NGS
What you need to know **BEFORE** starting analysis

Rie Shimizu-Inatsugi
Department of Evolutionary Biology and Environmental Studies
UZH
rie.inatsugi@ieu.uzh.ch

Whole Workflow with NGS

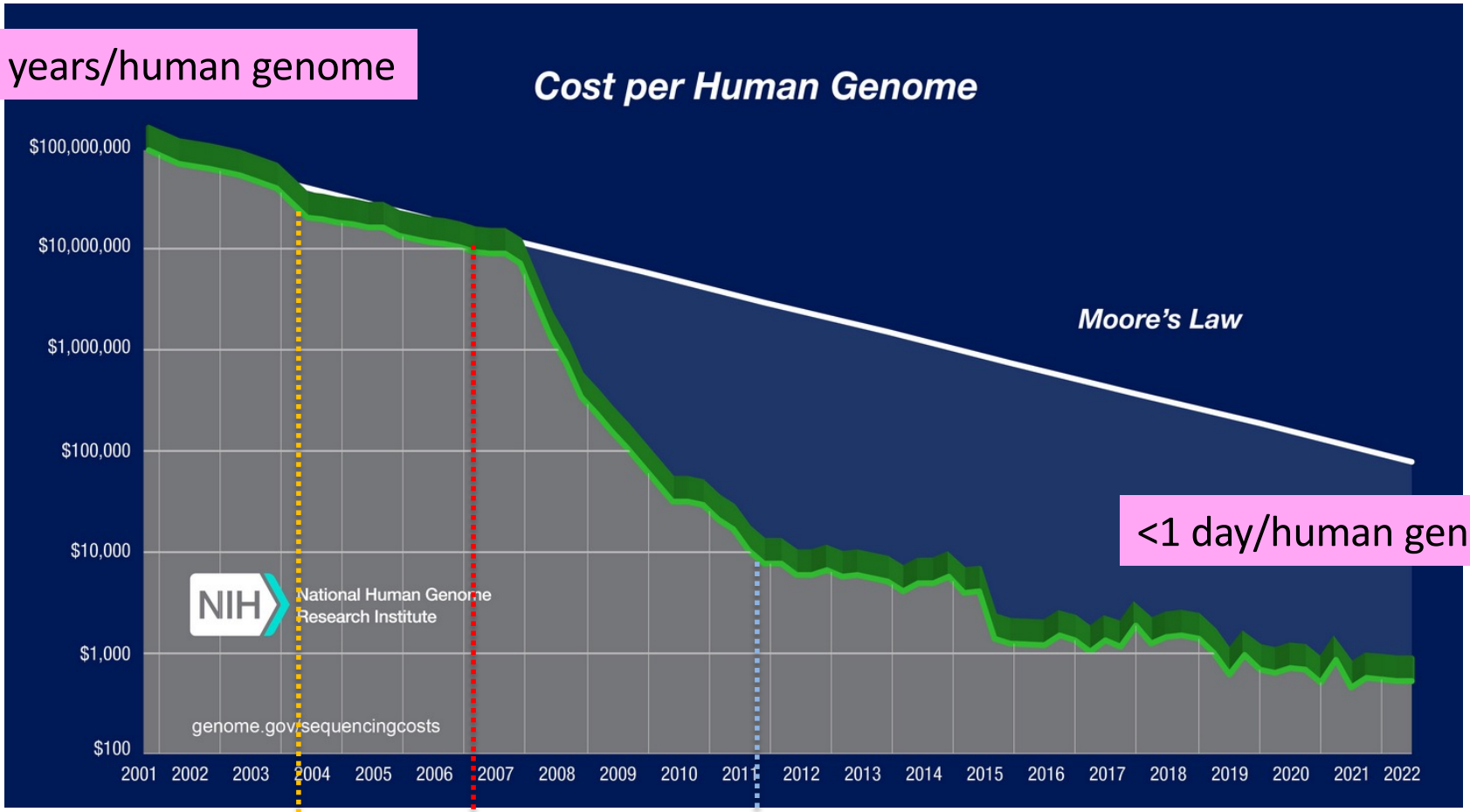
1. Question
2. Experimental Design
3. Field trip or lab culture
4. Sample collection
5. Sample storage
6. Extraction (DNA, RNA…)
7. Quality & quantity check
8. Library synthesis
9. NGS Sequencing
10. Data quality control (QC)
11. Data analysis (data mining)

This talk

Main part of BIO610

Sequencing
Classic vs. New

20< years/human genome



<1 day/human genome

Sanger ←→ NGS

(454)

Solexa/Illumina

PacBio

modified from:
https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

Many NGS platforms (machine types) …so far

FLX (454)

SOLiD

BioNano

PACBIO

Nanopore

Ion Torrent

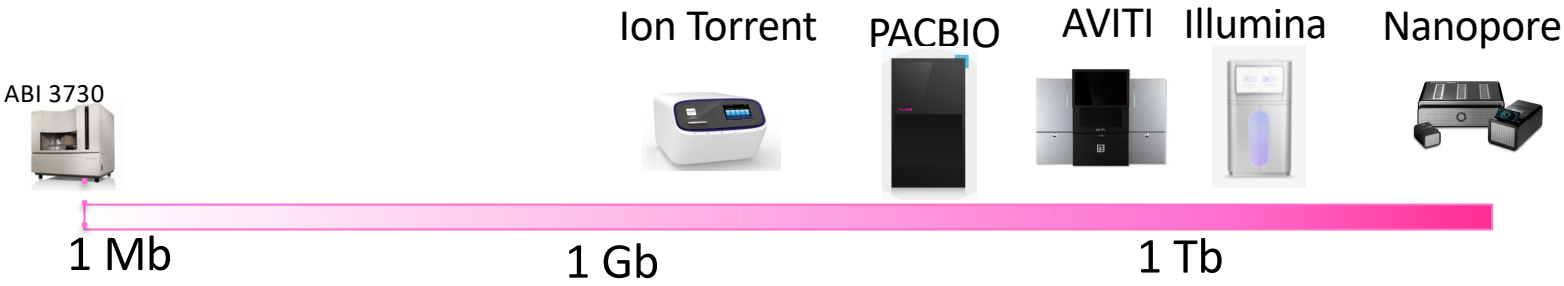
Illumina

AVITI

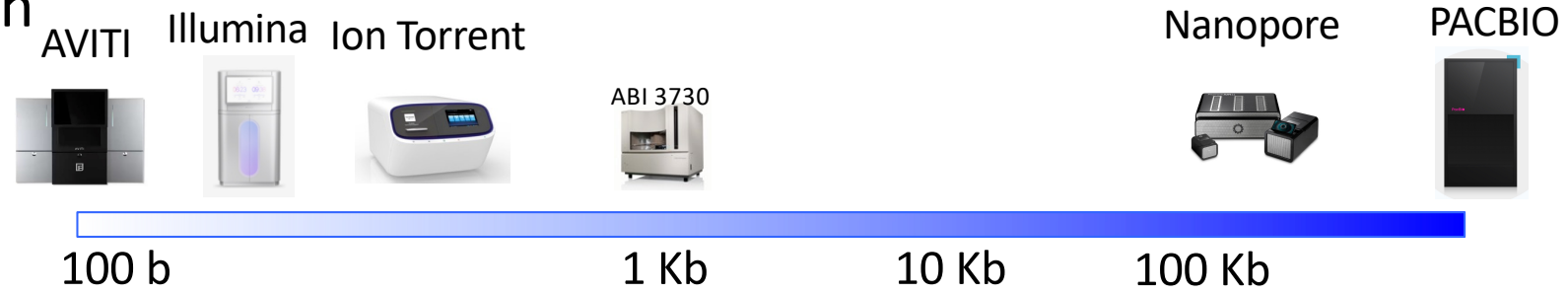
DNBSEQ

How are they different?

Output per day



Read length



Accuracy

What can we sequence by NGS?

