

PacBio, Oxford nanopore 기술에 대한 이해

(주)제노믹베이스
이사 남 문

nammoon8406@genomicbase.co.kr

010-2607-5344

NGS generation 개요

➤ 1세대 sequencing

- Sanger sequencing : Dideoxynucleotide chain termination

: ddNTP에 형광표지 부착 후 레이저로 확인하는 염기서열 분석 방법
분석 과정에 있어 비용과 시간이 많이 소요된다는 단점이 있음

➤ 2세대 NGS (short read sequencing)

- Roche 454 GS FLX, Thermo Fisher Ion PGM, Illumina genome analyzer, ABI SOLiD platform

: PCR 증폭 방식에 따라 구분 됨 ; Roche 454, Ion PGM – emulsion PCR (longest read length : 1,000bp, 400bp)

: Illumina, SOLiD solid-phase amplification (longest read length : 300bp, 75bp)

짧은 read 길이, 라이브러리 및 PCR 증폭 필요, 반복서열이 존재할 경우 유전체 염기서열의 assembly에 영향이 커짐

➤ 3세대 NGS (long read sequencing)

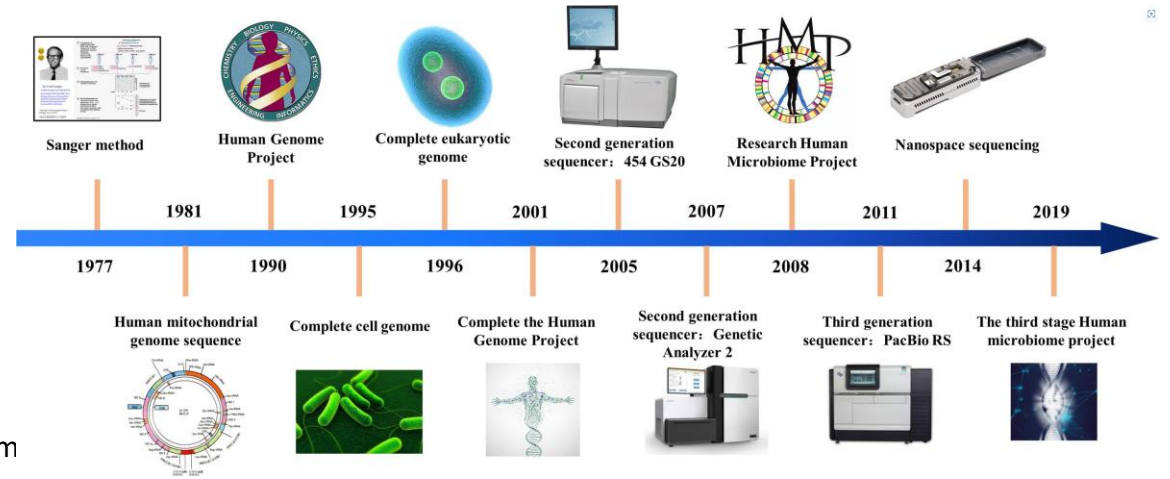
- Pacific Biosciences PacBio ; SMRT (longest read length : 20,000bp)

: Sequencing 진행 전 PCR 과정을 생략하고 DNA 단일 분자를 그대로 분석하기 때문에 시간 단축 및 PCR 과정 중 오류 억제
긴 read 길이, 라이브러리만 필요

- Oxford Nanopore (longest read length : 200,000bp)

: 1,2세대의 PCR을 통한 DNA 증폭이나 3세대의 형광을 사용한 분석 방법이 생략되어 준비시간이 매우 짧고 분석 역시 실시간으로 진행 됨

Homopolymer(동일염기반복서열)과 같은 부분에서 어려움이 상대적으로 높고, 이 때문에 다른 platform을 함께 이용함 (Hybrid genome seq)

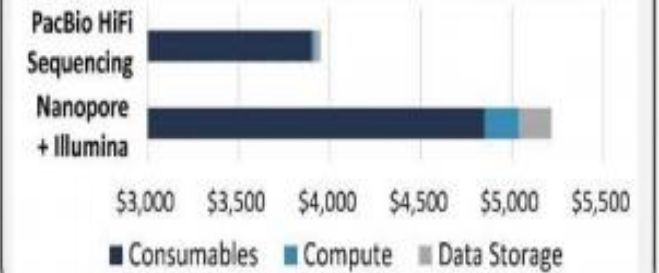


PacBio vs Nanopore comparison

Platform	세대	특징	Read length	PCR 여부	시퀀싱 비용	
PacBio sequencing	3 세대	DNA직접 이용 형광 검출	≤20,000	No PCR	CCS (20-30Gb)	-
					CLR (250-300Gb)	-
Oxford Nanopore		DNA직접 이용 전류량 변화	≤200,000	No PCR	-	-
					-	-

Head-To-Head Direct Cost Comparison

Sequencing Costs for a De Novo Human Genome

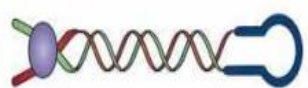


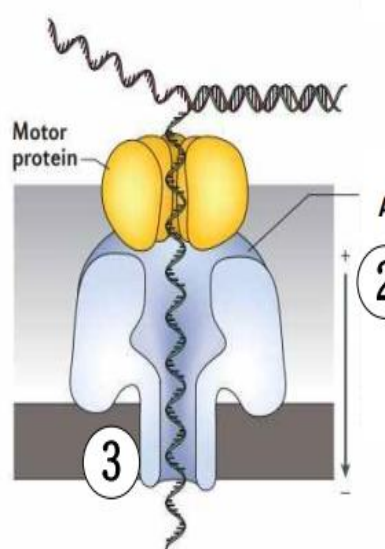
Human Genome Assembly Quality Metrics

	Nanopore + Illumina	PacBio HiFi Sequencing
Contiguity (N50)	32.3 Mb	98.7 Mb
Correctness (Quality Score)	Q34	Q51
Completeness (Genome Size)	2.8 Gb	3.1 Gb

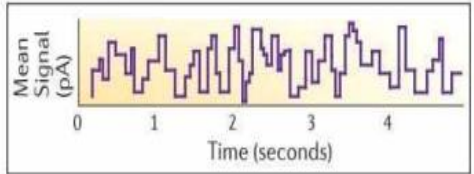
PacBio vs Nanopore comparison

Oxford nanopore

운동단백질 →  Leader-Hairpin 라이브러리 헤어핀 구조의 어댑터 및 운동단백질이 결합한 형태

 Motor protein Alpha-hemolysin (나노포어)


② 전류 나노포어를 통해 DNA가 이동하면서 전류의 차이가 발생함

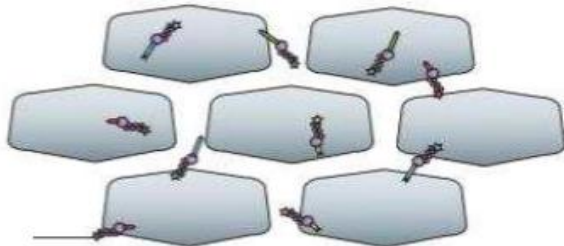
③  나노포어를 통해 DNA가 이동할 때 흐르는 이온전류 값의 차이로 염기의 종류를 분석함

④

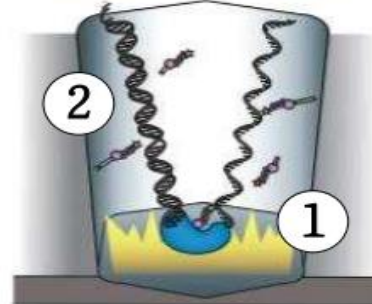
그림 10. Oxford nanopore Sequencing
Source : GENETICS, 2016, Vol.17:333-351

PacBio SMRT

두 개의 헤어핀 구조 어댑터가 결합한 형태의 SMRT 라이브러리 

분석이 진행되는 ZMW wells 

모든 염기는 형광 표지됨

염기는 중합효소(Polymerase)에 의해 라이브러리에 상보적으로 결합되면서 빛을 방출하고, 카메라가 이를 측정함 


모든 ZMW의 색상 변화를 측정함.  ③

그림 9. Pacific Biosciences sequencing
Source : GENETICS, 2016, Vol.17:333-351

Pacific Biosciences (PacBio) Sequel II Sequencing

❑ CCS Mode (Circular Consensus Sequencing)

“Circular consensus sequencing (CCS) read: The consensus sequence resulting from alignment between subreads taken from a single ZMW. Requires at least two full-pass subreads from the insert. CCS reads are advantageous for amplicon and RNA sequencing projects and are highly accurate (>99% accuracy, Q>20).”

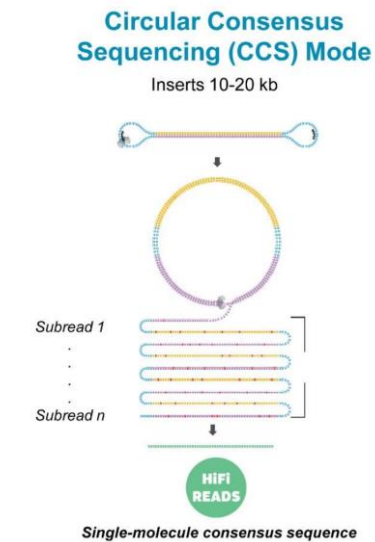
PacBio definition

Insert Size: 10-20 kb

❑ CLR Mode (Continuous Long Read)

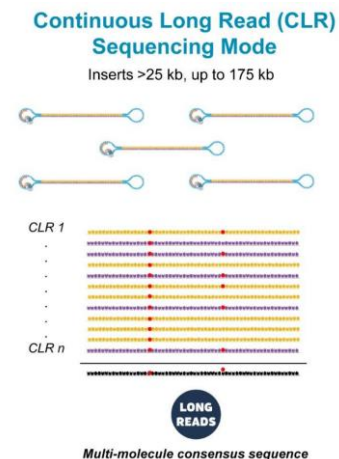
“Continuous long reads (CLR) read: Reads with a subread length approximately equivalent to the polymerase read length indicating that the sequence is generated from a single continuous template from start to finish. The CLR sequencing mode emphasizes the longest possible reads.” PacBio definition

