing out of the process
- Modelling approaches can improve base calling and homopolymer definition (see e.g. https://doi.org/10.1093/bioinformatics/btt212)

Pros

- Low complexity
- Low sequencing cost
- Fast turnaround
- Low error rate

Cons

- Problematic call with homopolymers
- Issue with indels and structural variation

- 2-7 h/run
- Read: 400bp
- 400.000- 5 M reads

- 2-4 h/run
- Reads: 200bp
- 60-80Mreads >=50bp

Ion GeneStudio S5 System

Scalable targeted NGS to support small and large projects

The Ion GeneStudio S5 system is a scalable, targeted-NGS workhorse with wide application breadth and throughput capability.

Ion Torrent Genexus System

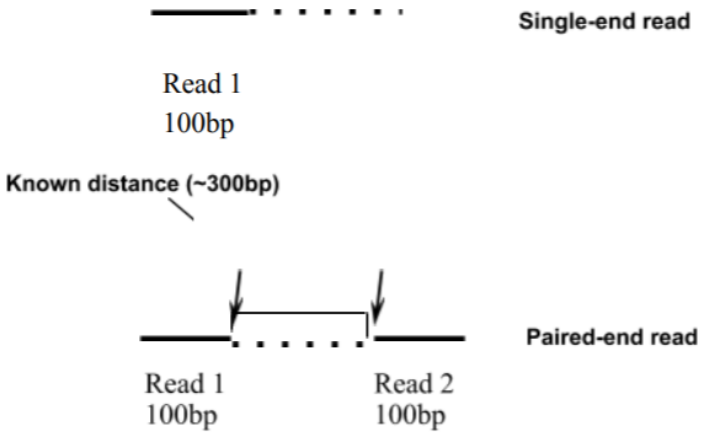
Specimen to report in a single day with a hands-off, automated workflow*

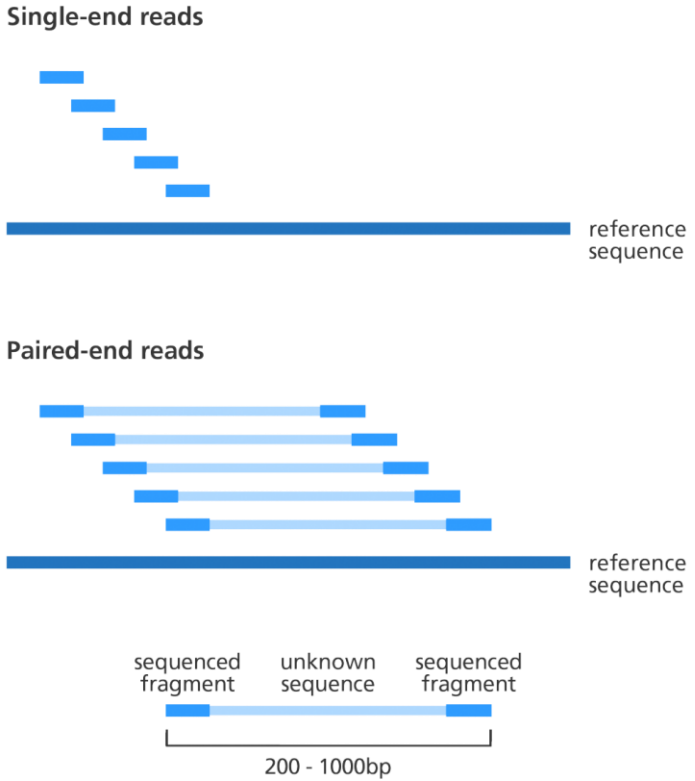
The Genexus System is the first turnkey NGS solution that automates the specimen-to-report workflow and delivers results in a single day with just two user touchpoints.*

Very used in clinical applications

Optimizing sequencing range via paired ended sequencing

- Read lenght may be extended using paired ended sequencing, i.e. including a fragment of unknown sequence but known size in between two reads
- It remains challenging to reconstruct complex genomes





NGS pros

- No *E.coli* subcloning
    - No cloning bias
    - Easier library preparation
    - Lower robotics, also for large genomes sequencing projects
- Each sequence comes from a single DNA molecule
    - Quantification possible by ‘digital counting’
    - Huge dinamic range
    - Rare variants detection
- Revolutionary cost decrease and very fast data production
- Non only DNA sequencing but a lot of different applications
    - DNA-seq
    - RNA-seq
    - CHIP-seq
    - amplicon-seq
    - target resequencing
    - BS-seq
    - Etc...

NGS cons

- Shorter reads with respect to Sanger sequencing
  - Illumina technology can reach now 300bp paired-end (e.g. 600bp/fragment)
  - Third and fourth generation sequencers will produce very long reads (Kbp)
- Big investments in terms of IT infrastructures (storage and RAM)
    - Tbytes of data are produced per run
    - Analysis pipelines out of the instrument
- Complex bioinformatics data analysis
    - Scripting in real time
    - The technology is going faster than the human capacity to give to the data a biological significance

Need for long reads technology

- Genomes are highly complex!
- Resolution of large structural features is needed
- Complex repetitive regions
- Chromothripsis (chromosome rearrangements, e.g. in cancer)
- Transcriptome research (mRNA transcripts, gene isoforms, connection exons)
- De novo genome reconstruction
- …

Example of complexity

- Wild emmer genome (Avni et al 2017, Science)

