UNIVERSITÄT BERN

De novo Genome & Transcriptome Assembly Course

Lecture 2 – Sequencing Technologies

Rémy Bruggmann – Interfaculty Bioinformatics Unit (IBU)

22.09.21

SIB Swiss Institute of Bioinformatics

Content

- Library preparation principles
- Sequencing technologies
  - Sanger
  - NGS
  - third generation (PacBio & Nanopore)
- Sources of Errors
- Hi-C method & Optical mapping
- Is perfect assembly possible?

New sequencing methods required

UNIVERSITÄT BERN

Requirements | Wish List

- much higher throughput
- much lower costs
- parallel library preparation,
  (i.e., no "cloning" of fragments)
- parallel sequencing
- error free (at least no systematic errors)
- ultra long reads

Next Generation Sequencing (NGS) Instruments

UNIVERSITÄT BERN

Thermo Fisher (LifeTechnologies)

- IonTorrent (PGM)
- IonProton
- Ion GeneStudio S5
- Ion Torrent Genexus




Illumina

- MiSeq
- NextSeq500
- NextSeq550
- HiSeq2500
- HiSeq X ten
- NextSeq550
- HiSeq3000
- HiSeq4000
- HiSeq X five
- MiniSeq
- NextSeq1000/2000
- NovaSeq6000

images not to scale

3rd Generation Sequencing Instruments

UNIVERSITÄT BERN

Pacific Biosciences (PacBio)

PacBio RS II

PacBio Sequel I & II

Oxford Nanopore Technology (ONT)

MinION

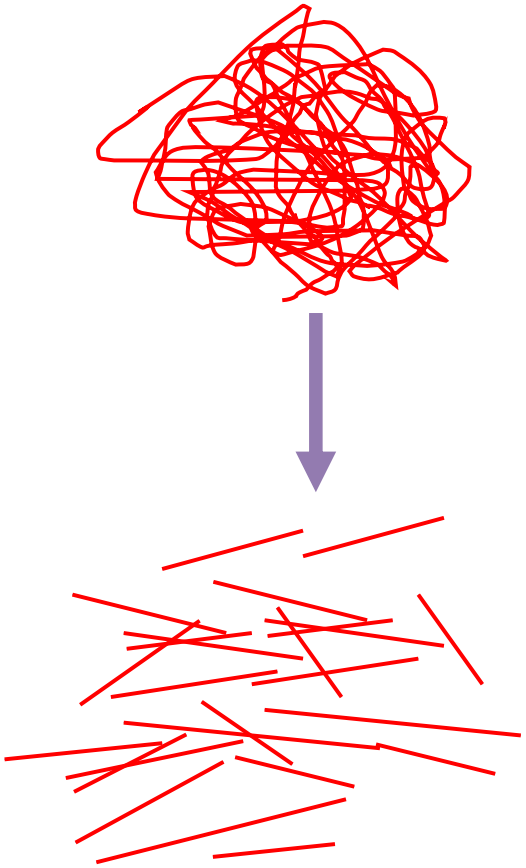
MinION Mk1C

GridION

PromethION

images not to scale

Library Preparation – (Short Reads)

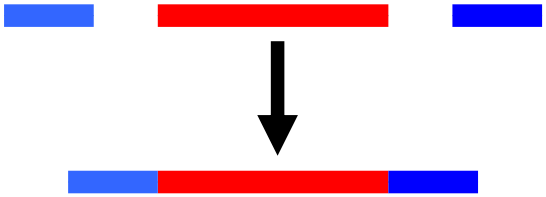


- Input material (DNA)
  - Input material mustbe broken into to smaller pieces suitable for sequencing (and PCR)
  - Not required for amplicon seq
- Fragmentation methods
  - Mechanical
  - Enzymatic
  - Chemical
- Size of fragments follows Poisson distribution

Library Preparation

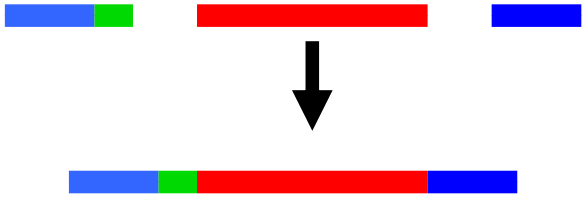
UNIVERSITÄT BERN

no barcodes



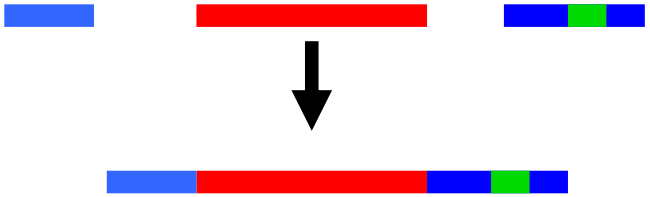
- Adapter ligation
  - no barcodes

single barcodes before Insert



- Adapter ligation
  - IonTorrent, 454
  - 10 bp long barcodes

single barcode in seq primer



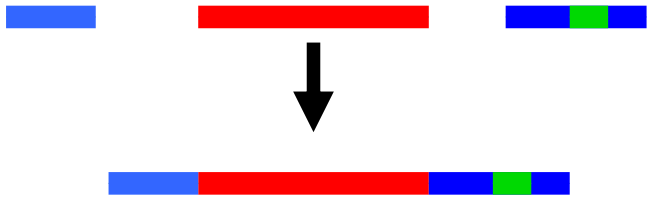
- Adapter ligation
  - Illumina
  - 6-12 bp long barcodes

primer A primer B insert (to be sequenced) barcode

Library Preparation | Barcoding

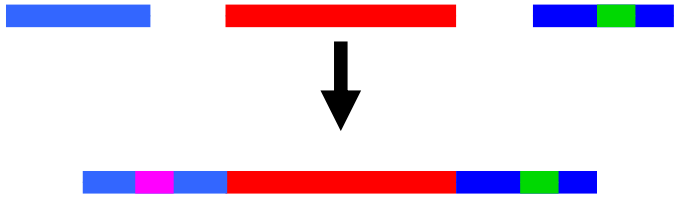
UNIVERSITÄT BERN

single barcode in sequencing primer



- Adapter ligation
  - Illumina
  - 6-12 bp long barcodes

dual barcodes in sequencing primer



- Adapter ligation
  - dual barcodes
  - up to12 bp
  - combinations (up to 384)

Sequencing of Libraries

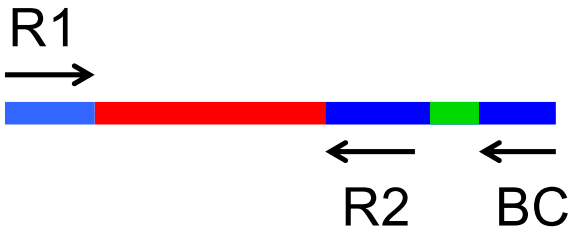
UNIVERSITÄT BERN

sequencing BC+Insert

- IonTorrent
  - sequencing of barcode and insert is together

→ 1 sequence

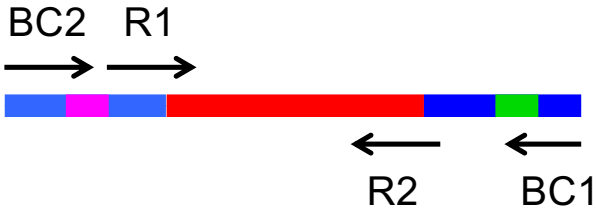
Paired-end, on BC



- Illumina
  - 3 sequencing reactions
    - R1
    - R2
    - BC

→ 3 individual sequences

Paired-end, two BCs

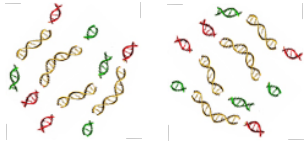
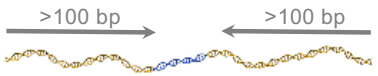
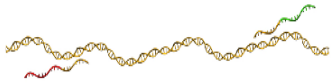


- Illumina
  - 4 sequencing reactions
    - R1
    - R2
    - BC1
    - BC2

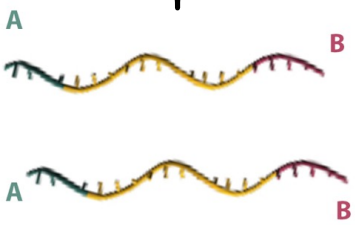
→ 4 individual sequences

primer A | primer B | insert (to be sequenced) | barcode1 | barcode2

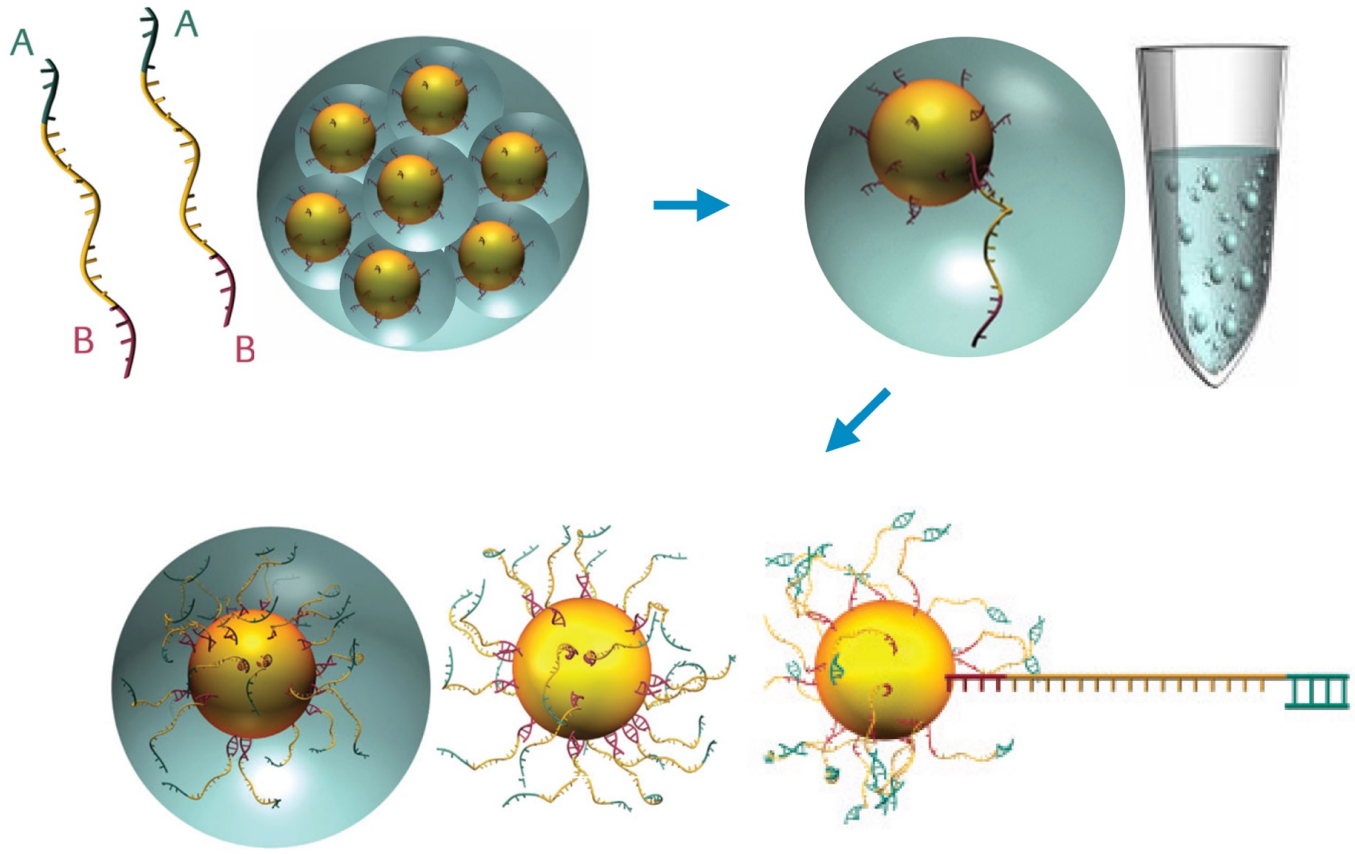
NGS Library Types

Shotgun	Blunt end	Mate pair	Amplicons
▪ whole genome ▪ full length transcripts ▪ BACs ▪ long range PCR	▪ ncRNA (e.g. miRNAs) ▪ ancient DNA ▪ short DNA fragments ▪ short PCR fragments	▪ de novo assembly ▪ structural variants	▪ short targeted region ▪ cancer gene panel ▪ HIV typing ▪ metagenomics
			

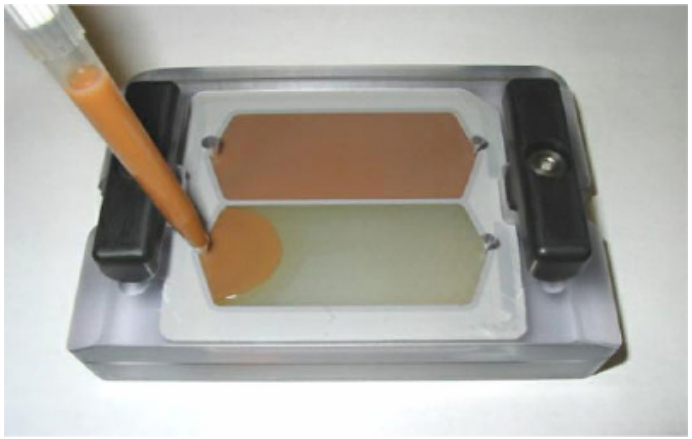
- ssDNA Library
- 20-800 bp insert size
- ready for sequencing

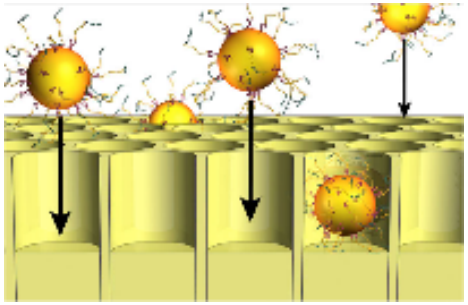


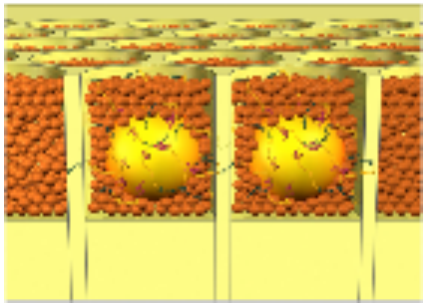
Preamplification of Libraries

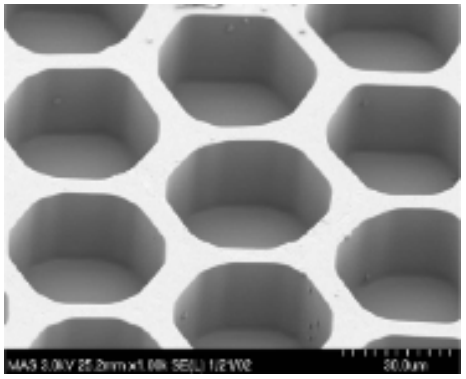


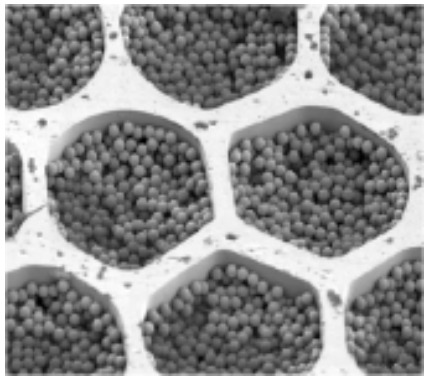
Picotiter Plate



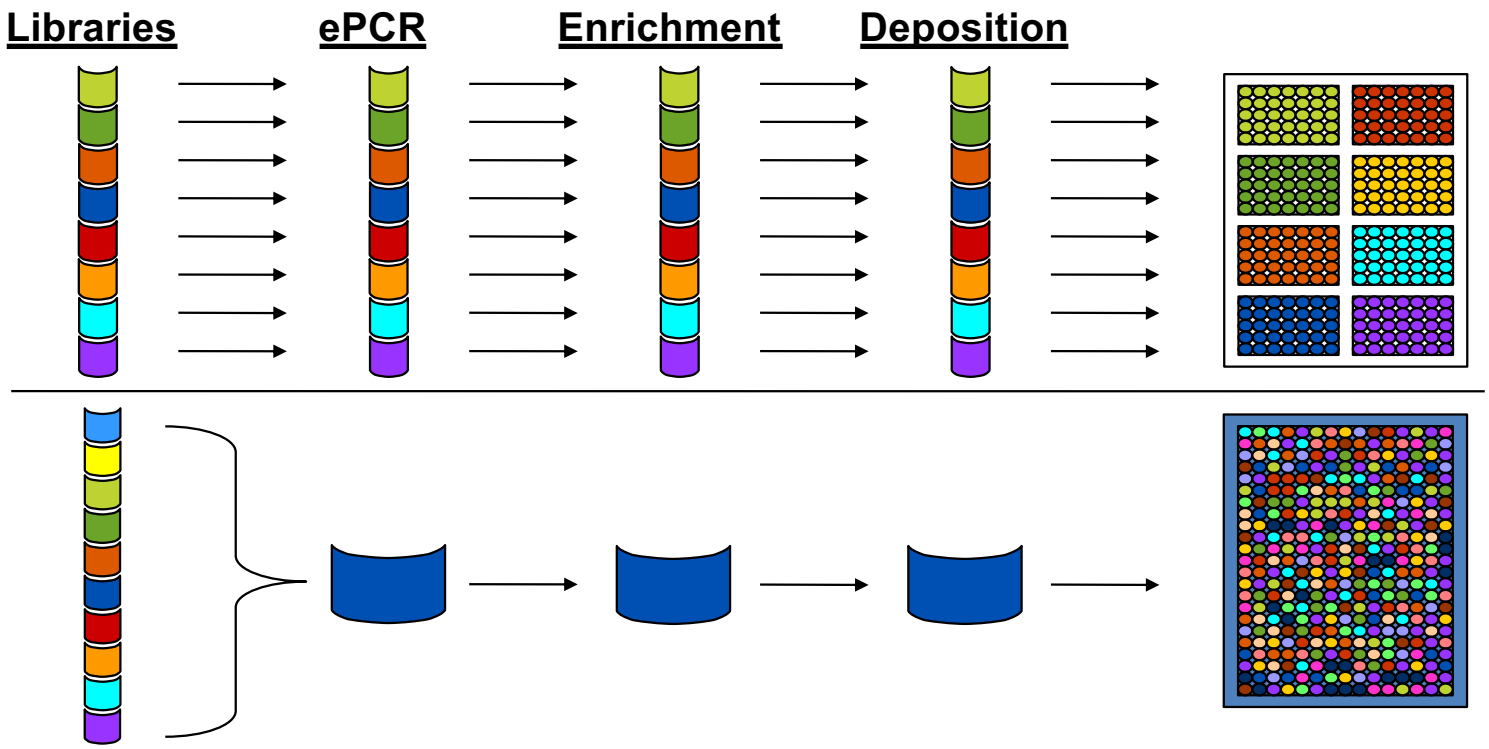








Multiplexing – Barcoding



Ion Torrent Sequencing | Technology

UNIVERSITÄT BERN

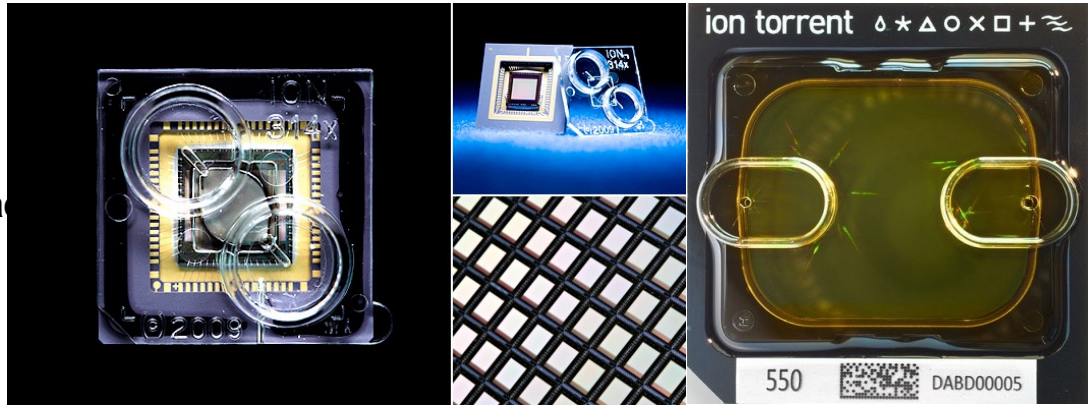
- Short read to long read technology (low to high throughput)
- Clonal amplification by emulsion PCR (ePCR)
- Sequencing by synthesis (SBS) – pyrosequencing
- Label free!

