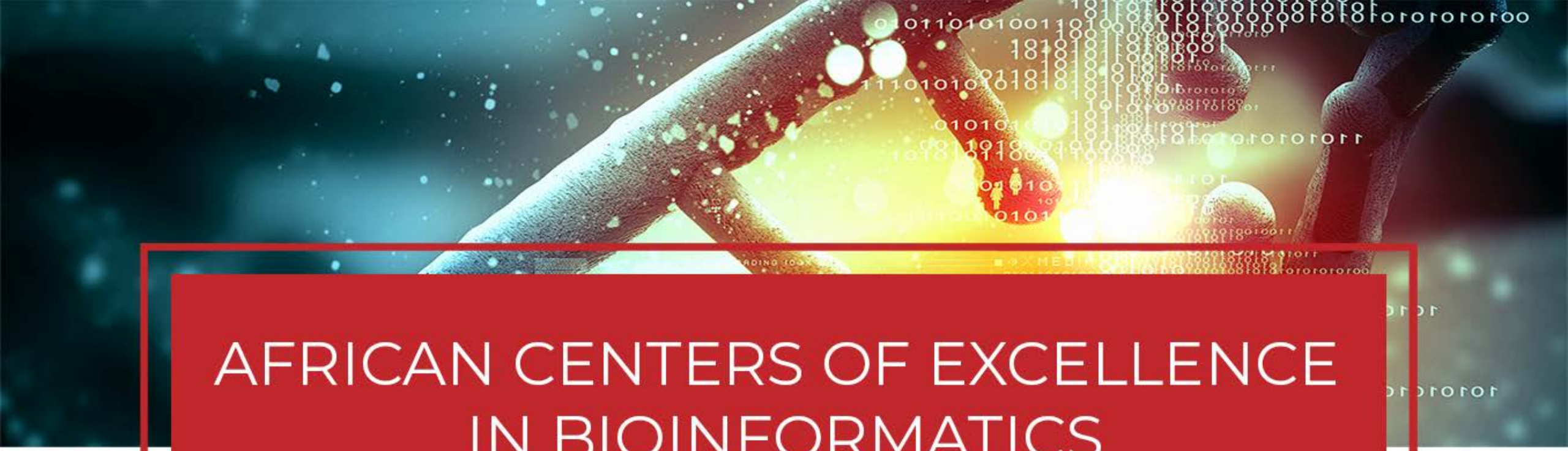


AFRICAN CENTERS OF EXCELLENCE IN BIOINFORMATICS

KAMPALA, UGANDA

Molecular Biology for Bioinformatics: An introduction to Sanger and Next Generation Sequencing

Today's Instructor

Dr. Hernán Lorenzi,
Ph.D. in Molecular Biology

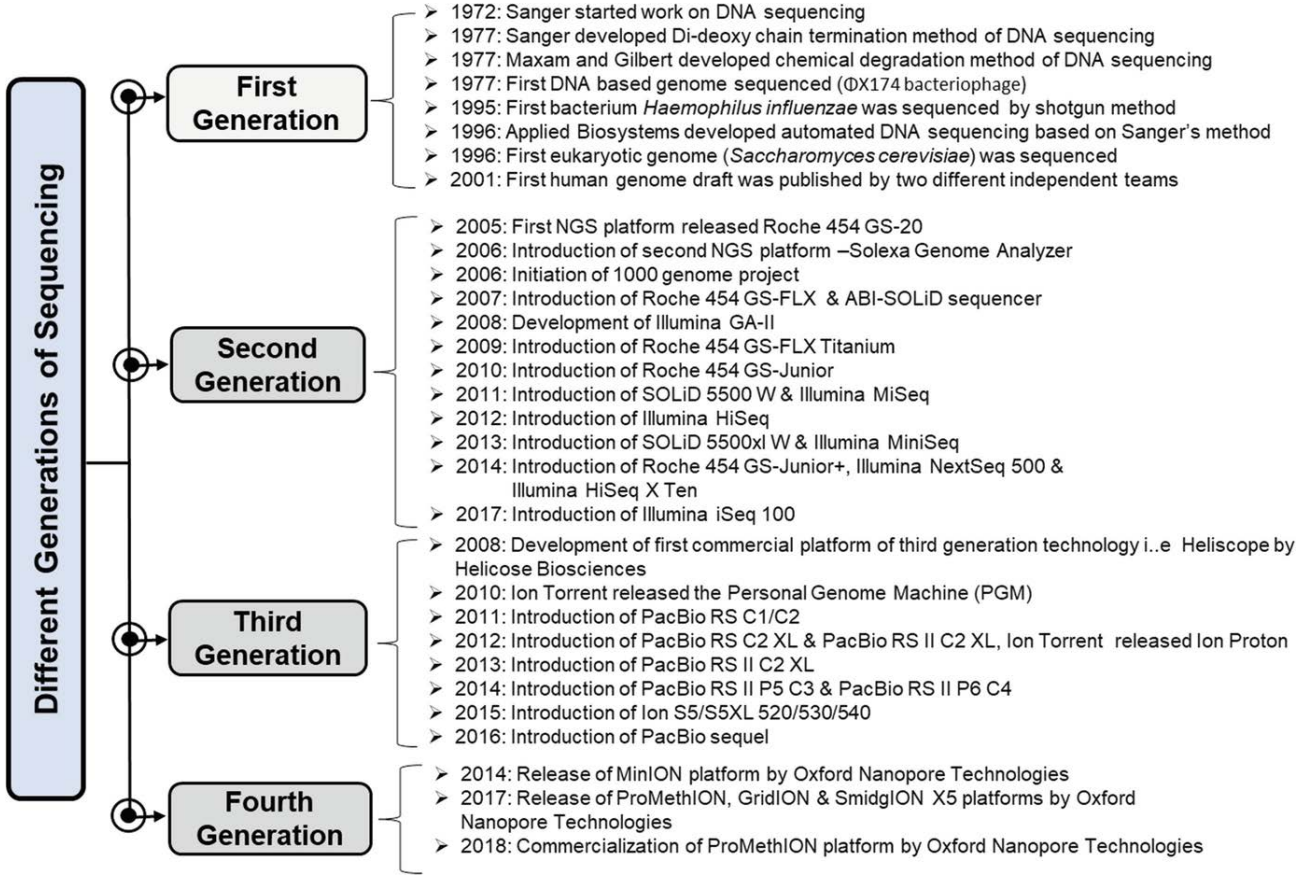
Ongoing Computational Biology projects:

- 16S microbiome/WGS
- Parasite genomics

- Bioinformatics and Computational Biosciences Branch (BCBB), NIAID
- National Institutes of Health, Bethesda, MD USA.
- Contact our team via email:
  - Email: bioinformatics@niaid.nih.gov
  - Instructor: Hernan.lorenzi@nih.gov

Agenda

- Sequencing technologies
  - Sanger Sequencing
  - Next Generation Sequencing (NGS):
    1. Roche 454/Ion Torrent
    2. Illumina
    3. PacBio
    4. MinION / Oxford Nanopore
- NGS biases and errors
- Sequencing data formats
  1. FASTA
  2. FASTQ
  3. SAM/BAM
- Preprocessing of sequencing reads
- NGS Applications



Gupta N., Verma V.K. (2019) Next-Generation Sequencing and Its Application: Empowering in Public Health Beyond Reality. In: Arora P. (eds) Microbial Technology for the Welfare of Society. Microorganisms for Sustainability, vol 17. Springer,

Sequencing vocabulary

Read: piece of sequenced DNA output by the sequencing machine.
DNA template: Fragment of DNA to be sequenced.
Read depth: number of times a given nucleotide in the genome (reference) has been **read** in a sequencing experiment. It can be calculated as (N x L / G) where N = number of reads; L = mean read length; G = genome size
Read coverage: percentage of the genome covered by the sequencing reads.
Contig: a contiguous piece of sequencing data generated from overlapping sequencing reads
Scaffold: set of 2 or more contigs linked together in the same order and orientation as they are in the chromosome.
Reference: anything containing DNA information (genome, chromosome, contigs, etc.)
Mapping: to place reads to one or more specific locations on the reference sequence based on sequence identity.
Base quality: number associated with the probability of error of a nucleotide call within a sequencing read.
Mapping quality: number reflecting how accurate is the mapping of a sequencing read to the reference.
Sequencing bias: preference shown by NGS technologies of sequencing DNA/RNA regions with specific nucleotide composition.
Homopolymer region: DNA sequence composed by a run of a single nucleotide.
Insertion: DNA sequence present in the sequencing data but not in the reference.
Deletion: DNA sequence present in the reference but missing in the sequencing data.
InDel: term referring to a DNA polymorphisms involving an insertion or deletion.
Single Nucleotide Variant (SNV): DNA polymorphism involving a single nucleotide.
Single Nucleotide Polymorphism (SNP): Special case of SNV involving a replacement of a nucleotide for another.
Copy Number Variation (CNV): polymorphism involving changes in the number of copies of a genomic feature with respect to the reference.
Structural Variation: DNA polymorphisms involving chromosomal/sequence rearrangements (inversions, translocations, etc.)
K-mer: nucleotide sequence of a certain length.

Sequencing technologies: Sanger

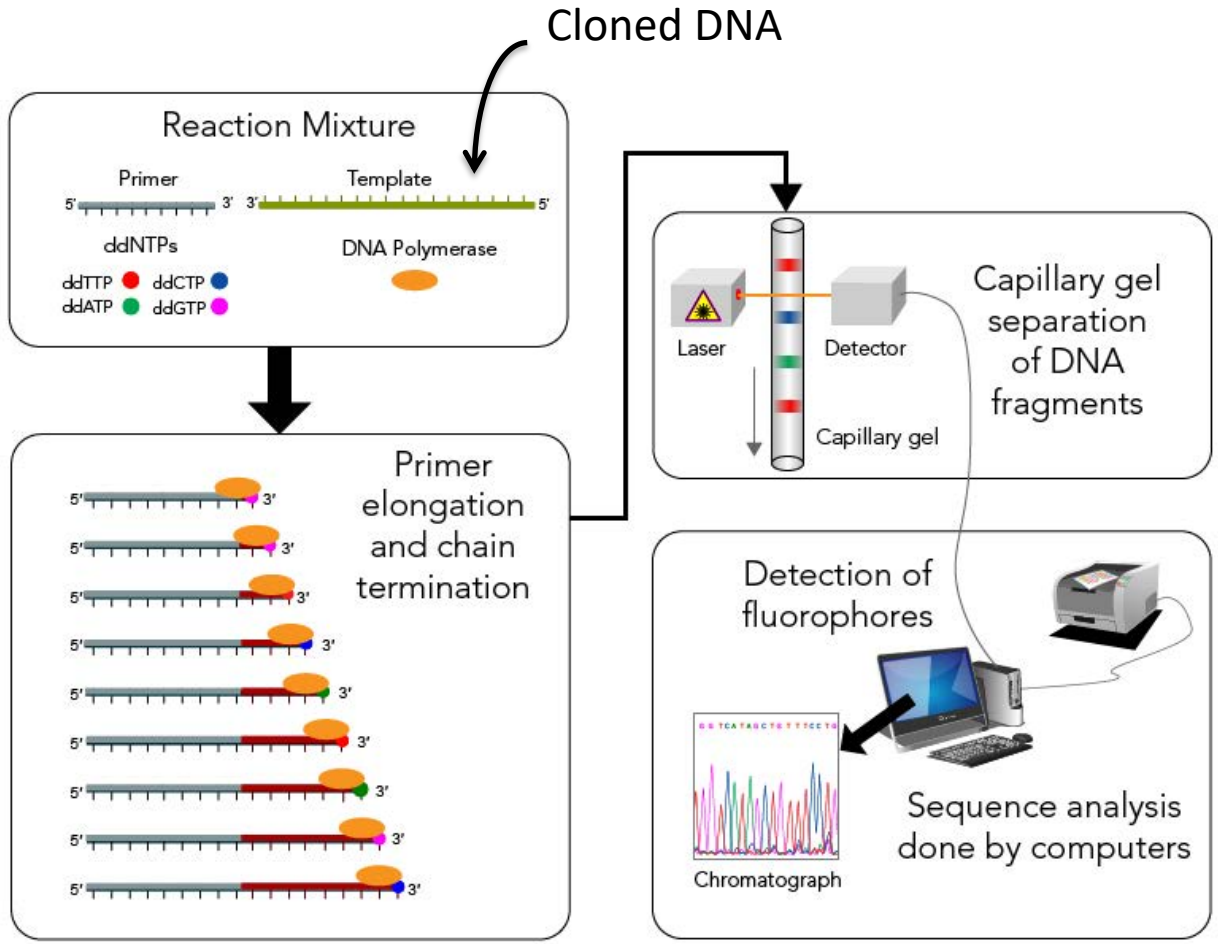
Sanger sequencing:
It was developed by Frederick Sanger and colleagues in 1977.

Most widely used sequencing method for ~40 years until the appearance of NGS.

Still in use for smaller-scale sequencing projects and small budgets.

It was first commercialized by Applied Biosystems in 1986

It uses selective incorporation of chain-terminating fluorescent dideoxynucleotides by DNA polymerase during in vitro DNA replication



Figures taken from Let's talk Science website < https://letstalkscience.ca >

Sequencing technologies: Sanger

For Sanger sequencing it is necessary to amplify or clone the DNA fragments you want to sequence

<u>Sequencing libraries:</u>

- Small fragment libraries (<10kb):
  - Single clone: PCR fragment, DNA fragment cloned in a plasmid.

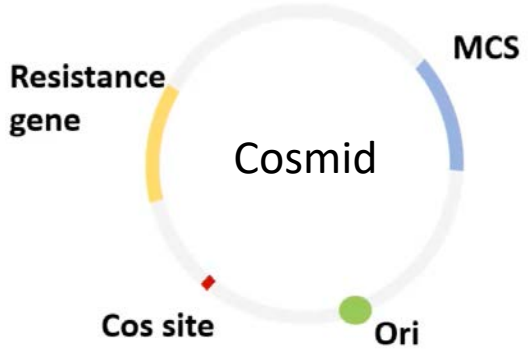
- Mid-size fragment libraries (10 kb – 50 kb):
  - Fosmids. Few copies / cell
  - Cosmids. Many copies / cell

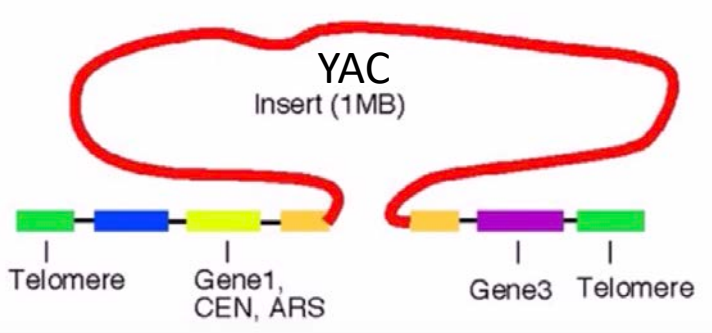
  Up to 50 kb

  - Bacterial Artificial Chromosomes (BACs). One copy / cell

  Up to 300 kb

- Large insert libraries (100 kb - 1000 kb):
  Yeast Artificial Chromosomes (YACs)





Approach selection will depend on the project: e.g. construct verification, whole genome sequencing, cDNA sequencing, characterization of transfection library, etc.

Sequencing technologies: Sanger

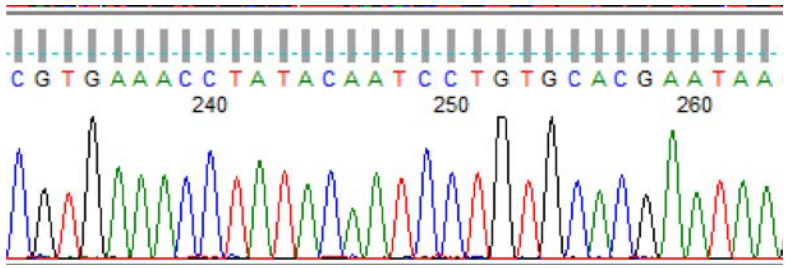
Sanger sequencing Pros:

- High quality
  (mean Phred Q-value = 20 => 99% accuracy)
- Reads longer than some NGS reads (500 – 1,000 bp)
- Low cost per sample
- Lower start-up cost compared to NGS
- Cost effective for low number of targets

Cons:

- Low throughput (1 to ~100 reads per run)
- Low sensitivity (e.g. detection of gene expression)
- Low speed (per Mb)
- Short reads (500 – 1,000 bp)
- Expensive (cost per base)
- Less automated

ABI 3730xl DNA Sequencer



Trace file

From: https://cgenetool.com/product/abi-3730xl-dna-sequencer/

Sequencing technologies: Sanger

When to use Sanger sequencing?

