Hum Mol Genet. 2010 Oct 15;19(R2):R227-40. Epub 2010 Sep 21.

A window into third-generation sequencing

Eric E. Schadt*, Steve Turner and Andrew Kasarskis
Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025, USA
Received September 15, 2010; Revised and Accepted September 17, 2010

First- and second-generation sequencing technologies have led the way in revolutionizing the field of genomics and beyond, motivating an astonishing number of scientific advances, including enabling a more complete understanding of whole genome sequences and the information encoded therein, a more complete characterization of the methylome and transcriptome and a better understanding of interactions between proteins and DNA. Nevertheless, there are sequencing applications and aspects of genome biology that are presently beyond the reach of current sequencing technologies, leaving fertile ground for additional innovation in this space. In this review, we describe a new generation of single-molecule sequencing technologies (third generation sequencing) that is emerging to fill this space, with the potential for dramatically longer read lengths, shorter time to result and lower overall cost.

SEQUENCING=

SAMPLE PREPARATION

+ SEQUENCING

+ RE-ASSAMBLY

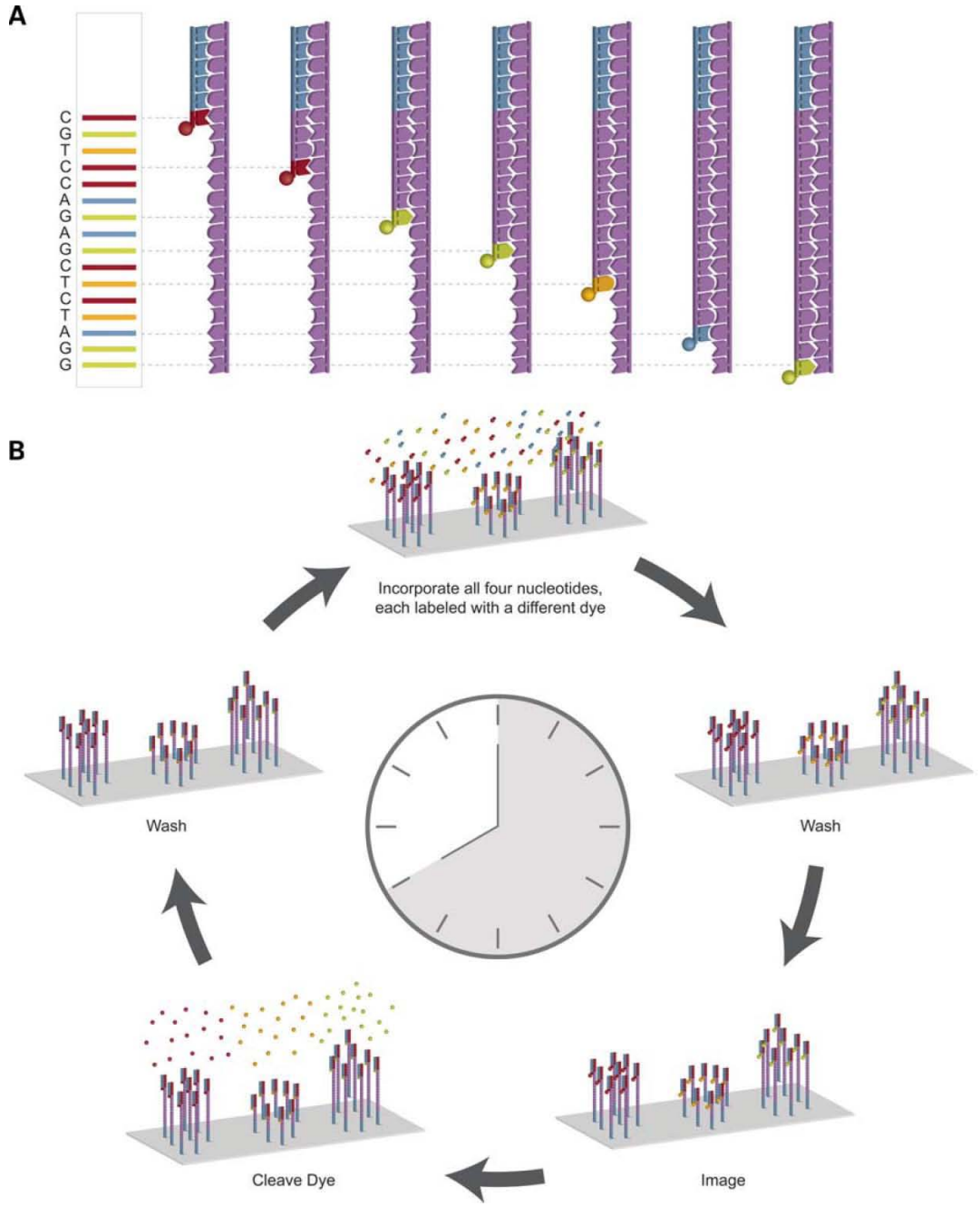


Figure 1. *How previous generation DNA-sequencing systems work.*

*(**A**) A modern implementation of Sanger sequencing is shown to illustrate differential labeling and use of terminator chemistry followed by size separation to resolve the sequence.*

*(**B**) The Illumina sequencing process is shown to illustrate the wash-and-scan paradigm common to second-generation DNA-sequencing technologies.*

The Nobel Prize in Chemistry 1980

"for his fundamental studies of the biochemistry of nucleic acids, with particular regard to recombinant-DNA"

"for their contributions concerning the determination of base sequences in nucleic acids"

Paul Berg

1/2 of the prize

USA

Stanford University
Stanford, CA, USA

Walter Gilbert

1/4 of the prize

USA

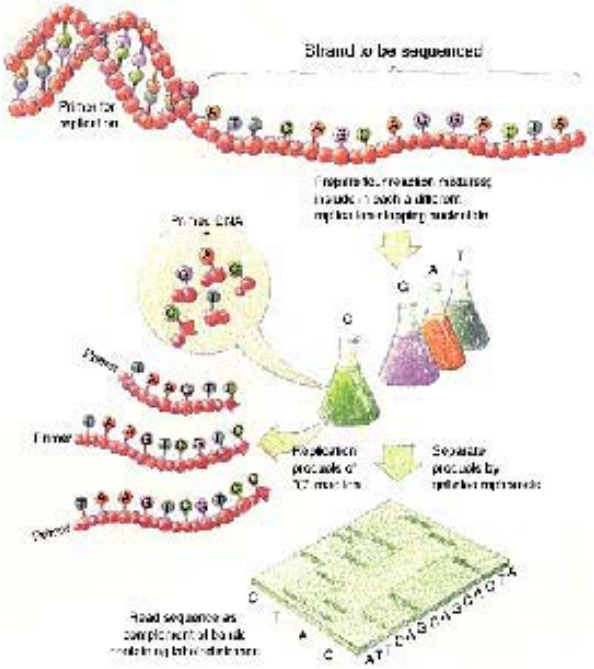
Harvard University,
Biological Laboratories
Cambridge, MA, USA

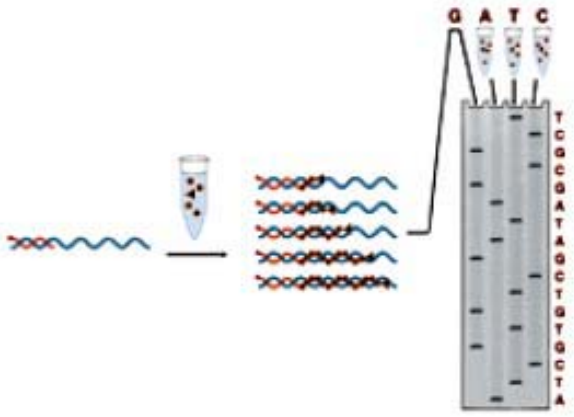
Frederick Sanger

1/4 of the prize

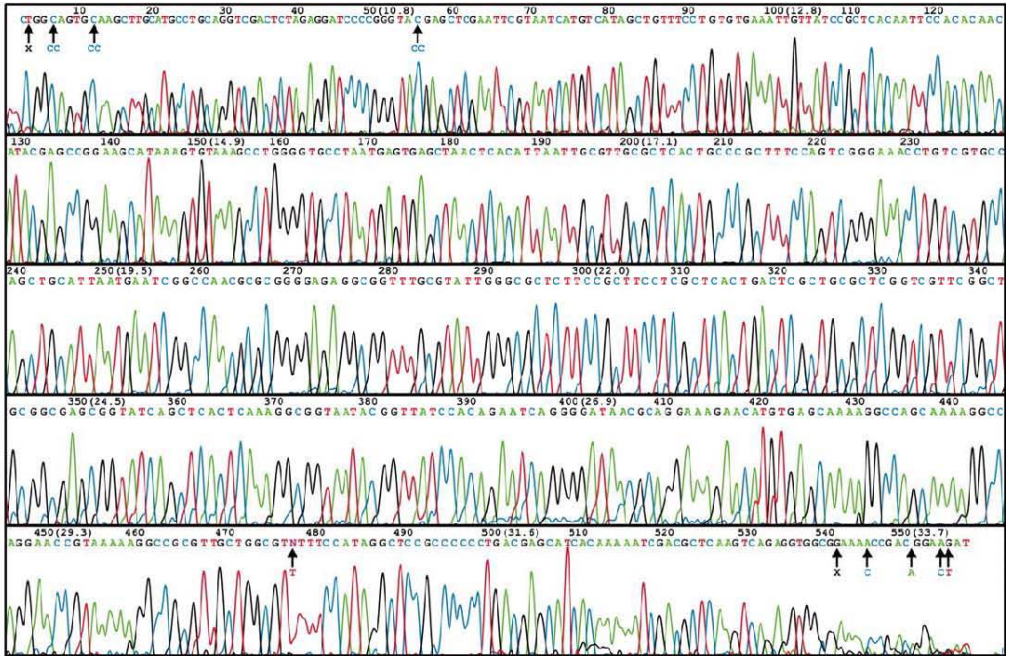
United Kingdom

MRC Laboratory of
Molecular Biology
Cambridge, United





ABI 3730 XL DNA Sequencer



96/384 DNA sequencing in 2 hrs, approximately 600-1000 readable bps per run.