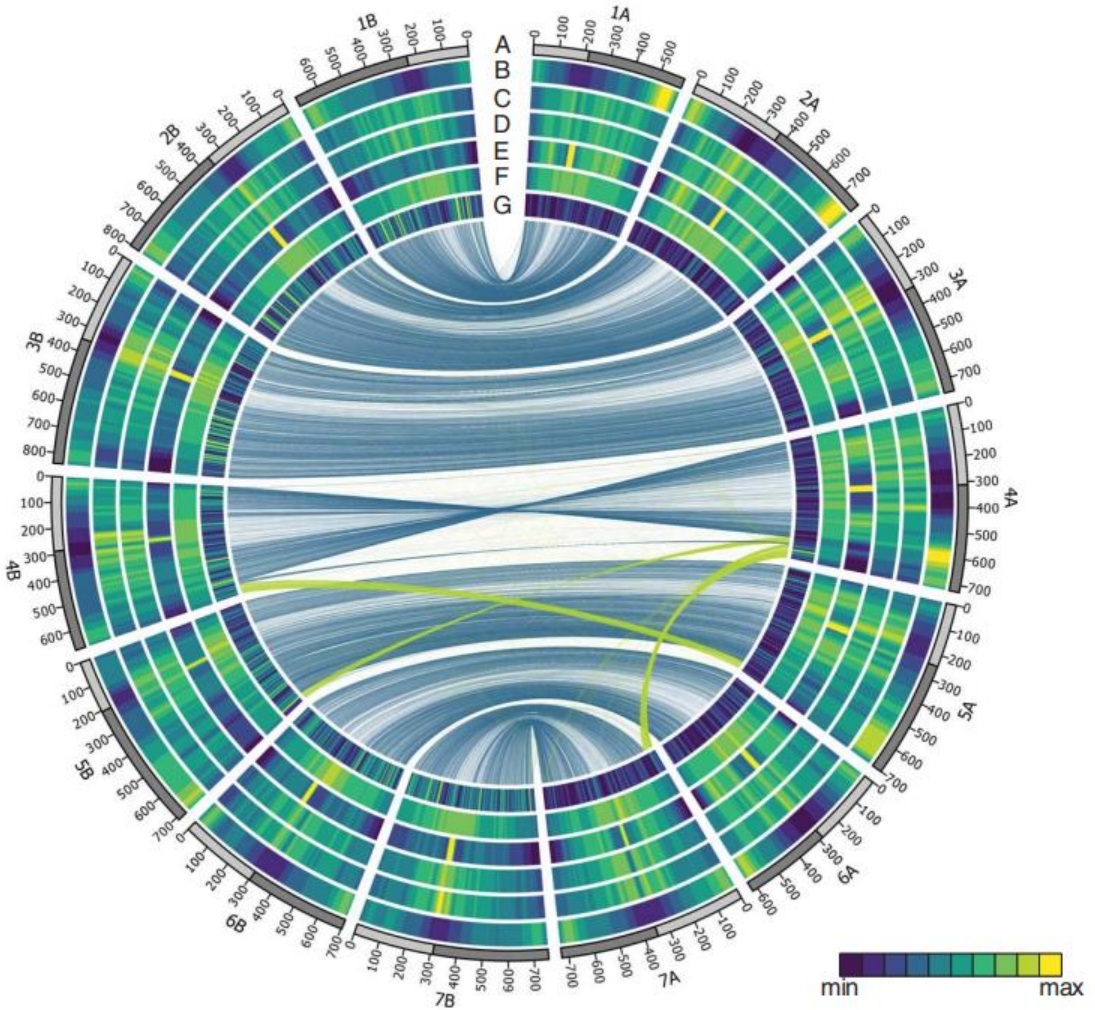
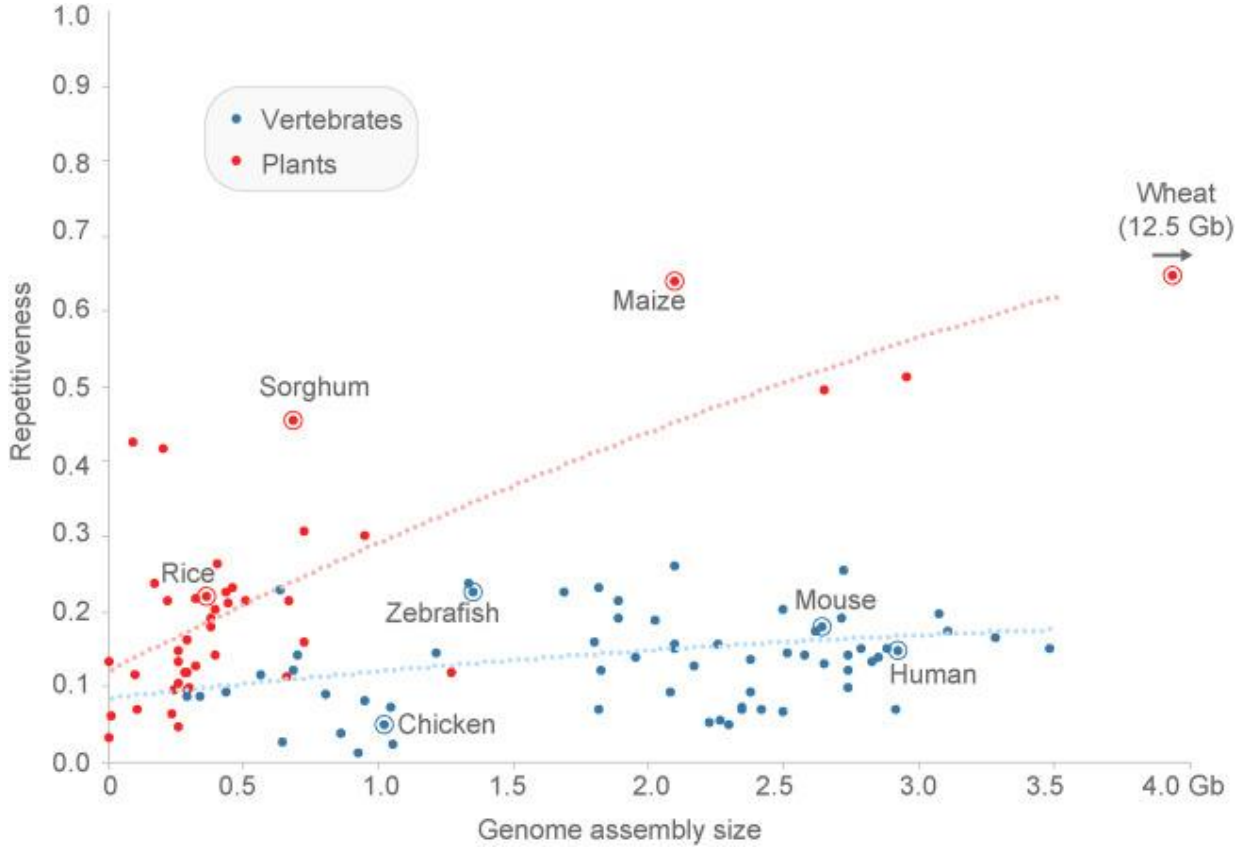


