

Biotechnology course

Sequencing Human genome sequencing

Dr. Stella Georgiou

sgeorgiou@bioacademy.gr

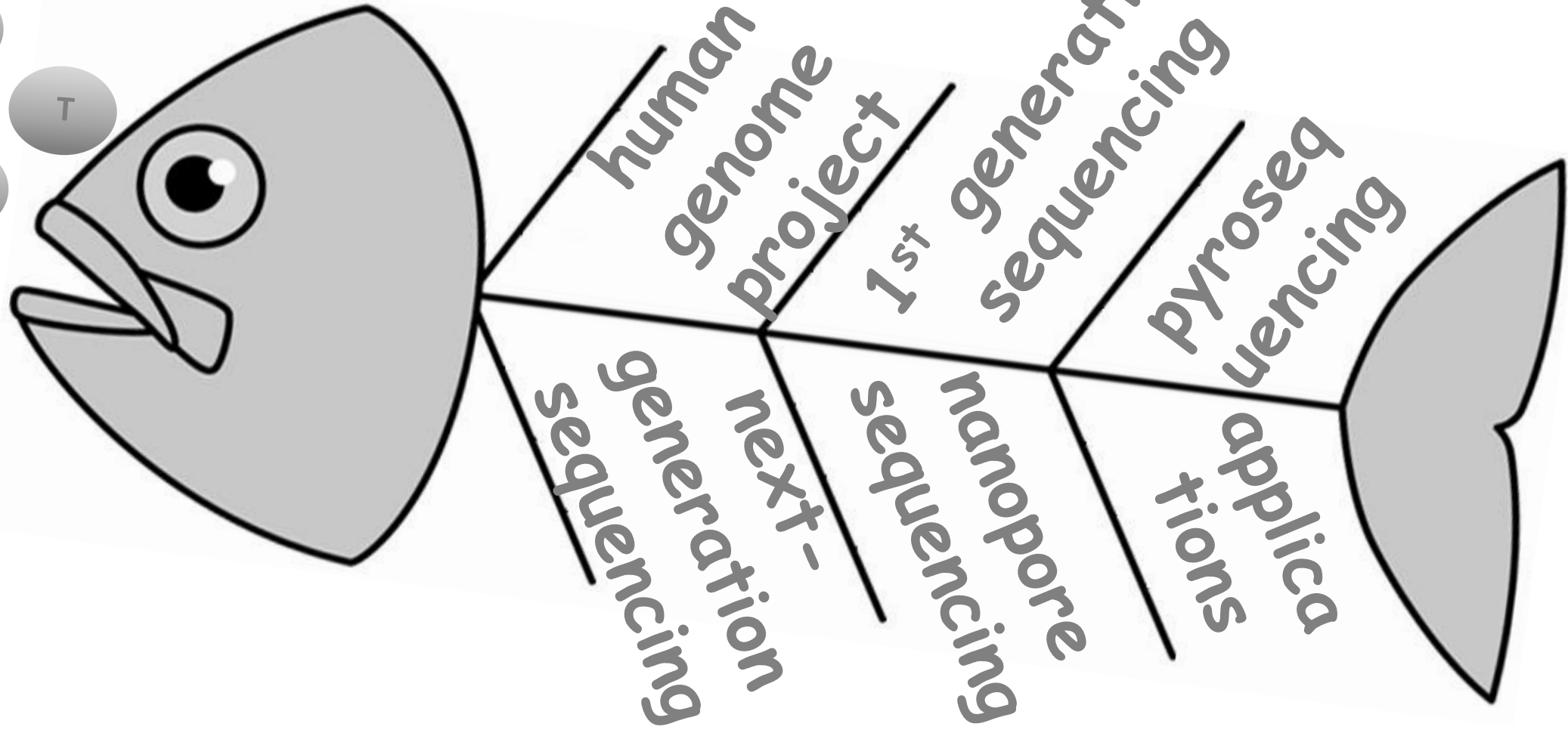
A

G

A

C

T



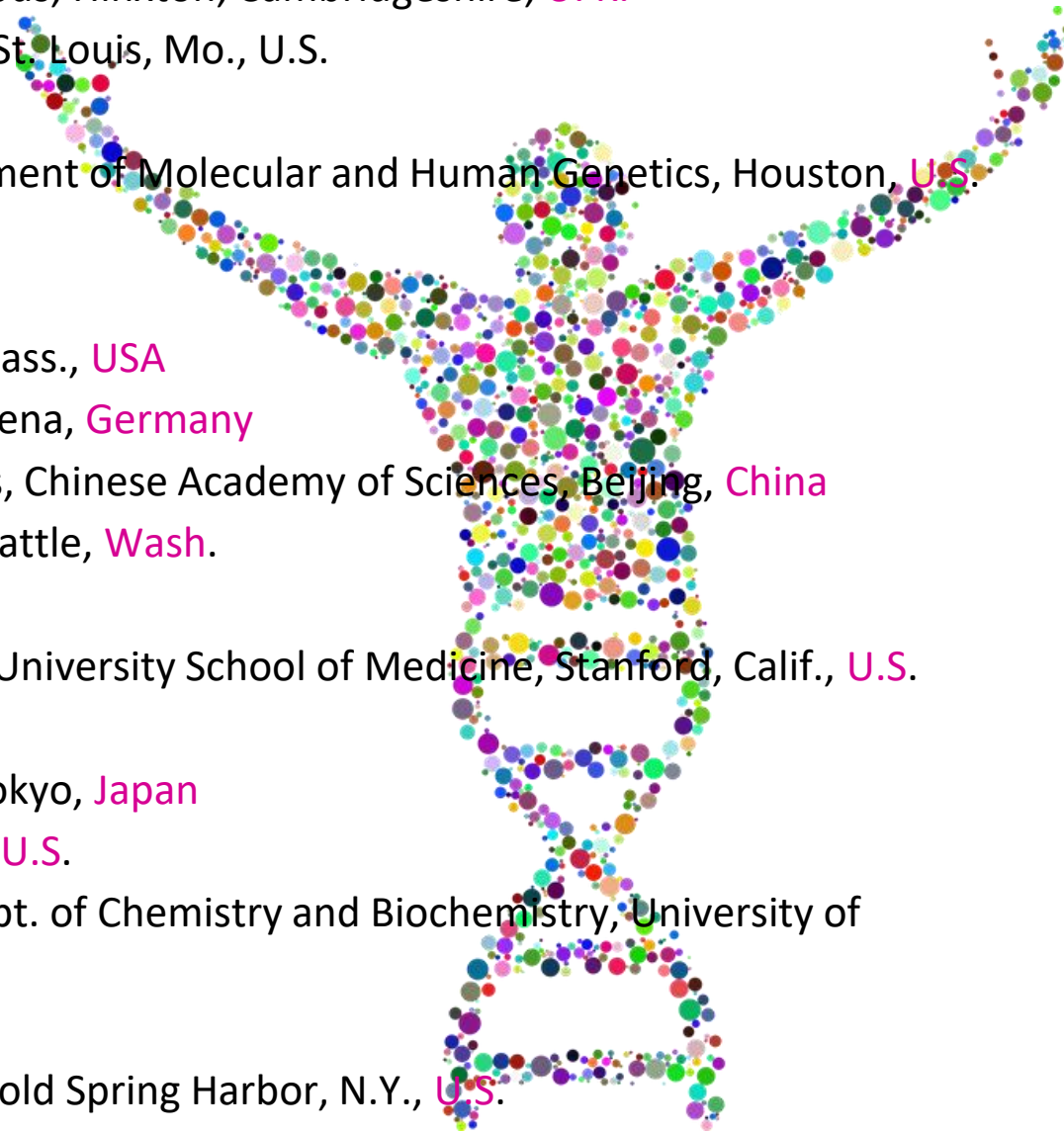


What is the Human Genome Project?

The Human Genome Project (HGP) was the international, collaborative research program whose goal was the complete mapping and understanding of all the genes of human beings. All our genes together are known as our "genome."

Who participated?

1. The Whitehead Institute/MIT Center for Genome Research, Cambridge, Mass., **U.S.**
2. The **Wellcome Trust Sanger Institute**, The Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, **U. K.**
3. **Washington University School of Medicine Genome Sequencing Center**, St. Louis, Mo., **U.S.**
4. **United States DOE Joint Genome Institute**, Walnut Creek, Calif., **U.S.**
5. **Baylor College of Medicine Human Genome Sequencing Center**, Department of Molecular and Human Genetics, Houston, **U.S.**
6. **RIKEN Genomic Sciences Center**, Yokohama, **Japan**
7. **Genoscope** and CNRS UMR-8030, Evry, **France**
8. **GTC Sequencing Center**, Genome Therapeutics Corporation, Waltham, Mass., **USA**
9. **Department of Genome Analysis**, Institute of Molecular Biotechnology, Jena, **Germany**
10. **Beijing Genomics Institute/Human Genome Center**, Institute of Genetics, Chinese Academy of Sciences, Beijing, **China**
11. **Multimegabase Sequencing Center**, The Institute for Systems Biology, Seattle, **Wash.**
12. **Stanford Genome Technology Center**, Stanford, Calif., **U.S.**
13. **Stanford Human Genome Center** and Department of Genetics, Stanford University School of Medicine, Stanford, Calif., **U.S.**
14. **University of Washington Genome Center**, Seattle, Wash., **U.S.**
15. **Department of Molecular Biology**, Keio University School of Medicine, Tokyo, **Japan**
16. **University of Texas Southwestern Medical Center at Dallas**, Dallas, Tex., **U.S.**
17. **University of Oklahoma's Advanced Center for Genome Technology**, Dept. of Chemistry and Biochemistry, University of Oklahoma, Norman, Okla., **U.S.**
18. **Max Planck Institute** for Molecular Genetics, Berlin, **Germany**
19. **Cold Spring Harbor Laboratory**, Lita Annenberg Hazen Genome Center, Cold Spring Harbor, N.Y., **U.S.**
20. **GBF - German Research Centre for Biotechnology**, Braunschweig, **Germany**



“The rate of progress is stunning. As costs continue to come down, we are entering a period where we are going to be able to get the complete catalog of disease genes. This will allow us to look at thousands of people and see the differences among them, to discover critical genes that cause cancer, autism, heart disease, or schizophrenia.”

Eric Lander, founding director of the Broad Institute of MIT and Harvard and principal leader of the Human Genome Project

Human genome project: probably about 20,500 human genes

detailed information about the structure, organization and function of the complete set of human genes

understand the blueprint for building a person

major impact in the fields of medicine, biotechnology, and the life sciences

human genome project

HGP researchers deciphered the human genome in three major ways:

1. determining the order, or "sequence," of all the bases in our genome's DNA;
2. Making physical maps that show the locations of genes on the chromosomes;
3. producing linkage maps, through which inherited traits (such as those for genetic disease) can be tracked over generations

This information can be thought of as the basic set of inheritable "instructions" for the development and function of a human being

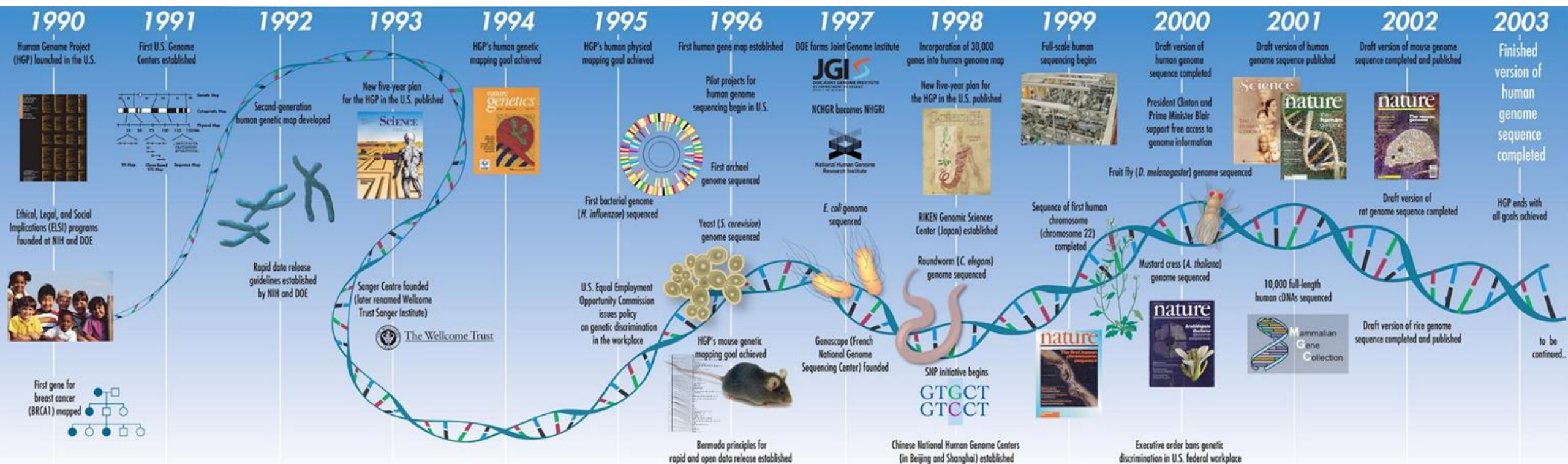
human genome project

The International Human Genome Sequencing Consortium published the first draft of the human genome in the journal Nature in February 2001 with the sequence of the entire genome's three billion base pairs some 90 percent complete.

A startling finding of this first draft was that the number of human genes appeared to be significantly fewer than previous estimates, which ranged from ***50,000 genes to as many as 140,000***. The full sequence was completed and published in **April 2003**.

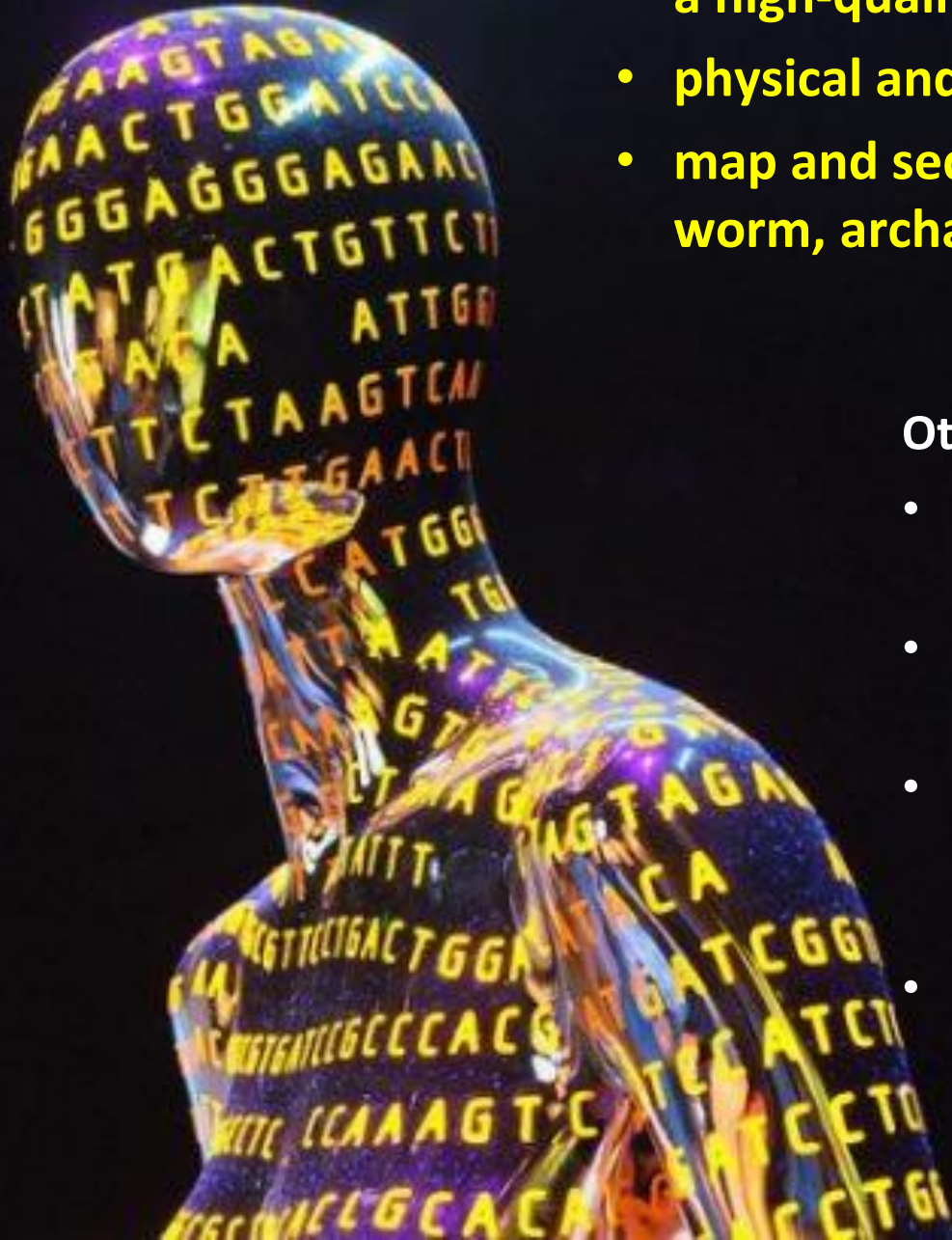
human genome project

Researchers have sequenced all **3.2 billion base pairs** in the human genome...



...within a span of only 13 years!