1-4 MB bps/day

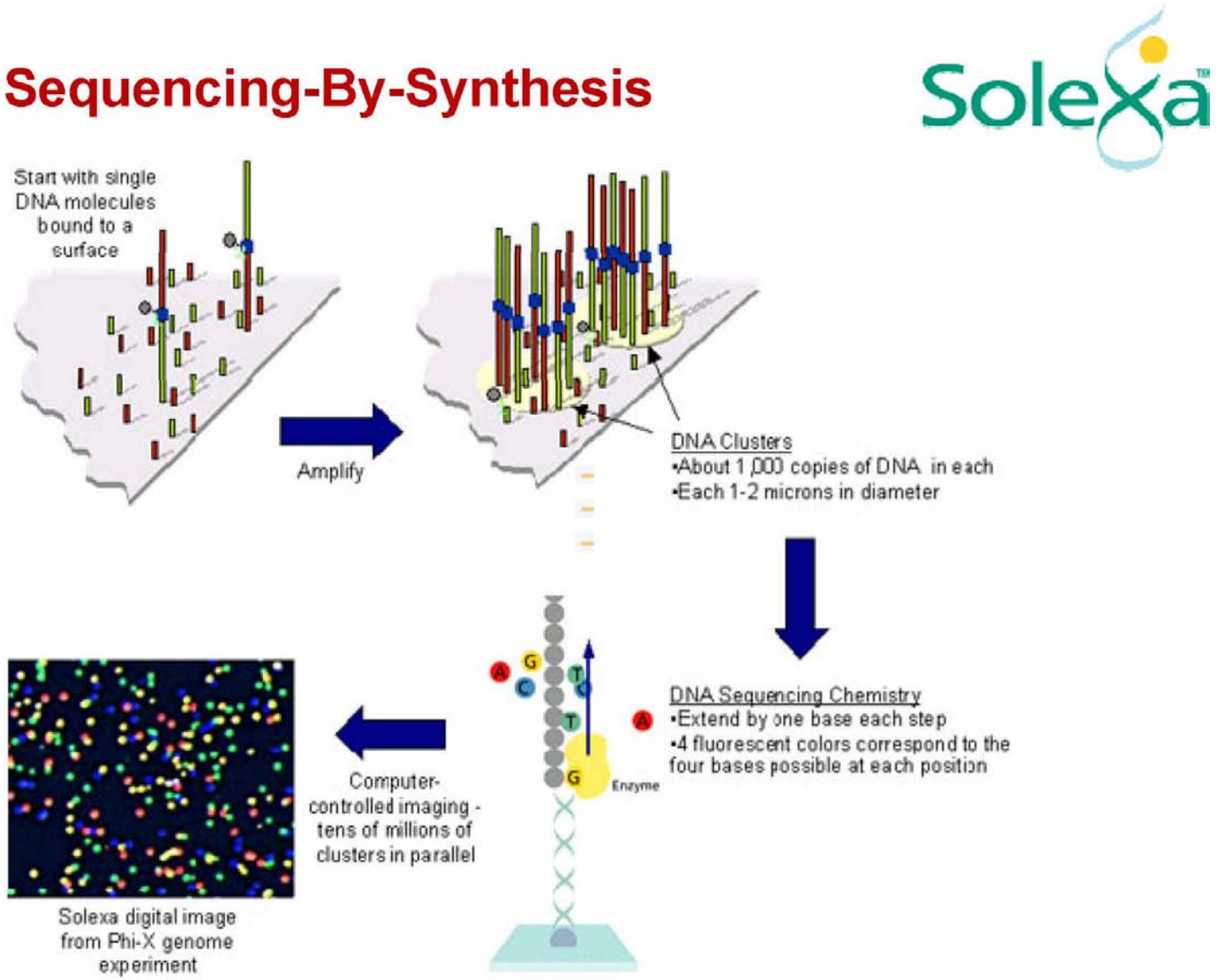
*A human genome of **3GB need 750 days to finish 1X coverage***

Second Generation Sequencing (SGS)

One of the hallmark features of the SGS technologies is their massive throughput at a modest cost, with hundreds of gigabases of sequencing now possible in a single run for several thousand dollars

- most SGS are “sequencing by synthesis”(SBS) technologies that rely on PCR to grow clusters of a given DNA template
- attaching the clusters of DNA templates to a solid surface
- imaged as the clusters are sequenced by synthesis in a phased approach

Illumina's sequencing by synthesis



Illumina's sequencing by synthesis (SBS) technology

http://www.illumina.com/technology/sequencing_technology.ilmn

Evolution of Sequencing Technology

Sanger dideoxy-sequencing

ABI 3730XL

Massive parallel sequencing

Roche 454 FLX, Illumina Genome Analyzer, Life Technologies SOLiD

Bead-based em-PCR and sequencing by ligation

Dover Systems' Polonator

Massive parallel sequencing and single molecule sequencing

Pacific Biosciences (single-molecule real-time DNA sequencing (SMRT) technology)
Helicos (true single-molecule-sequencing (tSMS) technology)
VisiGen Biotechnologies (real-time, single-molecule sequencing fluorescence resonance energy transfer (FRET) technology)

Single molecule sequencing and nanopore technology?

Oxford Nanopore Technologies (label-free, single-molecule sequencing (BASE) technology), Affymetrix, Reveo, Base4innovation, Genome Corp, and Complete Genomics.

Next Generation Sequencing Technology

Roche 454 GS FLX	Pyrosequencing
Illumina SOLEXA	Sequencing by synthesis
Applied Biosystems SOLID	Sequencing by ligation
VisiGen Biotechnologies	Single-molecule sequencing
Helicos BioSciences	Nanopore sequencing
Polonator	Sequencing by ligation

Questions addressed in the review

- How do these next–next-generation technologies work?
- What scales of data generation will be achieved with these new technologies?
- What types of ‘sequencing’ data can be generated? Will they ease analysis issues and/or create new ones?
- And, most importantly, what are the timelines for these technologies to become available?
- Will they really meet the above promises and what do we need to do to prepare???

New generation – “next- next generation sequencing” promises to deliver:

- higher throughput;
- faster turnaround time (e.g. sequencing metazoan genomes at high fold coverage in minutes);
- longer read lengths to enhance *de novo* assembly and enable direct detection of haplotypes and even whole chromosome phasing;
- higher consensus accuracy to enable rare variant detection;
- small amounts of starting material (theoretically only a single molecule may be required for sequencing);
- low cost, where sequencing the human genome at high fold coverage for less than $100 is now a reasonable goal for the community.

Glossary:

third-generation sequencing (TGS);

single-molecule sequencing (SMS) technologies

The most important qualifiers for TGS:

SMS technologies that do not purposefully pause the sequencing reaction after each base incorporation

=>

INCREASING

- sequencing rates
- throughput
- read lengths (facilitating re-assambly)