Fig. 2. Structural, functional, and conserved synteny landscape of the 14 chromosomes of wild emmer wheat (WEW). Tracks from outside to inside, with respective scales, are as follows: (**A**) Chromosome name and size (each tick mark is 100 Mb). (**B**) Density of high-confidence genes (HC; 0 to 14 genes per Mb). (**C**) Expression of HC genes [log(FPKM + 1)]; mRNA expression ranges from 0 to 2.5; mean value of all 20 conditions (i.e., tissue types + time points). (**D**) Expression density of HC genes (number of samples in which genes were expressed; from 0 to 20). (**E**) *K*-mer frequencies. (**F**) Percent identity between homeologous genes (90 to 100%). Central links connect homeologous genes between subgenomes. Color of links is blue between homeologous chromosomes and green in cases of large translocations. (**G**) Density of Ty1/copia-like insertions older than 1.2 Ma (insertions per Mb).



Wen Biao Jao et al 2017

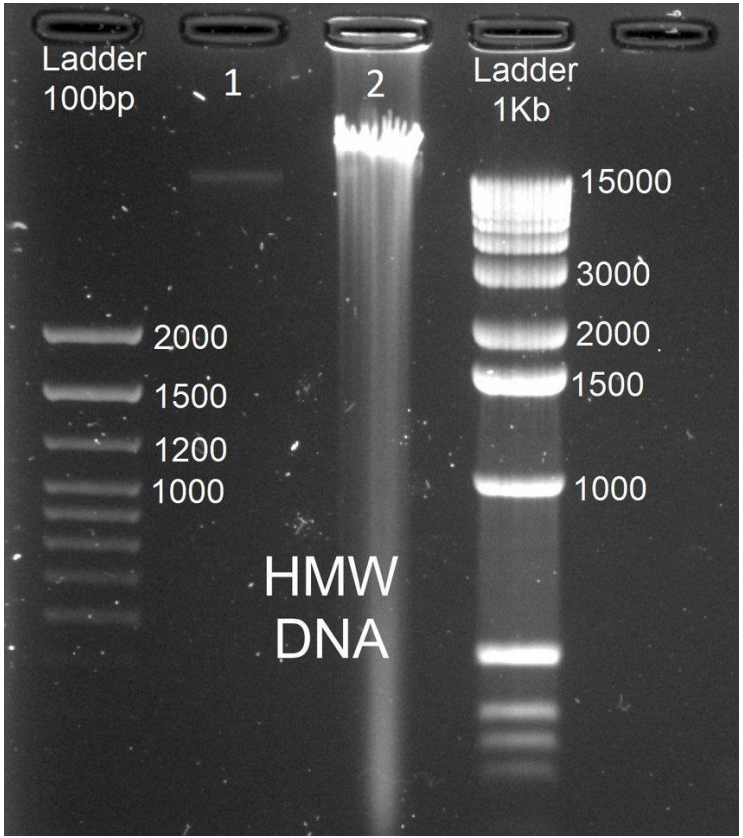
approach	company	Technology/challenge	Fragment length
Single Molecule Real Time (SMRT)	Pacific Biosciences	Zero Mode Waveguides (ZMW) Fixed polymerase	10+ kb
	Oxford Nanopore	Direct detection Detection of k-mers	
Synthetic approach (*in silico*)	Illumina	microtiter	10+ kb
	10x Genomics	Emultion PCR	100+ kb

SMRT

- Single molecule approach → do not rely on a clonal population of amplified DNA fragments
- Do not require chemical cycling for each dNTP added