Insert Size: 25-175 kb



99% Accurate

Single molecule;
Multiple reads

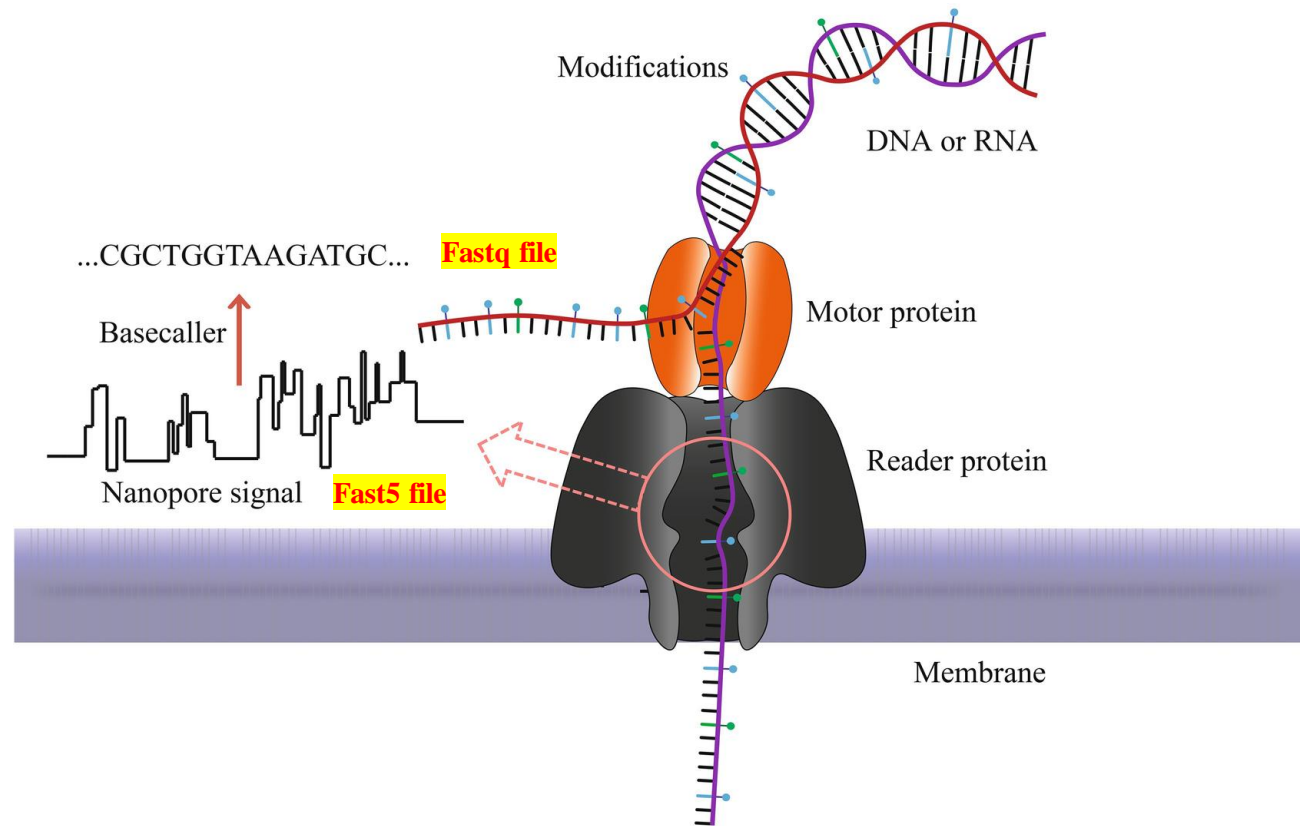


90% Accurate

Multiple molecule;
single reads

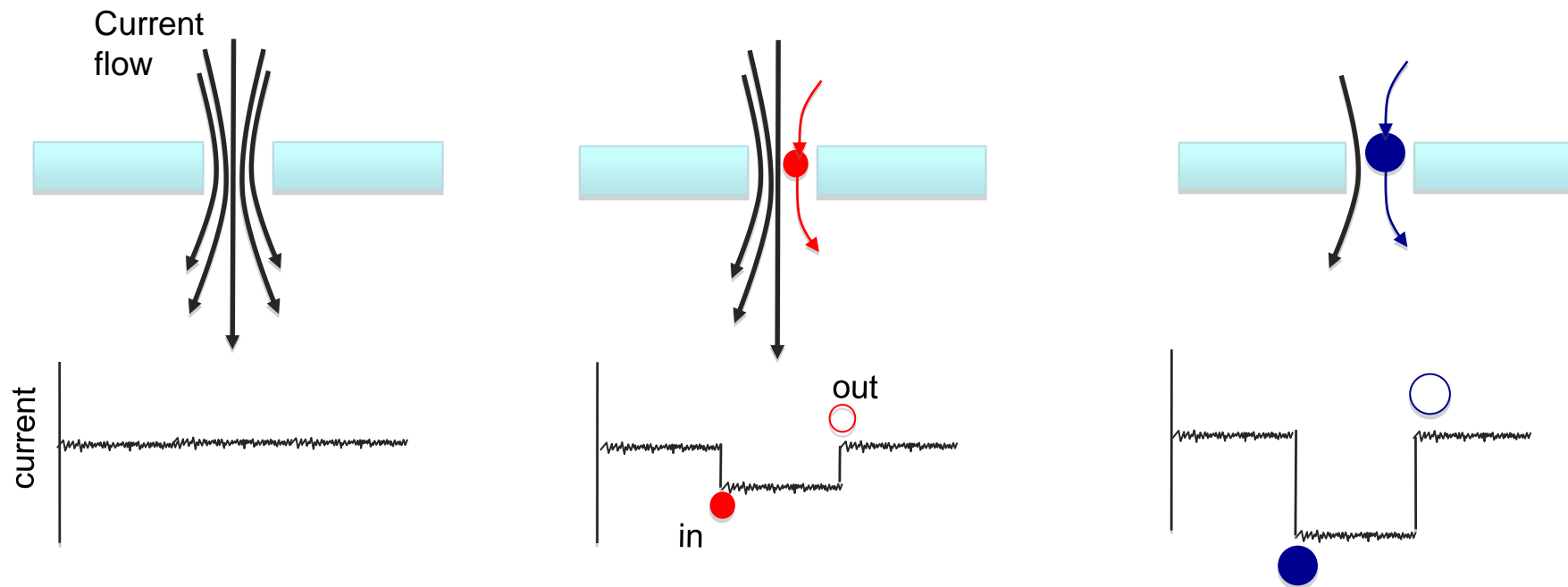
Introduction to Oxford Nanopore Technologies

- ‘Long read’ technology
- A DNA library is prepared (proteins are added)
- Nucleic acids are passed through a protein nanopore
- As the different bases move through the nanopore, it creates a different electrical signal
- These resulting changes in the electrical signal is decoded(Fast5 file) to provide the specific DNA or RNA sequence(Fastq file)



What is a nanopore?

- Nanopore = ‘very small hole’_약 1.5nm 정도의 크기
- Electrical current flows through the hole
- Introduce analyte of interest into the hole → identify “analyte” by the disruption or block to the electrical current



PacBio sequencing 원리 및 과정

Real-Time DNA Sequencing from Single Polymerase Molecules

John Eid,* Adrian Fehr,* Jeremy Gray,* Khai Luong,* John Lyle,* Geoff Otto,* Paul Peluso,* David Rank,* Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, Bidhan Chaudhuri, Frederick Christians, Ronald Cicero, Sonya Clark, Ravindra Dalal, Alex deWinter, John Dixon, Mathieu Foquet, Alfred Gaertner, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden, Gregory Kearns, Xiangxu Kong, Ronald Kuse, Yves Lacroix, Steven Lin, Paul Lundquist, Congcong Ma, Patrick Marks, Mark Maxham, Devon Murphy, Insil Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorenson, Austin Tomaney, Kevin Travers, Mark Trulson, John Vieceli, Jeffrey Wegener, Dawn Wu, Alicia Yang, Denis Zaccarin, Peter Zhao, Frank Zhong, Jonas Korfach,† Stephen Turner†

We present single-molecule, real-time sequencing data obtained from a DNA polymerase performing uninterrupted template-directed synthesis using four distinguishable fluorescently labeled deoxyribonucleoside triphosphates (dNTPs). We detected the temporal order of their enzymatic incorporation into a growing DNA strand with zero-mode waveguide nanostructure arrays, which provide optical observation volume confinement and enable parallel, simultaneous detection of thousands of single-molecule sequencing reactions. Conjugation of fluorophores to the terminal phosphate moiety of the dNTPs allows continuous observation of DNA synthesis over thousands of bases without steric hindrance. The data report directly on polymerase dynamics, revealing distinct polymerization states and pause sites corresponding to DNA secondary structure. Sequence data were aligned with the known reference sequence to assay biophysical parameters of polymerization for each template position. Consensus sequences were generated from the single-molecule reads at 15-fold coverage, showing a median accuracy of 99.3%, with no systematic error beyond fluorophore-dependent error rates.

The Sanger method for DNA sequencing (1) uses DNA polymerase to incorporate the 3'-dideoxynucleotide that terminates the synthesis of a DNA copy. This method relies

on the low error rate of DNA polymerases, but exploits neither their potential for high catalytic rates nor high processivity (2–4). Increasing the speed and length of individual sequencing reads beyond the current Sanger technology limit will shorten cycle times, accelerate sequence assembly, reduce cost, enable accurate sequencing analysis of repeat-rich areas of the genome, and reveal large-scale genomic complexity (5, 6). Alternative approaches that increase sequencing performance

have been reported [(7–10), reviewed in (11, 12)]. Several of these methods have been deployed as commercial sequencing systems (13–16), which have greatly increased overall throughput, enabling many applications that were previously unfeasible. However, because these methods all gate enzymatic activity, using various termination approaches, they have not yielded longer sequence reads (limited to ~400 nucleotides), nor do they exploit the high intrinsic rates of polymerase-catalyzed DNA synthesis.

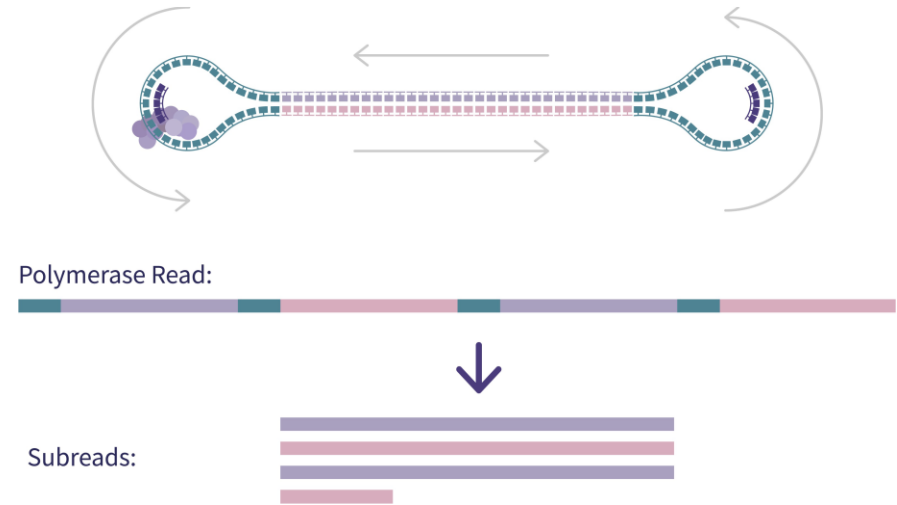
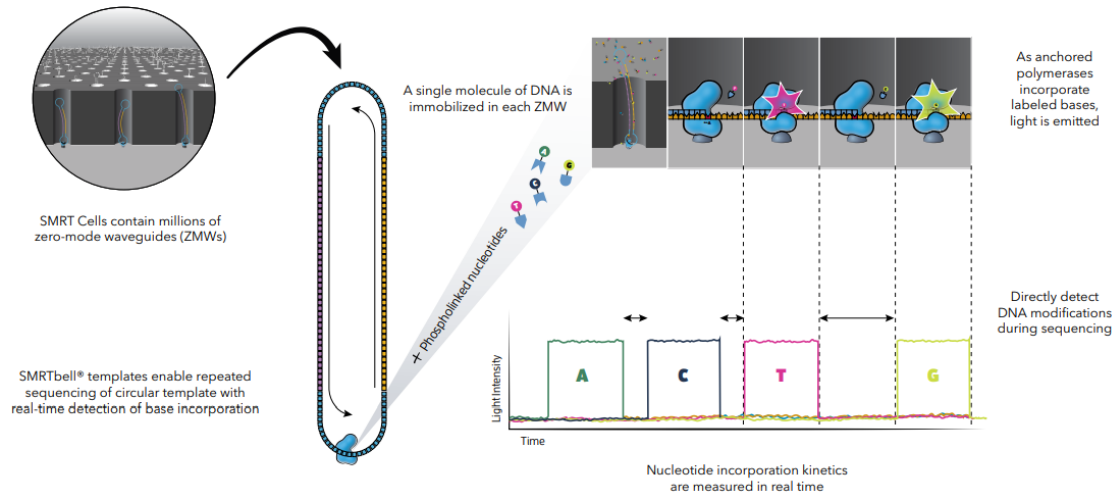
The use of DNA polymerase as a real-time sequencing engine—that is, direct observation of processive DNA polymerization with base-pair resolution—has long been proposed but has been difficult to realize (7, 8, 17–22). To fully harness the intrinsic speed, fidelity, and processivity of these enzymes, several technical challenges must be met simultaneously. First, the speed at which each polymerase synthesizes DNA exhibits stochastic fluctuation, so polymerase molecules would need to be observed individually while they undergo template-directed synthesis. Because of the high nucleotide concentrations required by DNA polymerases (20), a reduction in the observation volume beyond what is afforded by conventional methods, such as confocal or total internal reflection microscopy, directly improves single-molecule detection. Second, deoxyribonucleoside triphosphate (dNTP) substrates must carry detection labels that do not inhibit DNA polymerization even when 100% of the native nucleotides are replaced with their labeled counterparts. Third, a surface chemistry is required that retains activity of DNA polymerase molecules and inhibits nonspecific adsorption of labeled dNTPs. Finally, an instrument is required that can faithfully detect and distinguish incorporation of four different labeled dNTPs. Here, we provide proof-of-concept for an approach to highly

Pacific Biosciences, 1505 Adams Drive, Menlo Park, CA 94025, USA.