- Sequencing single genes/clones
- Sequencing 1-100 PCR amplicon targets at low cost
- Sequencing up to 96 samples at a time without barcoding
- Microbial Identification
- Fragment analysis, high throughput genotyping using, for example, SNaPshot
- Microsatellite or Short Tandem Repeat analysis
- Confirmation of Next Generation Sequencing results.

Sequencing technologies: Next Generation Sequencing Technologies (NGS)

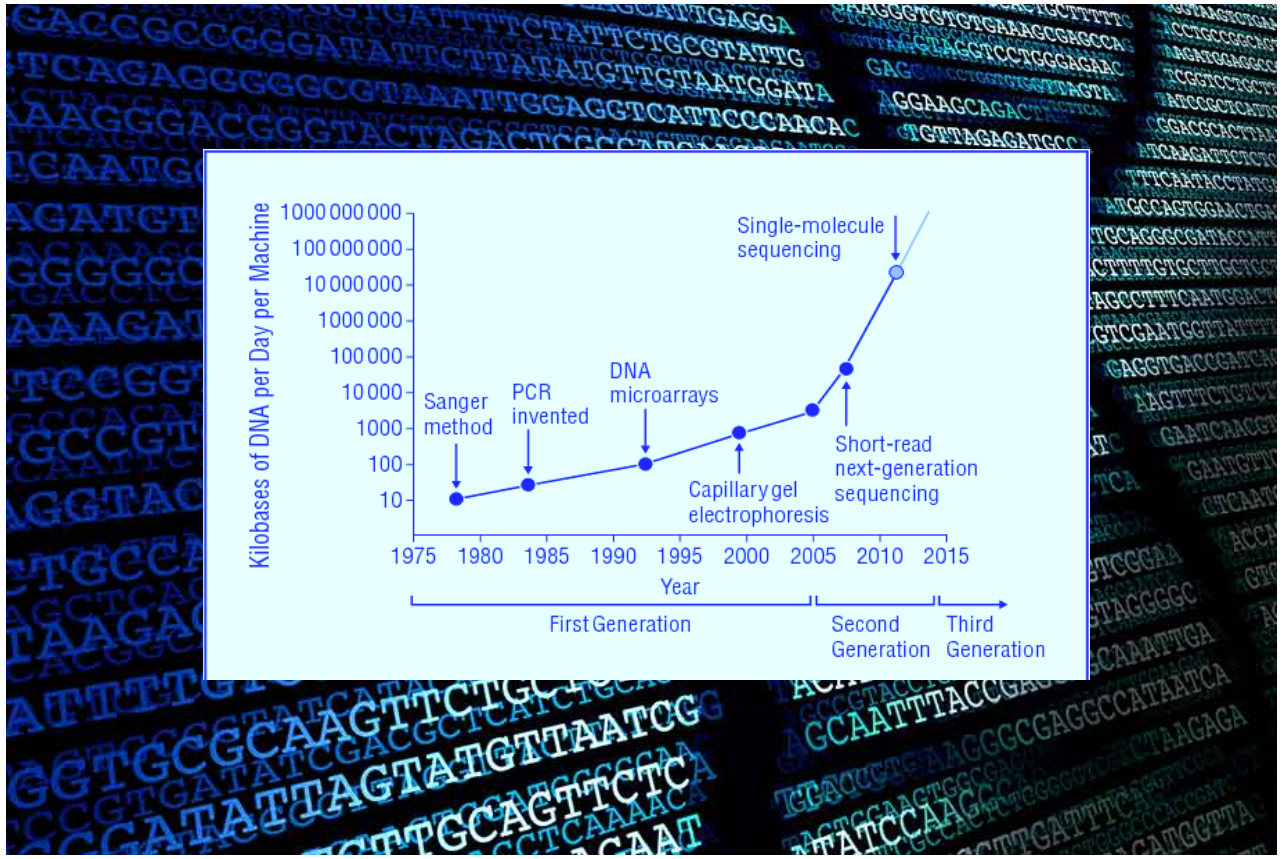
What is NGS?

Sequencing data generated with 2nd and 3rd generation sequencing technologies.

Also known as “High-Throughput NGS” or “Deep Sequencing”.

They use massively parallel sequencing by synthesis of millions of templates producing millions to billions of sequencing reads. Game changer!

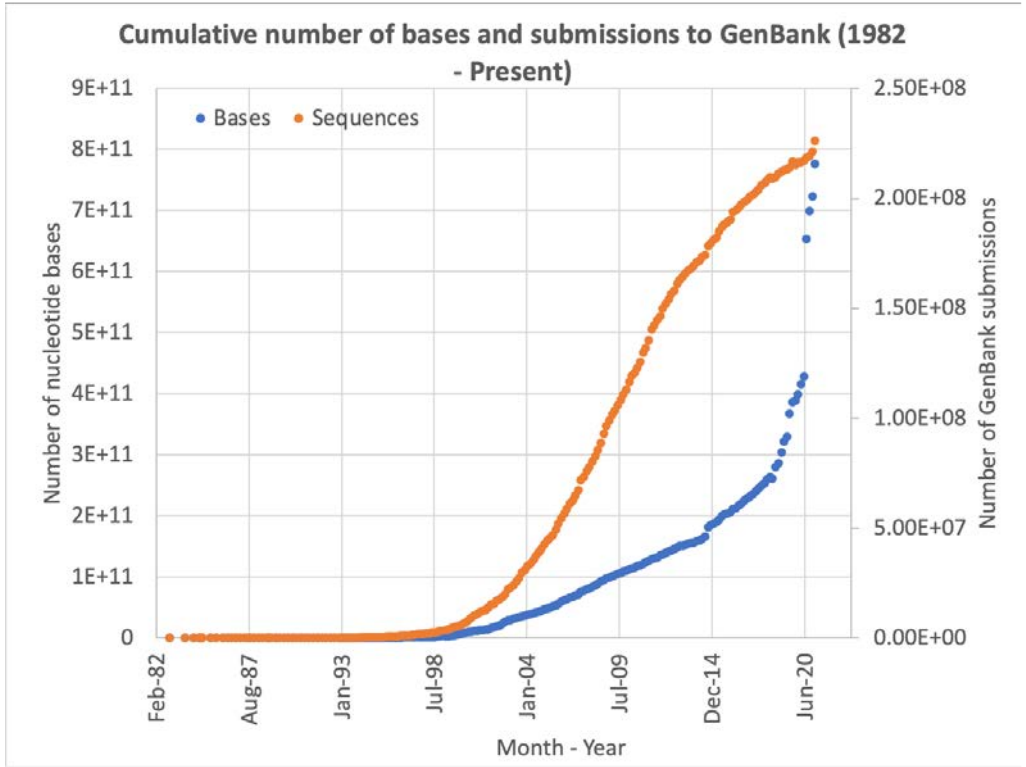
No DNA/cDNA cloning required



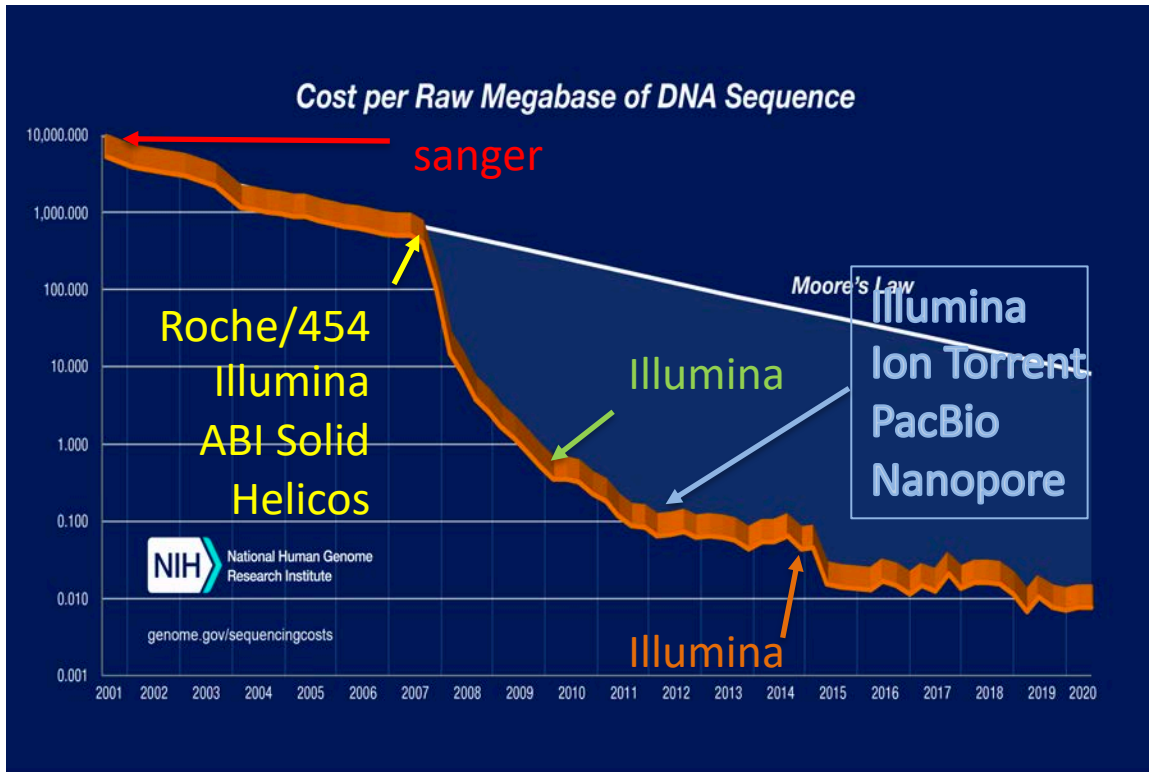
Adapted from Pittman et al, JAMA Neurol. 2013

Sequencing technologies: Next Generation Sequencing Technologies (NGS)

Some technologies may have higher sequencing error rate compared with Sanger.



Source: NCBI



Adapted from www.nih.gov

Sequencing technologies: Next Generation Sequencing Technologies (NGS)

Current NGS Technologies:

- Roche 454 / Ion Torrent
- Illumina

2nd generation

Short reads (75bp – 600bp)
Very Large number of reads (millions)
Sequencing of amplified single molecules

- PacBio
- Oxford Nanopore

3rd generation

Very long reads ($1x10^3$bp – $1x10^6$bp)
Large number of reads
Single molecule sequencing

illumina® ion torrent PACBIO® Oxford NANOPORE Technologies

NGS technologies: Ion Torrent and Roche 454

454 Roche GS FLX+ (out of production)
Throughput 400 Mb
Read length (Single end) ~400 bp – 600 bp
Read length (mate pair) ~150 bp

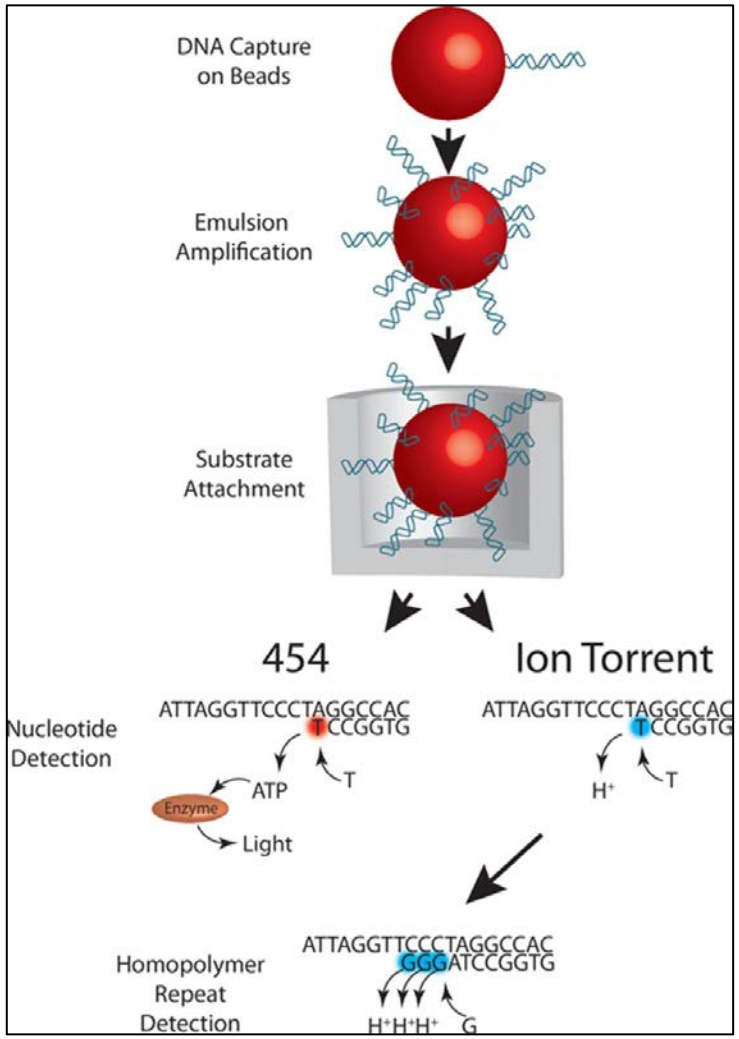
Sequencing error ~1%
Problems with long homopolymers (~>4)

Read output format:
FASTQ (Ion Torrent)
SFF (Standard Flowgram Format - 454)

Ion PGM Dx
Throughput up to 600 Mb – 1 Gb
Read length ~200 – 400 bp
Number of reads 4 – 5.5 million

Ion GeneStudio S5
Throughput 15 - 50 Gb
Read length ~200 – 400 bp
Number of reads 2 – 130 million

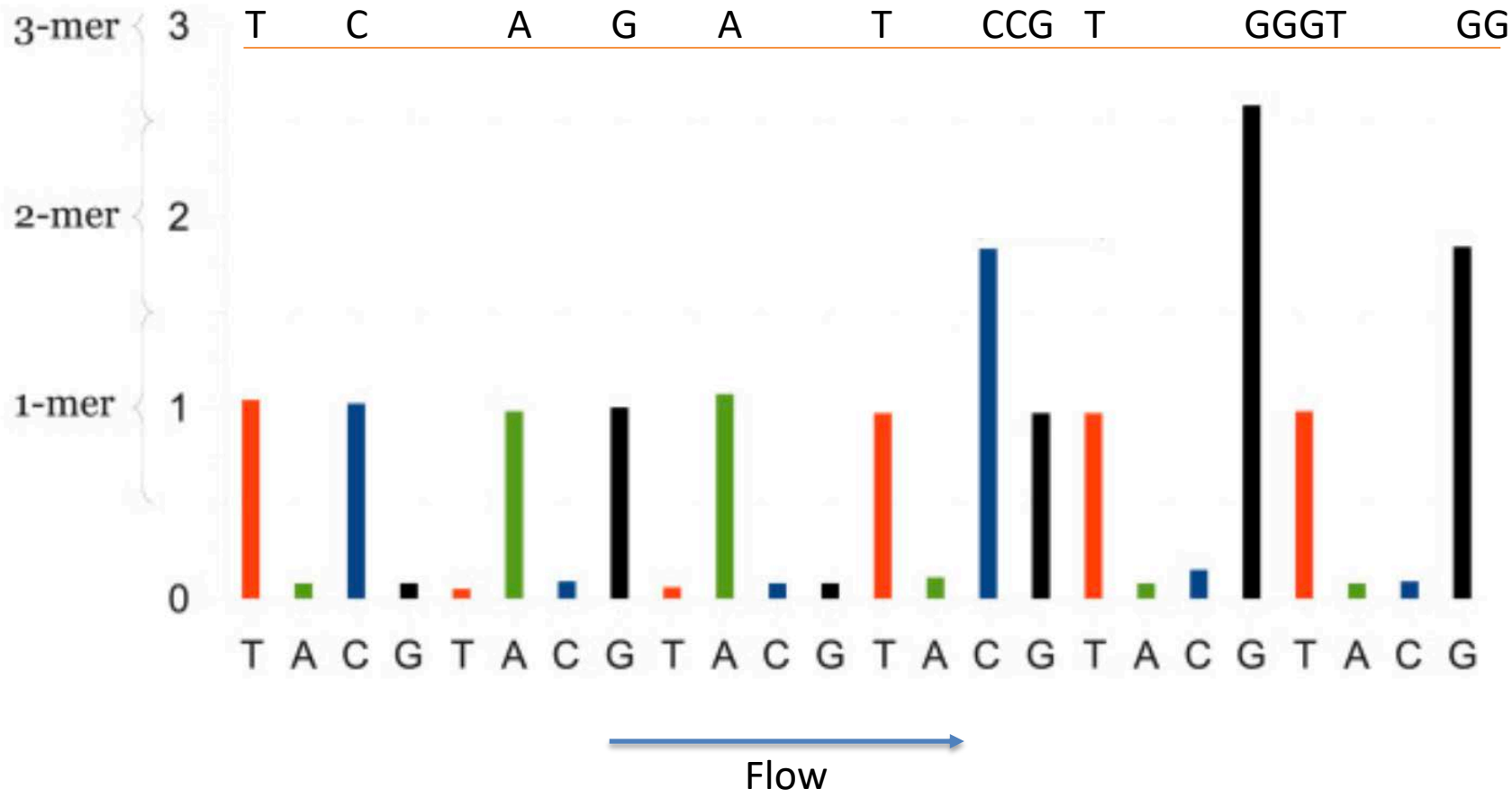
Indel error rate ~ 0.7%
Subst. error rate ~0.1%
Read accuracy > 99%



Churko J, et al. Circulation Research (2013)

NGS technologies: Ion Torrent and Roche 454

Output from Ion Torrent: Ionogram.