Genomics/Genetics (DNA)

-*de novo* genome sequencing
-re-sequencing
(SNP-calling of mass individuals)
-Genotyping by RAD-seq
(Restriction site Associated DNA Sequencing, non whole genome)

Epigenetics (DNA/RNA)

-small RNA
-DNA methylation
-ChIP sequencing
(histone methylation/acetylation)
-chromatin structure

Transcriptomics (mRNA)

-gene expression pattern
-exome
-isoform/splice variant discovery

Community genomics (DNA/RNA) (Metagenomics)

-microbes 16*S* rRNA
-total metagenome
-environmental DNA (eDNA)

Target-enriched sequencing (DNA/RNA)

Enrichment by **hybridization**
-ChIP-seq
(Chromatin Immunoprecipitation)
-Array capture (Nimblegen etc.)
-Beads capture (Myselect etc.)

Enrichment by **amplification** (PCR)
-Amplicon
(normal PCR product, Fluidigm etc.)
-Exon sequencing (Exome)

Application examples of NGS

Human health

-Personalized medicine (100$ genome coming)/Clinical test (pharmacogenomics, oncopanel test: detect mutation causing cancer)

Agricultural applications

-Breeding

Basic Research

-'non-model' species (diversity, evolution, ecology…)

-Ancient DNA (extinct species, ancient humans or animals…)

-Single cell analysis

Others

-Criminal investigation (forensic medicine)

-Parentage diagnosis

-Personal health support (diseases, dietary advice service etc.)

-Non-invasive prenatal testing 'trisomy' etc.

-Ancestry test

Platforms currently available in FGCZ

Illumina

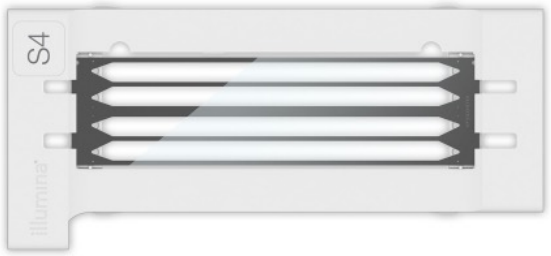
AVITI

short read

PACBIO

Nanopore

long read



Illumina

Short read
Massive output

RNA-seq
DNA-seq

Key specifications	iSeq 100 System	MiniSeq System	MiSeq System	NextSeq 550 System	NextSeq 1000 and 2000 Systems	NovaSeq 6000 System	NovaSeq X Series
Max output per flow cell	1.2 Gb[a]	7.5 Gb[b]	15 Gb[c]	120 Gb[b]	540 Gb[d]	3 Tb[b]	8 Tb[c]
Run time (range)[e]	~9.5–19 hr	~5–24 hr	~5–56 hr	~11–29 hr	~8–44 hr	~13–44 hr	~17–48 hr
Max reads per run (single reads)	4M[a]	25M[b]	25M[c]	400M[b]	1.8B[d]	10B (single flow cell)[b] 20B (dual flow cells)[b]	26B (single flow cell)[c] 52B (dual flow cells)[c,e]
Max read length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 300 bp	2 × 250 bp	2 × 150 bp

Wide selection of platforms

100$ genome is coming!
(human genome = 3Gb)

https://www.youtube.com/watch?v=fCd6B5HRaZ8

http://www.illumina.com/

AVITI (Element Biosciences)

Short read
Massive output

RNA-seq
DNA-seq

The same library as Illumina can be sequenced on this platform with relatively lower cost, in smaller scale.

Sequencing metrics at 800M Reads PF/flow cell; 1600M Reads PF combined when running both flow cells

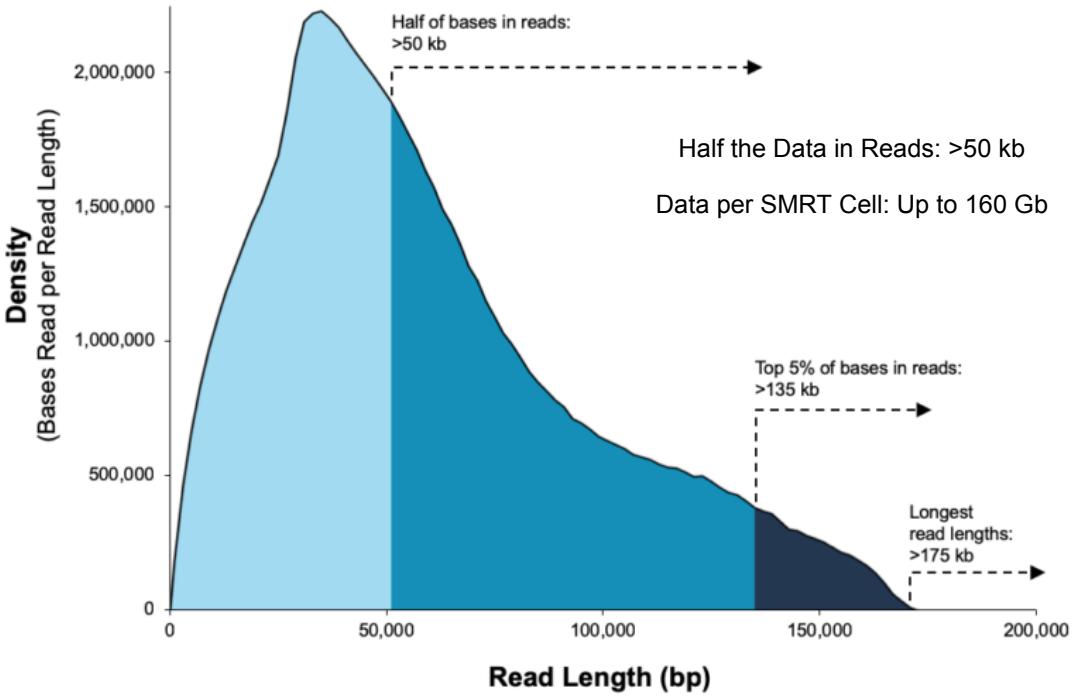
READ LENGTH	1 FLOW CELL DATA OUTPUT (GB)	2 FLOW CELLS DATA OUTPUT (GB)	SEQUENCING RUN TIME	DATA QUALITY
2x150	240	480	48hrs	%Q30 > 90
2x100	160	320	35hrs	%Q30 > 90
2x75	120	240	29hrs	%Q30 > 90
2x50	80	160	23hrs	%Q30 > 90
2x25	80	80	17hrs	%Q30 > 90

https://www.elementbiosciences.com/

PACBIO (Pacific Bioscience)

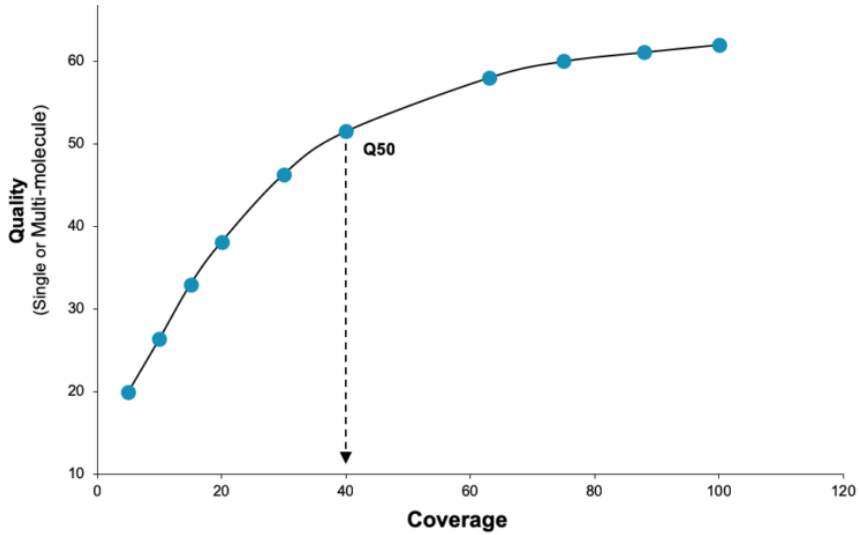
Sequel II

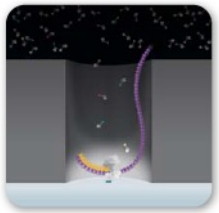
Revio
(15x scale)

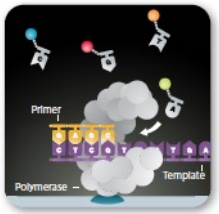


Long read
Single molecular seq
Middle output

de novo sequencing
DNA modification
(methylation)