*These authors contributed equally to this work.

†To whom correspondence should be addressed. E-mail: jkorfach@pacificbiosciences.com (J.K.); sturner@pacificbiosciences.com (S.T.)

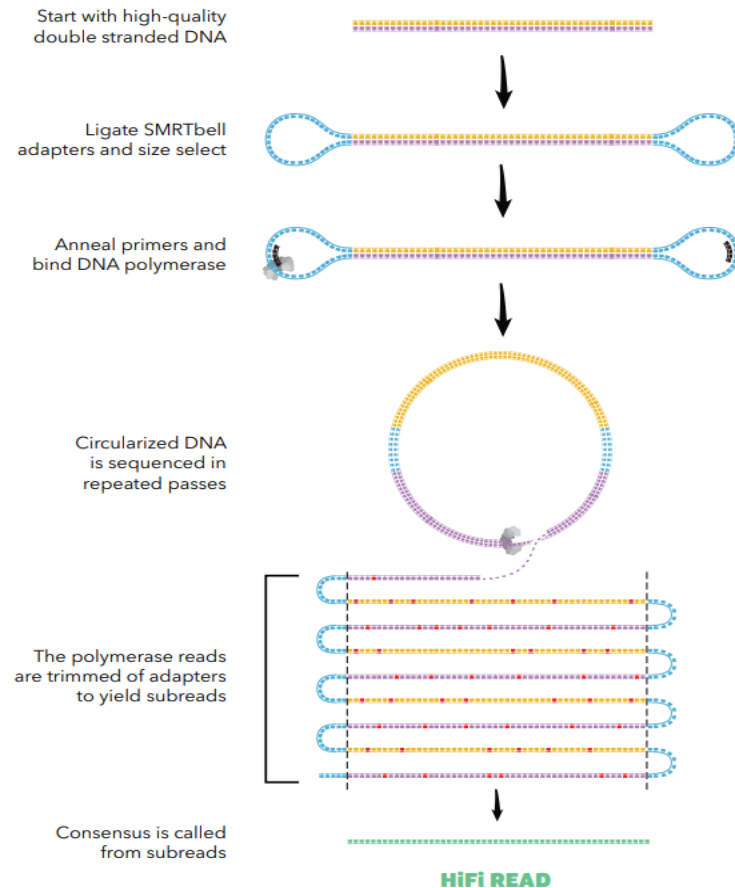
PacBio CLR read _ Continuous Long Reads



- Single molecule real-time (SMRT) sequencing
- Library preparation 과정에서 만들어지는 **SMRT bell**이 가장 중요한 기술
- 정제된 DNA를 일정 조각으로 잘라준 후, illumina와 마찬가지로 양 말단에 adapter를 붙여주는데 이 adapter는 일종의 **hairpin 구조**를 가지고 있음
- Library preparation이 끝나게 되면 종 두개가 붙은 끈 모양이 되는데 denaturation을 하게 되면 원형의 library가 만들어짐
- 이렇게 만들어진 library를 SMRT cell에 전개 시키게 되는데 cell은 작은 방 아래에 효소가 고정되어 있는 구조
- 이 효소가 DNA분자를 만나게 되어 합성을 진행하고 각 dNTP에 붙어 있는 signal molecule에 따라 신호가 발생하여 sequencing 진행
- 또한 DNA 분자가 원형이므로 효소는 동일 분자를 여러 번 sequencing하게 되며 이를 통해 error가 보정 됨
- Long-read seq의 장점 : contig assembly과정에서 repeat이나 error로 인해 contig가 끊어질 확률이 줄어들기 때문에 보다 적은 contig로 유전체 sequence 확인 가능
- 다만 illumina와 같은 platform에 비해 yield가 다소 낮음, Read length : up to ~100kb

PacBio CCS read _ Circular Consensus Sequencing

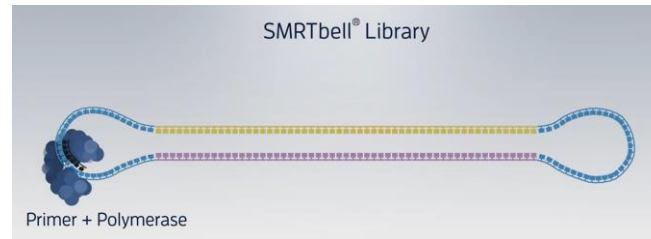
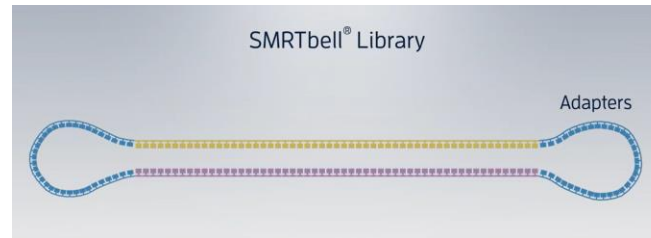
- provide base-level resolution with >99.9% single molecule read accuracy
- HiFi Read : high-fidelity long reads



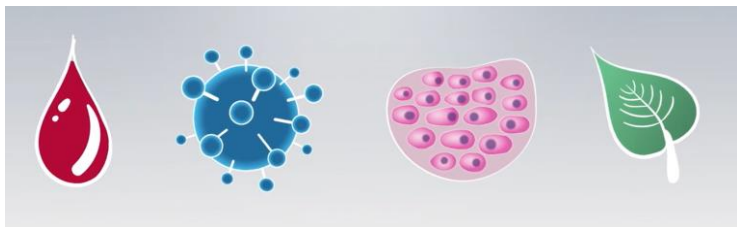
The Benefits of HiFi Reads

- ✓ Long read lengths up to 25kb
- ✓ High read accuracy >99.9%
- ✓ Easy library preparation
- ✓ Low coverage requirements
- ✓ Small file size to minimize computer time
- ✓ A single technology solution for a range of applications
- ✓ Unmatched data clarity for rapid interpretation

PacBio Sequencing – How it Works



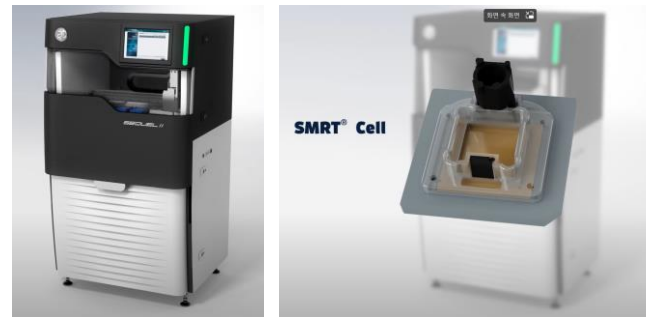
- The PacBio SequelII System, powered by Single molecule, real-time (SMRT) sequencing technology.



- Here's how SMRT sequencing works...
- From any sample type, ranging from viruses to vertebrates, DNA or RNA is isolated.

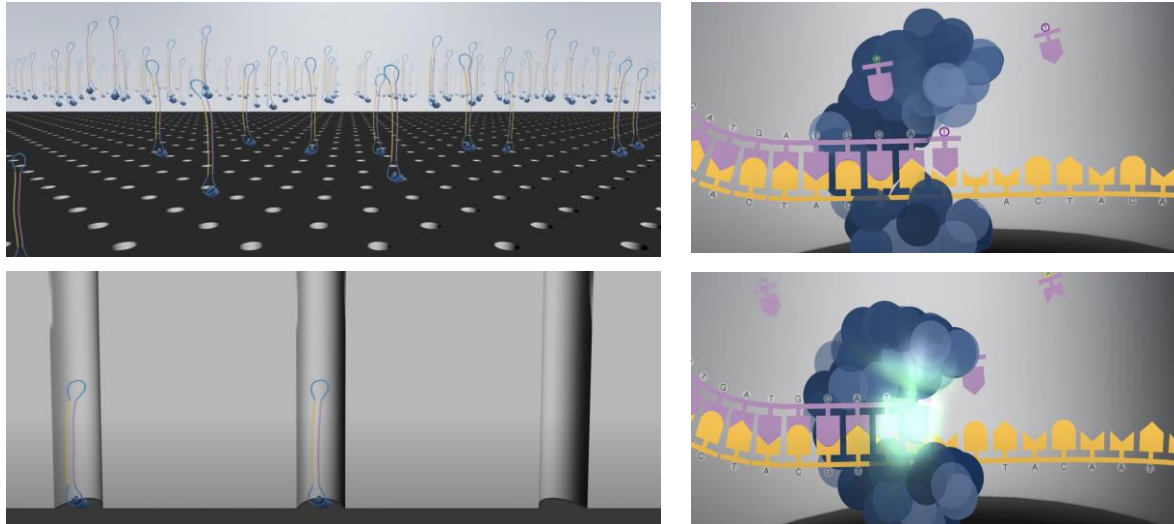
- Next, a **SMRTbell** library is created by ligating adapters to double stranded DNA, creating a circular template.

- Primer and polymerase are added to the library that is placed on the instrument for sequencing.

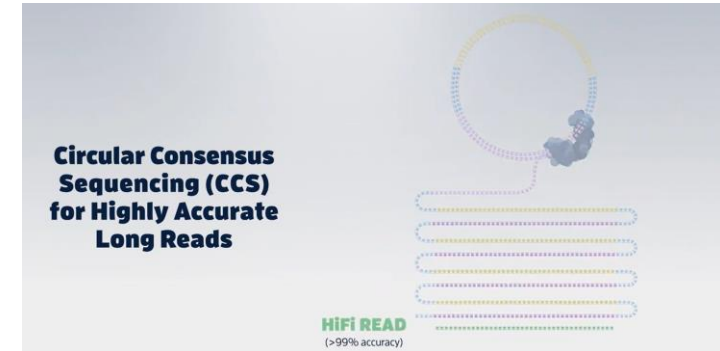


- At the core of SMRT sequencing is the SMRT Cell, which contains millions of tiny wells called zero-mode waveguides, or ZMWS.

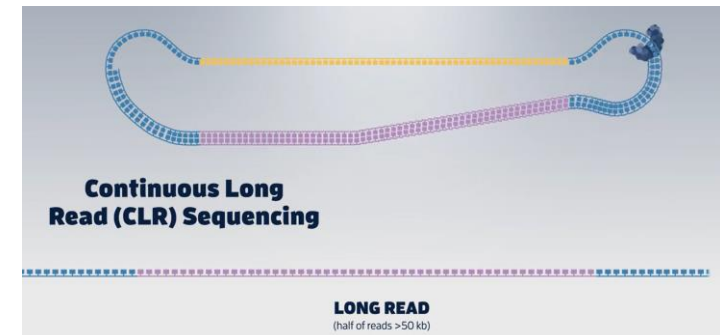
PacBio Sequencing – How it Works



- A single molecule of DNA is immobilized in the **ZMWs** and as the polymerase incorporates labeled nucleotides, light is emitted.
- With this approach nucleotide incorporation is measured in real time.
- With the Sequel system you can optimize your results with two sequencing modes.



