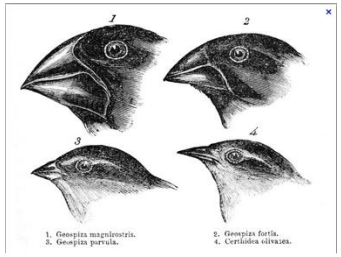
INTRODUCTION AU SÉQUENÇAGE À HAUT DÉBIT POUR LA GÉNOMIQUE

Ezechiel B. TIBIRI

INSTITUT DE L'ENVIRONNEMENT ET DE RECHERCHES AGRICOLES – INERA

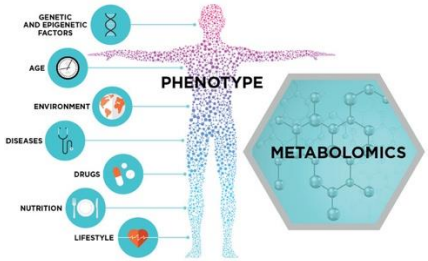
Adapted from: DUBii – C. THERMES

Rappels historiques



Https://fr.wikipedia.org/wiki/Pinsons_de_Darwin

De la théorie darwinienne (1859) à la métabolomique (exploration cellulaire en temps réel)



https://www.mtidx.com/our-technology/metabolomics

Structure De l'ADN 1953

1958 Dogme central

1957 FORTRAN

IBM 650

ENIAC 1945

Détermination d'une séquence protéique par ordinateur **1964**

Premiers arbres phylogénétiques **1967**

1965 Atlas of Protein Sequence and Structure

Margaret Oakley Dayhoff

Dayhoff the "mother and father of bioinformatics"

Algorithme pour l'alignement global de séquences **1970**

1972 Clonage

1969 UNIX et ARPANET

1971 Premier microprocesseur Intel 4004

prédiction de structures secondaires 1974

TPC/IP

Séquençages ADN 1977

1978 - 1980 EMBL, GenBank, PIR

Algo. alignement Local 1981

PCR 1984

1981 IBM-PC 8088

FASTA **1985**

1983 IBM-XT (10 Mb)

1990

1993

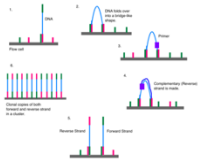
1988 NCBI

1989 INTERNET Python

11 génomes bactéries séq. 1997

NGS **2007**

2001 1erWGS humain



Miseq -Hiseq 2011

2015 MinION

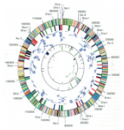
Médecine personnalisée >> milliers de génomes

2017

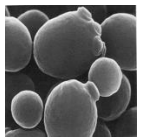
Premiers génomes entièrement séquencés

Médecine personnalisée >> milliers de génomes

2 bactéries



Levure



Ver

Plante

Mouche

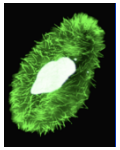
Humain

Souris

Poule

Chimpanzé

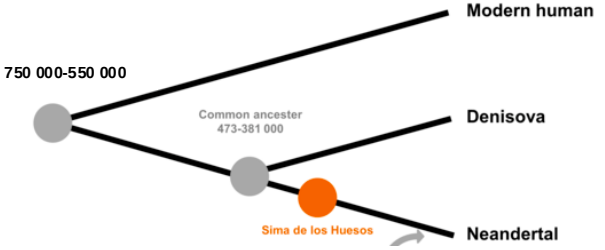
Paramécie



2 humains

1000 génomes humains

1990 91 92 93 94 95 96 97 98 99 2000 01 02 03 04 05 06 07 08 09 12 ... 20



1rst generation : Sanger sequencing

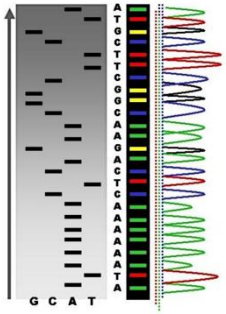
• Has been the major method up to 2005

Limitations

• Extremely high cost

• Long experimental set up times

• High DNA concentrations needed



2d generation

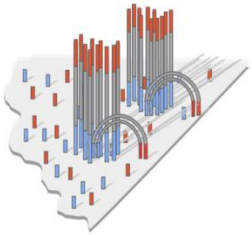
• Single DNA molecules replicated in clusters