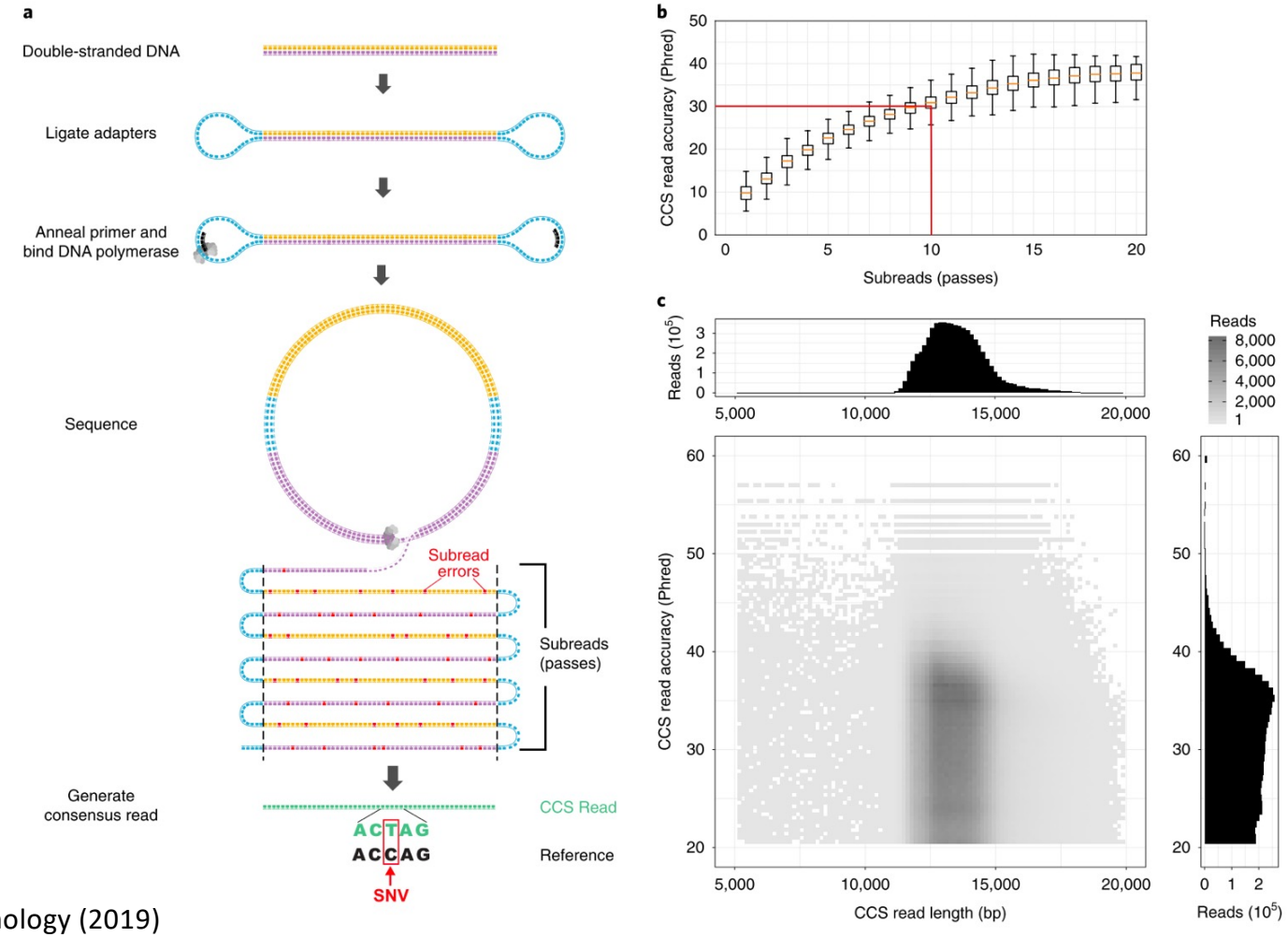
https://www.youtube.com/watch?v=v8p4ph2MAvI
https://www.youtube.com/watch?v=_ID8JyAbwEo

http://www.pacb.com/

Two sequencing modes of PACBIO

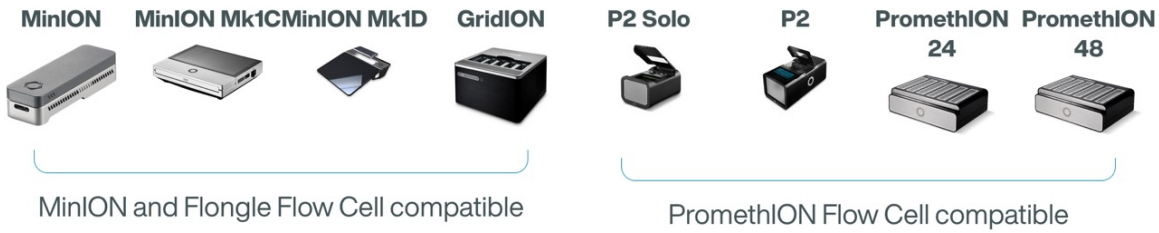
CLR (continuous long read) sequencing — 1 single DNA molecule = 1 read

CCS (circular consensus sequencing) = **HiFi** reads — 1 single DNA molecule = 10< read



Nature Biotechnology (2019)
(https://doi.org/10.1038/s41587-019-0217-9)

Nanopore (Oxford Nanopore Technologies)



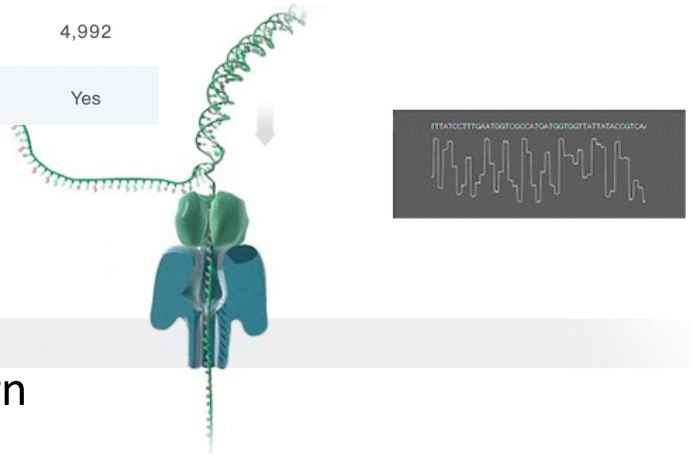
Long read
Single molecular seq

de novo sequencing
mRNA sequencing
DNA modification
(methylation)

Configuration	Platform			Techniques		Tech specifications	
Number of flow cells per device	1	1	1	5	2	2	24	48
Maximum number of channels per flow cell	512	512	512	512	2,675	2,675	2,675	2,675
Run time	72 Hours	72 Hours	72 Hours	72 Hours	72 Hours	72 Hours	72 Hours	72 Hours
Device TMO†	50 Gb	50 Gb	50 Gb	250 Gb	580 Gb	580 Gb	~7 Tb	~14 Tb
Maximum number of flow cells per year*	104	104	104	520	208	208	2,596	4,992
Offer sequencing as a service	No	No	No	Yes	Yes	Yes	Yes	Yes

read length
up to 300Kb

†Theoretical max output when system is run for 72 hours at 420 bases / second with all flow cells sequencing.