With **\$3 billion**



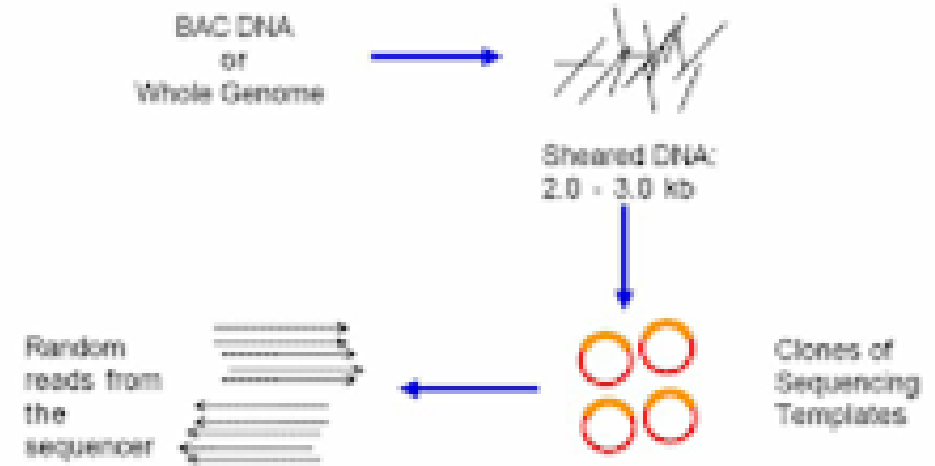
GOALS

- a high-quality version of the human sequence
- physical and genetic maps of the human genome
- map and sequence a set of five model organisms (mouse, worm, archaeon, yeast, plant)

Other outcomes

- an advanced draft of the mouse genome sequence (December 2002)
- an initial draft of the rat genome sequence (November 2002)
- the identification of more than 3 million human genetic variations, called single nucleotide polymorphisms (SNPs)
- the generation of full-length complementary DNAs (cDNAs) for more than 70 percent of known human and mouse genes

human genome project: 3 phases



1. Obtaining a DNA clone to sequence

2. Sequencing the DNA clone using the **SHOTGUN** approach

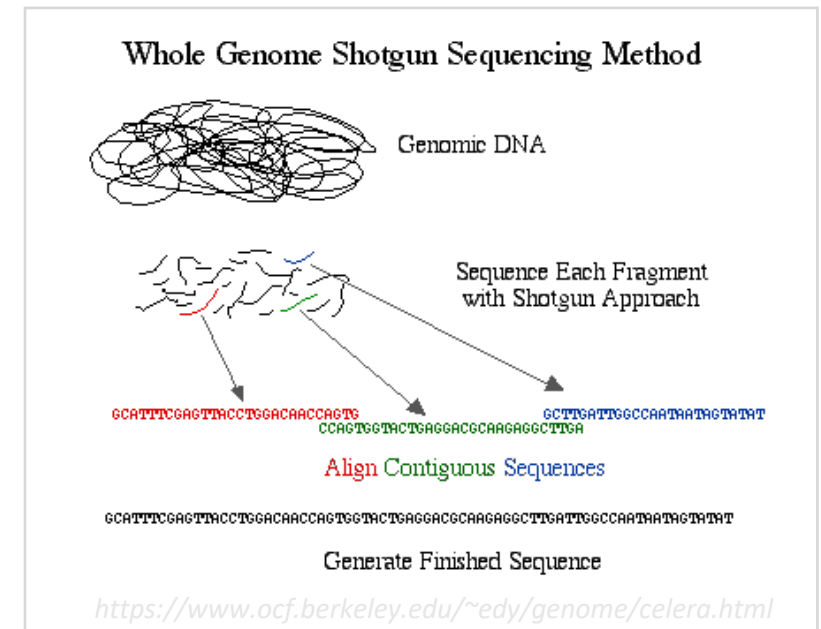
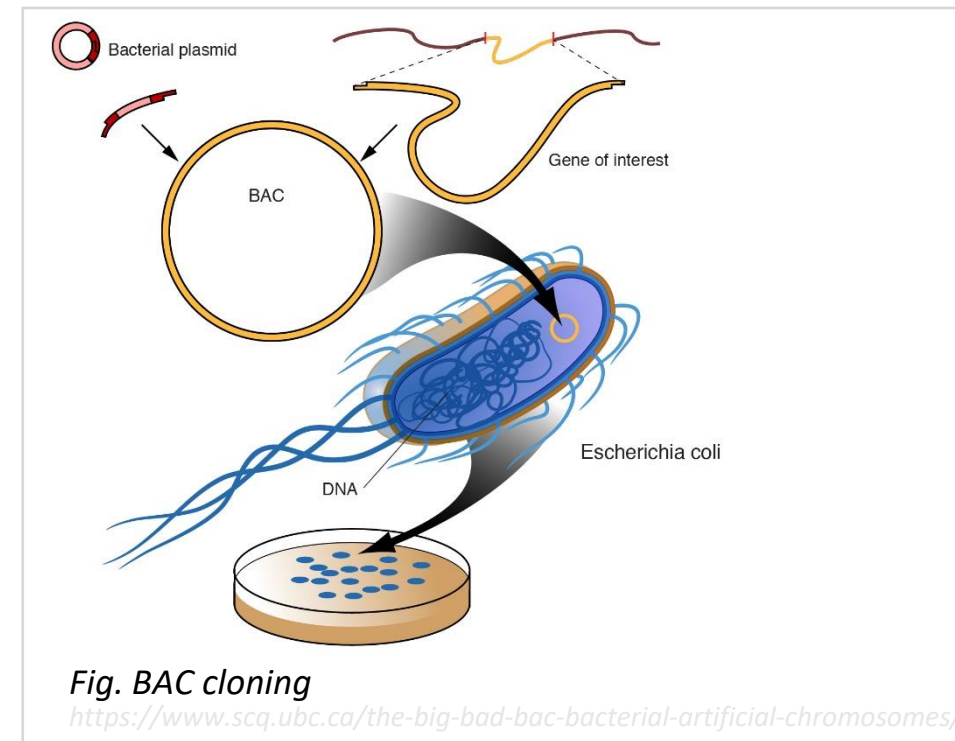
3. Assembling sequence data from multiple clones to determine overlap and establish a contiguous sequence

SHOTGUN sequencing

Individual BAC clones selected for DNA sequence analysis were further fragmented, and the smaller genomic DNA fragments were subcloned into vectors to generate a BAC-derived shotgun library.

The inserts were sequenced using primers matching the vector sequence flanking the genomic DNA insert, and overlapping shotgun clones were used to generate a DNA sequence spanning the entire BAC clone.

The term "shotgun" comes from the fact that the original BAC clone was randomly fragmented and sequenced, and the raw DNA sequence data was then subjected to computational analyses to generate an ordered set of DNA sequences that spanned the BAC clone.



whole genome assembly

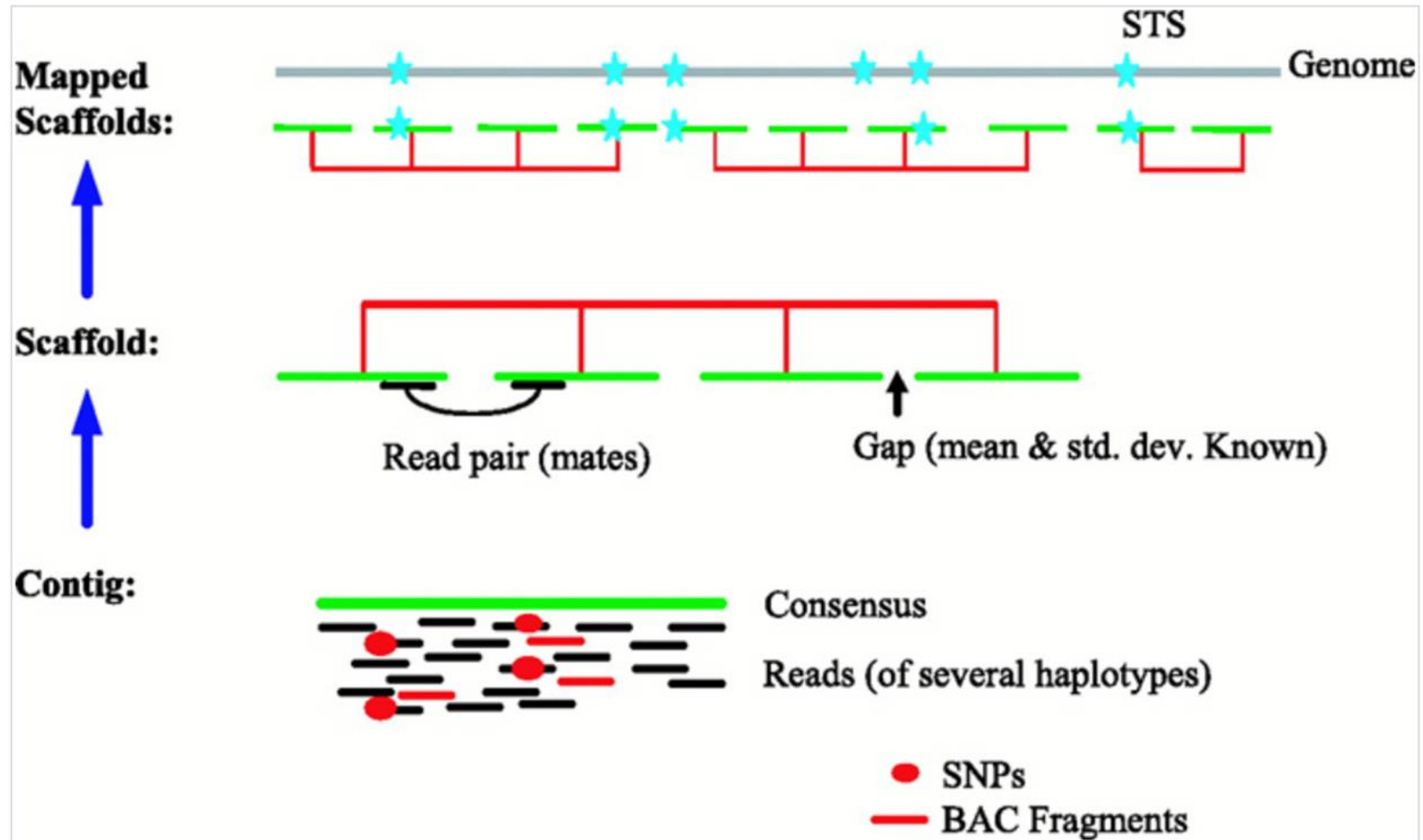



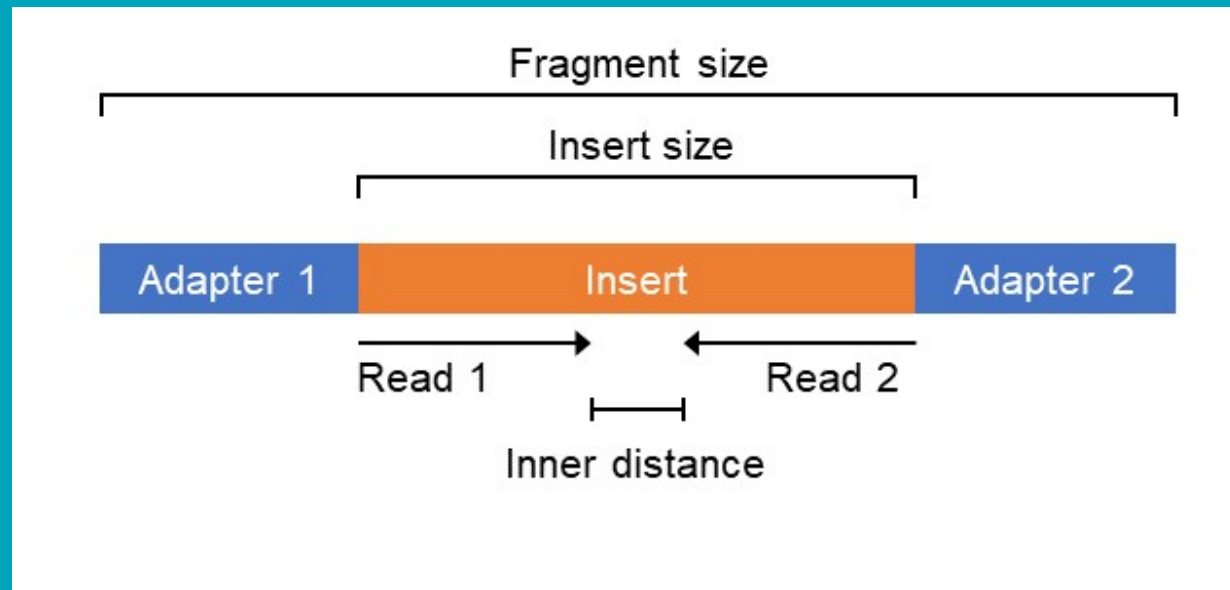
Figure 5: Anatomy of whole-genome assembly.

In whole-genome assembly, the BAC fragments (red line segments) and the reads from five individuals (black line segments) are combined to produce a contig and a consensus sequence (green line). The contigs are connected into scaffolds, shown in red, by pairing end sequences, which are also called mates. If there is a gap between consecutive contigs, it has a known size. Next, the scaffolds are mapped to the genome (gray line) using sequence tagged site (STS) information, represented by blue stars.

© 2001 **American Association for the Advancement of Science** Venter, C. *et al.* The sequence of the human genome. *Science* 291, 1304–1351 (2001). All rights reserved. 

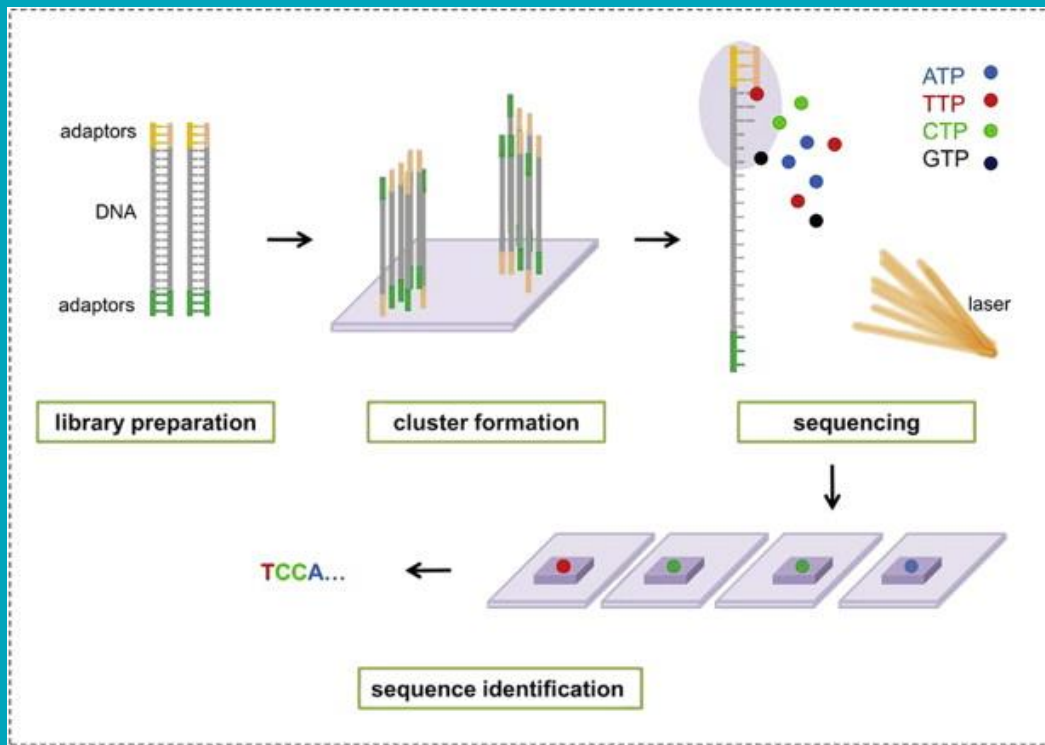
Glossary

adapters: The oligos bound to the 5' and 3' end of each DNA fragment in a sequencing library. The adapters are complementary to the lawn of oligos present on the surface of Illumina sequencing flow cells.



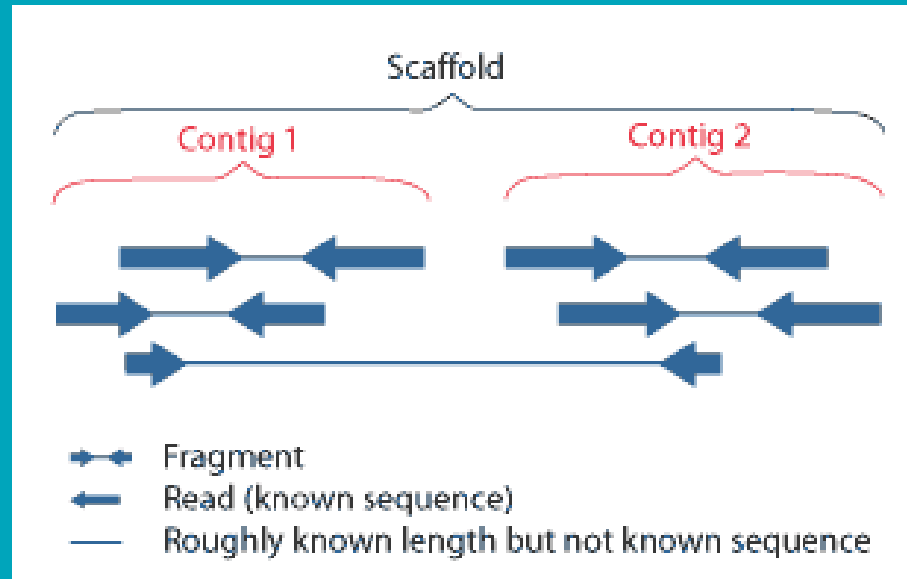


clusters: A clonal grouping of template DNA bound to the surface of a flow cell. Each cluster is seeded by a single template DNA strand and is clonally amplified through bridge amplification until the cluster has ~1000 copies. Each cluster on the flow cell produces a single sequencing read. For example, 10,000 clusters on the flow cell would produce 10,000 single reads and 20,000 paired-end reads.



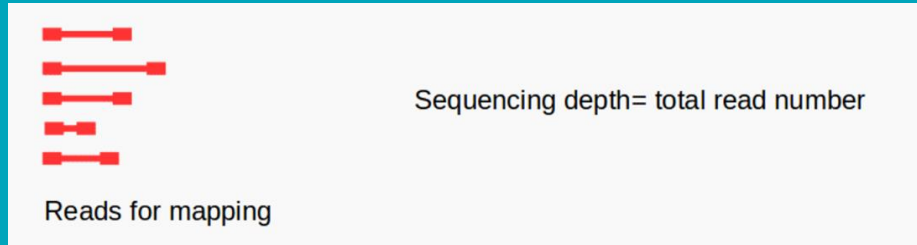
Sample multiplexing, also known as multiplex sequencing, allows **large numbers of libraries to be pooled and sequenced simultaneously during a single run** on Illumina instruments. Sample multiplexing is useful when targeting specific genomic regions or working with smaller genomes.

contigs: A stretch of continuous sequence, *in silico*, generated by aligning overlapping sequencing reads.



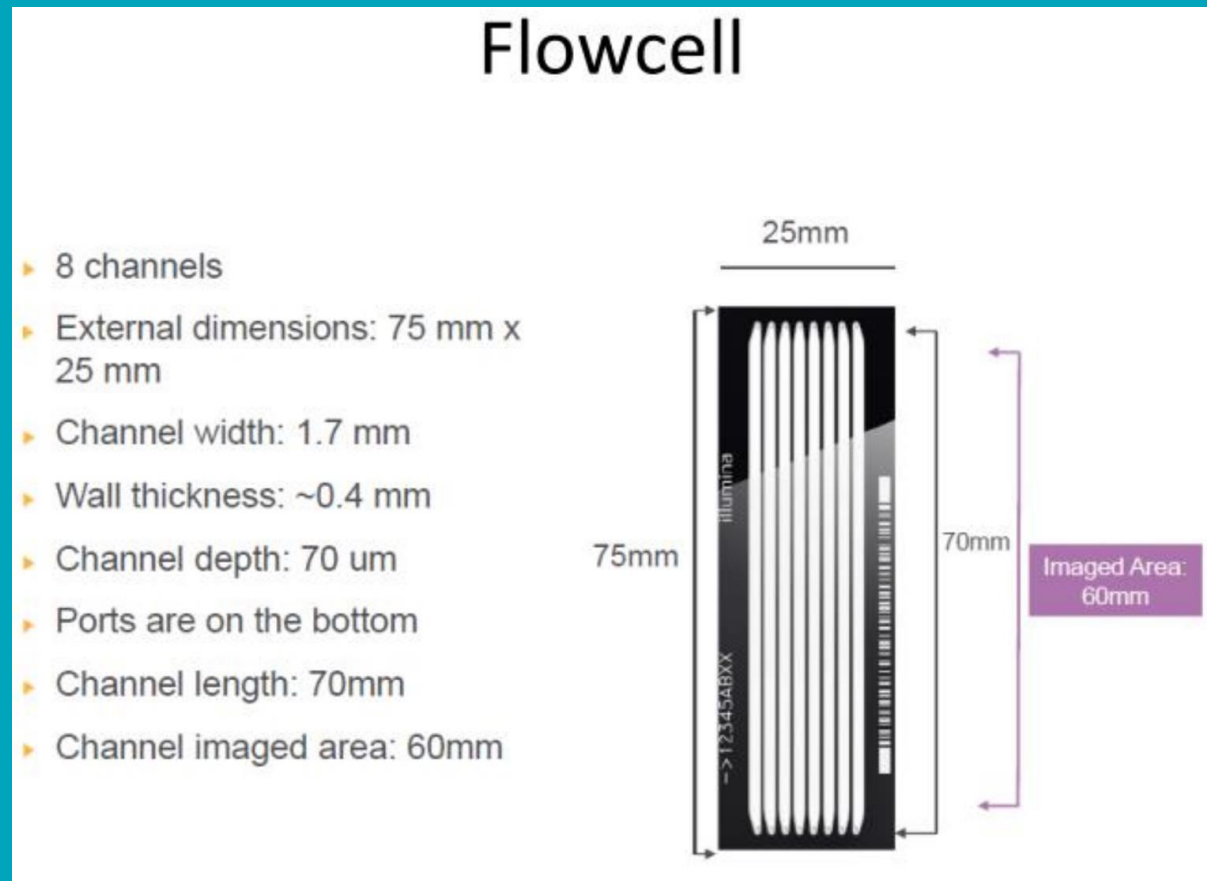


coverage level: The average number of sequenced bases that align to each base of the reference DNA. For example, a whole genome sequenced at 30x coverage means that, on average, each base in the genome was sequenced 30 times.