Key assett: HMW DNA

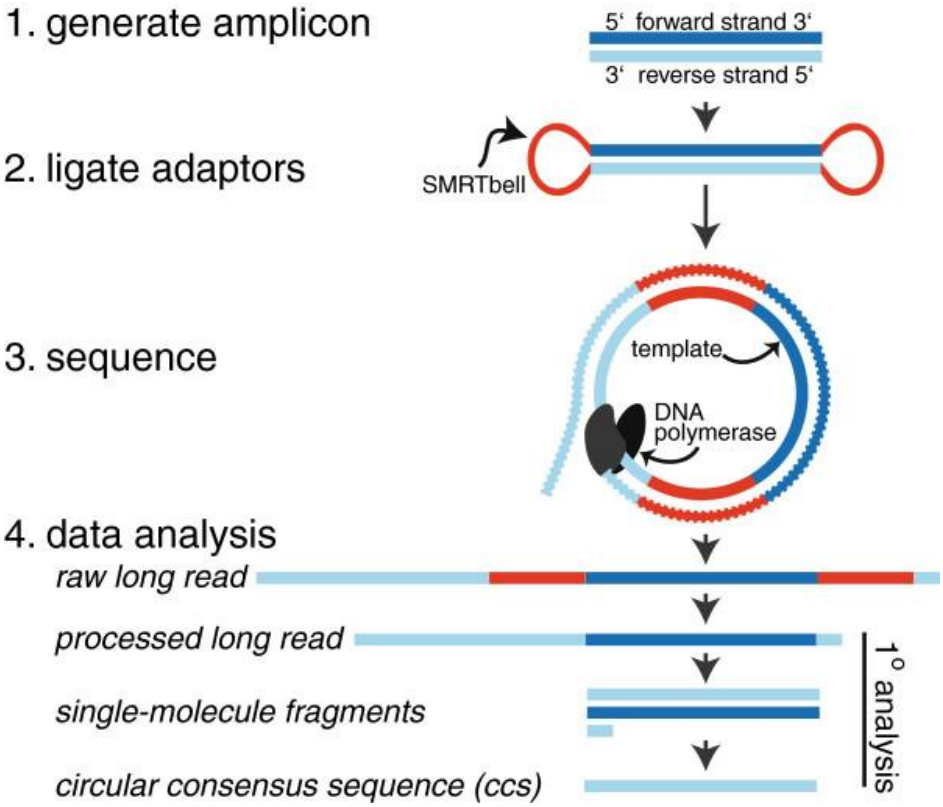
- To leverage long read sequencing, unfragmented, full length DNA molecules are desirable
- (human DNA is about 2m long!)



Pacific Biosciences (PacBio)

Single Molecule, Real-Time (SMRT) sequencing

- First step is looping the DNA with adapters (SMRTbell)
- The SMRTbell is topologically circular and structurally linear
- Polymerase is attached to it and used to replicate DNA in a circular loop
- Later on, adaptors are bioinformatically removed and original sequence is reconstructed

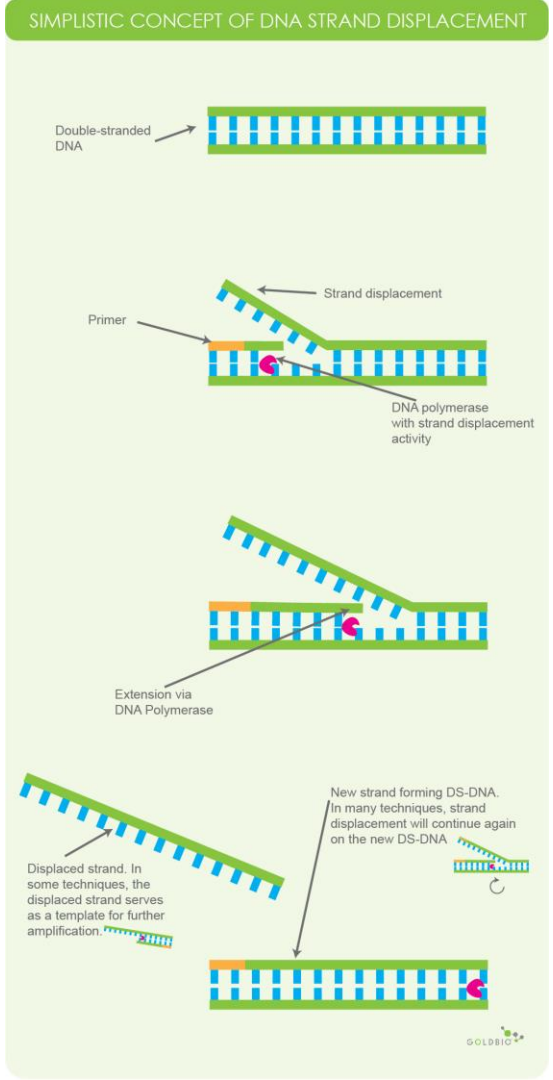


Due to the hairpin design of our DNA substrates, we were able to observe the DNA holoenzymes operating in both strand displacement and primer extension modes. During strand displacement synthesis, the enzyme needed to open one base pair of the DNA hairpin in order to incorporate one nucleotide resulting in a large increase in molecular extension, ~0.8 nm for each nucleotide incorporated at 10 pN of applied force (Supplementary Figure S2). Once the enzyme synthesized through the hairpin loop to completely open it (or the hairpin substrate was mechanically opened

3.4. DNA polymerase—the sequencing "engine"

Various DNA polymerases can be used in conjunction with SMRT DNA sequencing, and the sequencing performance will depend on their specific properties. We have applied wild-type and mutant DNA polymerases from bacteriophage ϕ29 to our SMRT DNA sequencing method, taking advantage of several favorable characteristics. ϕ29 DNA polymerase is extremely processive (tens of kilobases), relatively fast (~50–100 bases/s) and highly accurate (error rate of $\sim 10^{-5}$–10^{-6}) (Baner *et al.*, 1998; Blanco *et al.*, 1989; Esteban *et al.*, 1993). It is also very stable, maintaining constant enzymatic activities for up to several days (Dean *et al.*, 2001; Nelson *et al.*, 2002). The use of double-stranded DNA templates is possible by its efficient DNA strand displacement synthesis activity, thus simplifying sample preparation procedures.