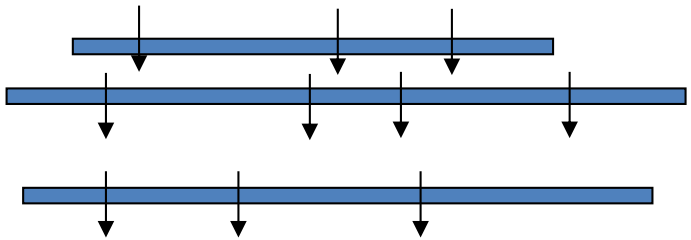
fluorescence-free method (low cost)

context-dependent interpretation of the pattern

https://www.youtube.com/watch?v=CGWZvHIi3i0

https://nanoporetech.com/

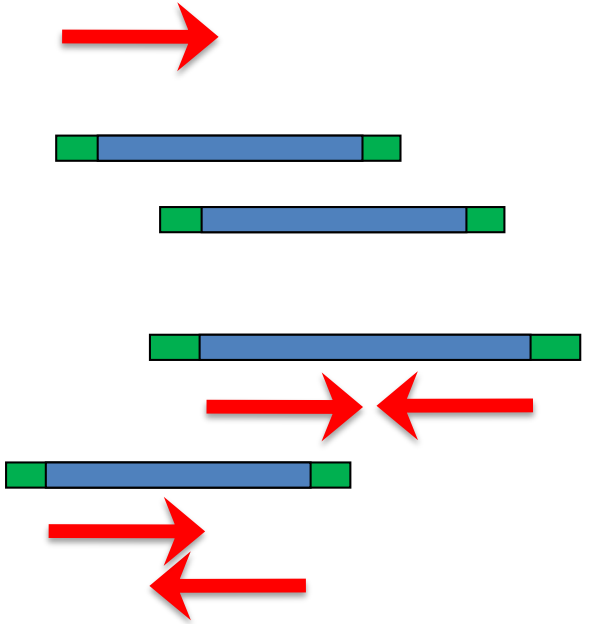
Application of short read sequence
Shotgun (DNA) library



Fragmentation (~500 bp)
mechanical (nebulization, sonication)
or enzymatic

Bind Adaptors

Single-end sequencing
(only from one side)

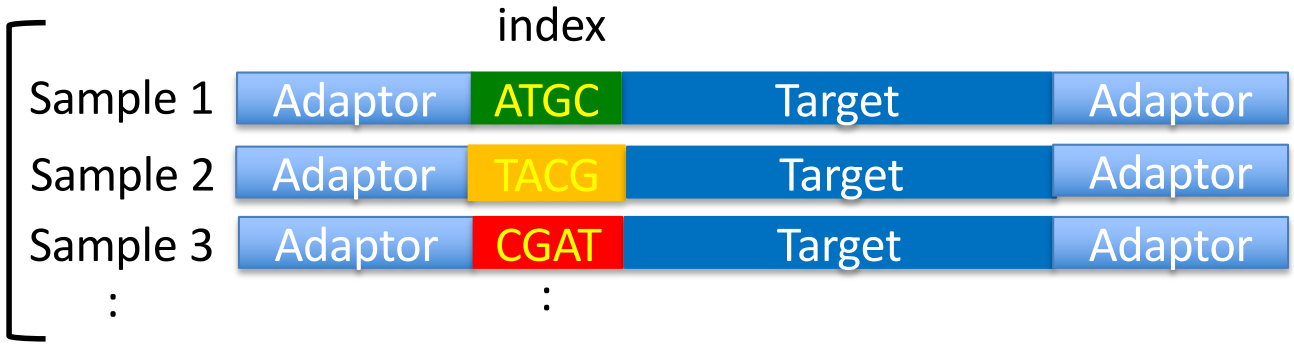


Paired-end sequencing
(from both side)

How many samples can be sequenced together?

Index (barcode, tag)

Tagging a group of fragments from one same sample to read them together in one sequencing run.



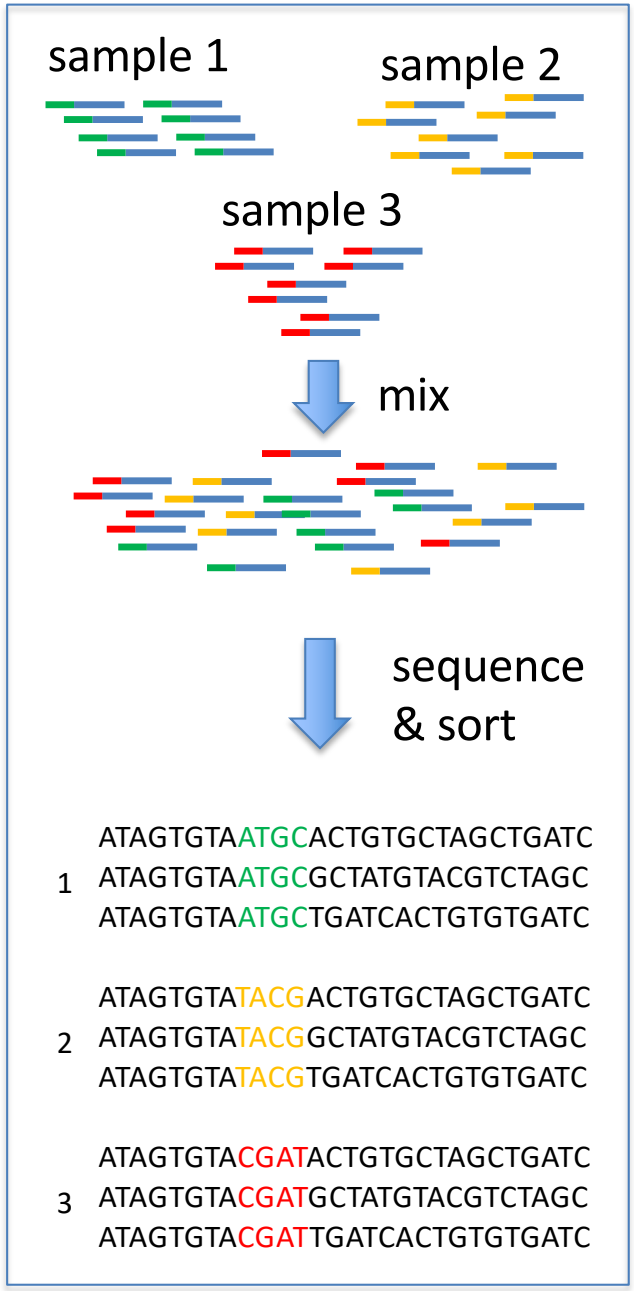
Pool to sequence **(multiplexing, pooling)**

Separate by bioinformatics (**demultiplexing**)

illumina index

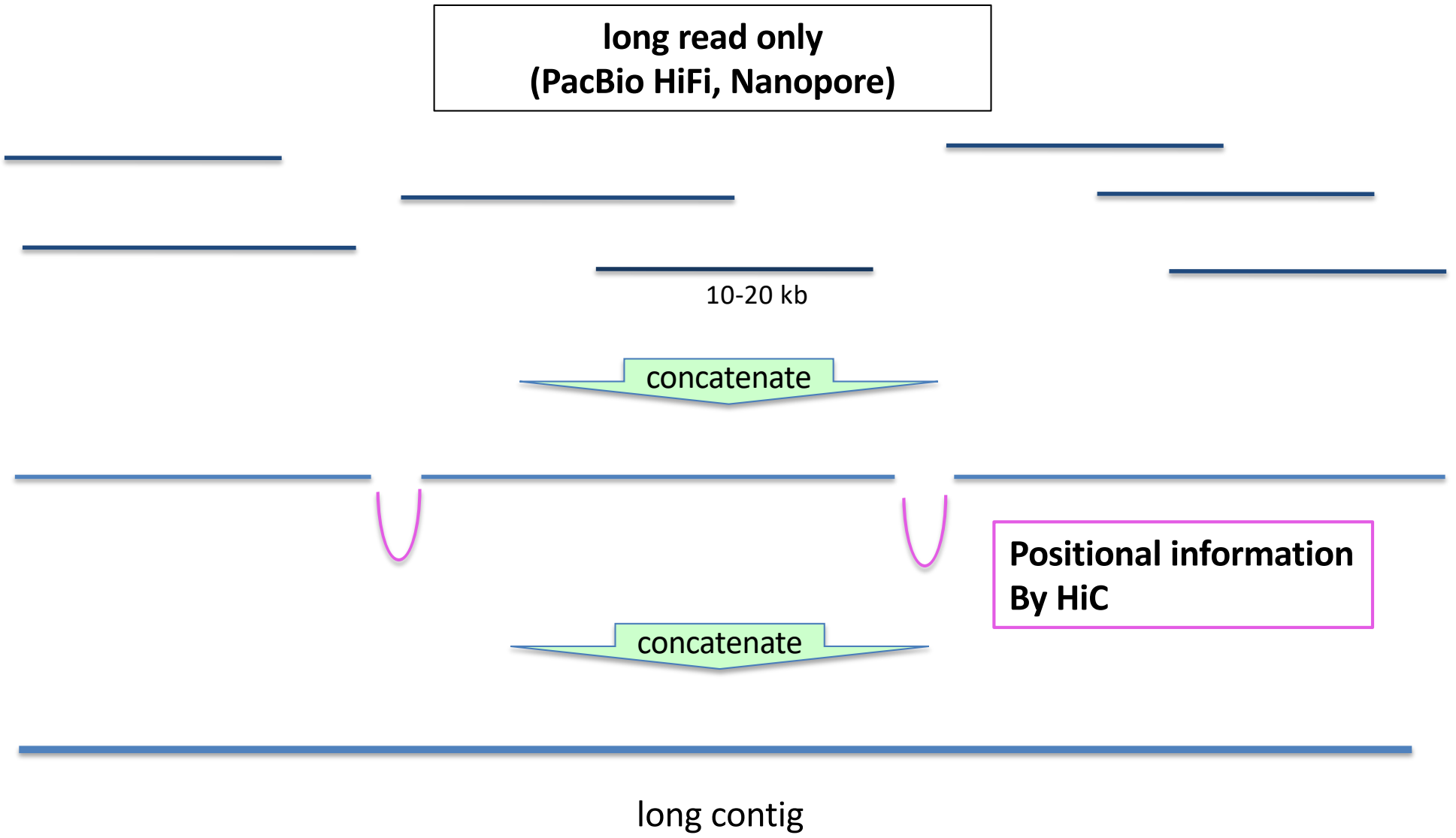
Table 13 TruSeq Stranded mRNA LT Sample Prep Kit Indexed Adapter Sequences