The flow cell is the where the sequencing chemistry occurs. The flow cell is a glass slide containing small fluidic channels, through which polymerases, dNTPs and buffers can be pumped. The glass inside the channels is decorated with short oligonucleotides complementary to the adapter sequences.

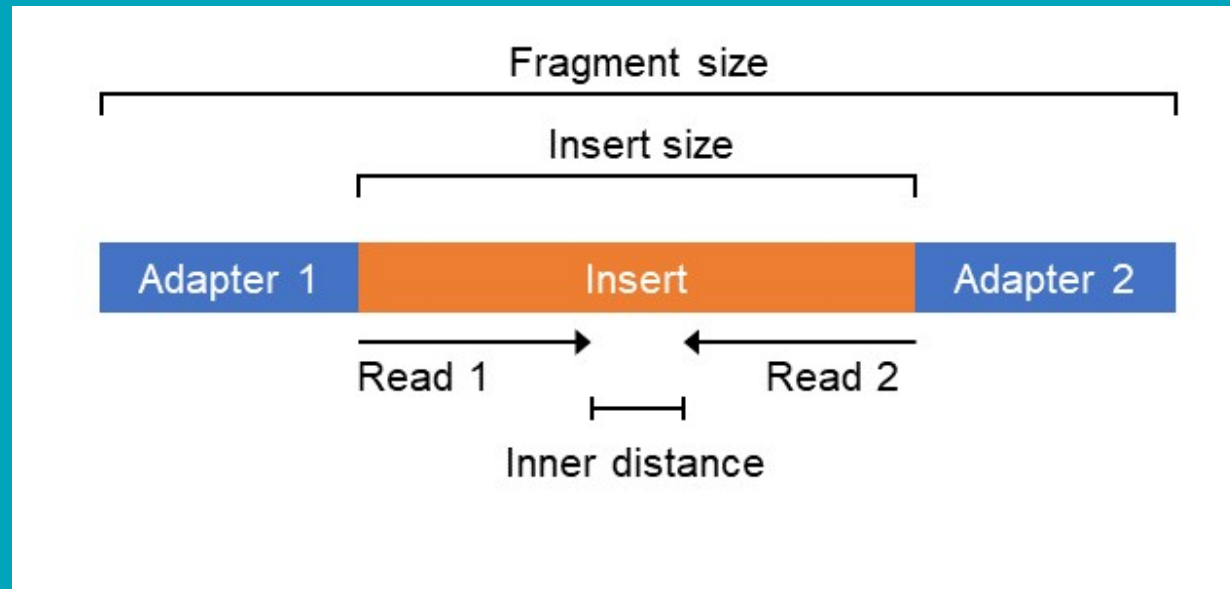




indexes/barcodes/tags: A unique DNA sequence ligated to fragments within a sequencing library for downstream *in silico* sorting and identification. Indexes are typically a component of adapters or PCR primers and are ligated to the library fragments during the sequencing library preparation stage. Illumina indexes are typically between 8–12 bp. Libraries with unique indexes can be pooled together, loaded into one lane of a sequencing flow cell, and sequenced in the same run. Reads are later identified and sorted via bioinformatic software. All together, this process is known as “multiplexing.”

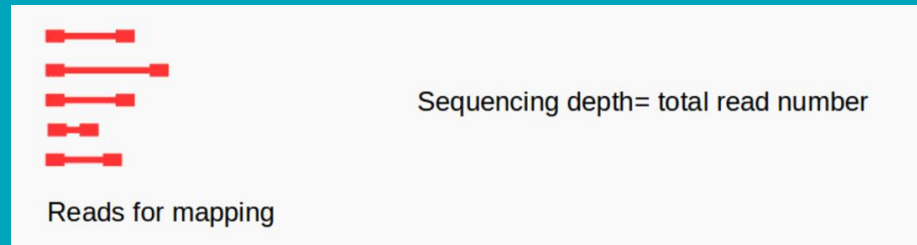


insert: During the library preparation stage, the sample DNA is fragmented, and the fragments of a specific size (typically 200–500 bp, but can be larger) are ligated or “inserted” in between two oligo adapters. The original sample DNA fragments are also referred to as “inserts.”



Glossary

read: NGS uses sophisticated instruments to determine the nucleotide sequence of a DNA or RNA sample. In general terms, a sequence “read” refers to the data string of A, T, C, and G bases corresponding to the sample DNA or RNA. With Illumina technology, millions of reads are generated in a single sequencing run.




Single-read sequencing involves sequencing DNA from only one end. This solution delivers large volumes of high-quality data, rapidly and economically.

Paired-end sequencing allows users to sequence both ends of a fragment and generate high-quality, alignable sequence data.



reference genome: A reference genome is a fully sequenced and assembled genome that acts as a scaffold against which new sequence reads are aligned and compared. Typically, reads generated from a sequencing run are aligned to a reference genome as a first step in data analysis.



RefSeq: NCBI Reference Sequence Database

A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein.

Using RefSeq

- [About RefSeq](#)
- [Human Reference Genome](#)
- [Prokaryotic RefSeq Genomes](#)
- [FAQ](#)
- [NCBI Handbook](#)
- [Factsheet](#)

RefSeq Access

- [Human Genome Resources and Download](#)
- [RefSeq FTP](#)
- [RefSeq genomes FTP](#)
- [New RefSeq genomic \(last 30 days\)](#)
- [New RefSeq transcripts \(last 30 days\)](#)
- [New RefSeq proteins \(last 30 days\)](#)
- [Searching for RefSeq records \(Queries\)](#)

RefSeq projects

- [Consensus CDS \(CCDS\)](#)
- [RefSeq Functional Elements](#)
- [RefSeqGene](#)
- [Targeted Loci](#)
- [Virus Variation](#)
- [RefSeq Select](#)
- [MANE](#)

Announcements

September 17, 2020
RefSeq Release 202 is available for FTP

This release includes:

Proteins:	186,755,483
Transcripts:	33,077,068
Organisms:	104,969

Available at: <ftp://ftp.ncbi.nlm.nih.gov/refseq/release/>
Documentation: [Release Notes](#)

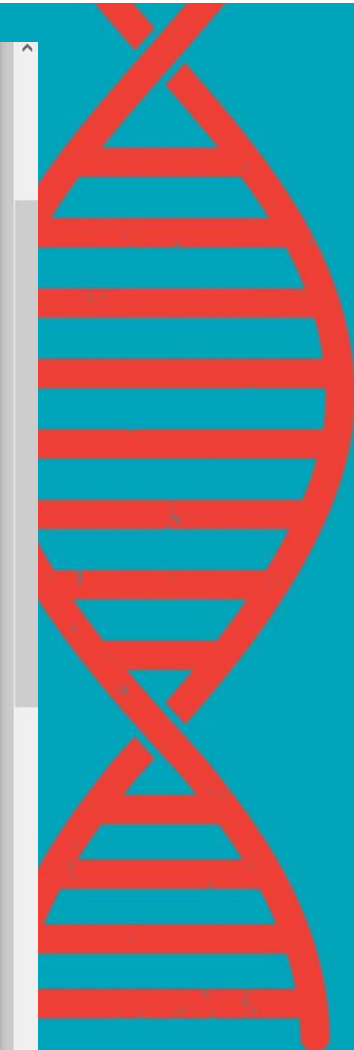
See [previous announcements](#), follow [NCBI on Twitter](#), or subscribe to [NCBI's refseq-announce mail list](#) to receive announcements

Related Links

- [Assembly](#)
- [Gene](#)
- [Genome](#)
- [Genome Data Viewer](#)
- [Annotated Eukaryotic Genomes](#)

Feedback & Credits

- [Publications and Citing RefSeq](#)
- [Contact RefSeq Help Desk](#)
- [Contact CCDS Help Desk](#)
- [Submit a GeneRIF](#)
- [Collaborators](#)





Library: The preparation of the sequencing library is the very first step in any sequencing analysis. A sequencing library can be made by starting from genomic **DNA** or from **RNA**.

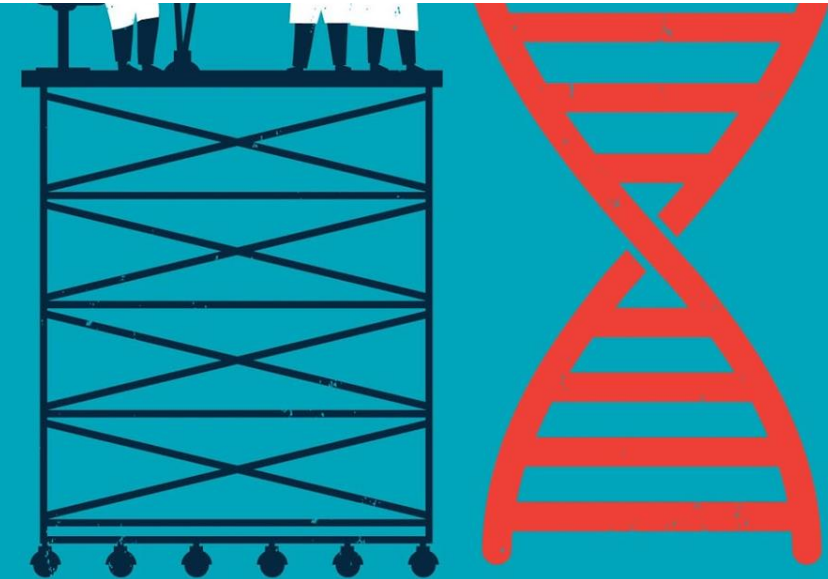
It allows DNA or RNA to adhere to the sequencing flowcell and allows the sample to be identified.

The workflow for the preparation of a DNA sequencing library consists of three fundamental steps:

1. **Fragmentation and sizing** of the nucleic acid (DNA or RNA) to obtain fragments of a predefined length
2. **Attachment of the adaptors (adapters)** to the extremities of the fragments
3. **Library quantification**

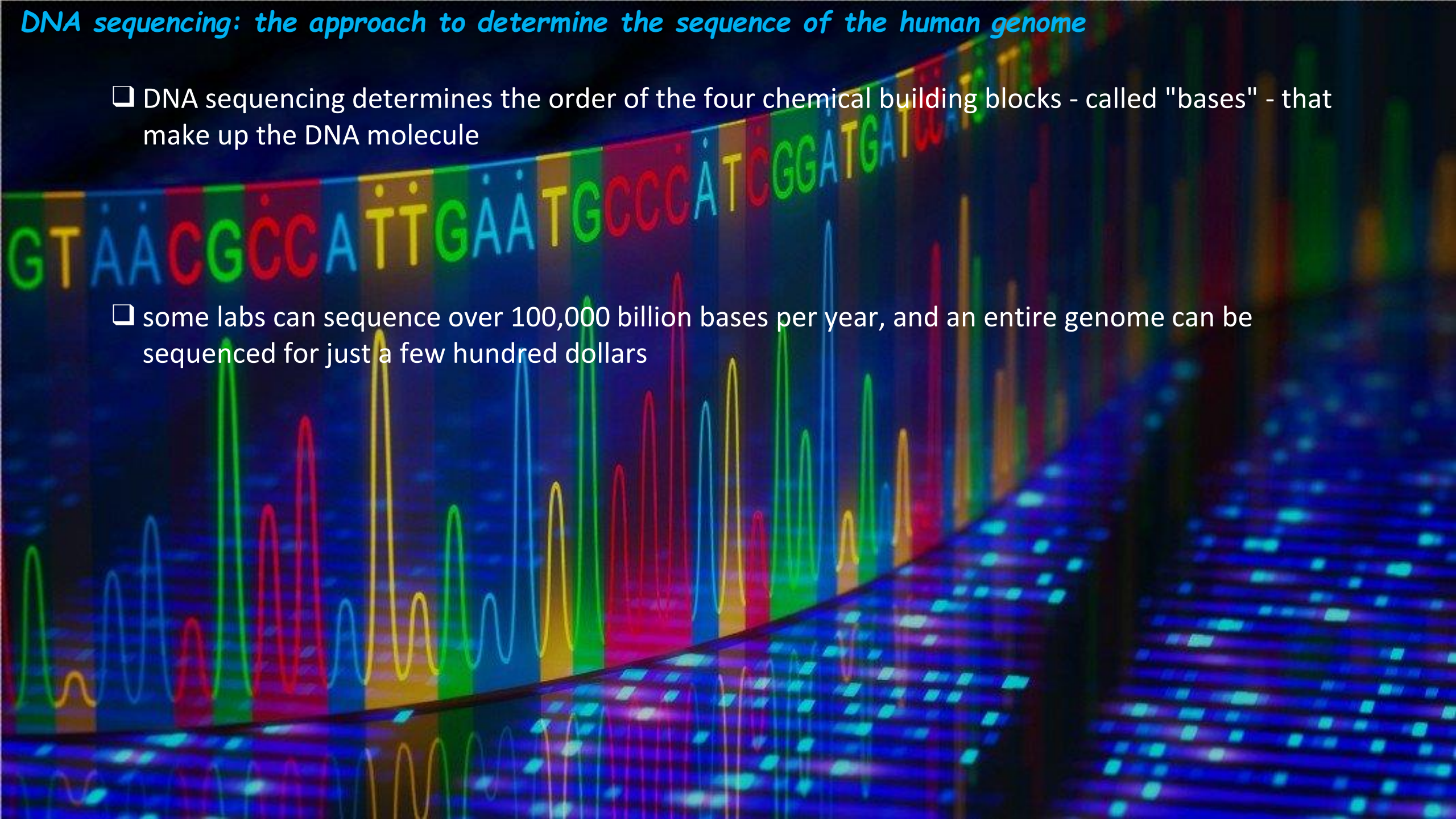
In any RNA sequencing library there's an additional step: the **RNA conversion in cDNA**. The fragmentation step can be done before or after the cDNA synthesis.

A sequencing library is, by definition, a pool of DNA fragments with adapters attached.

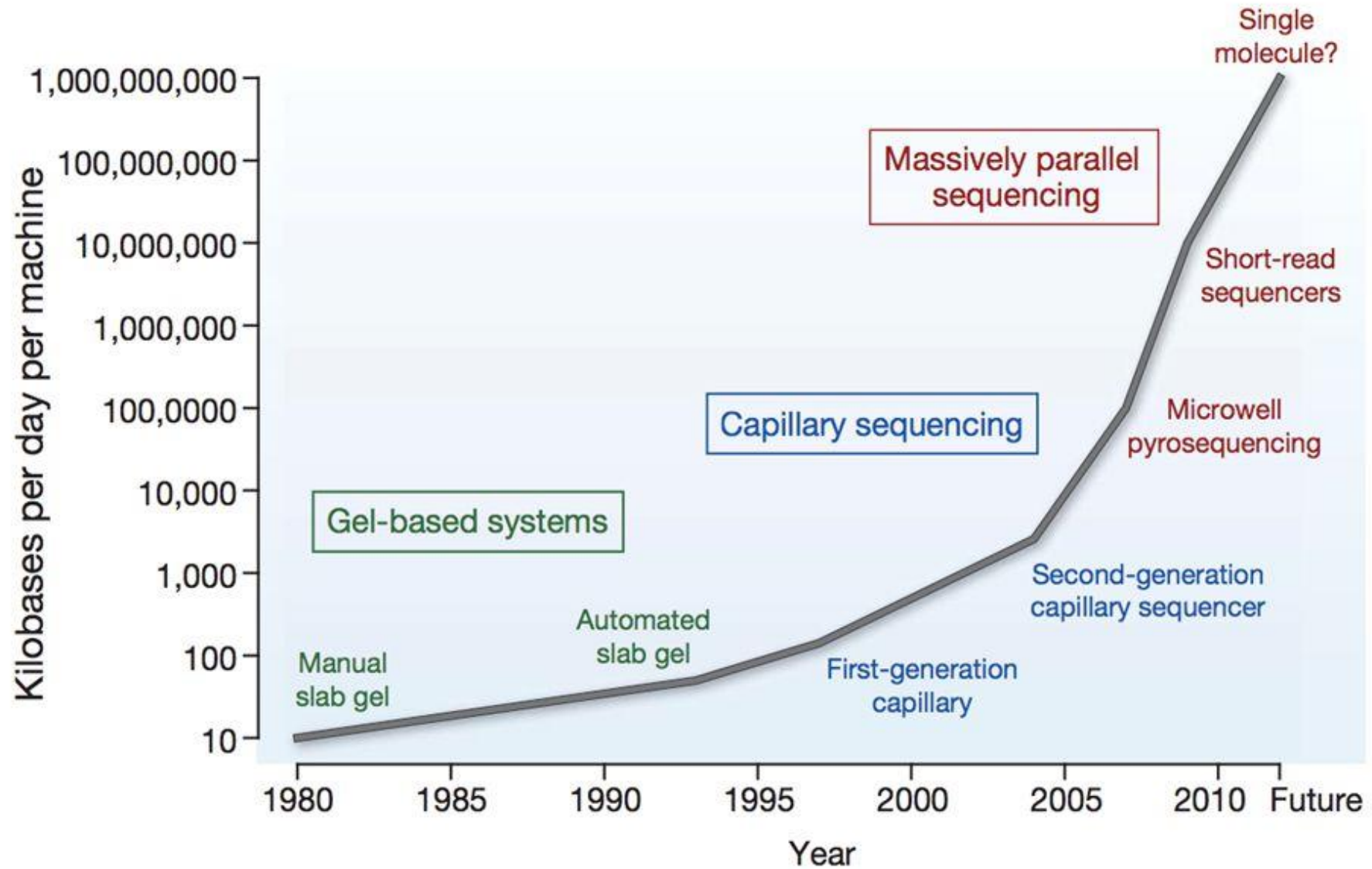


DNA sequencing: the approach to determine the sequence of the human genome

- DNA sequencing determines the order of the four chemical building blocks - called "bases" - that make up the DNA molecule
- some labs can sequence over 100,000 billion bases per year, and an entire genome can be sequenced for just a few hundred dollars