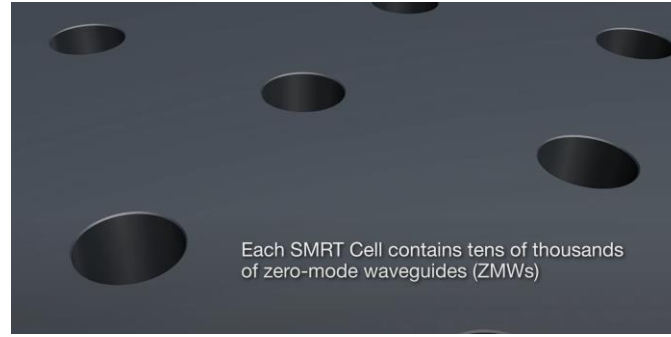
- Use circular consensus sequencing mode to produce highly accurate long reads, known as **HiFi reads**.



- Or use the **continuous long read sequencing** mode to generate the longest possible reads

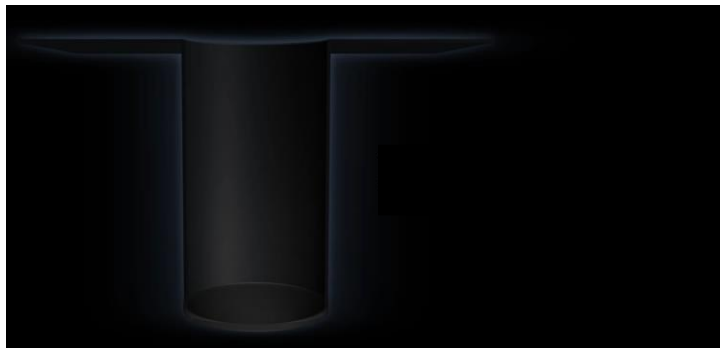
Introduction to SMRT Sequencing

SMRT™ Cell

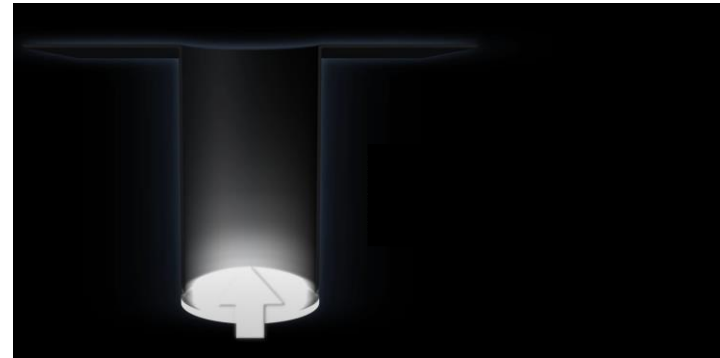


PacBio sequencing
- SEQUELII

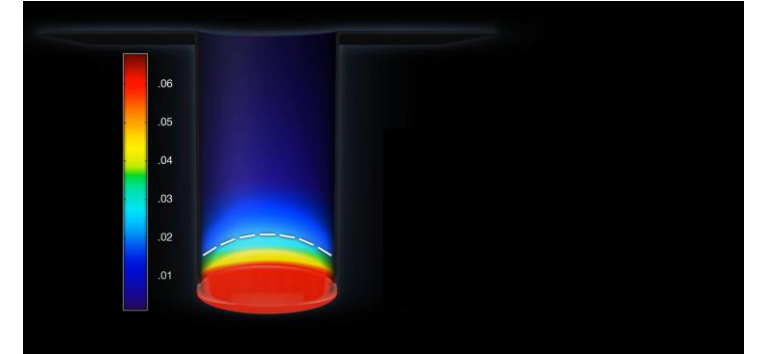
- Each SMRT Cell contains tens of thousands of zero-mode waveguides. (ZMWs)



- The ZMW provides the world's smallest detection volume.



- 각각의 ZMW의 아래에서 빛을 비추며, 빛의 파장이 커서 waveguide를 통해 밖으로 빛이 나가지 못함.

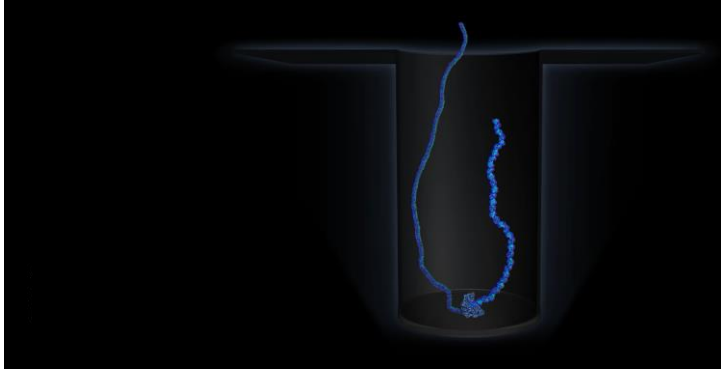


- Attenuated light from excitation beam penetrates the lower 20-30nm of each ZMW...
- ...creating the world's most powerful microscope with a detection volume of 20 zeptoliters (10^{-21} liters)

Introduction to SMRT Sequencing

PacBio sequencing process

1단계



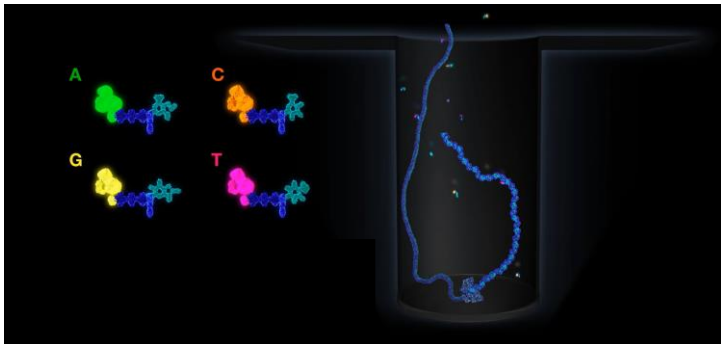
- A DNA template-polymerase complex is immobilized at the bottom of the ZMW

2단계



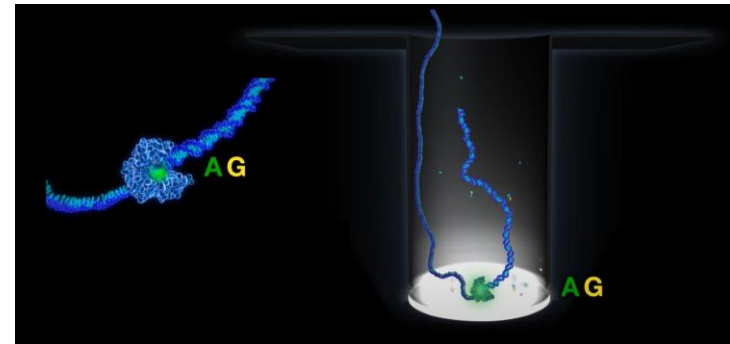
- Phospholinked nucleotides are introduced into the ZMWs chamber

3단계



- Each of the four nucleotides is labeled with a different colored fluorophore

4단계

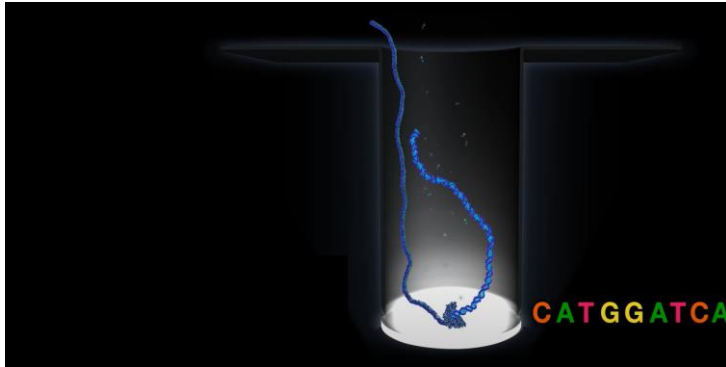


- As a base is held in the detection volume, a light pulse is produced
- This tiny detection volume provides 1000-fold improvement in the reduction of background noise

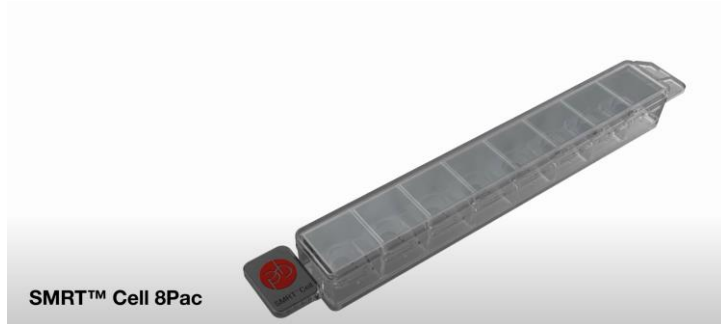
Introduction to SMRT Sequencing

PacBio sequencing process

5단계

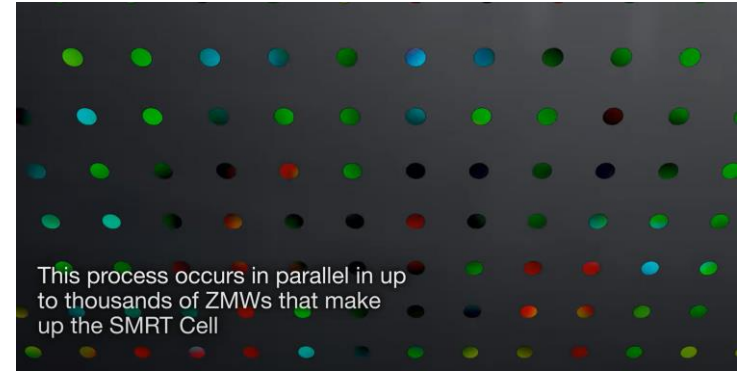


- This process occurs in parallel in up to thousands of ZMWs that make up the SMRT Cell

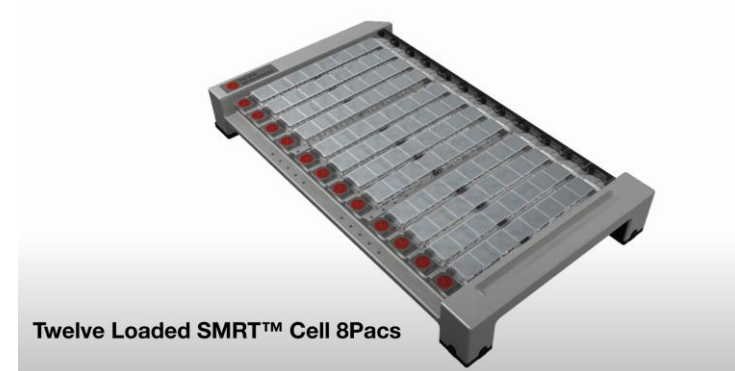


- SMRT™ Cell 8Pac

6단계



- This process occurs in parallel in up to thousands of ZMWs that make up the SMRT Cell