Adapter	Sequence	Set		Adapter	Sequence	Set
AR001	ATCACG(A)	B		AR013	AGTCAA(C)	A
AR002	CGATGT(A)	A		AR014	AGTTCC(G)	A
AR003	TTAGGC(A)	B		AR015	ATGTCA(G)	A
AR004	TGACCA(A)	A		AR016	CCGTCC(C)	A
AR005	ACAGTG(A)	A		AR018	GTCCGC(A)	A
AR006	GCCAAT(A)	A		AR019	GTGAAA(C)	A
AR007	CAGATC(A)	A		AR020	GTGGCC(T)	B
AR008	ACTTGA(A)	B		AR021	GTTTCG(G)	B



Genomics 1. De novo sequencing

producing the first ‘reference’ genome assembly

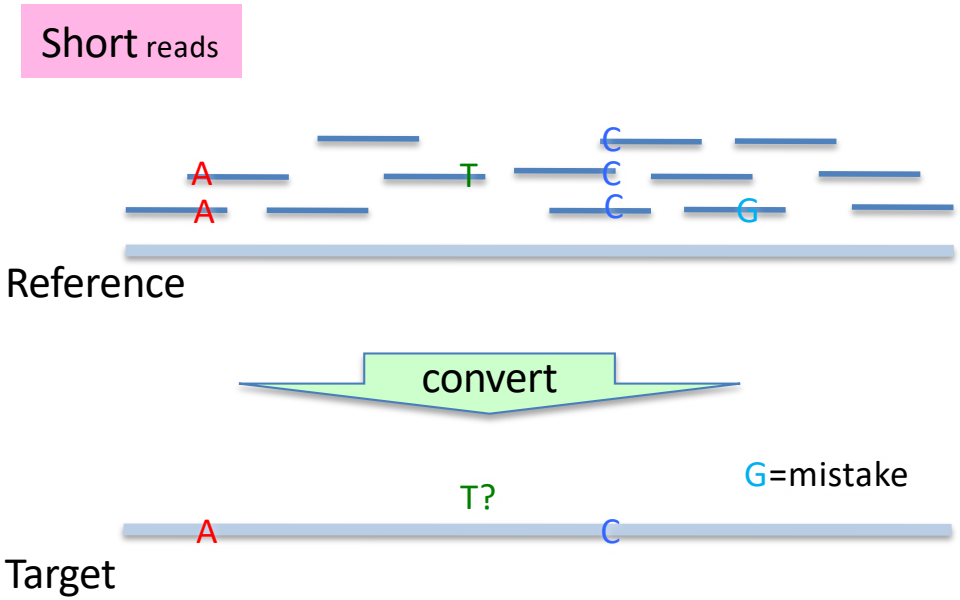


Genomics　2. Re-sequencing

sequence multiple individuals of the same species

-map on the reference genome,
and find the difference (variation)

= **intra-specific SNPs & Indels (0.1%)**



SNPs (single nucleotide polymorphisms)

Low coverage = not supported (declined)
Low frequency = mistake
High coverage/frequency = accurate (supported)

Basic Information for

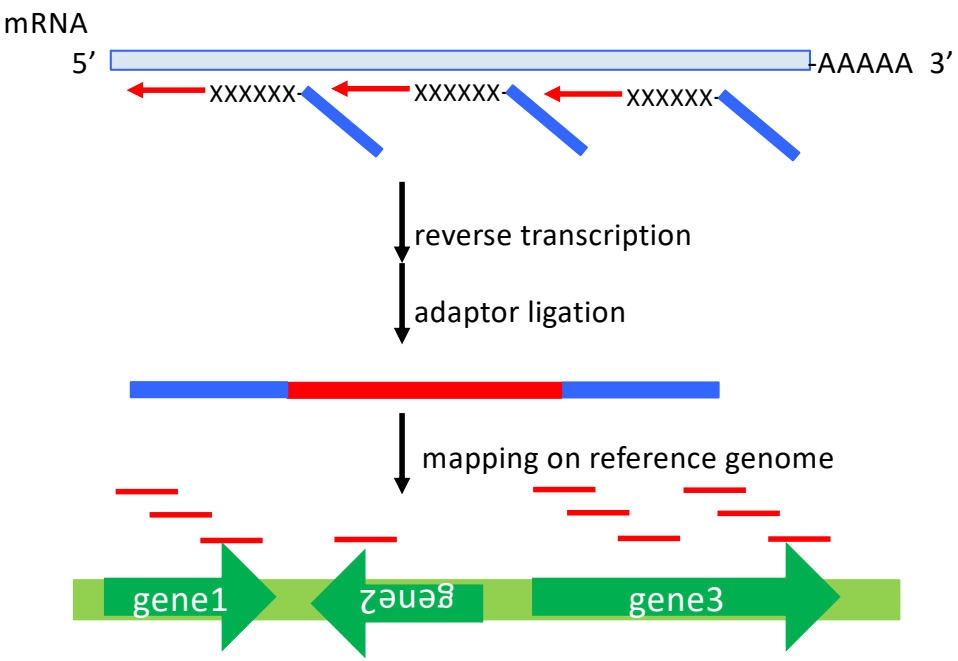
Evolutionary and Ecological Genomics

SNP calling

population structure,
history,
functional genes,
GWAS,
…

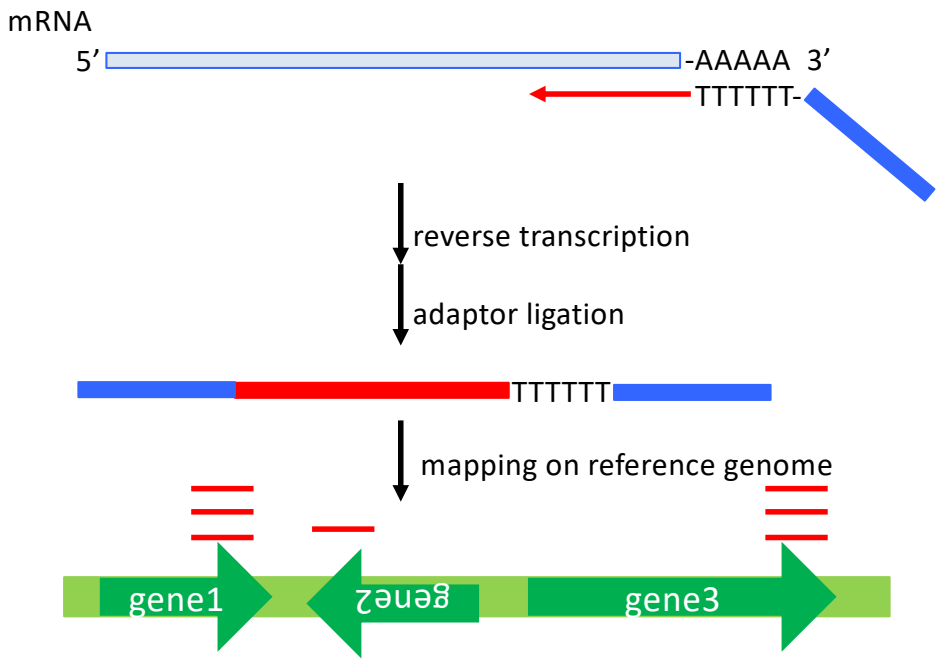
Transcriptomics
RNA seq (transcriptome/exome)