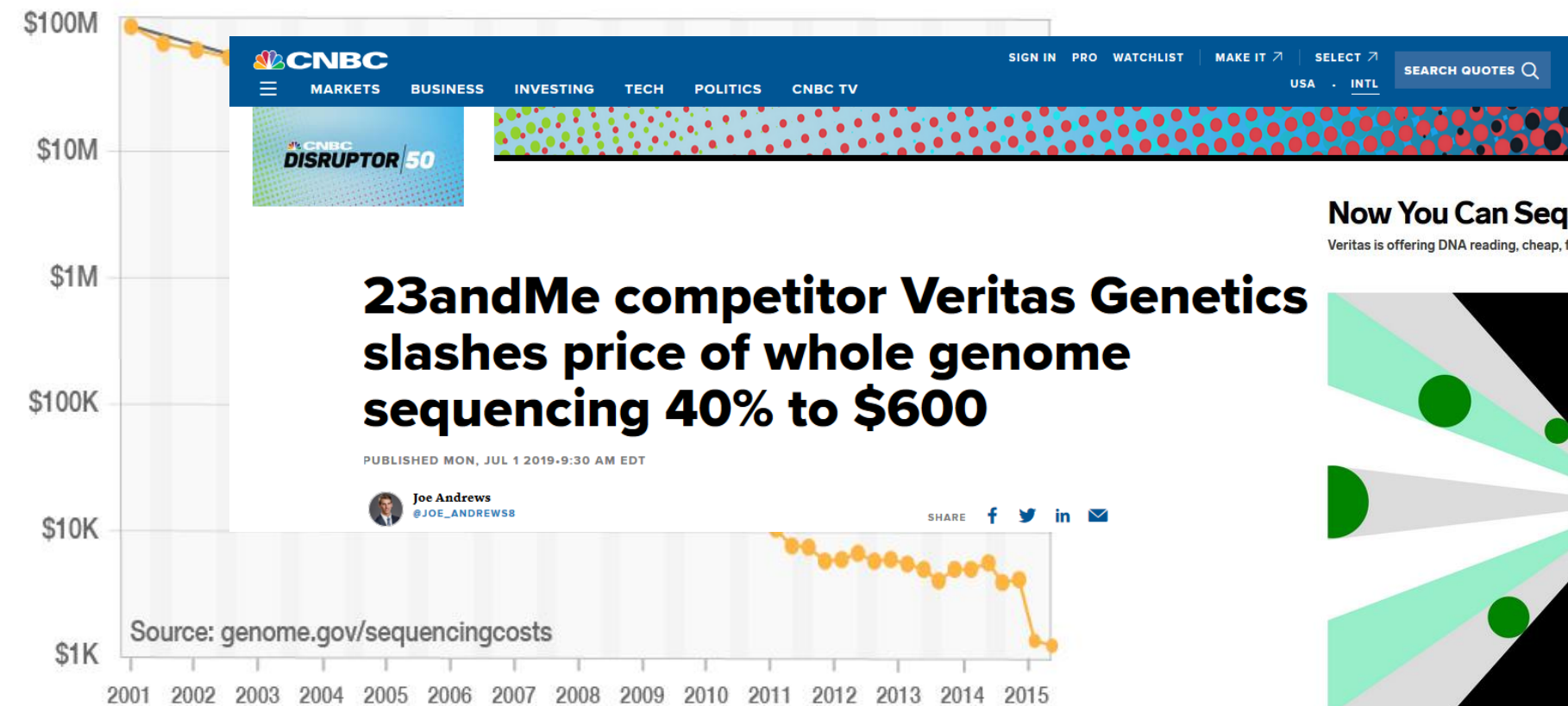
The History of DNA Sequencing Technology



cost vs time vs length during the years

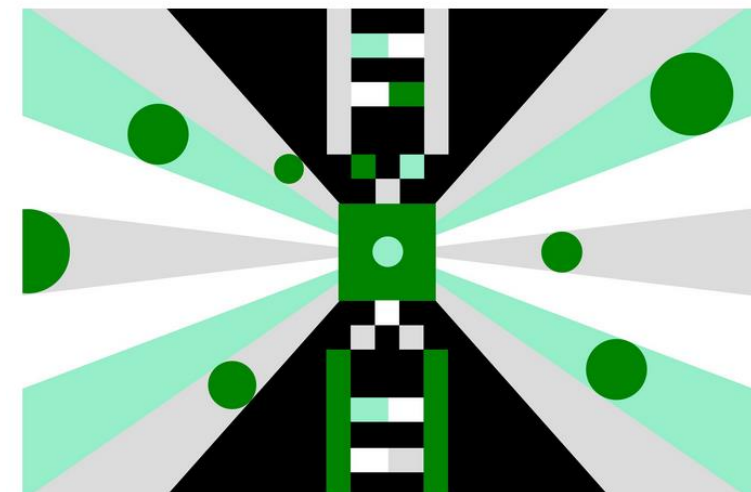
The following 'sequence coverage' values were used in calculating the cost per genome:

- Sanger-based sequencing (average read length=500-600 bases): 6-fold coverage
- 454 sequencing (average read length=300-400 bases): 10-fold coverage
- Illumina and SOLiD sequencing (average read length=75-150 bases): 30-fold coverage



Now You Can Sequence Your Whole Genome for Just \$200

Veritas is offering DNA reading, cheap, for two days. But most consumers don't understand the difference between that and a 23andMe test.

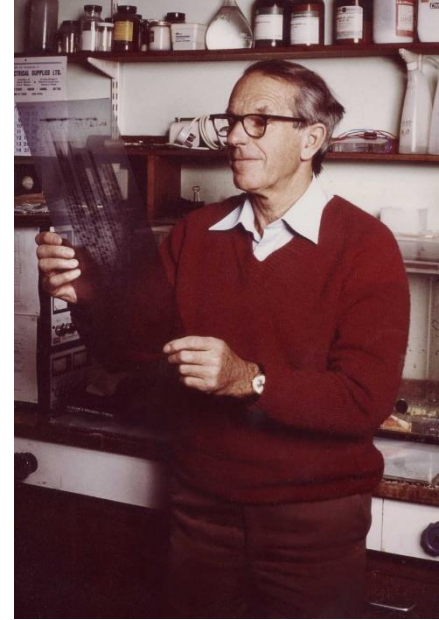


Sanger sequencing- "the chain termination method"

"... [A] knowledge of sequences could contribute much to our understanding of living matter."

[Frederick Sanger [1]]

[1] F. Sanger, Frederick Sanger – Biographical, 1980 (URL http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1980/sanger-bio.html).



- Fred Sanger won his second Nobel prize for the invention of Sanger sequencing in 1977
- Sanger sequencing (or sequencing by synthesis, SBS) was the main technology used to sequence genomic data until the mid 2000's when the technology was replaced by second-generation generation sequencing technologies

Why Sanger seq was that novel?

DNA polymerase naturally uses dNTPs to synthesize a new strand. The synthesis process occurs very quickly, making it hard to make any sort of measurement during synthesis.

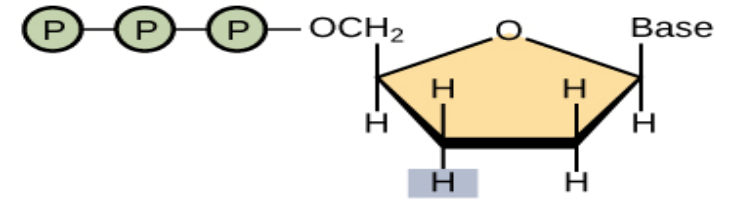
Sanger overcame this problem by figuring out a way to terminate synthesis using a modified version of dNTPs called *ddNTPs* (dideoxynucleotide triphosphates).

Sanger sequencing- "the chain termination method"

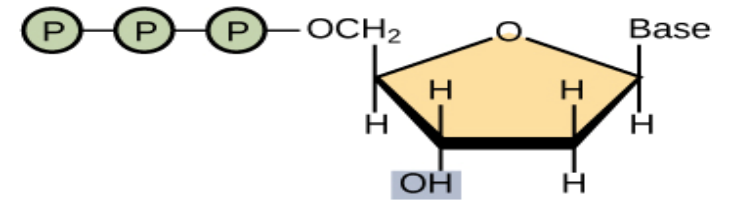
Ingredients for Sanger sequencing are similar to those needed for DNA replication in an organism, or for PCR

- A DNA polymerase enzyme
- A **primer**
- The four DNA nucleotides (dATP, dTTP, dCTP, dGTP)
- The template DNA to be sequenced

- **Dideoxy, or chain-terminating, versions of all four nucleotides (ddATP, ddTTP, ddCTP, ddGTP), each labeled with a different color of dye**



Dideoxynucleotide (ddNTP)

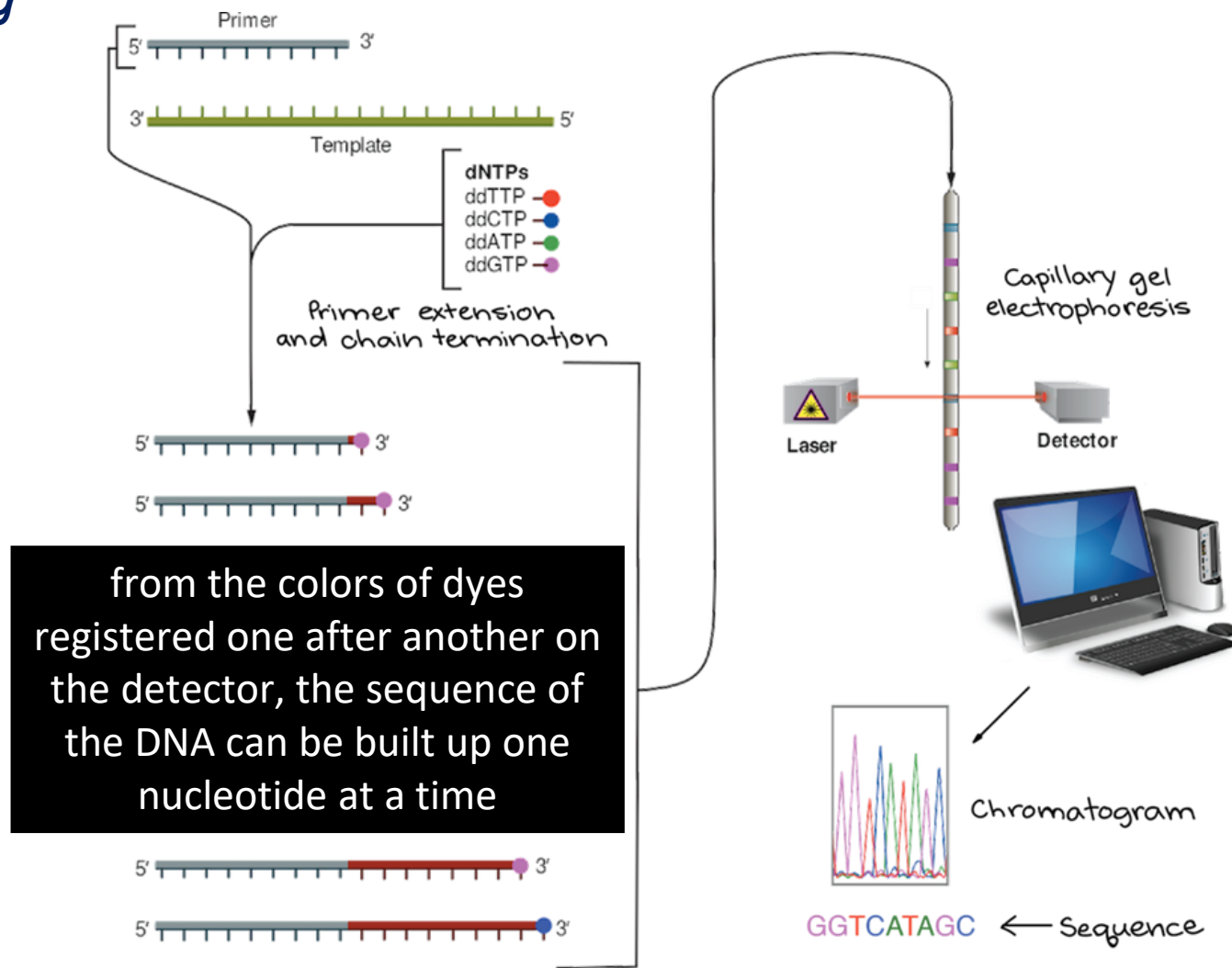


Deoxynucleotide (dNTP)

Principle

In a regular nt, the 3' hydroxyl group acts as a "hook," allowing a new nt to be added to an existing chain. Once a dideoxy nucleotide has been added to the chain, there is no hydroxyl available and no further nucleotides can be added. **The chain ends with the dideoxy nucleotide, which is marked with a particular color of dye depending on the base (A, T, C or G) that it carries.**

Sanger sequencing



from the colors of dyes registered one after another on the detector, the sequence of the DNA can be built up one nucleotide at a time

The DNA sequence is read from the peaks in the chromatogram

Sanger sequencing is expensive and inefficient for larger-scale projects, such as the sequencing of an entire genome or metagenome

Sanger sequencing- "the chain termination method"

- **Sequencing by Synthesis (SBS) method**
- **Regions of DNA up to about 900 base pairs in length**
- **The fragments were aligned based on overlapping portions to assemble the sequences of larger regions of DNA and, eventually, entire chromosomes**

Limitations

- **error rate is high (aprox. 0.001%)**
- **low-throughput (slow);** scientists can sequence DNA fragments up to 3000 bases per week

library preparation & amplification

Library preparation

- 1) DNA is fragmented either enzymatically or by sonication (excitation using ultrasound) to create smaller strands
- 2) Adaptors (short, double-stranded pieces of synthetic DNA) are then ligated to these fragments with the help of DNA ligase (an enzyme that joins DNA strands).

Library amplification

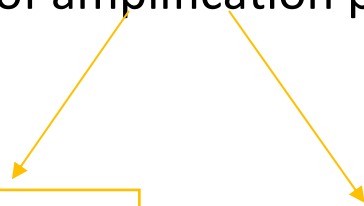
It is required so that the received signal from the sequencer is strong enough to be detected accurately

With enzymatic amplification, phenomena such as 'biasing' and 'duplication' can occur leading to preferential amplification of certain library fragments

Instead, there are several types of amplification processes which use PCR to create large numbers of DNA clusters

Emulsion PCR

Bridge PCR

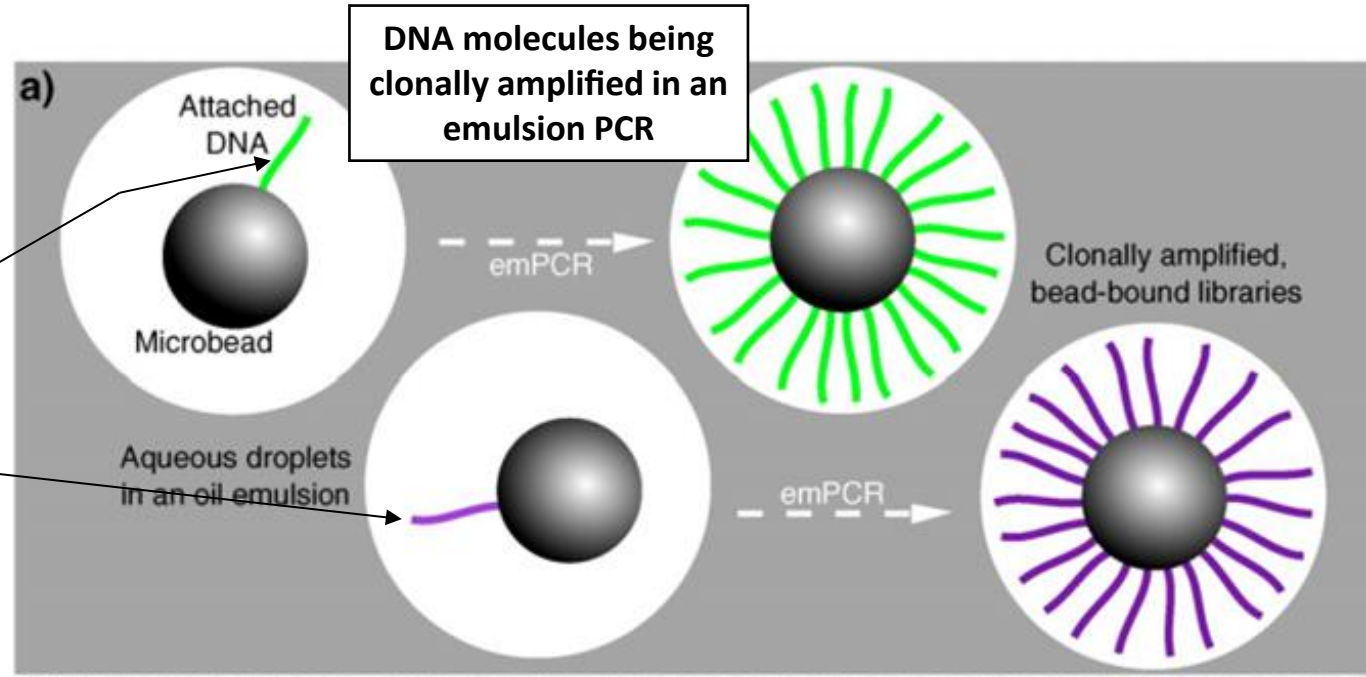


library amplification

emulsion PCR

Adapter ligation and PCR produce DNA libraries with 5' and 3' ends, which can then be made single stranded and immobilized onto individual suitably oligonucleotide-tagged microbeads.

Bead-DNA conjugates can then be emulsified using aqueous amplification reagents in oil, ideally producing emulsion droplets containing only one bead



454 seq (Roche)
Ion Torrent (Life Tech.)

Bridge amplification produces clusters of clonal DNA populations in a planar solid-phase PCR reaction. It happens for thousands of clusters all over the flow cell at once.

The flow cell is coated with two types of oligos, complementary to the two adapters on the fragment strand, respectively.

Once the fragment strand is added to the flow cell, it hybridizes to one of the oligos on the cell surface.