• Very high troughput

Limitations

• Maximum read length ≤ 300bp

Illumina



3rd generation

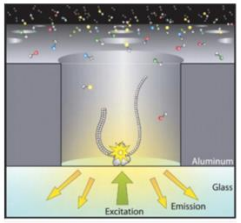
• Single molecules sequencing

• Very long reads

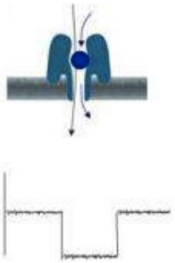
Limitations

• High error rates

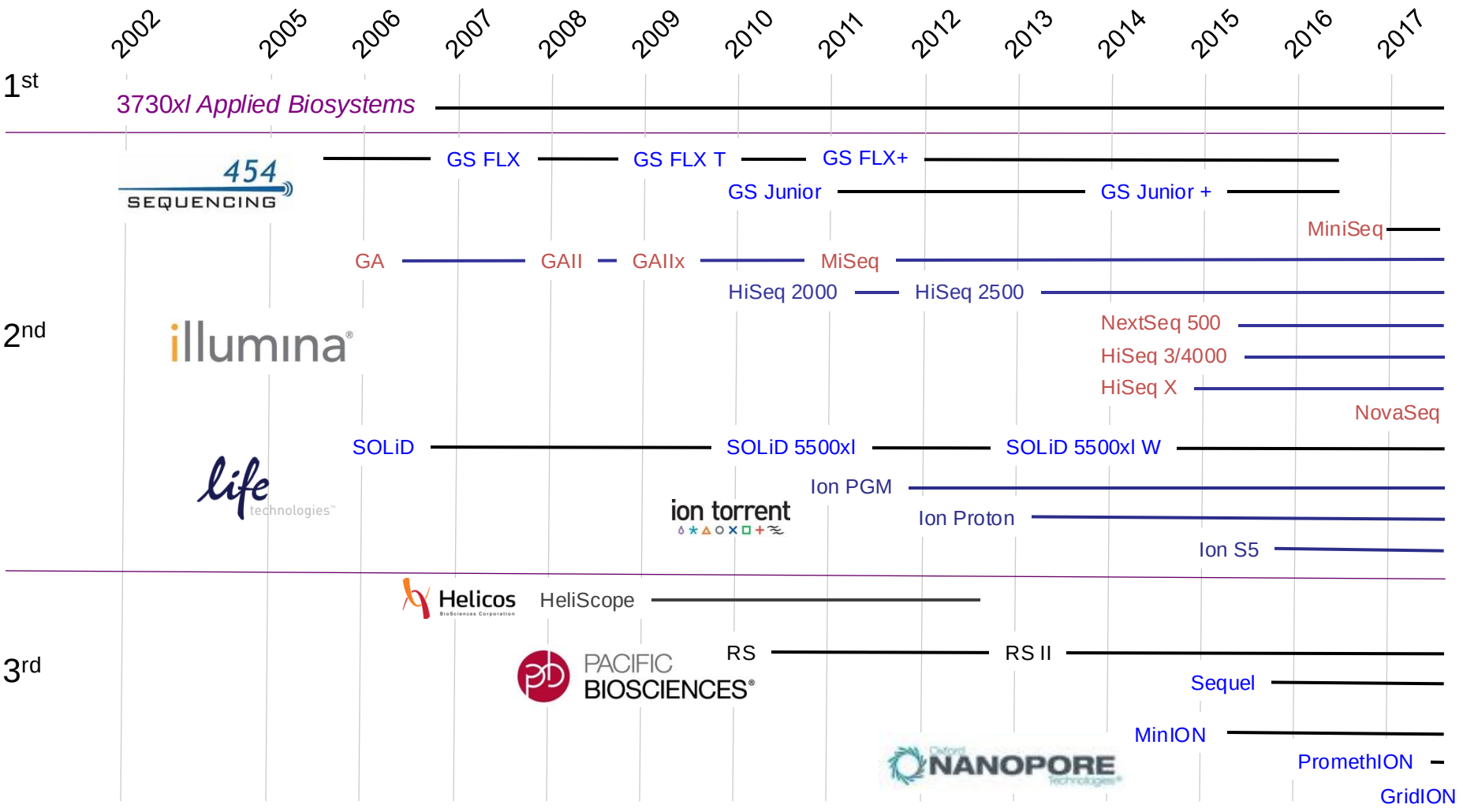
PacBio



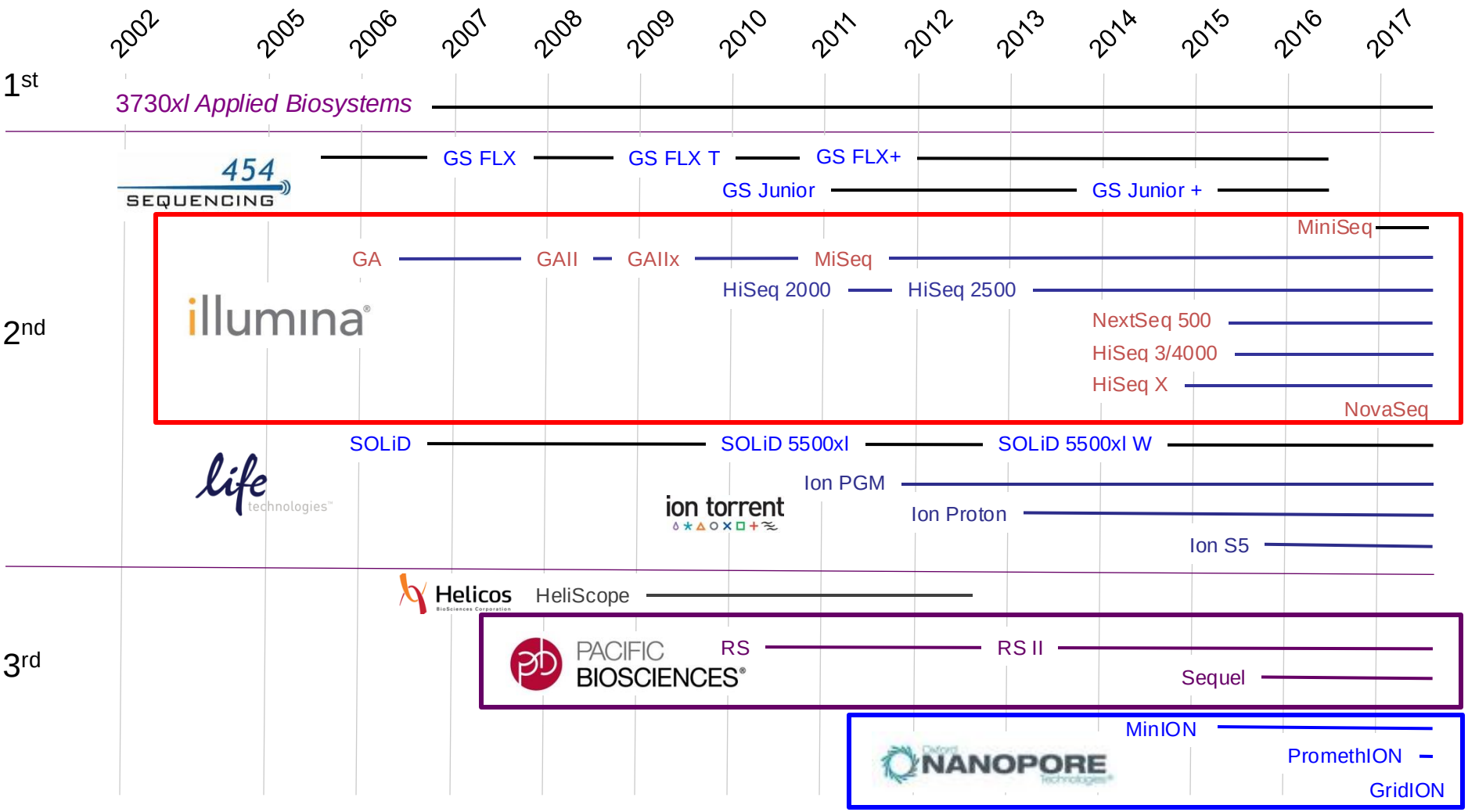
Oxford Nanopore



Sequencing technologies



Sequencing technologies



PART I

2^{d} GENERATION SEQUENCING

Illumina : the winning technology

MiniSeq
25 million reads

MiSeq
25 millions reads, 2 x 300 bp

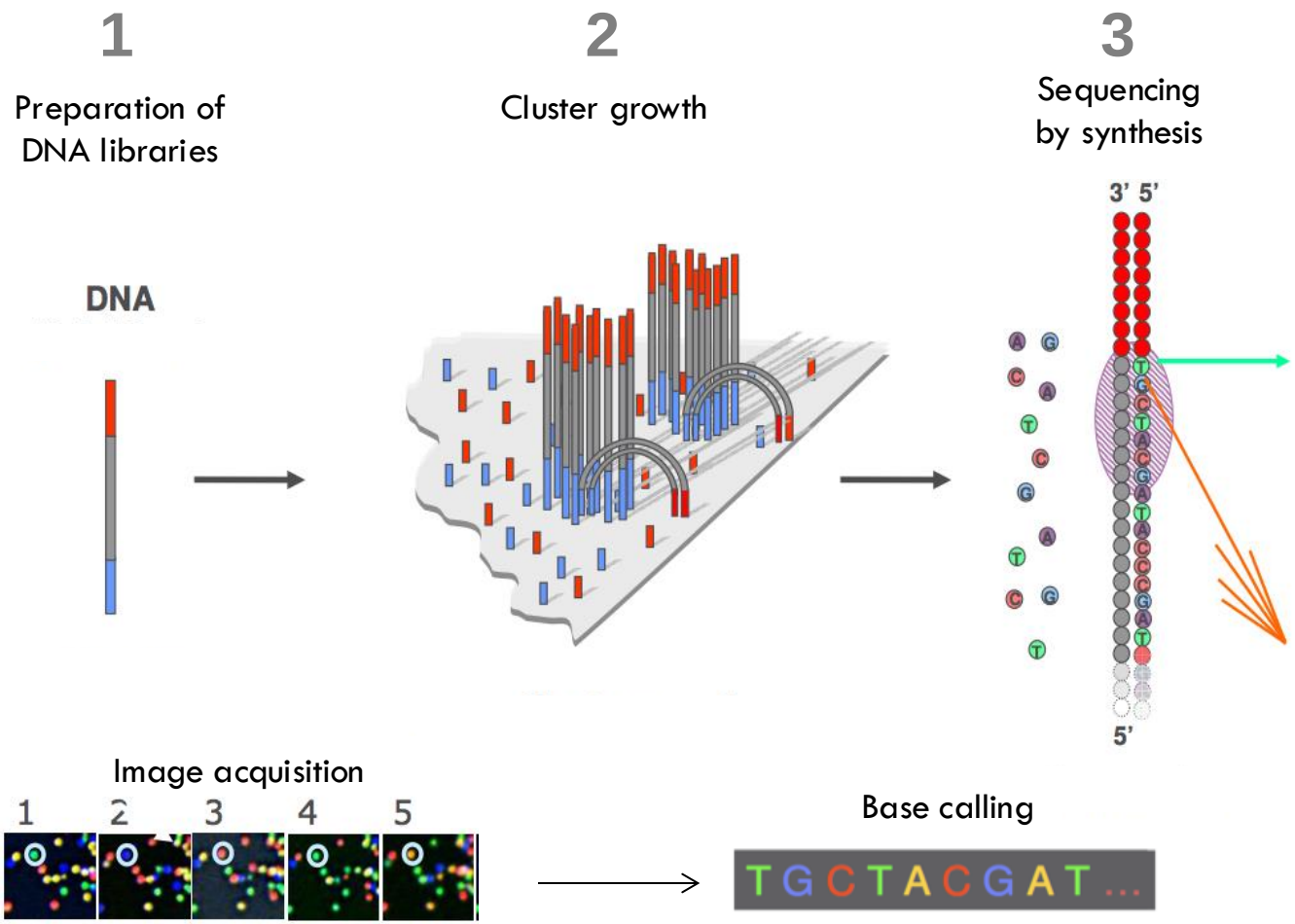
NextSeq
400 million reads

HisSeq 4000
5 billion reads

HisSeq X
6 billion reads

NovaSeq 6000
20 billion reads

General scheme of Illumina sequencing



1 - Preparation of DNA-seq Libraries

Illumina TruSeq technology

Genomic DNA

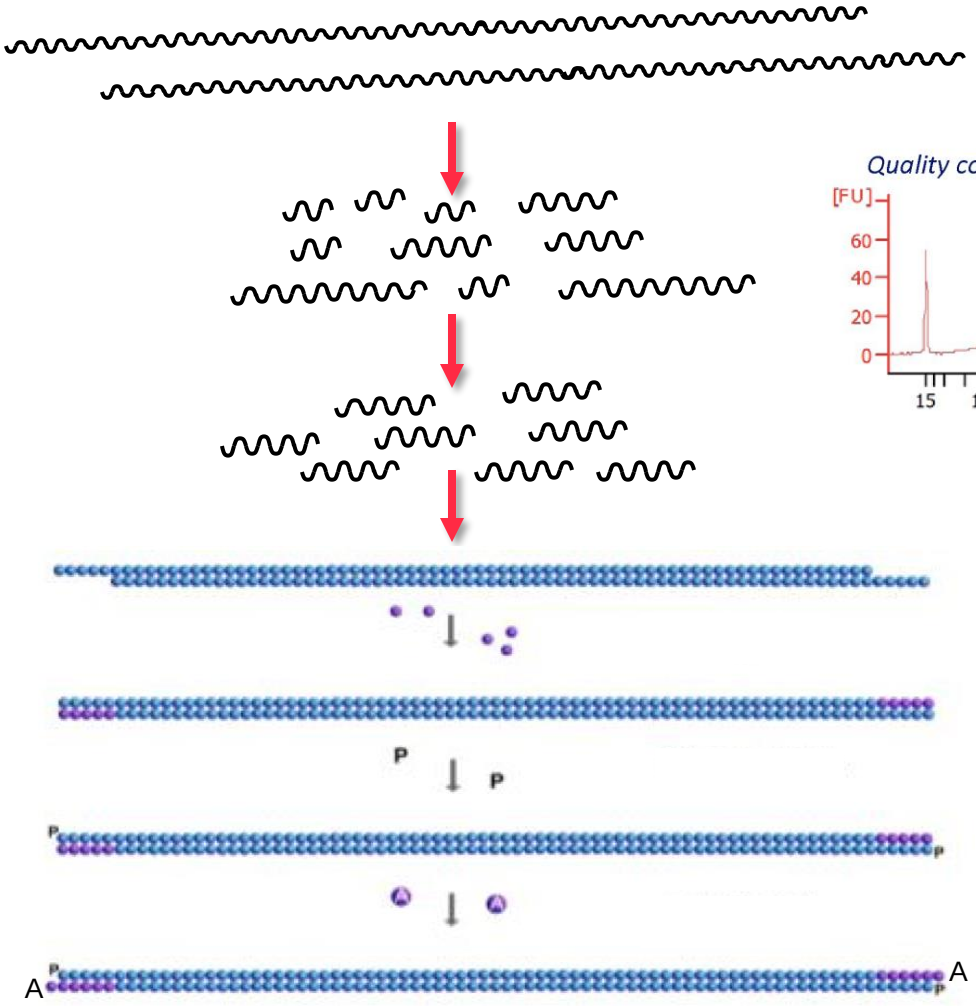
Sonication

Size selection

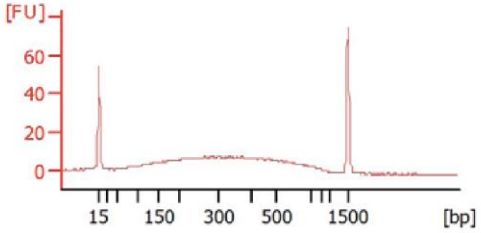
End repair

Phosphorylation

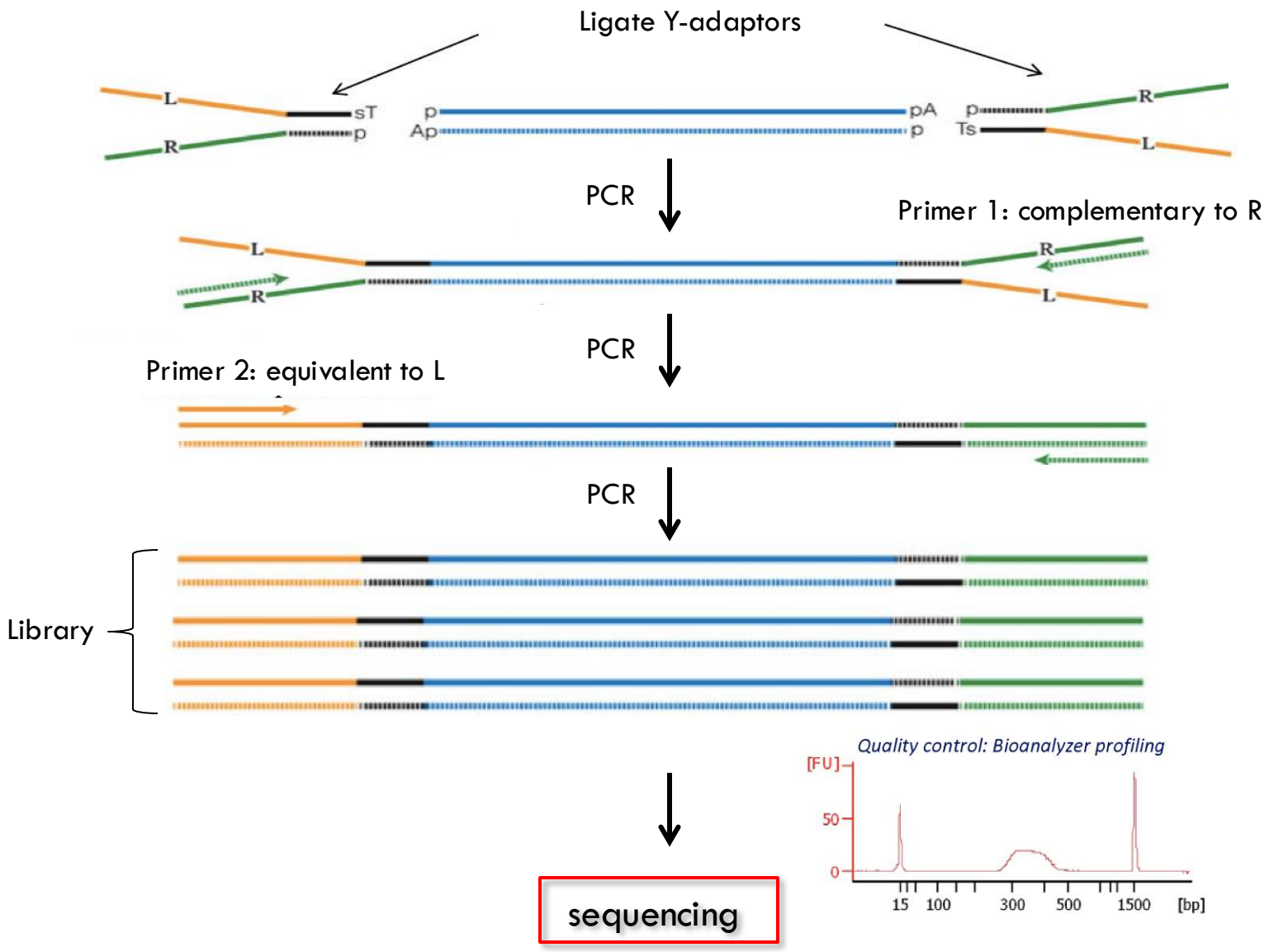
A - overhang



Quality control: Bioanalyzer profiling



1 - Preparation of DNA-seq Libraries



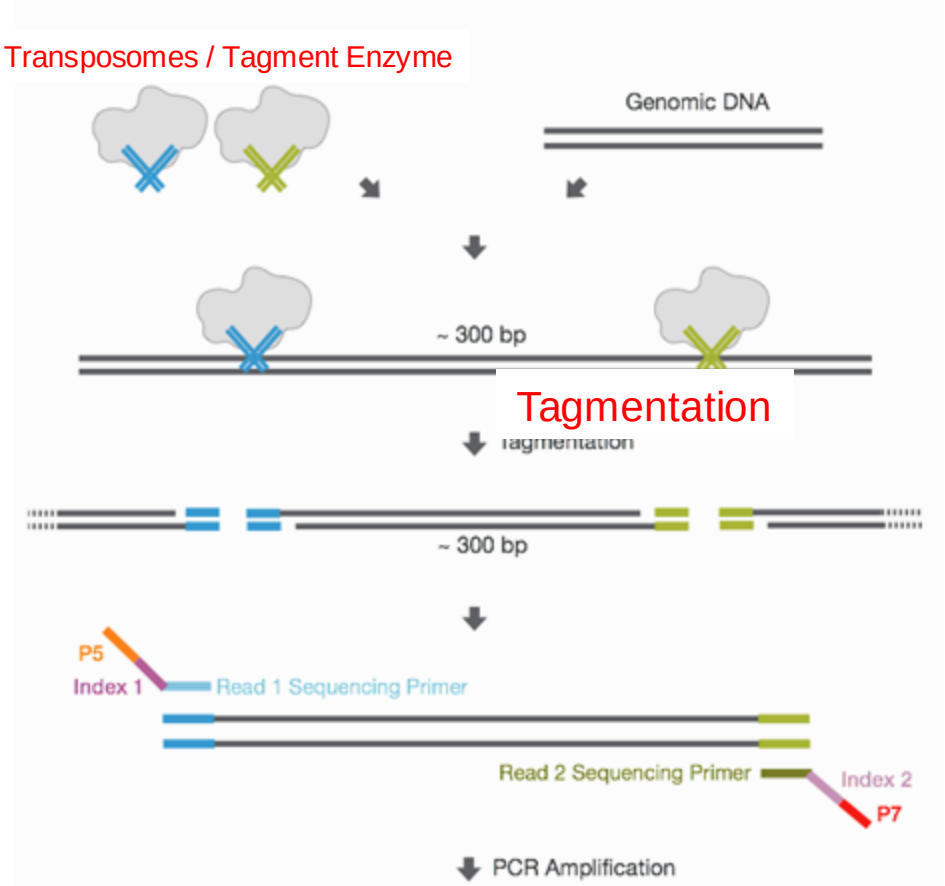
1 - Preparation of DNA-seq Libraries

Nextera "tagmentation"

Tagment Enzyme fragments DNA and attaches junction adapters (blue and green) to both ends of the tagmented molecule

Dual barcode approach

up to 96 indexed samples



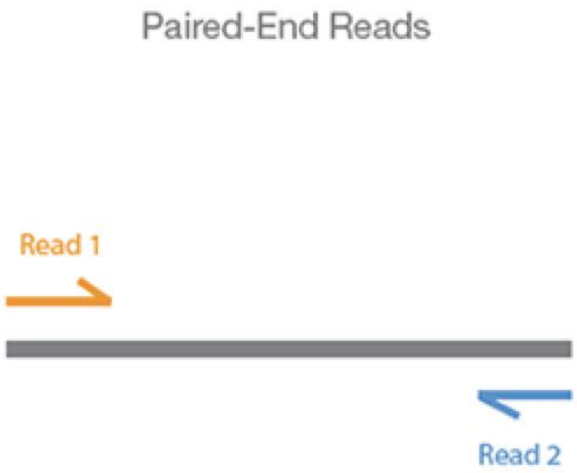
requires small quantities 1ng (bacteria) to 50 ng (human)

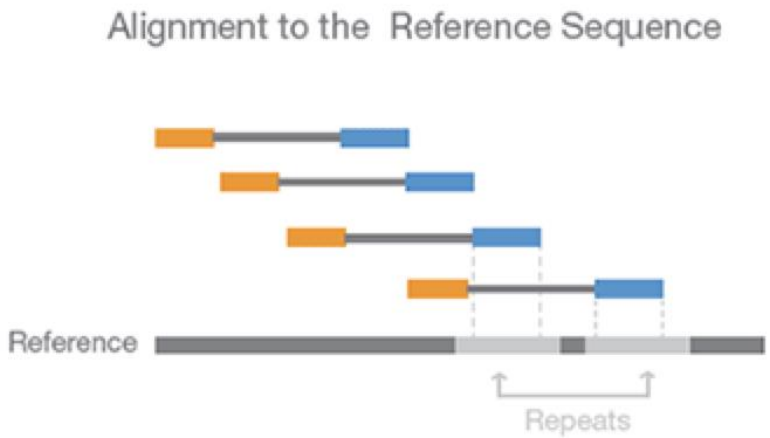
1 - Preparation of DNA-seq Libraries

SINGLE READ and PAIRED-END SEQUENCING

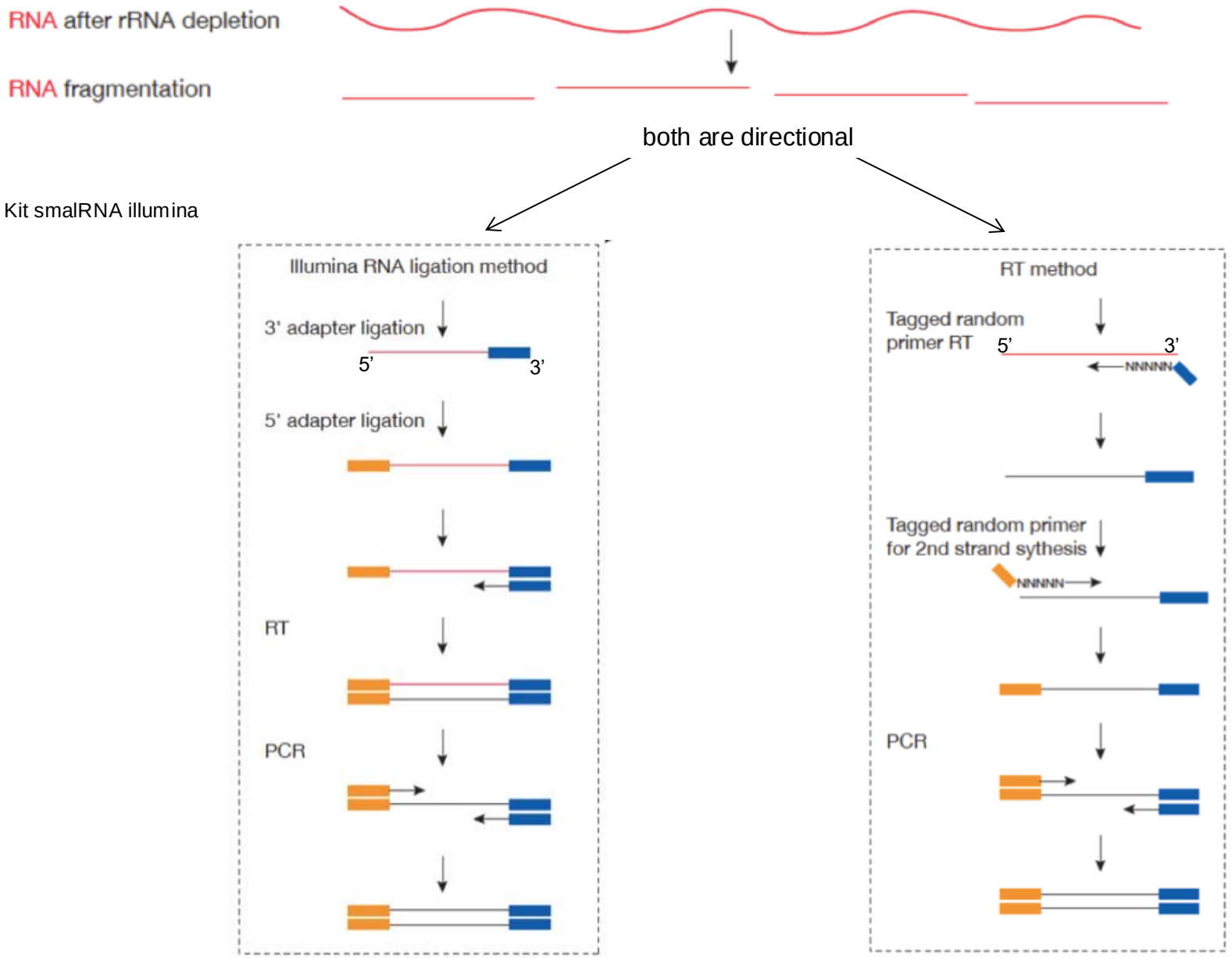
- **Single end**: Sequence one physical end of DNA fragment

- **Paired End:** Sequence both physical ends of DNA fragment
  - End distance: < 800nt





1 - Preparation of RNA-seq Libraries



Banques à partir d'ADN

Protocole	**Illumina Truseq** (nombreux kits disponibles dans le commerce, chez tous les fournisseurs de produits de bio mol) PRINCIPE : Ligation d'adaptateurs sur fragments d'ADN	**Nextera** (kit Nextera Illumina) PRINCIPE : ajout d'adpatateurs et fragmentation par Tagmentation
Matériel de départ	ADN génomique fragmenté (fragmentation enzymatique ou mécanique) Fragments d'ADN double brin (par ex ChIP) Toutes quantités de 1 à 5000 ng	ADN génomique, 50 ng (input fixe)
Avantages	• Adaptable à tout type d'ADN double brin • Bon contrôle de la taille finales (purif sur gel ou sur billes magnétiques) • Fonctionne également sans PCR si la quantité de matériel suffisante (>100ng) • Fonctionne bien même avec des petites quantités	Très rapide (3h)
inconvénients	• Protocole long si fragmentation mécanique • Possibilité de formation de dimères d'adaptateurs (perte de reads, perte de qualité) ; il existe des kits qui permettent de réduire la formation de dimères	• Très sensible à qualité de l'ADN de départ (intégrité, pureté) • Contrôle de la taille des fragments obtenus parfois difficile • PCR obligatoire • Extrémités des fragments de l'ADN de départ sont perdues (ex : génome de phage)

Banques à partir d'ARN

	TruSeq small RNA (Kit Illumina ou autres fournisseurs (NEB))	**TruSeq Stranded RNA** (Kit Illumina, autres fournisseurs possibles)	**SMART-Seq V4 (Takara)**
Protocole			
Matériel de départ	Petits ARNs pré-purifiés Peut être utilisé sur de l'ARN fragmenté	100-2000 ng ARN total	ARN total (selection oligodT) 10 pg à 1ng ARN NON DEGRADÉ
Principe	Ligation directe d'adaptateurs sur ARN Suivi de Reverse transcription et PCR Directionnalité conservée	• Sélection des messagers par polyA ou déplétion des ARNs ribosomiques par méthode RiboZero (au choix) • RT par random priming pour obtention d'un cDNA • Ajout des adaptateurs par ligation • Enrichissement PCR Directionnalité conservée	Synthèse du cDNA à partir de la queue polyA (oligo dT) puis amplification des messagers. Construction des banques à partir du cDNA obtenu (Nextera ou Truseq) Directionnalité non conservée
Avantages		• Petites quantités possibles • Possibilité d'utiliser de l'ARN dégradé avec l'option RiboZero	• RNA-seq possible même si très petites quantités d'ARN total. • cDNA longs
inconvénients	• Long : 2-3 jours de manip • Peu adapté aux messagers car faible rendement	• Sensible à contamination par ADN génomique	• ARN non dégradé seulement • Coûteux
Remarques	N'est plus utilisé à la PF pour le RNAseq classique		Peut être utilisé pour d'autres applications, comme la synthèse de cDNA longs pour séquençage nanopore


Many applications
FAIRE-SEQ
SHAPE-SEQ
MDA
RRBS-SEQ
RIBO-SEQ
ICE
OS-SEQ
CLASH-SEQ
TIF-SEQ/PEAT
IN-SEQ
TAB-SEQ
CHIP-SEQ
5-C
SMMIP
OXBS-SEQ
RIP-SEQ
4-C
TC-SEQ
NET-SEQ
UMI
CAP-SEQ
DNASE-SEQ
FAIRE-SEQ
DUPLEX-SEQ
PAR-CLIP-SEQ
BS-SEQ
MEDIP-SEQ
GRO-SEQ
MIRP-SEQ
DIGITAL
CHIA-PET
ATAC-SEQ
CIP-TAP
ICLIP
MBDCAP-SEQ
HITS-CLIP
PARE-SEQ/GMUT
HI-C/3-C
TRAP-SEQ
RC-SEQ
FRAG-SEQ

SINGLE CELL TECHNOLOGIES

Single cell transcriptomics allows to study transcriptome heterogeneity,

to investigate differences in transcript expression and gene regulation ***in individual cells*** :

- ❖ Differences in trancript abundance
- ❖ Alternative splicing and differential expression of isoforms

Most widely used device to study single-cell transcriptomics : ***Chromium controller (10x Genomics)***

Several applications :

Single cell Gene expression
Measures gene activity on a cell-by-cell basis, characterize cell populations, cell types, …

Linked read genomics
Performs diploid de novo assembly, phase haplotypes, genetic variations