Type 1. random



- More fragments from longer gene (normalization needed)

Type 2. strand-specific



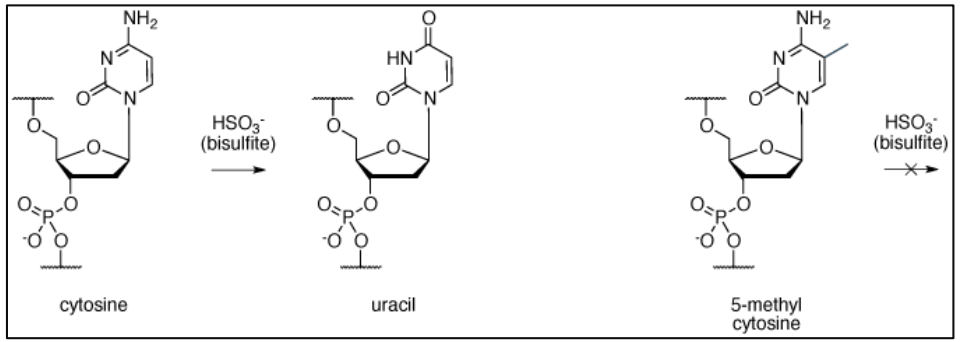
- 1 fragment from 1 mRNA (normalization not needed)
-mapped at 3' site of the gene

Units of RNA seq

RPKM	TPM
(Read Per Kilobase per Million mapped reads)	(Transcripts Per Million mapped reads)

Methylated DNA detection

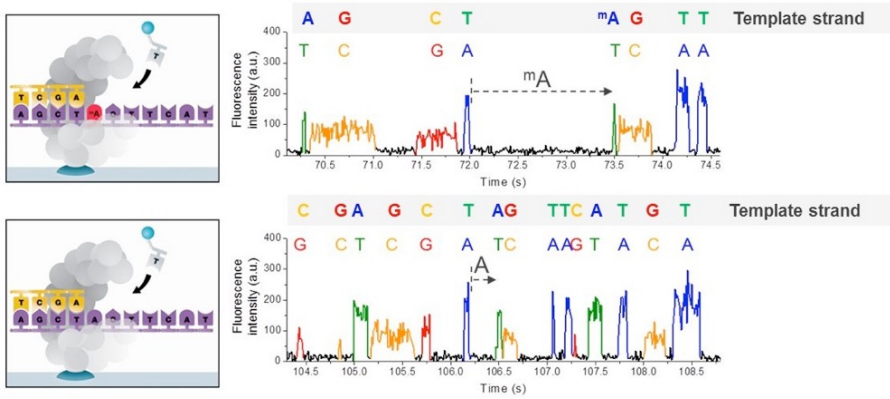
BS-DNA seq (illumina)
bisulfate sequencing (short read)



PacBio (long read)

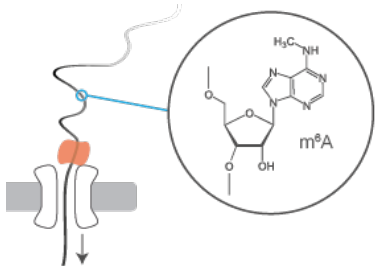
time lag of kinetics

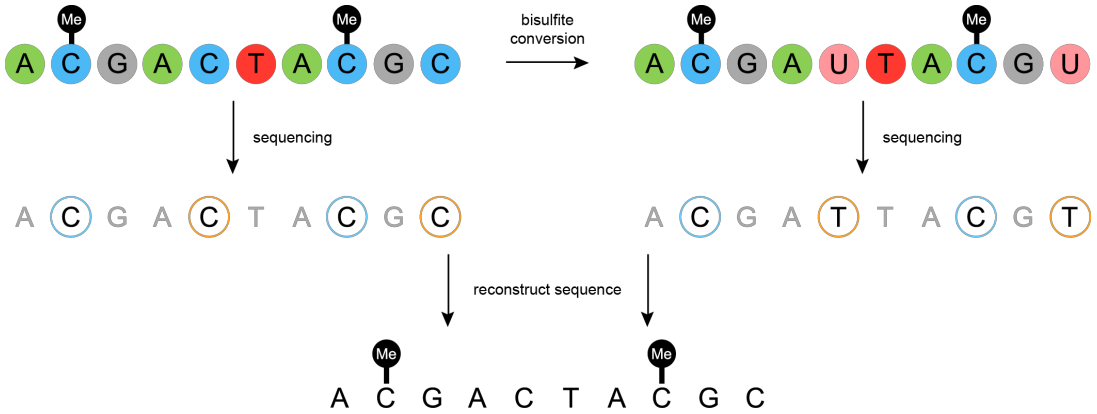
Example: N^6-methyladenine

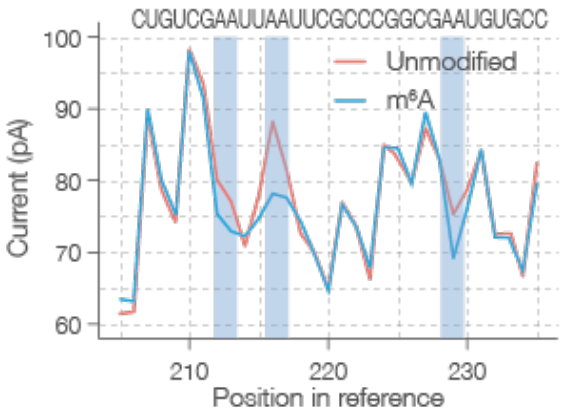


Nanopore (long read)

current difference
(accuracy?)