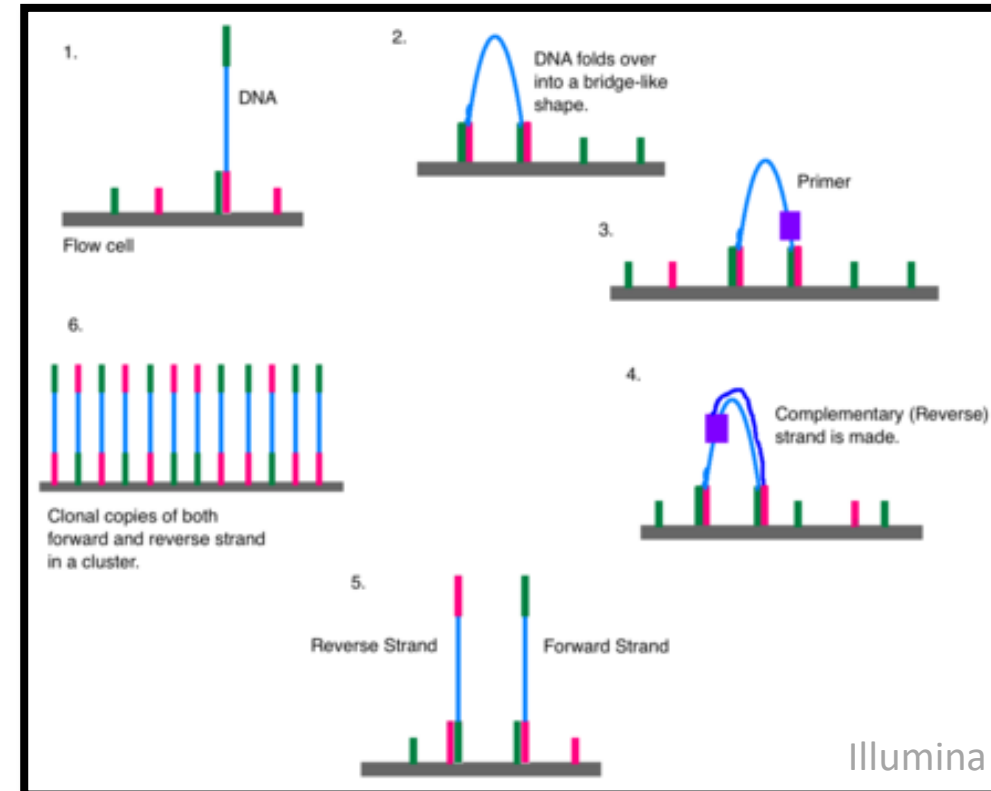
A polymerase then moves along the strand, creating its complementary DNA strand, i.e. the reverse strand. The double-stranded DNA is denatured and the original strand (forward strand) is washed away.

The remaining reverse strand then folds over and its adapter region hybridizes to the second type of oligo on the flow cell.

Polymerase attaches to the reverse strand and generates the complementary strand that is identical to the forward strand, forming a double-stranded bridge.

This bridge is then denatured, resulting in two single-stranded copies of the DNA, forward and reverse strand, anchored to the flow cell.

By repeating this denaturation and extension process, millions of fragments are amplified, forming localized **clusters** on the flow cell.



Second-generation sequencing- 454 SEQUENCING (ROCHE)

Pyrosequencing

- Does not infer nucleotide identity through using radio- or fluorescently-labelled dNTPs or oligonucleotides before visualising with electrophoresis
- Utilizes a luminescent method for **measuring pyrophosphate synthesis**

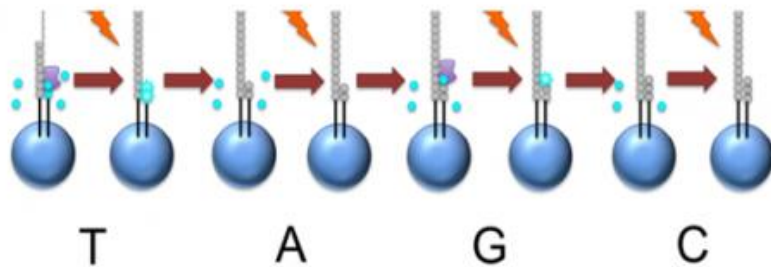
**Read lengths are around 200-300 bases.
400,000 reads of parallel sequencing
100mb of output per run
Run time 7.5 hours**



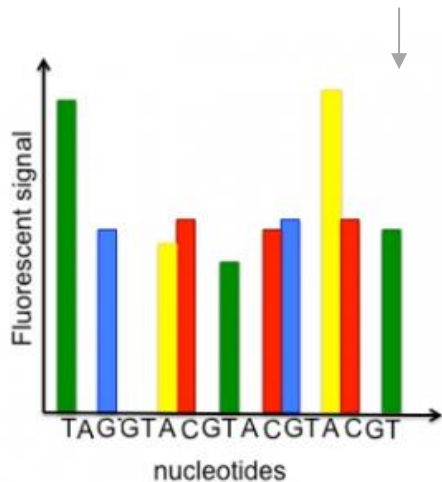
Second-generation sequencing- 454 SEQUENCING (ROCHE)

- 1) DNA or RNA is fragmented in up to 1kb reads
- 2) Generic adaptors are added to the ends and these are annealed to beads, one DNA fragment per bead
- 3) The fragments are then amplified by PCR using adaptor-specific primers
- 4) Each bead is then placed in a single well of a slide

Each well will contain a **single bead**, covered in many PCR copies of a **single sequence**. The wells also contain DNA polymerase and sequencing buffers.

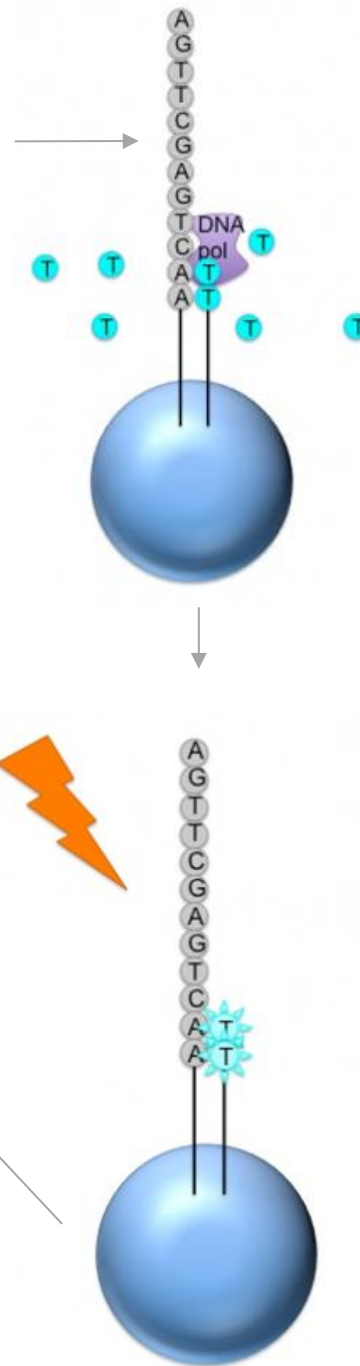


This NTP mix is washed away. The next NTP mix is now added and the process repeated, cycling through the four NTPs.



This kind of sequencing generates graphs for each sequence read, showing the signal density for each nucleotide wash. The sequence can then be determined computationally from the signal density in each wash.

All of the sequence reads we get from 454 will be different lengths, because different numbers of bases will be added with each cycle.

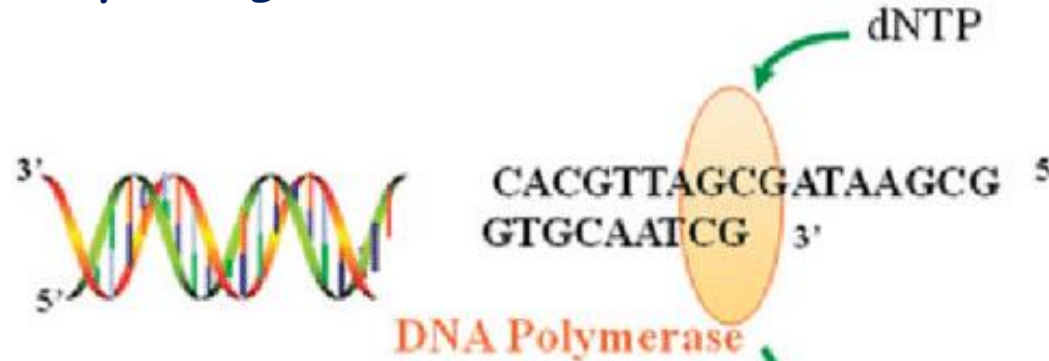


The slide is flooded with one of the four NTP species. Where this nucleotide is next in the sequence, it is added to the sequence read. If that single base repeats, then more will be added. So if we flood with Guanine bases, and the next in a sequence is G, one G will be added, however if the next part of the sequence is GGGG, then four Gs will be added.

The addition of each nucleotide releases a light signal. These locations of signals are detected and used to determine which beads the nucleotides are added to.

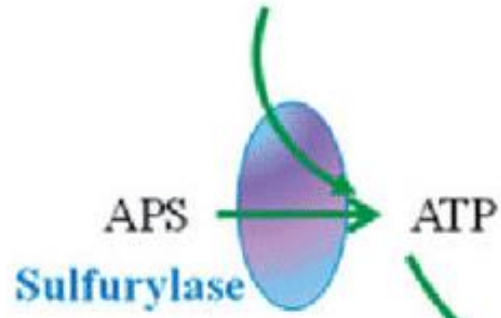
Second-generation sequencing- 454 SEQUENCING (ROCHE)

B
I
O
C
H
E
M
I
S
T
R
Y

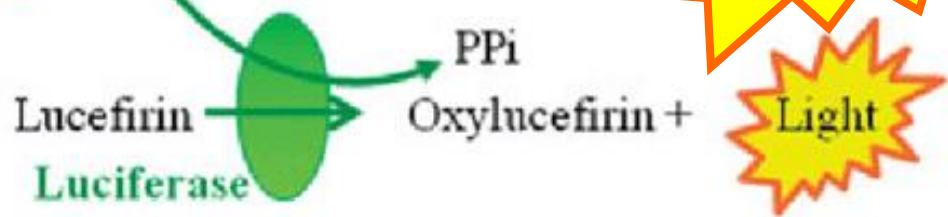


In DNA synthesis a dNTP is attached to the 3' end of the growing DNA strand. The two phosphates on the end are released as pyrophosphate (PPi)

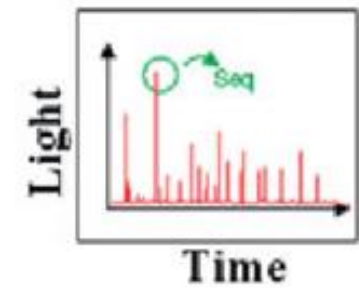
ATP sulfurylase uses PPI and adenosine 5' phosphosulfate to make ATP



Luciferase uses luciferin and ATP as substrates to produce oxyluciferin while releasing visible light



The amount of light is proportional to the number of nucleotides added



sequence-by-synthesis (SBS) techniques

Sanger's dideoxy and this pyrosequencing method are **'sequence-by-synthesis' (SBS) techniques**, as they both require the **direct action of DNA polymerase** to produce the observable output

BUT second-generation sequencing massively parallelizes Sanger sequencing, resulting in a **gain of roughly 6 orders of magnitude in terms of cost and speed**

Next-generation sequencing

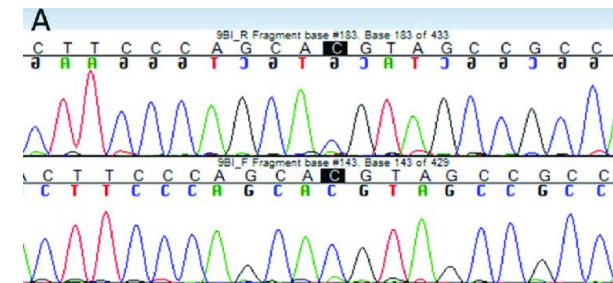
Next generation sequencing (NGS), massively parallel or deep sequencing are related terms that describe a DNA sequencing technology which has revolutionised genomic research

Using NGS an entire human genome can be sequenced within a single day

NGS can be used to sequence entire genomes or constrained to specific areas of interest

*Why use **NGS** compared to Sanger sequencing?*

- **Highly parallel:** many sequencing reactions take place at the same time
- **Micro scale:** reactions are tiny and many can be done at once on a chip
- **Fast:** because reactions are done in parallel, results are ready much faster
- **Low-cost:** sequencing a genome is cheaper than with Sanger sequencing
- **Shorter length:** reads typically range from 50 -700 nucleotides in length



NGS workflow

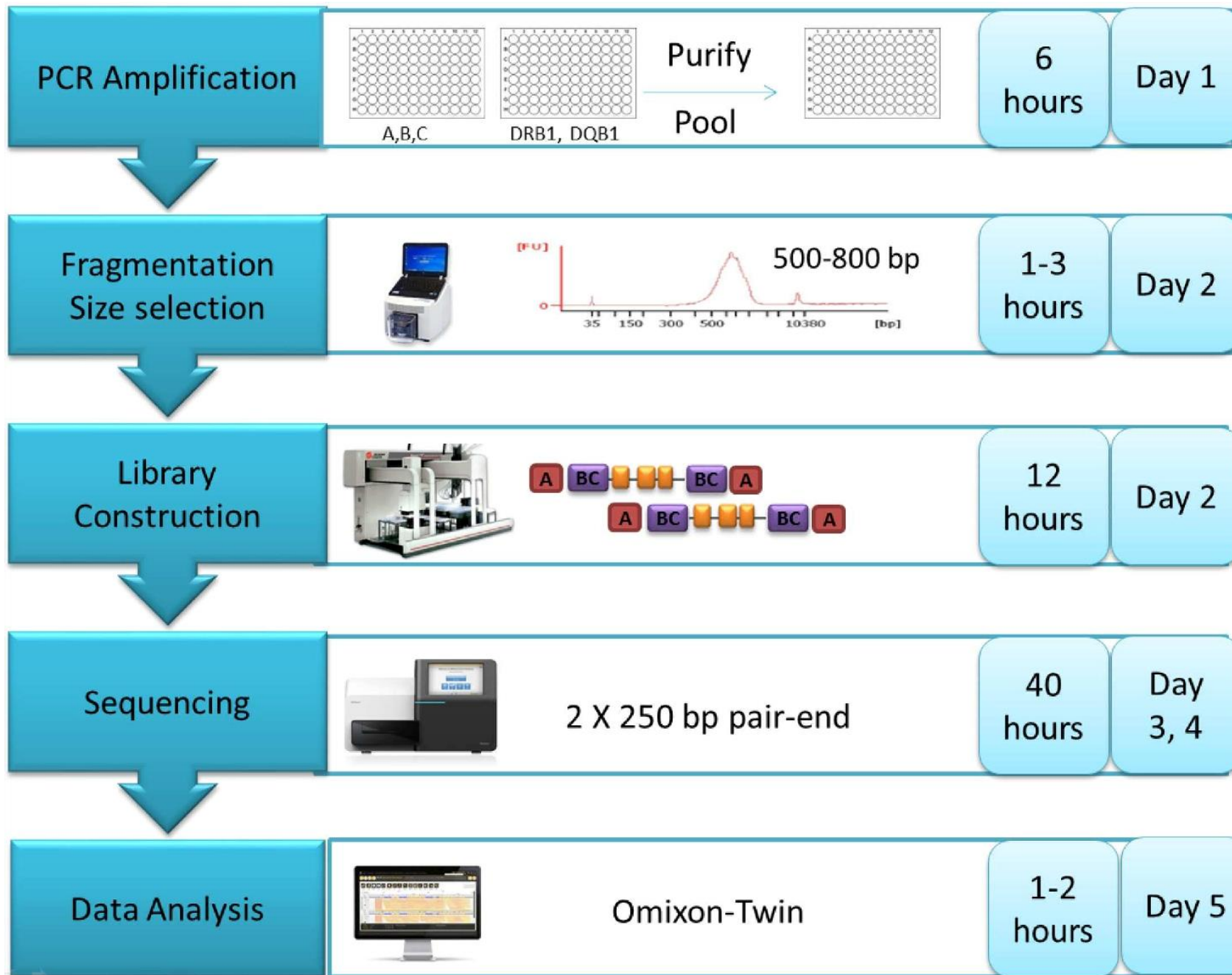


Fig. High-throughput NGS workflow. High-throughput NGS workflow begins with multiplex long range PCR of A, B, C and DRB1, DQB1. After PCR, amplicons are purified and pooled in equimolar concentrations. Sheared amplicons then undergo library preparation. To maximize throughput, each clinical sample is labeled with unique dual indices. 2×250 bp paired-end sequence data from the Illumina MiSeq are exported and analyzed, with 3.19.0 IMGT/HLA database serving as the reference.

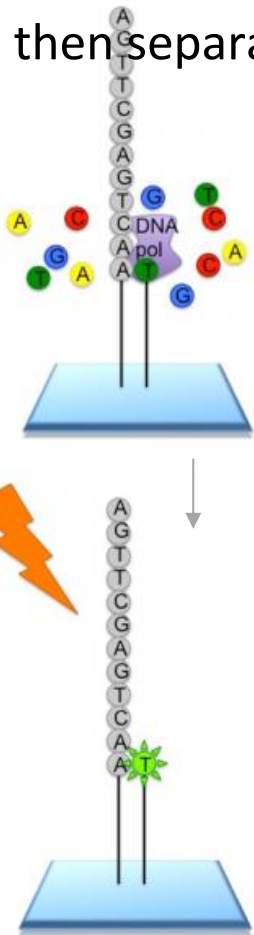
Next-generation sequencing- ILLUMINA

Illumina sequencing

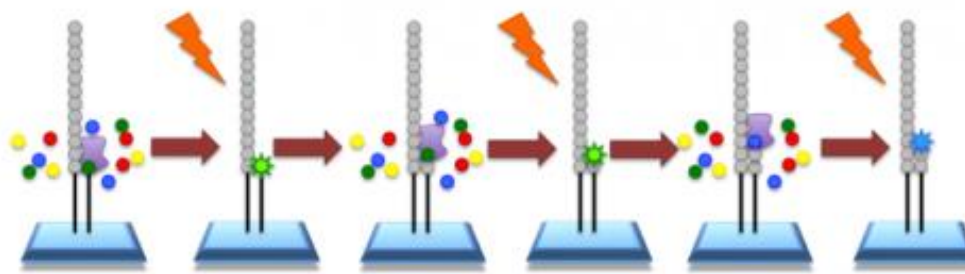
100-150bp reads are used. Somewhat longer fragments are ligated to generic adaptors and annealed to a slide using the adaptors. **Fluorescently labelled nucleotides with terminators are incorporated in the new strand.**

PCR is carried out to amplify each read, **creating a spot with many copies of the same read.** They are then separated into single strands to be sequenced.

B
I
O
C
H
E
M
I
S
T
R
Y



The slide is flooded with nucleotides and DNA polymerase. These nucleotides are fluorescently labelled, with the colour corresponding to the base. They also have a terminator, so that only one base is added at a time.



The slide is then prepared for the next cycle. The terminators are removed, allowing the next base to be added, and the fluorescent signal is removed, preventing the signal from contaminating the next image.

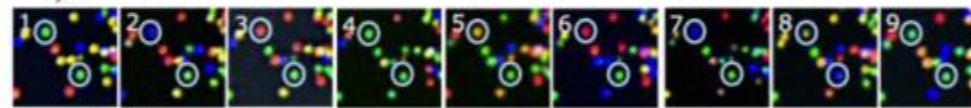
The process is repeated, adding one nucleotide at a time and imaging in between.

An image is taken of the slide. In each read location, there will be a fluorescent signal indicating the base that has been added.



Illumina Flow Cell

TGCTACGAT...