- Twelve Loaded SMRT™ Cell 8Pac

PacBio CCS read _ Circular Consensus Sequencing

Specifications

100M
ZMW / run

360 Gb
HiFi yield per run

24 hr
Sequencing time



15-18 kb
Read length

5mC
DNA methylation

90% ≥Q30
Base quality



Scale: 15x more throughput than Sequel IIe system

	 Sequel IIe system	 Revio system	Increase
Higher density	8 million ZMWs	25 million ZMWs	3×
Independent stages	1	4	4×
Shorter run times	30 hours	24 hours	1.25×
30× human HiFi genomes / year	88	1,300	15× overall

Four key benefits of the Revio system



Scale

1,300 human HiFi genomes per year



Ease of use

Simplified consumables and flexible run setup



Compute power

Google DeepConsensus and more on board

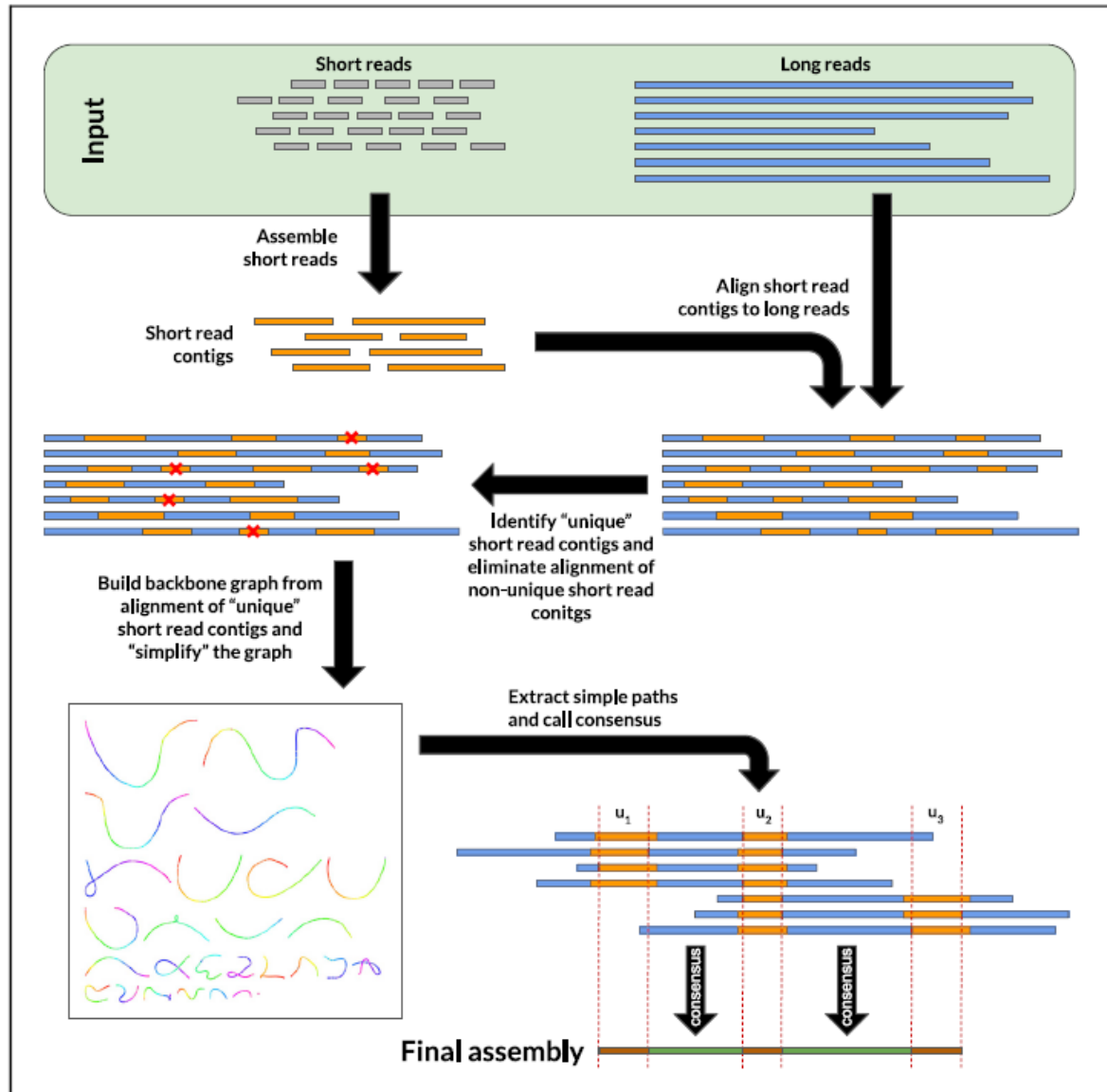


Affordability

Cost-effective human HiFi genome

PacBio sequencing 사례

HASLR: Fast Hybrid Assembly of Long Reads_Long read assembly method



iScience

CellPress
OPEN ACCESS

Article

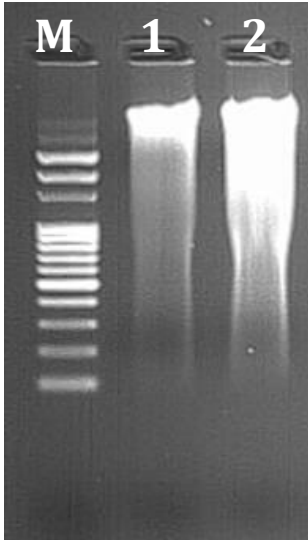
HASLR: Fast Hybrid Assembly of Long Reads

The read assembly process of HASLR consists of the following steps:

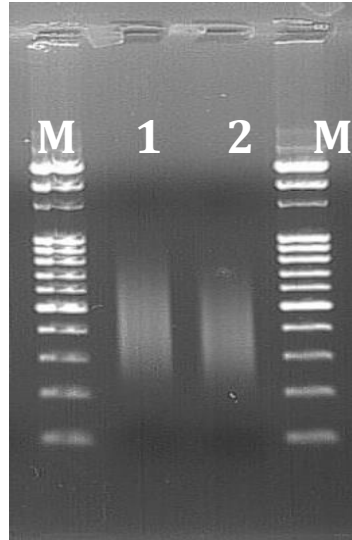
1. Assemble single reads (SR) using a fast SR assembler called Minia.
2. Identify unique SR contigs that are estimated to appear only once in the genome.
3. Map long reads (LR) to unique SR contigs and build a new data structure called backbone graph with edges connecting them.
4. Simplify the backbone graph to reduce the impact of misassemblies and LR mappings to SR contigs.
5. Compute consensus sequences that fill gaps between adjacent SR contigs for each edge.
6. Generate the final assembly using all SR contigs and consensus sequences.

Pacbio long read sequencing 1차 (A작물)

gDNA QC with agarose gel



Library QC with agarose gel



No	Sample	Qubit flex fluorometer		Denovix Nanodrop	
		Con (ng/ul)	260/280	260/230	
1	A 작물	42.4	1.79	1.23	
2	A' 작물	87.0	1.80	1.62	

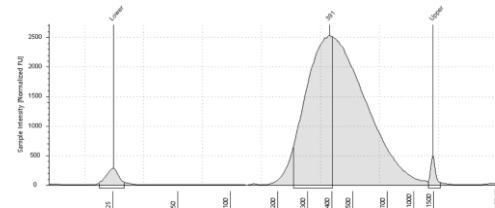
1.5% Agarose gel
200v, 35min
100bp marker(M) 5ul
5ul of library

Library QC Result of DNA

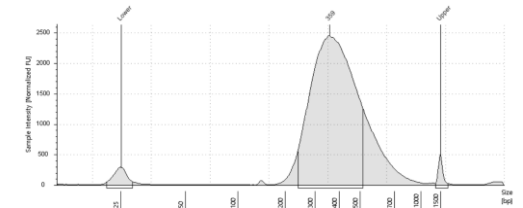
#	Library Name	Library Type	Conc. (ng/ul)	Conc. (nM)	Size (bp)	Result*	
1	A 작물	ETC	46.29	158.62	449	Pass	
2	A' 작물	ETC	43.24	162.26	410	Pass	

✓ Library Size Check

- To verify the size of PCR enriched fragments, we check the template size distribution by running on a **Agilent Technologies 2100 Bioanalyzer** using a DNA 1000 chip.



A 작물 (WGS-lib)



A' 작물 (WGS-lib)

3ul used for library QC

Pacbio long read sequencing (A작물)

Statistics of sequencing raw data (short read)

Samples	Sequencing file	No. of reads	Avg. length (bp)	Total length (bp)	GC(%)*1	Q30(%)*2	Genome cov.*3
A 작물	A 작물 _1.fastq	175,321,935	151	26,473,612,185	37.73	91.79	≒ 62.29X
	A 작물 _2.fastq	175,321,935	151	26,473,612,185			
Total	2ea	764,234,058		115,399,342,758			

1) GC (%): GC content.

2) Q30 (%): Ratio of bases that have phred quality score of over 30.

3) Genome cov.: 각 샘플의 Total read length를 expected genome size (850Mb)로 나눈 값.

Subread Stats (long read)

Cell Name	Subread Bases	Subreads	Subread N50	Average Read Length
A 작물	229,508,101,506	18,459,167	13,715	12,433
	258,085,627,470	23,257,062	12,467	11,097
Total	487,593,728,976	41,716,229	26,182	23,530

. Subread Bases : The number of bases in the Subread.

. Subreads : The number of reads in the Subread.

. Subread N50 : 50% of all bases come from Subreads longer than this value.

. Average Read Length : The mean length of the Subreads.

Statistics of sequencing raw data (long read)

Sample	No. of reads	Avg. length(bp)	Total length(bp)	N50	Avg. Pass	Avg. Quality	Genome cov.*2	
A 작물	1차	750,952	13,117	9,850,531,514	13,036	11	Q32	≒ 11.59X
	2차	955,034	11,795	11,265,292,873	11,965	15	Q35	≒ 13.25X
Total	1,705,986		21,115,824,387					

. HiFi Bases : Total bases of HiFi reads.

. HiFi reads : The number of reads in the HiFi read.

. HiFi N50 : A N50 means that half of all bases reside in reads of this size or longer.

. Avg Reads Length : The mean length of HiFi reads.

. Avg Pass : The mean passes of HiFi reads.

. Avg Quality : The mean quality of HiFi reads.

Pacbio long read sequencing (A작물)

Statistics of A작물 genome assembly

Sample	Total contigs	Assembled genome size (bp)	GC (%)	MIN (bp)	MAX (bp)	AVG (bp)	N50 (bp)
A작물	46,185	2,531,609,098	36.71	464	1,196,160	54,814	83,023

Statistics of A작물 genome polishing

Sample	Total contigs	Assembled genome size (bp)	GC (%)	MIN (bp)	MAX (bp)	AVG (bp)	N50 (bp)
A작물	46,135	2,531,362,472	36.71	513	1,196,142	54,868	83,014

- Assembly polishing 과정까지 거친 최종 assembly 결과, 46,135개 contig로 구성된 약 2.5Gb의 서열을 확보.

Pacbio Sequel Sequencing_General terminology

- **SMRT Cell**: Consumable substrates comprising arrays of zero-mode waveguide nanostructures.
- **Adapters**: Exogenous nucleic acids that are ligated to a nucleic acid molecule to be sequenced. For example, SMRTbell™ adapters are hairpin loops that are ligated to both ends of the double stranded DNA insert to produce a SMRTbell™ sequencing template. When adapter sequences are removed from a HiFi read, the read is split into multiple subreads.
- **Movie**: Real-time observation of a SMRT Cell.
- **zero-mode waveguide (ZMW)**: A nanophotonic device for confining light to a small observation volume. This can be, for example, a small hole in a conductive layer whose diameter is too small to permit the propagation of light in the wavelength range used for detection. Physically part of a SMRT Cell.
- **Sequencing ZMW**: A ZMW (zero-mode waveguide) that is expected to be able to produce a sequence if it is populated with a polymerase. ZMWs used for automated SMRT Cell alignment are not considered sequencing ZMWs.
- **Run**: Specifies.
 - The wells and SMRT Cells to include in the sequencing run.
 - The collection and analysis protocols to use for the selected wells and cells.