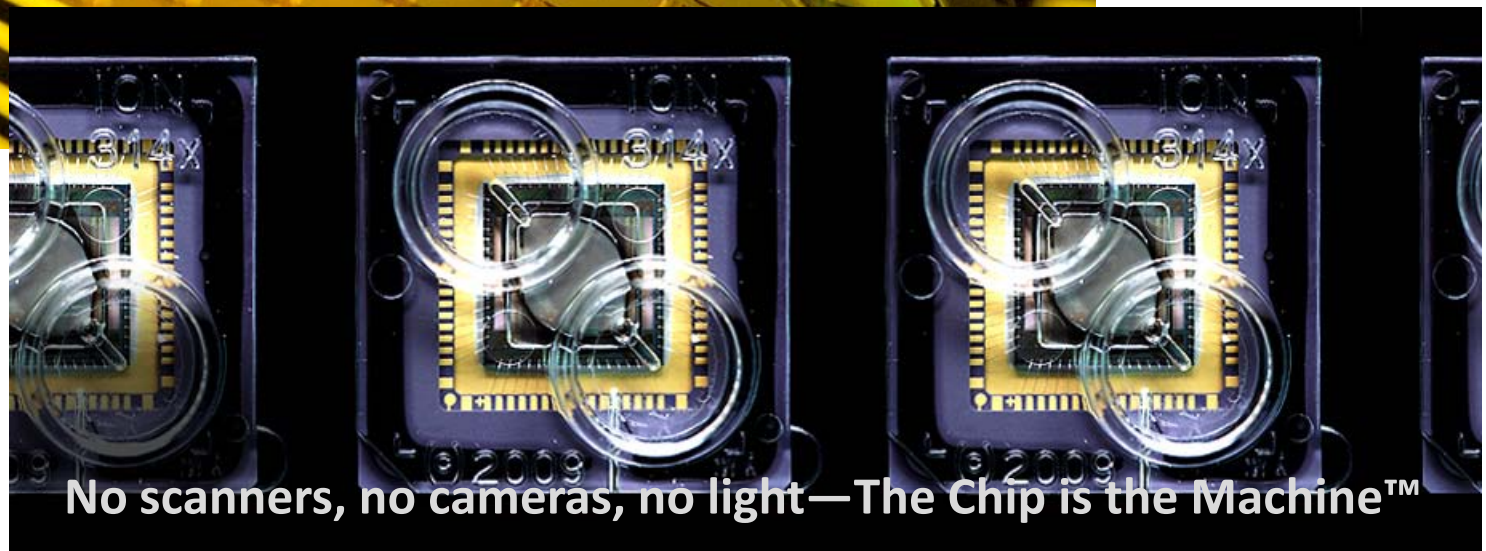
LOWERING

- the complexity of sample preparation
- ultimately decreasing cost

The simplest sequencing chemistry—
natural nucleotides, no enzymatic cascade

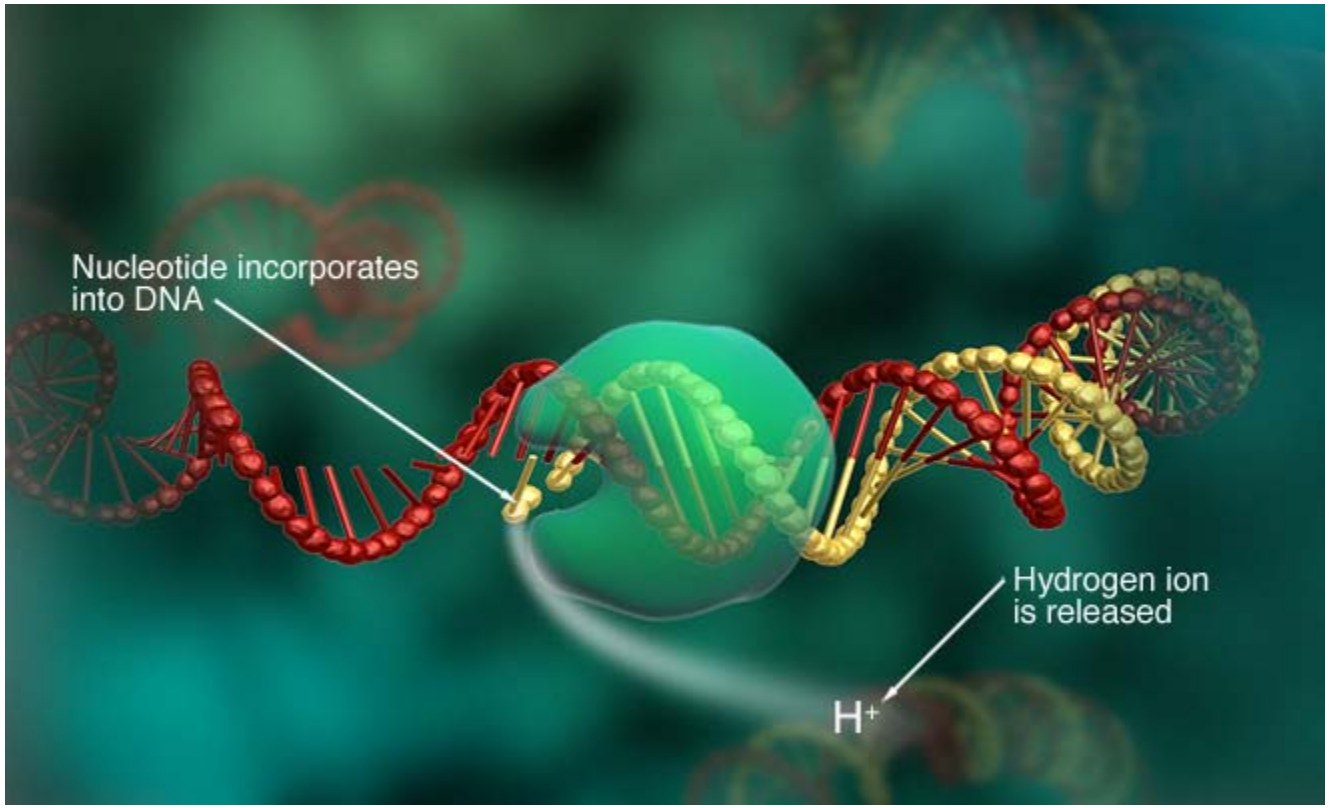
True semiconductor sequencing—one platform, 1000X scalability

ion torrent
by life technologies™



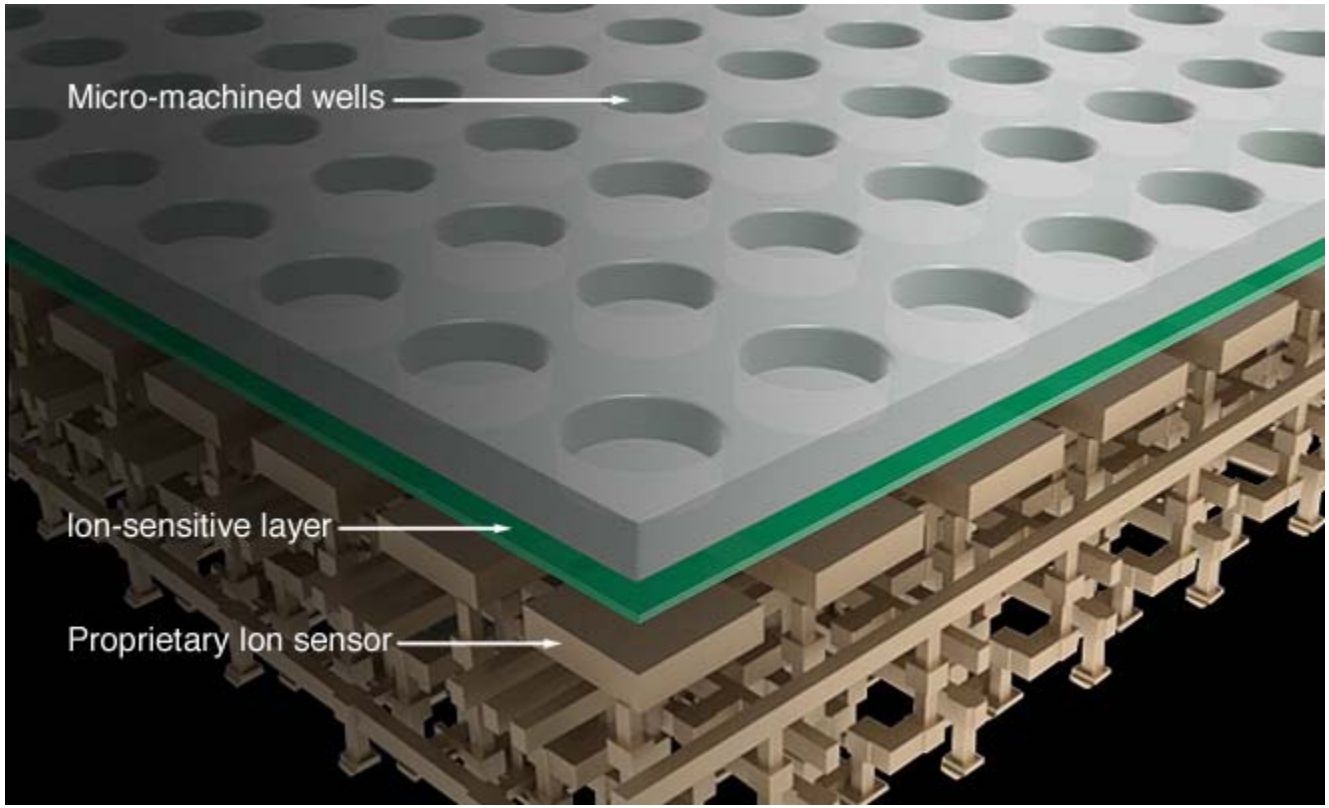
No scanners, no cameras, no light—The Chip is the Machine™

When a nucleotide is incorporated into a strand of DNA by a polymerase, a hydrogen ion is released as a byproduct:



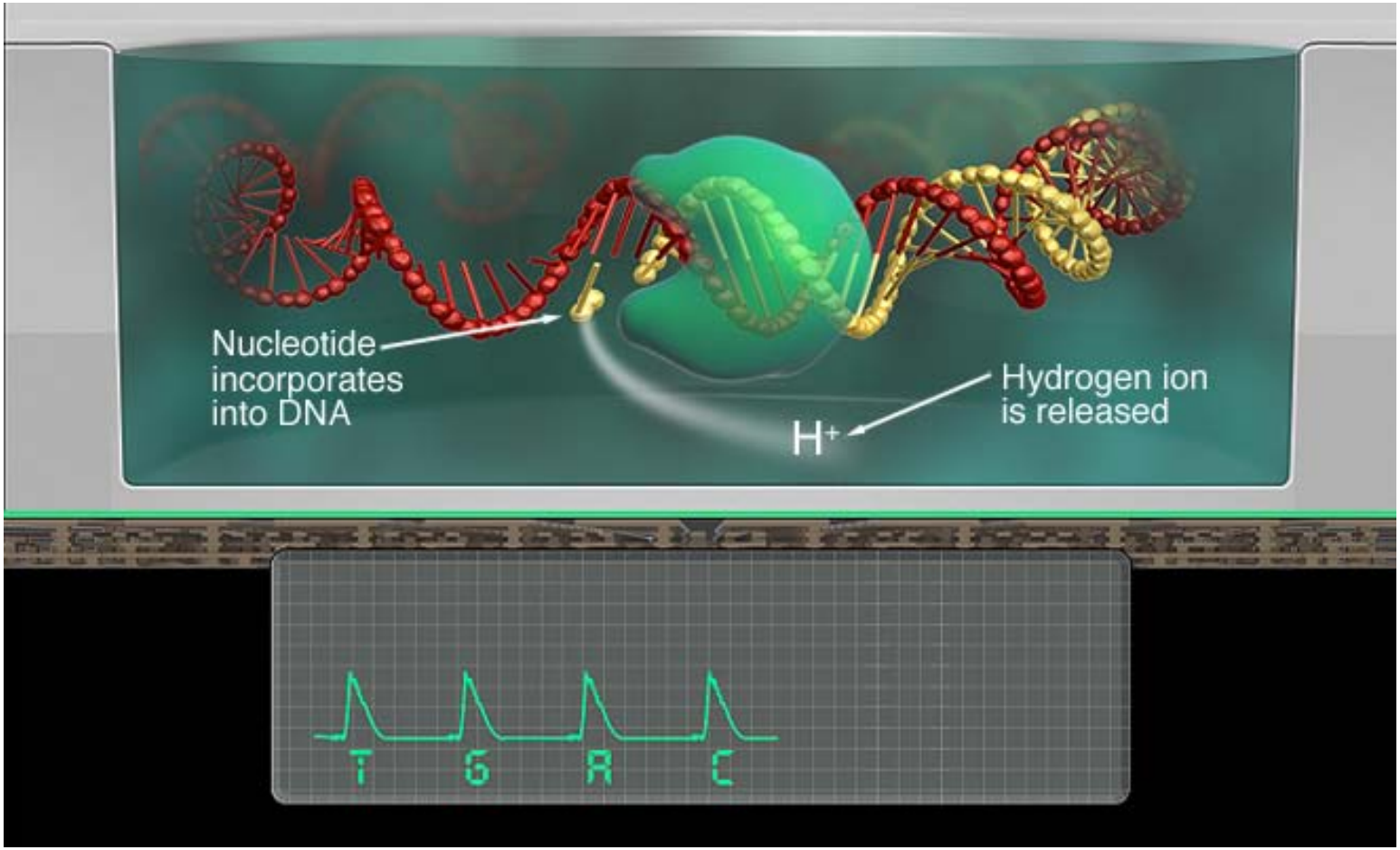
ion torrent
by life technologies

Ion Torrent uses a high-density array of micro-machined wells to perform this biochemical process in a massively parallel way. Each well holds a different DNA template. Beneath the wells is an ion-sensitive layer and beneath that a proprietary Ion sensor.



ion torrent
by life technologies™

Here's how the technology is used to call a base:



ion torrent
by life technologies™

Helicos Genetic Analysis Platform:

- The Helicos® Genetic Analysis System - which includes the HeliScope™ Analysis Engine, the HeliScope™ Sample Loader and the HeliScope™ Single Molecule Sequencer –

"The world's first DNA Microscope instrument" - performs the tSMS chemistry and directly analyzes images of single molecules, producing accurate sequences of billions of templates at a time.