Single cell ATAC
Measures epigenetics by detecting open chromatin regions

PART 2

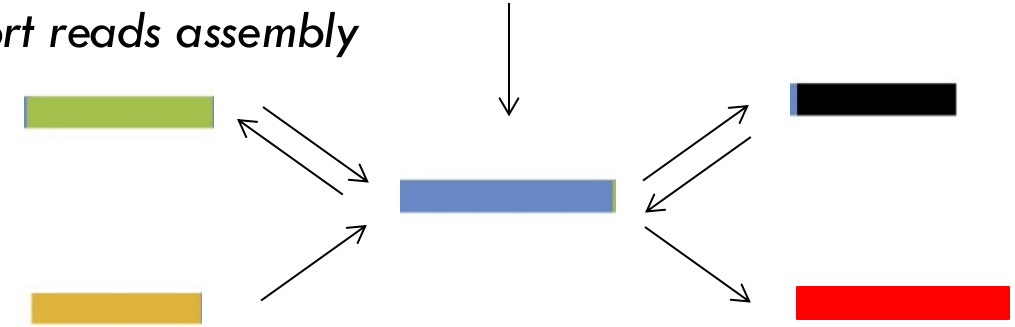
3rd GENERATION SEQUENCING

LONG READS

LONG-READS VERSUS SHORT-READS

Assembly of DNA fragments with repeated sequences

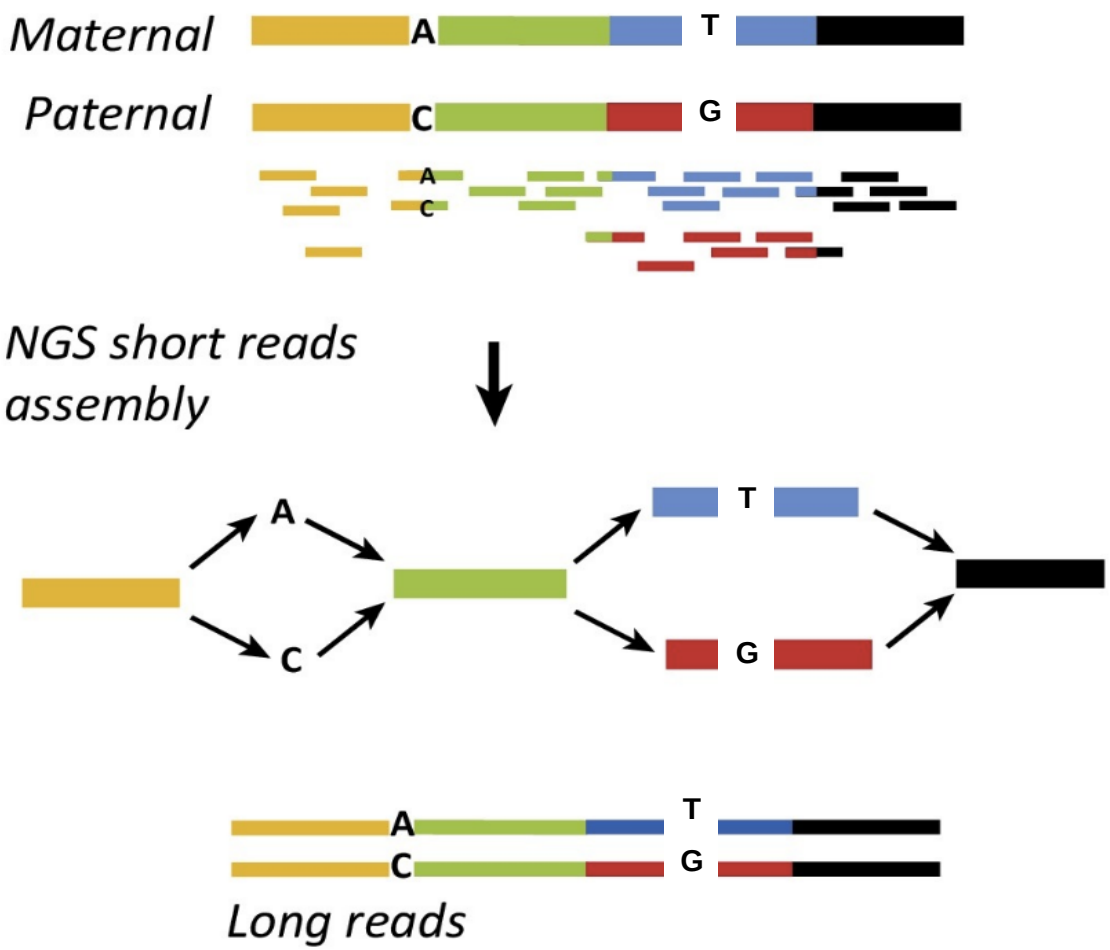
NGS short reads assembly



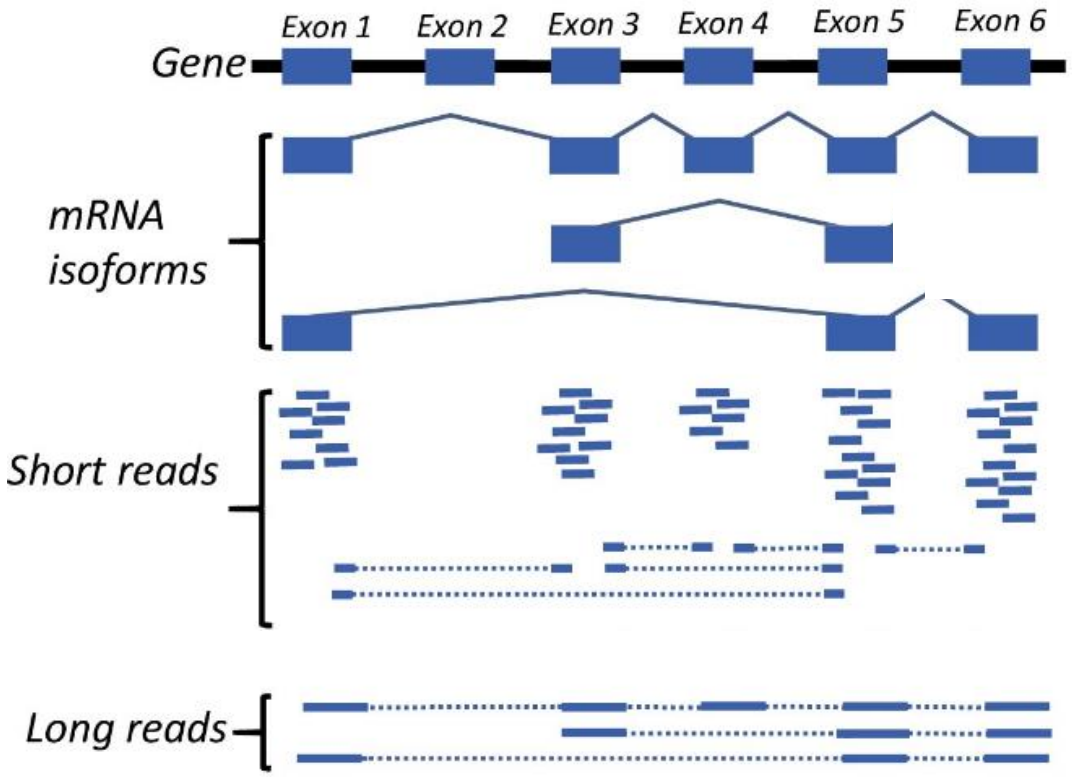
Several contigs → incomplete assembly, underestimation of repeats

Long reads assembly

Haplotype phasing



Detection of splicing isoforms



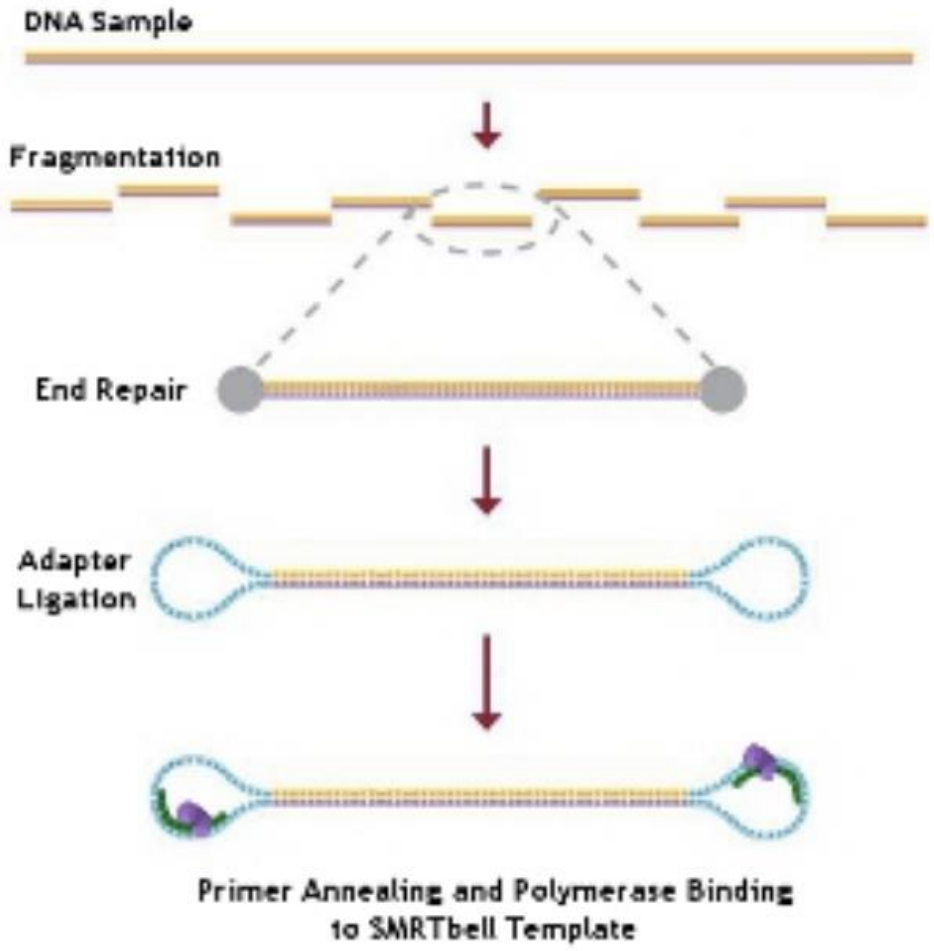
The 3rd generation winning technologies

Sequel - Pacific Biosciences
Single molecules
Up to 80,000 bp long
Error rate ≈ 10-15 % - CCS: <1%
Compensated by coverage

MinION - Oxford Nanopore
Single molecules
> 200 000 bp long
Error rate ≈ 10-15 %
Compensated by coverage

PacBio : Single Molecule Real Time (SMRT) sequencing

PacBio DNA-seq library

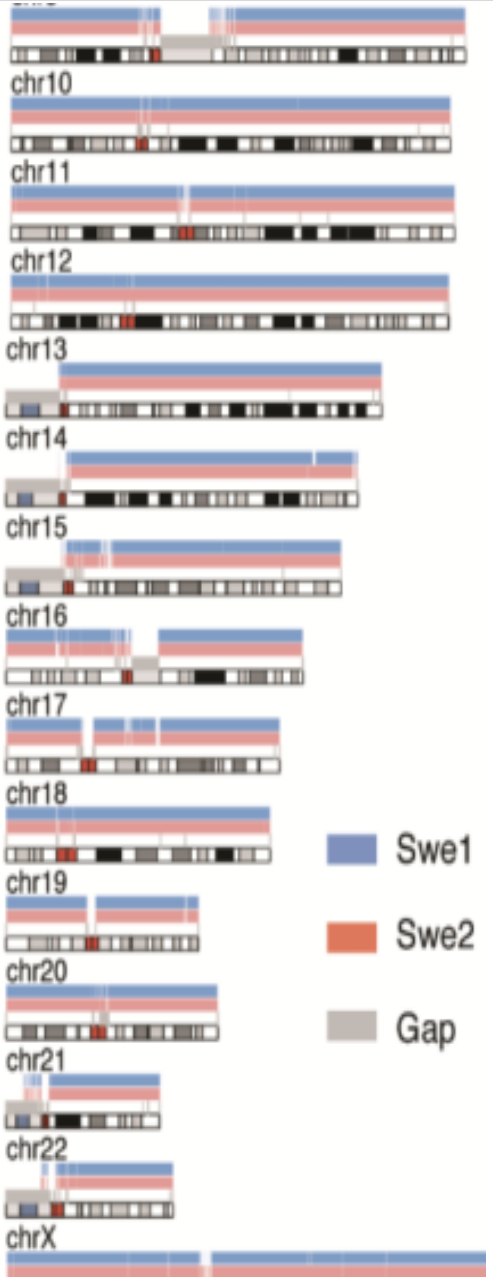


PacBio GENOME ASSEMBLY

De novo Assembly of Two Swedish Genomes Reveals Missing Segments from the Human Reference (hg38)

Ameur et al. *Genes*, 2018

- 10 Mb of the 2 genomes are absent from hg38 reference
- 1 Mb are assigned to chr. Y
- 6 Mb are shared with a Chinese personal genome
- Inclusion of these sequences in GRCh38 genome radically improves alignment and variant calling from short-read data :
- re-analysis :
  - yields > 75,000 putative novel single nucleotide variants (SNVs)
  - removes > 10,000 false positive SNV calls per individual
- It becomes possible to represent specific population groups by assembly of representative genomes from different populations.



Next Generation Sequencing

Oxford NANOPORE Technologies®

Next Generation Sequencing

Oxford NANOPORE Technologies®

Nanopore sequencing platforms

Next Generation Sequencing

Oxford NANOPORE Technologies®

Flongle Adapter

MinION Mk1B

Low cost flow cells
Delivering up to 2.8 Gbases of data

Next Generation Sequencing

Oxford NANOPORE Technologies®

MinION Mk1D

MinION

Read length

Nanopore sequencing reads the entire length of DNA or RNA fragment included in the libraries from short to ultralong (longest >4Mb)

Dimensions

- Size: W 125mm H 13 mm D 55mm

Temperature control

New for Mk1D Peltier based temperature control enables reliable and robust sequencing at ambient temperatures between 10-35° C

Suitable applications

MinION Flow Cells provide output up to 48Gb* making them ideal for small genome and targeted application sequencing such as:

- Bacterial metagenomics
- Bacterial isolate
- Bacterial whole genome
- Viral amplicon
- Full length 16s rRNA
- Whole exome
- Whole transcriptome (cDNA)
- Smaller transcriptomes (direct RNA)

* Theoretical maximum output when system is run for 72 hours at 400 bases/second outputs may vary according to library type and run conditions.

Connectivity

Weighs 130 g and plugs into a laptop or MacBook® using a high-speed USB-C cable

Cost-effective

MinION Mk1D pack from $4,950 including sequencing consumables and 5 MinION Flow Cells

- $990 per Flow Cell
- Multiplexing kits for higher sample throughput

Next Generation Sequencing

Oxford NANOPORE Technologies®

GridION

Read length

Nanopores read the length of DNA or RNA presented to them — from short to ultra-long (longest >4 Mb)

Dimensions

Compact benchtop device with integrated compute suitable for any lab

- Size: W 370 mm, H 220 mm, D 365 mm
- Weight: 11 kg

240 Gb output*

Up to 48 Gb per MinION Flow Cell / 2.8 Gb per Flongle Flow Cell*

Suitable applications

- Bacterial metagenomics
- Bacterial isolate
- Bacterial whole genome
- Viral amplicon
- Full length 16s rRNA
- Whole exome
- Whole transcriptome (cDNA)
- smaller transcriptomes (direct RNA)

- Theoretical max output when system is run for 72 hours (or 16 hours for Flongle) at 400 bases / second. Outputs may vary according to library type, run conditions, etc.

Next Generation Sequencing

Oxford NANOPORE Technologies®

PromethION

On-demand sequencing

Run up to 24 independently addressable, high-capacity PromethION Flow Cells.

* Theoretical max output when system is run for 72 hours at 400 bases / second. Outputs may vary according to library type, run conditions, etc.

Ultra-high throughput

Generate terabases of data — streamed in real time for immediate analysis.

Powerful compute

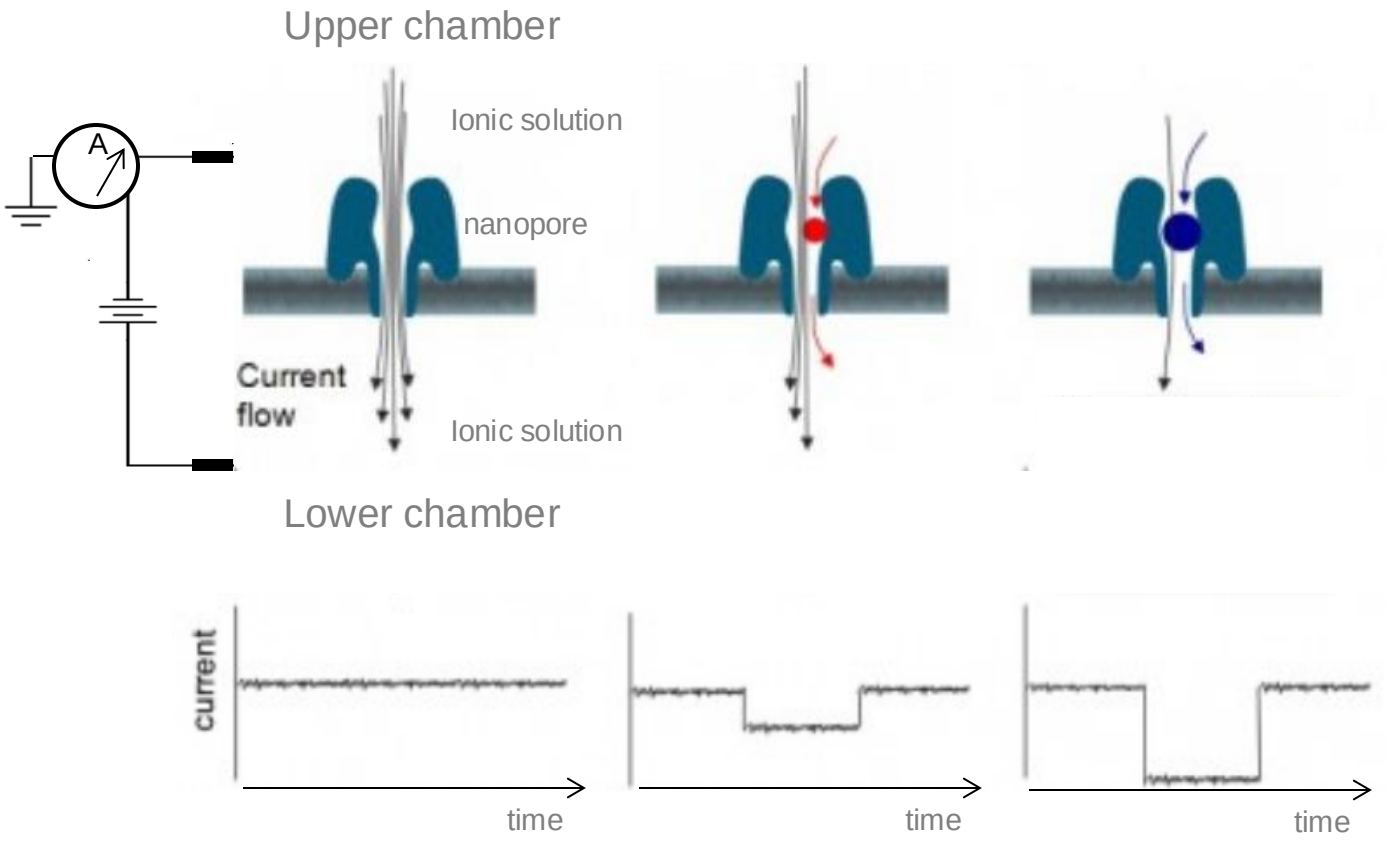
Alleviate data analysis bottlenecks.