IonTorrent | Overview

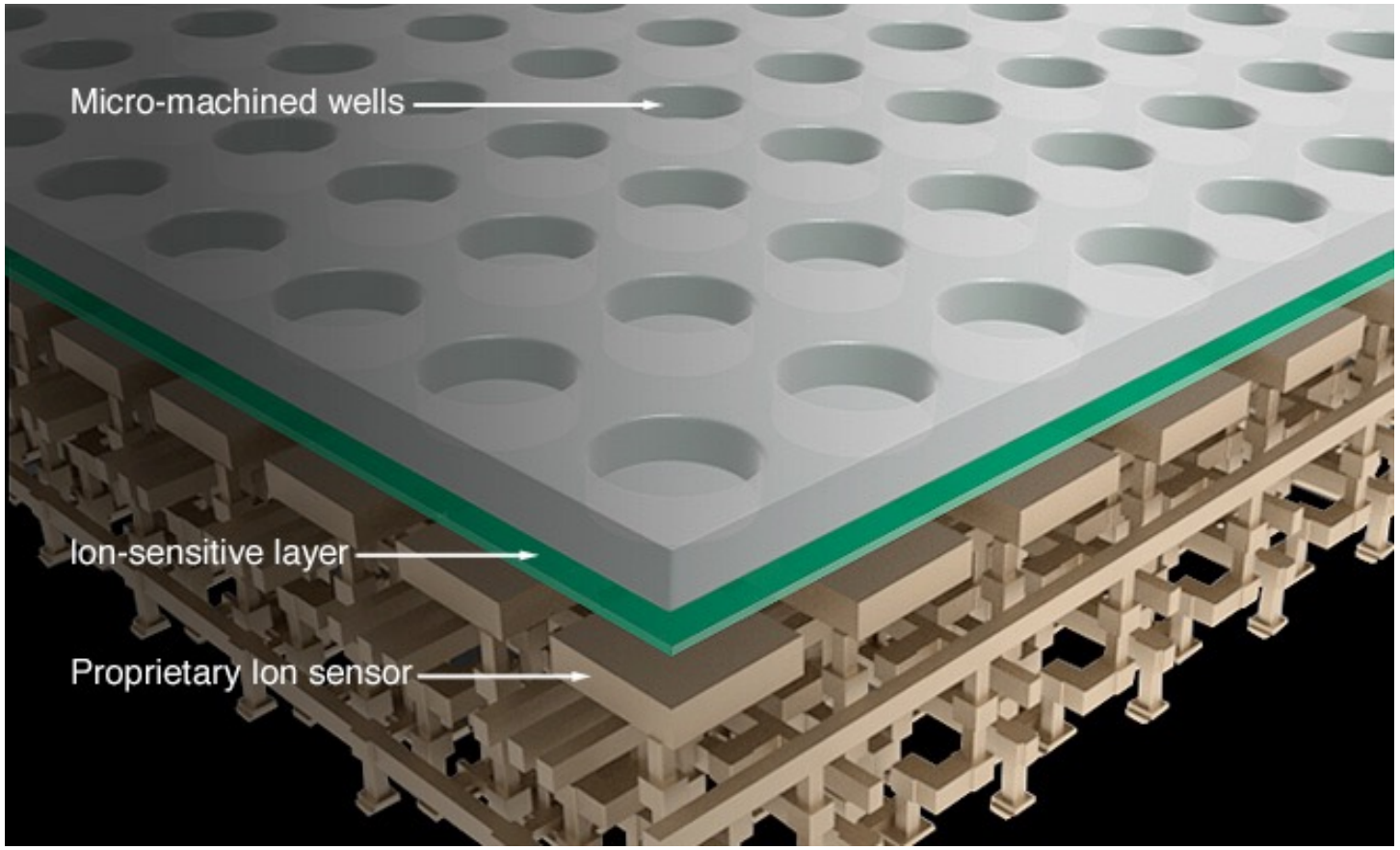
- Ion semiconductor sequencing chips
- The simplest sequencing chemistry
  – natural nucleotides, no enzymatic casca
- True semiconductor sequencing
- No scanners, no cameras, no light
  —The chip is the machine.

- ***emulsion-PCR still required!***

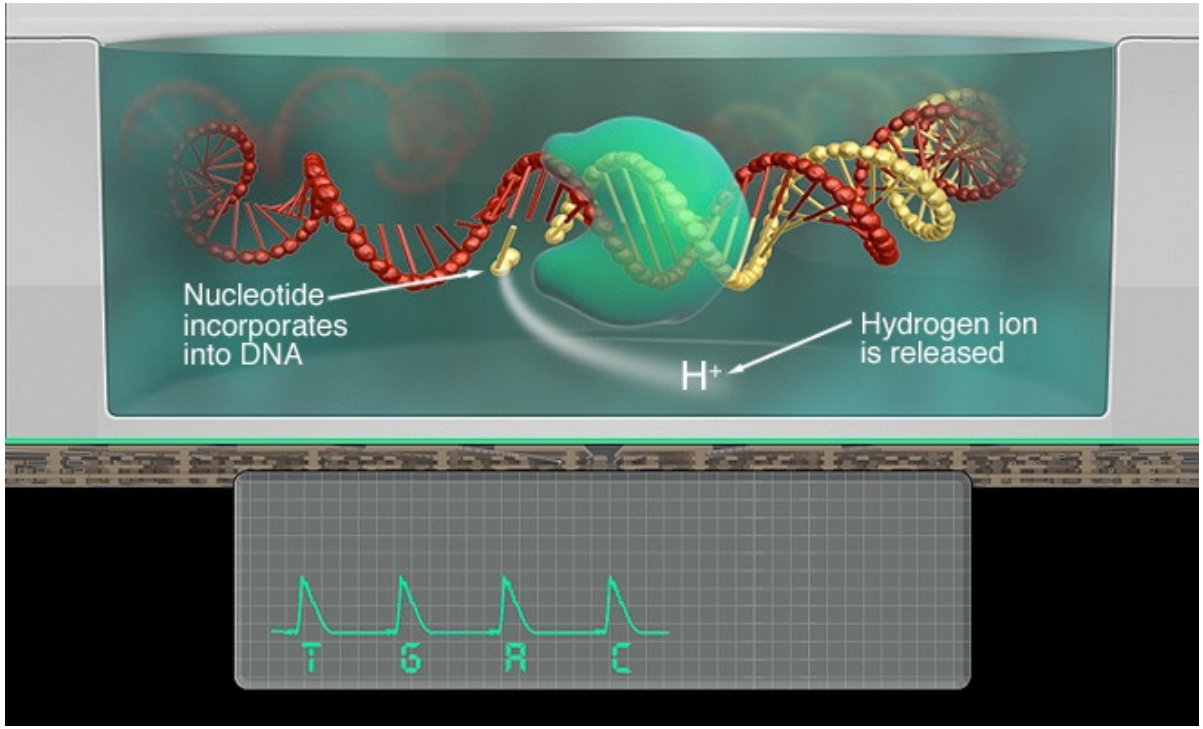
- Instrument costs:
  1/5 of traditional sequencers ($125'000)



World's Smallest pH Meter

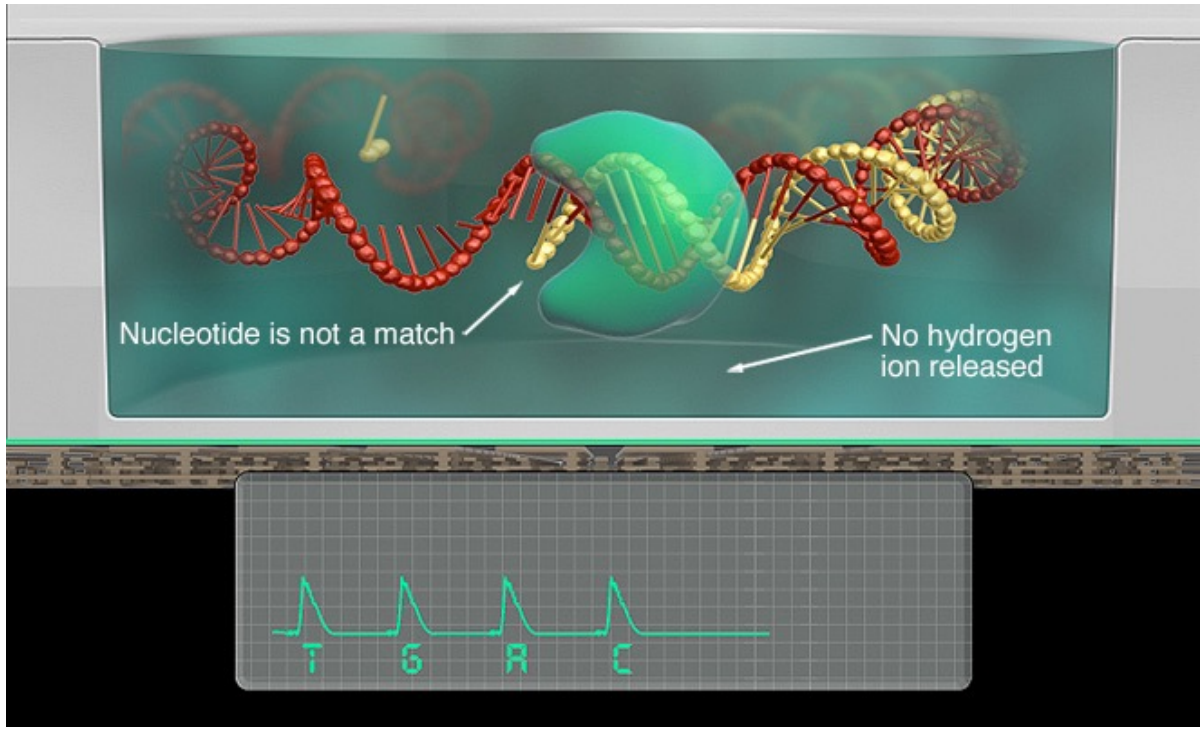


World's Smallest pH Meter

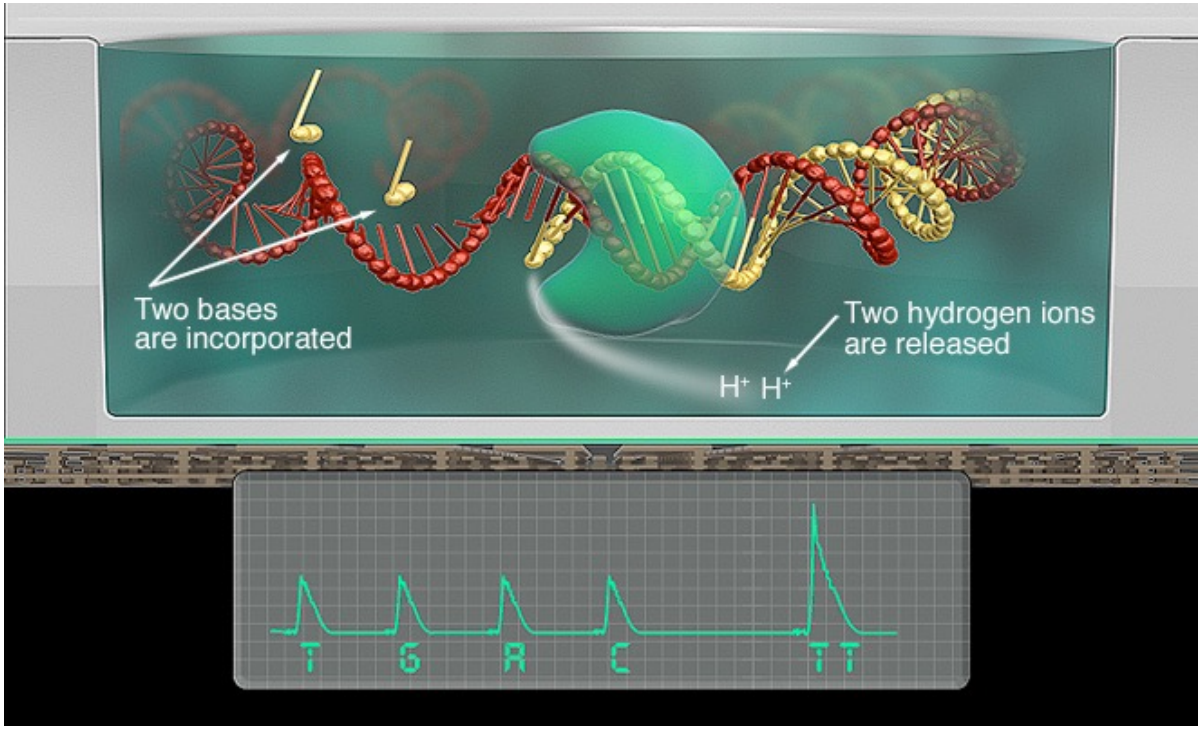


Chemical information is directly transferred into digital information

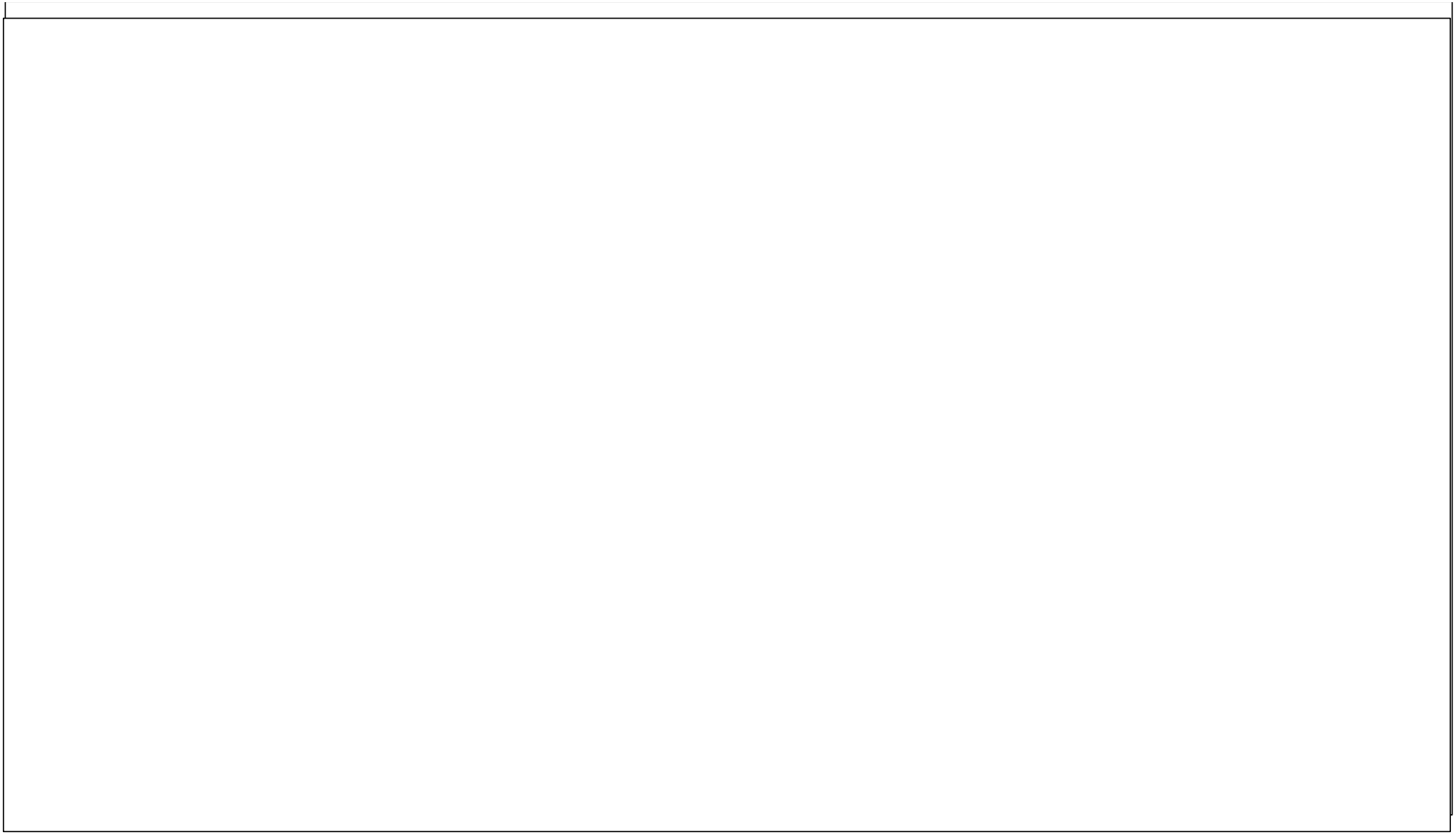
World's Smallest pH Meter

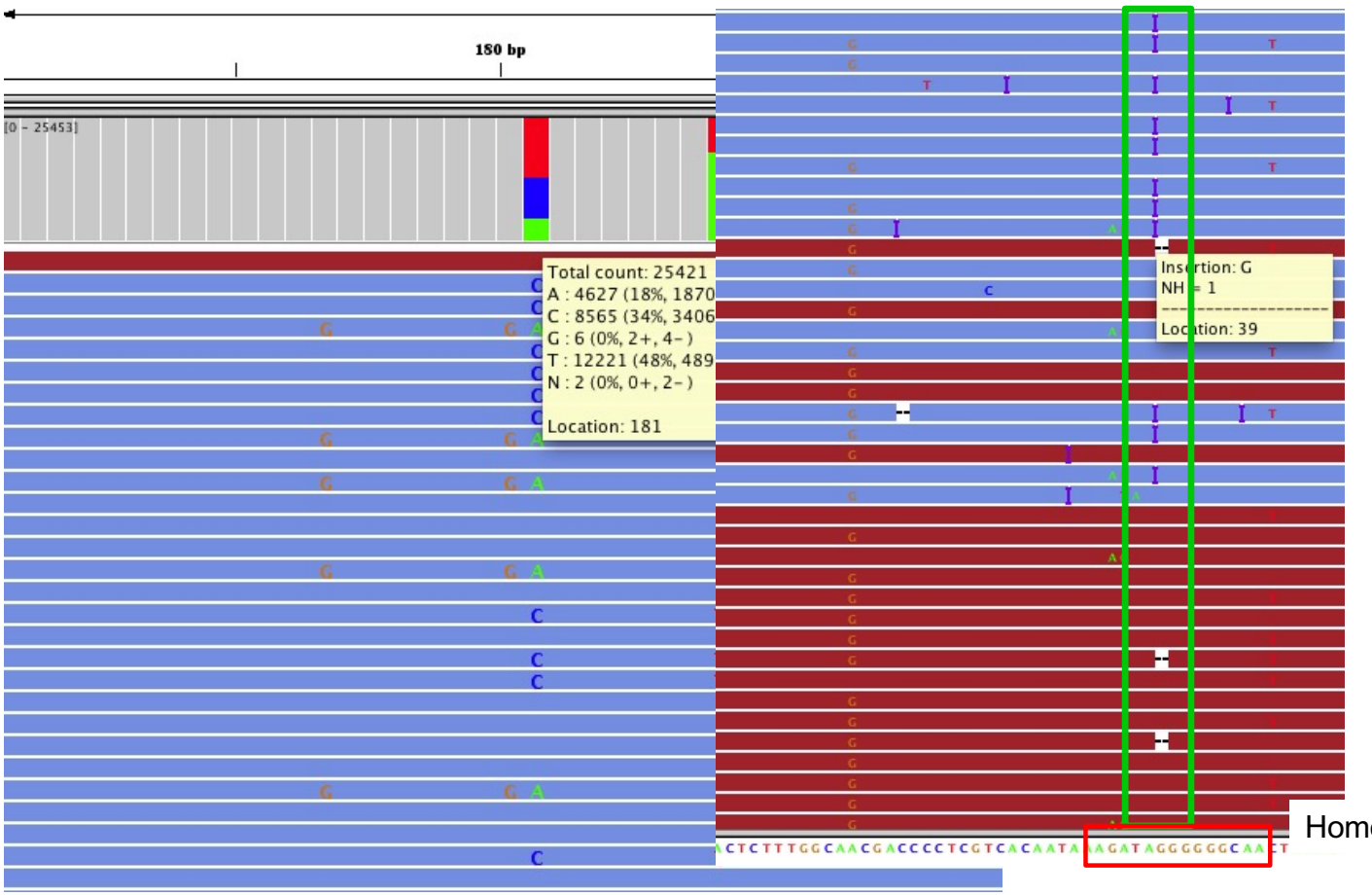


World's Smallest pH Meter



Homopolymers are problematic and the major source of errors





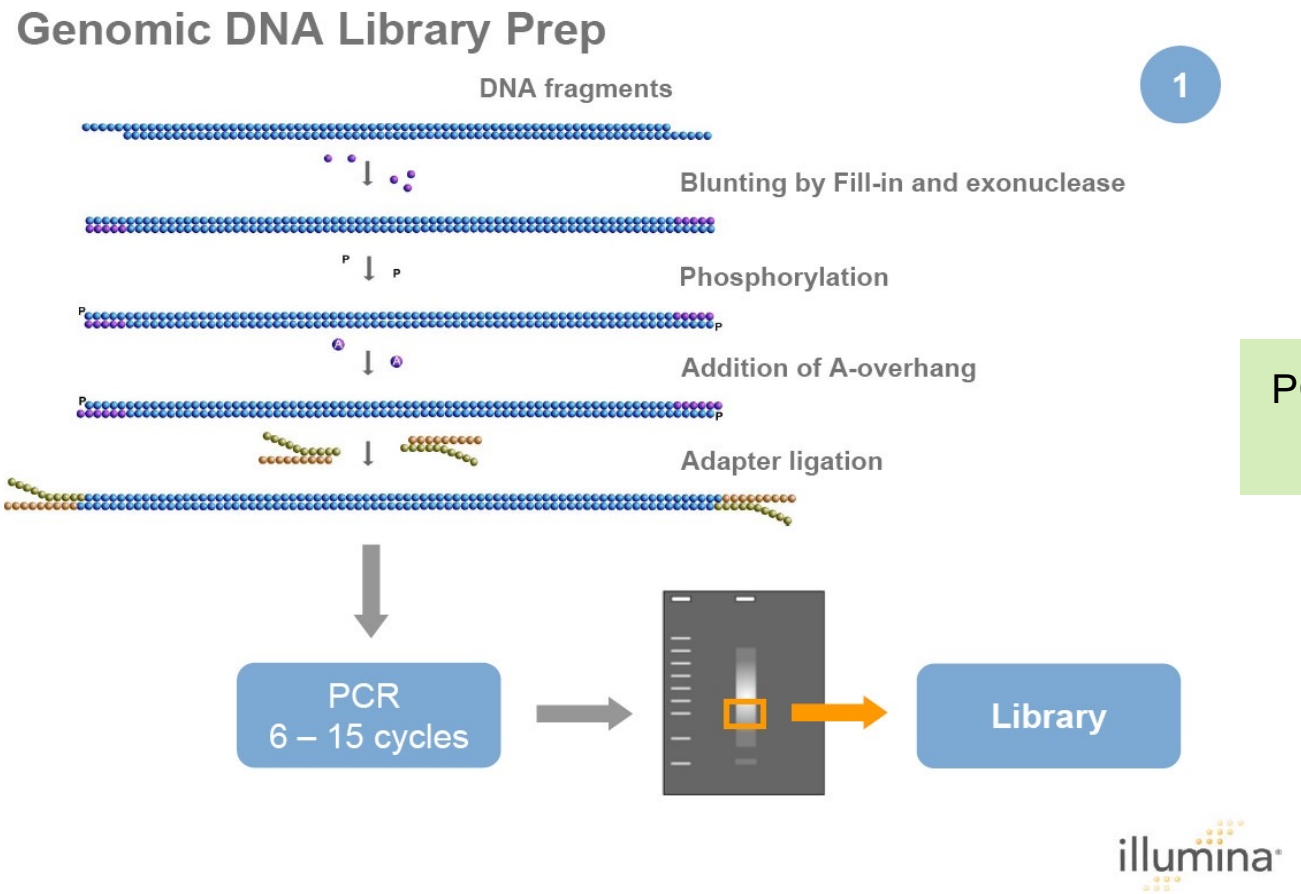
Homopolymer

Illumina Sequencing | Technology

UNIVERSITÄT BERN

- Short read sequencing technology | low to medium to extremely high throughput
- Clonal amplification by cluster/bridge PCR
- Sequencing by synthesis (SBS)– one base at a time