Nanopore sequencing 원리 및 과정



NIH Public Access

Author Manuscript

Nat Biotechnol. Author manuscript; available in PMC 2009 May 18.

Published in final edited form as:

Nat Biotechnol. 2008 October ; 26(10): 1146–1153. doi:10.1038/nbt.1495.

The potential and challenges of nanopore sequencing

Daniel Branton¹, David W Deamer², Andre Marziali³, Hagan Bayley⁴, Steven A Benner⁵, Thomas Butler⁶, Massimiliano Di Ventra⁷, Slaven Garaj⁸, Andrew Hibbs⁹, Xiaohua Huang¹⁰, Stevan B Jovanovich¹¹, Predrag S Krstic¹², Stuart Lindsay¹³, Xinsheng Sean Ling¹⁴, Carlos H Mastrangelo¹⁵, Amit Meller¹⁶, John S Oliver¹⁷, Yuriy V Pershin⁷, J Michael Ramsey¹⁸, Robert Riehn¹⁹, Gautam V Soni¹⁶, Vincent Tabard-Cossa³, Meni Wanunu¹⁶, Matthew Wigginn²⁰, and Jeffery A Schloss²¹

¹Department of Molecular and Cell Biology, Harvard University, Cambridge, Massachusetts 02138, USA

²Department of Chemistry and Biochemistry, University of California, Santa Cruz, California 95064, USA

³Department of Physics and Astronomy, University of British Columbia, Vancouver, British Columbia V6T 1Z1, Canada

⁴Department of Chemical Biology, Oxford University, Oxford OX1 3TA, UK

⁵Foundation for Applied Molecular Evolution, Gainesville, Florida 32604, USA

⁶Department of Physics, University of Washington, Seattle, Washington 98195, USA

⁷Department of Physics, University of California at San Diego, La Jolla, California 92093, USA

⁸Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA

⁹Electronic BioSciences, San Diego, California 92121, USA

¹⁰Department of Bioengineering, University of California at San Diego, La Jolla, California 92093, USA

¹¹Microchip Biotechnologies Inc., Dublin, California 94568, USA

¹²Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA

¹³Departments of Physics and Chemistry and the Biodesign Institute, Arizona State University, Tempe, Arizona 85287, USA

¹⁴Department of Physics, Brown University, Providence, Rhode Island 02912, USA

¹⁵Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, Ohio 44106, USA

¹⁶Biomedical Engineering, Boston University, Boston, Massachusetts 02215, USA

¹⁷NABsys, Inc., Providence, Rhode Island 02906, USA

¹⁸Department of Chemistry, University of North Carolina, Chapel Hill, North Carolina 27599, USA

¹⁹Department of Physics, North Carolina State University, Raleigh, North Carolina 27695, USA

Correspondence to: Daniel Branton.

Correspondence should be addressed to D.B.(E-mail: dbranton@harvard.edu).

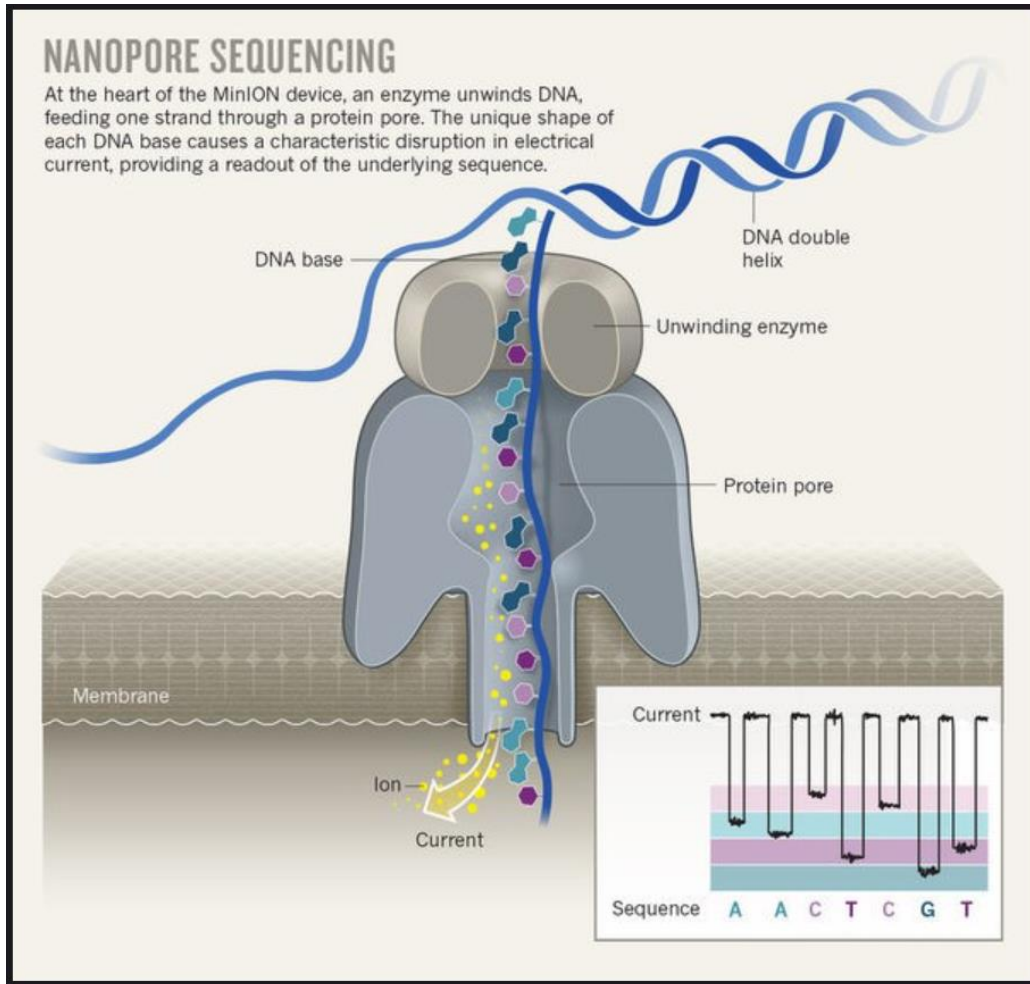
AUTHOR CONTRIBUTIONS D.B. wrote this review, with additions and editorial assistance from D.W.D., A.Marziali, and H.B. S.A.B., T.B., M.D., S.G., A.H., X.H., S.B.J., P.S.K., S.L., X.S.L., C.H.M., A.Meller, J.S.O., Y.V.P., J.M.R., R.R., G.V.S., V.T.-C., M.Wanunu, and M.Wigginn, contributed some of the text and read drafts of the manuscript for accuracy. J.A.S. proposed the idea for the review and read the manuscript for accuracy.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

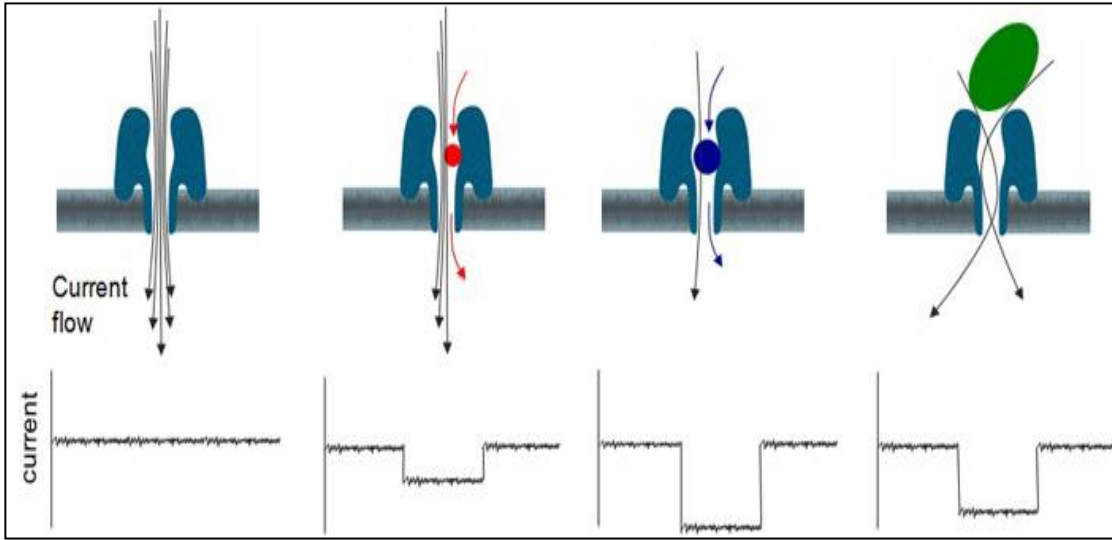
NIH-PA Author Manuscript

Long-read sequencing _ Nanopore



- 직경이 1.5nm인 α -hemolysin 나노포어를 이용하여 DNA 혹은 RNA가 나노포어를 통과하는 것을 측정
- 나노포어 양쪽에 전압을 걸어주면 용액 내의 이온들에 의해 전류가 흐르는데 (-)전극 쪽 챔버에 음전하를 띠고 있는 DNA 용액을 넣으면 DNA가 나노포어를 통과하여 (+)극 쪽으로 이동하게 됨
- DNA에 의해 이온들이 통과할 수 있는 공간이 줄어들기 때문에 이온전류가 감소하게 되며, 통과 후에는 초기 상태의 이온전류값으로 돌아온다.
- 전류 pulse 혹은 blockade 전류 측정을 통해 DNA length, structure 등을 분석할 수 있다

Long-read sequencing _ Nanopore



- Membrane에 나노미터 수준의 채널을 형성한 후 DNA 단일 가닥을 통과 시킬 때, 염기서열에 따라 막 간의 전위차가 변하는 것을 측정하여 sequencing 함
- Membrane 바깥 쪽에서 이중 나선의 DNA가 helicase에 의해 단일 가닥으로 풀리며 채널속으로 일정한 속도로 이동하도록 구성 되어 있으며 채널 안에 약 4개 염기 정도가 들어가기 때문에 실제로 측정한 전위차는 한 개 염기에 해당하는 것이 아니라 4개 염기단위로 얻어짐
- 이러한 방식으로 얻은 data로부터 단일 염기서열을 복구하는 소프트웨어가 중요한 역할을 함

- 다른 sequencing과 달리 template DNA에 상보적인 화학 반응을 하지 않아 DNA를 있는 그대로 측정할 수 있으며 한번에 수만의 염기를 읽을 수 있음
- 특별한 기기가 필요하지 않고 실제 sequencing이 일어나는 membrane을 포함한 flowcell이 있는 카트리지만 필요로 함
- Sequencing의 특징 상 특히 반복 서열에 취약하며 error rate가 높은 편임
- 실시간으로 시퀀싱 되는 것은 전류의 흐름을 측정하여 Fast5 파일로 저장되며, Fast5 파일은 인코딩을 통해 최종적으로 Fastq 파일로 저장 됨