Helicos BioSciences Corporation

How Helicos tSMS Works?

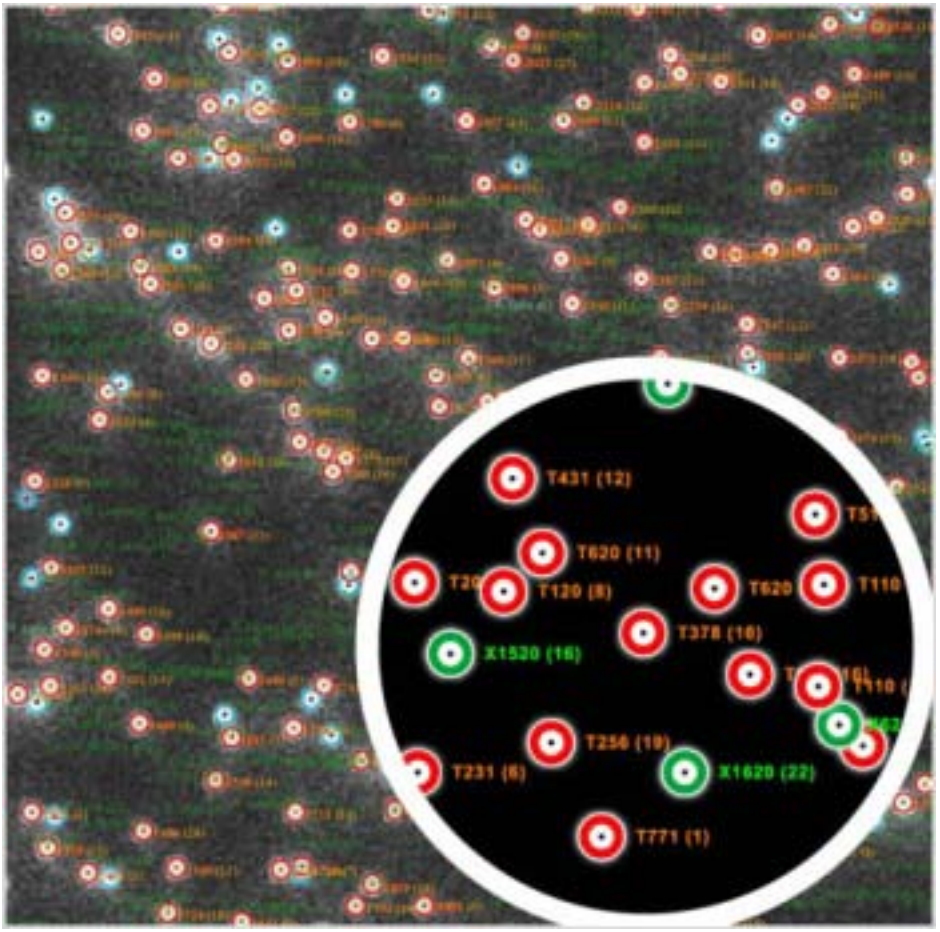
- Within two flow cells, billions of single molecules of sample DNA are captured on an application-specific proprietary surface. These captured strands serve as templates for the sequencing-by-synthesis process:
- Polymerase and one fluorescently labeled nucleotide (C, G, A or T) are added.
- The polymerase catalyzes the sequence-specific incorporation of fluorescent nucleotides into nascent complementary strands on all the templates.
- After a wash step, which removes all free nucleotides, the incorporated nucleotides are imaged and their positions recorded.
- The fluorescent group is removed in a highly efficient cleavage process, leaving behind the incorporated nucleotide.
- The process continues through each of the other three bases.
- Multiple four-base cycles result in complementary strands greater than 25 bases in length synthesized on billions of templates—providing a greater than 25-base read from each of those individual templates.

Image Analysis

Helicos BioSciences Corporation

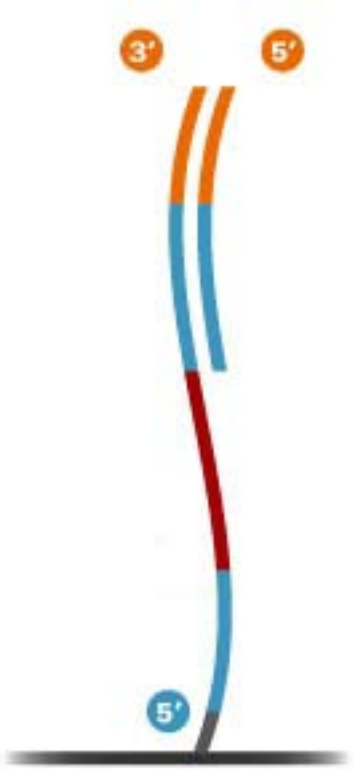
Using the latest in high-performance-computing technology and state of the art optics, the system's image analysis software, located on the HeliScope™ Analysis Engine, identifies and extracts millions of nucleotide base incorporations from the 48 images produced by the HeliScope™ Single Molecule Sequencer each second - **in near real time**.

Features are identified according to stringent nucleotide incorporation criteria and catalogued in an “object table” for subsequent base calling and sequence formation.



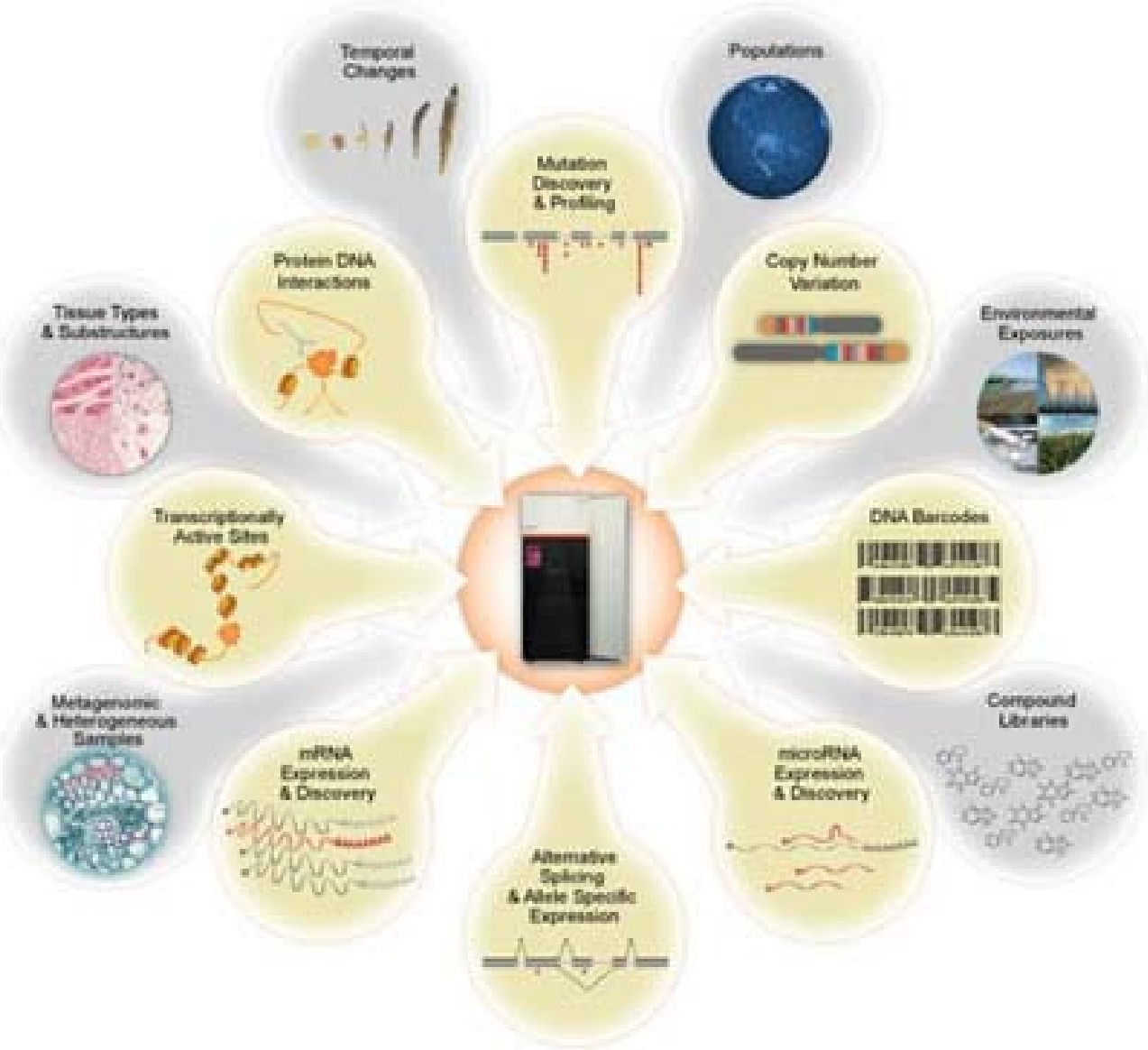
Prep-Free Paired-End Reads from Single Molecules

Helicos BioSciences Corporation



- The genomic sample is digested/sheared.
- Fragments within a desired size range are selected.
- An adaptor sequence, containing a universal priming site, is ligated to the 5’ ends of the fragments.
- Poly(A) tails are generated on the 3’ ends of the fragments.
- The sample is now ready for paired-end sequencing.