Illumina Sequencing | Technology



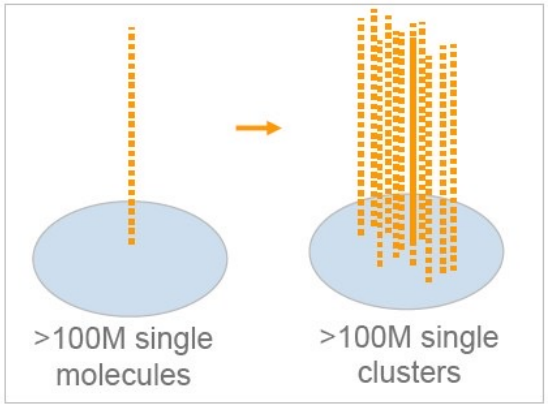
PCR-free Genomic DNA Libraries are now standard for high quality DNA

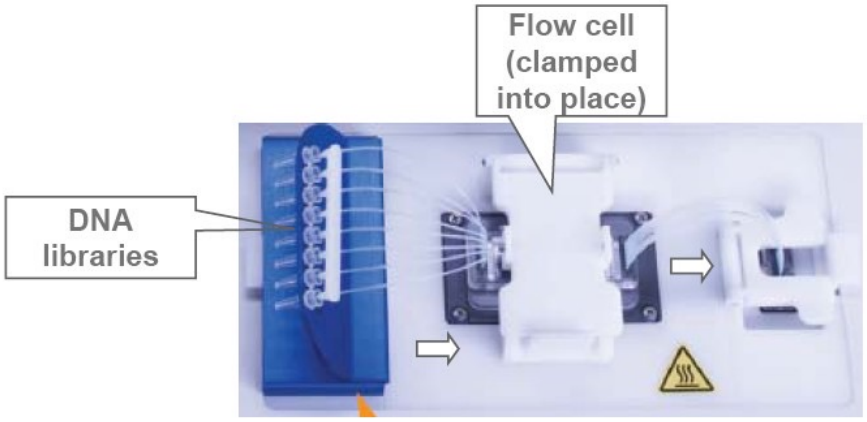
Illumina Sequencing | Technology

Cluster Generation

Cluster station

- Aspirates DNA samples into flow cell
- Automated amplified clonal clusters





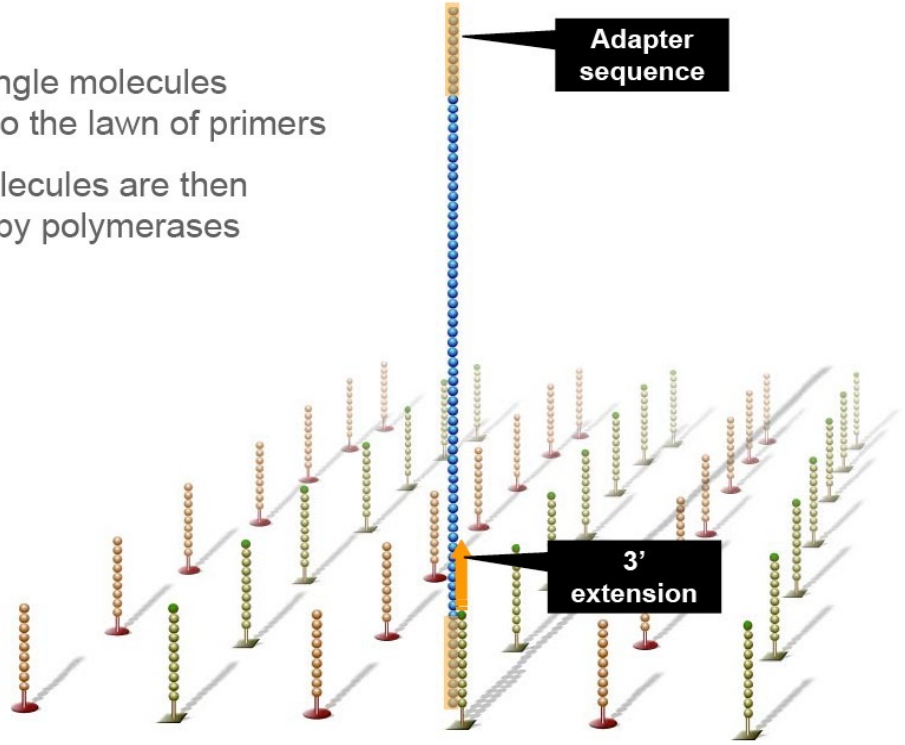
Cluster generation is integrated in the new instruments

Illumina Sequencing | Cluster Generation/PCR

Cluster Generation

Hybridize Fragment & Extend

- >150 M single molecules hybridize to the lawn of primers
- Bound molecules are then extended by polymerases

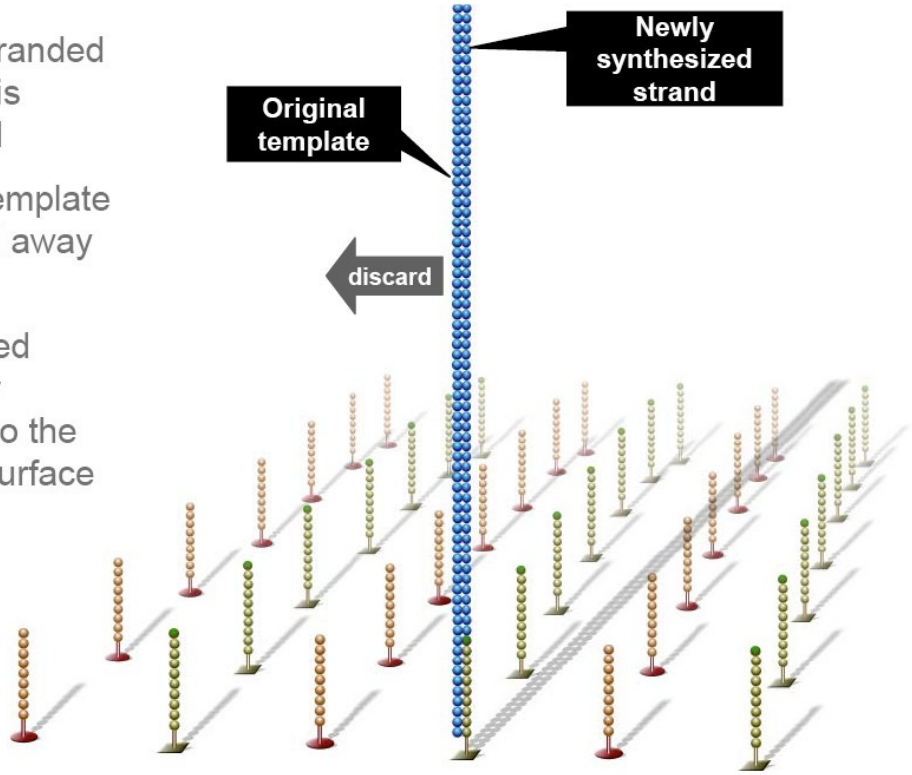


Illumina Sequencing | Cluster Generation/PCR

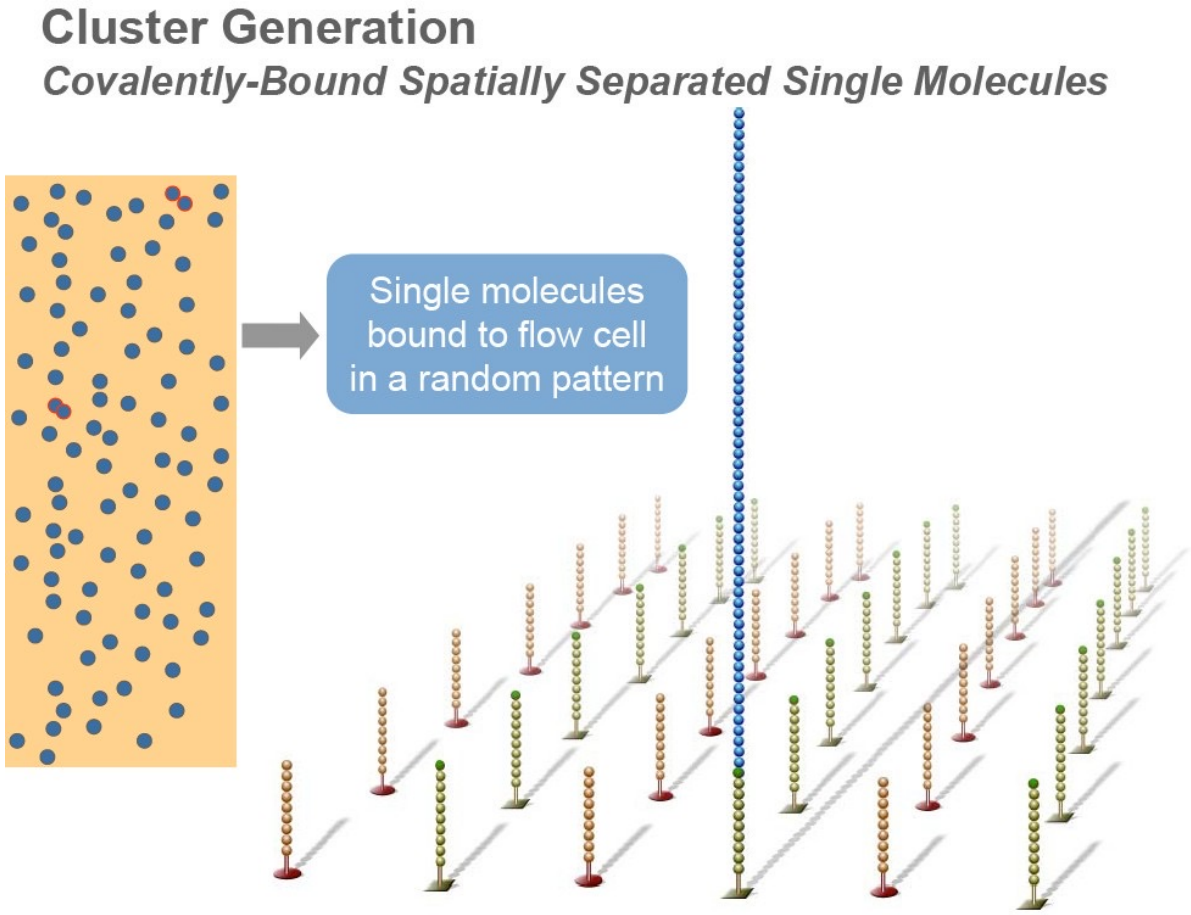
Cluster Generation

Denature Double-stranded DNA

- Double-stranded molecule is denatured
- Original template is washed away
- Newly synthesized covalently attached to the flow cell surface



Illumina Sequencing | Cluster Generation/PCR

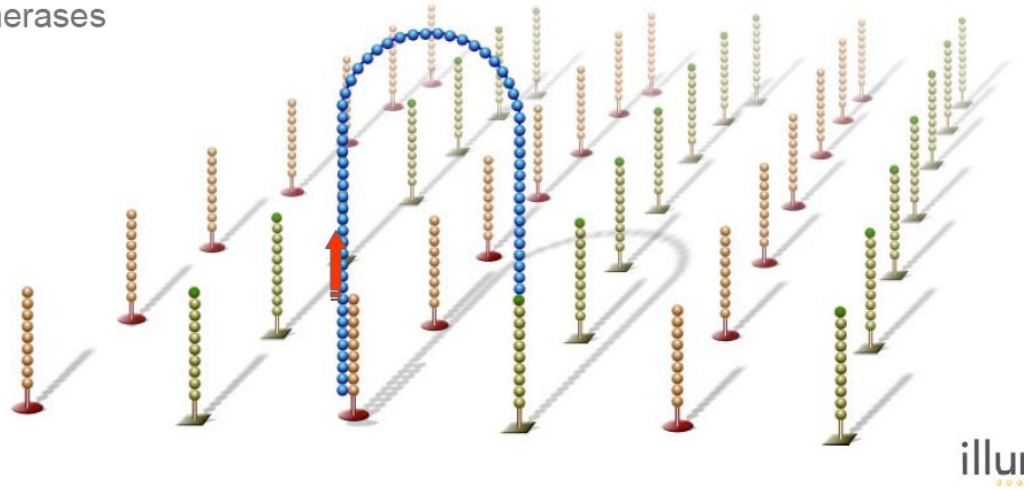


Illumina Sequencing | Cluster Generation/PCR

UNIVERSITÄT BERN

Cluster Generation
Bridge Amplification

- Single-strand flips over to hybridize to adjacent primers to form a bridge
- Hybridized primer is extended by polymerases

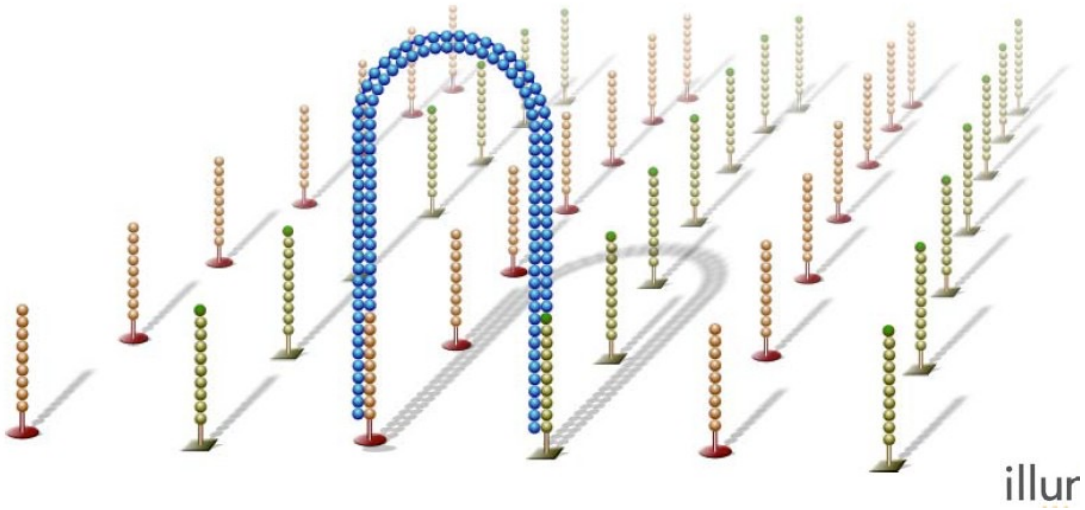


Illumina Sequencing | Cluster Generation/PCR

Cluster Generation

Bridge Amplification

- Double-stranded bridge is formed



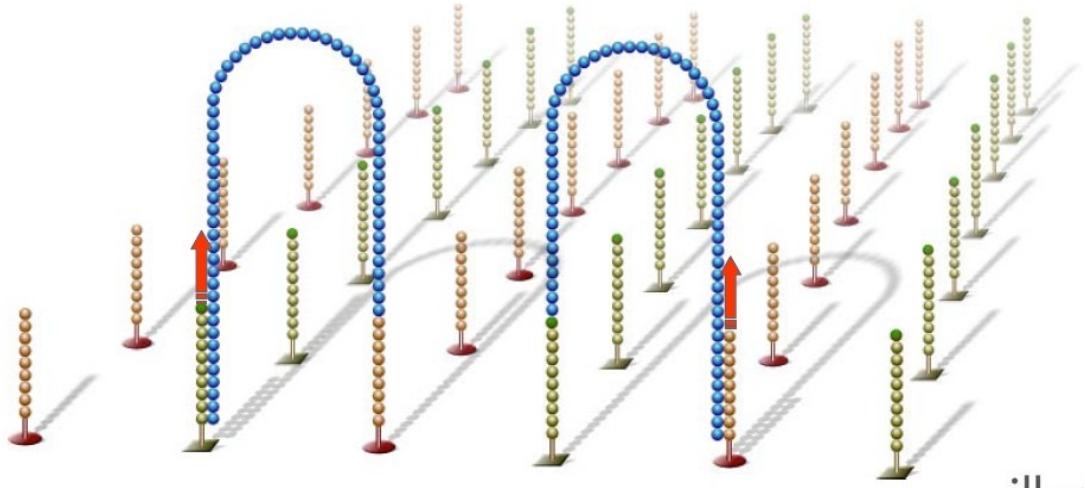
Illumina Sequencing | Cluster Generation/PCR

UNIVERSITÄT BERN

Cluster Generation

Bridge Amplification

- Single-strands flip over to hybridize to adjacent primers to form bridges
- Hybridized primer is extended by polymerase