Template sequence

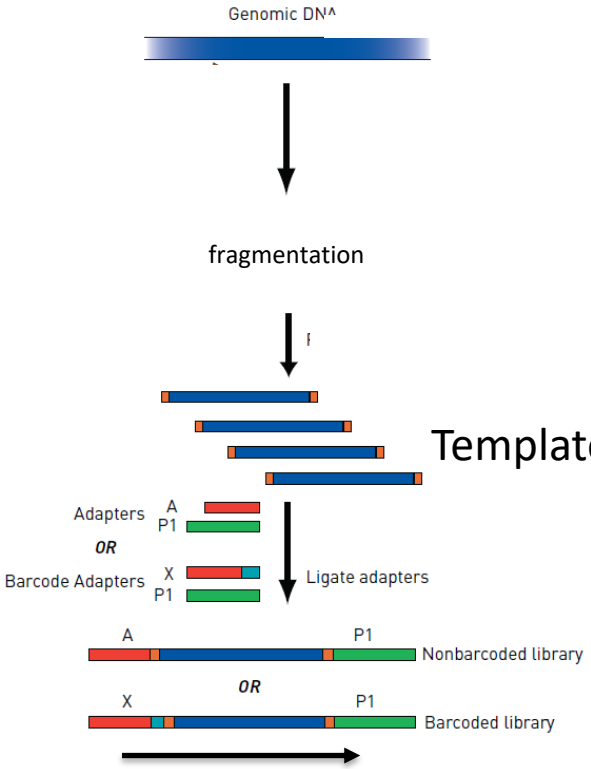
Repetitive sequential base flow [TACG][TACG]…
Not all bases are incorporated

Adapted from Lysholm F, et al. BMC Bioinf. (2011)

NGS technologies: Ion Torrent and Roche 454

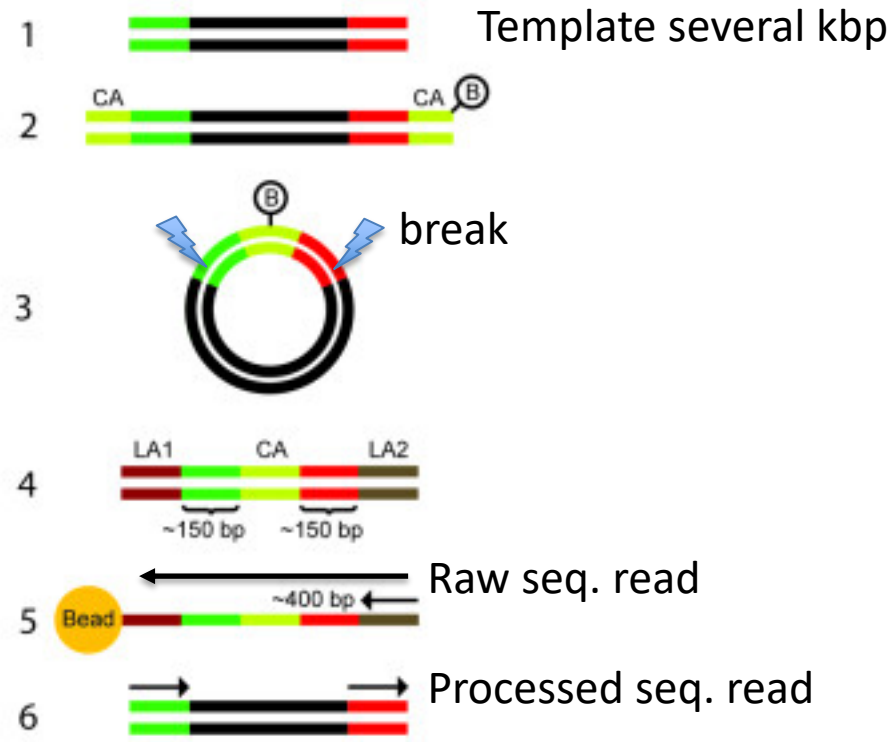
Types of sequencing libraries:

Fragment or single end



Fragment read

Mate pair

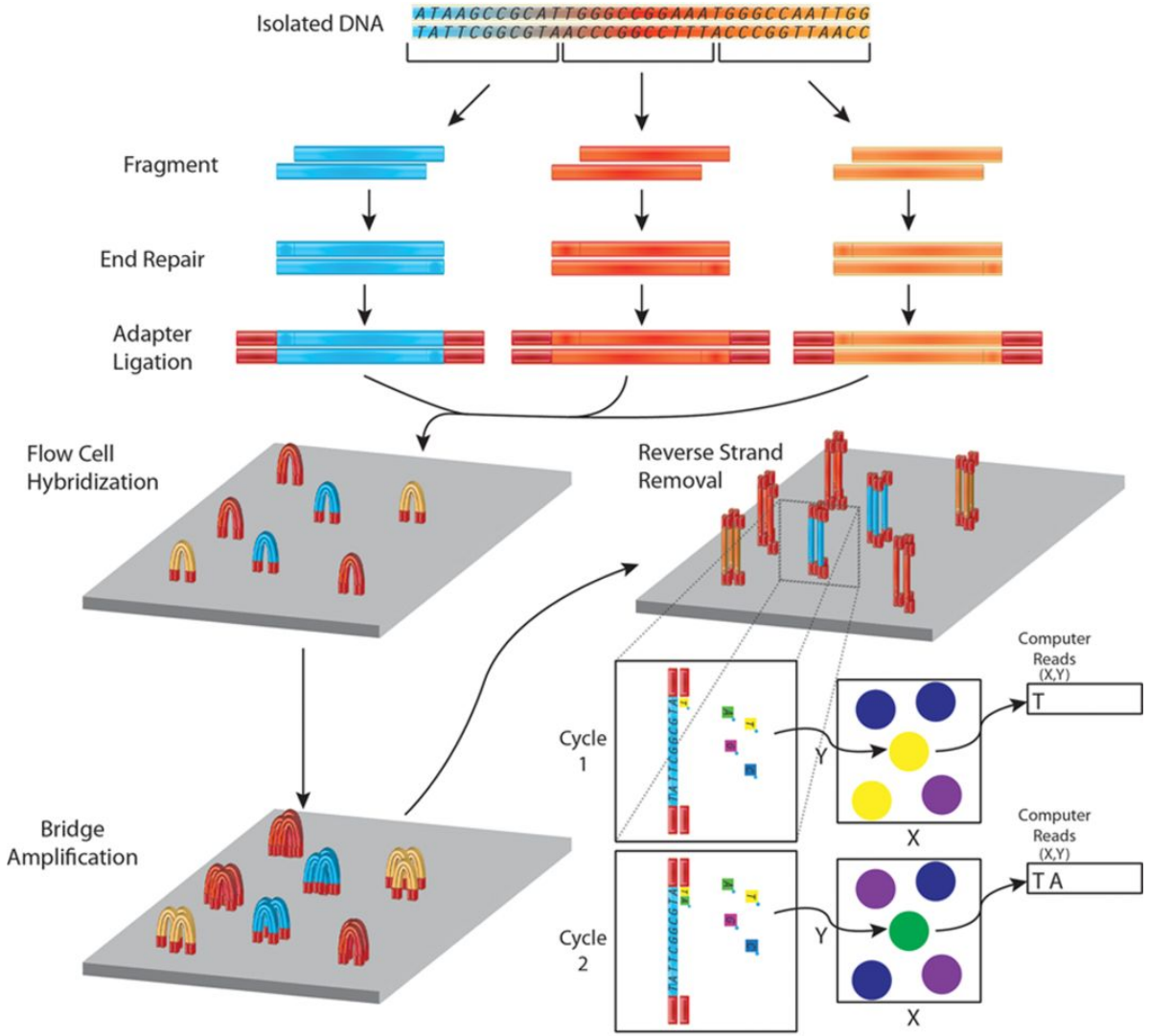


Mate pair reads (shorter)

labmedicineblog.com

Berglund E, et al. Invest. Genet. (2011)

NGS technologies: Illumina

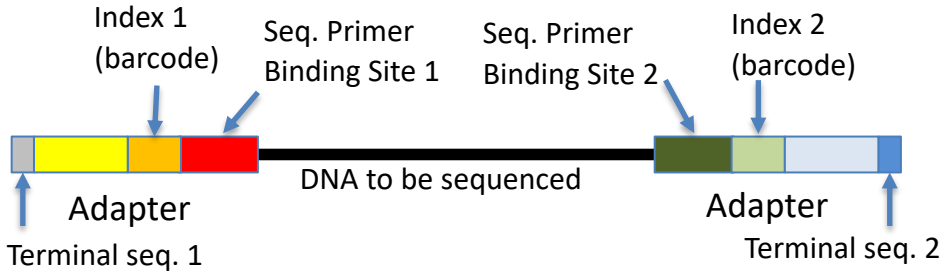


- Sequencing error (substitutions): ~0.1%; can be lowered by software.
- Read length (75 bp – 300 bp)
- Throughput up to 6,000 Gb
- Read output format: FASTQ

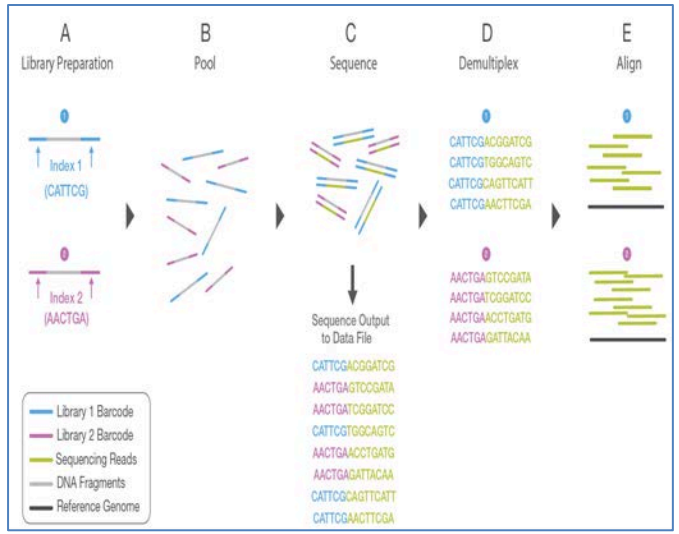
Churko J, et al. Circulation Research (2013)

NGS technologies: Illumina

Barcoded sequencing libraries and multiplexing



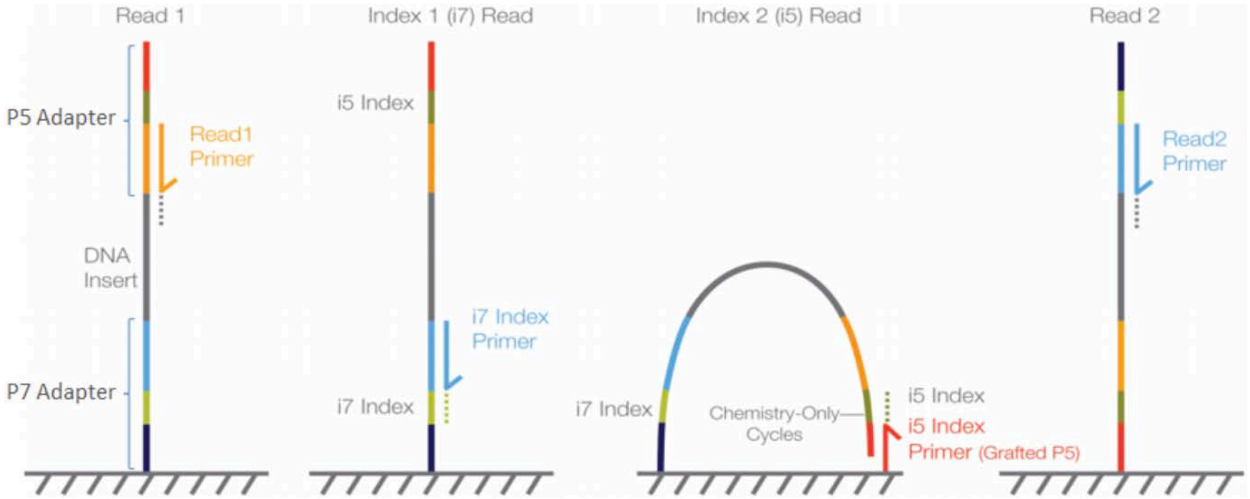
Barcoded library



www.illumina.com

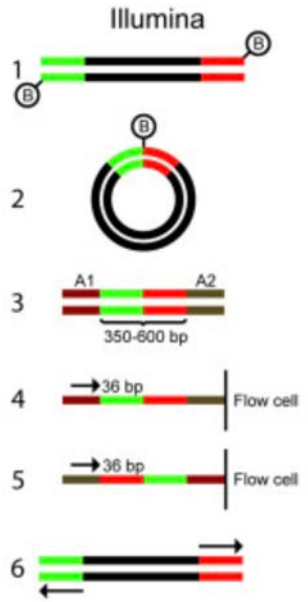
Paired-end library

Order of sequencing: Read 1 -> Index 1 -> Index 2 -> Read 2



High-Quality bases Read 1 > # High-Quality bases Read 2

Mate-pair library

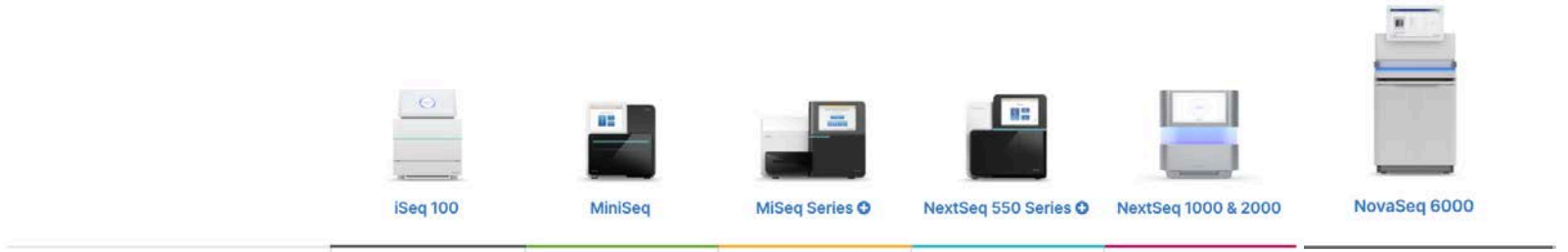


www.illumina.com

Berglund E, et al. Invest. Genet. (2011)