Oxford Nanopore Technologies

➤ MinION



➤ PromethION



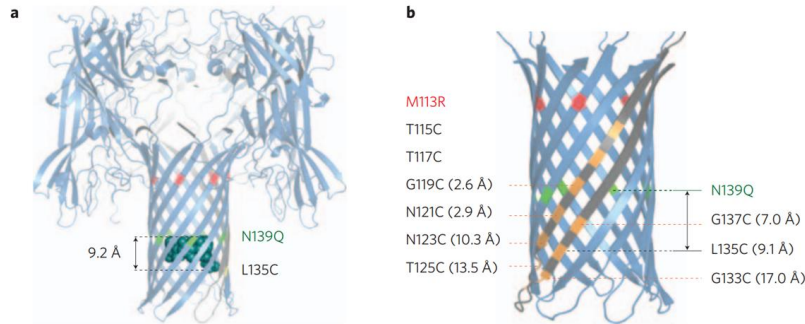
➤ GridION



Device	No. flow cells per run	Throughput	Theoretical maximum output	Cost	국내가격
MinION	1	512 channels	50 Gb	From \$1,000 (starter pack)	250만원 정도 (flow cell 1개)
Mk1C	1	512 channels	50 Gb	From \$4,000 (starter pack)	
GridION	5	5 x 512 channels	250 Gb	From \$49,000 (starter pack)	740만원 정도 (flow cell 4개)
PromethION	24 or 48	10,700+ channels	2,596+ Gb	From \$100,455 (starter pack)	

Continuous base identification for single-molecule nanopore DNA sequencing

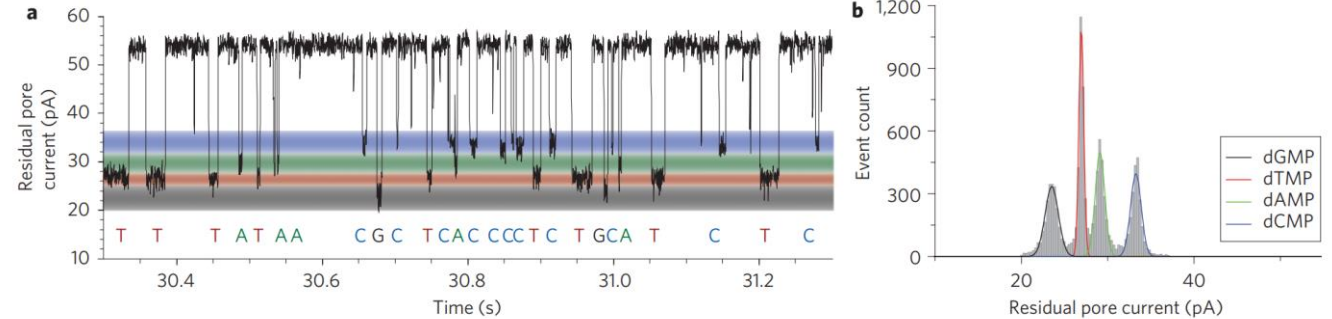
James Clarke¹, Hai-Chen Wu², Lakmal Jayasinghe^{1,2}, Alpesh Patel¹, Stuart Reid¹ and Hagan Bayley^{2*}



$\text{\AA} = 0.1 \text{ nm}$

Structures of haemolysin mutants.

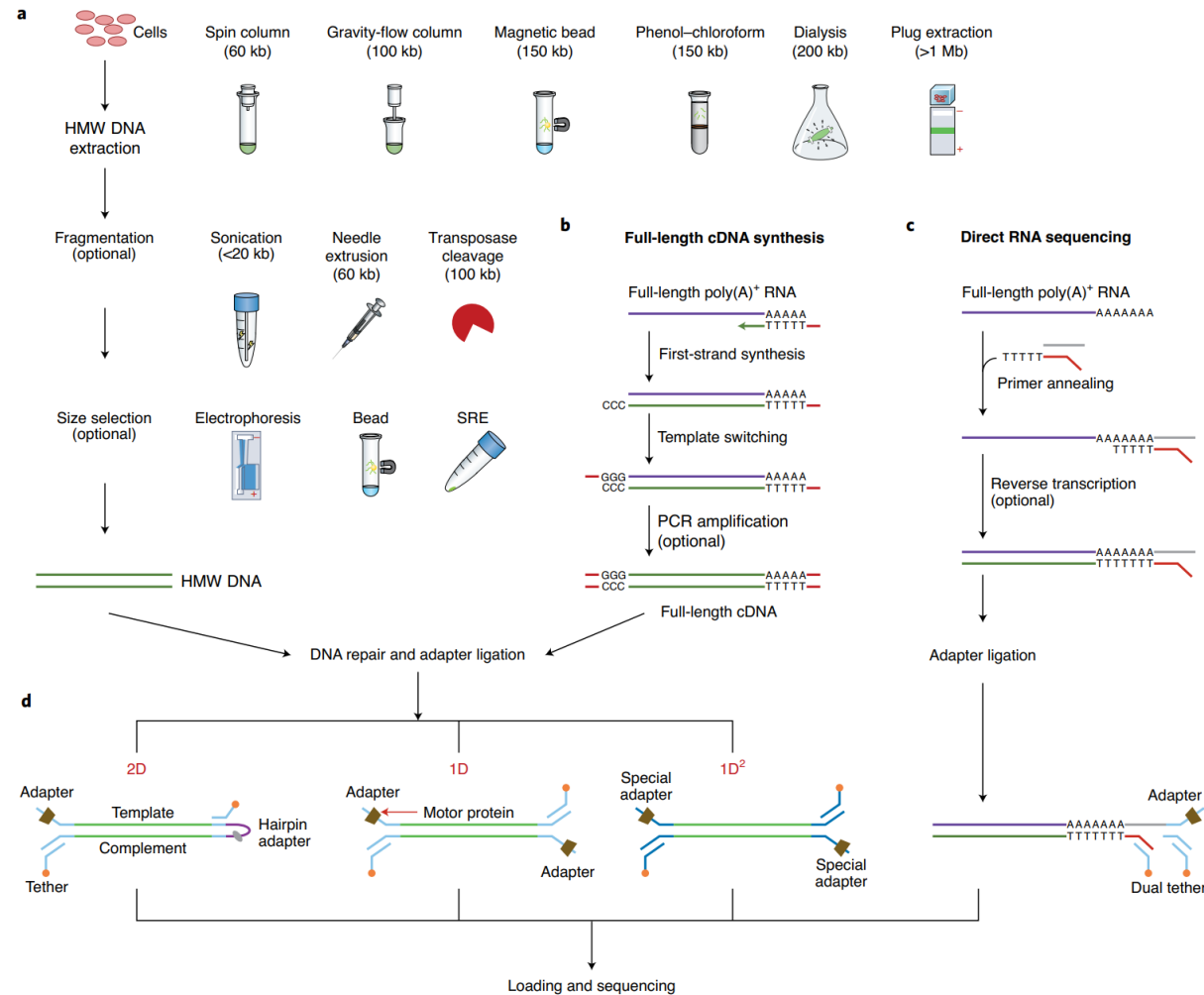
- ▶ 그림 a는 WT-(M113R/N139Q)₆(M113R/N139Q/L135C)₁ 변이형의 구조를 보여주고 있음.
- ▶ 위치 135에 공유 결합된 사이클로덱스트린과, 139번 잔기의 글루타민, 그리고 사이클로덱스트린의 이산화황 결합부와 이차수산화물 간의 수직 거리를 보여줌.
- ▶ 그림 b는 b 배열의 근접한 부분을 확대하여, 이 연구에서 검사된 변이체들의 나이트로젠-나이트로젠 거리를 보여줌.
- ▶ 변이체들은 표준 단일 문자 아미노산 코드를 사용하여 기술됨.



Nucleotide event distributions with the permanent adapter.

- ▶ WT-(M113R/N139Q)₆(M113R/N139Q/L135C)₁-am₆amDP₁bCD 포어에서 dGMP, dTMP, dAMP 및 dCMP의 분별력을 나타내는 색상이 추가된 핵산 이벤트 분포가 포함되어 있음.
- ▶ 각 뉴클레오타이드의 잔류 전류 분포를 나타내기 위해 가우시안 피팅의 중심에서 세 표준 편차를 더한 색상이 추가된 dGMP, dTMP, dAMP 및 dCMP 결합 이벤트의 잔류 전류 히스토그램도 포함되어 있음.
- ▶ 이 데이터는 400mM KCl, 25mM Tris HCl, pH 7.5에서 +180mV의 전압과 함께 10mM dGMP, 10mM dTMP, 10mM dAMP 및 10mM dCMP가 존재하는 상황에서 측정함.

Library preparation workflow for ONT sequencing



a, **Special experimental techniques** for ultralong genomic DNA sequencing, including HMW DNA extraction, fragmentation and size selection.

b, **Full-length cDNA synthesis** for direct cDNA sequencing (without a PCR amplification step) and PCR-cDNA sequencing (with a PCR amplification step).

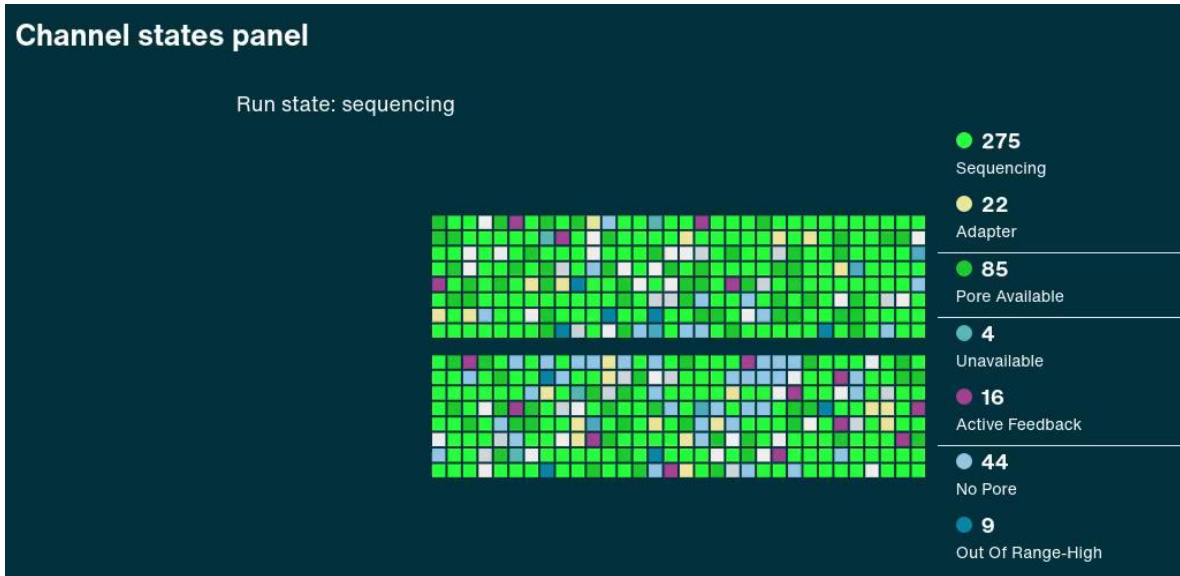
c, **Direct RNA-sequencing library preparation** with or without a reverse transcription step, where only the RNA strand is ligated with an adapter and thus only the RNA strand is sequenced.

d, **Different library preparation strategies** for DNA/cDNA sequencing, including 2D (where the template strand is sequenced, followed by a hairpin adapter and the complement strand), 1D (where each strand is ligated with an adapter and sequenced independently) and 1D² (where each strand is ligated with a special adapter such that there is a high probability that one strand will immediately be captured by the same nanopore following sequencing of the other strand of dsDNA); SRE, short read eliminator kit