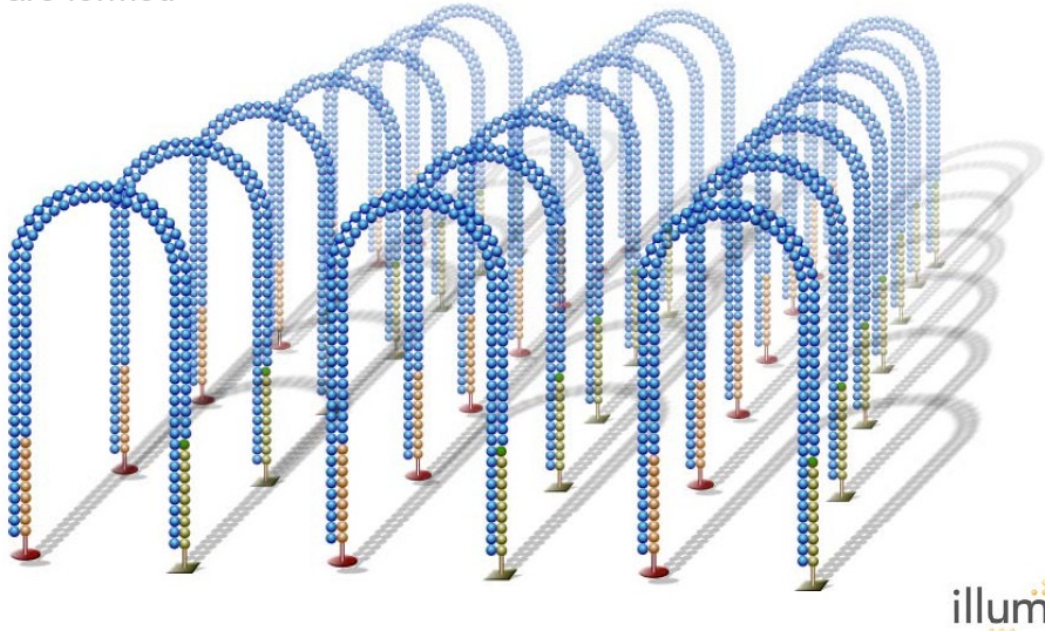


Illumina Sequencing | Cluster Generation/PCR

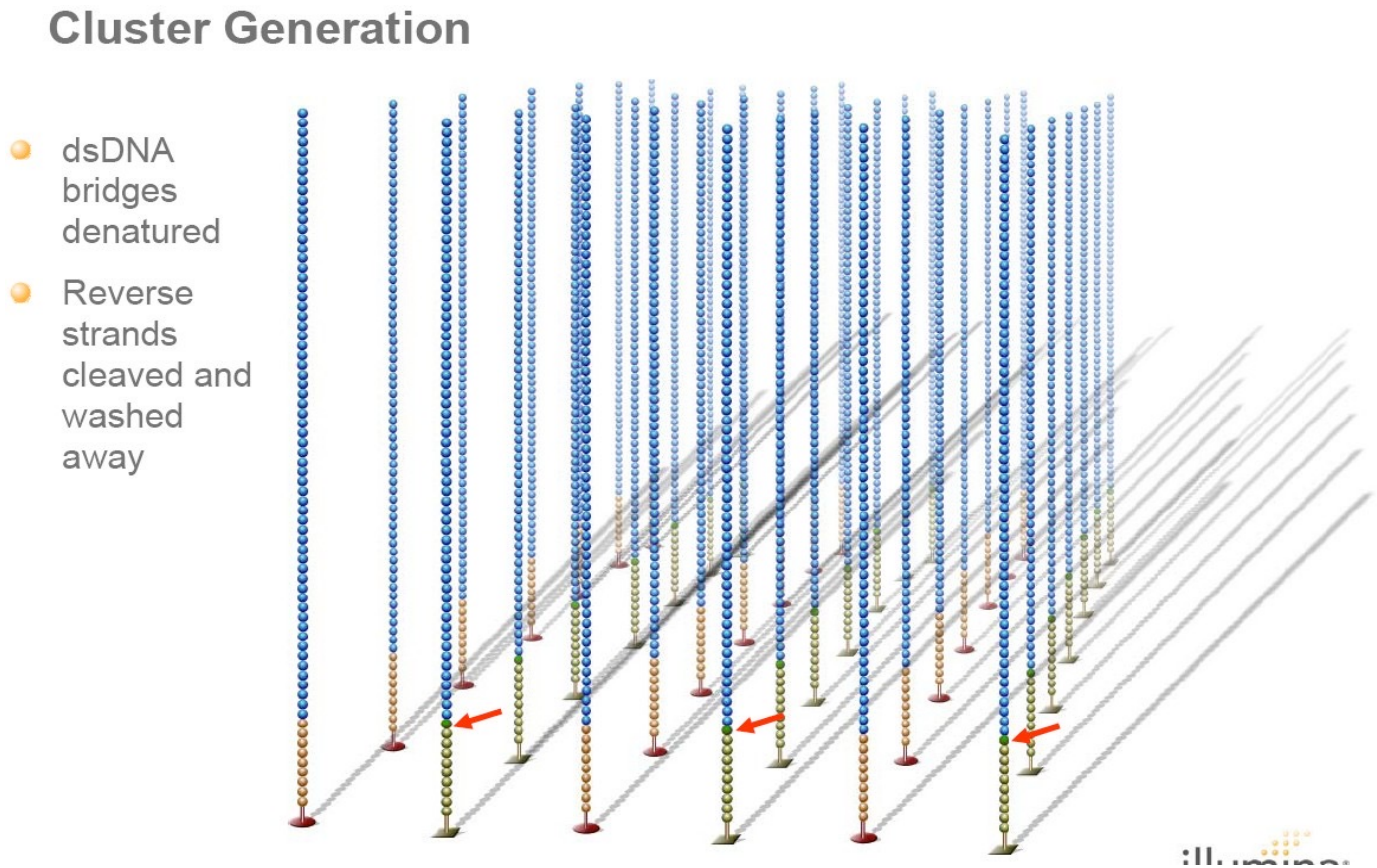
Cluster Generation

Bridge Amplification

- Bridge amplification cycle repeated until multiple bridges are formed

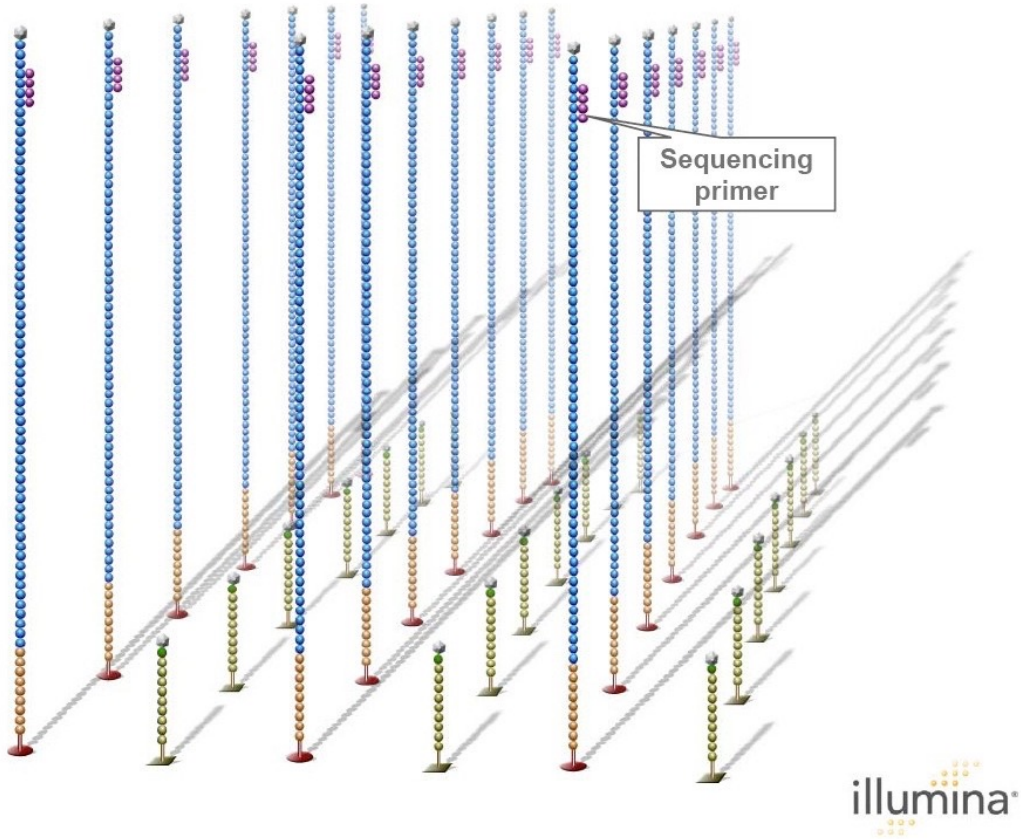


Illumina Sequencing | Cluster Generation/PCR

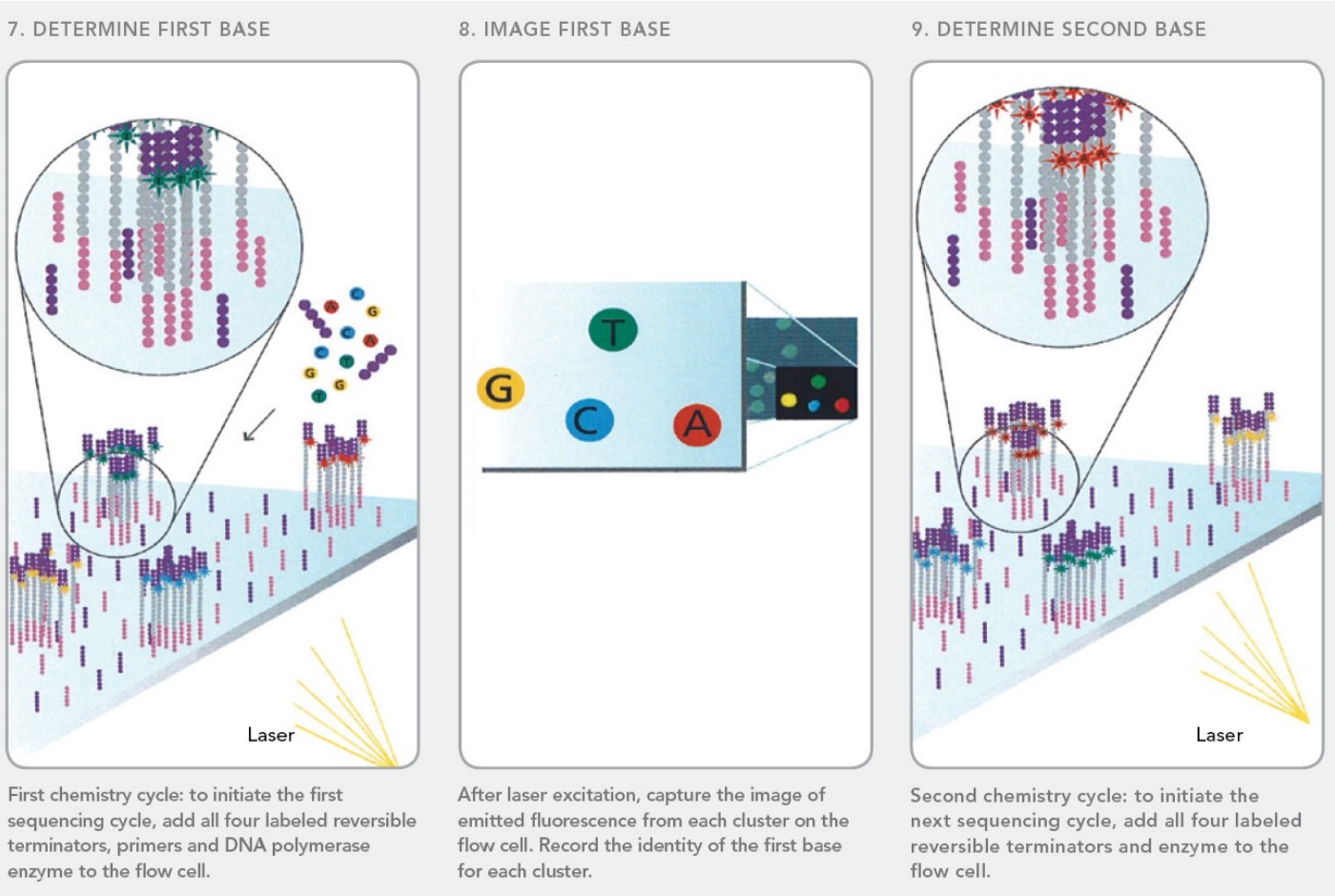


Illumina Sequencing | Sequencing Reaction

- Sequencing primer is hybridized to adapter sequence



Illumina Sequencing | Sequencing Reaction

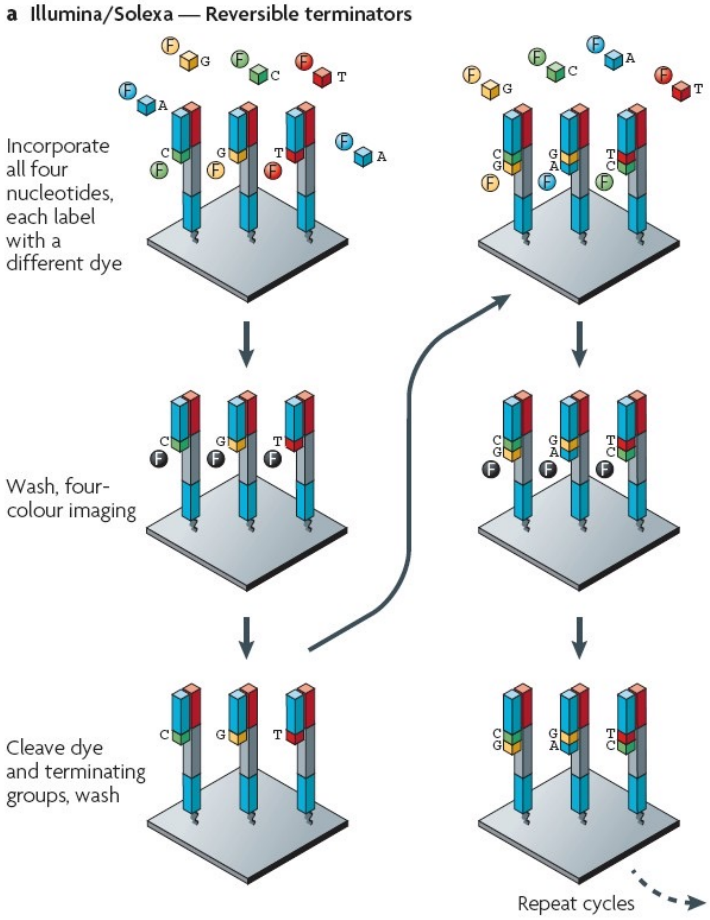


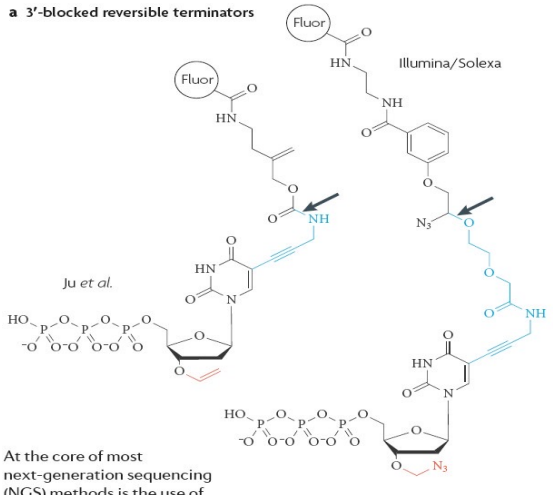
First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

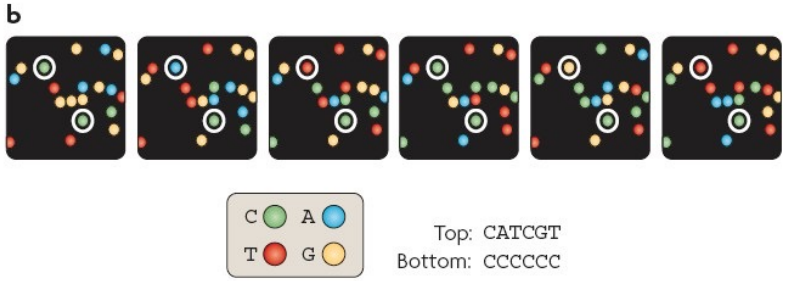
Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

Illumina Sequencing | Sequencing Reaction

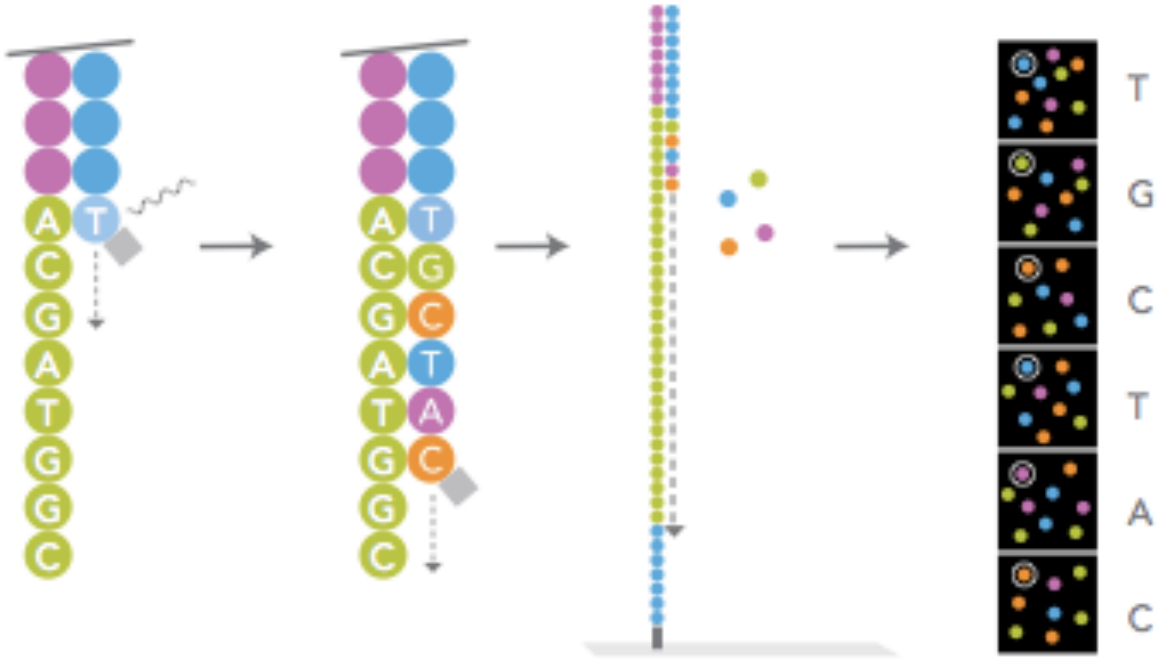




Illumina uses 3‘-O-Azidomethyl as reversible blocking group



Illumina Sequencing | Sequencing Reaction



Sequencing – one base per cycle

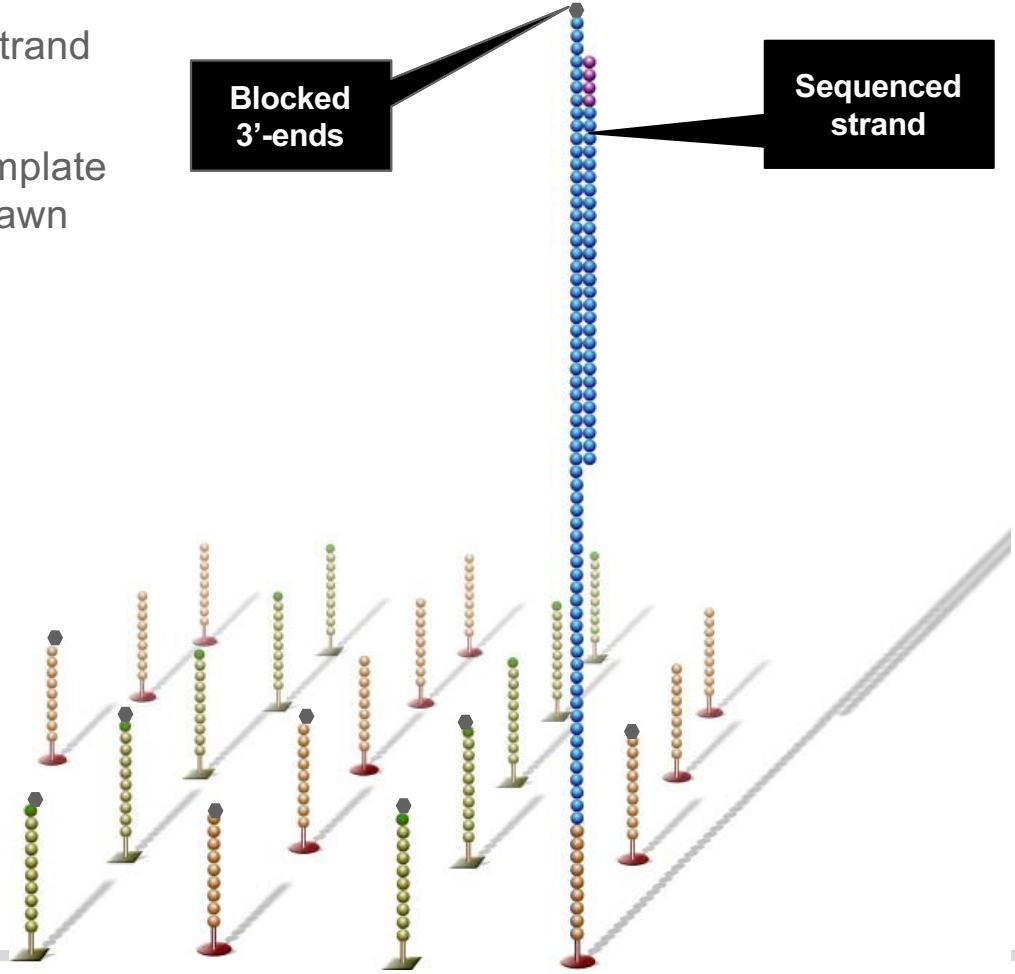
Illumina Sequencing | Paired End Sequencing

At the moment, we only have sequenced from one end of the library,

though we want to sequence from both ends

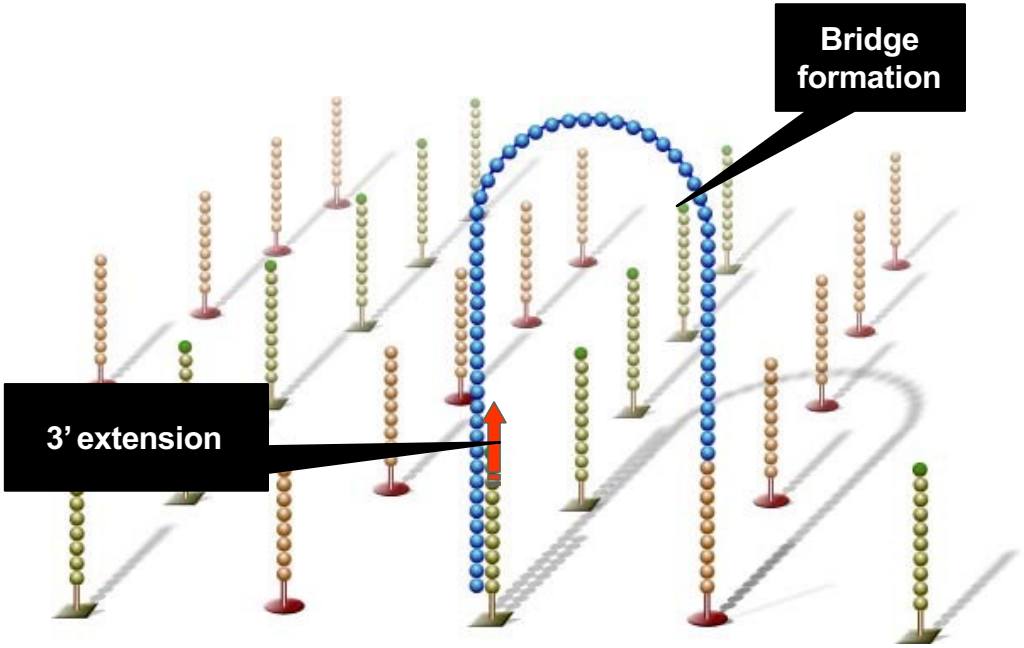
Illumina Sequencing | Paired End Sequencing

- Sequenced strand is tripped off
- 3'-ends of template strands and lawn primers are unblocked



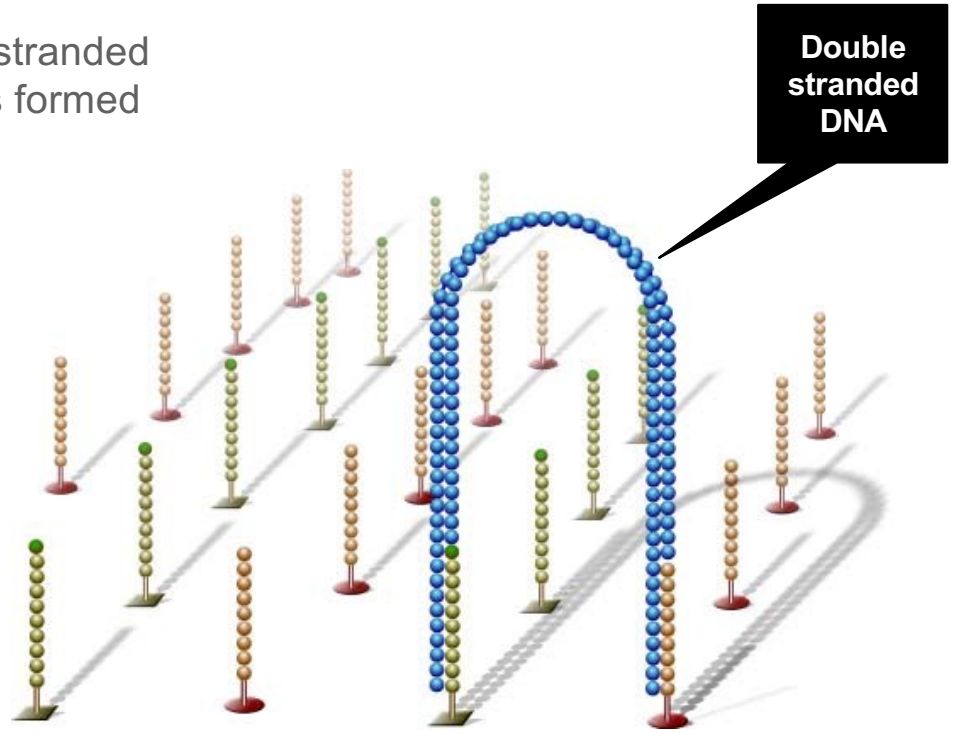
Illumina Sequencing | Paired End Sequencing

- Single-stranded template loops over to form a bridge by hybridizing with a lawn primer
- 3'-ends of lawn primer is extended