Sequencing technologies: Illumina

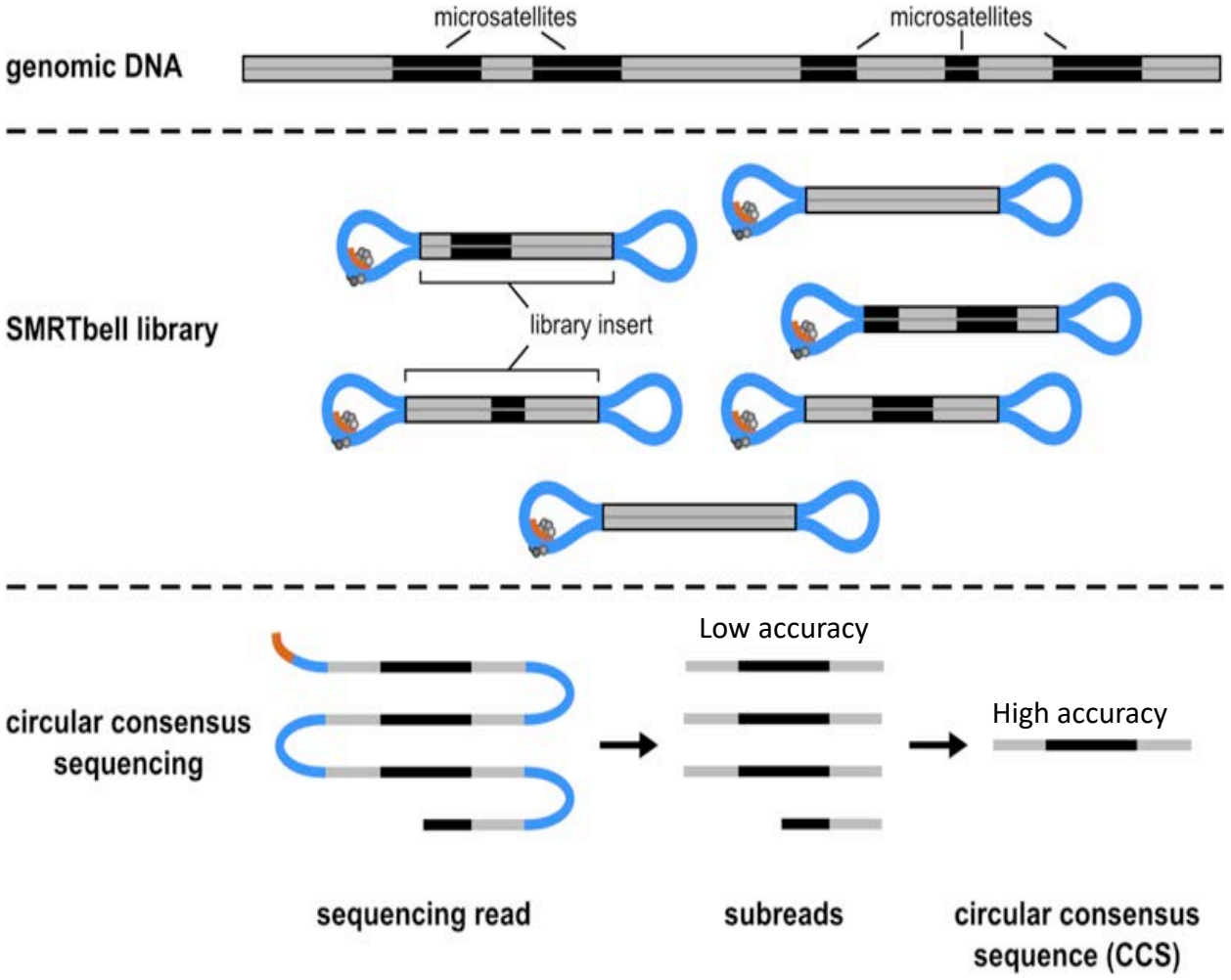
There are different Illumina sequencers available to accommodate distinct applications and budgets:



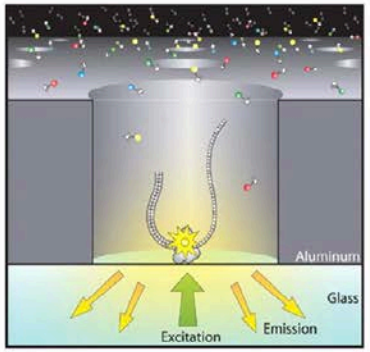
Run time	9.5-19hs	4-24hs	4-55hs	12-30hs	11-48hs	13-44hs
Max. Output	1.2 Gb	7.5 Gb	15 Gb	120 Gb	330 Gb	6,000 Gb
Max. Reads/Run	4 million	25 million	25 million	400 million	1,000 million	20,000 million
Max. Read Length	2 x 150 bp	2 x 150 bp	2 x 300 bp	2 x 150 bp	2 x 150 bp	2 x 250 bp
Applications						
Transcriptomics (total RNAseq, mRNAseq)				X	X	X
Single-Cell profiling (scDNA-seq, scRNA-seq)				X	X	X
Metagenomics sequencing				X	X	X
16S rRNA taxonomic profiling			X	X	X	X
Small whole genome seq.	X	X	X	X	X	X

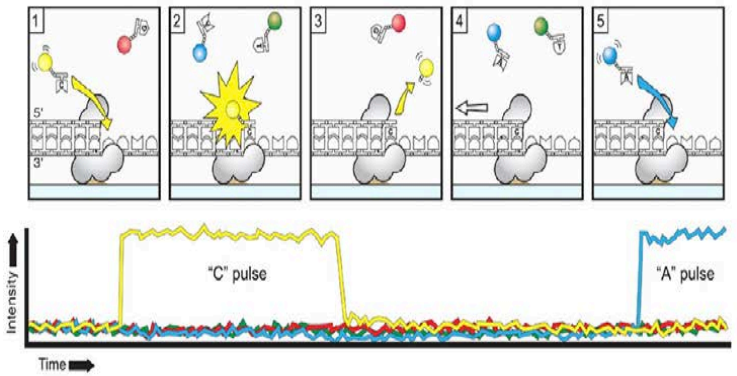
Source www.illumina.com

NGS technologies: PacBio

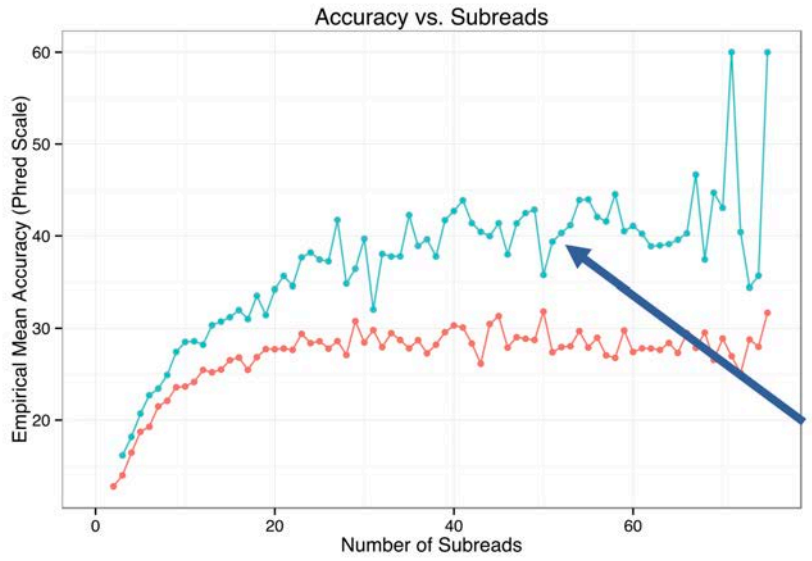


Grohme M, et al. BioTechniques. (2018)





Rhoads A. Prot. Genom. & Bioinf. (2015)



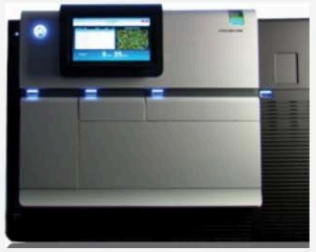
New CCS Algorithm removes bias that restricted accuracy.

www.pacb.com

NGS technologies: PacBio

PacBio Specifications

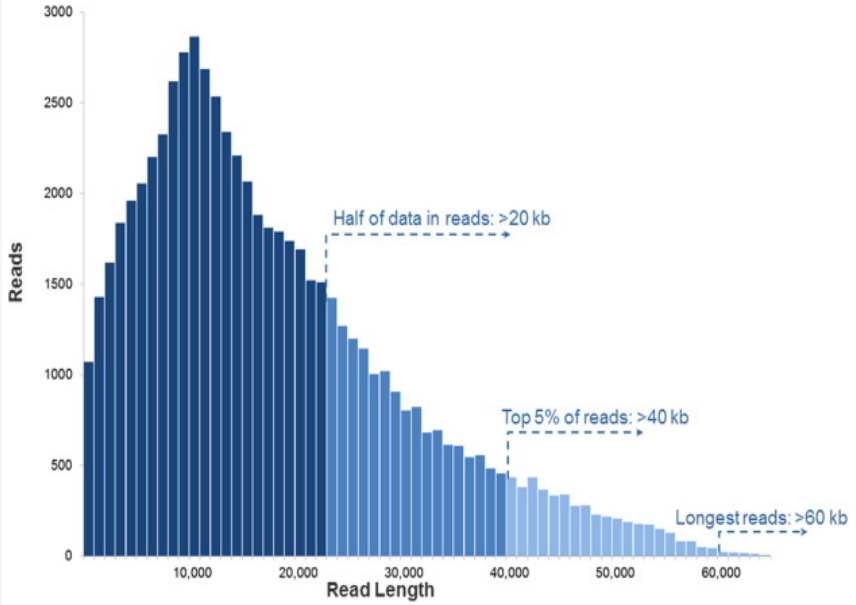
RS II



RS II

Sequel

Sequel



Rhoads A. Prot. Genom. & Bioinf. (2015)

	RS II (P6-C4)	Sequel
Run time	up to 240 min	up to 240 min
Total output	~500 Mb - 1 Gb	5 Gb - 10 Gb
Output/day	~2 Gb	20 Gb
Mean read length	10 -15 kb	10 -15 kb
Single pass accuracy	~86%	~86%
Consensus (30X) accuracy	>99.999%	>99.999%
# of reads	~50k	~500k

Pros:

- No DNA amplification required
- Fast
- Longer reads (11kb – 15kb average)
- Can be used to detect/characterize long structural variants (large inversions, translocations, tandem amplifications).
- It detects base modifications

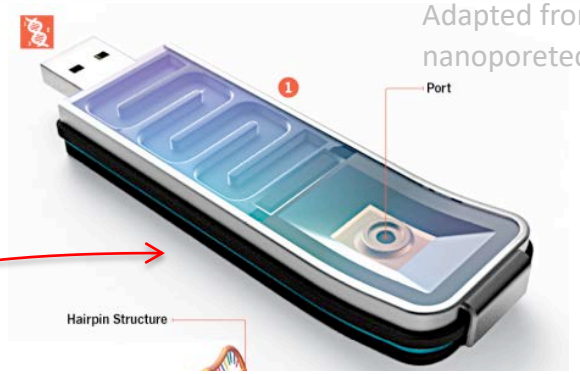
Cons:

- High error rate ~14% (but can be improved to <0.1% at the consensus level)

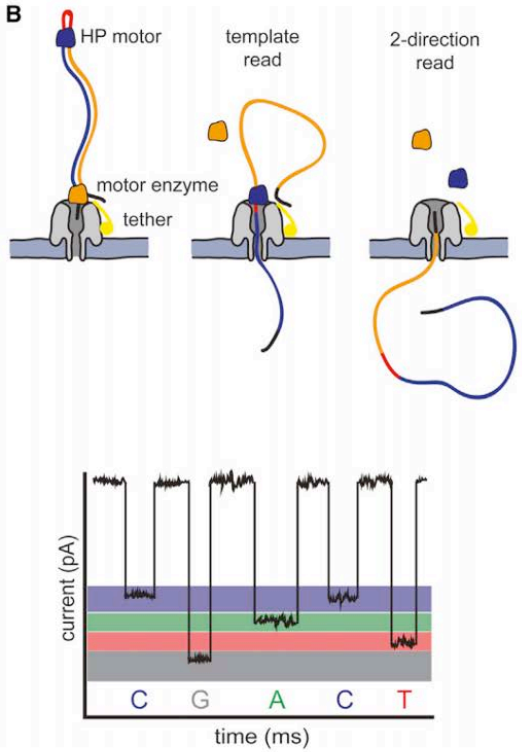
	Sequel IIe System	Sequel II System	Sequel System
Supported SMRT Cell	SMRT Cell 8M	SMRT Cell 8M	SMRT Cell 1M
Number of HiFi Reads >99%* Accuracy	Up to 4,000,000	Up to 4,000,000	Up to 500,000
Sequencing Run Time per SMRT Cell	Up to 30 hrs	Up to 30 hrs	Up to 20 hrs

www.pacb.com

NGS technologies: MinION Oxford Nanopore



Adapted from nanoporetech.com



Kchouk M, et al. Biol and Medicine. (2017)

Pros:

- Real time sequencing data
- No amplification required
- Highly portable and minimum hardware requirement
- Can sequence both DNA and RNA molecules
- Longer reads (entire DNA/RNA fragment length)
- Can be used to detect/characterize long structural variants (large inversions, translocations, tandem amplification).
- Can detect base modifications.

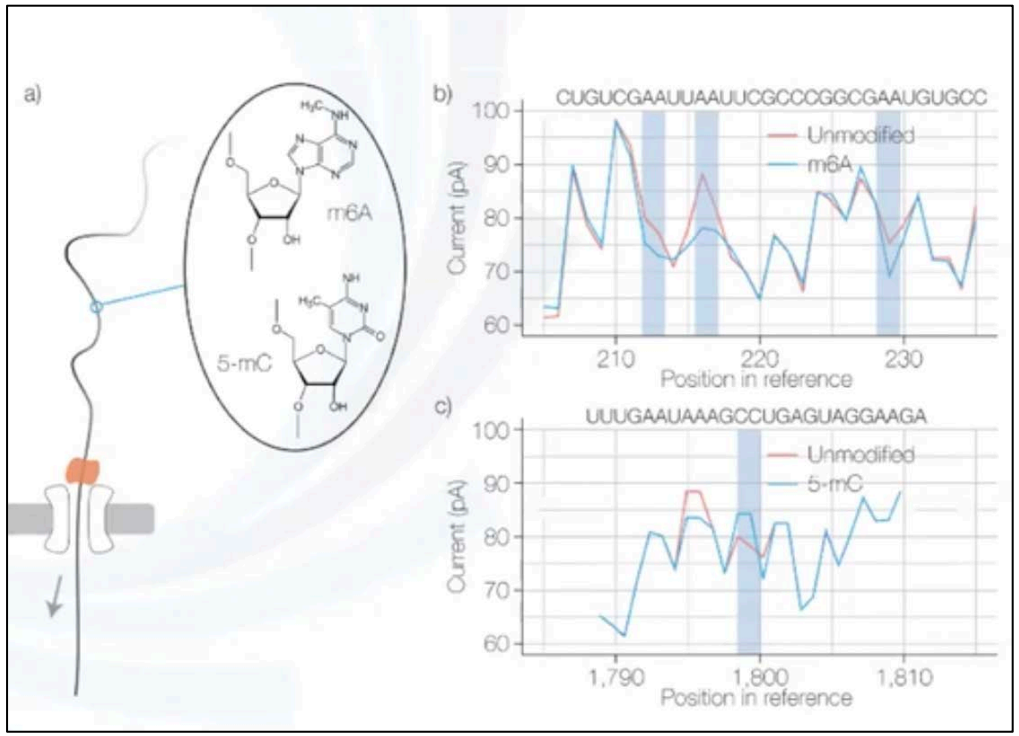
Cons:

- High error rate (median ~3 %). It could be improved with error correction algorithms and complementary Illumina sequencing data

NGS technologies: MinION Oxford Nanopore

Library types and applications:

- Sequencing of DNA or RNA molecules
- Barcoded libraries by PCR or ligation (12/24 barcodes)
- Room-temp library preparation kits (portability)
- PCR-free RNA/cDNA sequencing (portability, less bias)
- PCR-free targeted sequencing with CRISPER/Cas9 technology
- Detection of base modifications for both DNA and RNA.