Population-scale sequencing

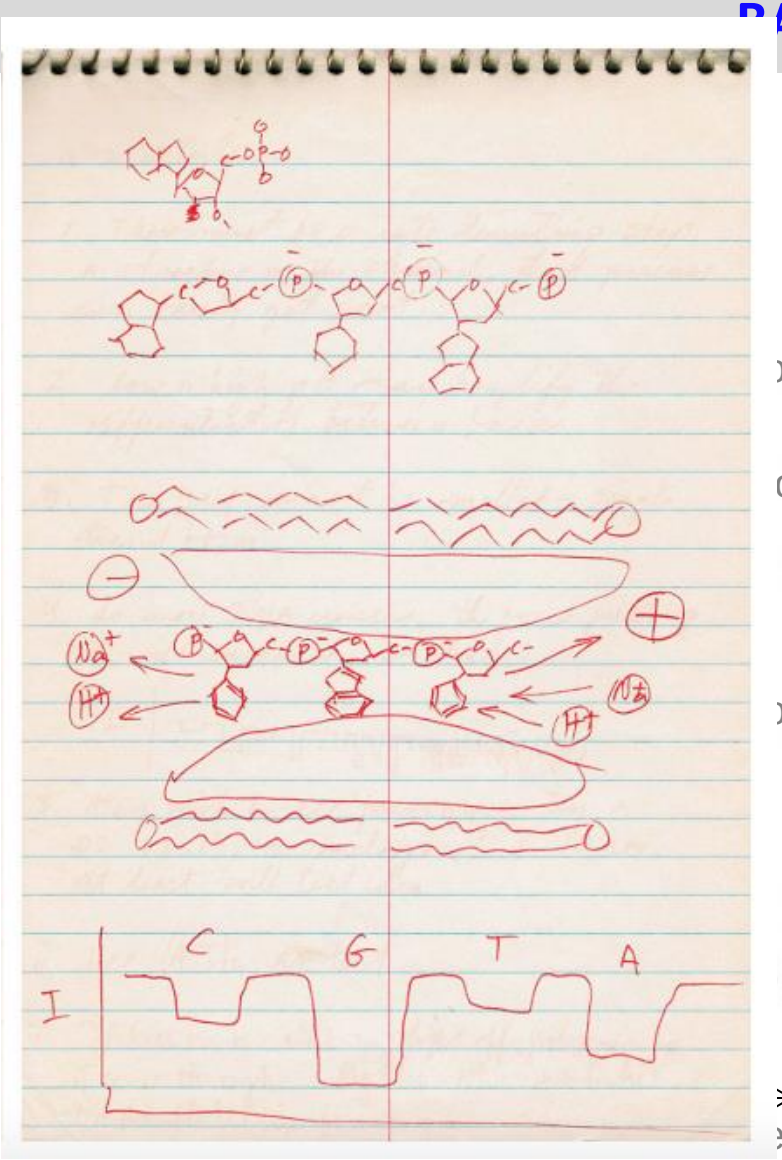
Generate human genomes from as little as $720/genome for all consumables and instrumentation. Excludes third party reagents

PromethION 2 Solo & 2 Integrated > PromethION 24 >

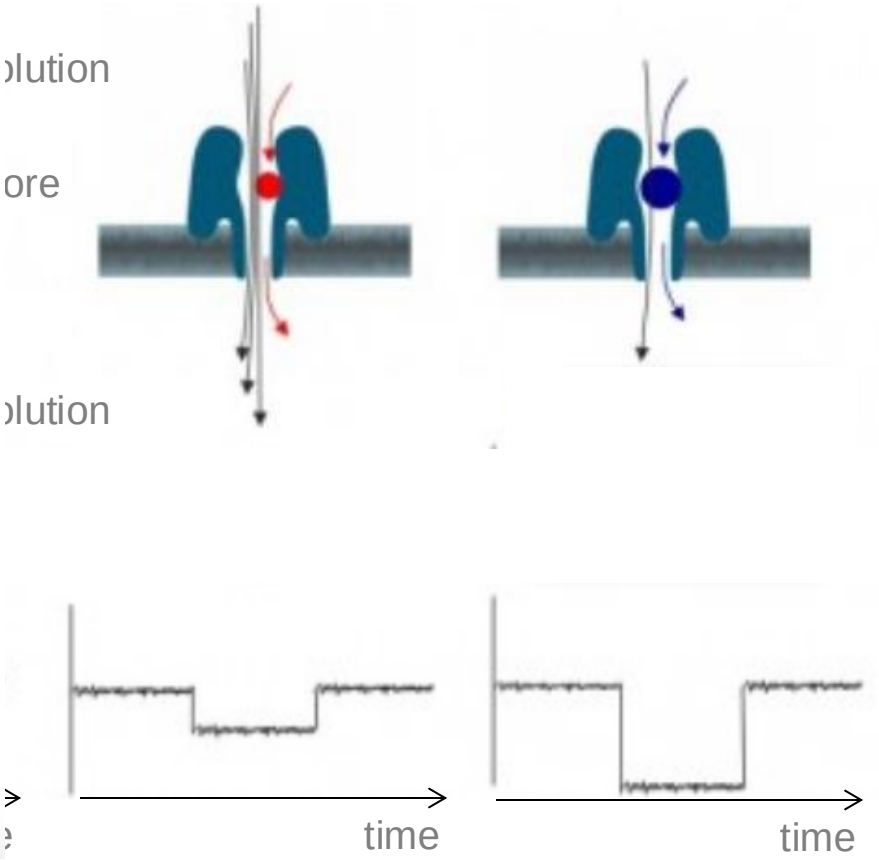
BASIC CONCEPTS



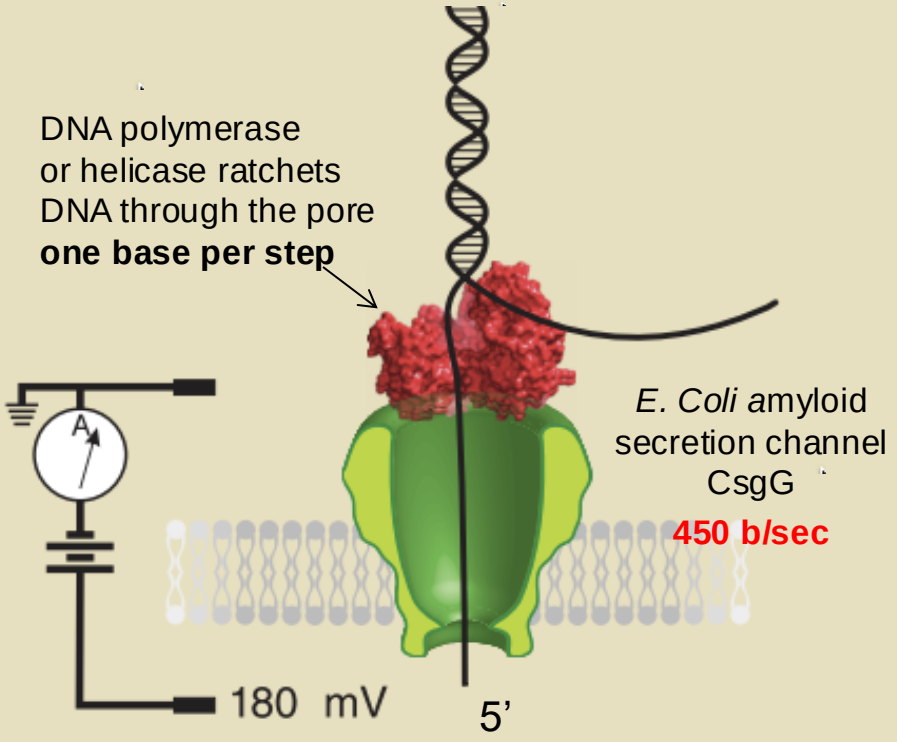
BASIC CONCEPTS

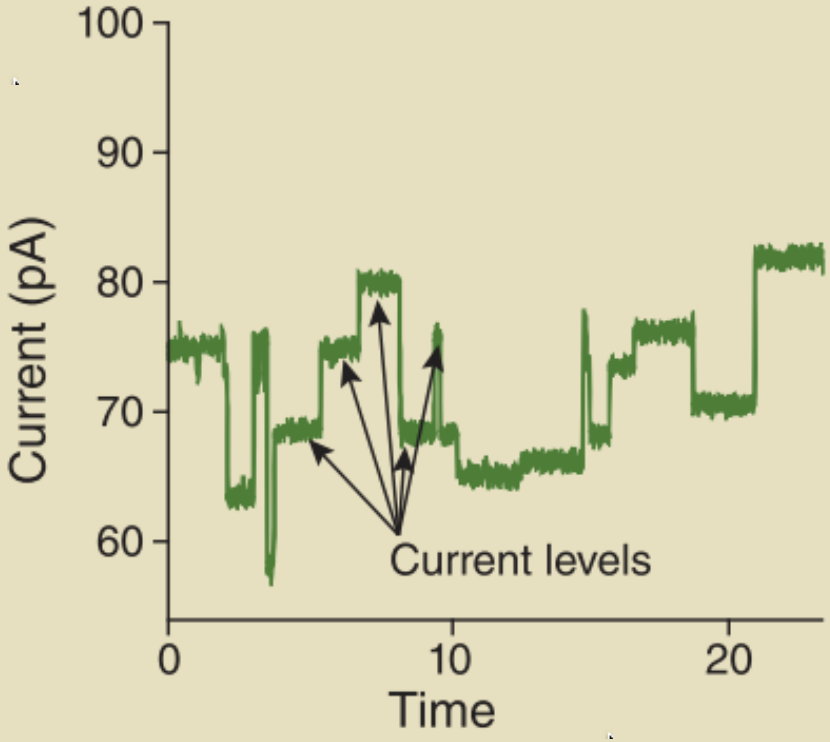


1989 David Dreamer's notebook



BASIC CONCEPTS

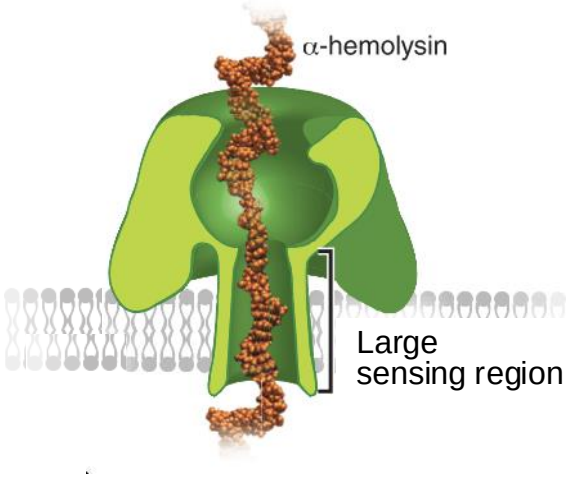




NANOPORES USED IN THE MinION

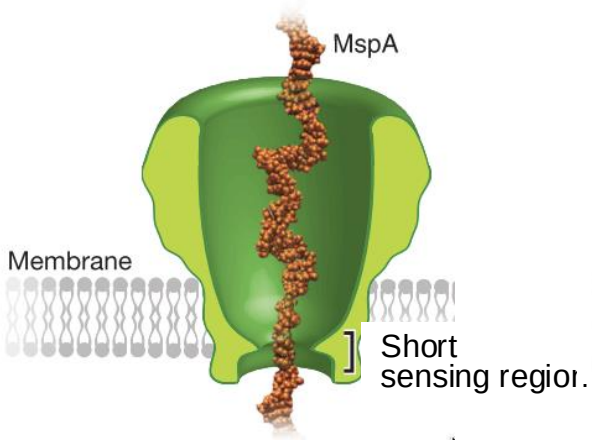
1996

alpha-hemolysin from *Staphylococcus aureus*



2012

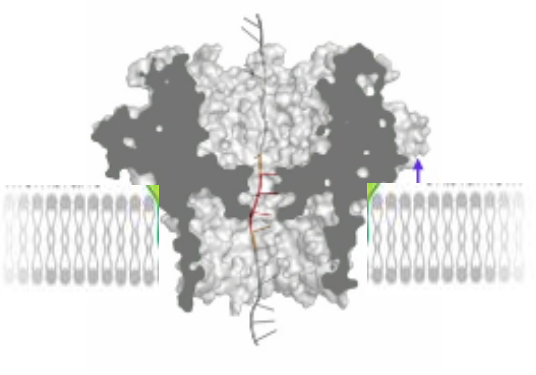
Porin from *Mycobacterium smegmatis*



2017

E. Coli amyloid secretion channel CsgG

450 b/sec



MinION available

DNA transport through alpha-hemolysin

Single nucleobase discrimination in a DNA strand

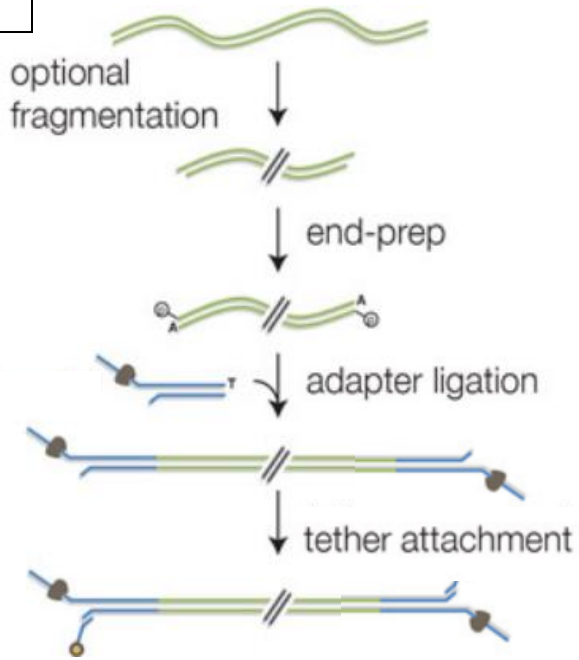
DNA sequencing with phi29 DNA pol. and MspA pore

E. coli assembled *de novo*

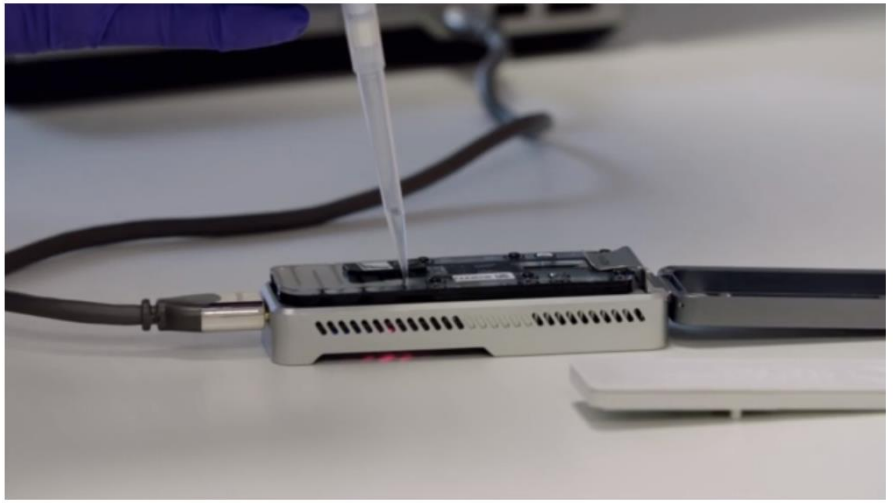
1995 1996 1999 2001 2005 2008 2012 2014 2015

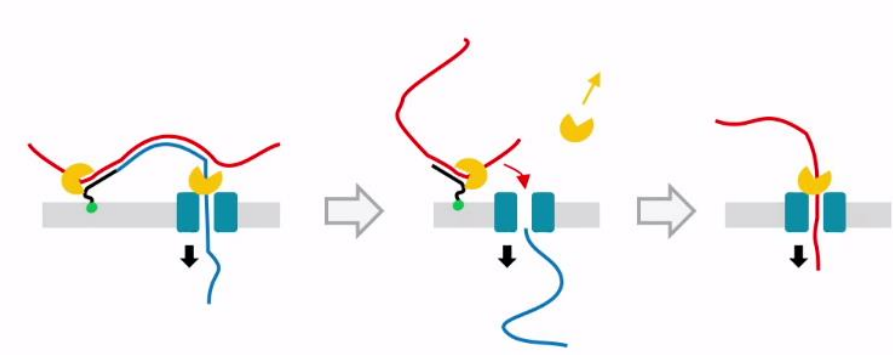
SEQUENCING PROCESS

1D^2 Library
(2017)