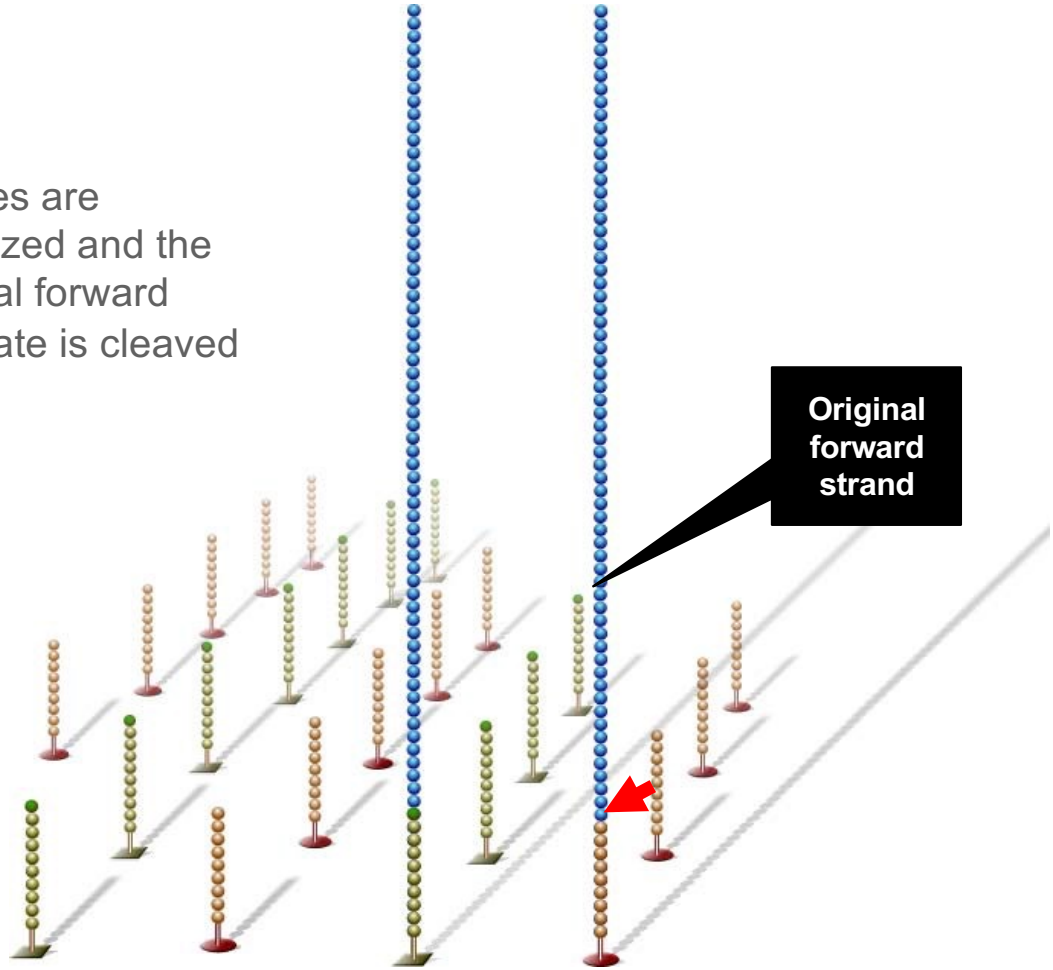
Illumina Sequencing | Paired End Sequencing

- Double-stranded bridge is formed



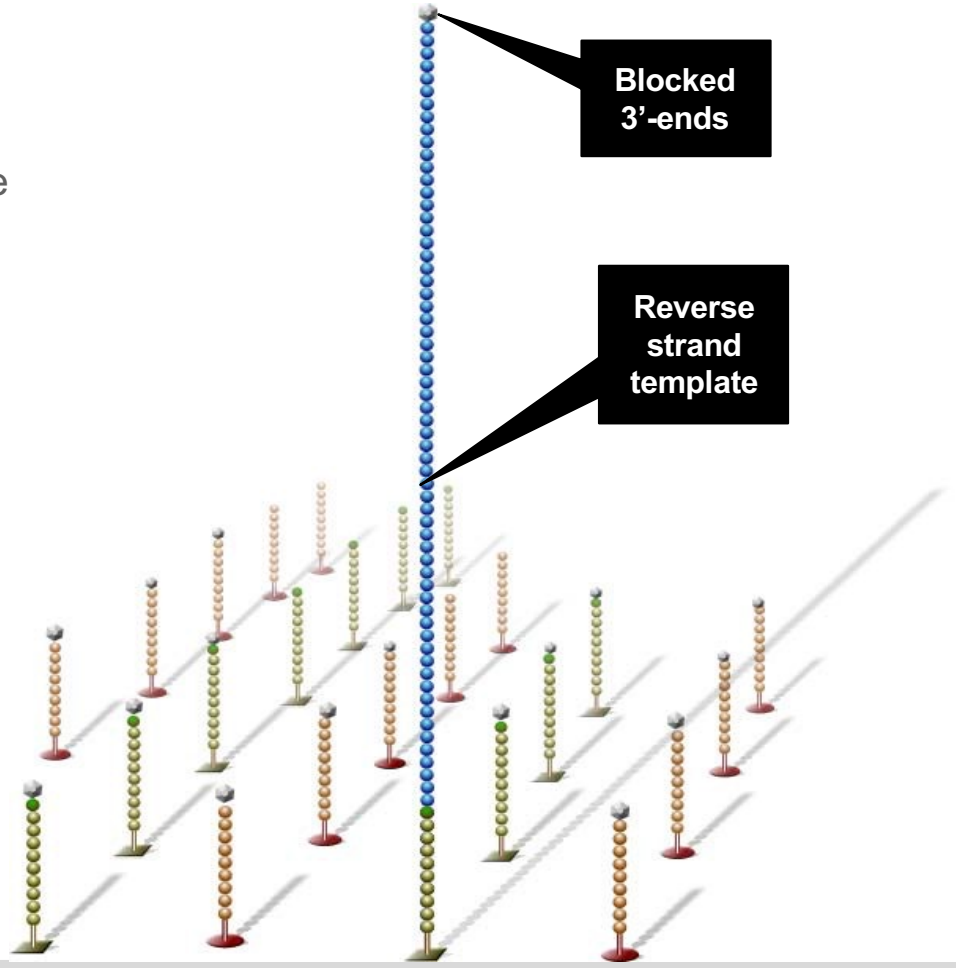
Illumina Sequencing | Paired End Sequencing

- Bridges are linearized and the original forward template is cleaved off

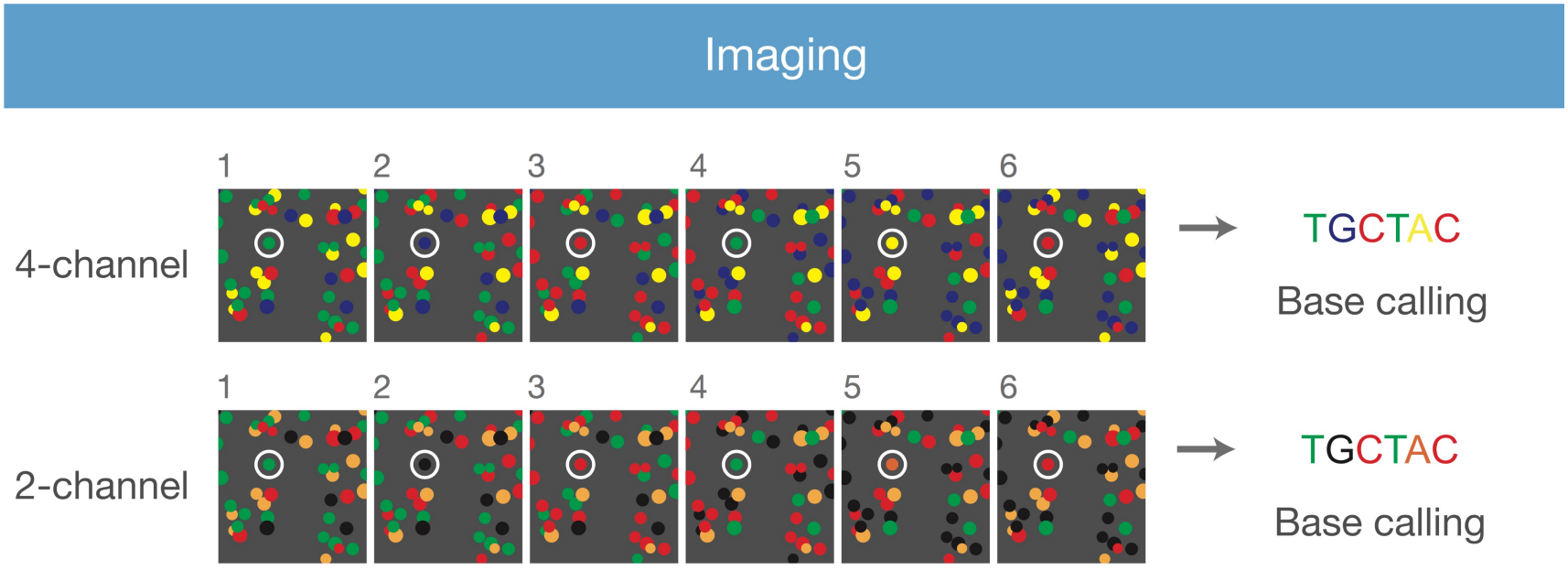


Illumina Sequencing | Paired End Sequencing

- Free 3’ ends of the reverse template and lawn primers are blocked to prevent unwanted DNA priming



Illumina Sequencing | 4-channel vs. 2-channel



Advantages/disadvantages 4-channel vs. 2-channel

UNIVERSITÄT BERN

4-channel

- 4 images taken per cycle
- Twice the data
- Proven accuracy/reliability
- Each base has its own color/signal

all new Instruments use 2-channel

2-channel

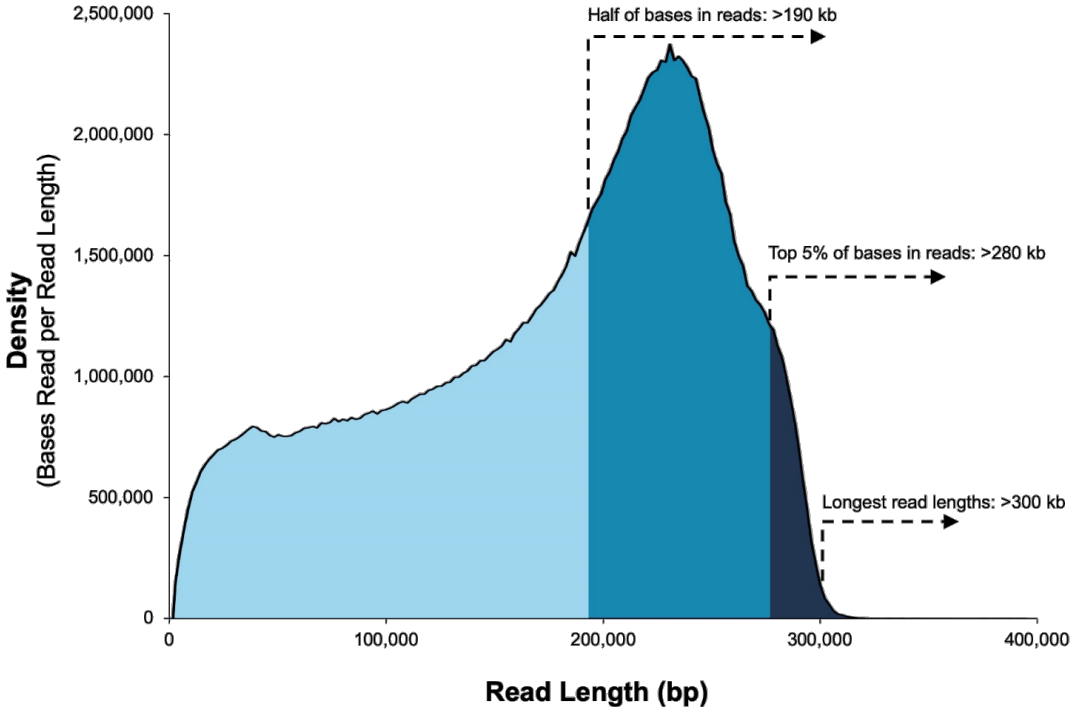
- Only 2 images taken per cycle (2-channel)
- Less storage and computation
- Faster
- Cheaper (camera)
- No-color: problematic → same as background

Pacific Bioscience (PacBio)

UNIVERSITÄT BERN

- 3rd generation sequencing technology
- **single molecule sequencing!** (relatively large amounts of DNA required as input!)
- measures polymerase activity in real time
  (**S**ingle **M**olecule **R**eal **T**ime: **SMRT**)
- can directly detect modified (methylated) nucleotides
- high error rate (random!)
- Very long reads

HIFI Sequencing (20 KB INSERT) – Raw Data



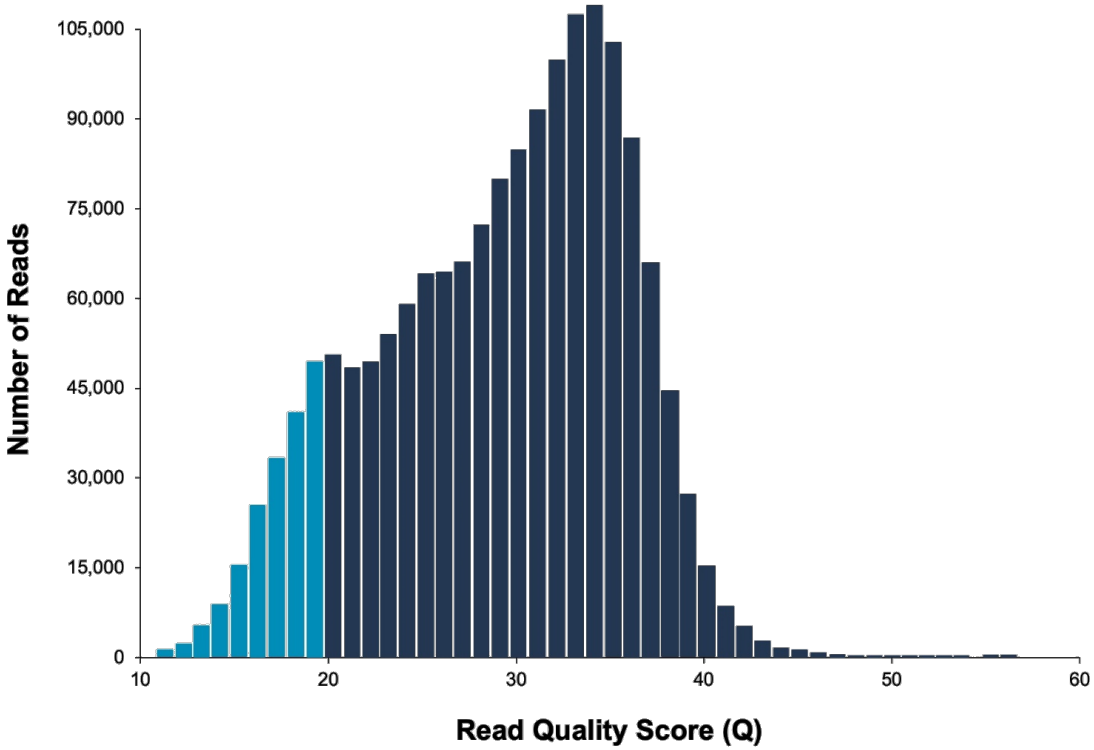
Up to 500 Gb
Customer average: 325 Gb

Metrics	
Insert Size	20 kb
Number of Raw Bases (Gb)	377
Total Reads	4,266,403
Half of Bases in Reads	>193,403
Longest read lengths	>300,000

Data shown above from a 20 kb size-selected human library using the SMRTbell Template Prep Kit on a Sequel II System (2.0 Chemistry, Sequel II System Software v8.0, 30-hour movie). Read lengths, reads/data per SMRT Cell 8M and other sequencing performance results vary based on sample quality/type and insert size.

from PacBio

HIFI Sequencing – Post-CSS processing



Q20 (99%) HiFi accuracy

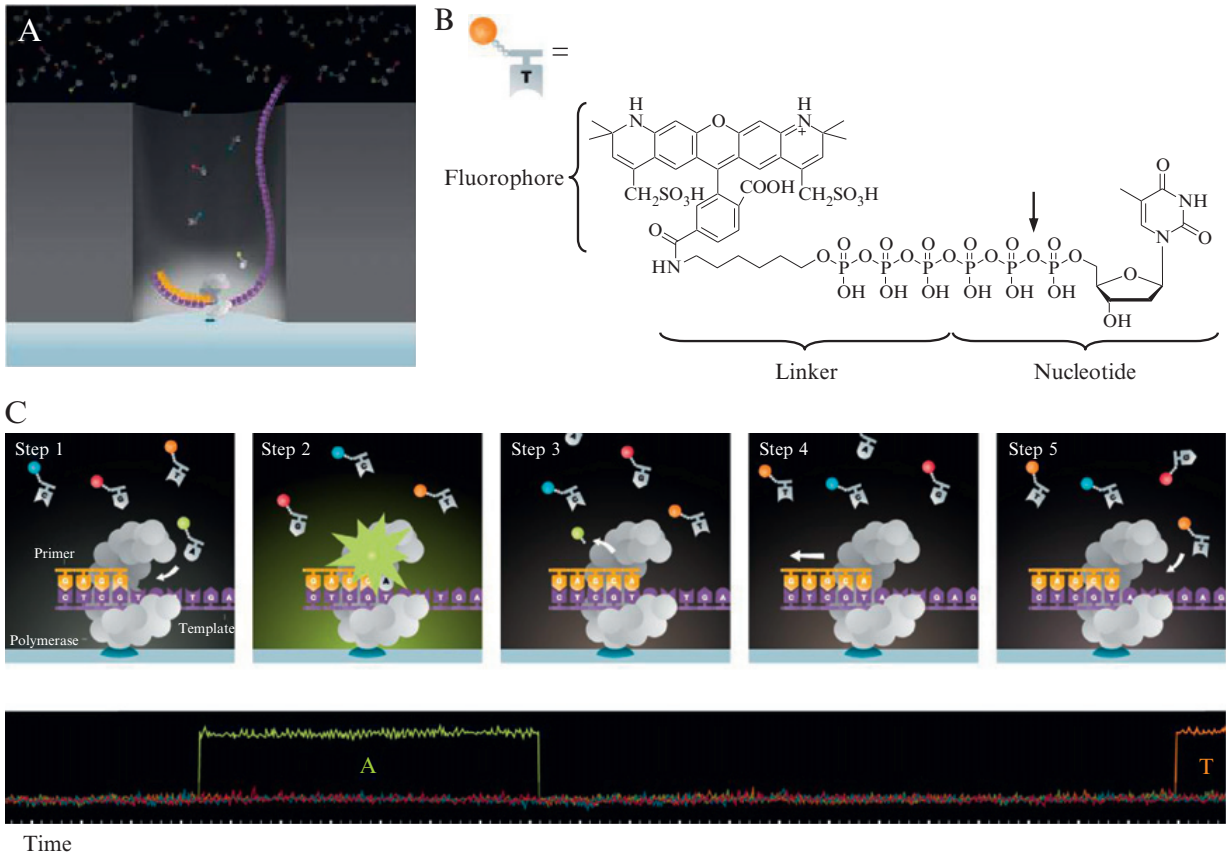
Metrics	
Insert Size	20 kb
Number of CCS Bases (Gb)	30 Gb
Number of >Q20 Bases	26 Gb
Number of >Q20 Reads	1,423,277
Accuracy (Mean)	**99.92%**

Customer average: 25-30 Gb

Data shown above from a 20 kb size-selected human library using the SMRTbell Template Prep Kit on a Sequel II System (2.0 Chemistry, Sequel II System Software v8.0, 30-hour movie). Read lengths, reads/data per SMRT Cell 8M and other sequencing performance results vary based on sample quality/type and insert size.