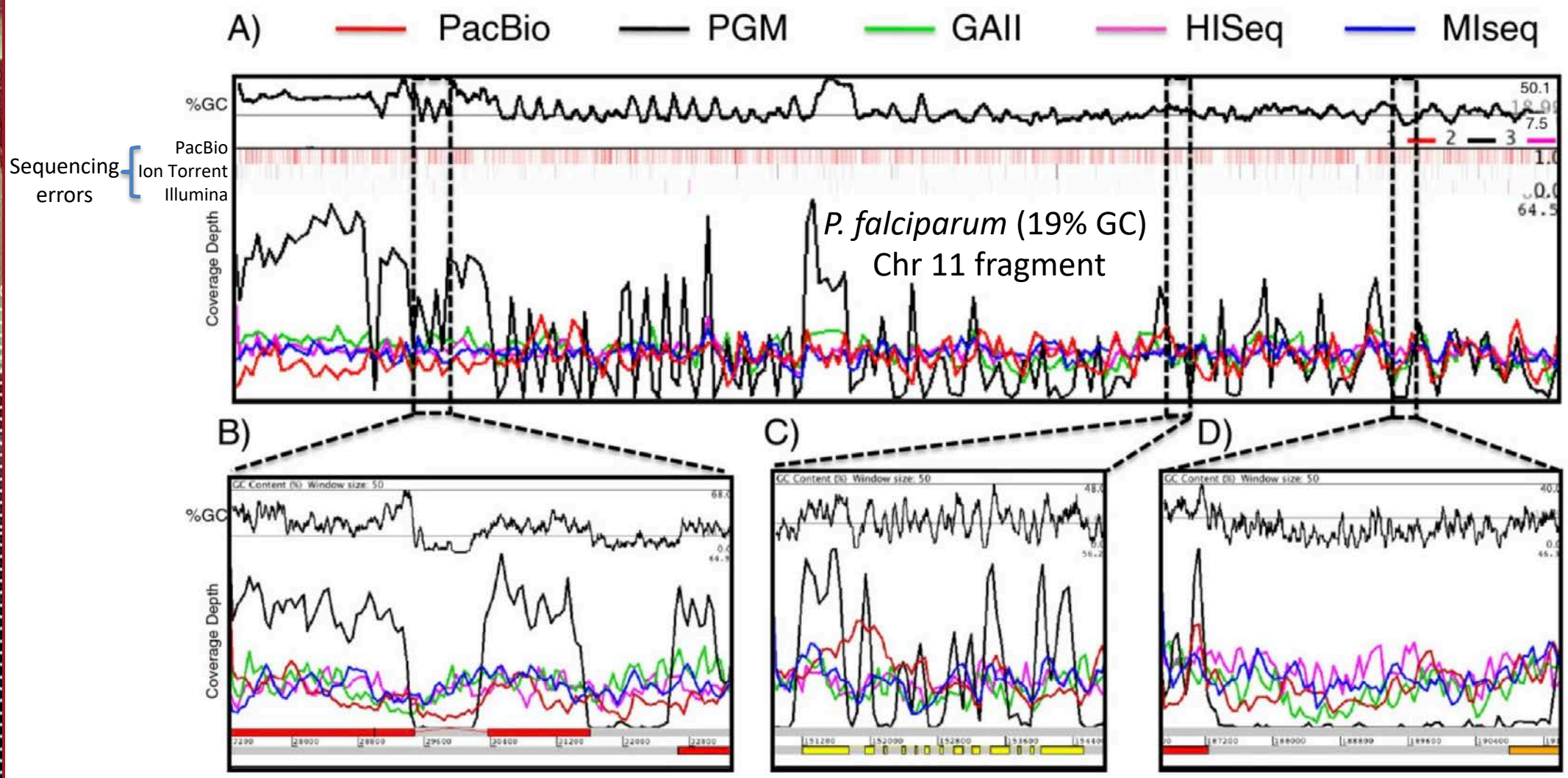
Adapted from nanoporetech.com slide by Folkard D.

Sequencing technologies: Oxford Nanopore

	Flongle	MinION Mk1B	MinION Mk1C	GridION Mk1	PromethION 24	PromethION 48
Max. yield / flow cell	2 Gb	44 Gb	44 Gb	44 Gb	242 Gb	242 Gb
# Flow Cells / device	1	1	1	5	24	48
Max. yield / device	2 Gb	44 Gb	44 Gb	220 Gb	5 Tb	10 Tb
Best in field yield / flow cell	1-1.8 Gb	42 Gb	42 Gb	42 Gb	245 Gb	245 Gb
Cost	$1,460	$1,000	$4,900	$49,995	$195,455	$265,455

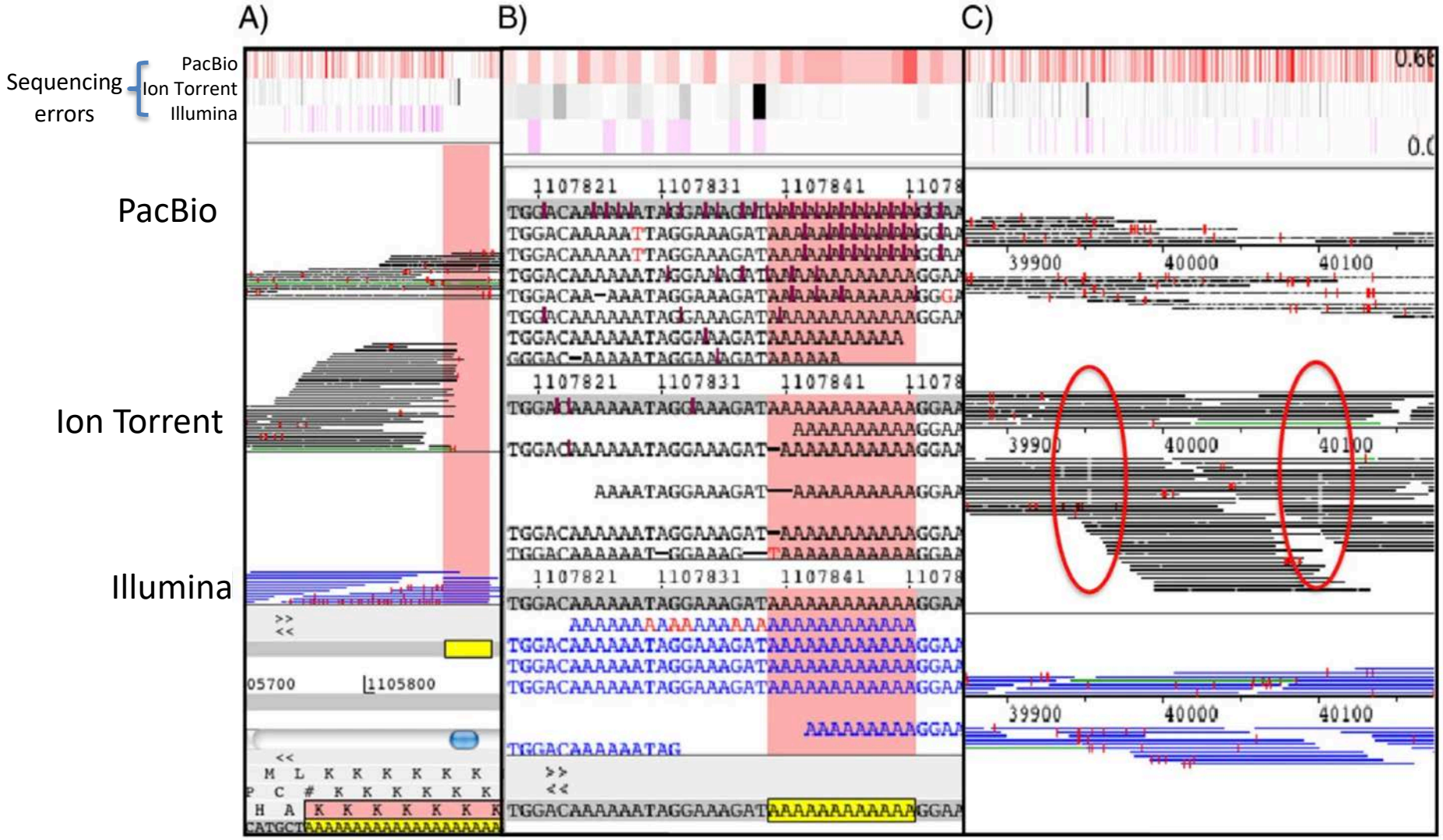
Read length: Longest read so far > 4 Mb.

Comparative analysis of sequencing vias across sequencing technologies



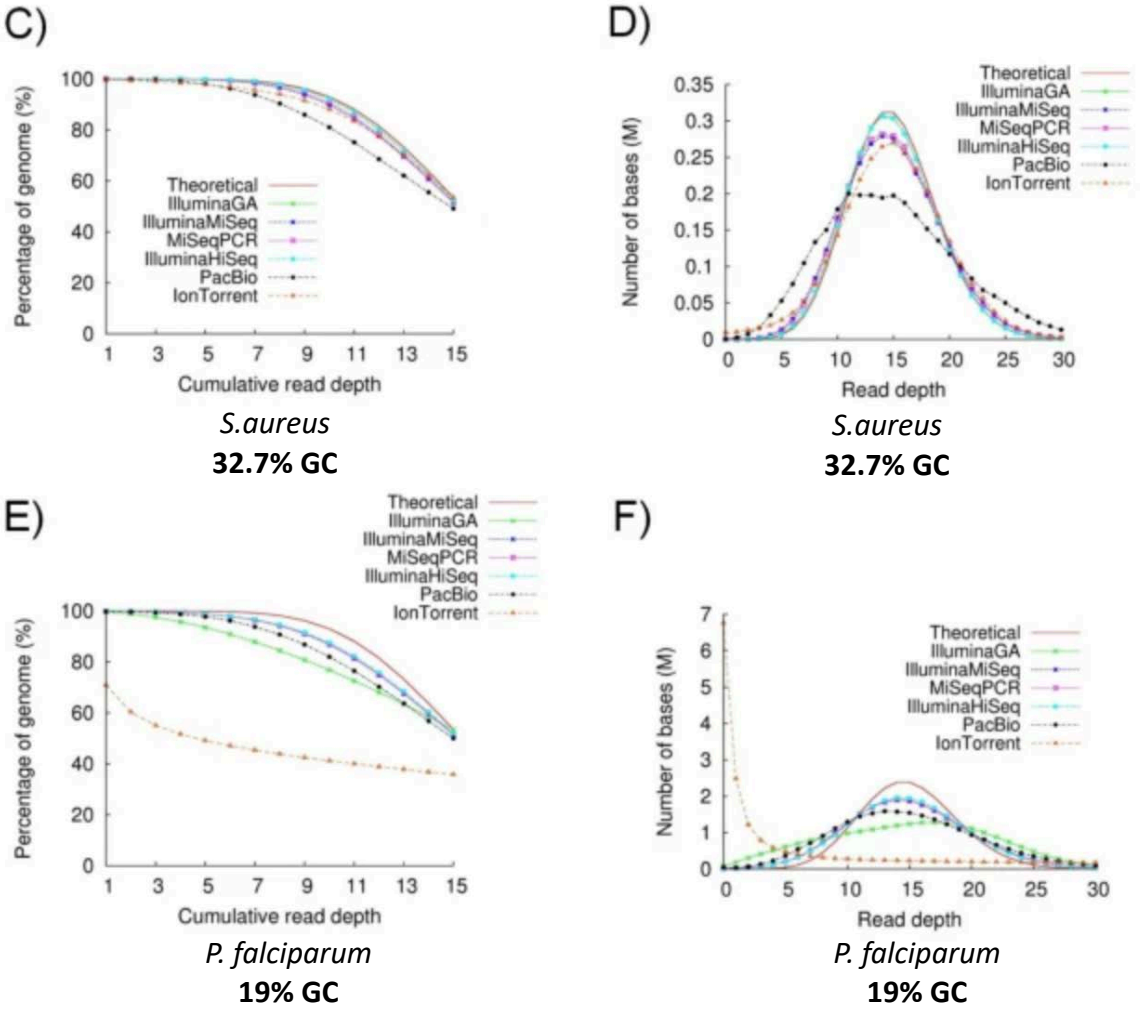
From Quail et al, BMC Genomics 2012

Comparative analysis of platform-specific sequencing errors



From Quail et al, BMC Genomics 2012

Comparative analysis of platform-specific sequencing errors



S.aureus
32.7% GC

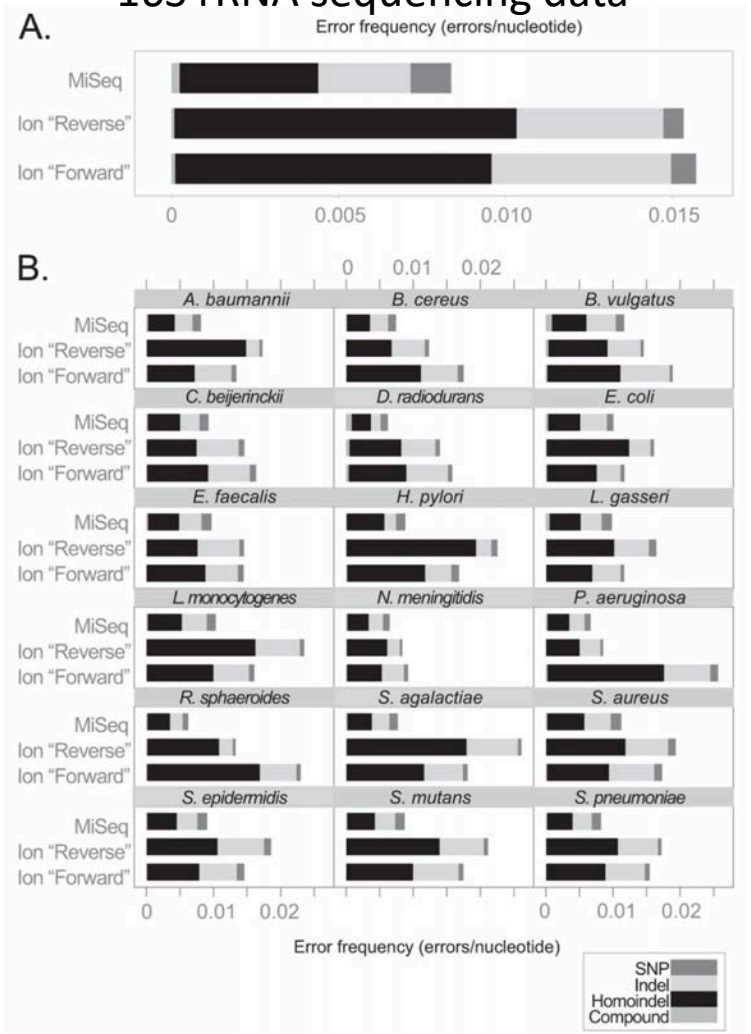
S.aureus
32.7% GC

P. falciparum
19% GC

P. falciparum
19% GC

From Quail et al, BMC Genomics 2012

16S rRNA sequencing data



From Salipante et al, App. Env. Microbiol. 2014

Sequencing data formats: FASTA

The FASTA format is used to represent sequence information. The format is very simple:

- A > symbol on the FASTA header line indicates a fasta record start.
- A string of characters called the **sequence id** follows the > symbol.
- The header line may contain an arbitrary amount of text (including spaces and tabs) on the same line.
- Subsequent lines contain the sequence (DNA, protein).

Examples

>MY_seq_ID Genomic element description
ATGCTAGGCGCGTCGCTGCTAGTTTTAGTACGT

>bar other optional text could go here
CCGTA

>Sequence_1
ACTGCAGT
TTCGNNNNNAT

>repeatmasker
ATGTGTcgggggATTTT

>prot2; my_favourite_prot
MTSRRSVKSGPREVPRDEYEDLYYTPSSGMASP

Sequencing data formats: FASTQ

FASTQ syntax

FASTQ is a standard format by which all sequencing instruments represent sequencing data. It may be thought of as an enriched FASTA format that includes quality measures for each sequence base: **FASTA with QUALITIES**.