Versatility of the potential applications:



SMS technologies can roughly be binned into three different categories:

- SBS technologies in which single molecules of DNA polymerase are observed as they synthesize a single molecule of DNA;
- nanopore-sequencing technologies in which single molecules of DNA are threaded through a nano-pore or positioned in the vicinity of a nanopore, and individual bases are detected as they pass through the nanopore;
- direct imaging of individual DNA molecules using advanced microscopy techniques.

Single-molecule real-time sequencing (SMRT) by PACIFIC BIOSCIENCES

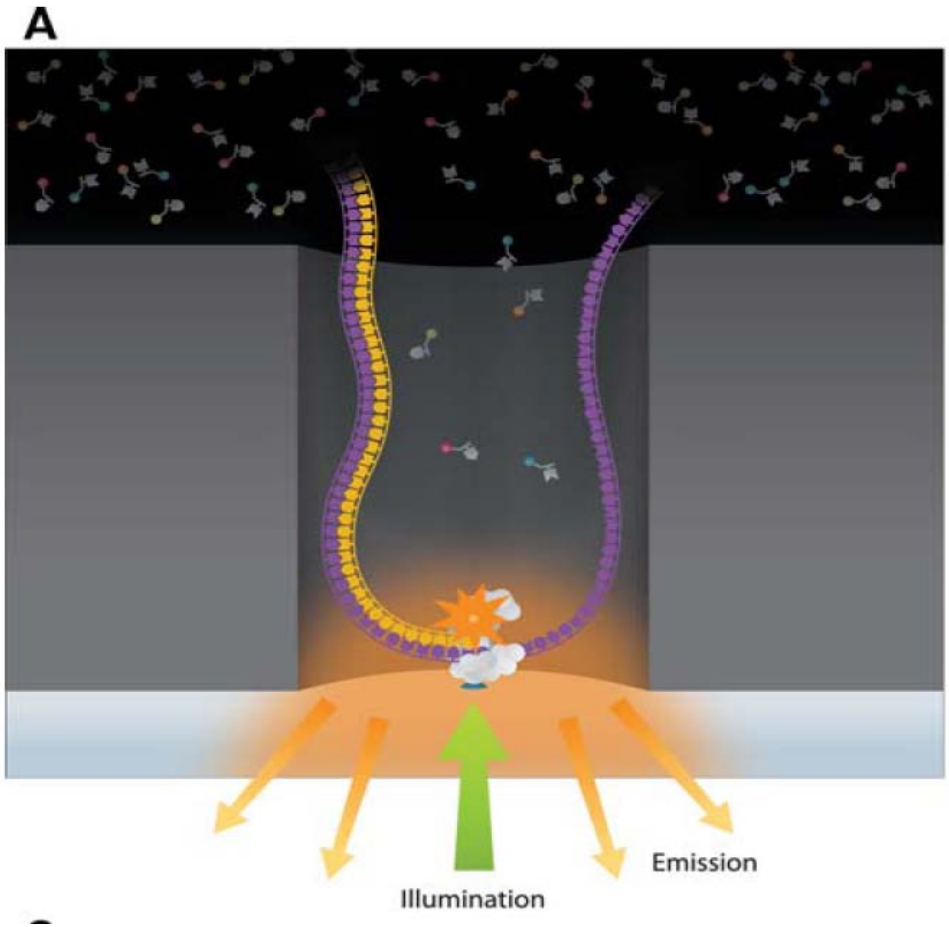
Obstacles to be overcome for direct observation of DNA synthesis:

- confining the enzyme **to a very small observation volume**
- labeling the nucleotides to be incorporated in the synthesis process such that the **dye–nucleotide linker is cleaved after completion of the incorporation process** and a nonmodified strand of DNA remains for continued synthesis

Technical approach:

- Using a **ZMV** , i.e. a zero – mode waveguide
- The SMRT sequencing approach **attaches the fluorescent dye to the phosphate chain of the nucleotide** rather than to the base. Upon incorporation of a phospholinked nucleotide, the DNA polymerase naturally frees the dye molecule from the nucleotide.

Pacific Biosciences technology for direct observation of DNA synthesis on single DNA molecules in real time.



PACIFIC BIOSCIENCES

All the smart things about the SMRT

- The SMRT sequencing platform requires minimal amounts of reagent and sample preparation to carry out a run
- There are no time-consuming scanning and washing steps, enabling time to result in a matter of minutes as opposed to days .
- SMRT sequencing does not require routine PCR amplification needed by most SGS systems, thereby avoiding systematic amplification bias.
- Because the processivity of the DNA polymerase is leveraged, SMRT sequencing realizes longer read lengths than any other technology at present, having the potential to produce average read lengths 1000 bp and maximum read lengths in excess of 10 000 bp, enabling *de novo* assembly, direct detection of haplotypes and even providing for the possibility of phasing entire chromosomes.
- However, with SMRT sequencing, the sample prep-aration consists of fragmenting the DNA into desired lengths, blunting the ends, ligating hairpin adaptors and then sequencing

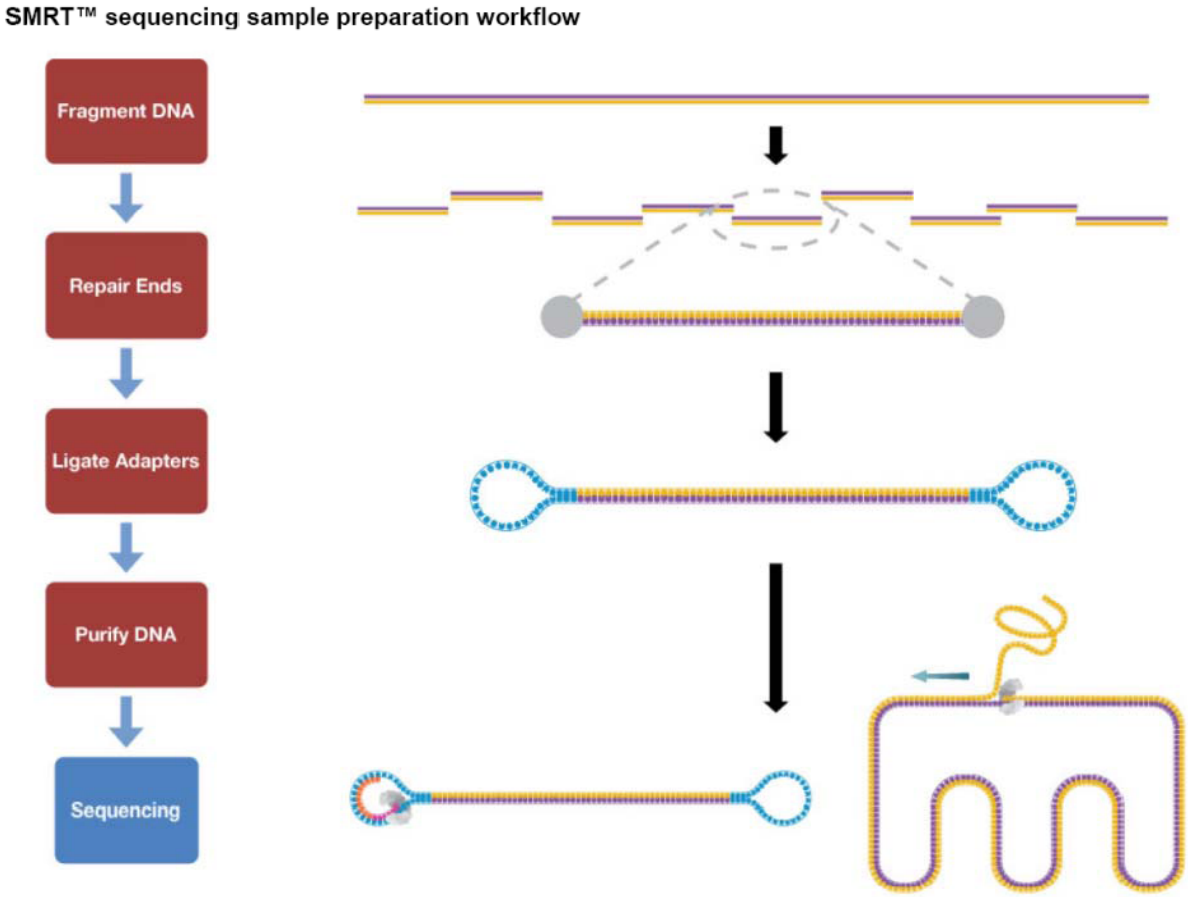


Figure 17. Sample Prep Workflow.
The input sample is first fragmented to the desired size. The ends are repaired and the hairpin structures are ligated to the ends of each fragment. A size selection and purification step selects those fragments with the adaptors attached to both ends. The SMRTbell templates then can go through the sequencing reaction. A strand displacing polymerase enzyme opens the SMRTbell into a circular template and can generate independent reads, both forward and reverse of the same DNA molecule. The quality score increases linearly with the number of times the molecule is sequenced.