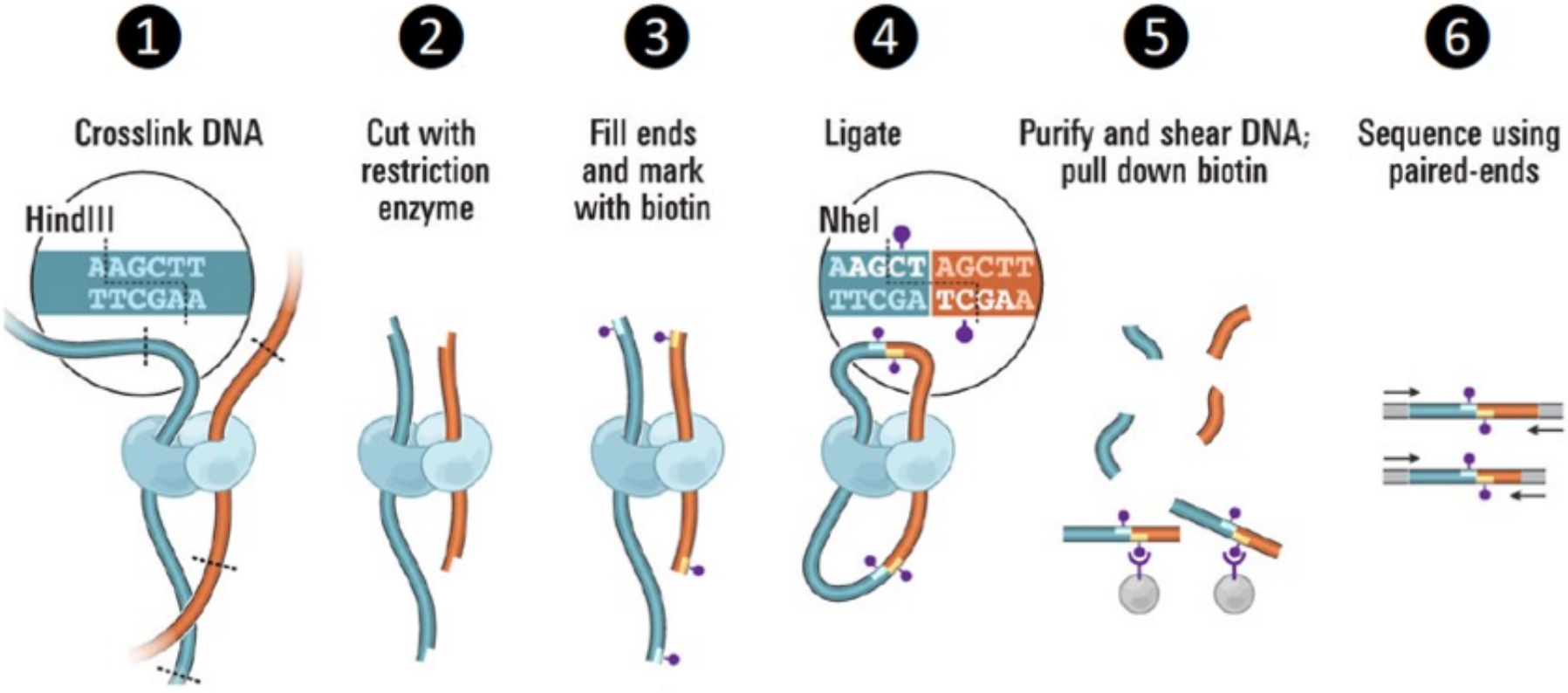




3-D genome architecture
Hi-C by Dovetail

Genome Assembly, chromatin conformation analysis
and structural variation detection

3D structure of nucleotides in cell nucleus?
Which sequences are located close to each other?



https://dovetailgenomics.com

3-D genome architecture

How DNA/chromosome is packed in a nucleus?

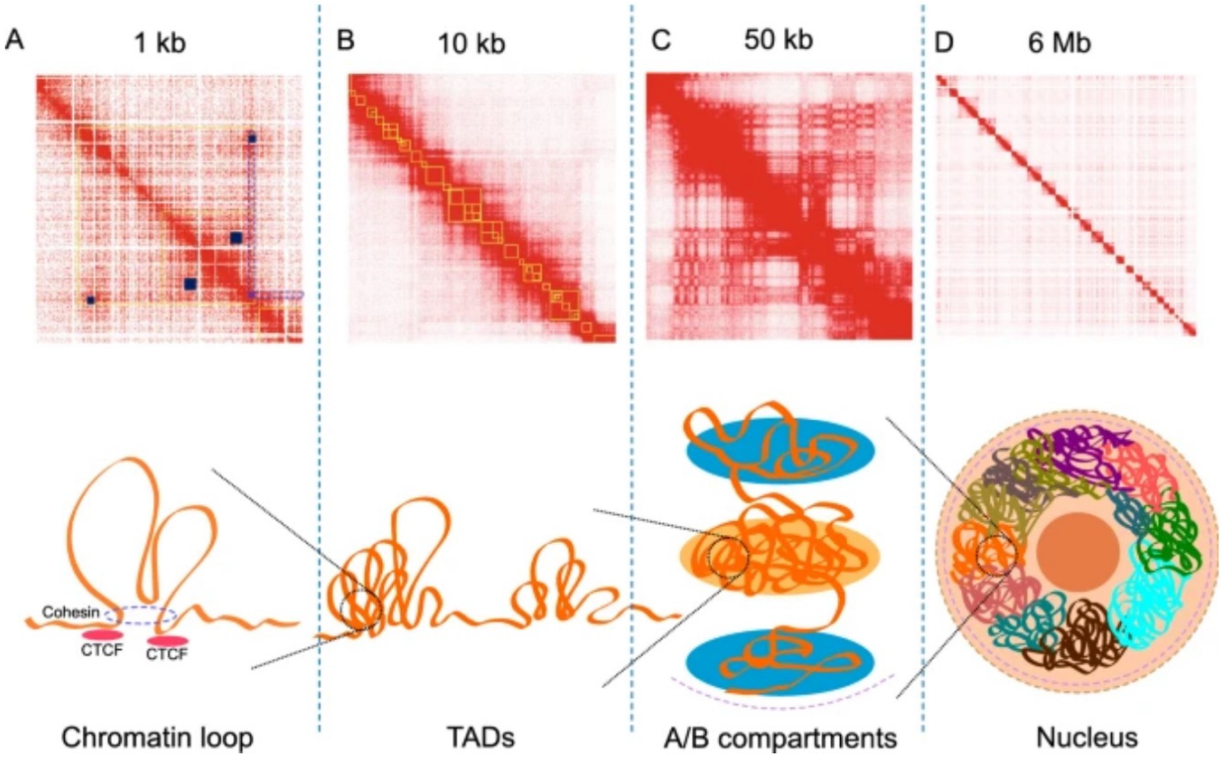


Illustration of genome architecture and the corresponding Hi-C interaction maps. Top panel: interaction heatmaps A, B, C, D are in different scales (kb or Mb per pixel) to correlate with the diagrams of 3D structures in the bottom panel, yellow boxes in A and B are identified TADs and small blue boxes in A indicate chromatin loops. The purple box in A is a frequently interacting region, with its classical "V" shape pattern coloured in purple dotted lines. Heatmaps were generated using Juicebox [29] with published Hi-C data of GM12878 [3]. Bottom panel: diagrams of 3D structures in the genome

3-D structure of genome in nucleus

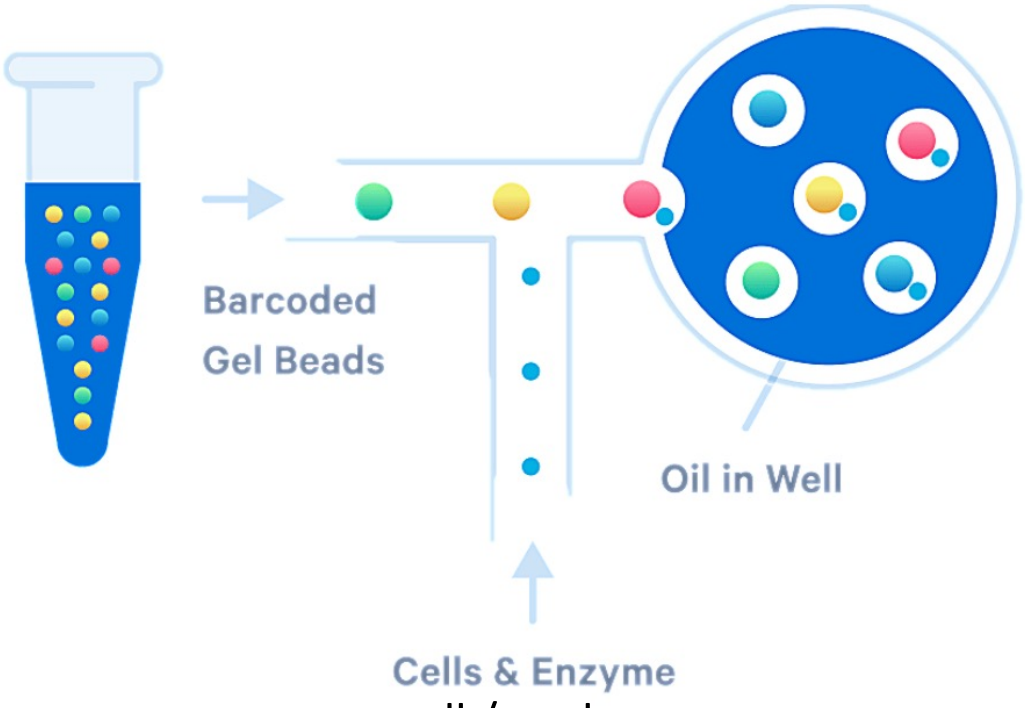
Utilized for genome assembly

Liu et al (2021) *Epigenetics & Chromatin*
https://doi.org/10.1186/s13072-021-00417-4

Single cell analysis

High-throughput library synthesis in microdroplet

transcriptome
epigenome
disease (cancer)



cell / nucleus

Droplet
accumulation/pooling

illumina sequencing

Chromium

Our Next GEM technology enables analysis of individual biological components at scale.