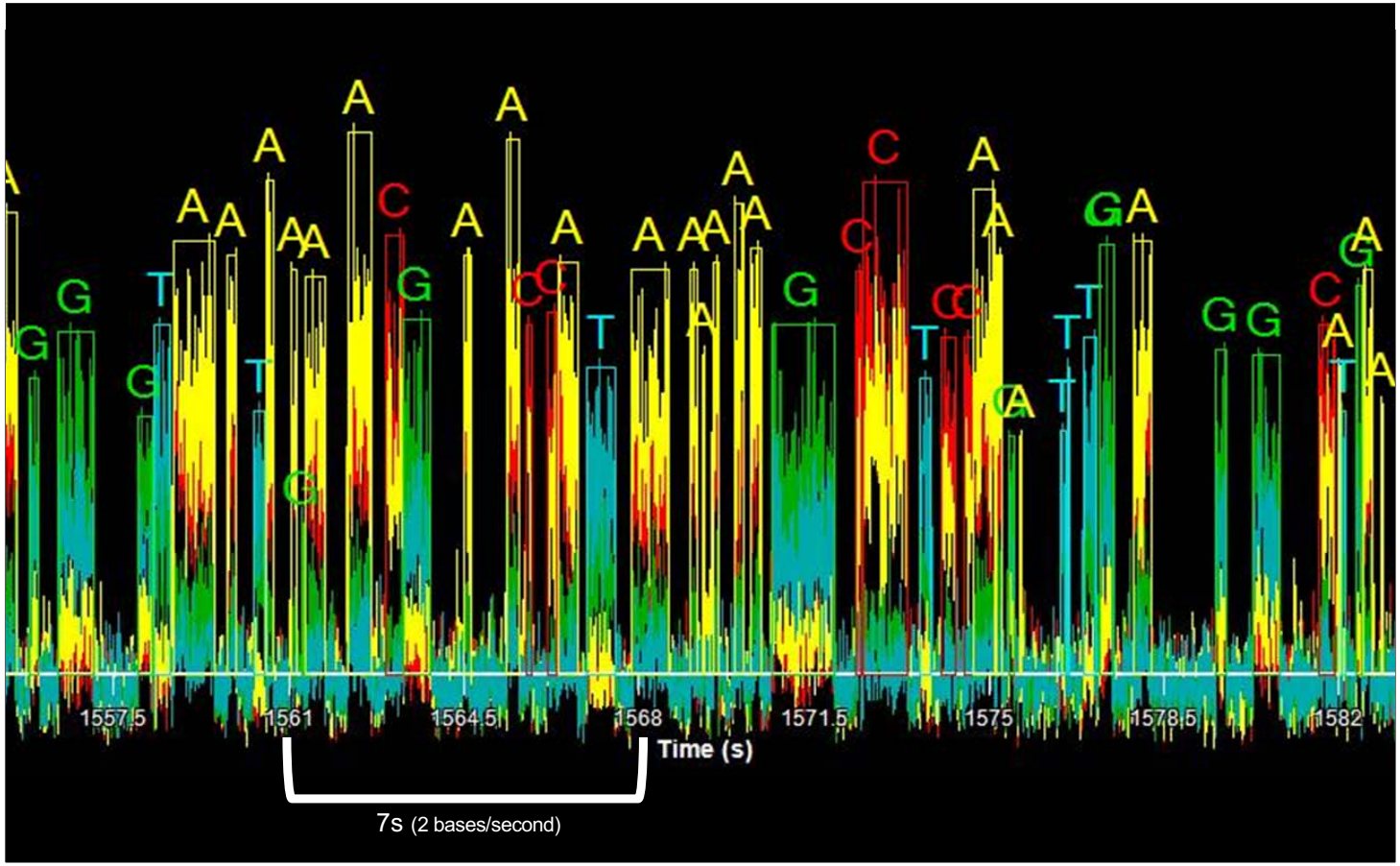
from PacBio

PacBio Sequencing Principle

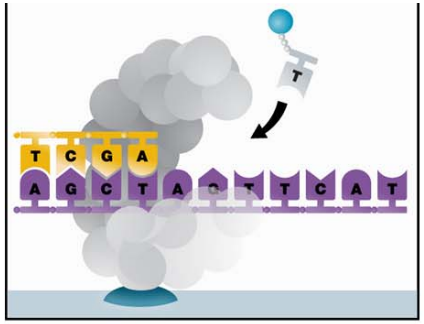


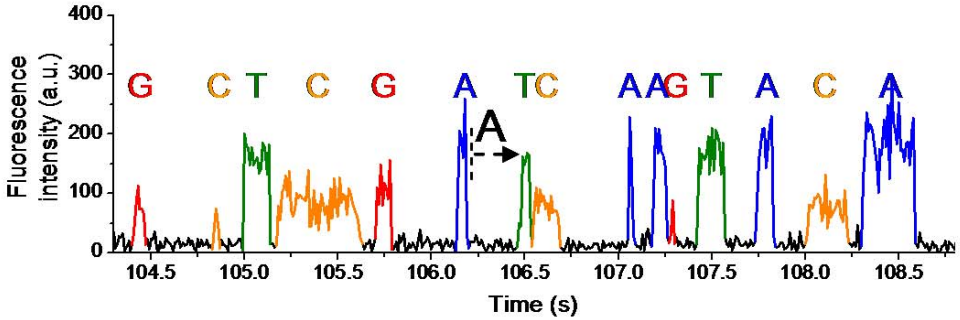
PacBio Sequencing Principle

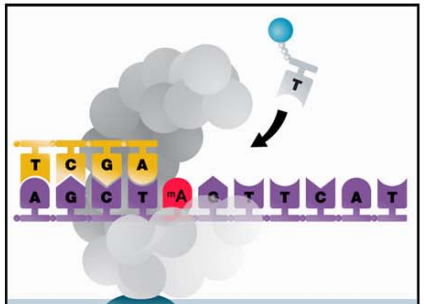
Natural Process Monitored

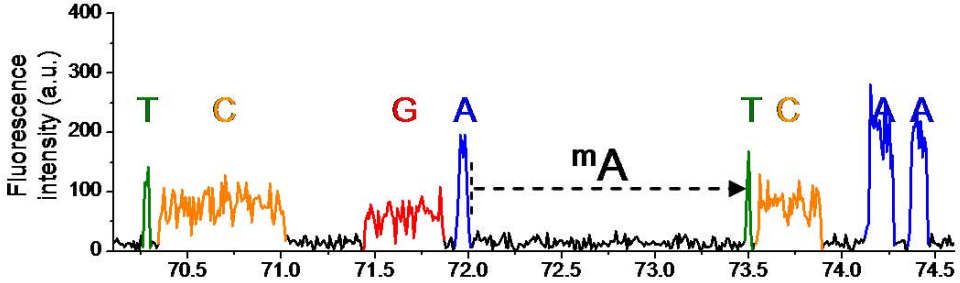


Kinetic Detection Enables Study of Modified Bases

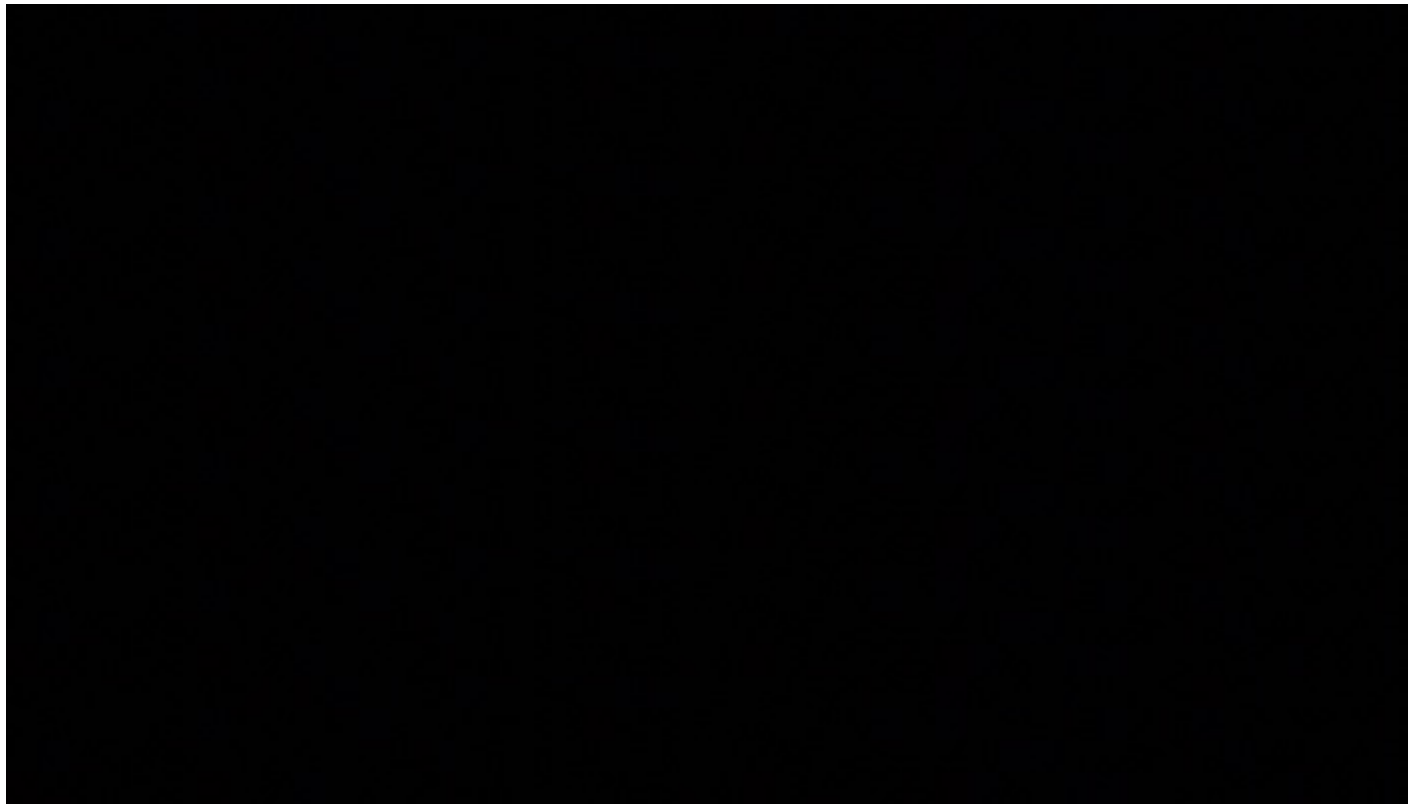




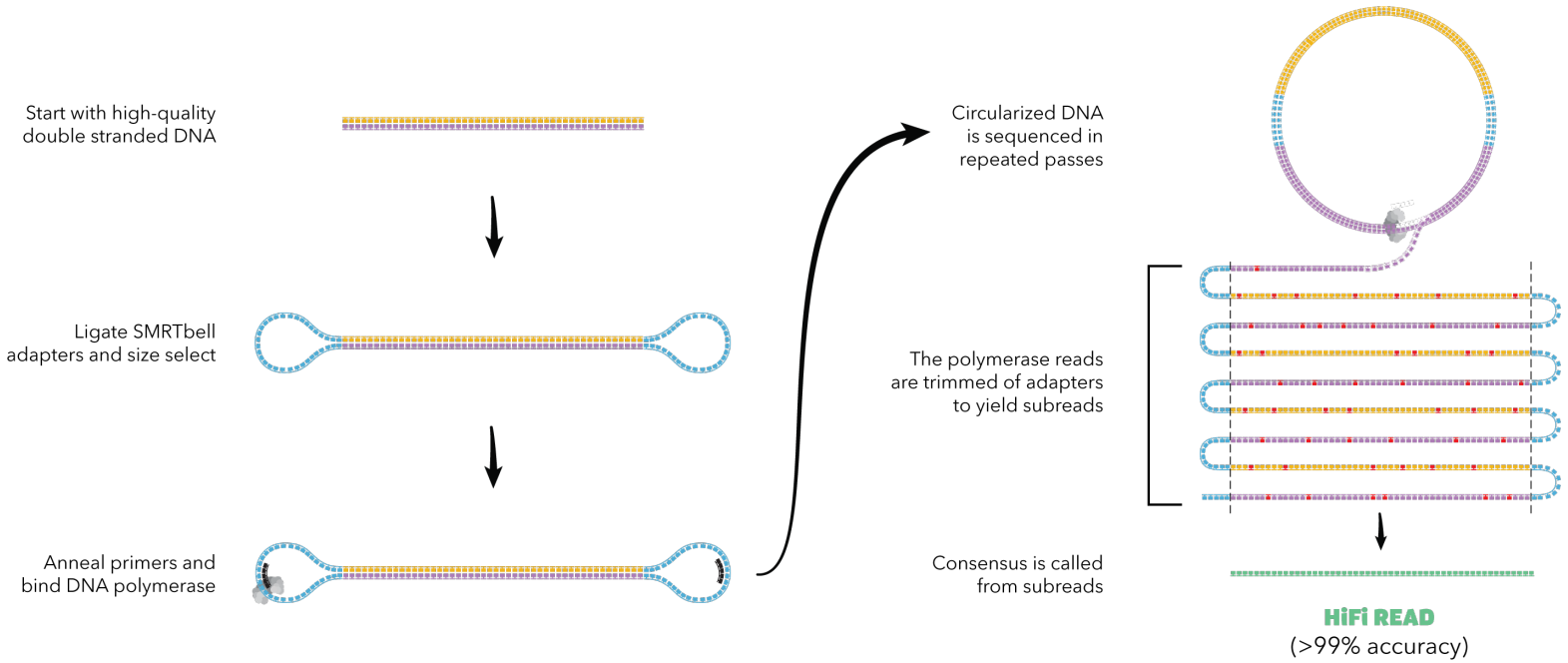




HIFI – SMRT-Bell

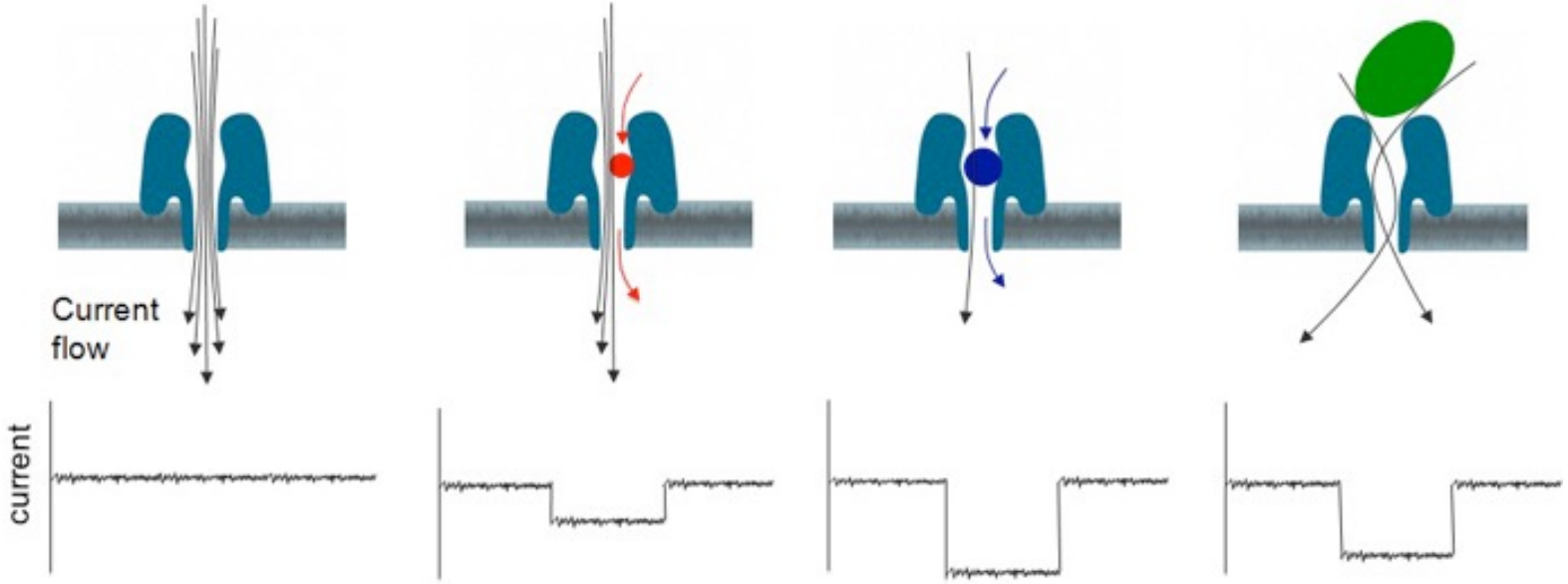


Pacbio HiFi Reads

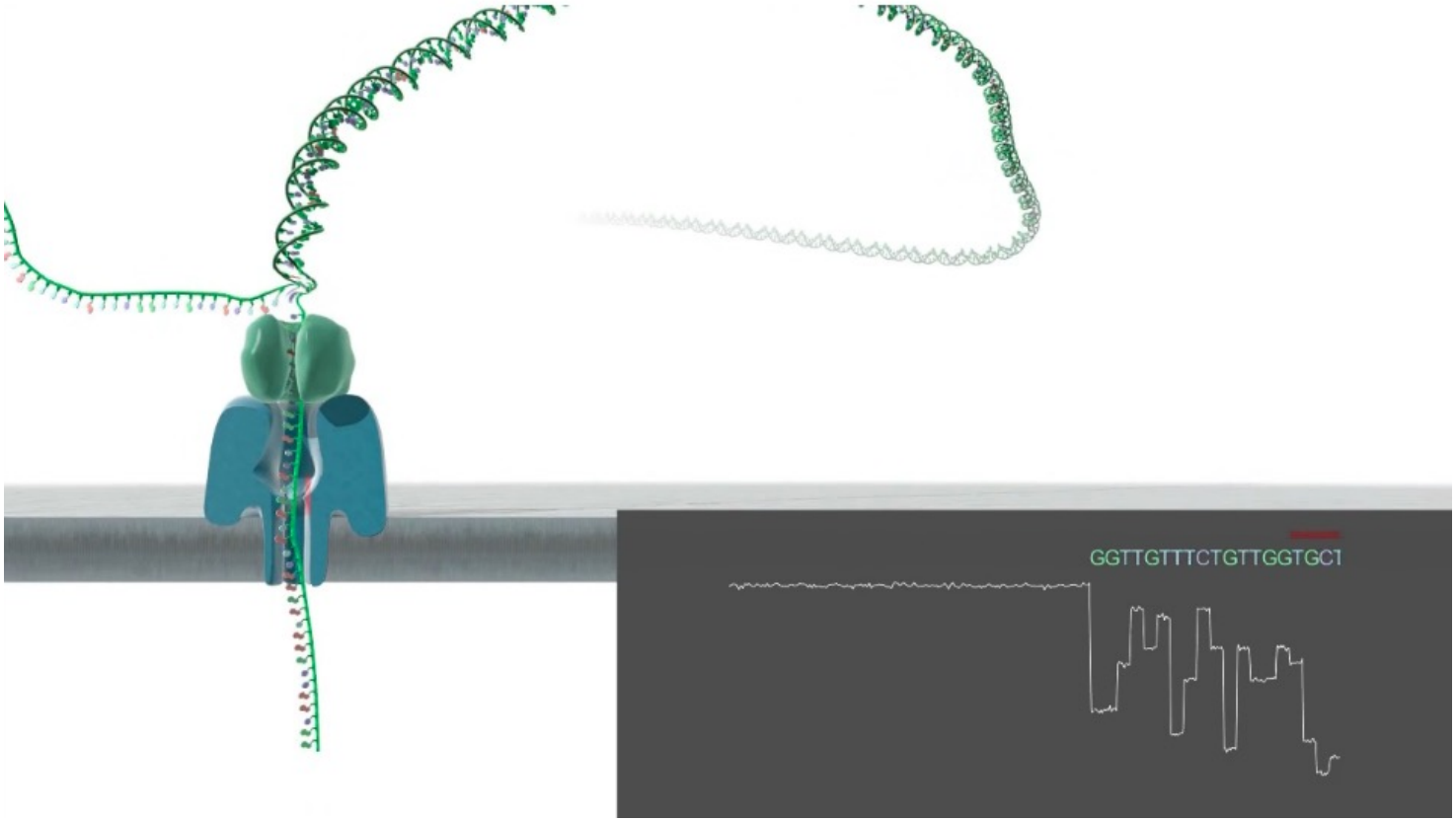


~8 errors per 20 kb read

Oxford Nanopore Technology | ONT



Oxford Nanopore Technology | ONT



ONT Sequencing Principle

https://www.youtube.com/watch?v=E9-Rm5AoZGw

ONT Sequencing Principle

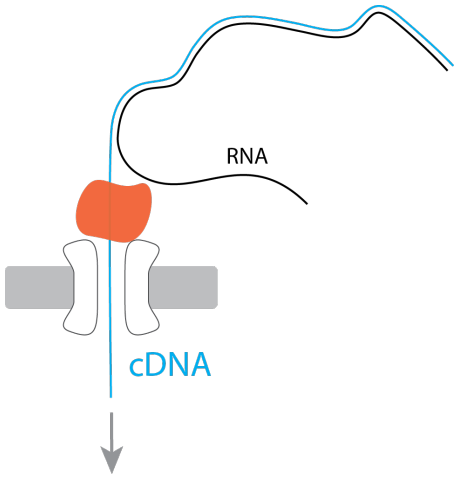
https://vimeo.com/337258910

ONT | RNA Sequencing

UNIVERSITÄT BERN

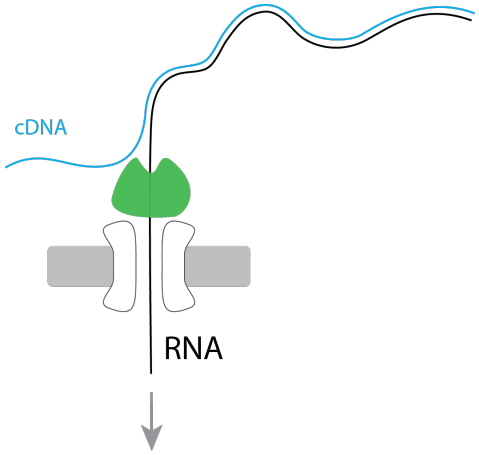
cDNA

- Isoform identification
- Compatible with cDNA workflows
- PCR free
- amplification for low input



direct RNA

- **sequence native RNA**
- no PCR bias
- **detection of modified bases**
- suitable for transcripts that fail RT



ONT | Features

UNIVERSITÄT BERN

- Real time sequencing/data analysis
  - first sequences appear minutes after the run start
  - provides live feedback on how the experiment is running
  - interesting for e.g. diagnostics applications where time matters

- no fixed run time
  - once the result/objective is reached, one can stop and wash/reuse the flowcell
  - The system can run until sufficient data has been collected, analysing the data in real time

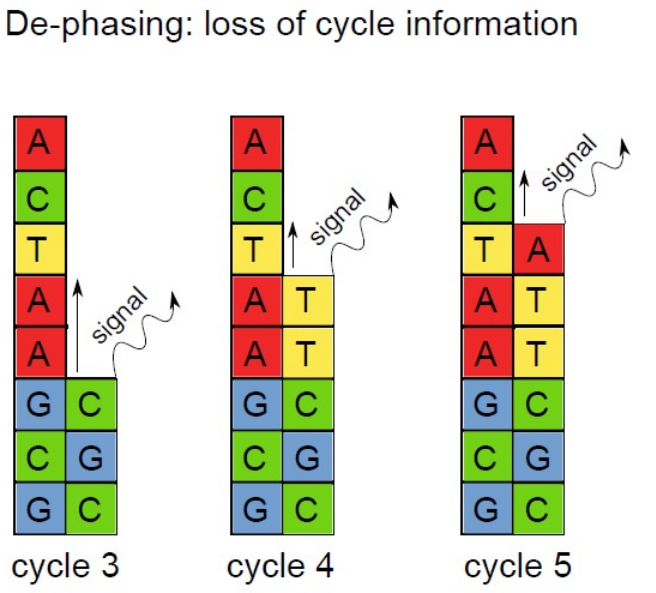
ONT | Features

Read length	Nanopores read the entire length of the fragment of DNA/RNA presented to them. Longest read so far: > 4Mb. Accuracy is maintained throughout the fragment.
Single Molecule Accuracy	R9 modal >98.3%, New chemistry >99% (coming soon)
Consensus Accuracy	R9.4.1: Current best Q45 (>99.99%) R10: Current Best Q50 (99.999%)
Sample Preparation time	Rapid Kit: 10 minutes, Ligation Kit: 60 minutes, other protocols and timings also available
Modified base detection	Yes - Base modification information available in raw signal
Time to 1st usable data	2 minutes

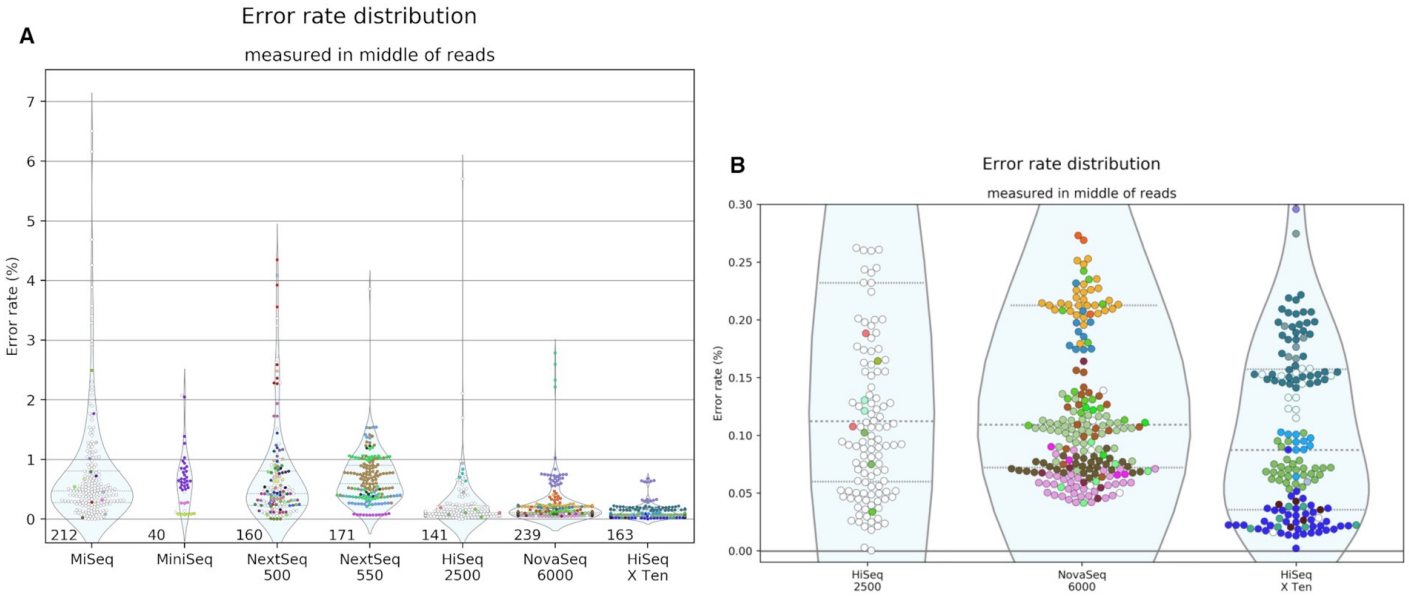
Error Sources

Sequencing Errors

- Cloning artifacts
- Errors in PCR amplification
- GC-Bias, enrichment bias
- De-phasing
- Base misincorporation
- Instrument specific errors
- ... lots we still don't know ...