Single Molecule Real-Time (SMRT) Sequencing

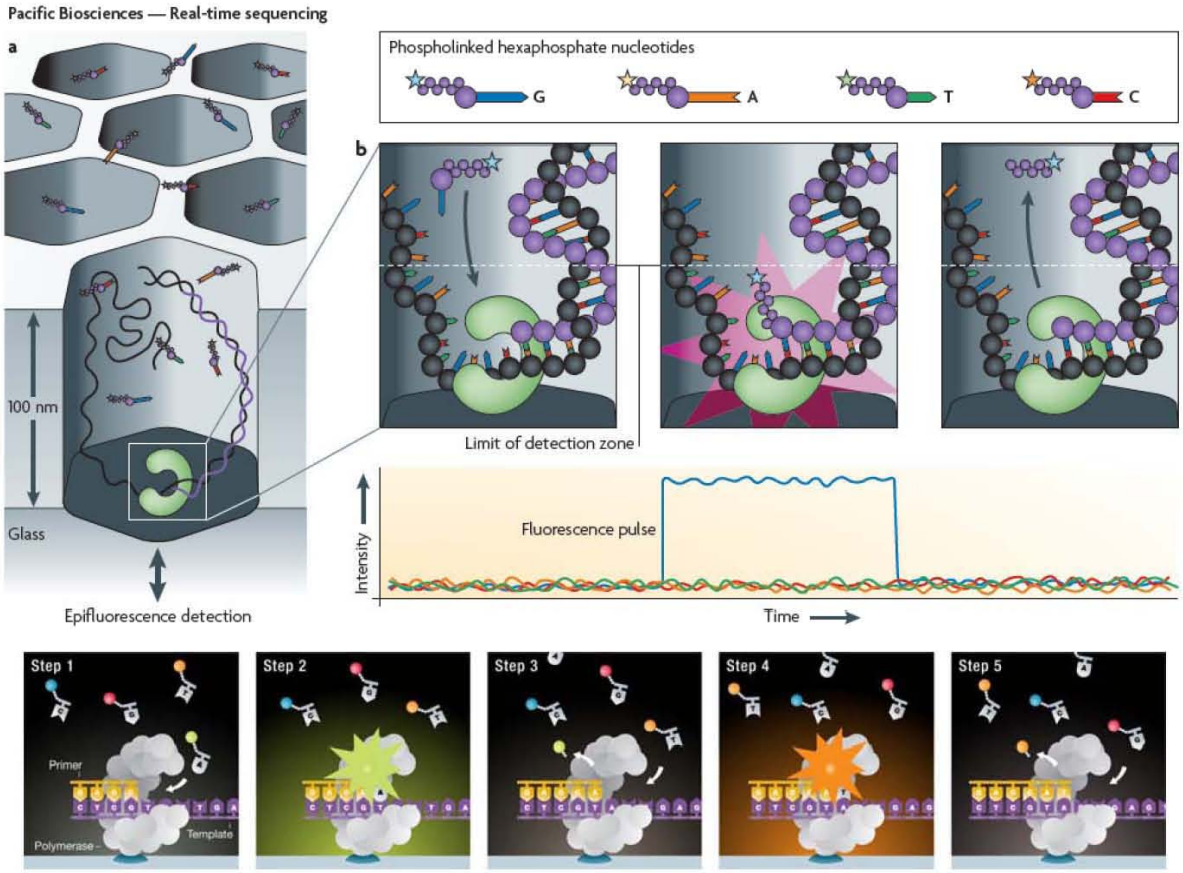
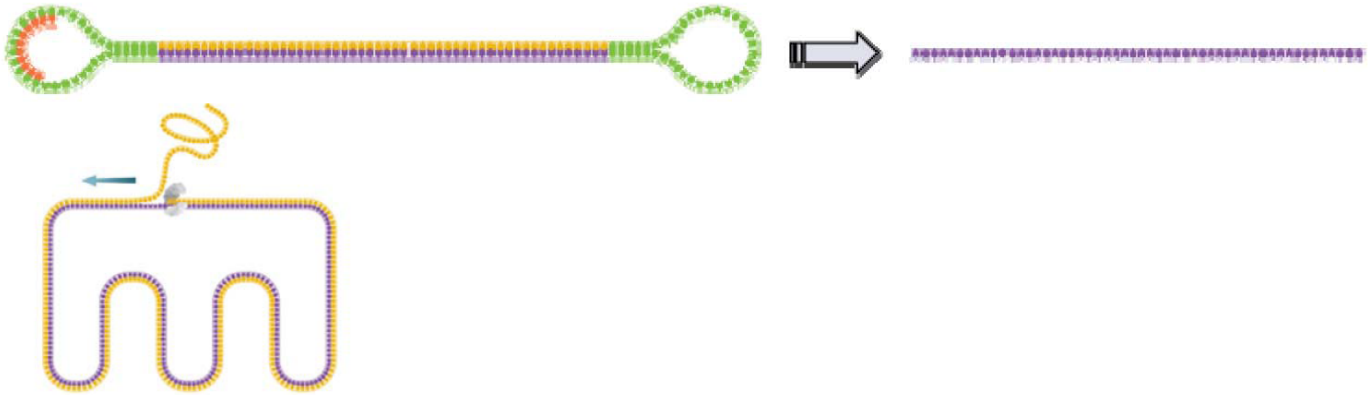


Figure 10. Processive Synthesis with Phospholinked Nucleotides.

Step 1: Fluorescent phospholinked labeled nucleotides are introduced into the ZMW.
Step 2: The base being incorporated is held in the detection volume for tens of milliseconds, producing a bright flash of light.
Step 3: The phosphate chain is cleaved, releasing the attached dye molecule.
Step 4-5: The process repeats.

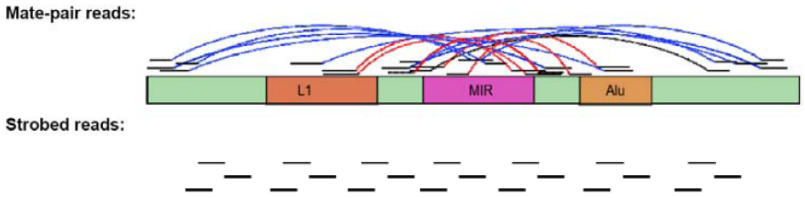
Standard sequencing protocol.

Enzyme processivity enables long readlengths while the speed of synthesis drives fast time to results.



Strobe sequencing protocol.

Strobe sequencing offers greater flexibility and eliminates the need to create multiple libraries of different sizes.



The development potential of the PacBio SMRT:

Extra information can be collected:

- Via the SMRT sequencing process, changes in the kinetics of incorporation associated with chemi-cal modifications to bases, such as methylation, can be detected in the normal course of collecting sequence data.

The instrument itself is capable of many more applications:

- real-time observation of the ribosome as it translated mRNA
- reverse transcriptase for RNA-sequencing applications

Challenges remain:

- Like the Helicos technol-ogy, the raw read error rates can be in excess of 5%, with error rates dominated by insertions and deletions, particularly pro-blematic errors when aligning sequences and assembling genomes. In addition, the throughput of SMRT sequencing will not initially match what can be achieved by SGS.

State of the art

Virtues:

- the SMRT sequencing instrument will consist of an array of 75 000 ZMWs. Each ZMW is capable of containing a DNA polymer-ase loaded with a different strand of DNA sample. As a result, the array enables the potential detection of 75 000 SMS reactions in parallel.

Shortcomings:

- At present, because the DNA polymerase and DNA template to be sequenced are delivered to ZMWs via a random diffusion process, approximately a third of the ZMWs of a given array are active for a given run

Real-time DNA sequencing using fluo-rescence resonance energy transfer (FRET):

- Leading developer: VisiGen Biotechnologies ; acquired by Life Technologies recently

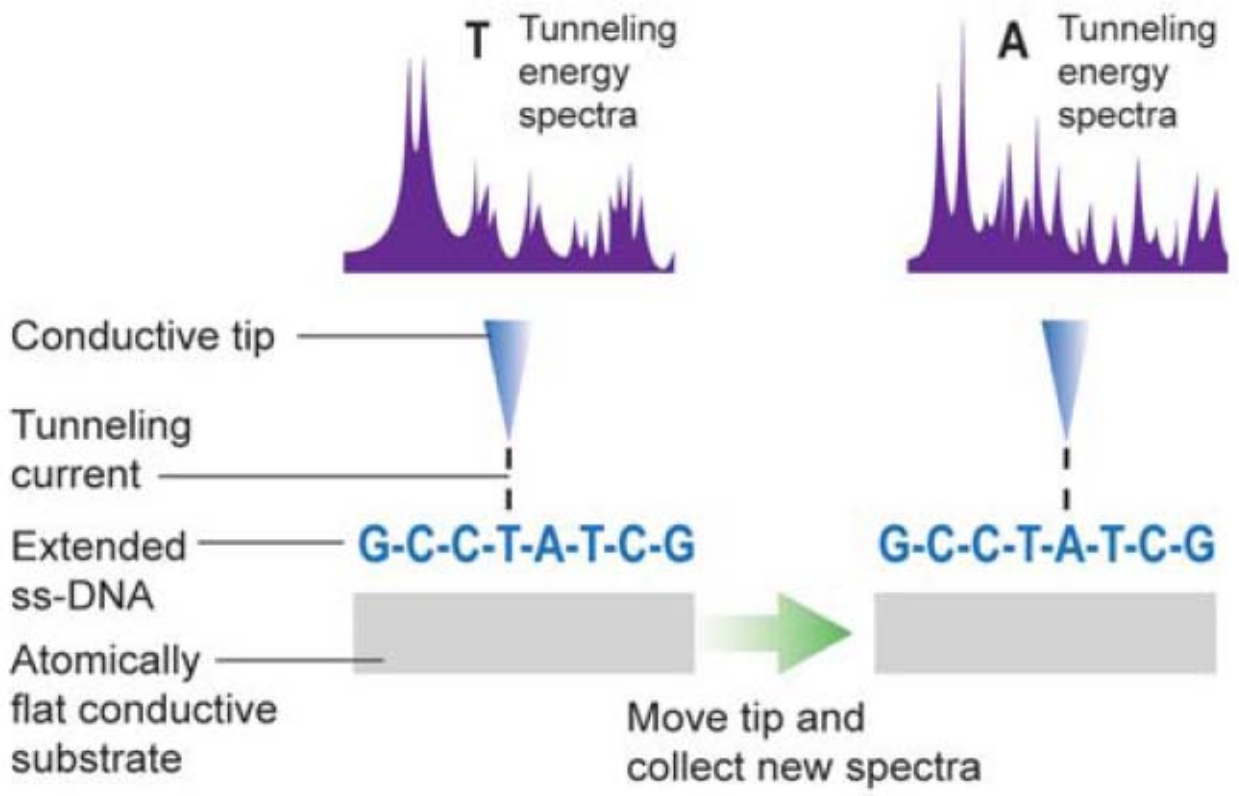
- The DNA polymerase is tagged with a fluorophore that when brought into close proximity to a nucleotide, tagged with an acceptor fluorophore, would emit a fluorscence resonance energy transfer (FRET) signal.
- After incorporation, the fluorophore label on the nucleotide can be released.
- Has the poten-ial to move at millions of bases per second, given potential for high multiplex

Tunneling and transmission-electron-microscopy-based approaches for DNA sequencing

- **Halcyon Molecular** is pioneering an SMS approach using transmission electron microscopy (TEM) to directly image and chemically detect atoms that would uniquely identify the nucleotides comprising a DNA template. The approach being pursued has been shown to reliably detect atoms in a non-periodic material on a planar surface, using annular dark-field imaging in an aberration-corrected scanning TEM.
- **The promise:** megabase single molecule reads in milliseconds!