Korlach, Jonas (2010). *[Methods in Enzymology] Single Molecule Tools: Fluorescence Based Approaches, Part A Volume 472 || Real-Time DNA Sequencing from Single Polymerase Molecules. , (), 431–455.* doi:10.1016/S0076-6879(10)72001-2



- Detection of sequence happens through imaging of fluorescence emitted by different nucleotides (somewhat similarly to what happens with Illumina)
- The zero-mode waveguide (ZMW) is a nanophotonic confinement structure that consists of a circular hole in an aluminum cladding film deposited on a clear silica substrate
- Light emission is telling of the incorporated nucleotide

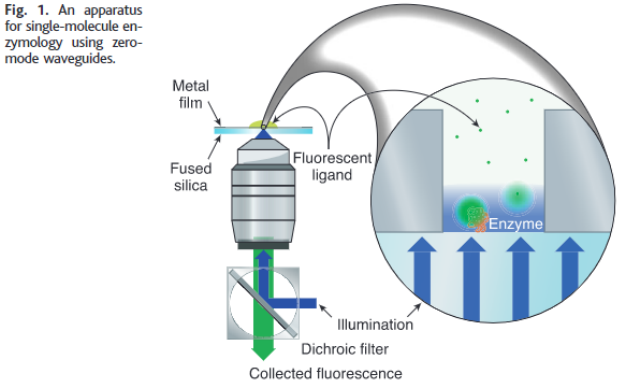


Fig. 1. An apparatus for single-molecule enzymology using zero-mode waveguides.

Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations

M. J. Levene,[1] J. Korlach,[1,2] S. W. Turner,[1]* M. Foquet,[1] H. G. Craighead,[1] W. W. Webb[1]†

Optical approaches for observing the dynamics of single molecules have required pico- to nanomolar concentrations of fluorophore in order to isolate individual molecules. However, many biologically relevant processes occur at micromolar ligand concentrations, necessitating a reduction in the conventional observation volume by three orders of magnitude. We show that arrays of zero-mode waveguides consisting of subwavelength holes in a metal film provide a simple and highly parallel means for studying single-molecule dynamics at micromolar concentrations with microsecond temporal resolution. We present observations of DNA polymerase activity as an example of the effectiveness of zero-mode waveguides for performing single-molecule experiments at high concentrations.

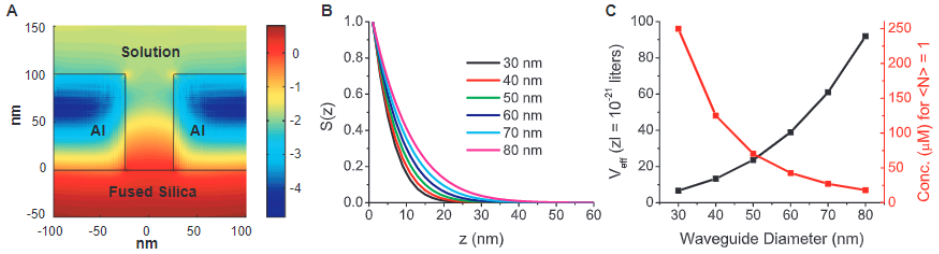


Fig. 2. (A) Three-dimensional finite-element time-domain simulation of the intensity distribution (log scale) for a zero-mode waveguide 50 nm in diameter and 100 nm long. (B) $S(z)$ curves for different waveguide diameter, d. (C) V_{eff} and the corresponding concentration for which there is, on average, one molecule in the volume ($\langle N \rangle = 1$).