10X Genomics
https://www.10xgenomics.com/

Spatial transcriptomics

gene expression analysis
in consecutive cell layers

10X Genomics Visium

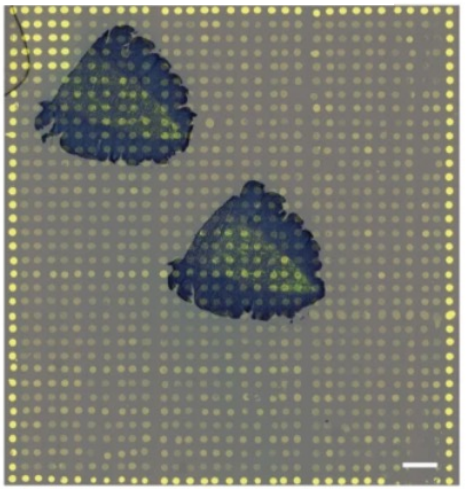
6mm x 6mm array

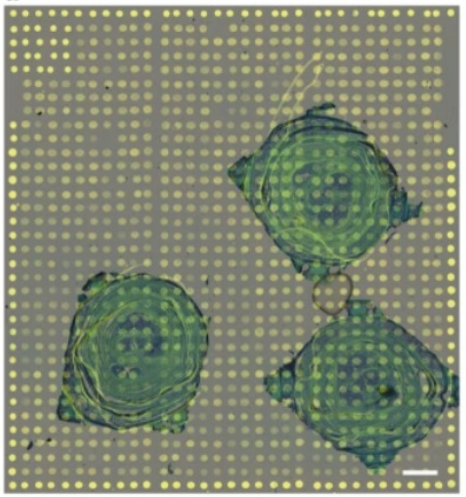
on-array-library synthesis

each spot has dT-oligo with individual index

minimum resolution (spot diameter) 55 µm

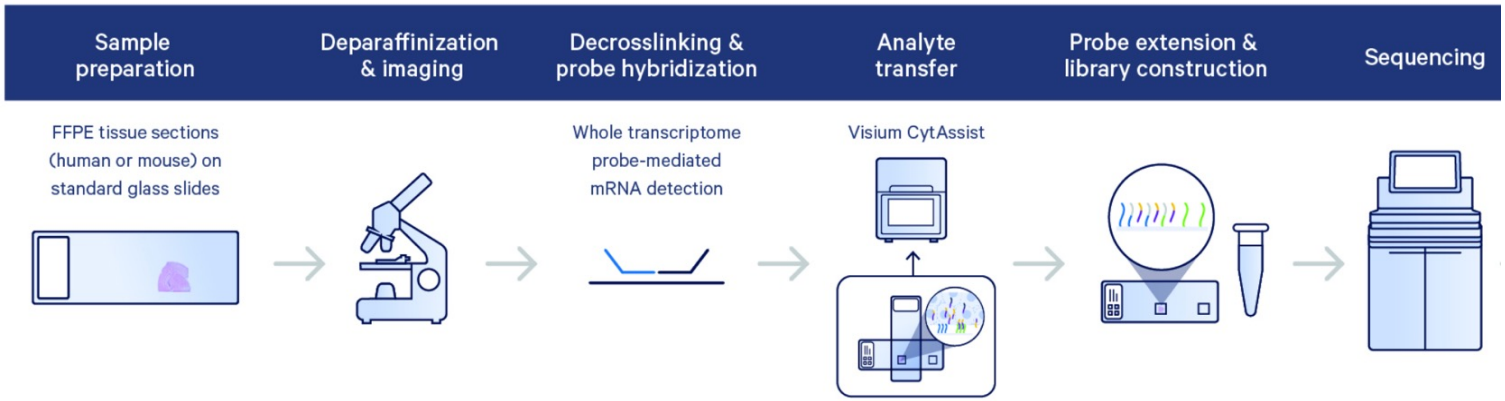
longitudinal section of
Picea abies female cones





cross-section of
Picea abies leaf bud

500um



https://www.nature.com/articles/s41596-018-0046-1

Community genomics

-**Metagenomics**

ex1. Microbe community in the soil

Extract total DNA from soil → sequencing (short or long)
→ 16S rRNA amplification and sequencing

ex2. Whole organism community in river/lake water

-**Metatranscriptomics (eRNA)**

ex. Seasonal transition of the transcriptome in Zurich lake

Extract total RNA from ecological sample (lake water)
→ sequencing (short)

Applications of long read sequence

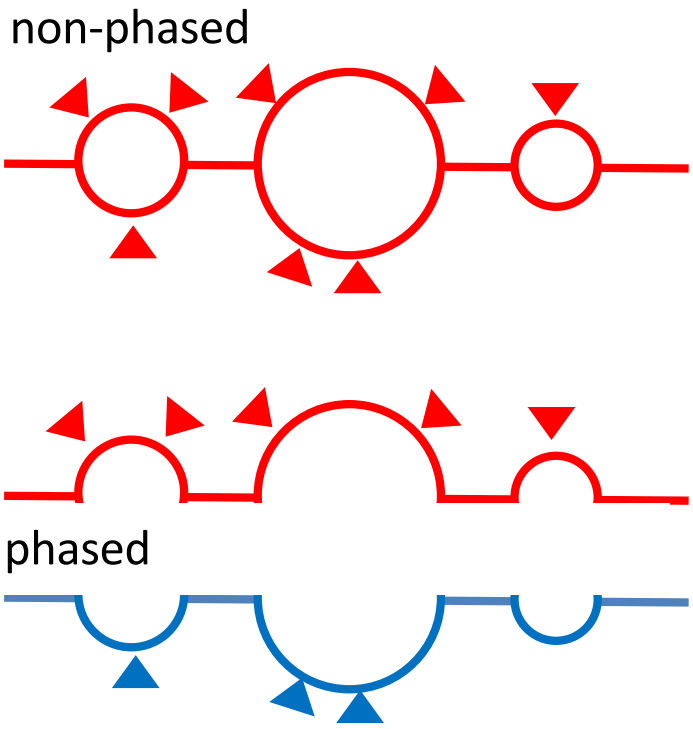
-*de novo* genome sequencing (HiFi)

-phased genome

two alleles of diploids, subgenomes of polyploid

-DNA methylation

-splice variant / isoform of mRNA



Other important terms in NGS

sequencing **coverage**

1 b : 1 nucleic acid base

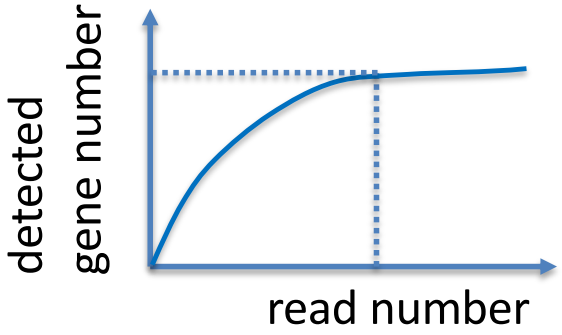
1 Kb = 1000 b
1 Mb = 1000 Kb = 10^6 b
1 Gb = 1000 Mb = 10^9 b
1 Tb = 1000 Gb = 10^{12} b

ex) species A with genome size 1 Gb

100 Gb **DNA** sequencing output = 100 x coverage

de novo genome sequencing 100x < coverage
SNP calling (re-sequencing) 10x < coverage

RNA sequencing (gene expression) 10M< **reads**



Chemical techniques of NGSs

you can learn more by yourself.

illumina https://www.youtube.com/watch?v=fCd6B5HRaZ8

PacBio https://www.youtube.com/watch?v=_lD8JyAbwEo

https://www.youtube.com/watch?v=NHCJ8PtYCFc

HiFi https://www.youtube.com/watch?v=DDbeyf1FEEU

Nanopore https://www.youtube.com/watch?v=CGWZvHIi3i0

Check-points to choose appropriate platform(s).

✅ Read length

-Depend on target (from microRNA to *de novo* genome assembly)

✅ Output data amount

-coverage (sequence depth) vs. cost

✅ Sequence accuracy

-Trade-off between accuracy and read length.

✅ Library type / index (barcode) attachment

✅ Combination of multiple platforms and/or methods

Functional Genomics Center Zurich

http://www.fgcz.ch/

Open user account → Apply project → Consultation with the team

Sample submission → Sequencing service → Bioinformatics support

Available equipment:
NGSs
Mass Spectrometers
Liquid Chromatography
Protein Sequencers
Amino Acid Analyzers
and more