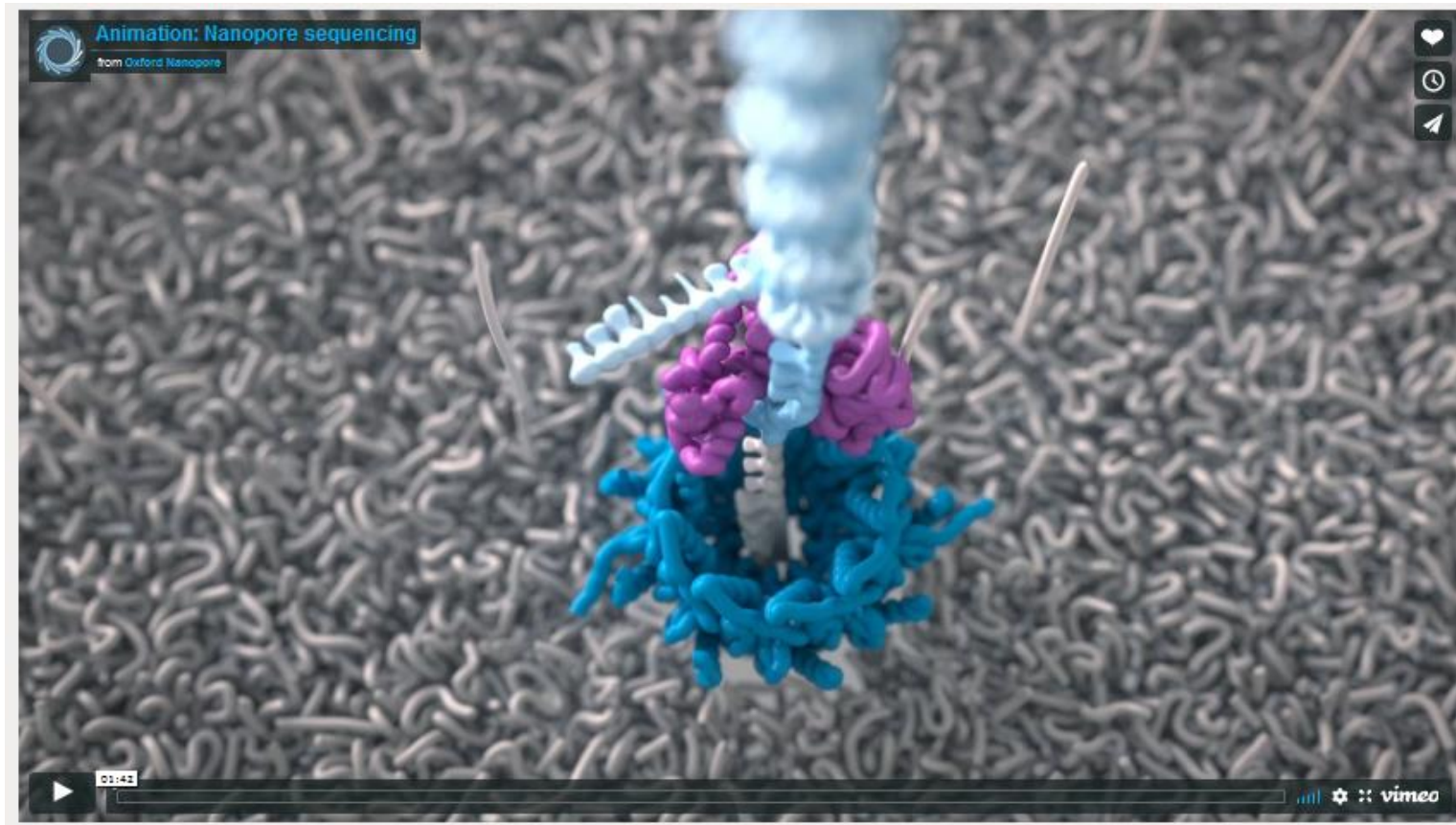
TTTTTTTGT...

Computers are then used to detect the base at each site in each image and these are used to construct a sequence.

All of the sequence reads will be the same length, as the read length depends on the number of cycles carried out.

nanopore sequencing

Nanopore sequencing works by **monitoring changes to an electrical current** as nucleic acids are passed through a protein nanopore. The resulting signal is decoded to provide the specific DNA or RNA sequence.



After library preparation, individual molecules are loaded into a flow cell, where **motor proteins**, which are attached during adaptor ligation, **dock with nanopores**.

The **motor protein** controls the translocation of the RNA strand through the nanopore, **causing a change in current that is processed to generate sequencing reads of 1–10 kb**

Generation of long-reads: longer than 10 kb
e.g. more than >5 Gb in each run have been obtained with MinION technology

1) **Paired-End Sequencing:** sequencing both ends of the DNA fragments in a library and aligning the forward and reverse reads as read pairs



- Paired-end sequencing enables both ends of the DNA fragment to be sequenced
- Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely
- **This results in better alignment of reads, especially across difficult-to-sequence, repetitive regions of the genome**

2) Tunable Coverage and Unlimited Dynamic Range

By increasing or decreasing the number of sequencing reads (coverage), researchers can **tune the sensitivity** of an experiment to accommodate various study objectives

- ❑ Dynamic range (i.e. signal range) with NGS is adjustable and nearly unlimited: researchers can quantify subtle gene expression changes with much **greater sensitivity** than traditional microarray-based methods
- ❑ Sequencing runs can be tailored to zoom in with **high resolution** on particular regions of the genome, or provide a more expansive view with **lower resolution**

Examples:

- **Detection of low-frequency mutations within a mixed cell population**: somatic mutations may only exist within a small proportion of cells in a given tissue sample. Using mixed tumor–normal cell samples, the region of DNA harboring the mutation must be sequenced at extremely high coverage, often upwards of 1000×, to detect these low-frequency mutations within the mixed cell population
- **Genome-wide variant discovery**: requires a much lower coverage level. The study design involves sequencing many samples (hundreds to thousands) at lower resolution, to achieve greater statistical power within a given population

3) Advances in Library Preparation

a) High-throughput

The first NGS library prep protocols involved random fragmentation of the DNA or RNA sample, gel-based size selection, ligation of platform-specific oligonucleotides, PCR amplification, and several purification steps } 1-2 days

Current NGS protocols have reduced the library prep time } 90 min to 6 hrs

b) High accuracy

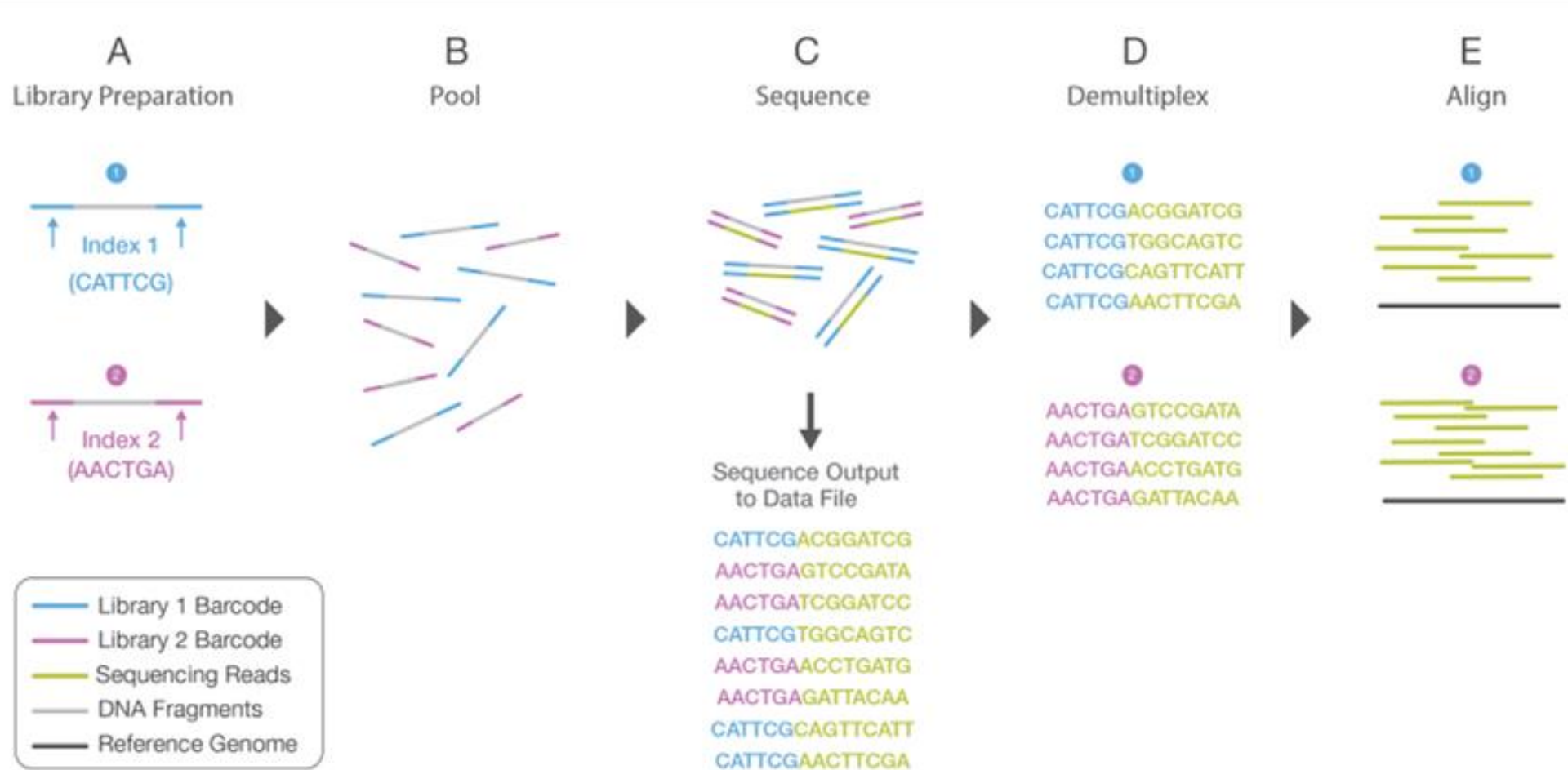
PCR-free (!) library preparation kits result in superior coverage of traditionally challenging areas, such as high AT/GC-rich regions, promoters, and homopolymeric regions (regions that include stretches of the same nucleotide (e.g. AAAAA or TTTTTTTT))

Next-generation sequencing - Advances

4. Multiplexing

allows large numbers of libraries to be pooled and sequenced simultaneously during a single sequencing run

NGS has dramatically reduced the time to data for multisample studies and enabled researchers to go from experiment to data quickly and easily



(A) Unique index sequences are added to two different libraries during library preparation. (B) Libraries are pooled together and loaded into the same flow cell lane. (C) Libraries are sequenced together during a single instrument run. All sequences are exported to a single output file. (D) A demultiplexing algorithm sorts the reads into different files according to their indexes. (E) Each set of reads is aligned to the appropriate reference sequence.

Next-generation sequencing

Sample indexing (or barcoding):

- Sample indexes enable multiple samples to be sequenced together (i.e., multiplexed) on the same instrument flow cell or chip
- Each sample index, typically 8–10 bases, is specific to a given sample library and is used for demultiplexing during data analysis to assign individual sequence reads to the correct sample
- Adapters may contain single or dual sample indexes depending on the number of libraries combined and the level of accuracy desired

Next-generation sequencing

Molecular barcoding with Unique Molecular Identifiers (UMIs) provide the highest levels of error correction and accuracy

- UMIs are short sequences that incorporate a unique barcode onto each molecule within a given sample library
 - UMIs enable precise quantification, since PCR duplicates will share the same insert and UMI tag
 - UMI deduplication is useful for RNA-seq gene expression analysis and other quantitative sequencing methods
 - UMIs have also been shown to reduce the rate of false-positive variant calls and increase sensitivity of variant detection
-
- By incorporating individual barcodes on each original DNA fragment, variant alleles present in the original sample (true variants) can be distinguished from errors introduced during library preparation, target enrichment, or sequencing
 - Any identified errors can be removed by bioinformatics methods before final data analysis

Next-generation sequencing- APPLICATIONS

Sequencing methods differ primarily by how the DNA or RNA samples are obtained (eg, organism, tissue type, normal vs. affected, experimental conditions, etc) and by the data analysis options used

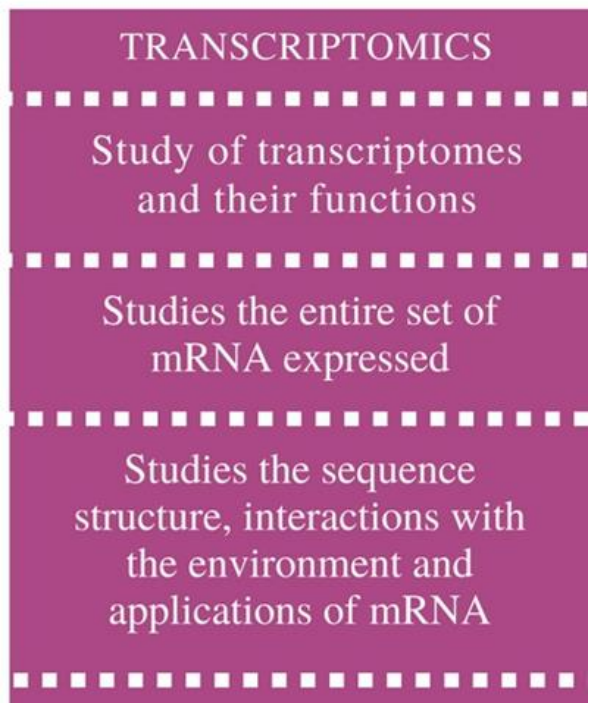
After the sequencing libraries are prepared, the actual sequencing stage remains fundamentally the same, regardless of the method

a. Genomics

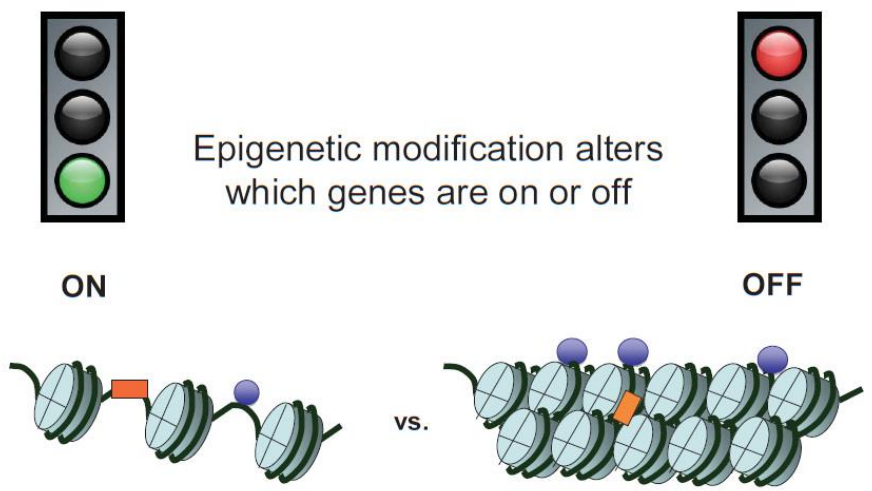


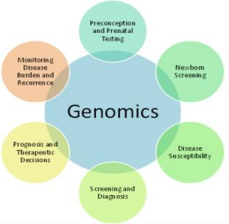
library preparation

b. Transcriptomics



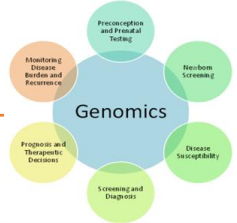
c. Epigenomics





De novo Sequencing

- sequencing a novel genome where there is no reference sequence available for alignment
- sequence reads are assembled as contigs and the coverage quality of *de novo* sequence data depends on the size and continuity of the contigs (i.e. the number of gaps in the data)

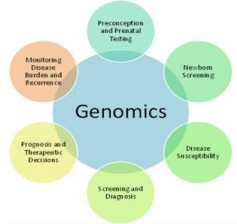


Targeted sequencing

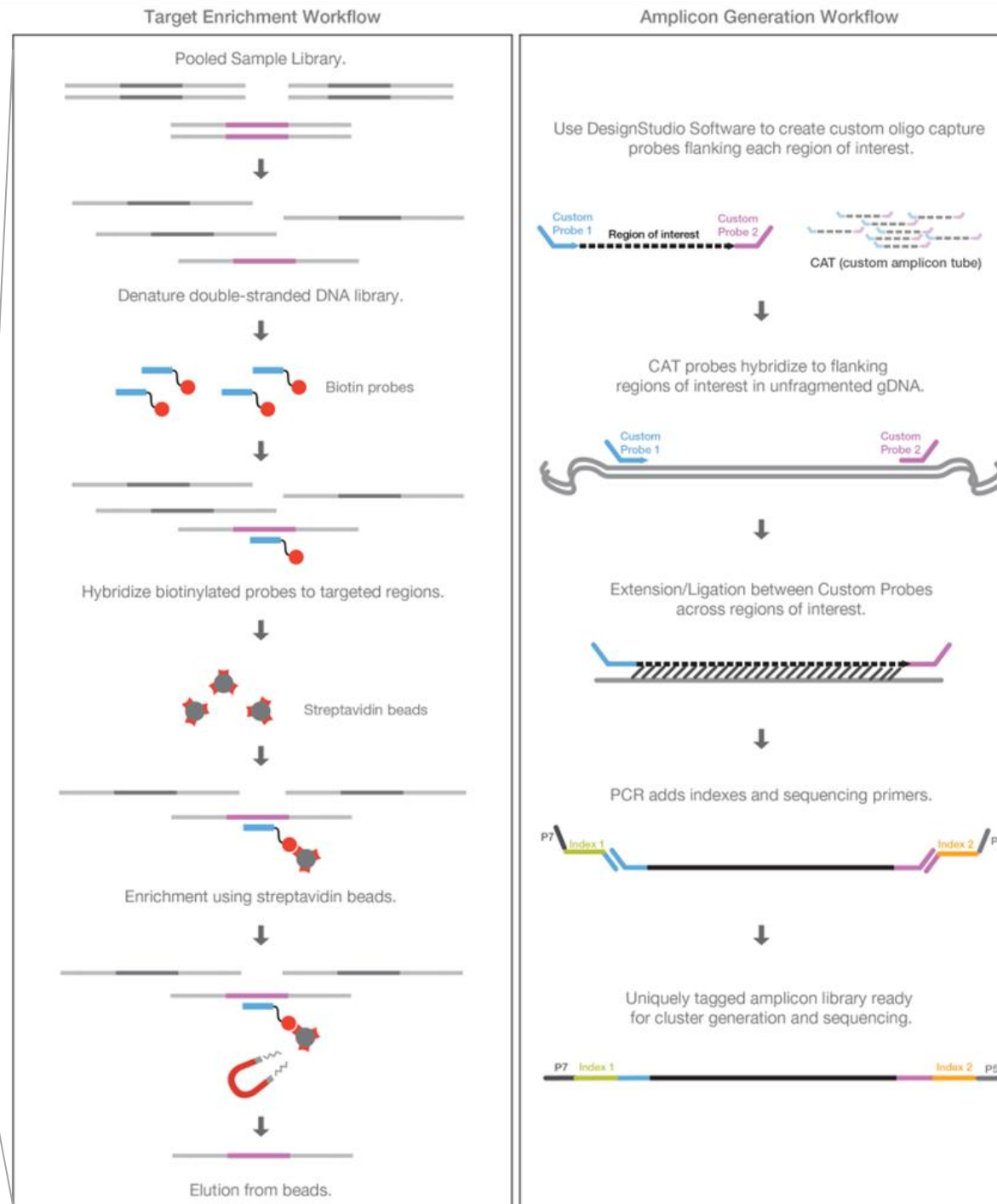
- a subset of genes or regions of the genome are isolated and sequenced
- allows researchers to focus time, expenses, and data analysis on specific areas of interest
- enables sequencing at much higher coverage levels
- target enrichment captures between 10 kb–62 Mb regions or sequence 16–1536 targets at a time depending on the library prep kit parameters
- **amplicon sequencing** (PCR products) is useful for discovery of rare somatic mutations in complex samples (e.g. cancerous tumors mixed with germline DNA), sequencing the bacterial 16S rRNA gene across multiple species used for phylogeny and taxonomy studies, particularly in diverse metagenomic samples
 - With amplicon sequencing, you do PCR with two primers flanking the region you care about. Presumably, you'd try to multiplex a bunch of them together per sample. Then you make a sequencing library of all those amplicons