The FASTQ format consists of 4 sections:

1. A FASTA-like header, but starting with the @ symbol followed by a **read ID** and more optional text.
2. The second section contains the nucleotide sequence, typically on a single line, but it may span several lines.
3. The third section is marked by the + sign and, optionally, followed by the same sequence id and header as the first section.
4. The fourth section encodes the quality values for the sequence in section 2, and must be of the same length as the sequence in section 2.

```
@NS500647:141:HFGTMAFXX:1:11101:11683:1016 1:N:0:TGAAGAGA
CCCGCNGCTCTGCCTCGTCTGCTGCGAGGGCAAGCAGCGAAACGAAGGCGCCGCAGCCGCTTCTCTGGTGCA
+
AAAAA#EEEEEEEEEEEEEE!"#$%&'()*+,-./0123456789:;<=>?@AEEEE/EEEEEEEAEEEEEE
```

Sequencing data formats: FASTQ

FASTQ quality values

Each of the ASCII characters in the 4th section represents a Phred score, encoded via a single letter encoding.

Phredd+33 scale

```
Quality values: !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
                |    |    |    |    |    |    |    |    |
Phred scores:   0....5...10...15...20...25...30...35...40
    (PS)        |                                        |
                worst                                  best
```

Phred scores represent the error probabilities of each base call in the sequence based on the formula:

$$\text{Error} = 10^{-PS/10}$$

- PS = 0 => 1 error / 1 bp or 100% error probability
- PS = 10 => 1 error / 10 bp or 10% error probability
- PS = 20 => 1 error / 100 bp or 1% error probability
- PS = 30 => 1 error / 1,000 bp or 0.1% error probability
- PS = 40 => 1 error / 10,000 bp or 0.01% error probability

Sequencing data formats: FASTQ

- There was a time when instrumentation makers could not decide at what ASCII character to start the Phred scores. The **current standard** shown in the previous slide is called Sanger (Phred+33) format where Phred scores start at ASCII character number 33 = “!”.
- Previous versions of Illumina/Solexa (< version 1.8) used another set of ASCII characters to represent quality values starting at or nearby ASCII value 64 (Phred+64).
- PacBio uses Phred+33 scale.

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
.................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
PPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                     |
33                        59   64       73                            104                   126
0........................26...31.......40
                         -5....0........9.............................40
                               0........9.............................40
                                  3.....9..............................41
0.2......................26...31........41
0..................20........30........40........50..........................................93

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
P - PacBio        Phred+33,  HiFi reads typically (0, 93)
```

Raw read quality values will depend on the sequencing technology and the base caller program used.

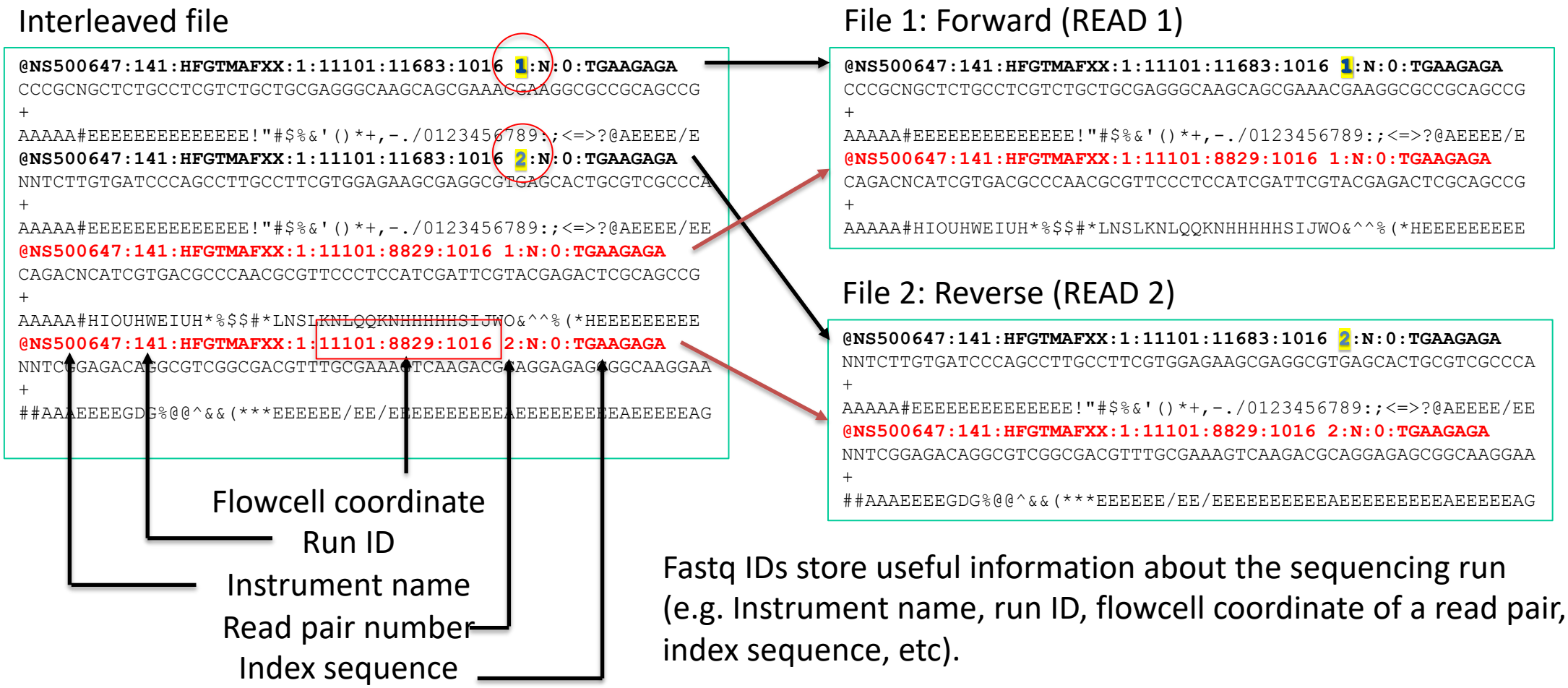
Typically for Illumina they go up to 41.

Larger quality values can be reached (e.g. PacBio consensus reads or mapped data)

Adapted from https://en.wikipedia.org/wiki/FASTQ_format

Sequencing data formats: FASTQ

Paired-End reads can be stored in the same fastq file (interleaved) one after the other, or as two different fastq files, in the same order.



Fastq IDs store useful information about the sequencing run (e.g. Instrument name, run ID, flowcell coordinate of a read pair, index sequence, etc).

Sequencing data formats: SAM/BAM

- The Sequence Alignment Map (SAM) format is usually used to represent the results of aligning a FASTQ file to a reference FASTA file.
- It is basically a tab-delimited text file containing sequencing read and mapping information.
- A BAM file is a binary version of a SAM file.
- PacBio processed sequencing data can also be stored as a BAM file.
- It has 2 sections:
  1. Header
  2. Alignment

1. Header (optional but necessary for some applications):
   Lines start with "@".
   Harbors information about the file itself, the reference and the sample.