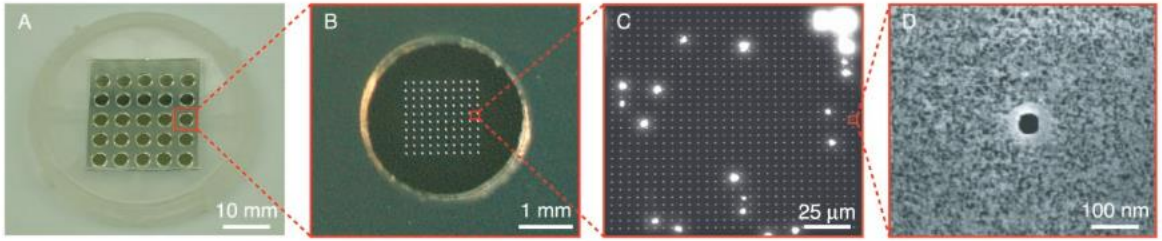
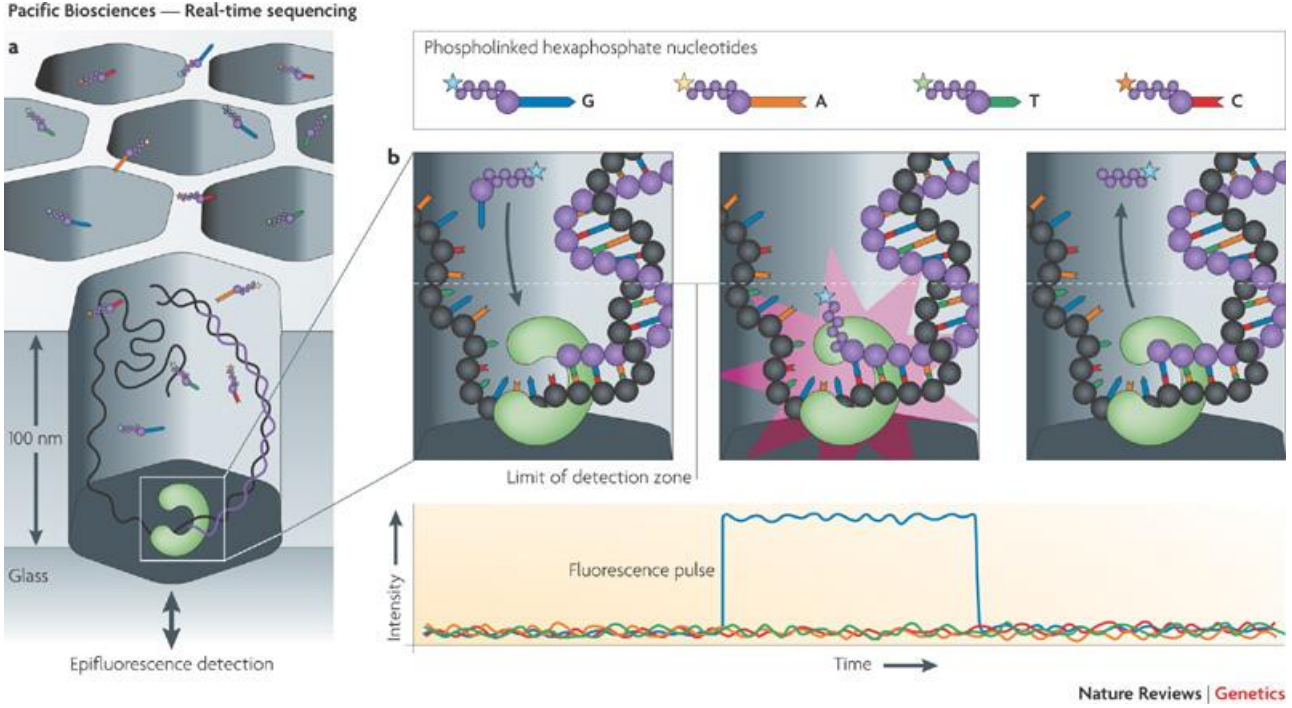
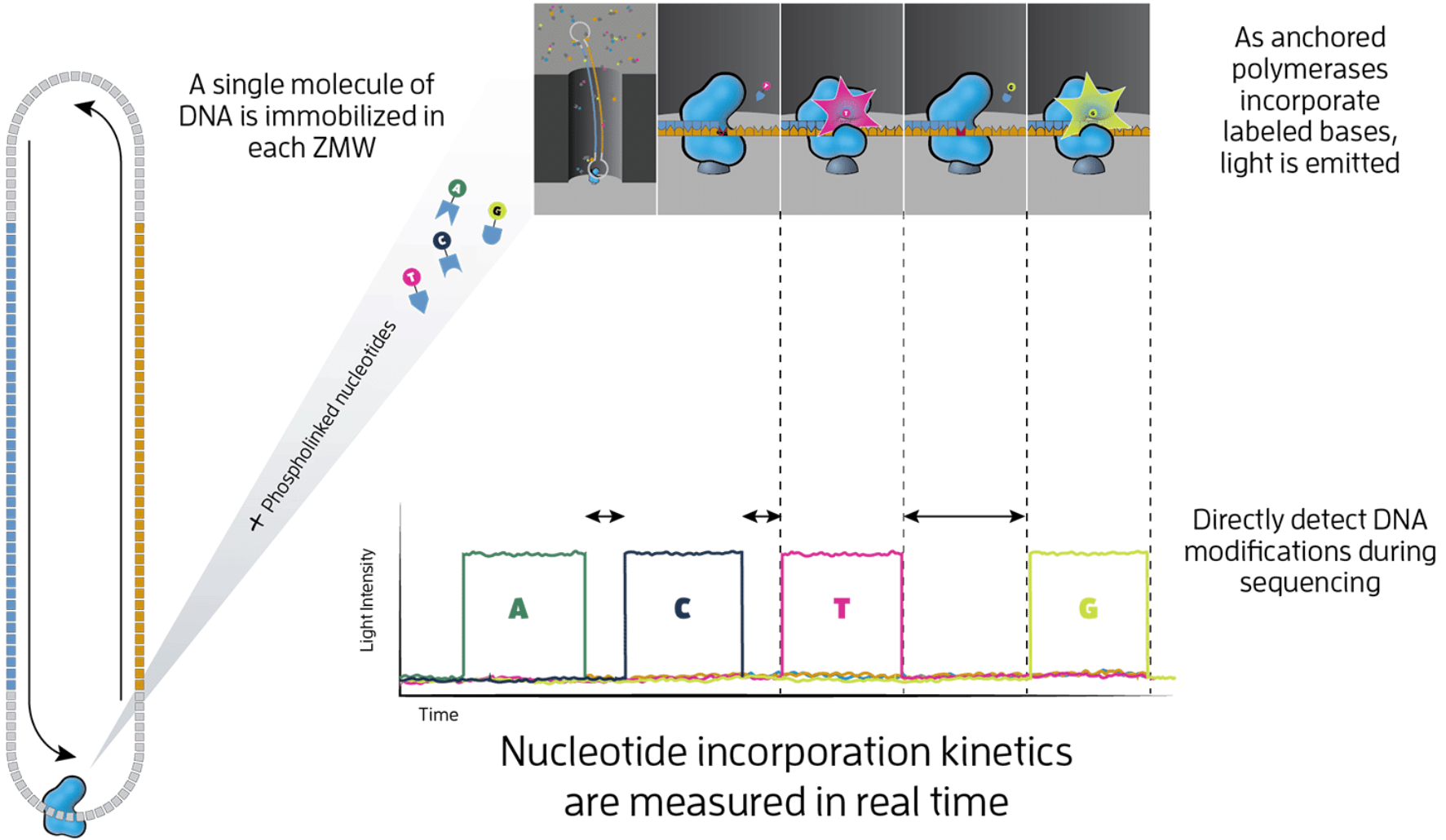
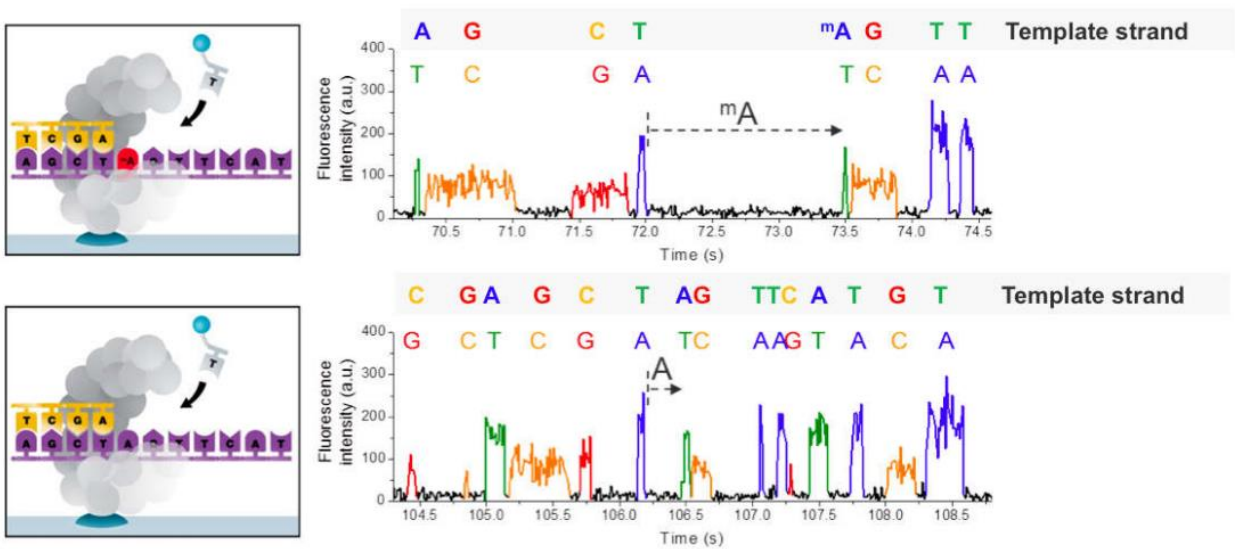