- **ZS genetics** is developing another TEM-based DNA sequencing instrument to directly image the sequence. With this technology, labeled atoms within the nucleotides of DNA are imaged using a high-resolution (subangstrom) electron microscope, where individual bases are detected and identified based on their size and intensity differences between the different labeled bases.
- **No publications as yet**, but claims that the technology is capable of producing 10 000– 20 000 base reads at a rate of 1.7 billion bases per day

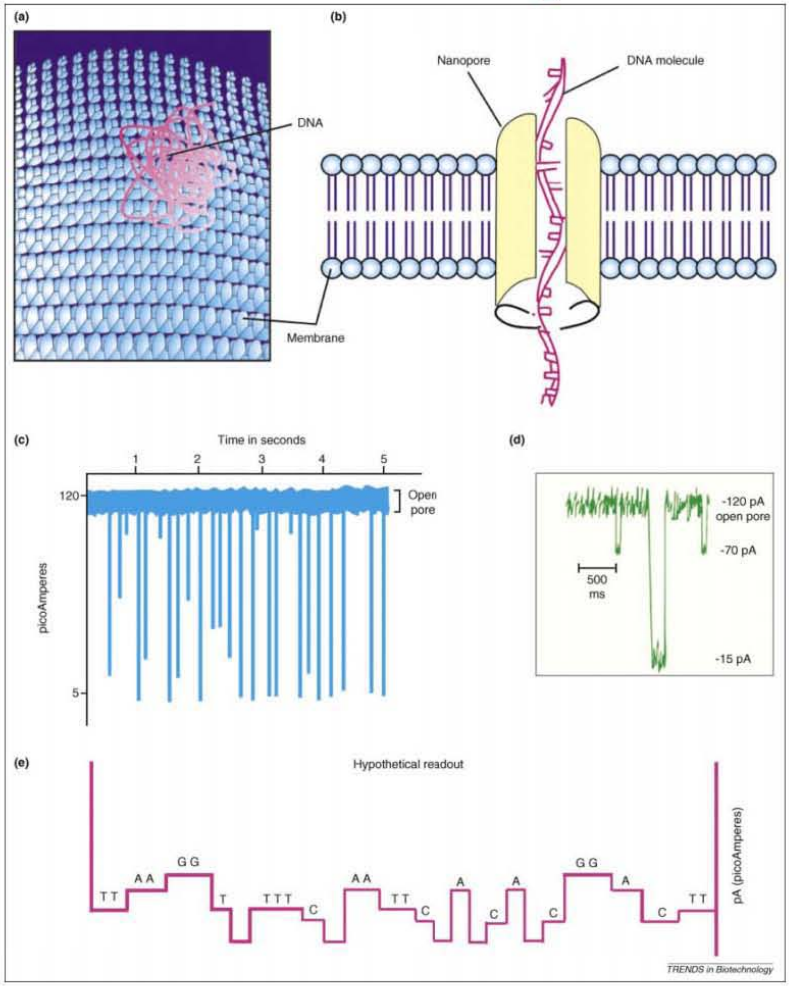
Direct imaging of DNA sequences using scanning tunneling microscope tips.



(B) Several companies seek to sequence DNA by direct inspection using electron microscopy similar to the **Reveo technology** pictured here, in which an ssDNA molecule is first stretched and then examined by STM.

DNA sequencing with nanopores

Single Molecule Nanopore Sequencing

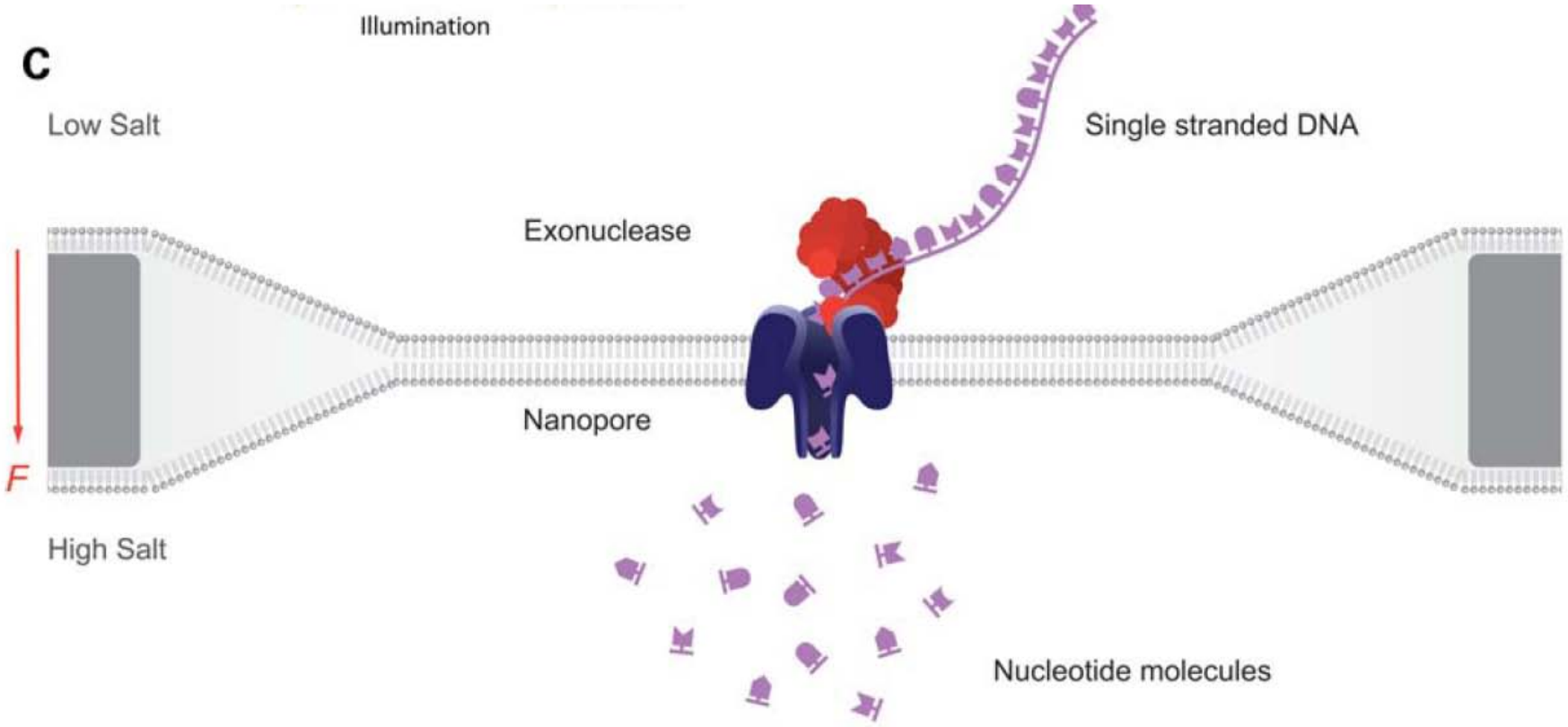


Nanopore sequencing

Nucleic acids driven through a nanopore.

Differences in conductance of pore provide readout.

Direct, electrical detection of single DNA molecules.



(C) **Oxford Nanopore technology** for measuring translocation of nucleotides cleaved from a DNA molecule across a pore, driven by the force of differential ion concentrations across the membrane.

Other nanopore technologies:

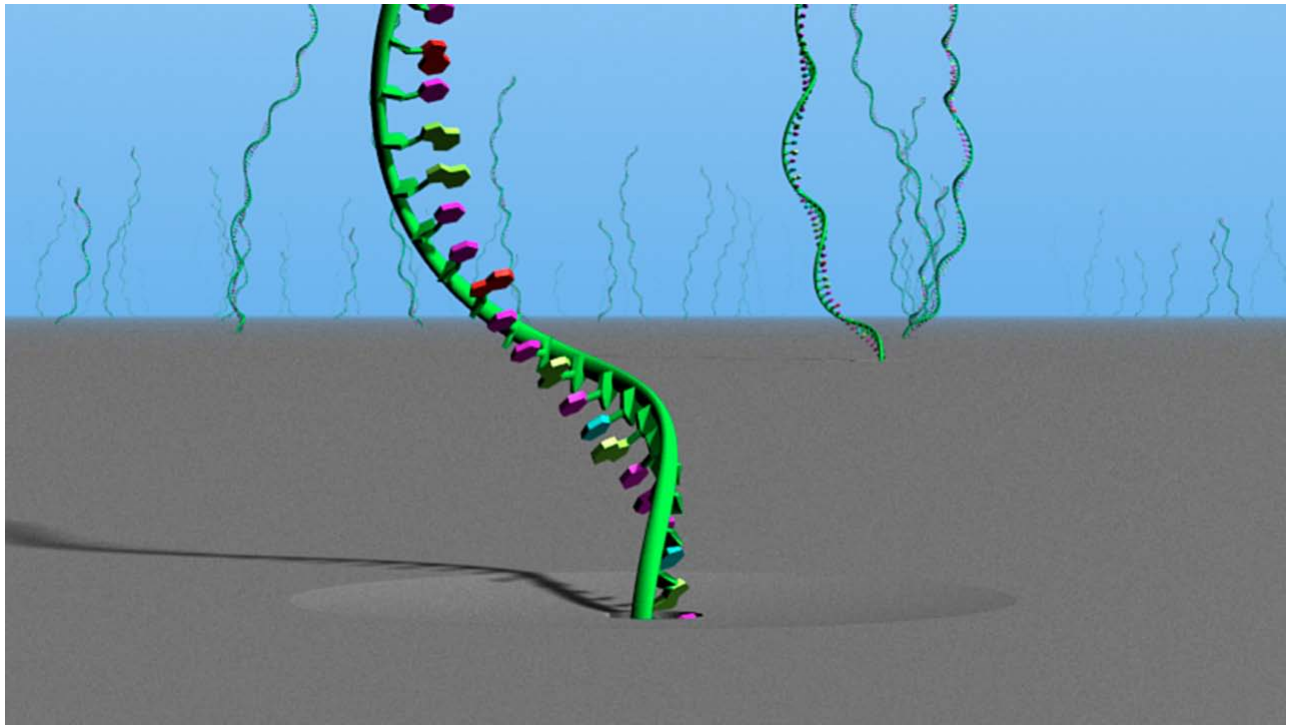
- **Nanopore DNA sequencing with MspA:**