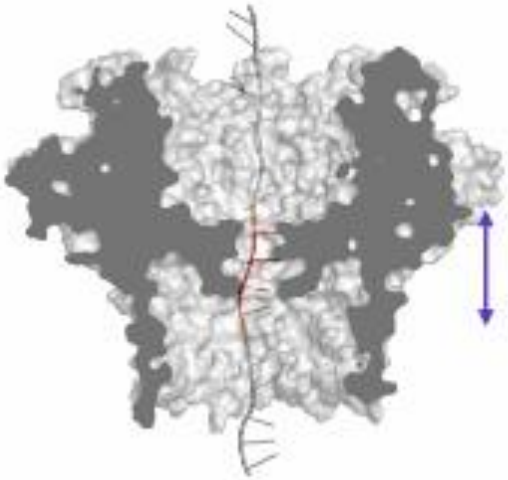


SEQUENCING



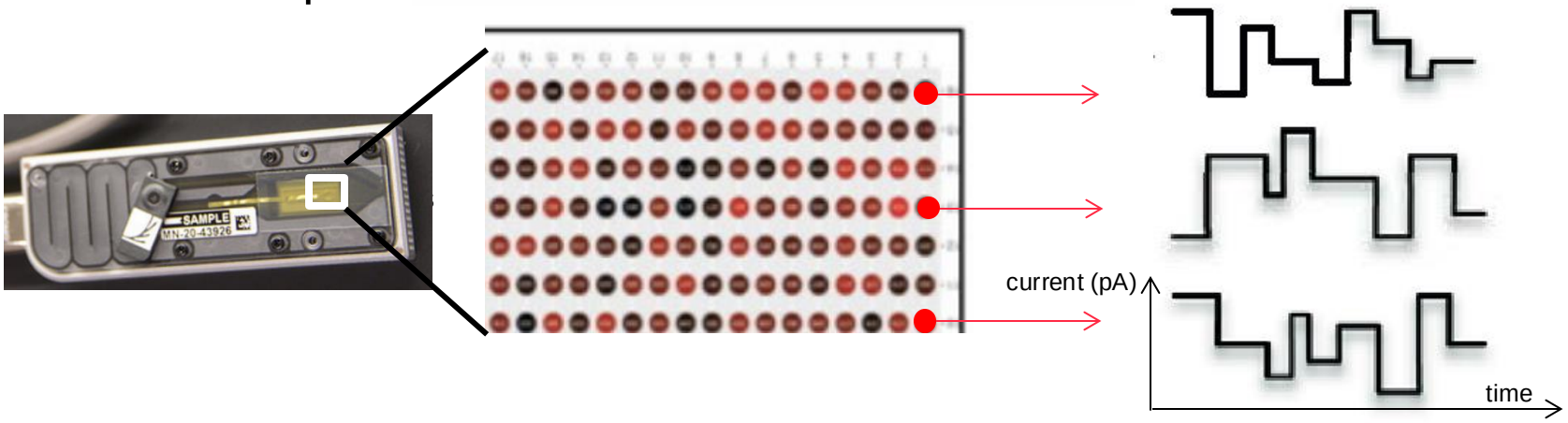


SEQUENCING PROCESS : MinION FLOW CELL



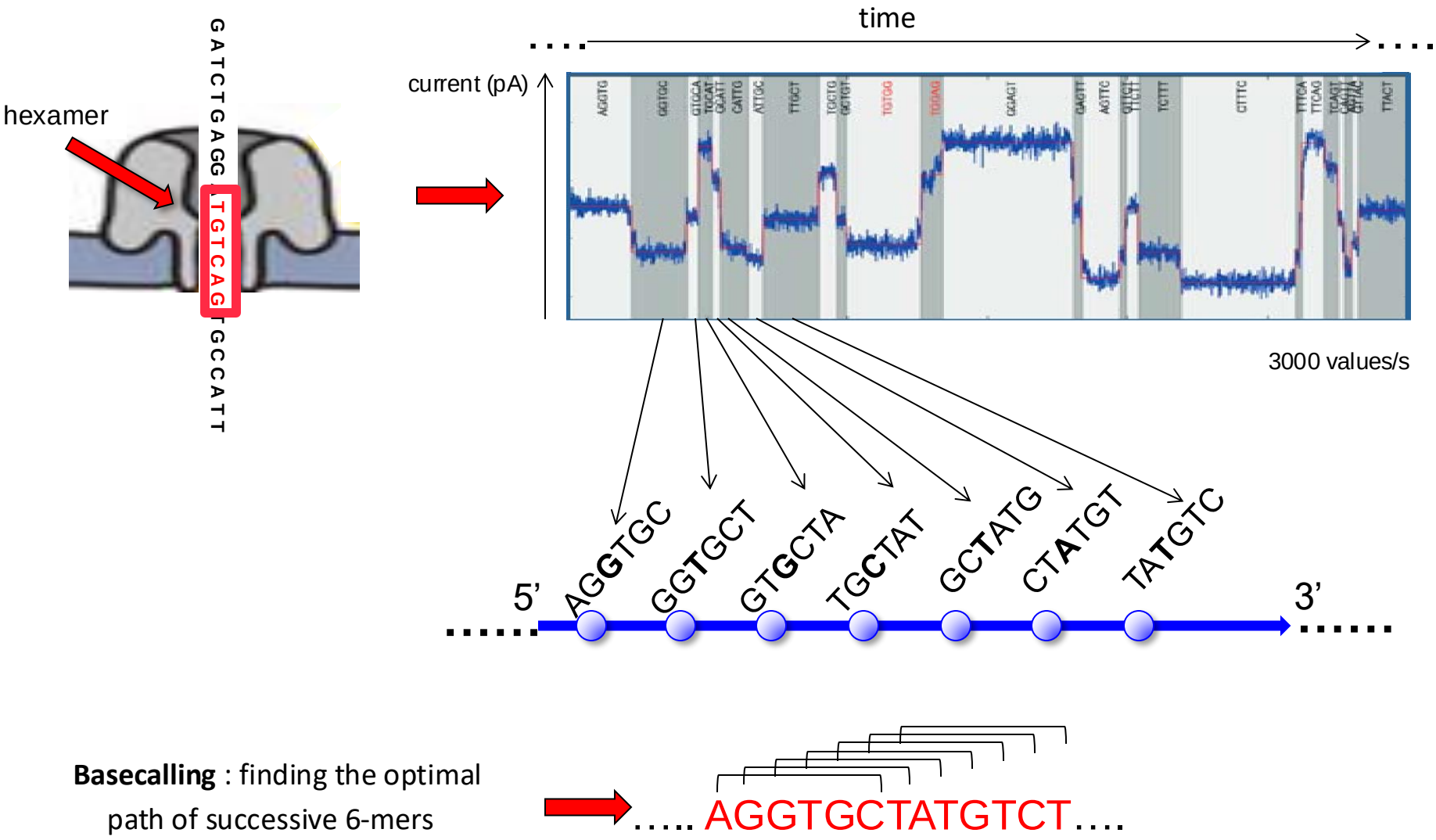
5-6 bases dominate the current signal

MinION : 512 pores



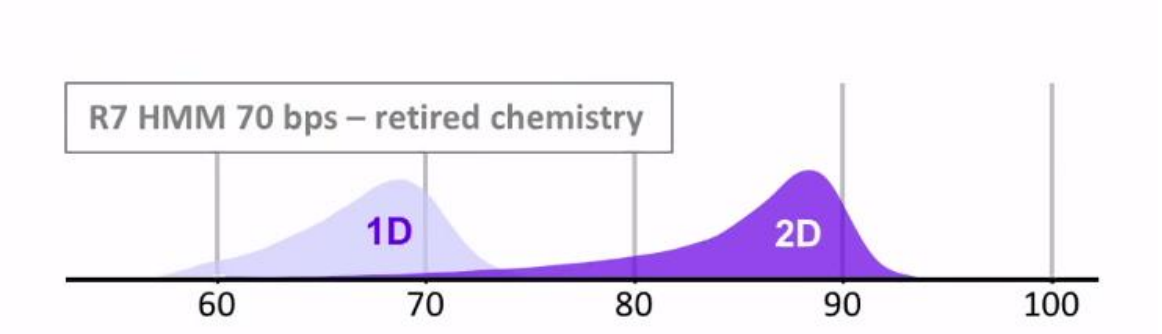
PromethION : 144000 pores (48 x 3000)

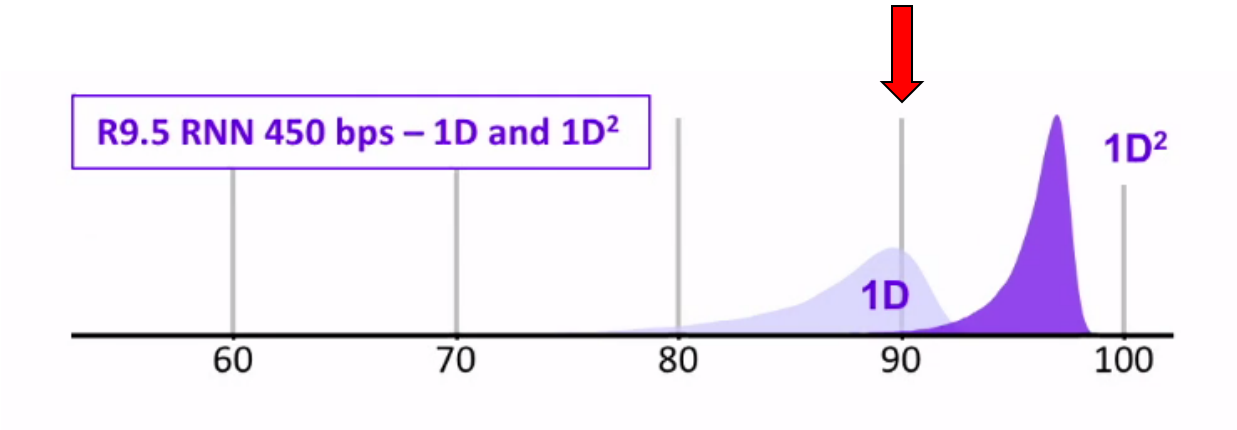
BASECALLING

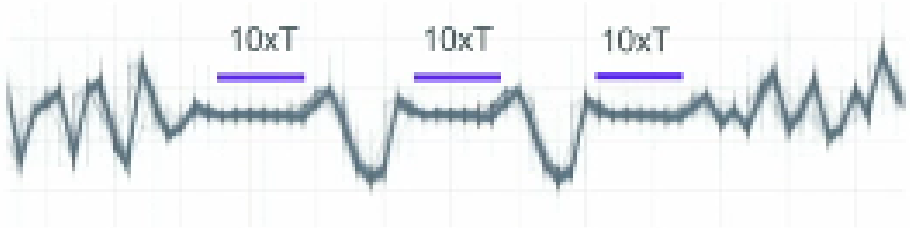


Basecalling : finding the optimal path of successive 6-mers

….. AGGTGCTATGTCT….



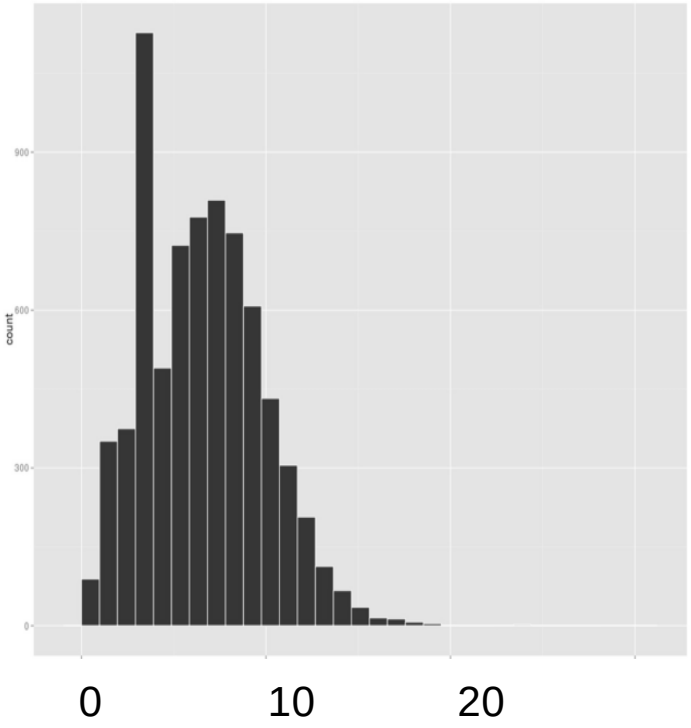




Homopolymers difficult to sequence

SIZE OF SEQUENCED DNA FRAGMENTS

Typical profile of fragment size

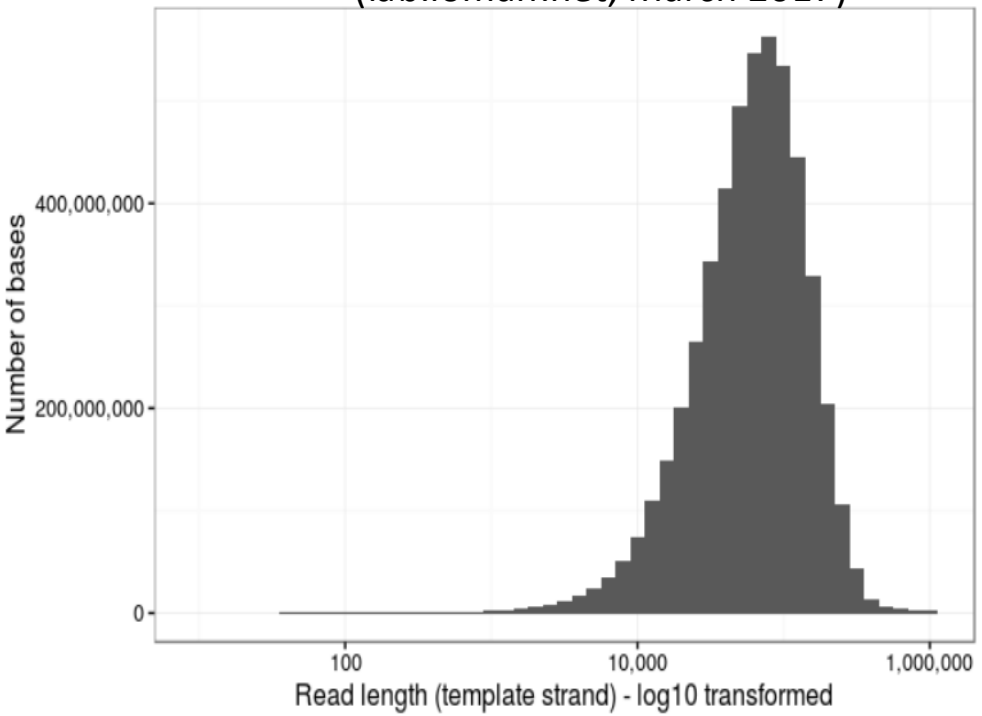


Length of sequenced fragments (kb)

(Risse et al. *GigaScience,* 2015)

"Ultra long" reads

(lab.loman.net, March 2017)

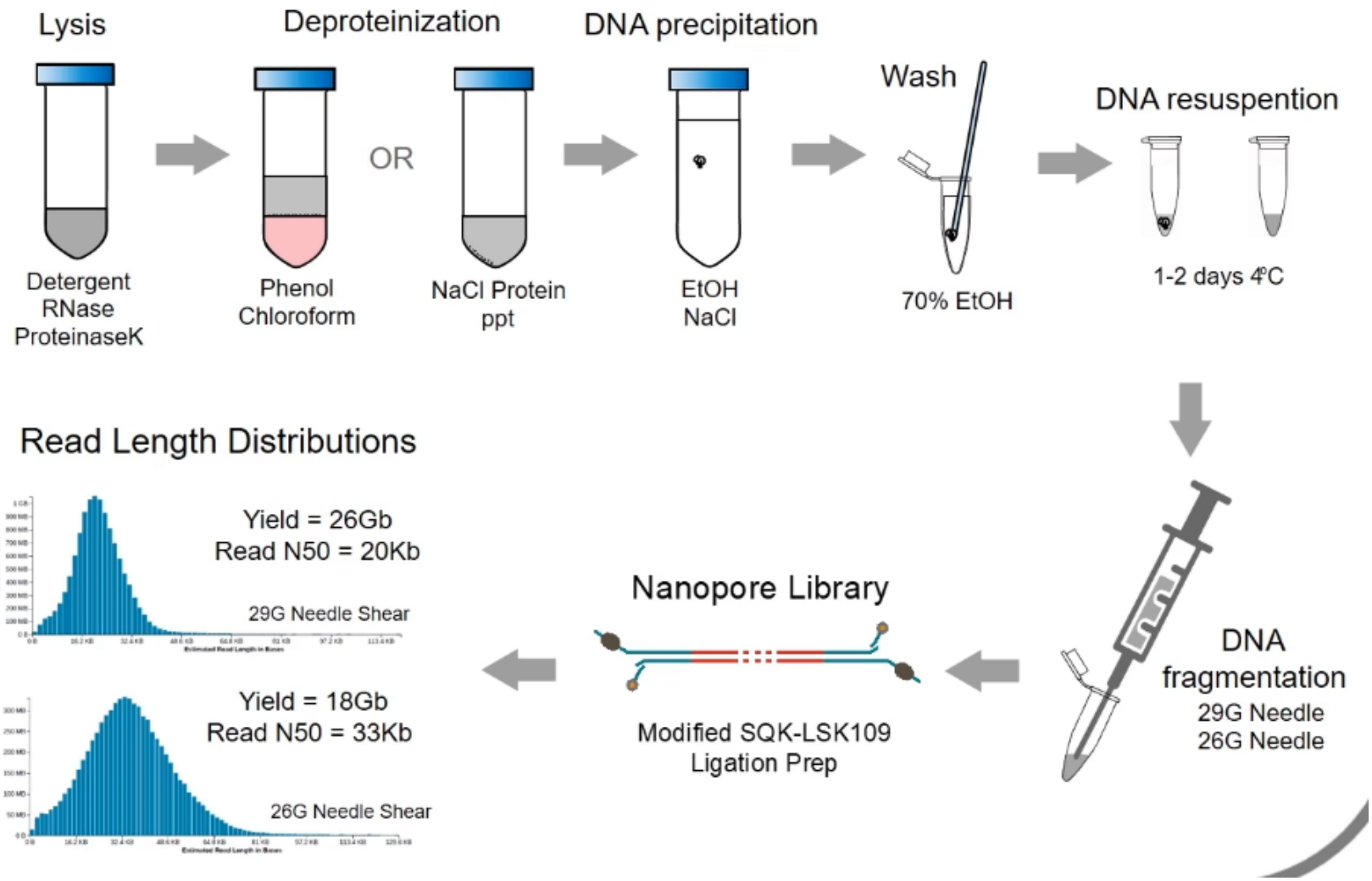


Size of the longest read : 778 kb

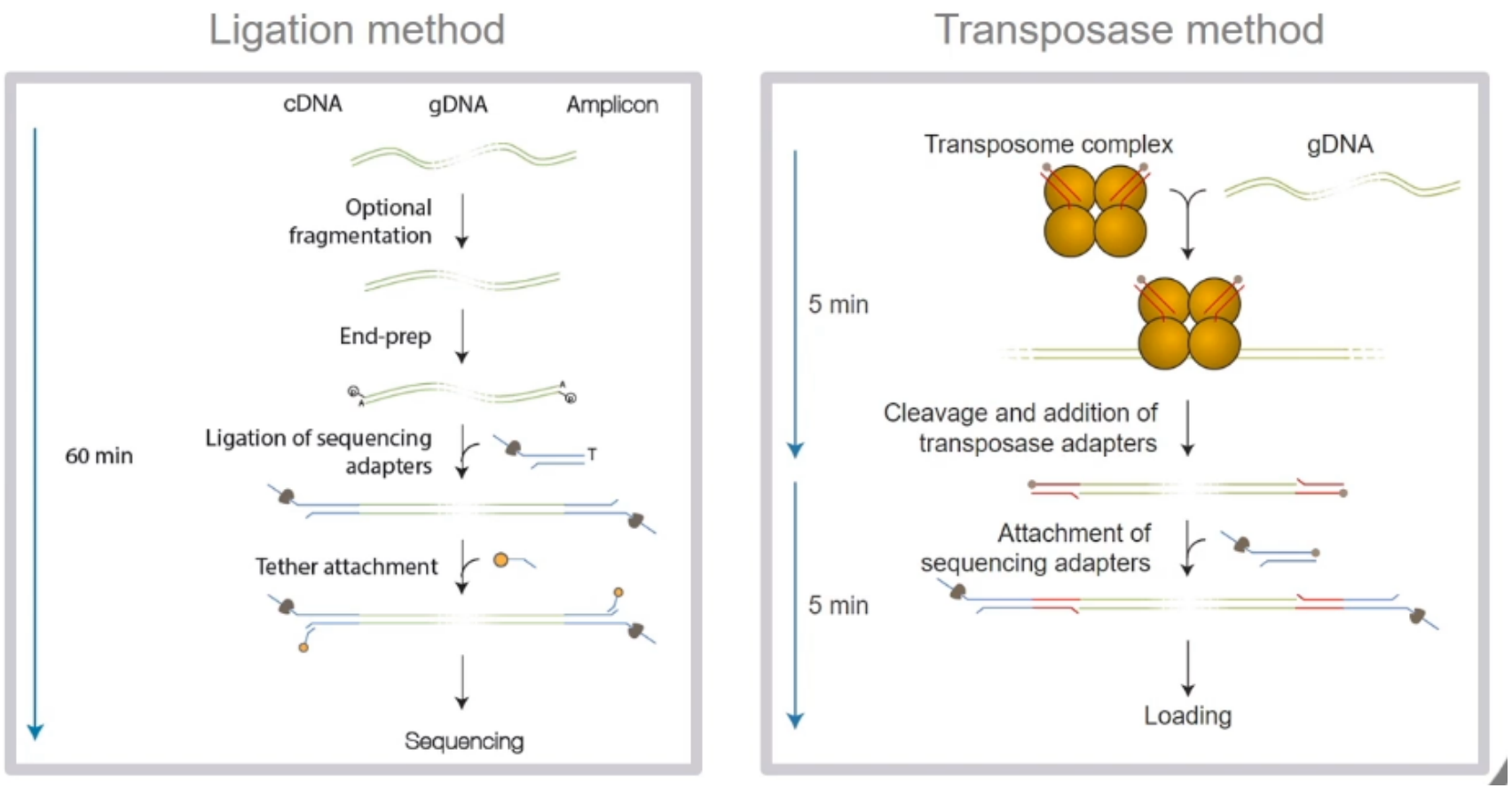
1 contig of the 4.6Mb chromosome of *E. coli*