The other common way to sequence a particular subset of a genome is to use a capture probe that binds to your sequence of interest, then you sequence that library. Large sets of probes designed to capture exonic sequence are commonly used

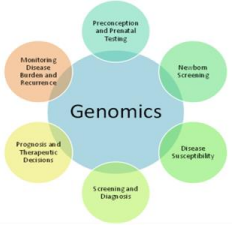
Next-generation sequencing- GENOMICS



With **target enrichment**, specific regions of interest are **captured** by hybridization to biotinylated probes, then isolated by magnetic pulldown



Amplicon sequencing involves the amplification and purification of regions of interest using highly multiplexed PCR oligos sets



Target Enrichment vs Amplicon Sequencing

- Larger gene content, typically >50 genes vs Smaller gene content, typically <50 genes
- More comprehensive profiling for all variant types vs Ideal for analyzing single nucleotide variants and insertions/deletions (indels)
- More comprehensive method, but with longer hands-on time and turnaround times
- More affordable, easier workflow

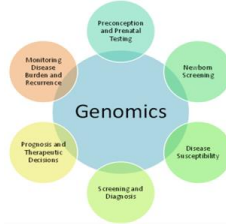


Table 5.2 Comparison of next-generation sequencing platforms

Sequencing instruments	Method	Read length (bp)	Error (%); primary error	Output (million reads/run)	Remarks
454	Emulsion PCR and Pyrosequencing	700	1; indels	1	Longer read but comparatively high cost/Mb data and high error rate compared to MiSeq®/HiSeq™. Ok for marker gene-based metagenomic sequencing
Ion PGM	Emulsion PCR and semiconductor Sequencing	400	≈1; indels	0.4–0.5 (314 chip)	Faster and reasonably longer read than MiSeq/HiSeq but low throughput than to MiSeq/HiSeq. Also too much hands-on time and high error rate associated with indels, OK for marker gene-based metagenomic sequencing but strict size selection is a problem
Ion proton	Emulsion PCR and semiconductor sequencing	200	≈1, indels	2–3 (316 chip) 4–5.5 (318 chip) 60–80 (PI chip)	Same as 314 chip Same as 314 chip Higher throughput than Ion PGM but lower throughput than HiSeq, shorter reads than Ion PGM/MiSeq/HiSeq/PacBio
MiSeq	Bridge amplification and reversible dye terminator sequencing	300+300	≈0.1; substitution	25	High throughput, low error rate, and comparable read length in paired-end sequencing make this platform best among all for marker gene-based metagenomic sequencing
HiSeq 2500	Bridge amplification and reversible dye terminator sequencing	250+250	≈0.1; substitution	600 (rapid run v2 kit)	Same as MiSeq but with much higher throughput and slightly shorter read length. Good for shotgun Meta-omic sequencing
PacBio® RS II	Single molecule real-time (SMRT) sequencing	125+125 8500 bp	≈11; indels	4000 0.05 (per smart cell per run)	No amplification step, longest read among all sequencing platforms (up to 20kb) but with low throughput and very high error rate in a single pass. Good for shotgun sequence assembly in meta-omic analysis using hybrid approach. Also works better for hard to sequence DNA template (e.g., AT/GC rich DNA, highly repetitive sequences, sequence with long homonucleotide stretch, etc.)

Next-generation sequencing- TRANSCRIPTOMICS

Library preparation methods for RNA-Seq

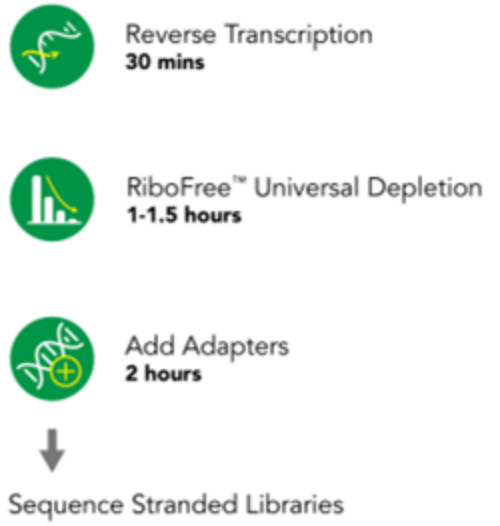
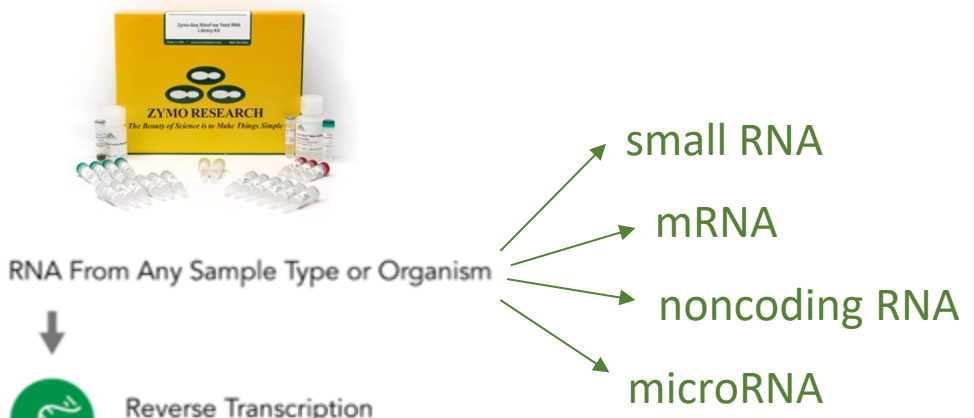
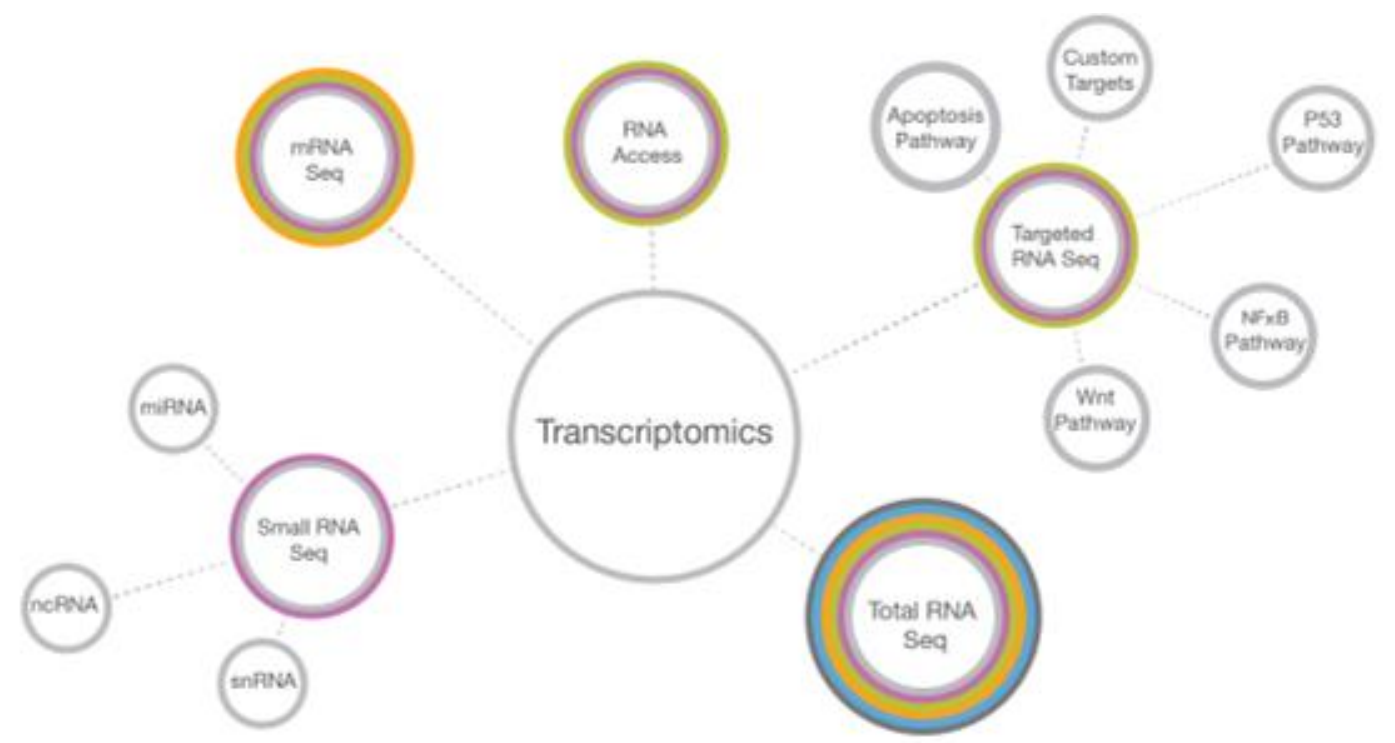


Figure 1: The Zymo-Seq RiboFree™ Total RNA Library Kit is the fastest and easiest Total RNA-Seq workflow. This kit minimizes the number of reagents and steps needed to generate stranded rRNA-depleted total RNA libraries in as little as 3.5 hours.

Library preparation methods for RNA-Seq typically begin with total RNA sample preparation followed by a ribosome removal step. The total RNA sample is then converted to cDNA before standard NGS library preparation. RNA-Seq focused on mRNA, small RNA, noncoding RNA, or microRNAs can be achieved by including **additional isolation or enrichment steps** before cDNA synthesis.



Total RNA and mRNA Sequencing

- transcriptome sequencing is a major advance in the **study of gene expression** because it allows a **snapshot of the whole transcriptome** rather than a predetermined subset of genes
- provides a comprehensive view of a **cellular transcriptional profile at a given biological moment**

Targeted RNA sequencing

Method for **measuring transcripts of interest** for detecting differential expression, allele-specific expression, detection of gene-fusions, isoforms, and splice junctions

e.g. **sequencing kits include preconfigured, experimentally validated panels focused on specific cellular pathways or disease states such as apoptosis, cardiotoxicity, NFκB pathway, and more**

e.g. **custom content can be designed and ordered for analysis of specific genes of interest**

e.g. **custom content for detecting small, noncoding RNA, or microRNAs**



Epigenetics is the study of heritable changes in gene activity caused by mechanisms other than DNA sequence changes. Mechanisms of epigenetic activity include DNA methylation, small RNA-mediated regulation, DNA-protein interactions, histone modification, etc.

Methylation sequencing

- is the study of cytosine methylation (5mC) states across specific areas of regulation, such as promoters or heterochromatin
- Cytosine methylation can significantly **modify temporal and spatial gene expression and chromatin remodeling**

ChIP Sequencing

- ✓ **Protein-DNA or protein-RNA interactions** have a significant impact on many biological processes and disease states. These interactions can be surveyed with NGS by combining chromatin immunoprecipitation (ChIP) assays and NGS methods. ChIP-Seq protocols begin with the chromatin immunoprecipitation step (vary widely as they must be specific to the species, tissue type, and experimental conditions)

Ribosome Profiling

- **Deep sequencing of ribosome protected-mRNA fragments.** Purification and sequencing of these fragments provides a **snapshot of all the ribosomes active in a cell at a specific time point.** This information can determine what **proteins are being actively translated in a cell,** and can be useful for investigating translational control, measuring gene expression, determining the rate of protein synthesis, or predicting protein abundance

single-cell sequencing

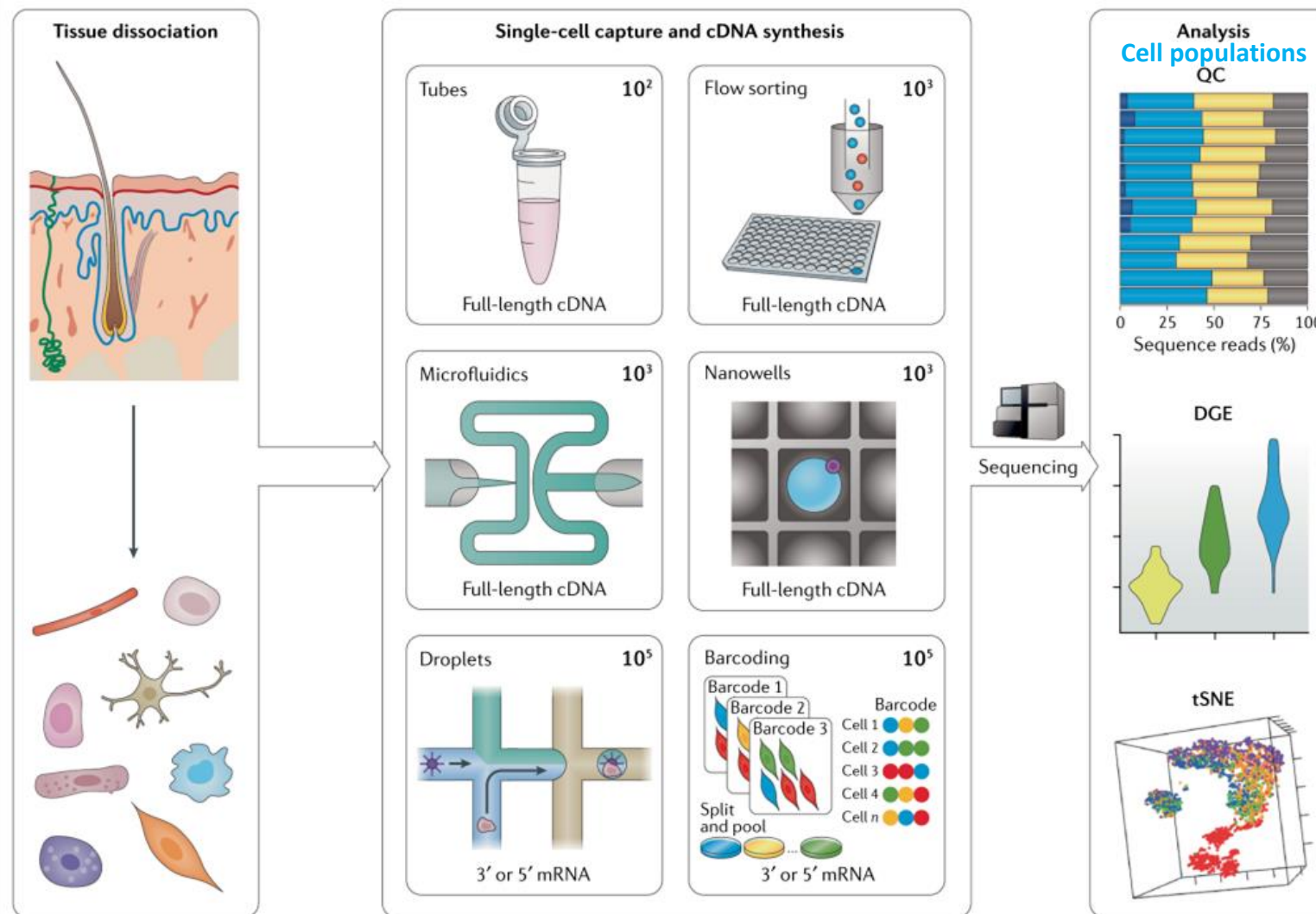
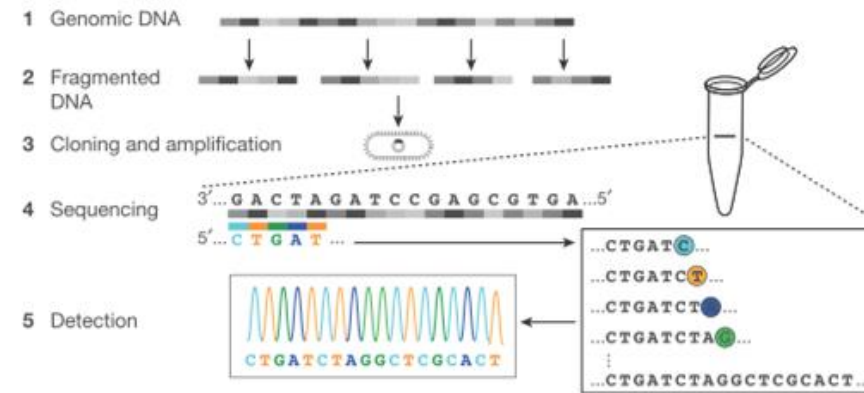
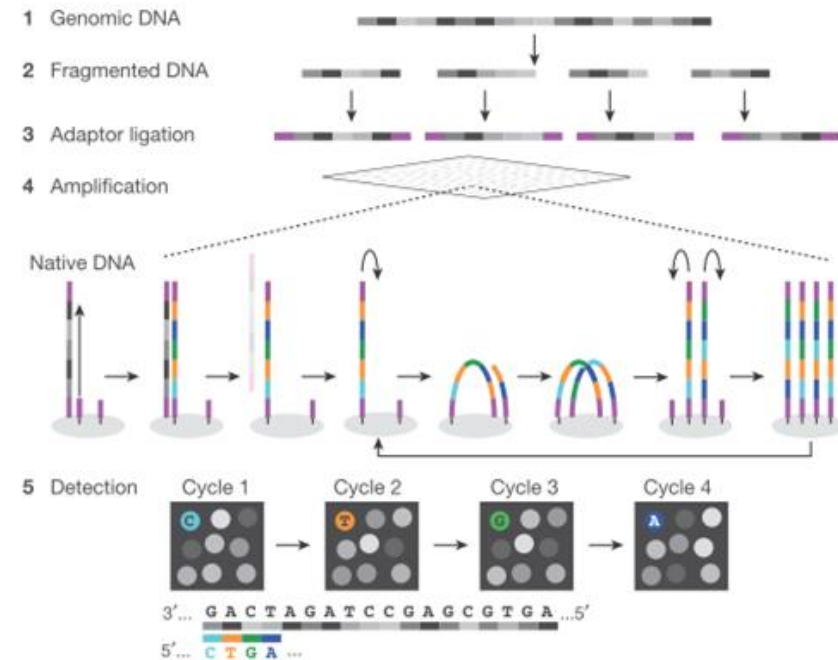


Fig. Single-cell sequencing begins with the isolation of single cells from a sample, such as dissociated skin tissue. Cells are reverse transcribed in order to produce cDNA (usually tagged with unique molecular identifiers (UMIs)) for RNA-seq library preparation and sequencing. Quality control (QC), differential gene expression (DGE) and 2D visualization, along with unsupervised clustering and network analysis, of the single-cell RNA-seq data are used to **determine discrete cell populations**. The number of cells usually profiled is indicated alongside each technology, as is the RNA-seq strategy — for example, 3' or 5' mRNA or full-length cDNA. <https://doi.org/10.1038/s41576-019-0150-2>

First generation sequencing (Sanger)



Second generation sequencing (massively parallel)



Third generation sequencing (Real-time, single molecule)

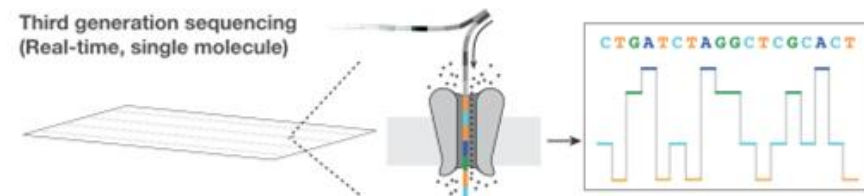


Fig. DNA sequencing technologies.