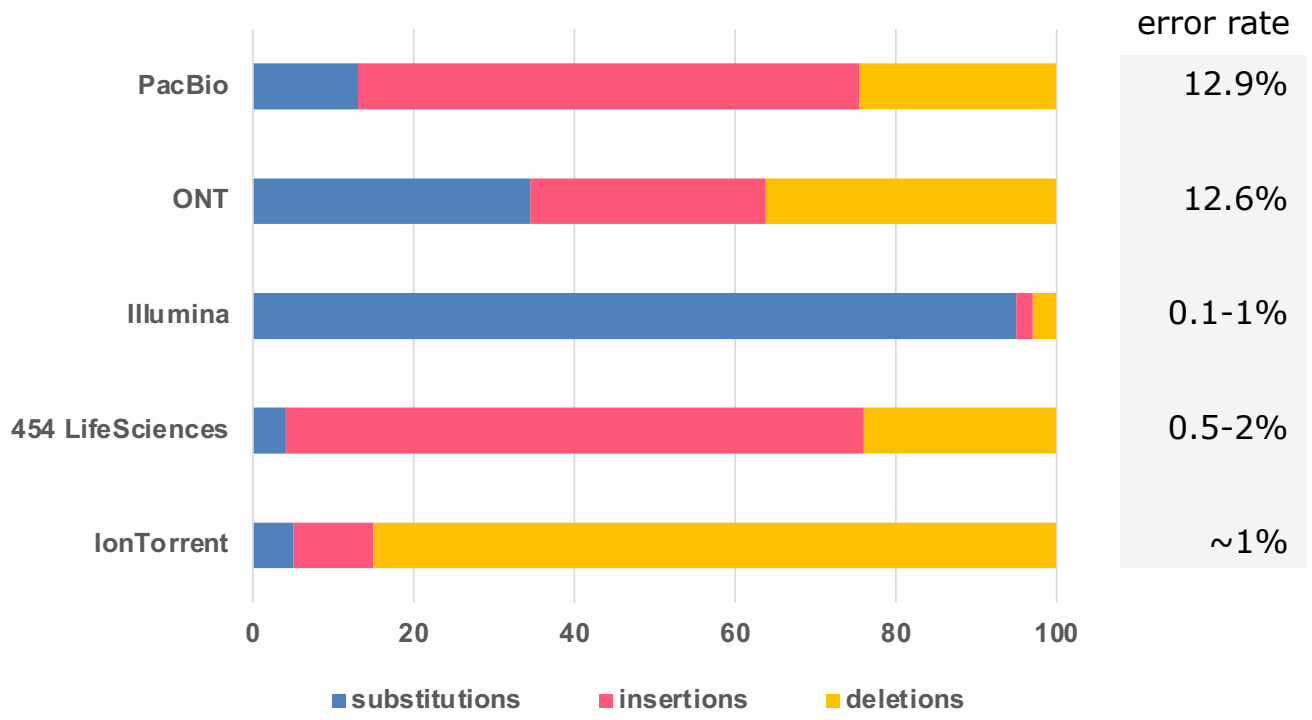
Error Rate of Illumina Instruments



		Error rate (%)	
Platform	**Number of samples**	**Median**	**Standard deviation**
MiSeq	212	0.473	0.938
MiniSeq	40	0.613	0.459
NextSeq 500	160	0.429	0.827
NextSeq 550	171	0.593	0.435
HiSeq 2500	141	0.112	0.544
NovaSeq 6000	239	0.109	0.350
HiSeq X Ten	163	0.087	0.126

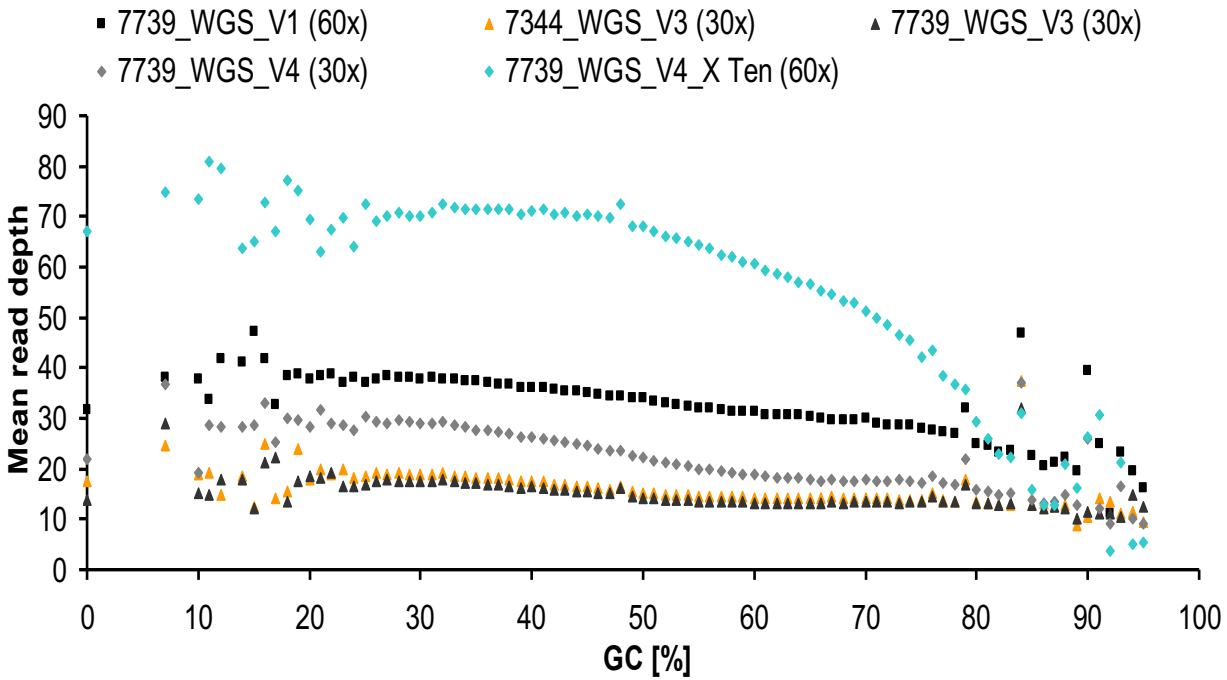
NAR Genom Bioinform, Volume 3, Issue 1, March 2021, lqab019, https://doi.org/10.1093/nargab/lqab019

Platform Specific Error Types

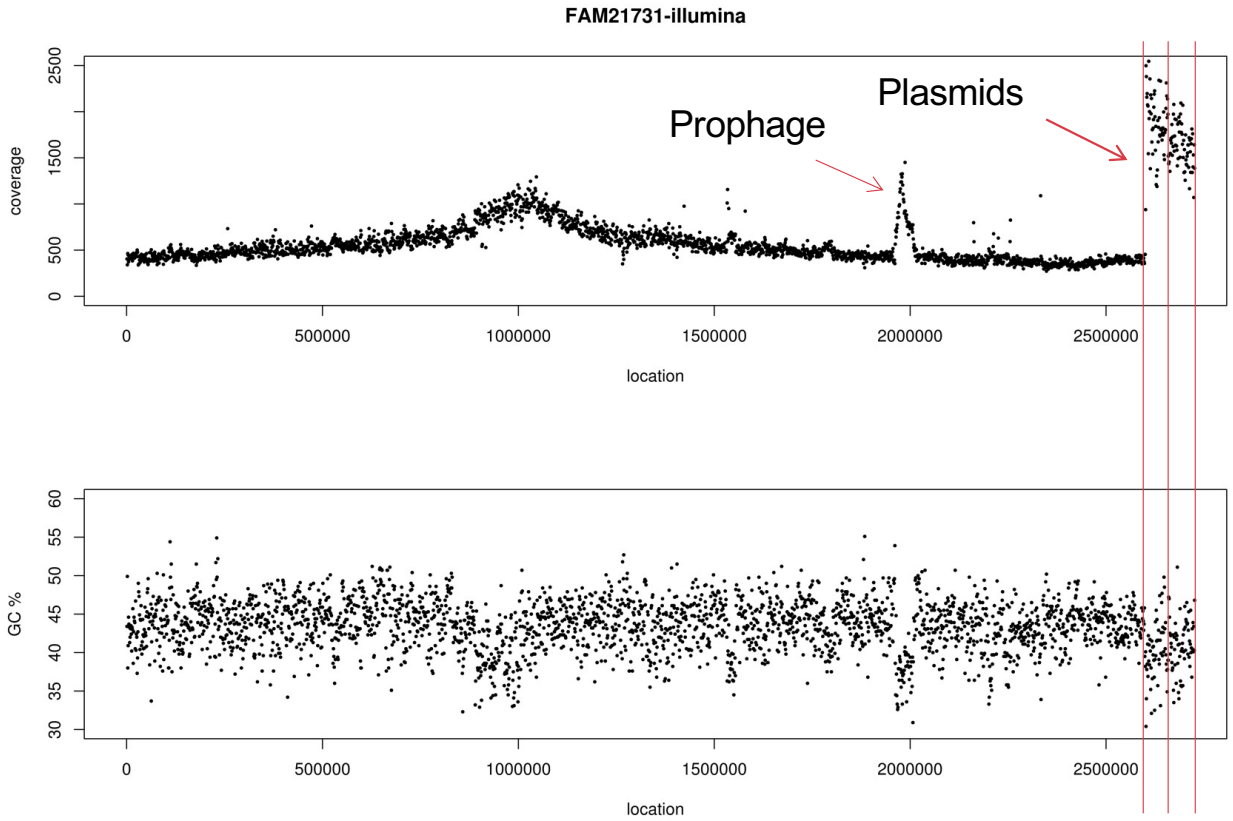


NAR Genom Bioinform, Volume 2, Issue 2, June 2020, lqaa037, https://doi.org/10.1093/nargab/lqaa037

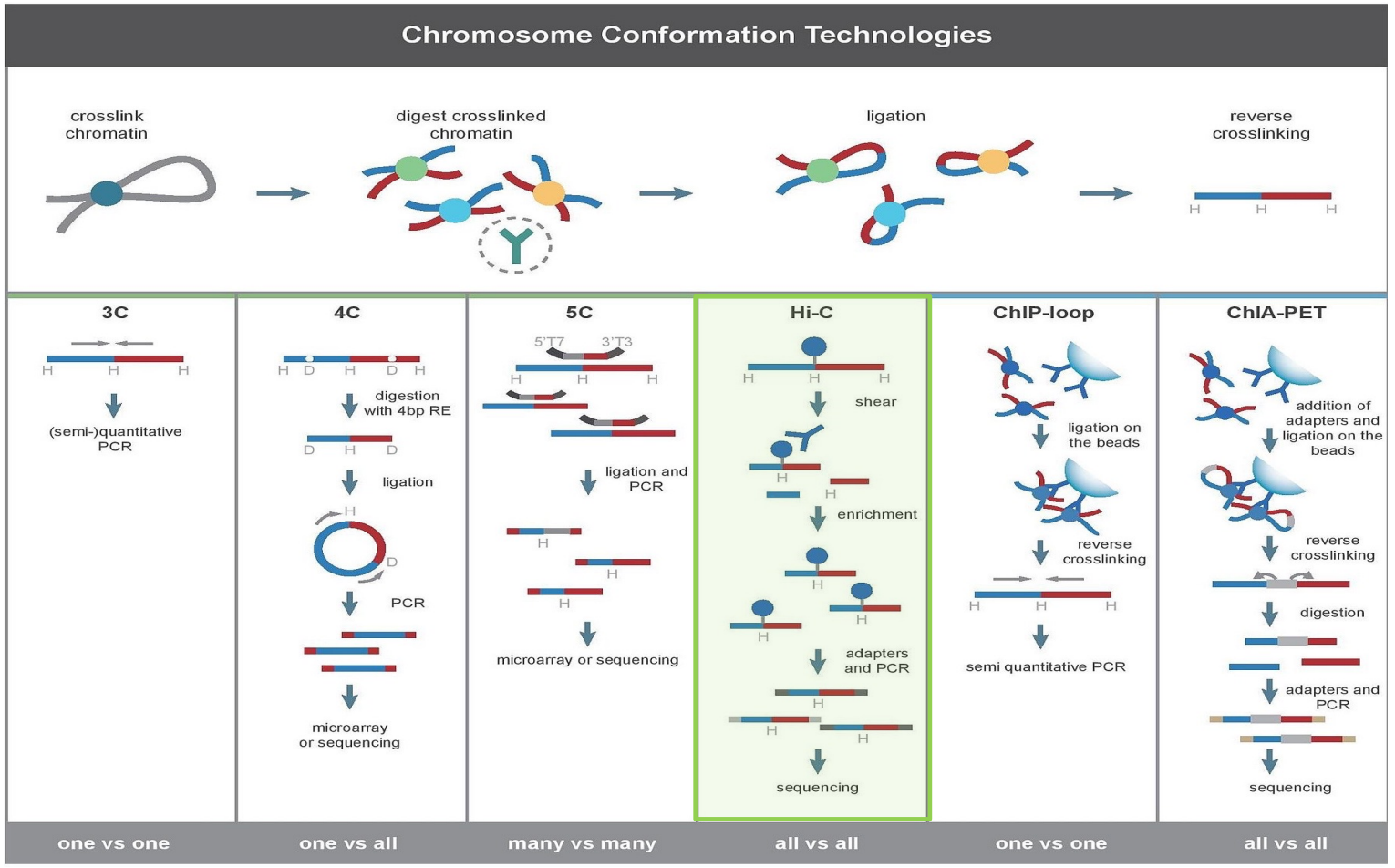
- Bias PCR amplification due to differences in the GC content



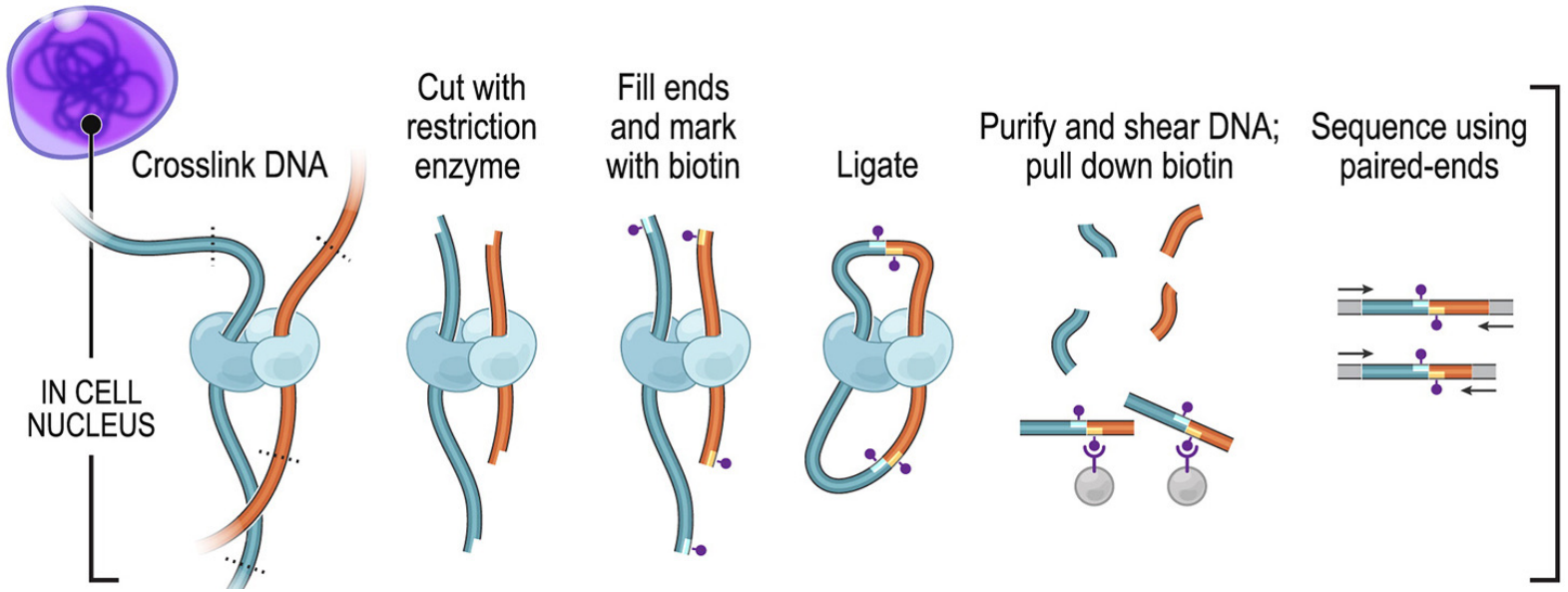
Meienberg J, Zerjavic1 K, Keller I, Okoniewski M, Patrignani A, Ludin K, Xu Z, Steinmann B, Carrel T, Röthlisberger B, Schlapbach R, Bruggmann R and Matyas G; Nucleic Acids Res . 2015 Jun 23;43(11):e76. doi: 10.1093/nar/gkv216. Epub 2015 Mar 27.


FAM21731-illumina
Prophage
Plasmids
coverage
0
500
1500
2500
location
0
500000
1000000
1500000
2000000
2500000
GC %
30
35
40
45
50
55
60

Hi-C | Chromosome Conformation Capturing



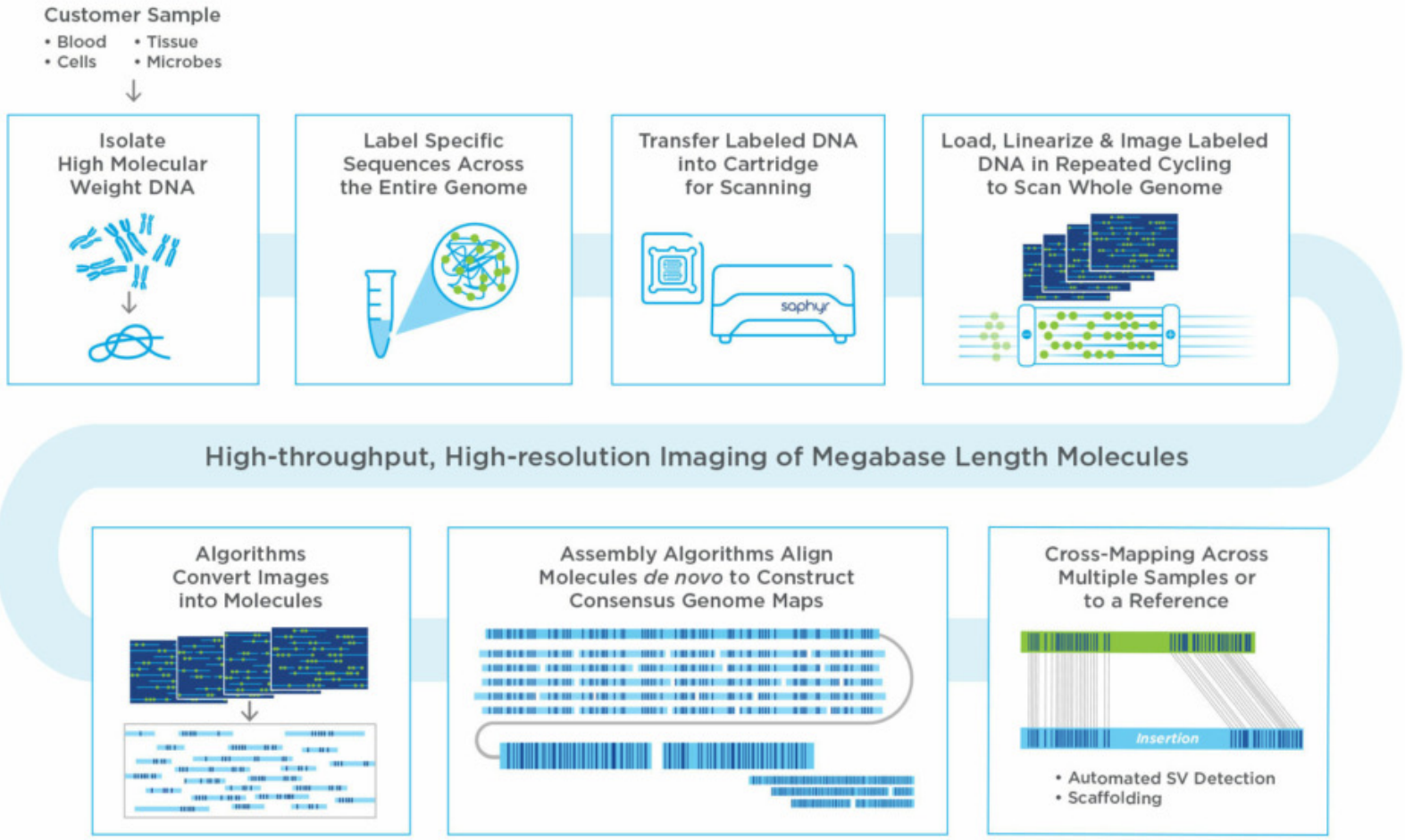
Hi-C | Chromosome Conformation Capturing



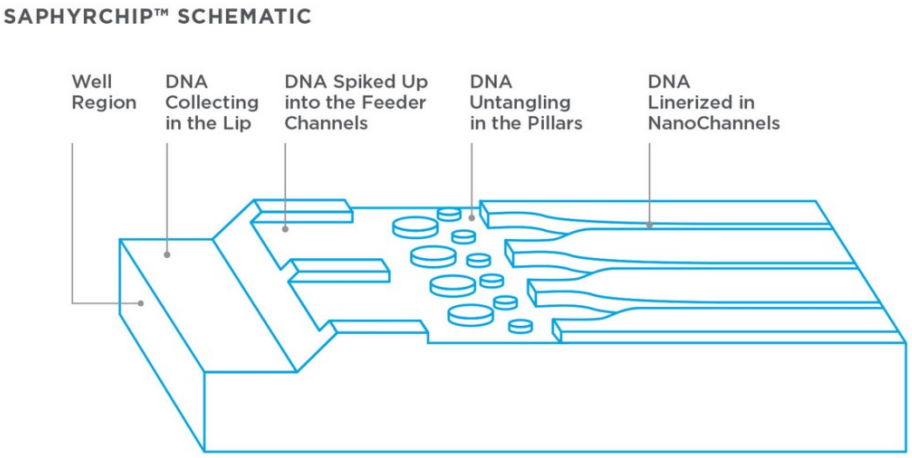
Distance between reads can be up to 500 Mb

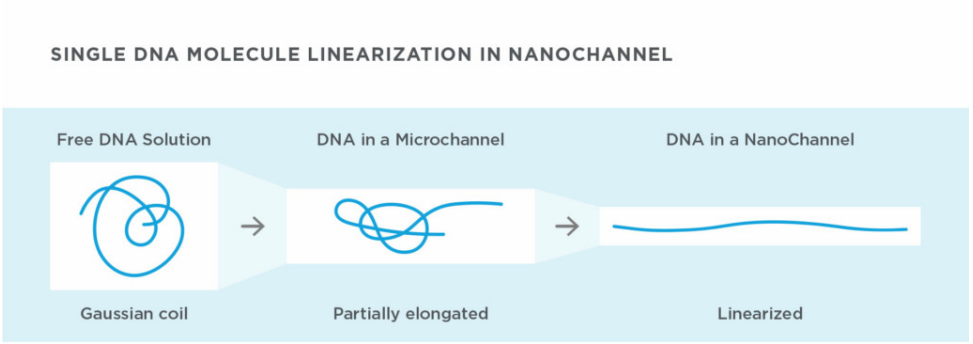
Rao et al., 2014. Cell 159: 1665-1680.