obtained with just the 7 longest reads

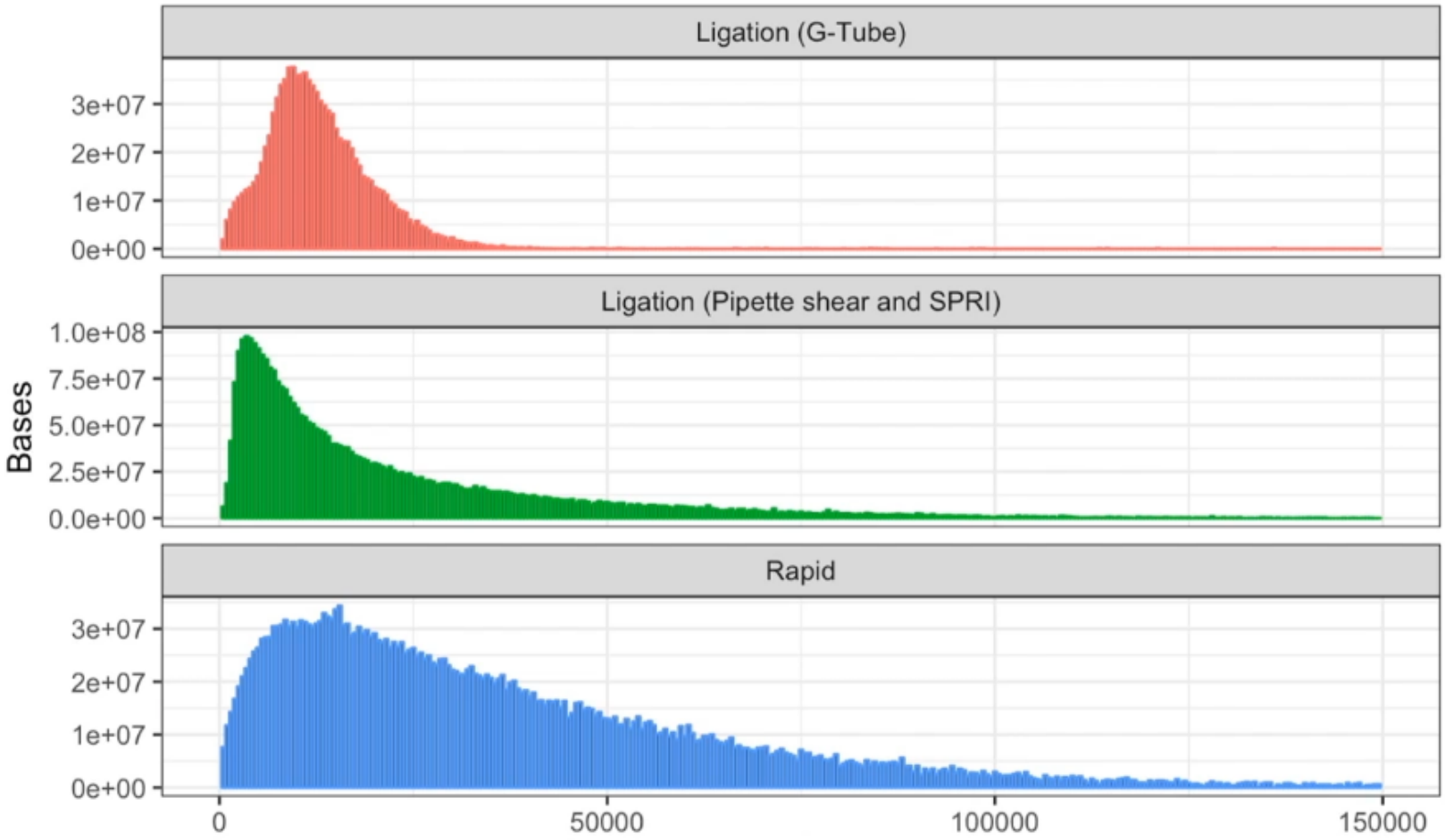
SIZE OF SEQUENCED DNA FRAGMENTS



SIZE OF SEQUENCED DNA FRAGMENTS



SIZE OF SEQUENCED DNA FRAGMENTS



Josh Quick, Nick Loman

see John Tyson's video (ONT website)

HYBRID GENOME ASSEMBLY : NANOPORE AND ILLUMINA DATA

Acinetobacter baylyi (data from Oxford Nanopore)

Assemblies	Illumina only	Illumina + MinION
Input Coverage	50X	13X
# contigs	20	1
Assembly size (Mb)	3.59	3.62
N90 size (Kb)	326	3 621
NA75 size (Kb)	194	1 002
Genome fraction (%)	99.73	99.997
# misassemblies	4	2
# local misassemblies	3	4
# mismatches per 100 Kb	6.49	3.11
# indels per 100 Kb	0.33	0.14

NANOPORE GENOME SEQUENCING

ARTICLES

nature biotechnology

OPEN

Jan. 2018

Nanopore sequencing and assembly of a human genome with ultra-long reads

Miten Jain[1,13], Sergey Koren[2,13], Karen H Miga[1,13], Josh Quick[3,13], Arthur C Rand[1,13], Thomas A Sasani[4,5,13], John R Tyson[6,13], Andrew D Beggs[7], Alexander T Dilthey[2], Ian T Fiddes[1], Sunir Malla[8], Hannah Marriott[8], Tom Nieto[7], Justin O'Grady[9], Hugh E Olsen[1], Brent S Pedersen[4,5], Arang Rhie[2], Hollian Richardson[9], Aaron R Quinlan[4,5,10], Terrance P Snutch[6], Louise Tee[7], Benedict Paten[1], Adam M Phillippy[2], Jared T Simpson[11,12], Nicholas J Loman[3] & Matthew Loose[8]

Using nanopore reads alone assembly of a human genome :

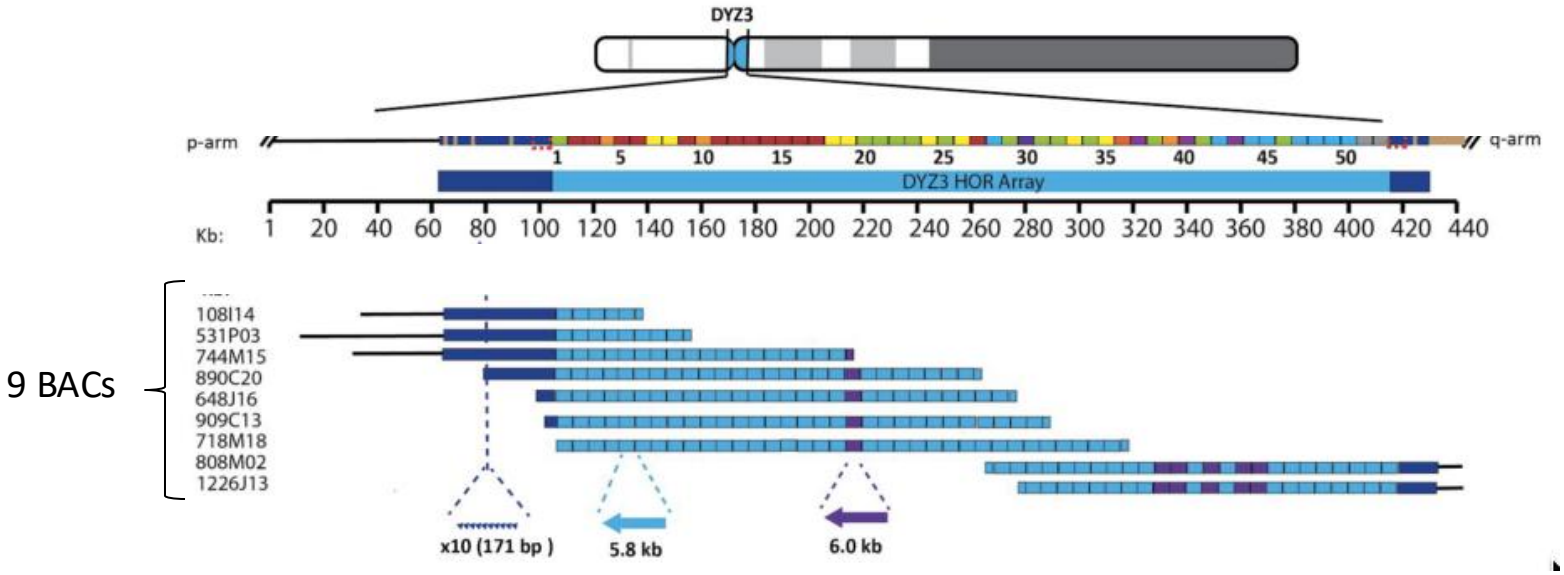
- NG50 contig size of ~6.4 Mb
- covers >85% of the reference
- 99.88% accuracy
- MHC locus on a single contig, phased over its full length
- closure of 12 large (>50 kb) gaps in the reference human genome

ASSEMBLY OF A HUMAN Y CENTROMERE

(Jain et al., *bioRxiv*, 2017)

300 kb array of 5.8 kb sequence repeated in an uninterrupted head-to-tail orientation

To date, no technology has been capable of sequencing centromeres due to requirement for extremely high-quality long reads

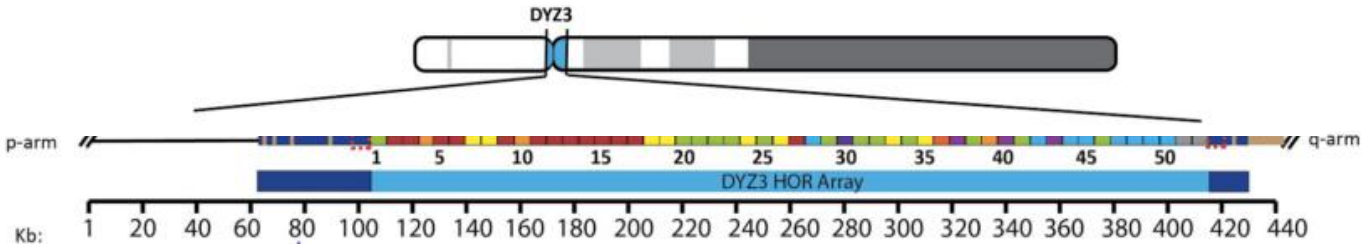


ASSEMBLY OF A HUMAN Y CENTROMERE

(Jain et al., *bioRxiv*, 2017)

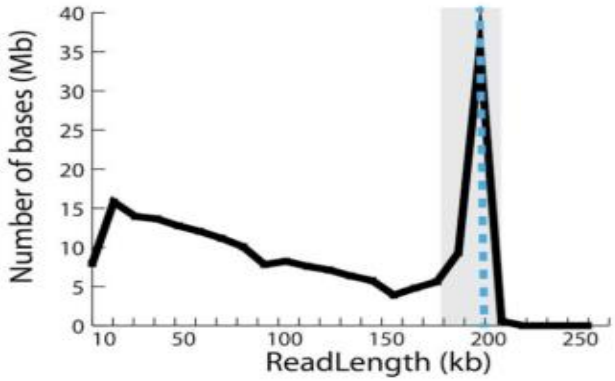
300 kb array of 5.8 kb sequence repeated in an uninterrupted head-to-tail orientation

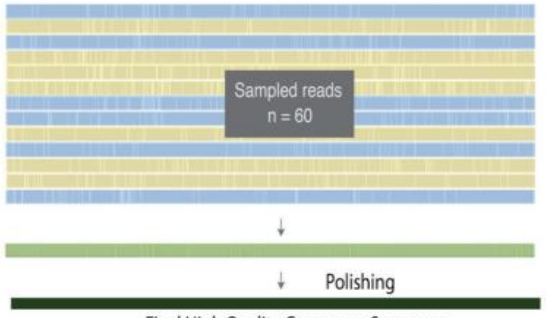
To date, no technology has been capable of sequencing centromeres due to requirement for extremely high-quality long reads