Fig. 3. A fused silica coverslip with zero-mode waveguides arrays. **(A)** The coverslip, with overlying gasket to isolate arrays for individual experiments. Successive increases in scale are shown in **(B)** to **(D)**. A scanning electron microscope image of an individual waveguide is shown in (D). The bright spots in (C) correspond to defects in the metal film. The large bright pattern in the upper right corner is a coded orientation marker.



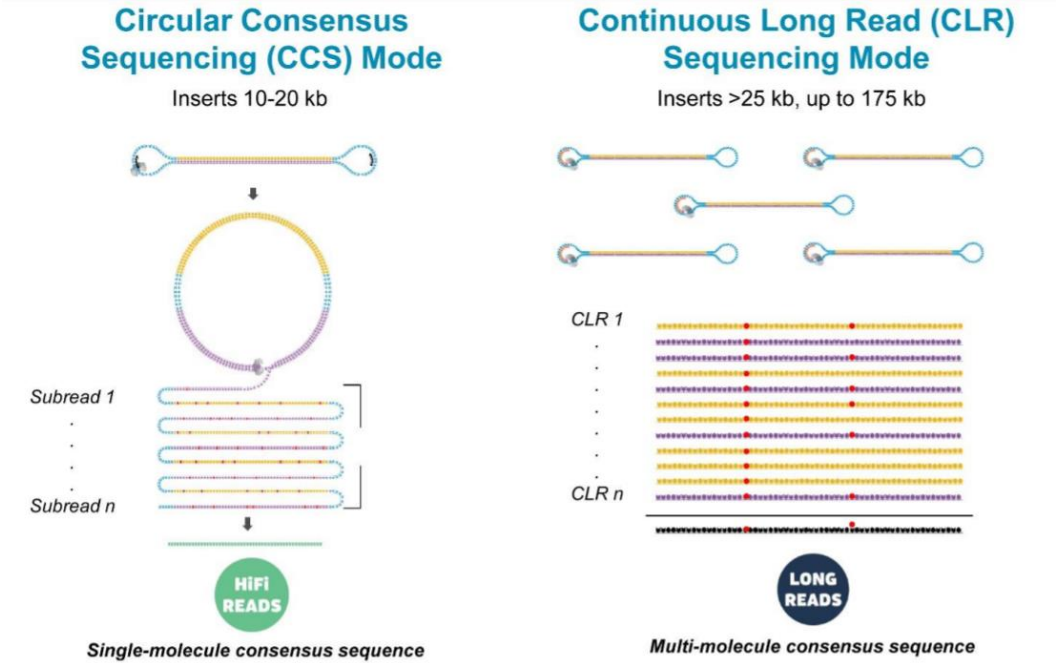

A single molecule of DNA is immobilized in each ZMW
+ Phospholinked nucleotides
As anchored polymerases incorporate labeled bases, light is emitted
Light Intensity
A
C
T
G
Time
Directly detect DNA modifications during sequencing
Nucleotide incorporation kinetics are measured in real time

- Interpulse duration (IPD) can be used to detect modification of Nts, e.g. methylation
- The presence of methylation results in a delayed incorportaion of the corresponding nucleotide, with a delay proportional to the modification