Nanopore sequencing 사례

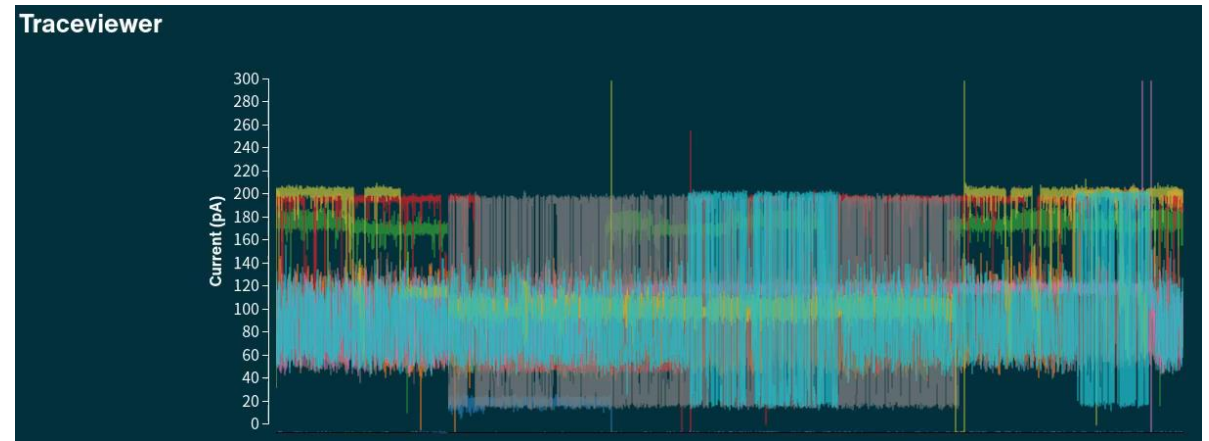
Nanopore long read sequencing

- Nanopore sequencing - 실시간 channel activation



총 pore의 수는 512개/1 채널(총 4채널)

- Nanopore sequencing - 실시간 전류의 흐름을 측정하여 Fast5 포맷으로 변환 중



Nanopore long read sequencing (효모 예시)

Run Info

Experiment Name	yeast
Sample ID	KJS_yeast
Run ID	0e86de01-732d-4aa3-8af2-6df3594f52e3
Flow Cell Id	FAP60700
Start Time	March 17, 15:32
Run Length	2d 20h 49m

Run Summary

Reads Generated	1.34 M
Estimated Bases	4.53 Gb

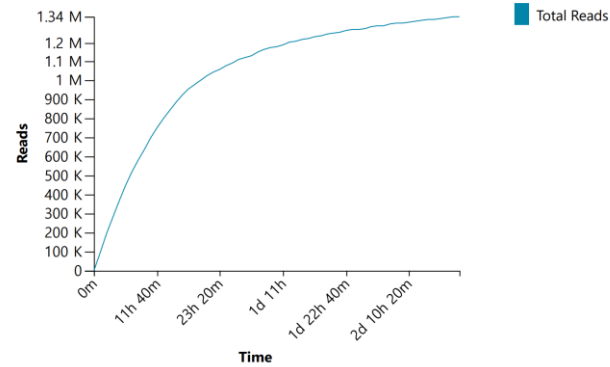
Run Parameters

Flow Cell Type	FLO-MIN106
Kit	SQK-LSK109
Basecalling	off
Specified Run Length	72 hours
Initial Bias Voltage	-180 mV
FAST5 Output	Enabled
FAST5 Output Options	zlib_compress,raw
FAST5 Reads per File	4000
Active Channel Selection	Enabled
Mux Scan Period	1 hour 30 minutes
Reserved Pores	0 %

Versions

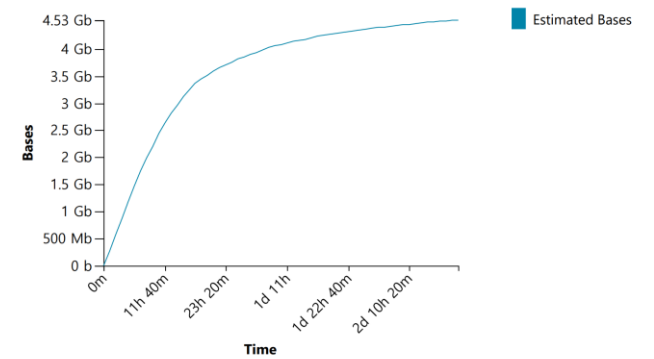
MinKNOW Core	3.6.5
Bream	4.3.16
Guppy	3.2.10

Cumulative Output Reads



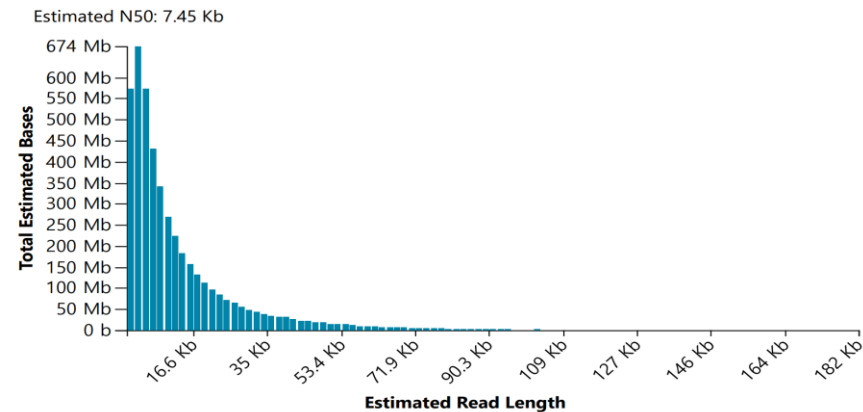
➤ 시간당 생산되는 누적 output reads

Cumulative Output Bases



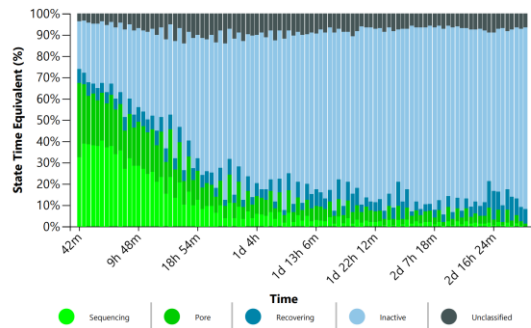
➤ 시간당 생산되는 누적 output bases

Read Length Histogram Estimated Bases

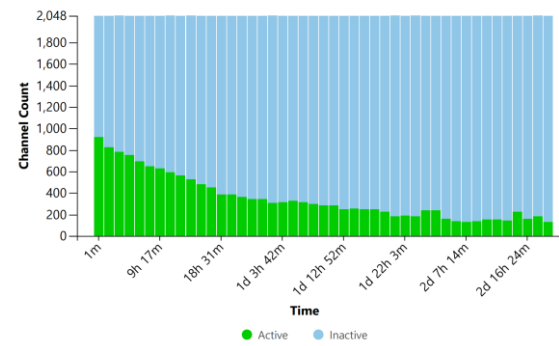


Nanopore long read sequencing (효모 예시)

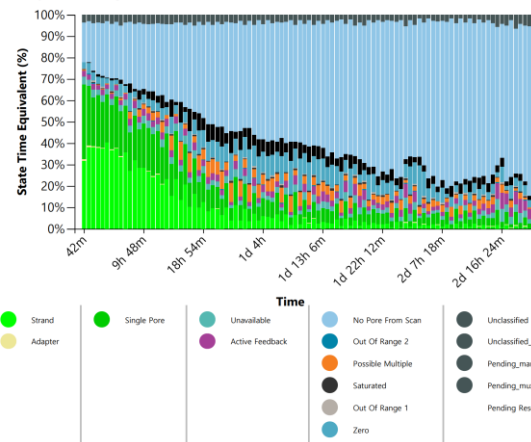
Duty Time Grouped



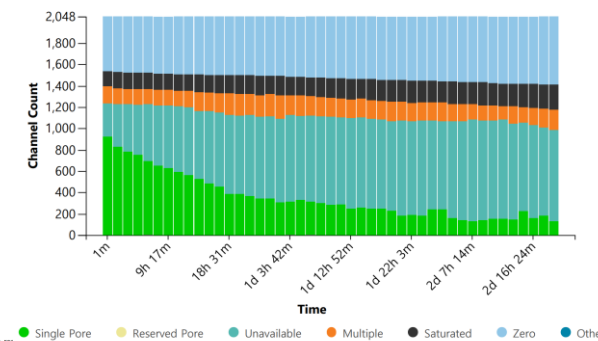
Mux Scan Grouped



Duty time Categorized



Mux Scan Categorized



Channel States

- **Strand** = Currently sequencing
- **Single Pore** = Channel with the characteristics of a pore but not currently sequencing
- **Unavailable** = Stalled chemistry or contaminant in pore
- If channel is 'unavailable' then MinKNOW can reverse potential ('active feedback') for individual channels to free them to sequence
- **Saturated** = popped membrane
- **Multiple** = second pore inserted (occasionally happens), channel switched off, turns zero
- **Zero** = zero current flow. Switched off channel or one never chosen. Spatial zero may be a bubble!
- **Out of range 1/2** = bucket states for extras
- **Remaining classifications (unclassified, pending_MUX_change etc)** = MinKNOW yet to make a decision

- Nanopore flowcell : max 2048 pore를 가짐
- 1차로 check flowcell에서 pore의 수가 800개 이상이면 시퀀싱 진행이 가능하며 평균 1000개 이상의 pore 수를 가짐
- 시퀀싱이 진행될 때는 약 200여 개의 pore가 교대로 돌아가면서 활성화 됨.

Nanopore long read sequencing (효모 예시)

- Statistics of sequencing raw data (short read)

Sample	File Name	No. of reads	Avg. Lenth (bp)	Total length (bp)	Trimmed/raw (%)	Genome cov.
Yeast	Paired 1.fastq	15,216,459	121.73	1,852,238,616	69.51%	308.91X
	Paired 2.fastq	15,216,459	121.89	1,854,733,374	69.60%	
Total		30,432,918		3,706,971,990		

- Statistics of sequencing raw data (long read)

Sample	File Name	No. of reads	Avg. Lenth (bp)	Total length (bp)	MIN (bp)	MAX (bp)	GC (%)	Genome cov.
Yeast	Nanopore	1,148,484	3,484	4,001,375,261	1	181,293	49,35%	333.45X

- Statistics of Hybrid genome assembly

-	Sample	Total contigs	Assembled genome size (bp)	GC (%)	MIN (bp)	MAX (bp)	AVG (bp)	N50 (bp)
Hybrid genome assembly	Yeast	14	27,244,164	50.78	525,906	2,615,776	1,946,011	2,241,297
Assembly polishing	Yeast	14	27,233,536	50.78	525,634	2,614,699	1,945,251	2,240,454

➤ Assembly polishing 과정까지 거친 최종 assembly 결과는 14개의 contig로 구성된 약 27Mb의 서열을 확보

PacBio vs Nanopore comparison

HASLR: Fast Hybrid Assembly of Long Reads_Long read assembly method

PacBio	Sequel-II	Revio	Nanopore
생산량	10-15Gb 보장	70-90Gb 생산	10-20Gb 생산
1cell 비용(VAT별도)	7백만원	5백만원	2백만원
라이브러리 제작	70만원	70만원	1백만원
Base quality	99.9% ≥ Q30	90% ≥ Q30	99.9% ≥ Q30
Run time(hours)	30	24	72

(실제 정보와 약간의 차이가 있을 수 있음)

- 최근 3세대 시퀀싱 기술인 long read를 이용한 분석의 중요성이 높아지면서 PacBio와 Nanopore사의 시퀀싱을 활용한 연구가 매우 활발히 진행되고 있음.
- Long read sequencing의 발전은 가장 큰 단점으로 얘기하는 base quality가 과거에 비해 좋은 품질로 개선 되어가고 있으며, short read와 함께 assemble을 진행하는 방식인 hybrid assembly tool **HASLR 프로그램** 등의 개발로 인하여 게놈 draft 서열 연구에 보다 효과적으로 사용할 수 있음.