Optical Mapping | Bionano

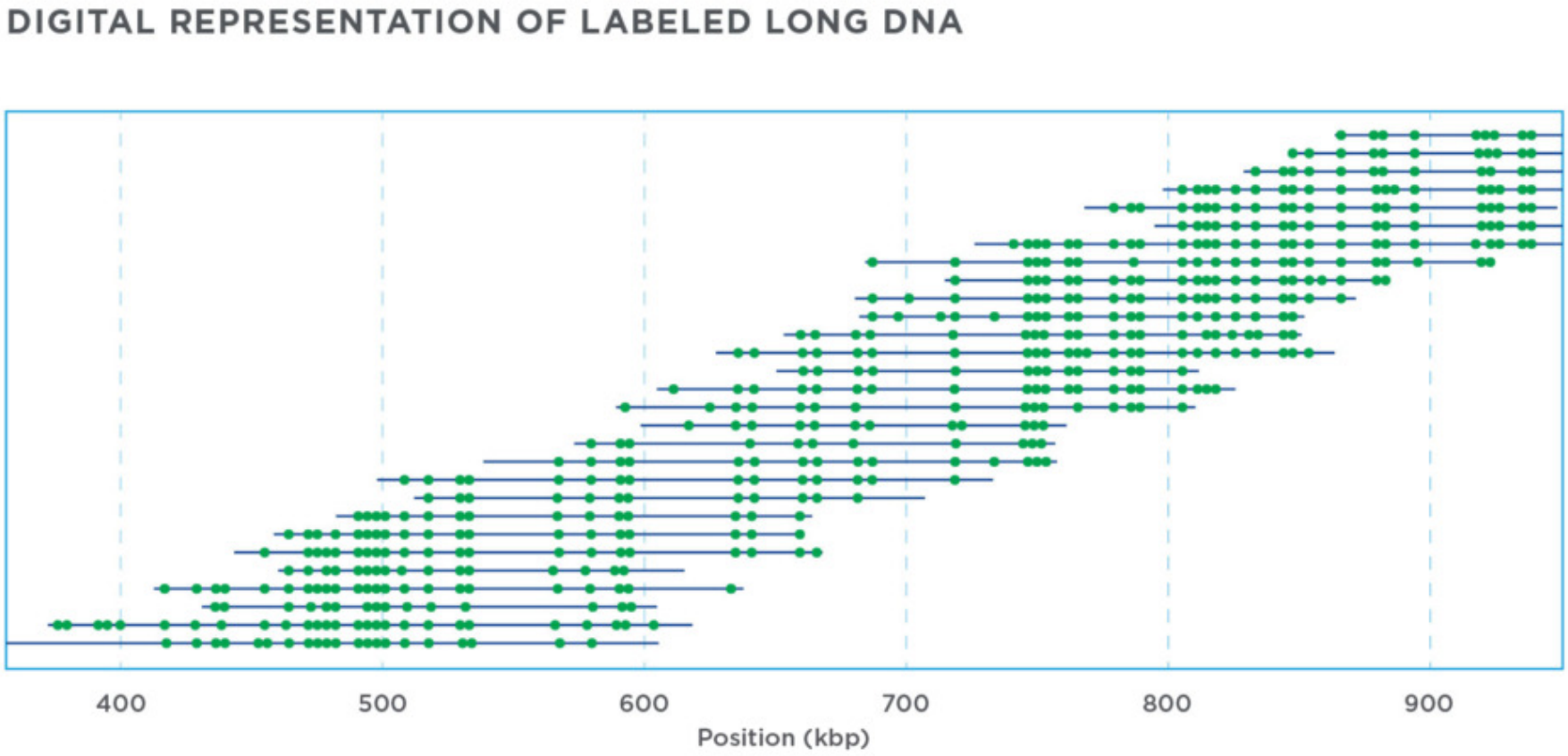


Optical Mapping | Bionano





Optical Mapping | Bionano



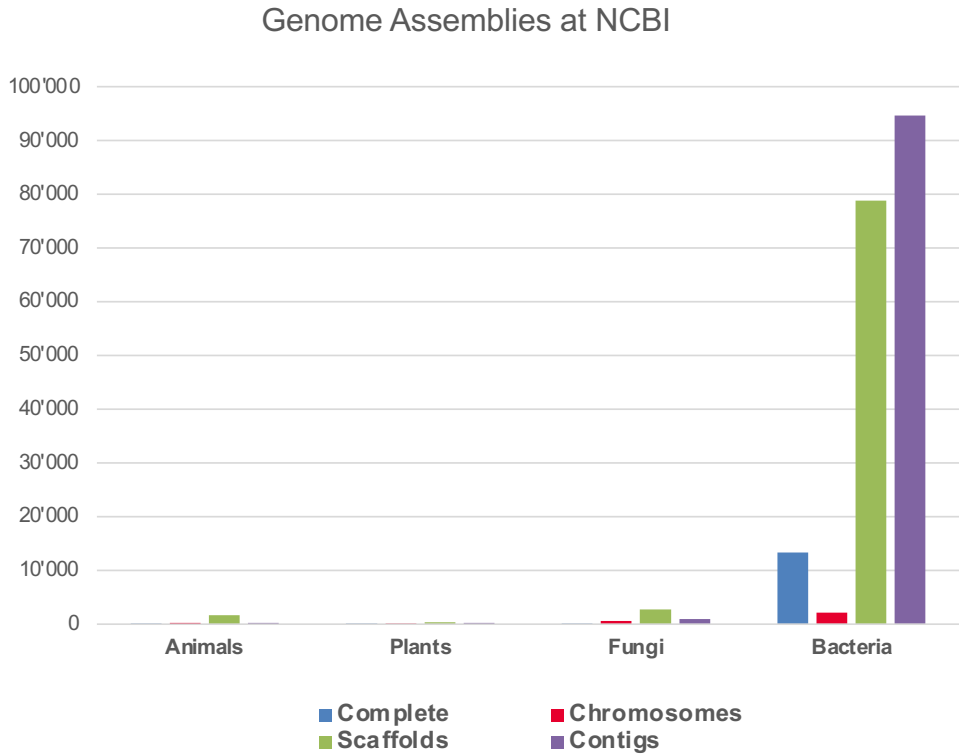
Towards high Quality Assemblies

Genomes in Public Databases (NCBI)

UNIVERSITÄT BERN

Assembly Level	Total	Complete		Chromosomes		Scaffolds		Contigs	
Animals	3'853	3	(0.1%)	695	(18.0%)	2'773	(72.0%)	382	(9.9%)
Plants	1'488	3	(0.2%)	416	(28.0%)	665	(44.7%)	404	(27.2%)
Fungi	6'886	86	(1.2%)	826	(12.0%)	4'390	(63.8%)	1'584	(23.0%)
Protists	943	18	(1.9%)	119	(12.6%)	593	(62.9%)	213	(22.6%)
Eukaryotes	13'170	110	(0.8%)	2'056	(15.6%)	8'421	(63.9%)	2'583	(19.6%)
Bacteria	276'806	19'884	(7.2%)	3'270	(1.2%)	102'252	(36.9%)	151'400	(54.7%)
Archaea	5'615	401	(7.1%)	25	(0.4%)	2'163	(38.5%)	3'026	(54.7%)
∑ Prokaryotes	282'421	20'285	(7.2%)	3'295	(1.2%)	104'415	(37.0%)	154'426	(54.7%)
Overall	295'591	20'395	(6.9%)	5'351	(1.8%)	112'836	(38.2%)	157'009	(53.1%)

*data accessed October 2020 at NCBI genomes
https://www.ncbi.nlm.nih.gov/genome/browse#!/overview/



The Human Genome

UNIVERSITÄT BERN

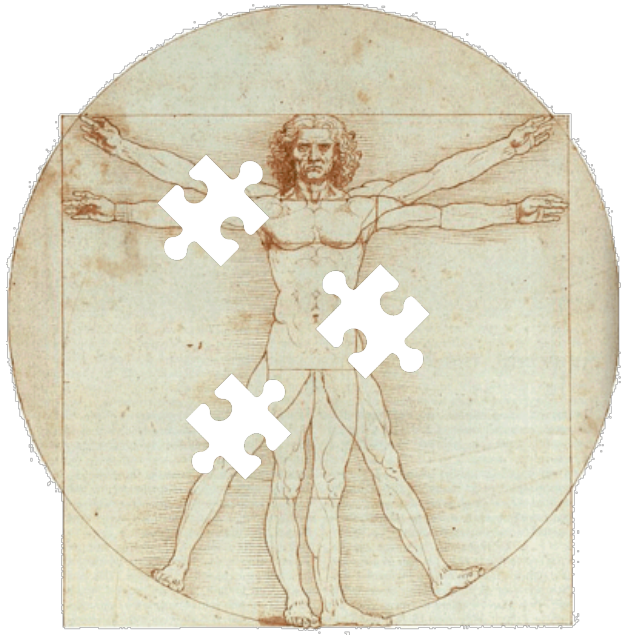
Psst, the human genome was never completely sequenced. Some scientists say it should be

By SHARON BEGLEY @sxbegle / JUNE 20, 2017

"As a matter of truth in advertising, **the 'finished' sequence isn't finished**," said Eric Lander, who led the lab at the Whitehead Institute that deciphered more of the genome for the government-funded Human Genome Project than any other. "I always say 'finished' is a term of art."

"**It's very fair to say the human genome was never fully sequenced**," Craig Venter, another genomics luminary, told STAT.

"**The human genome has not been completely sequenced and neither has any other mammalian genome** as far as I'm aware," said Harvard Medical School bioengineer George Church, who made key early advances in sequencing technology.



still missing sequences of the human genome
4-9%
(Georg Church, Karen Miga)

Telomere-to-telomere Initiative

Article

Telomere-to-telomere assembly of a complete human X chromosome

https://doi.org/10.1038/s41586-020-2547-7

Received: 30 July 2019

Accepted: 29 Ma

Published online:

Open access

Check for upd

Karen H. Miga[1,24], Sergey Koren[2,24], Arang Rhie[2], Mitchell R. Vollger[3], Ariel Gershman[4], Andrey Bzikadze[5], Shelise Brooks[6], Edmund Howe[7], David Porubsky[3], Glennis A. Logsdon[3], Valerie A. Schneider[8], Tamara Potapova[7], Jonathan Wood[9], William Chow[9], Joel Armstrong[1],

Article

The structure, function and evolution of a complete human chromosome 8

https://doi.org/10.1038/s41586-021-03420-7

Received: 4 September 2020

Accepted: 4 March 2021

Published online: 07 April 2021

Open access

Check for updates

Glennis A. Logsdon[1], Mitchell R. Vollger[1], PingHsun Hsieh[1], Yafei Mao[1], Mikhail A. Liskovykh[2], Sergey Koren[3], Sergey Nurk[3], Ludovica Mercuri[4], Philip C. Dishuck[1], Arang Rhie[3], Leonardo G. de Lima[5], Tatiana Dvorkina[6], David Porubsky[1], William T. Harvey[1], Alla Mikheenko[6], Andrey V. Bzikadze[7], Milinn Kremitzki[8], Tina A. Graves-Lindsay[8], Chirag Jain[3], Kendra Hoekzema[1], Shwetha C. Murali[1,9], Katherine M. Munson[1], Carl Baker[1], Melanie Sorensen[1], Alexandra M. Lewis[1], Urvashi Surti[10], Jennifer L. Gerton[5], Vladimir Larionov[2], Mario Ventura[4], Karen H. Miga[11], Adam M. Phillippy[3] & Evan E. Eichler[1,9]

The complete assembly of each human chromosome is essential for understanding human biology and evolution[1,2]. Here we use complementary long-read sequencing technologies to complete the linear assembly of human chromosome 8. Our assembly resolves the sequence of five previously long-standing gaps, including a 2.08-Mb centromeric α-satellite array, a 644-kb copy number polymorphism in the β-defensin gene cluster that is important for disease risk, and an 863-kb variable number tandem

b UNIVERSITÄT

The complete sequence of a human genome

Sergey Nurk[1,*], Sergey Koren[1,*], Arang Rhie[1,*], Mikko Rautiainen[1,*], Andrey V. Bzikadze[2], Alla Mikheenko[3], Mitchell R. Vollger[4], Nicolas Altemose[5], Lev Uralsky[6,7], Ariel Gershman[8], Sergey Aganezov[9], Savannah J. Hoyt[10], Mark Diekhans[11], Glennis A. Logsdon[4], Michael Alonge[9], Stylianos E. Antonarakis[12], Matthew Borchers[13], Gerard G. Bouffard[14], Shelise Y. Brooks[14], Gina V. Caldas[15], Haoyu Cheng[16,17], Chen-Shan Chin[18], William Chow[19], Leonardo G. de Lima[13], Philip C. Dishuck[4], Richard Durbin[21], Tatiana Dvorkina[3], Ian T. Fiddes[22], Giulio Formenti[23,24], Robert S. Fulton[25], Arkarachai Fungtammasan[18], Erik Garrison[11,26], Patrick G.S. Grady[10], Tina A. Graves-Lindsay[27], Ira M. Hall[28], Nancy F. Hansen[29], Gabrielle A. Hartley[10], Marina Haukness[11], Kerstin Howe[19], Michael W. Hunkapiller[30], Chirag Jain[1,31], Miten Jain[11], Erich D. Jarvis[23,24], Peter Kerpedjiev[32], Melanie Kirsche[9], Mikhail Kolmogorov[33], Jonas Korlach[30], Milinn Kremitzki[27], Heng Li[16,17], Valerie V. Maduro[34], Tobias Marschall[35], Ann M. McCartney[1], Jennifer McDaniel[36], Danny E. Miller[4,37], James C. Mullikin[14,29], Eugene W. Myers[38], Nathan D. Olson[36], Benedict Paten[11], Paul Peluso[30], Pavel A. Pevzner[33], David Porubsky[4], Tamara Potapova[13], Evgeny I. Rogaev[6,7,39,40], Jeffrey A. Rosenfeld[41], Steven L. Salzberg[9,42], Valerie A. Schneider[43], Fritz J. Sedlazeck[44], Kishwar Shafin[11], Colin J. Shew[20], Alaina Shumate[42], Yumi Sims[19], Arian F. A. Smit[45], Daniela C. Soto[20], Ivan Sović[30,46], Jessica M. Storer[45], Aaron Streets[5,47], Beth A. Sullivan[48], Françoise Thibaud-Nissen[43], James Torrance[19], Justin Wagner[36], Brian P. Walenz[1], Aaron Wenger[30], Jonathan M. D. Wood[19], Chunlin Xiao[43], Stephanie M. Yan[49], Alice C. Young[14], Samantha Zarate[9], Urvashi Surti[50], Rajiv C. McCoy[49], Megan Y. Dennis[20], Ivan A. Alexandrov[3,7,51], Jennifer L. Gerton[13], Rachel J. O'Neill[10], Winston Timp[8,42], Justin M. Zook[36], Michael C. Schatz[9,49], Evan E. Eichler[4,24,†], Karen H. Miga[11,†], Adam M. Phillippy[1,†]

[1-51] Affiliations are listed at the end

[*] Equal contribution

[†] Corresponding authors: Evan E. Eichler (eee@gs.washington.edu); Karen H. Miga (khmiga@ucsc.edu); Adam M. Phillippy (adam.phillippy@nih.gov)

Abstract

In 2001, Celera Genomics and the International Human Genome Sequencing Consortium published their initial drafts of the human genome, which revolutionized the field of genomics. While these drafts and the updates that followed effectively covered the euchromatic fraction of the genome, the heterochromatin and many other complex regions were left unfinished or erroneous. Addressing this remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium has finished the first truly complete 3.055 billion base pair (bp) sequence of a human genome, representing the largest improvement to the human reference genome since its initial release. The new T2T-CHM13 reference includes gapless assemblies for all 22 autosomes plus Chromosome X, corrects numerous errors, and introduces nearly 200 million bp of novel sequence containing 2,226 paralogous gene copies, 115 of which are predicted to be protein coding. The newly completed regions include all centromeric satellite arrays and the short arms of all five acrocentric chromosomes, unlocking these complex regions of the genome to variational and functional studies for the first time.

Nurk S, Koren S, Rhie A, Rautiainen M, et al. The complete sequence of a human genome. bioRxiv, 2021.

Is Perfect Assembly Possible?

Theorem: Perfect assembly possible if *(Gene Myers)*

i. errors random
ii. sampling is Poisson
iii. reads long enough to solve repeats

(Note: e-rate not needed)

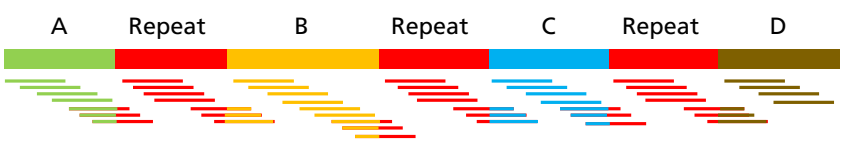
→ computationally expensive

→ depth of coverage ~100x required

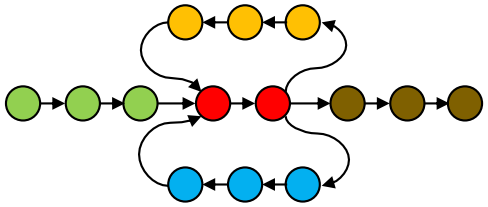
- error correction
  - self-alignment of long reads to correct errors
  - error correction using high accurate illumina reads (method based on variable length k-mers)

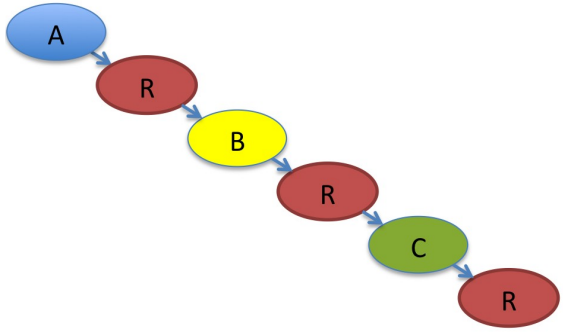
Repeats in Genome Assembly

a)



b)





adapted from Michael Schatz

(a) Genomic region with a triple repeat and 4 unique regions A, B, C, D. Mapped short reads are shown below this region.

(b) Corresponding de Bruijn graph representation of the genomic region if the triple repeat is longer than the k-mer size k. Multiple paths are possible through the graph.

Complete Genomes

- for universal applicability technologies should deliver
- extremely long reads → ideally chromosome size
- extremely low-cost data
- extremely fast and real-time
- very flexible instruments instead of monolithic ones
- extremely sensitive / low input requirements
- error free

European Reference Genome Atlas (ERGA)

UNIVERSITÄT BERN

- Aim of ERGA is to generate *de novo* genome assemblies that are
  - reference-quality
  - complete
  - error-free
- for species across the whole of European biodiversity
  - includes at least 200,000 species of higher organisms

Richard Durbin

*"We are **entering a new era of genomics**, in which reference sequences for all species will change the way we interact with the whole living world. The ERGA will coordinate European activities in this area and ensure that Europe plays a leading role in this fundamental initiative for 21st century science."*

Gene W. Myers

"An European-wide initiative like the ERGA is absolutely necessary given the scale of the proposed atlas. I belief this atlas **will revolutionize of conservation science, agricultural and animal science, and even contribute significantly to medicine**."

European Reference Genome Atlas (ERGA)

UNIVERSITÄT BERN

- Methods used to achieve the goal:
  - PacBio HiFi Reads
  - HiC reads

- Challenges to sequence > 200k genomes to completion
  - sampling and sampling permissions
  - DNA extraction (HMW DNA)
  - vast sequencing capacity
  - vast compute and storage resources required
  - standards to ensure quality and comparability
  - choice of methods
  - gene annotation not even discussed yet
  - how to assess assembly quality

Now it is your turn – Your Project in this Course

UNIVERSITÄT BERN

- Perform all steps of a de novo genome assembly by yourself
- Start are subsampled read data which are from a Nature Communications publication
  - Nat Commun. 2020 Feb 20;11(1):989. doi: 10.1038/s41467-020-14779-y.
- Each student gets their own data set (PacBio reads)
- Each student will write an own report.
  - report has two parts: (i) assembly & (ii) annotation
    (annotation will be performed in the next course by Christian Parisod (6 weeks)
- At the end of this course (last day), each student presents the own work (15 minutes + discussion)