9 BACs
100kb to 210kb

210 kb



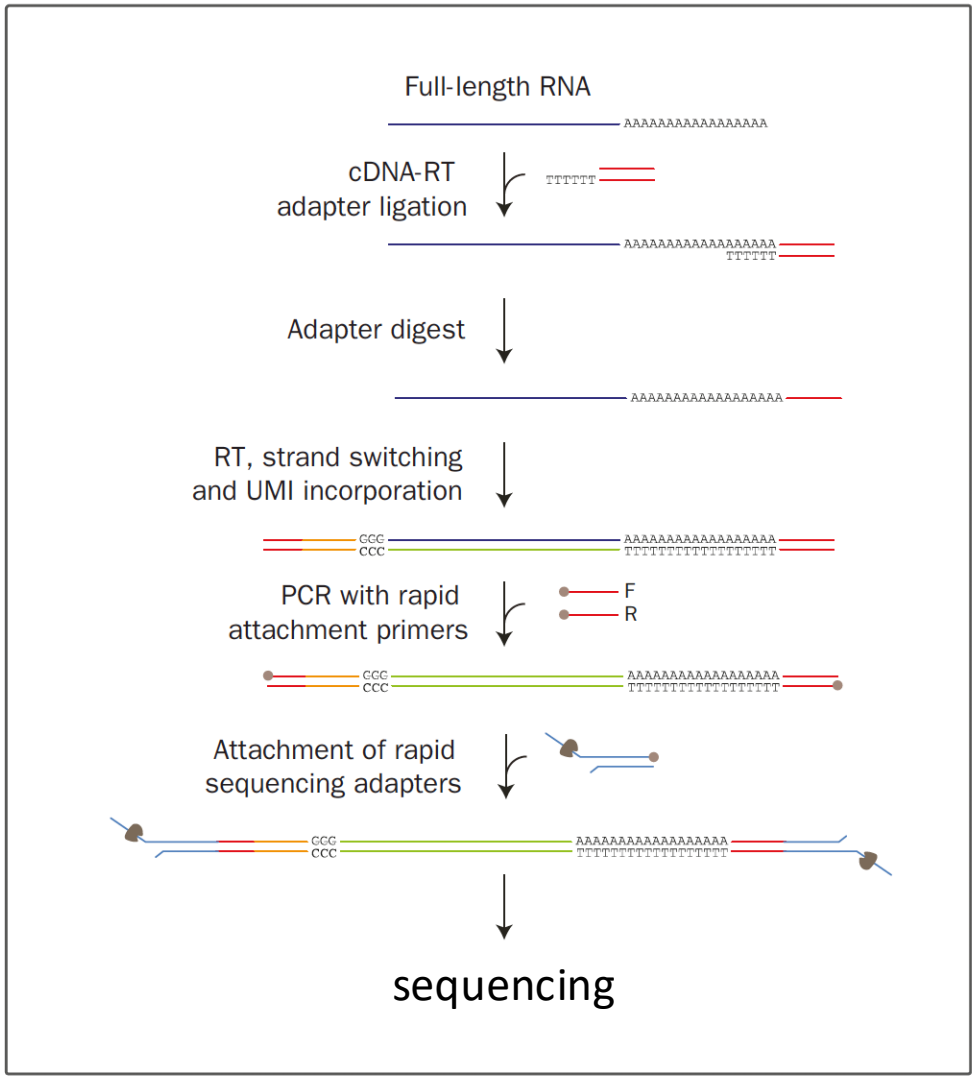


Final high quality consensus BAC sequence

FIRST COMPLETE SEQUENCE OF A HUMAN CENTROMERE

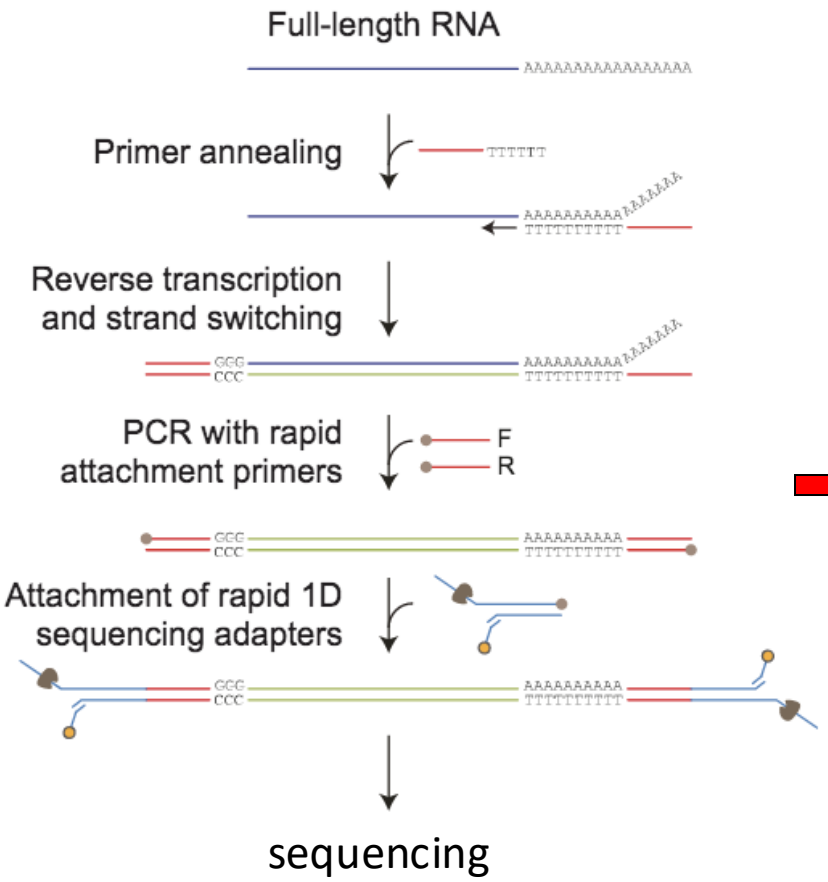
cDNA SEQUENCING

Library preparation



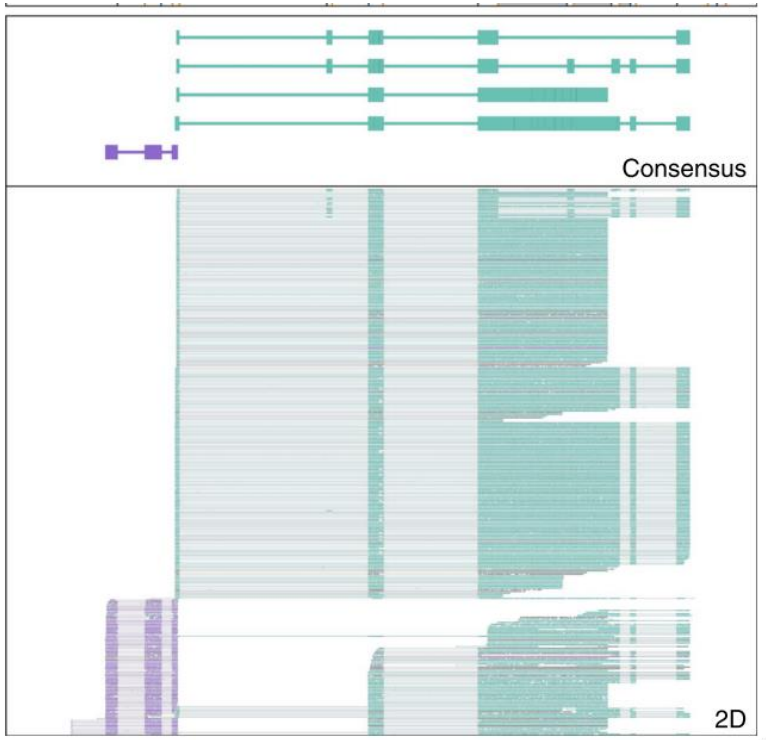
cDNA SEQUENCING

Library preparation



Detection of splice variants in surface receptor of B cells

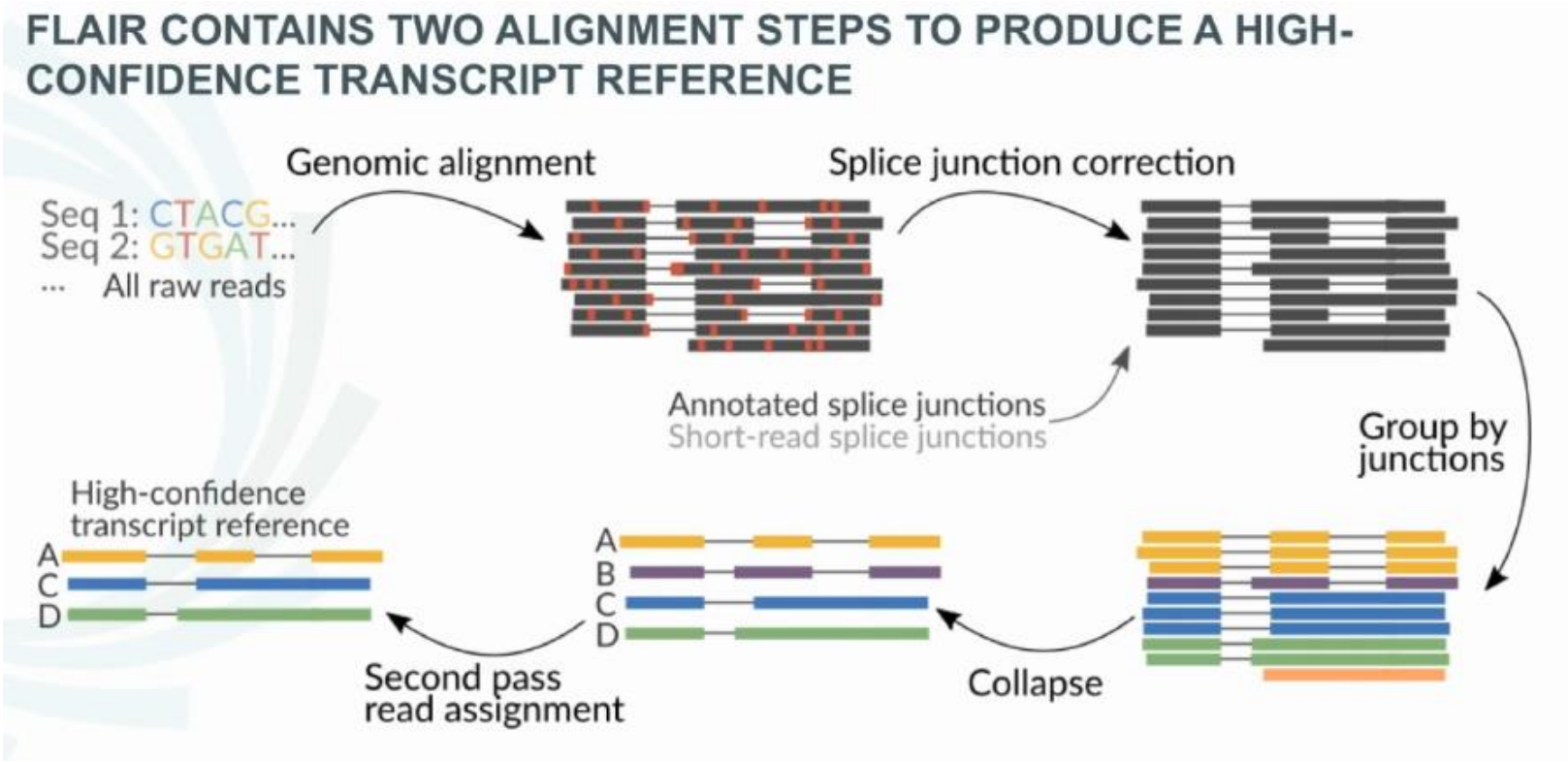
(Byrne et al. *Nat. Comm.* 2017)



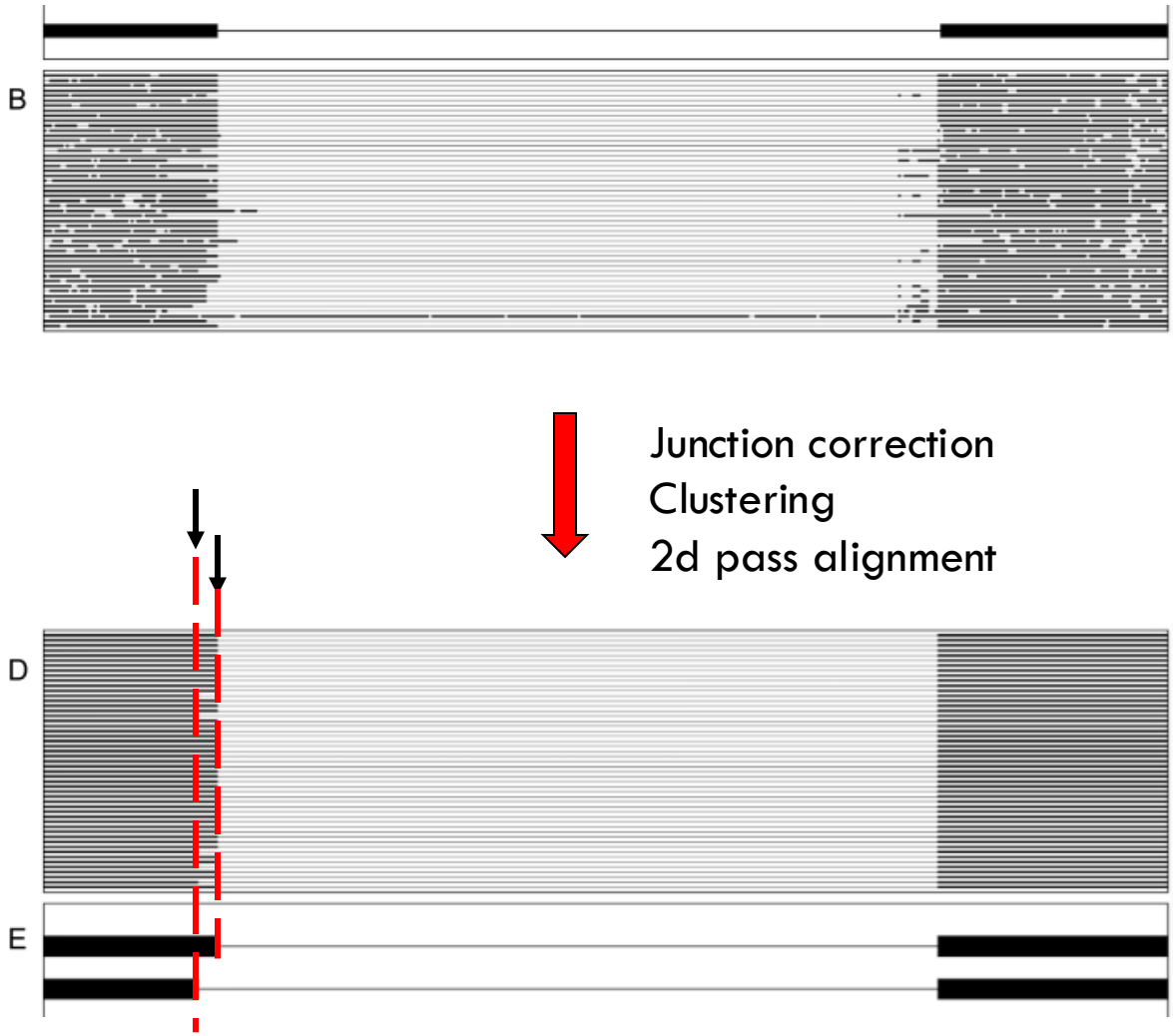
- Splice alignment difficult due to high (5-10%) error rate
- Reads are frequently truncated from 5' end

FLAIR : a pipeline for splicing isoform determination

Tang et al. *bioRxiv* 2018



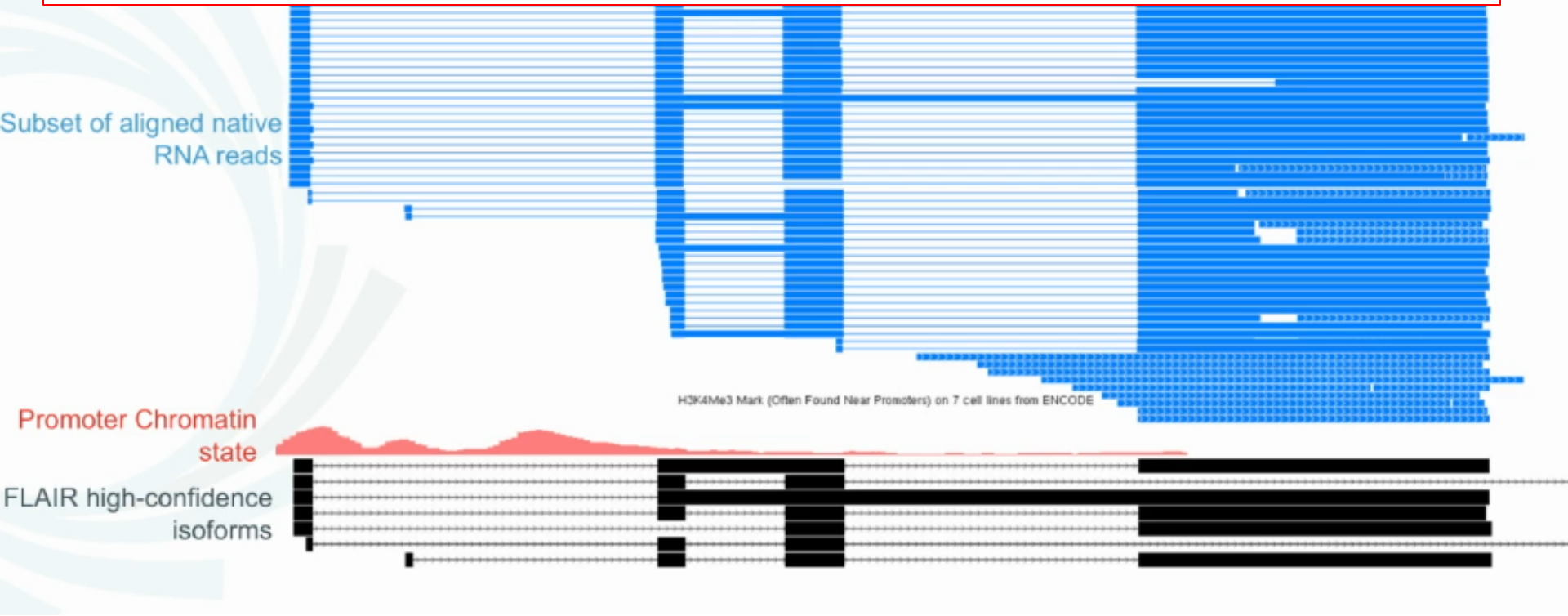
CHALLENGES OF NANOPORE TRANSCRIPTOME ANALYSIS



Tang et al. *bioRxiv* 2018

CHALLENGES OF NANOPORE TRANSCRIPTOME ANALYSIS

FLAIR pipeline incorporates promoter chromatin states to distinguish 5' truncations from true novel start sites



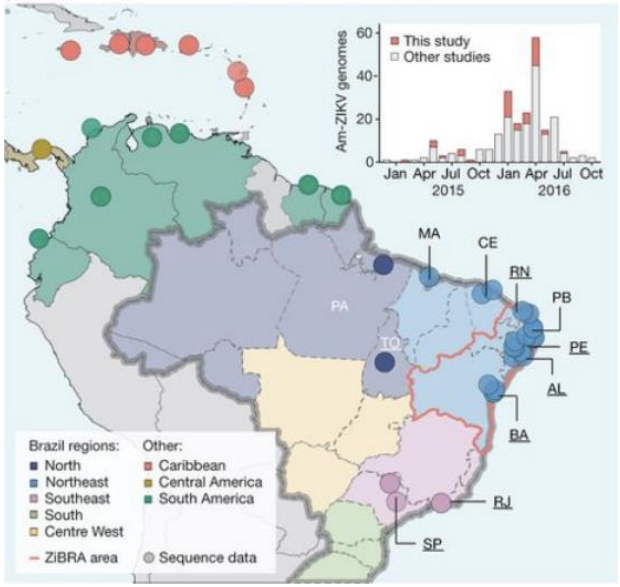
cDNA from chronic lymphocytic leukemia (CLL)

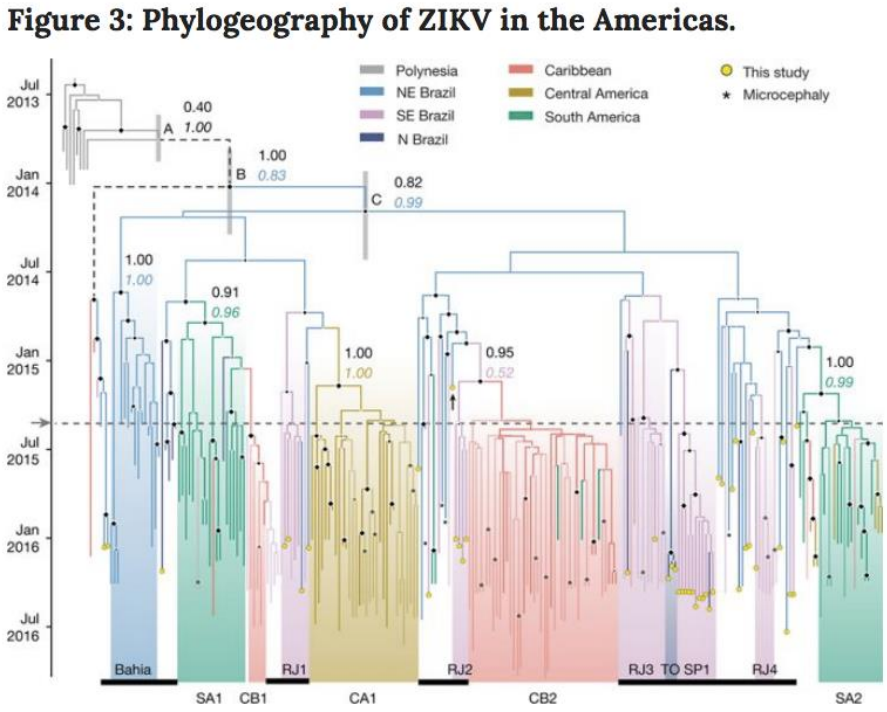
Tang et al. *bioRxiv* 2018

cDNA SEQUENCING

Faria et al. *Nature* (2017)

THE ZIBRA PROJECT : Establishment and cryptic transmission of Zika virus in Brazil and the Americas
Mobile genomics laboratory that travelled through northeast Brazil during June 2016.
The ZiBRA laboratory screened 1,330 samples (blood) from patients in 82 municipalities across 5 federal states
The MinION protocol does not require an Internet connection for analysis, making it suitable for field applications
Viral RNA genome : + sense, 10 kb
Viral consensus sequences can be achieved in 1-2 days.





ZIKV was present in northeast Brazil by February 2014 and is likely to have disseminated from there, nationally and internationally, before the first detection of ZIKV in the Americas.