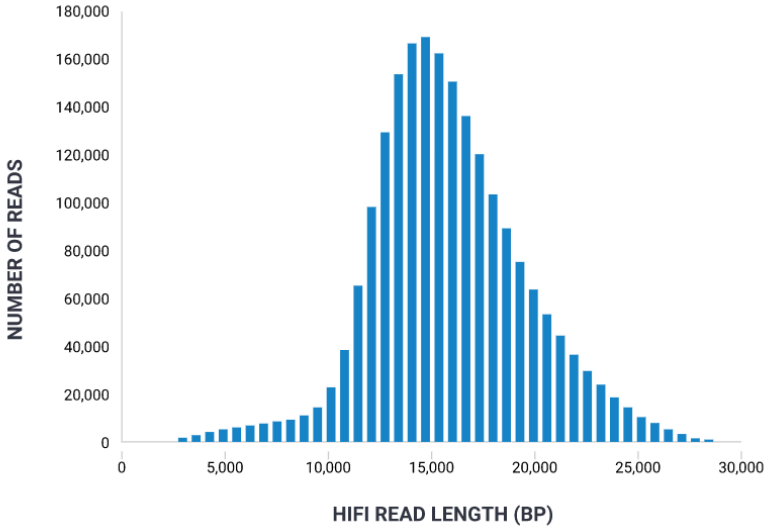


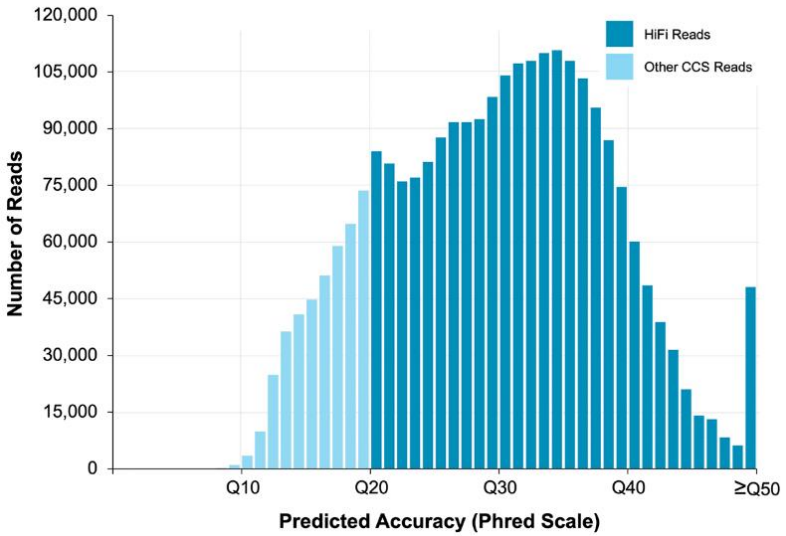
Two modes of operation:

- Circular consensus sequencing (CSS), in which the DNA fragment is read multiple times to improve accuracy (akak HiFi mode)
- Continuous long read sequencing (CLR) to get reads as long as possible (with lesser accuracy)

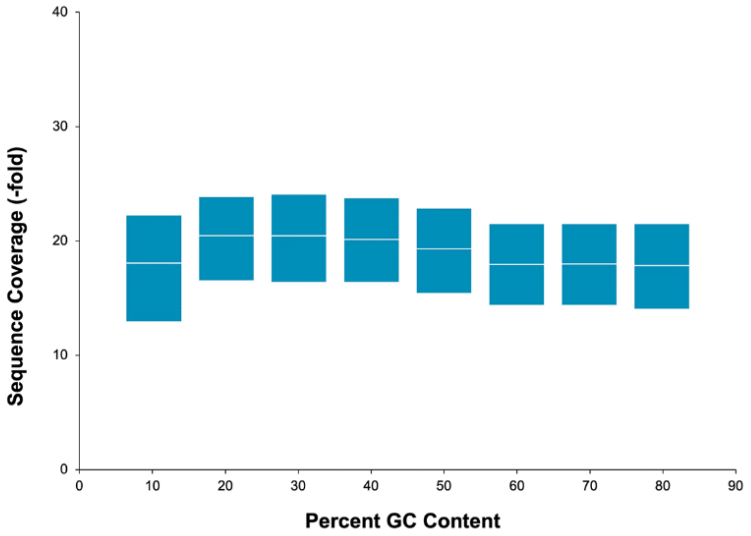


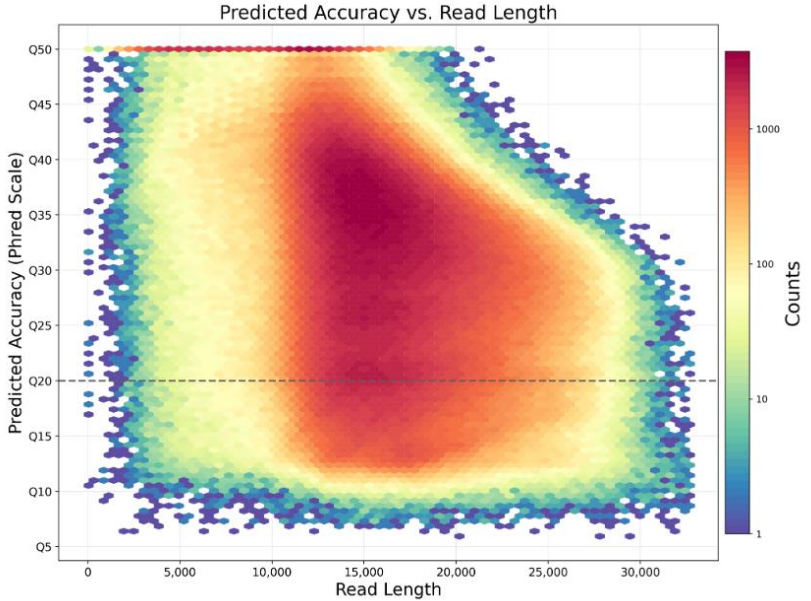
[The machine]





Long reads, > 99.999% accuracy





No GC bias, high single-molecule accuracy