```
@HD    VN:1.0  SO:coordinate
@SQ    SN:chr1 LN:249250621
@SQ    SN:chr10      LN:135534747
@SQ    SN:chr11      LN:135006516
@RG ID:My_sample_ID PL:Illumina SM:My_sample_name
@PG ID:bowtie2 PN:bowtie2 VN:2.2.4 CL:"bowtie2-align"
```

Mapped reads sorted by coordinate on the reference

Reference sequence IDs and their length in bp.

Read Group

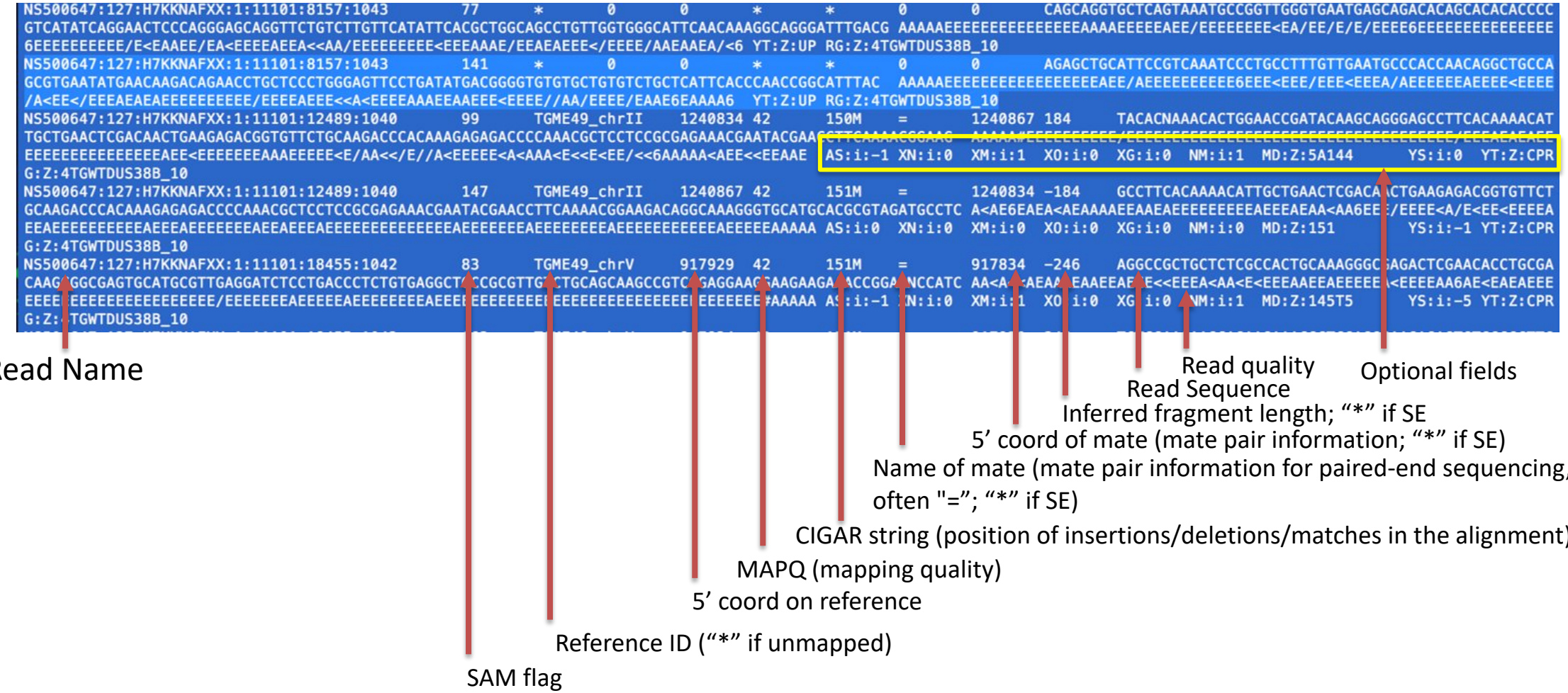
Information about the program and parameters used to generate the SAM/BAM file

Sequencing data formats: SAM/BAM

2. Alignment (required):

   Aligned read information (1 read and its alignment info per row – cigar format):

   Eleven mandatory fields with alignment and sequence information plus additional optional fields.



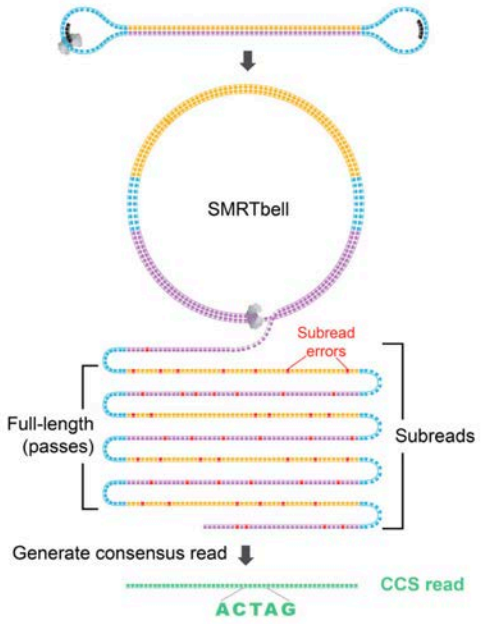
Sequencing data formats: SAM/BAM

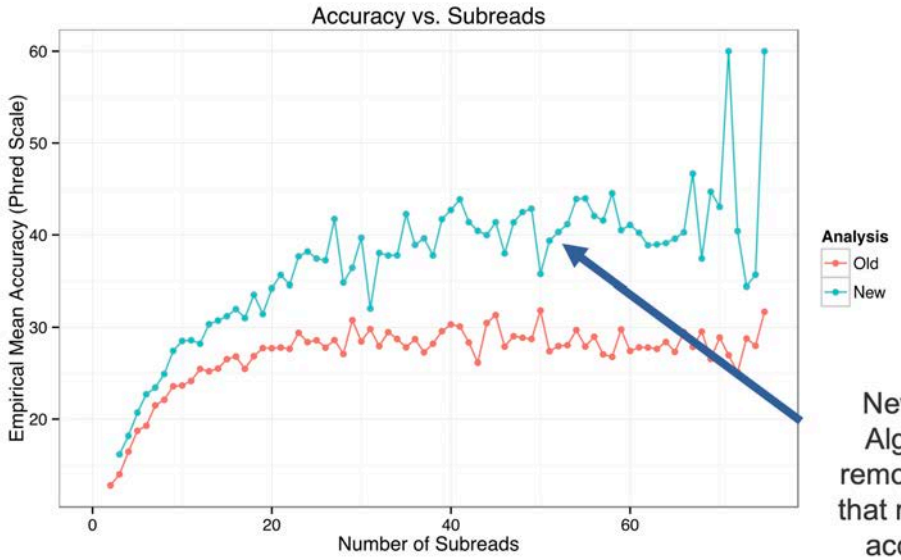
PacBio BAM files:

BAM files are also used to store PacBio subread information that can be used to build High-Fidelity CCS consensus reads.

Each row contains information for one subread (position within the whole PacBio read, quality values, etc.).

Subreads are sorted by their position in the PacBio read.





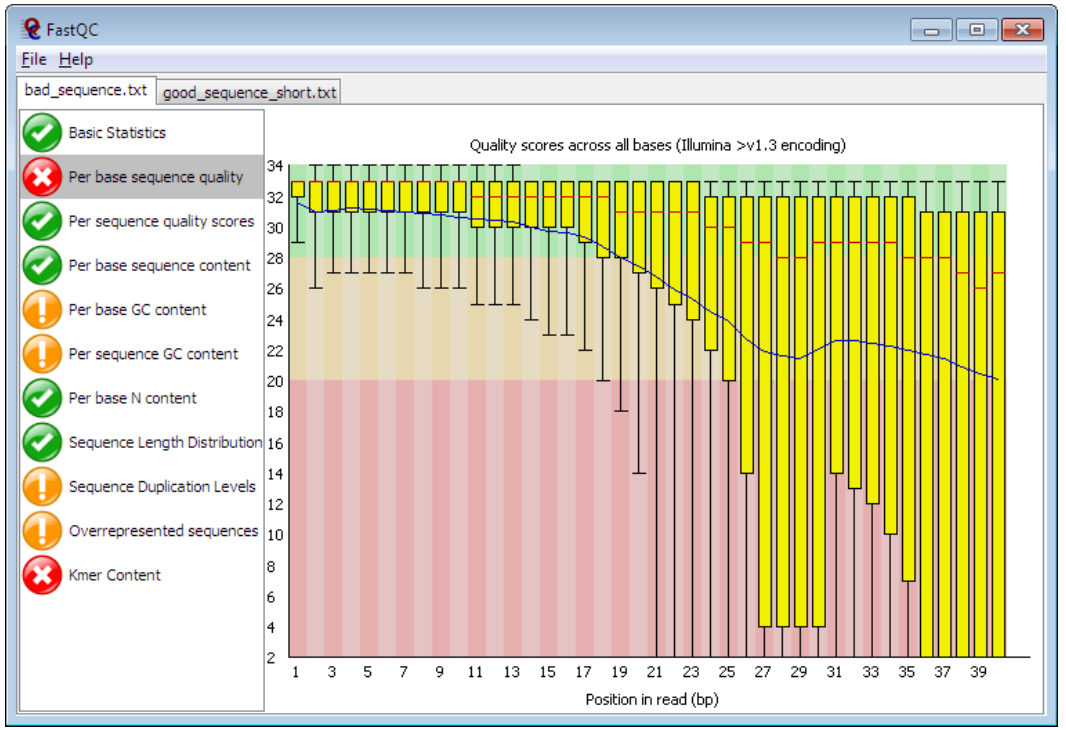
New CCS Algorithm removes bias that restricted accuracy.

www.pacb.com

Preprocessing of sequencing reads: Read Quality Check

Goal: to assess if sequencing reads contain:

- Unexpected short read size
- Low number of reads
- Low quality of bases
- Presence of adapters or indexes
- Any sequencing bias (overrepresentation of specific k-mers)

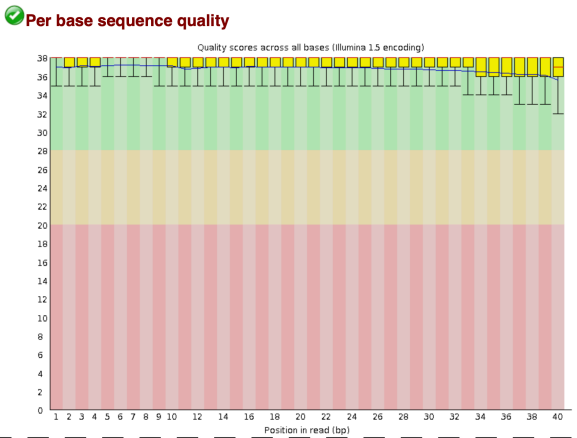


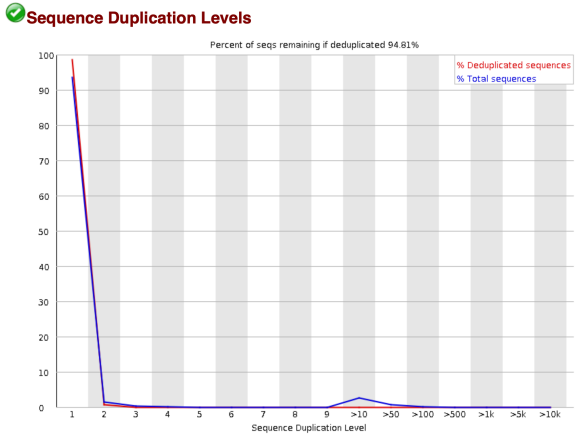
www.bioinformatics.babraham.ac.uk/projects/fastqc/

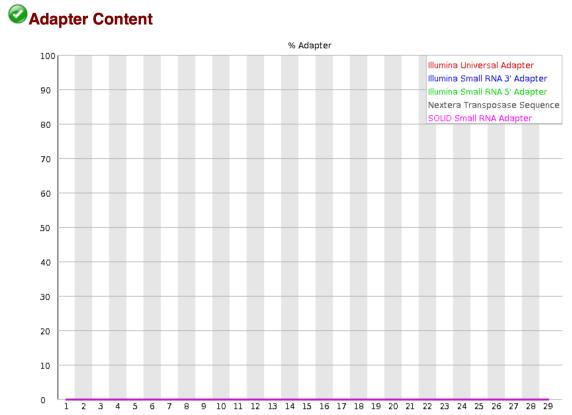
Preprocessing of sequencing reads: Read Quality Check

FastQC (Andrews, 2010) is a quality control tool for high throughput sequence data in Fastq format (Illumina/PacBio/454).
Command line fastqc or graphical interface
It generates complete HTML report to spot problem originating from sequencer, library preparation, contamination.
Report contains summary graphs and tables to quickly assess your data.

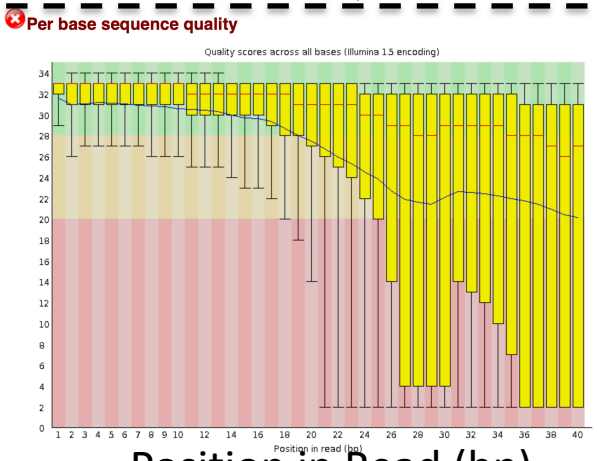
Good quality reads



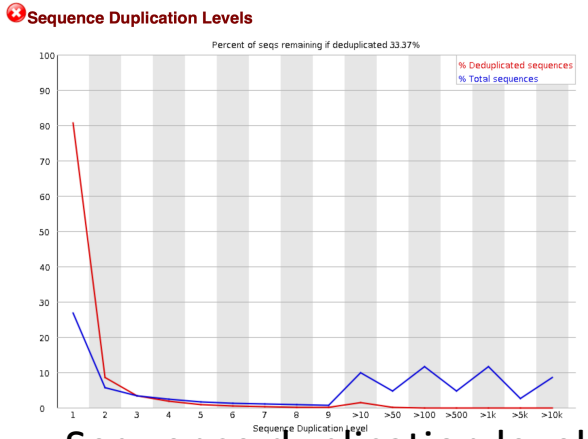




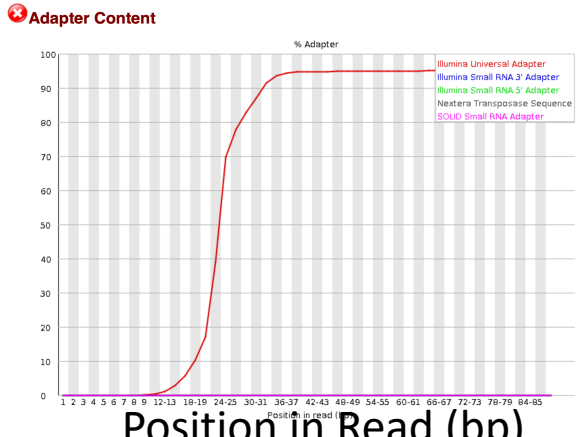
Bad quality reads



Position in Read (bp)



Sequence duplication level



Position in Read (bp)

Preprocessing of sequencing reads: Cleaning of Fastq files

Objective:

Detect and remove sequencing adapters (still) present in the FastQ files
Filter / trim reads according to quality (as plotted in FastQC)

Highly recommended step before performing any analysis with the sequencing data.

There are several tools available:

Sickle: trims sequences based on quality values (https://github.com/najoshi/sickle).

cutadapt: detect and removes adapter sequences, primers, poly-A tails and other types of unwanted sequences from high-throughput sequencing reads (https://cutadapt.readthedocs.io/en/stable/index.html).

Trimmomatic: a highly configurable tool for trimming adapters and removing low-quality bases (http://www.usadellab.org/cms/?page=trimmomatic).

fastp: an all-in-one and ultra-fast trimming tool. It supports multithreading (https://github.com/OpenGene/fastp#examples-of-report).

NGS Methods

- **Genomics**
  - Reference-based Whole Genome Sequencing
  - Targeted Sequencing: Exome capture
  - De Novo Genome Sequencing
- **Transcriptomics**
  - Total RNA and mRNA sequencing
  - Targeted RNA Sequencing
  - Small RNA and Noncoding RNA Sequencing

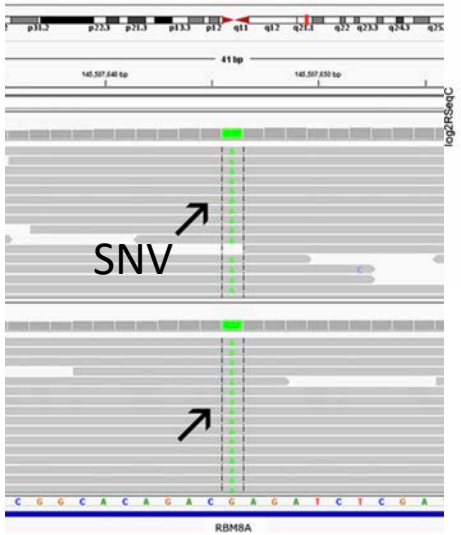
- **Epigenomics**
  - Methylation Sequencing
  - Chromatin Immunoprecipitation (ChIP) Sequencing

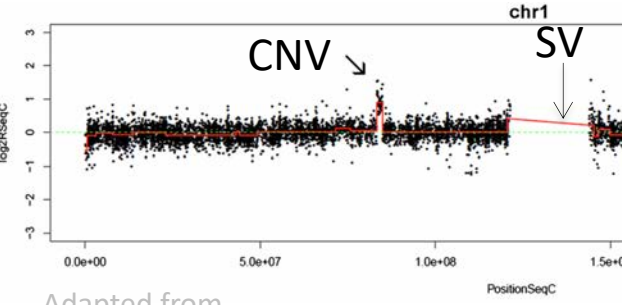
nextgenseek.com

NGS Methods – Reference-based Whole Genome Sequencing

Reference-based Whole Genome Sequencing (pre-existent reference genome sequence available):

1. Identification of Single Nucleotide Variants (SNV), Copy Number Variations (CNV) and Structural Variations (SV).
2. Genome-wide association Studies (GWAS).
3. Characterization of mutants.