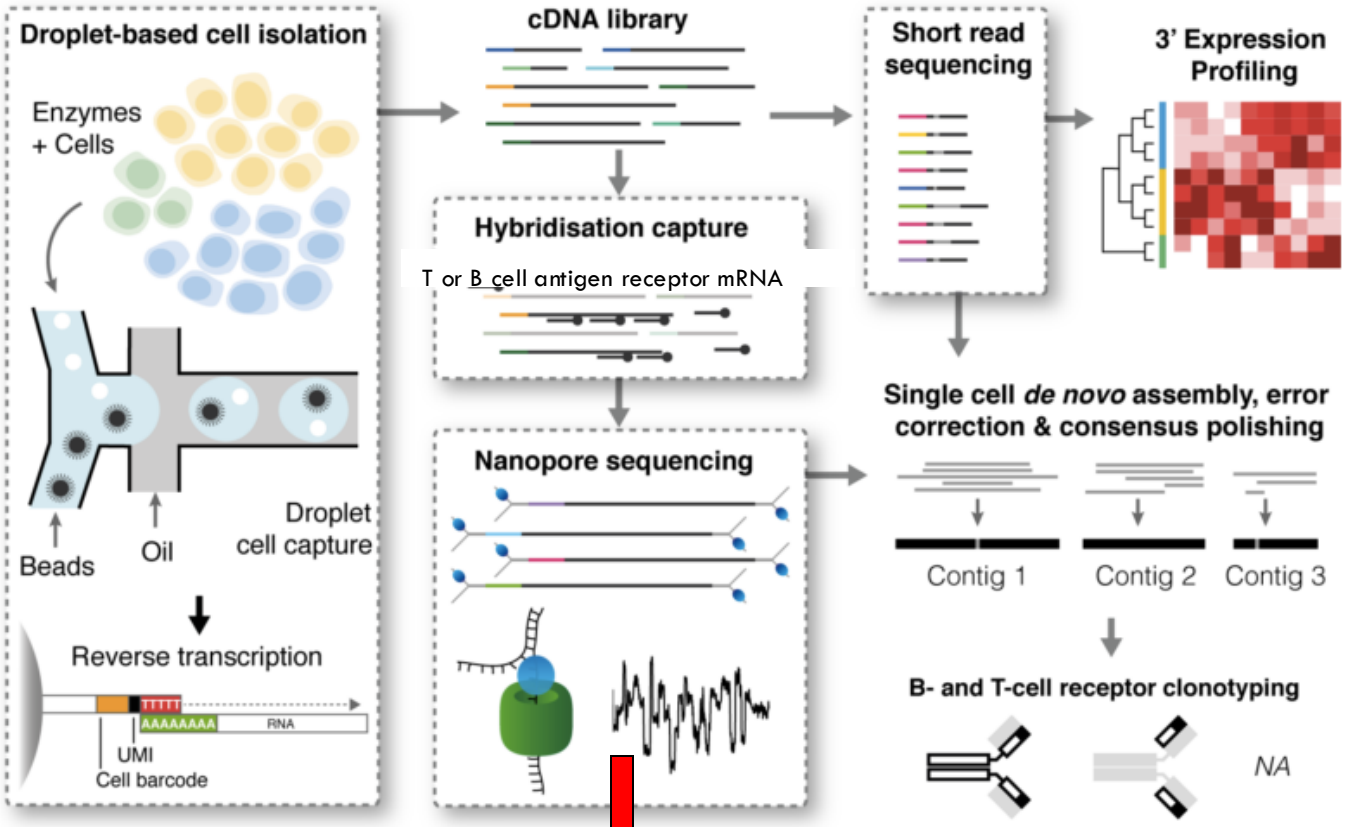
COUPLING NANOPORE and SINGLE CELL cDNA SEQUENCING

High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes

Singh et al., *bioRxiv*, 2018

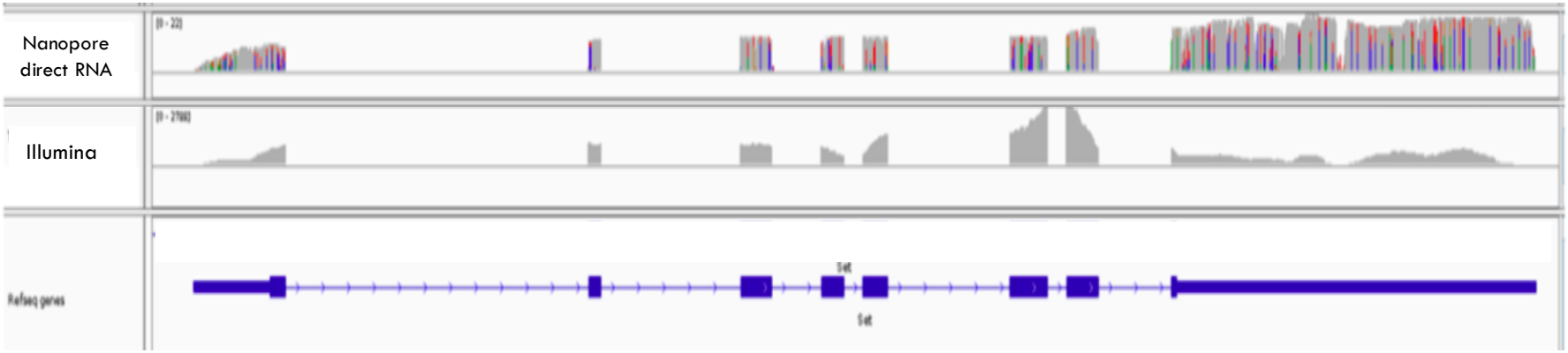
RAGE-seq (Repertoire And Gene Expression sequencing): high-throughput deep single cell profiling combines targeted long-read sequencing with short-read transcriptome of barcoded single cell libraries

10x Chromium Single Cell 3' (V2 chemistry)

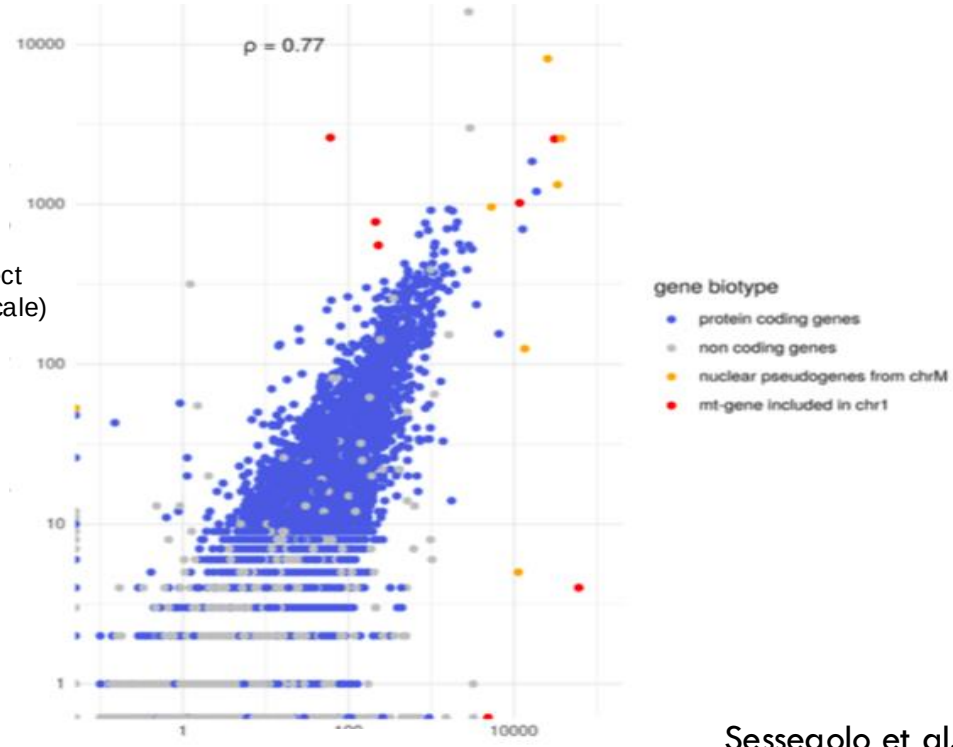


Tracking of somatic mutation, alternate splicing and clonal evolution of T and B lymphocytes

ILLUMINA cDNA vs NANOPORE DIRECT RNA



Nanopore RNA direct
(read number ; Log scale)



Illumina Truseq

Sessegolo et al. *Sci. Reports* 2019

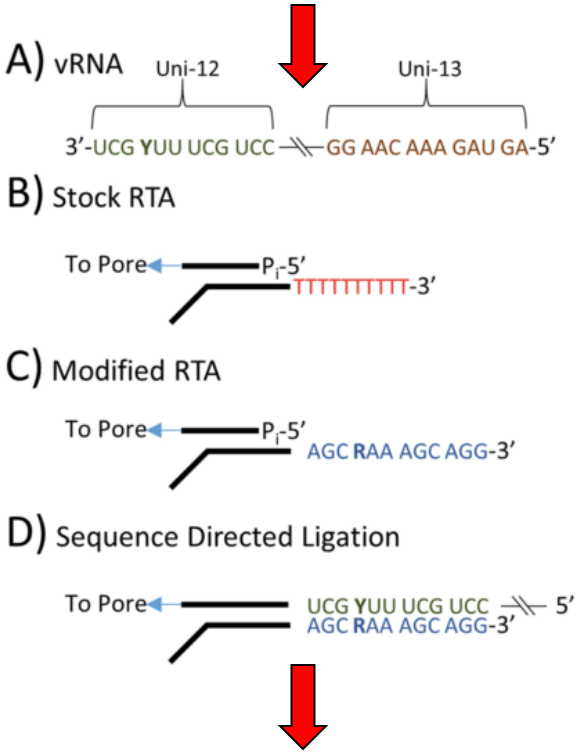
DIRECT RNA SEQUENCING

Direct RNA Sequencing of the complete Influenza A Virus Genome

Keller et al. *Scientific Reports*, Sept. 2018

For the first time a complete genome of an RNA virus sequenced in its original form

Influenza A viruses are negative-sense segmented RNA viruses (8 segments)

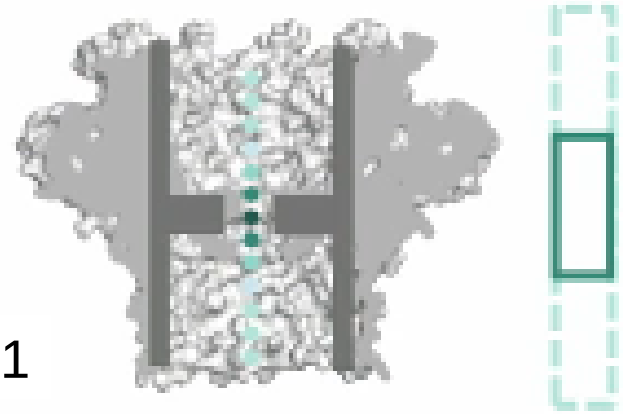


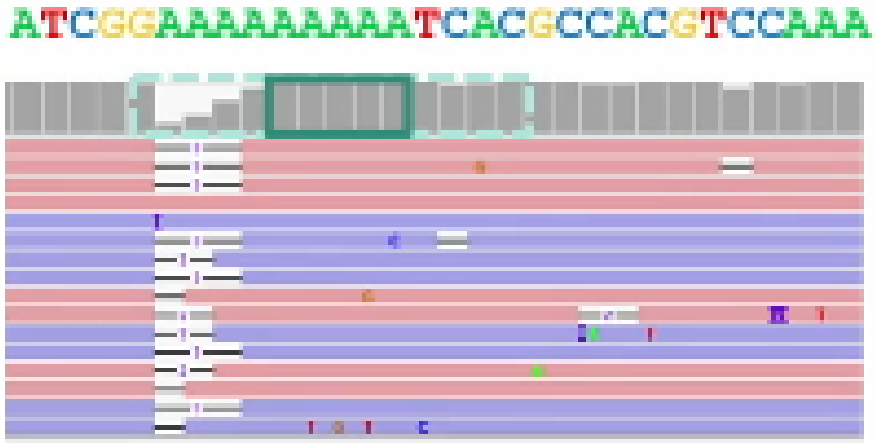
sequencing of complete genome with 100% nucleotide coverage, 99% consensus identity

Potential to identify and quantify splice variants, base modifications
not practically measurable with current methods

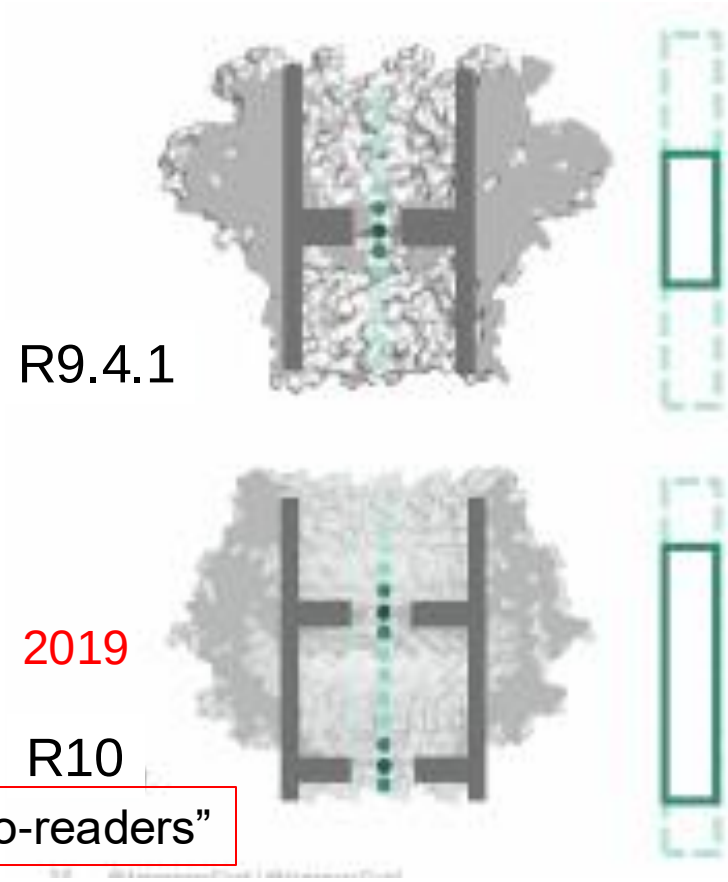
“one-reader” pore has difficulty to read homopolymers

R9.4.1





RECENT IMPROVEMENTS

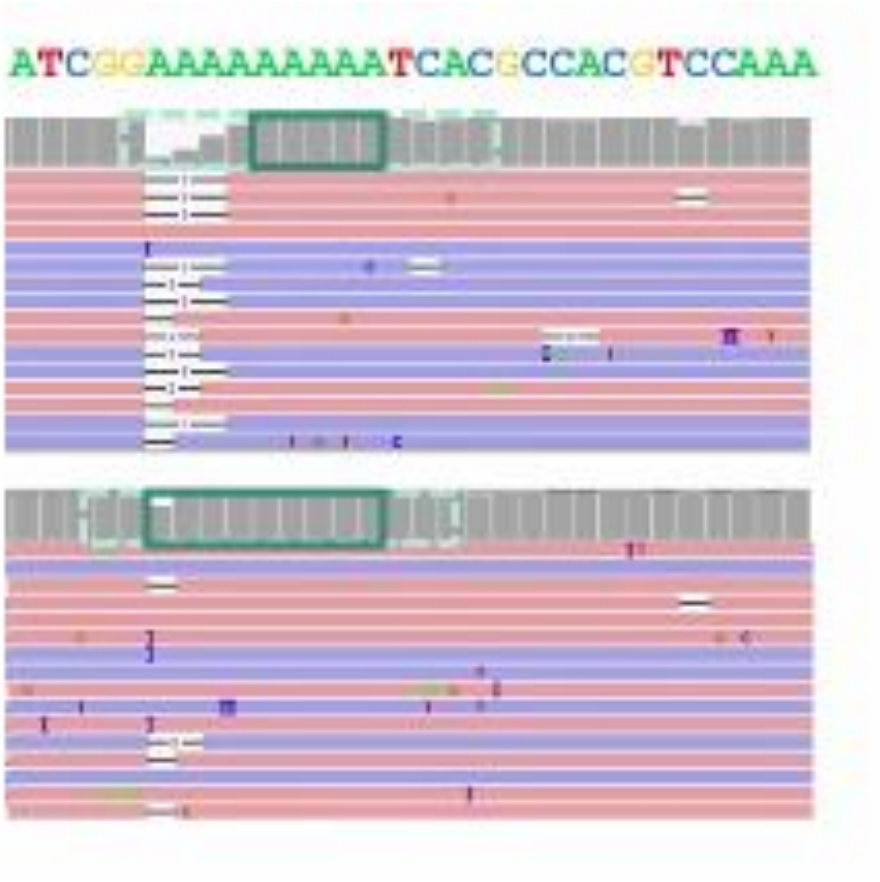


R9.4.1

2019

R10

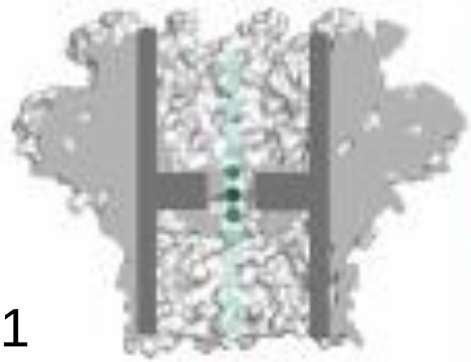
"two-readers"



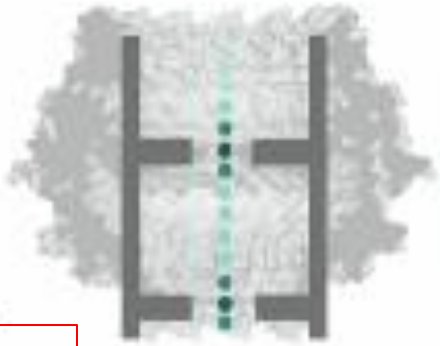
New pore accurately calls homopolymers

- A pore with a longer or multiple "readers" has more bases dominating the signal
- Longer homopolymers are "seen" by the pore and can be decoded with high accuracy

RECENT IMPROVEMENTS



R9.4.1

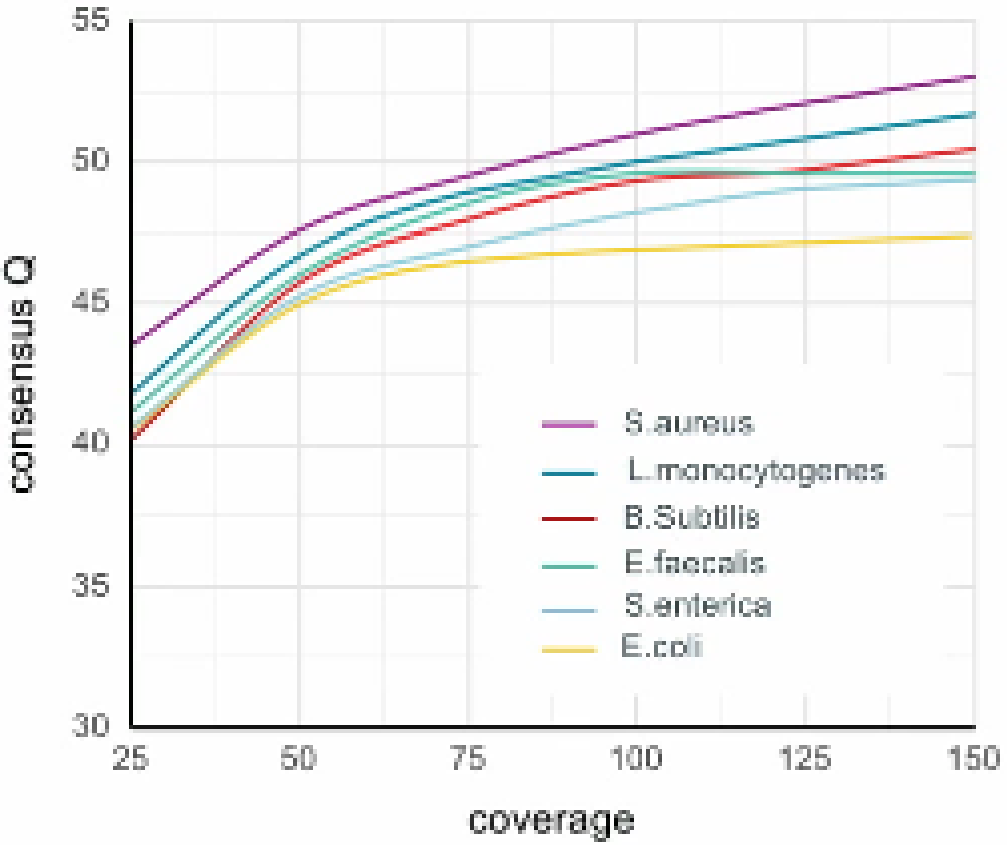


2019

R10

"two-readers"

Consensus accuracy (R10 flow cell)



"two-readers" flow cell can reach Q>50

Conclusions

Very fast-evolving technologies – Strong competition

PacBio

- Maximum read length : 200 kb
- Error rate compensated by highly accurate circular consensus sequencing (CCS) reads
- Sequencing of cDNAs (resolution of alternative splicing)
- Detection of modified DNA with context effects (preferentially 6mA)

Nanopore

- Very light sequencing system
- Very long reads : maximum length >> 200 kb
- Problems with homopolymers : solution with “two-readers” pore
- Sequencing of cDNAs (resolution of alternative splicing)
- Detection of modified DNA with context effects (preferentially 5mC)
- Direct sequencing of RNA
- Direct detection of modified RNA (6mA)