The *Mycobacterium smegmatis* Porin A (MspA) protein, which has a shorter blockade region and thus a better resolution, is used as the pore and the effect of a linear molecule of single-stranded DNA (ssDNA) on the current transiting the pore is measured

- **Nanopore sequencing with optical readout:**

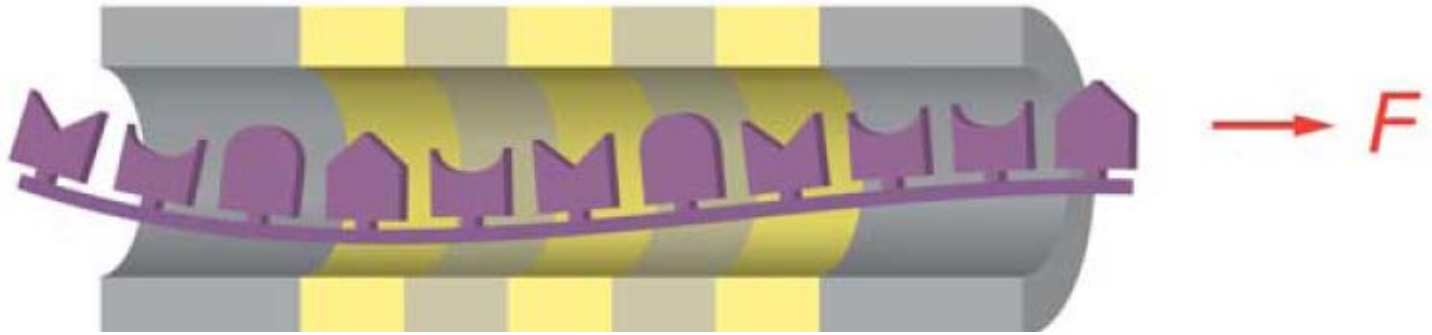
In this approach, the contrast between the four bases is first increased off-line through a biochemical process that converts each base in the DNA into a specific, ordered pair of concatenated oligo-nucleotides. Subsequently, two different fluorescently labeled molecular beacons are hybridized to the converted DNA. The beacons are then sequentially unzipped from the DNA molecules as they are translocated through a nanopore. Each unzipping event unquenches a new fluorophore, resulting in a series of dual-color fluorescence pulses that are detected by a high-speed CCD camera with a conventional total internal reflection fluorescence microscopy setup.

DNA sequencing (solid state nanopores)



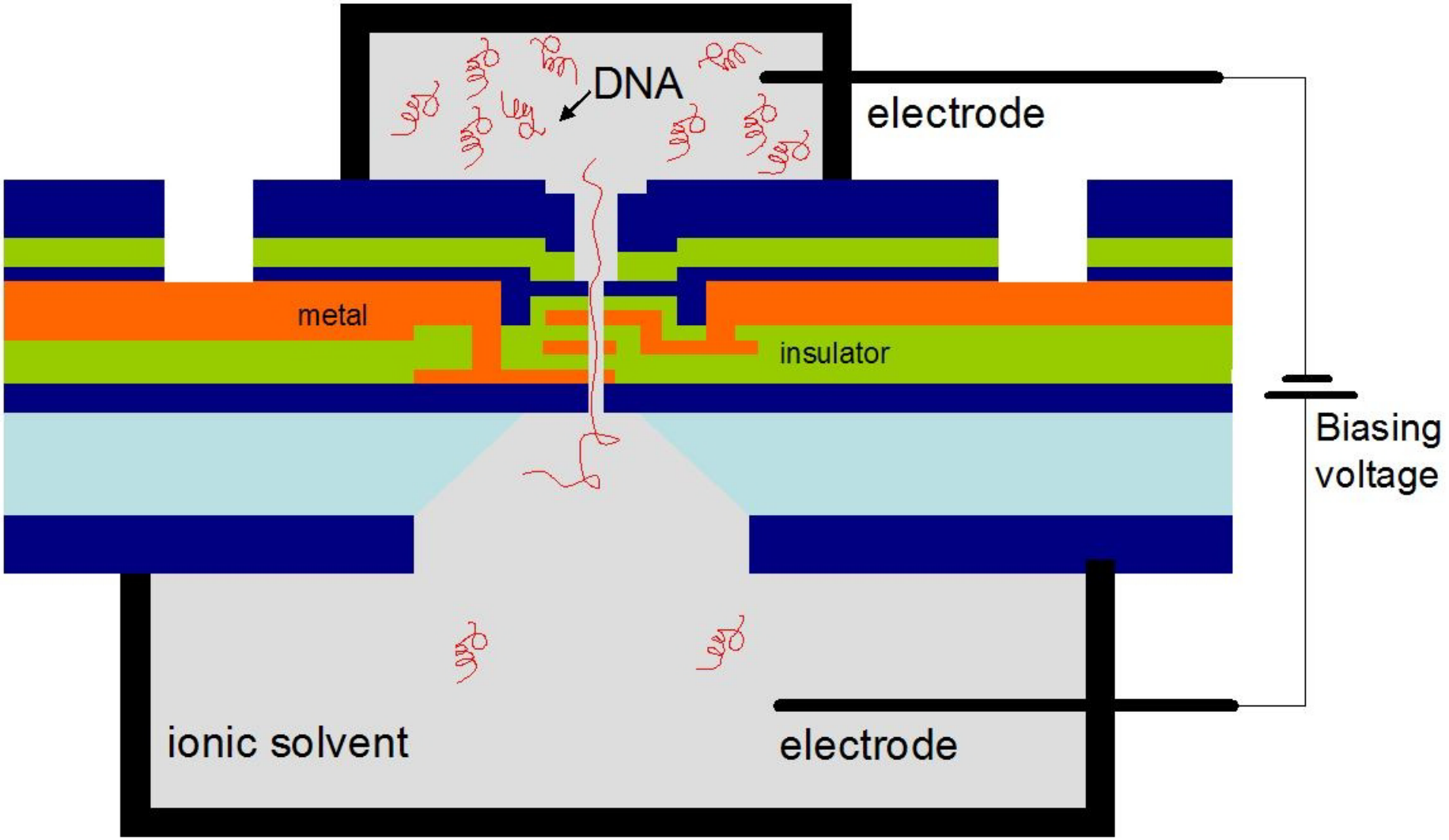
- **A future generation of nanopore technology is 'solid-state' nanopores.** These are man-made holes in synthetic materials, for example silicon nitride. As fabrication methods improve towards the ability to consistently manufacture nano-scale pores in thin materials, there is potential to further improve cost and yield of nanopore sequencing.

Transistor-mediated DNA sequencing.



IBM's DNA transistor technology reads individual bases of ssDNA molecules as they pass through a narrow aperture based on the unique electronic signature of each indi-vidual nucleo-tide. Gold bands represent metal and gray bands dielectric layers of the transistor.

In an effort to build a nanoscale DNA sequencer, IBM scientists are drilling nano-sized holes in computer-like chips and passing DNA strands through them in order to read the information contained within their genetic code.



TGS informatics opportunities

- Most of the TGS technologies discussed address (or have the potential to address) the limitations of SGS technologies with respect to assembly quality, given the read lengths and mate-pair distances in TGS are not only significantly beyond those realized with SGS, but with Sanger sequencing as well. Longer reads can span repeat regions that make assembly difficult and can obviate the need for more complex mate-pair strategies required to scaffold SGS reads.

Its strengths notwithstanding, TGS will come with its own set of challenges:

Because a TGS system by definition assays a single molecule, there is no longer any safety in numbers to minimize raw read errors =>

- The fraction of unlabelled nucleotides present equals the percentage of the rate of deletion artefacts in the raw data
- If a base fails to progress through a nanopore or a DNA transistor as intended and gets counted twice, there will be an insertion in the raw data.

The increased information content will demand new types of mathematical models and algorithms to get the most from the data:

- The kinetic information now provides more data on the nature of the template and this valuable information should be retained during raw data processing
- Error structure and distribution will be different
- The phasing correction and chastity filtering are no longer necessary, given the asynchronous nature of SMS.
- A number of issues arise that lead to uncertainty around the number and identity of bases read for a given template

Challenges:

- significant stochastic components of SMS complicate the relationship between observations and interpretations of those observations => provide for alternative base calls at any given position or identify more general structural variants?
- hudge amount of data generated => how to collapse it, while keeping the valuable extra information?

CONCLUSION/PERSPECTIVE

- Will TGS bring the true, realized advance over SGS?
- Many of the TGS platforms will also have a more general utility beyond DNA sequencing, including:
  - identification of patterns of methylation,
  - comprehensive characterization of transcriptomes,
  - comprehensive characterization of translation.
- **Are we ready to master that data and translate the results into managable bits of information ???**

“Only by marrying information technology to the life sciences and biotechnology will we realize the astonishing potential of the vast amounts of biological data we will be capable of generating with TGS coming on line …”

"We need time to dream, time to remember,
and time to reach the infinite.
Time to be."

—Gladys Taber

VIDEOS

454 SEQUENCING

http://www.454.com/products-solutions/multimedia-presentations.asp

Genome Sequencer FLX Multimedia Presentation

Genome Sequencer FLX Standard Series Workflow Presentation

Genome Sequencer FLX Amplicon Sequencing Presentation

Solexa

http://www.illumina.com/technology/sequencing_technology.ilmn

http://marketing.appliedbiosystems.com/images/Product/Solid_Knowledge/flash/102207/solid.html

Helicos BioSciences Corporation

http://www.helicosbio.com/Technology/TrueSingleMoleculeSequencing/tSMStradeHowItWorks/tabid/162/Default.aspx

VISIGEN BIOTECHNOLOGIES, INC

http://visigenbio.com/technology_movie_streaming.html

PACIFIC BIOSCIENCES

http://www.pacificbiosciences.com/video_lg.html