4. Population studies.
5. Characterization of new strains (pathogenicity islands, plasmids, virulence factors).





Adapted from
Vetro A. et al, Clinical Gen (2017)

Typical workflow

Genomic DNA from sample

↓

Fragmentation

↓

Sequencing library preparation and sequencing

↓

Raw Fastq files

↓

Read quality check (e.g. FastQC)

↓

Read cleaning (trimming of adaptors and low quality reads/bases)

↓

Clean Fastq files

↓

Map reads to reference genome (e.g. BWA, bowtie2)

↓

Raw SAM/BAM alignment file

↓

Removal of duplicated reads (coming from the same DNA template)

↓

Realign reads around complex genomic regions (e.g. di/trinucleotide tandem repeats)

↓

Final alignment BAM file

↓

Analyses (e.g. SNV/InDels, reference-based assembly, base modifications, etc.)

NGS Methods – Targeted Sequencing

Targeted Sequencing: Sequencing of specific genome regions/transcripts of interest.

Example: Exome Sequencing
Sequencing of a subset of protein-coding sequences of a genome.

Approach:

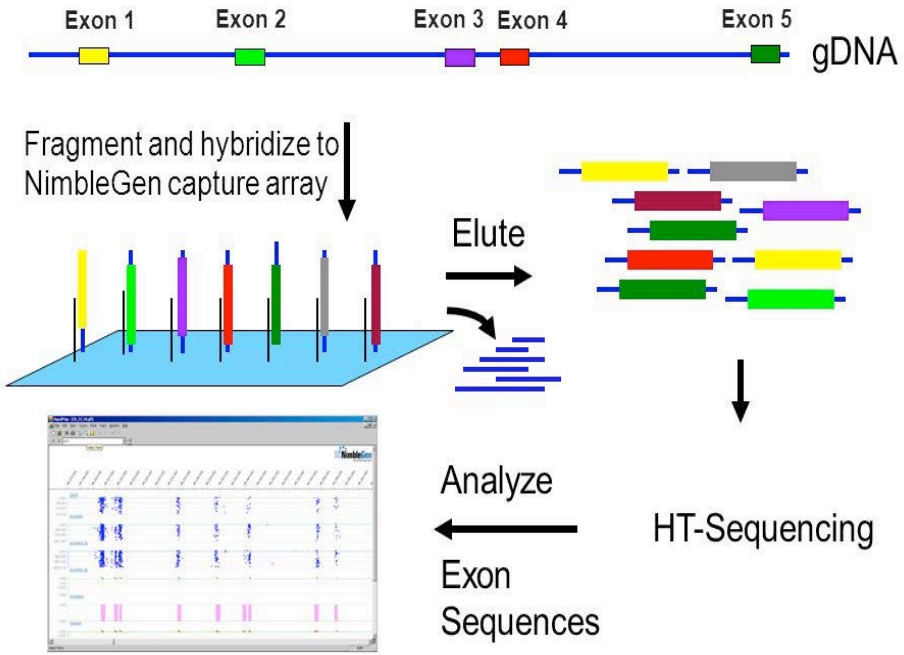
- PCR
- DNA/RNA capture

1. gDNA fragmentation

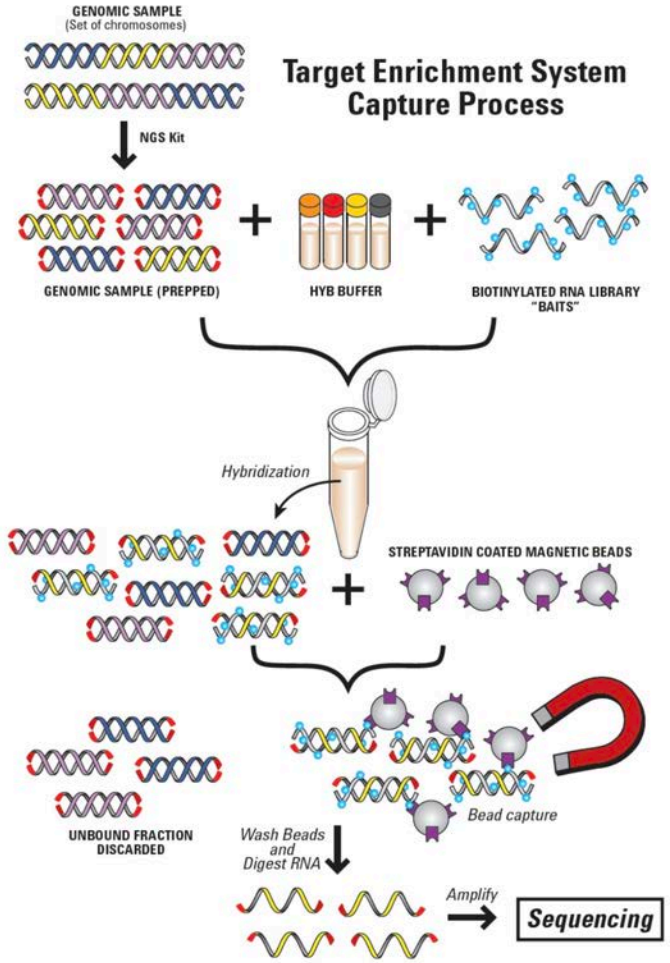
2. Enrichment of exon-containing fragments.
   - PCR amplification
   - Hybrid Capture
   - In-solution Capture
   - Others

3. NGS Sequencing (Usually Illumina sequencing).

Hybrid capture[b]



In-solution Capture[a]



(a) https://www.ddw-online.com/bringing-cost-and-process-efficiency-to-next-generation-sequencing-731-200908/
(b) https://www.slideserve.com/nuru/next-generation-sequencing-platforms-sequencing-by-synthesis-sbs-454-pyrosequencing

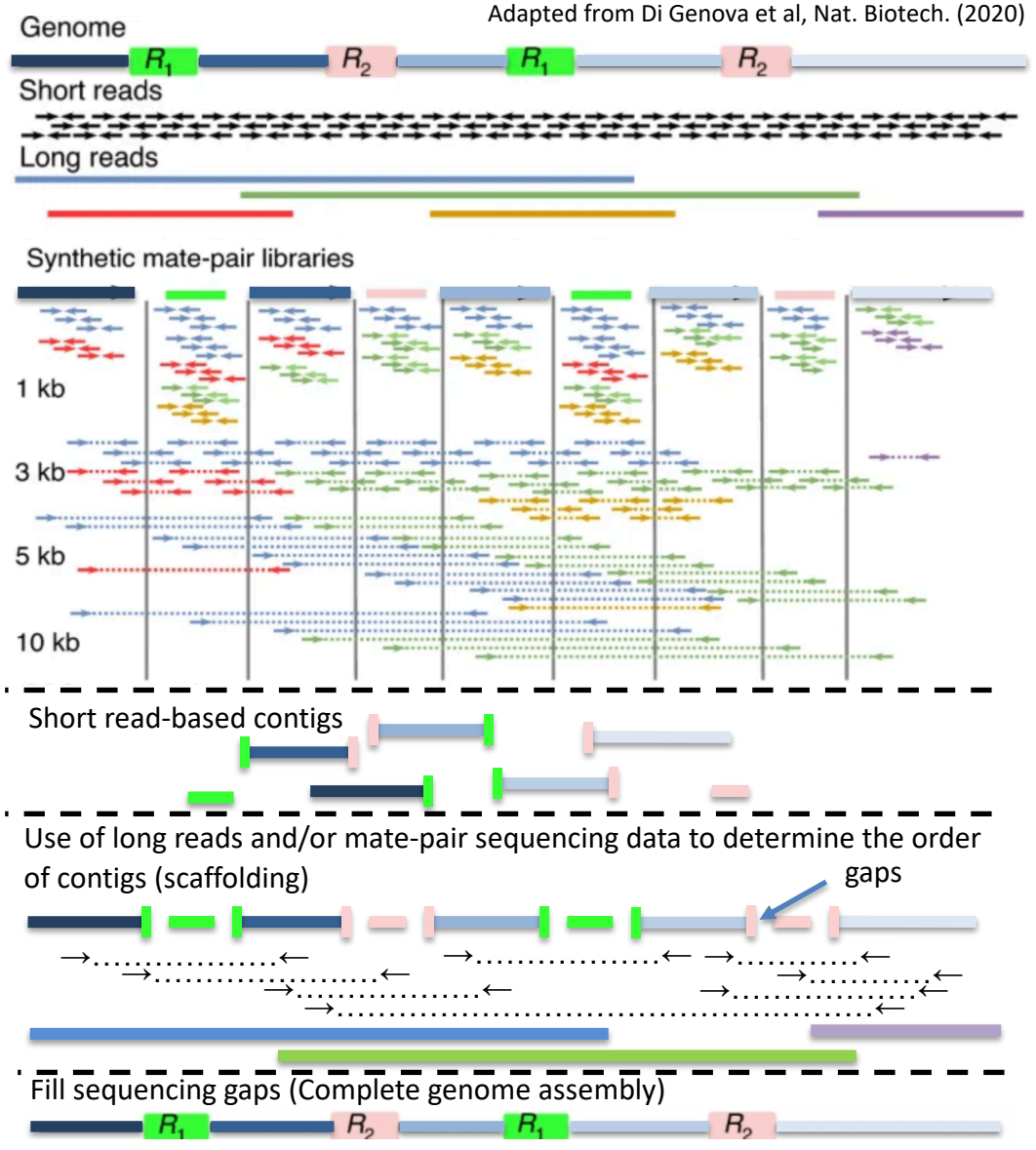
NGS Methods – De-novo Whole Genome Sequencing

De-novo Whole Genome Sequencing (No reference needed):
Sequencing and assembly of novel genome without the guidance of a reference genome sequence.

Steps

1. gDNA fragmentation.
2. Choose of sequencing technology(ies) (depends on genome properties)
   - AT-rich / GC-rich genomes
   - Repetitive genomes
   - Genome size
3. Sequencing library preparation:
   - Different insert sizes (depending on the nature of the genome).
   - Paired-end.
4. Assemble sequencing reads:
   - Depends on genome properties.
   - Sequencing libraries chosen
   - Sequencing technology
5. Structural and functional annotation of the genome assembly.

Adapted from Di Genova et al, Nat. Biotech. (2020)



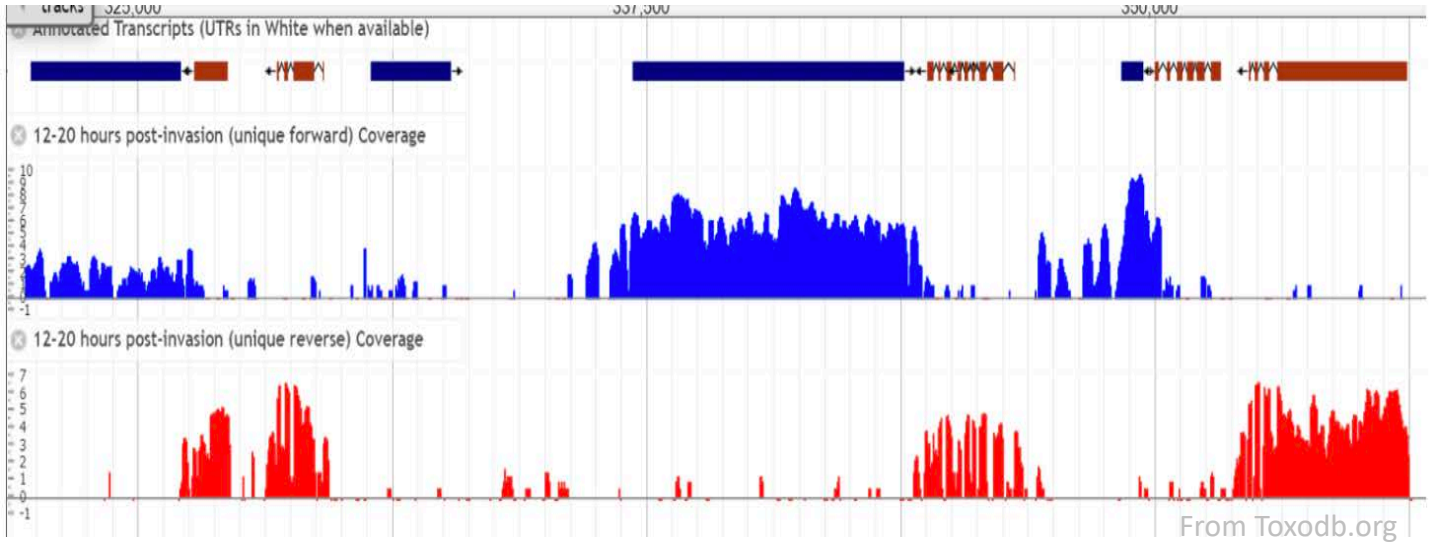
NGS Methods – RNA-seq

Total RNA and mRNA sequencing

- Differential gene expression analysis
- RNA (gene) discovery.
- Identification of common, rare and novel transcripts.
- Detection of isoforms.
- Detection of splice junctions.
- Identification of transcribed strands.
- Structural (gene) annotation.
- RNA base modifications.

Library Preparation and sequencing technology will depend upon source of RNA and project goal (Bacterial or eukaryotic RNA, blood RNA, etc.).

Strand-specific RNA-seq analysis



NGS Methods - Transcriptomics

Targeted RNA sequencing

- Measuring expression of specific genes, pathways or diseases.

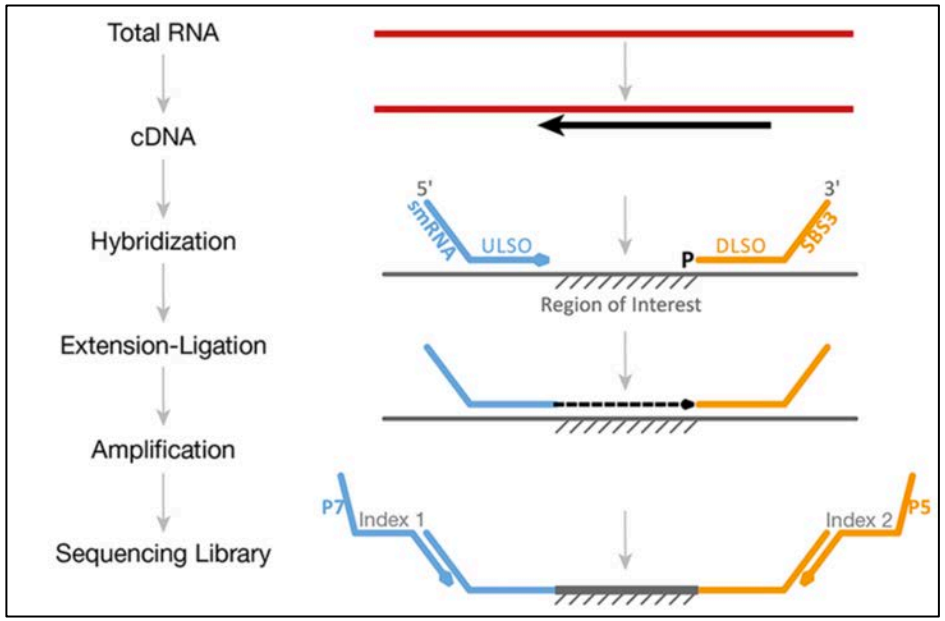
Small RNA and Noncoding RNA Sequencing

- Measuring expression of 18-22b RNA population.

RNA sequencing approaches:
Illumina sequencing:
75 bp vs 150 bp vs 250 bp reads
Fragment reads vs paired-end reads.

Short reads (Illumina) vs long reads (PacBio, Oxford Nanopore)

Illumina TruSeq® Targeted RNA Sequencing kits



www.illumina.com

NGS Methods - Epigenomics

Identification of nucleotide modifications along the genome

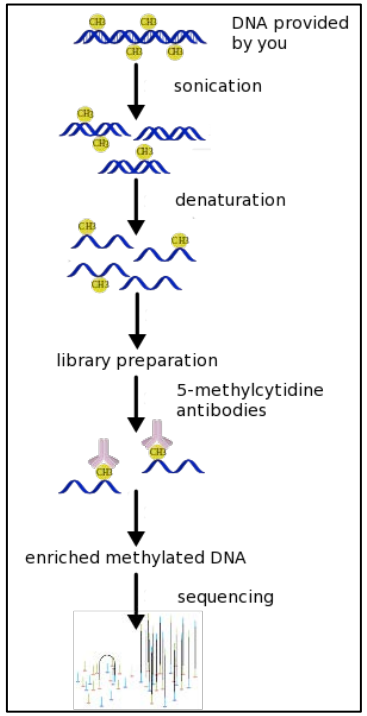
Indirect methods:

- Whole- genome bisulfite sequencing (WGBS).
- Methylated DNA Immunoprecipitation sequencing (MeDIP-seq) and Illumina sequencing.

Direct methods:

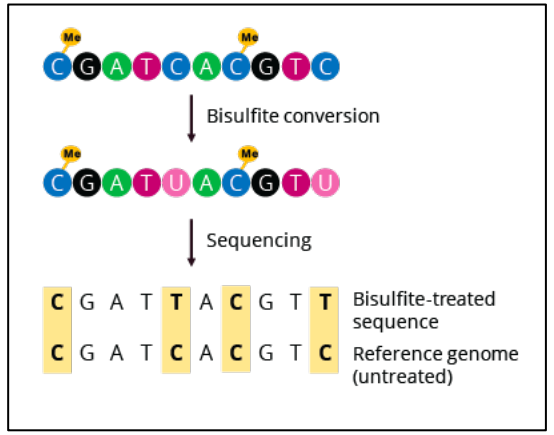
- PacBio sequencing
- Nanopore MinION

MeDIP-seq

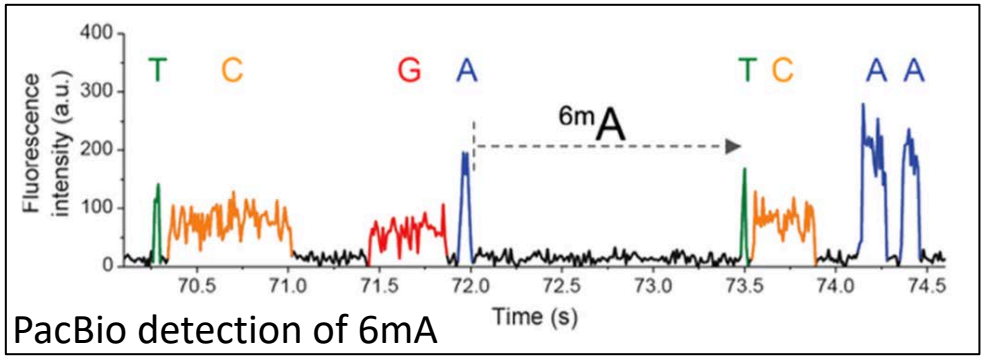


www.nxtgnt.ugent.be

WGBS



www.genewiz.com

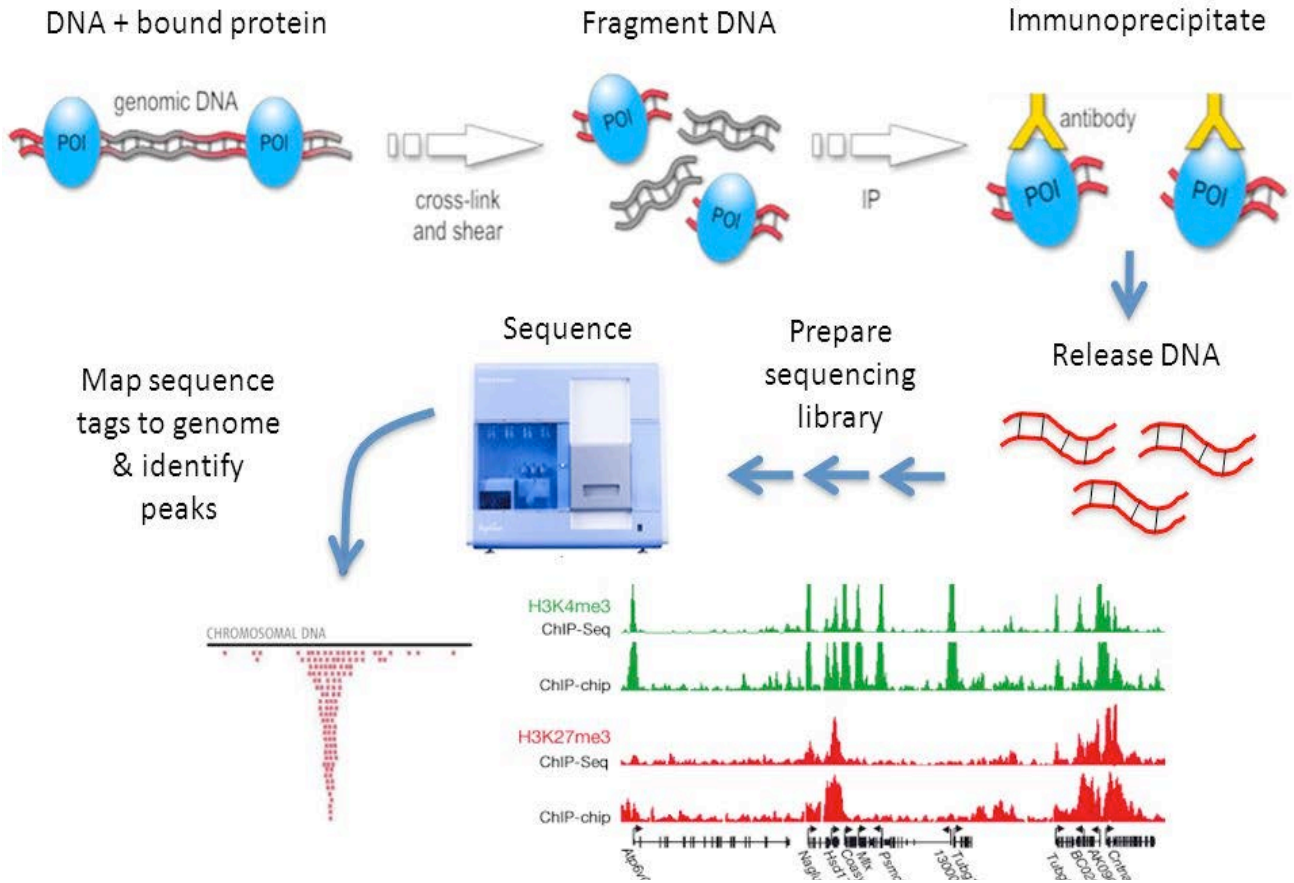


PacBio detection of 6mA

www.pacb.com/basemods

NGS Methods - Epigenomics

Chromatin Immunoprecipitation sequencing (ChIP-seq)



Thank you!