For extra info

- RNA sequencing: the teenage years. doi: <https://doi.org/10.1038/s41576-019-0150-2>
- DNA sequencing at 40: past, present and future <https://doi.org/10.1038/nature24286>
- Understanding the Basics of NGS: From Mechanism to Variant Calling. doi [10.1007/s40142-015-0076-8](https://doi.org/10.1007/s40142-015-0076-8)
- The sequence of sequencers: The history of sequencing DNA. doi [10.1016/j.ygeno.2015.11.003](https://doi.org/10.1016/j.ygeno.2015.11.003)
- Computational Biology Methods and Their Application to the Comparative Genomics of Endocellular Symbiotic Bacteria of Insects doi: [10.1007/s12575-009-9004-1](https://doi.org/10.1007/s12575-009-9004-1)
- Application of High-Throughput Next-Generation Sequencing for HLA Typing on Buccal Extracted DNA: Results from over 10,000 Donor Recruitment Samples. DOI:[10.1371/journal.pone.0165810](https://doi.org/10.1371/journal.pone.0165810)
- An introduction to Next-Generation Sequencing Technology www.illumina.com/technology/next-generation-sequencing.html
- Basics of DNA and sequencing by synthesis <http://data-science-sequencing.github.io/Win2018/lectures/lecture2/>
- Technology Spotlight: Illumina Sequencing
- <https://www.genome.gov/human-genome-project>
- <http://data-science-sequencing.github.io/Win2018/lectures/lecture2/>
- <https://www.khanacademy.org/science/high-school-biology/hs-molecular-genetics/hs-biotechnology/a/dna-sequencing>
- <https://www.slideshare.net/AshfaqAhmad52/pyrosequencing-87620649>



Sequencing
Human genome sequencing

questions?

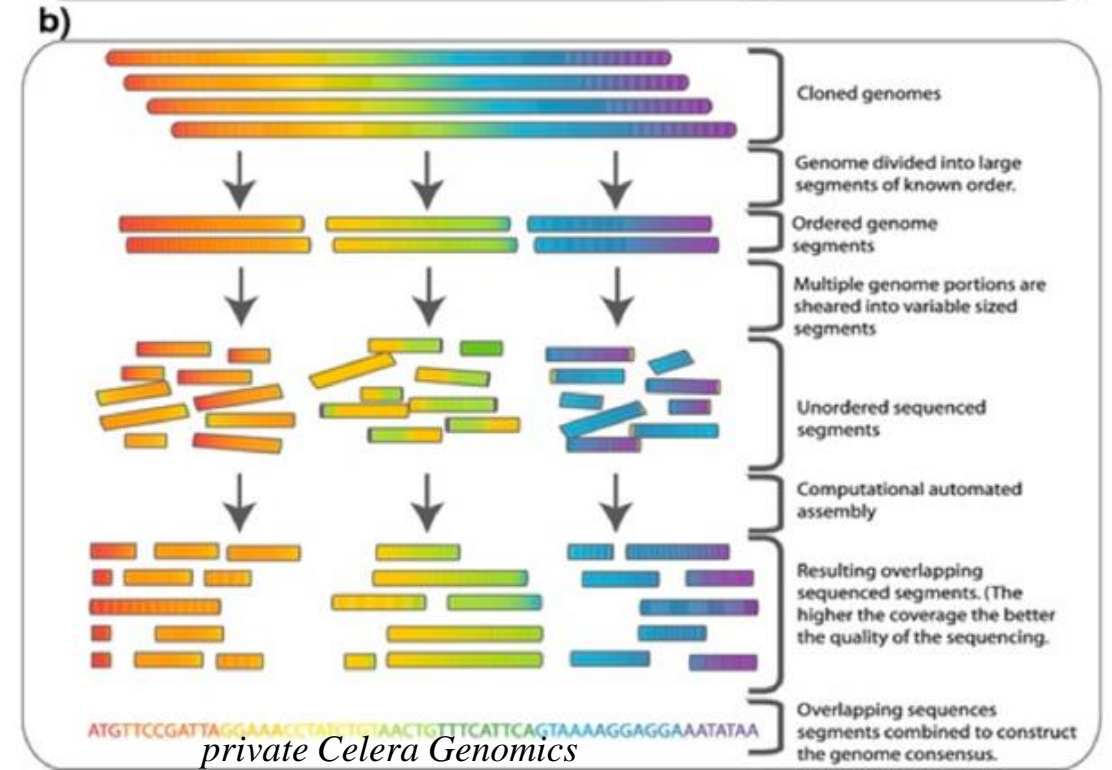
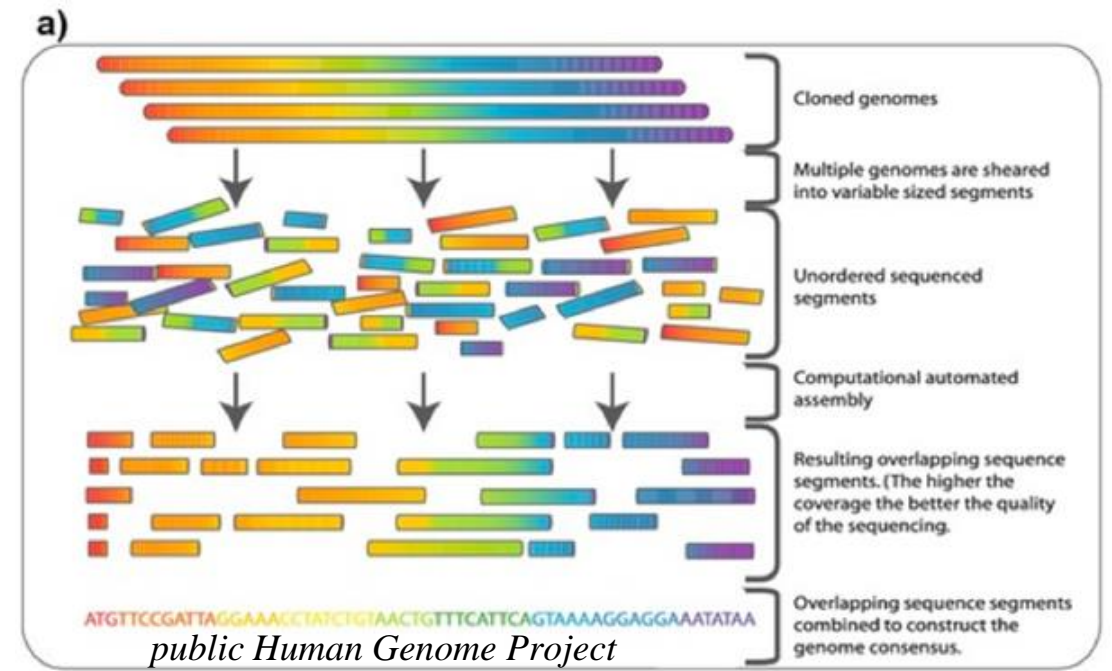
ТНАОК



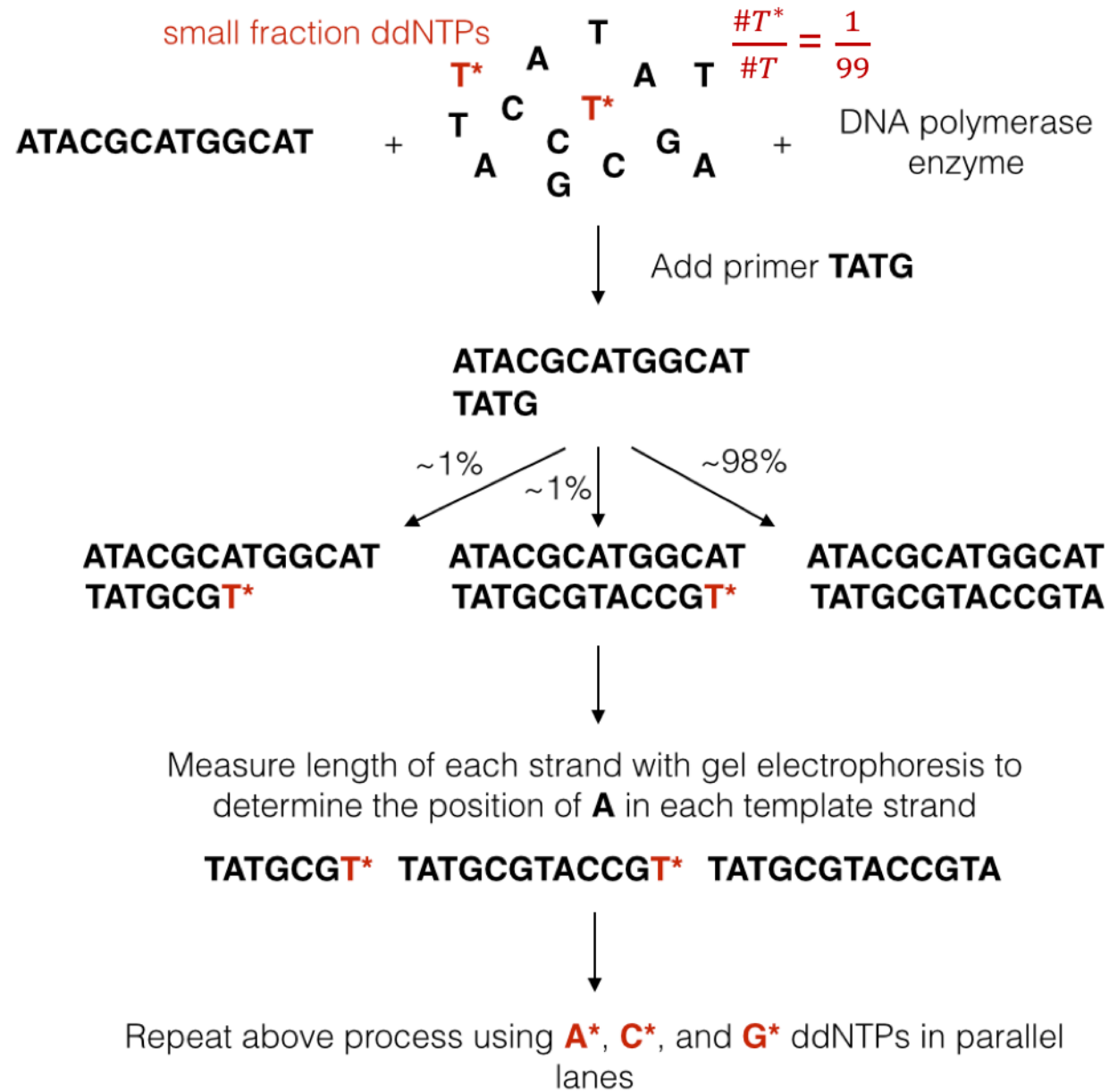
two sequencing strategies were pursued to assemble the human genome for SHOTGUN sequencing

A) In **whole genome shotgun sequencing**, the entire genome is sheared randomly into small fragments (appropriately sized for sequencing) and then reassembled

B) In **hierarchical shotgun sequencing**, the genome is first broken into larger segments. After the order of these segments is deduced, they are further sheared into fragments appropriately sized for sequencing



Sanger sequencing- "the chain termination method"



Sanger sequencing

The image shows a video player displaying a Sanger sequencing gel and a DNA sequence. The gel is a 4-lane gel with lanes labeled A, T, C, and G. The bands in the gel are as follows:

Lane	Band 1	Band 2	Band 3	Band 4	Band 5	Band 6	Band 7
A							
T							
C							
G							

The DNA sequence shown is:

```
5' G T G C T C A  
3' C A C G A G T
```

The video player interface includes a red progress bar, a play button, a volume icon, a timestamp of 5:31 / 5:59, and various control icons (full screen, settings, etc.) at the bottom.

Sanger sequencing- "the chain termination method"

Limitations of Sanger sequencing

1) **Read limitation:** as the length L of a sequence increases, distinguishing between the mass of a length L sequence and the mass of a length $L+1$ sequence becomes increasingly harder.

e.g. a tolerance of 0.1% in measurement would make it impossible to distinguish a sequence of length 1000 from one of length 1001 even if all bases had the same molecular weight.

This is a reason for **errors in Sanger sequencing**, though the **error rate is around 0.001%**.

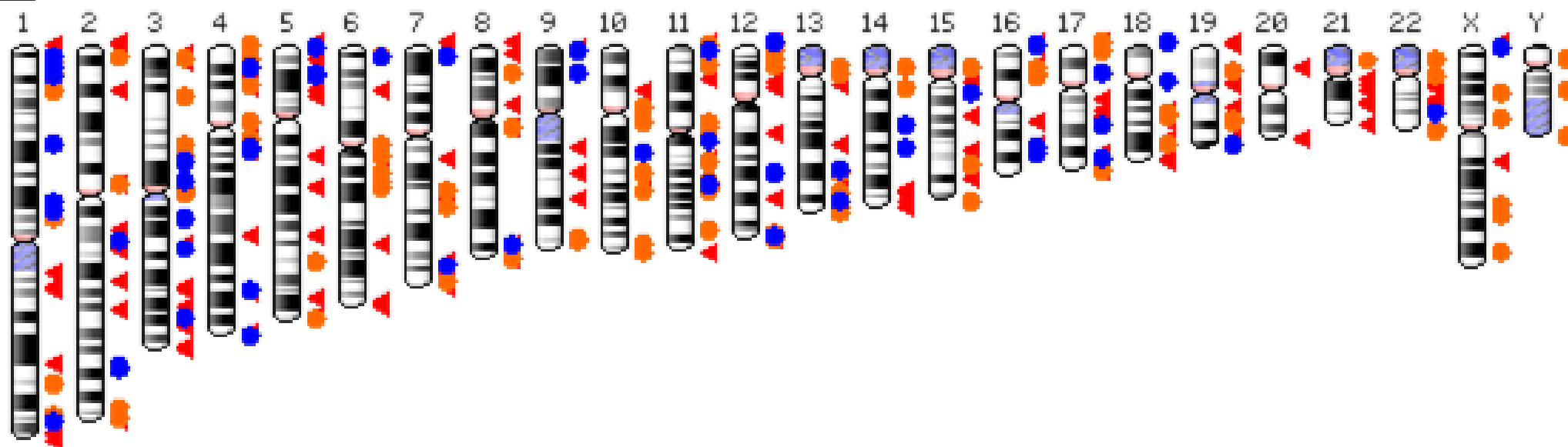
2) Sanger sequencing is **low-throughput (slow)** because the mass measuring process is time consuming; scientists can sequence DNA fragments up to 3000 bases per week.



Gel electrophoresis during Sanger sequencing.

Human Genome Overview

Information about the continuing improvement of the human genome



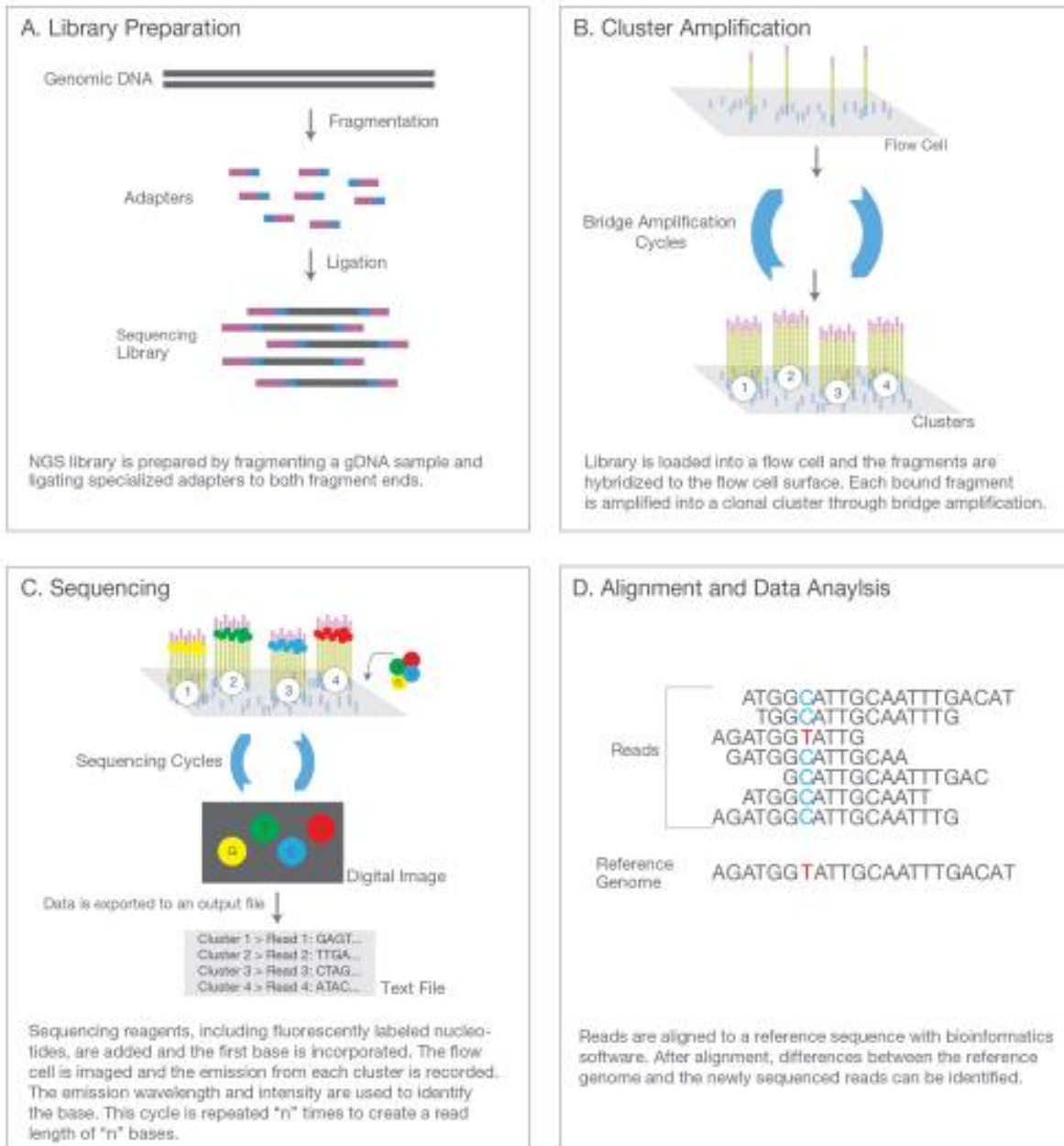
◀ Region containing alternate loci

● Region containing fix patches

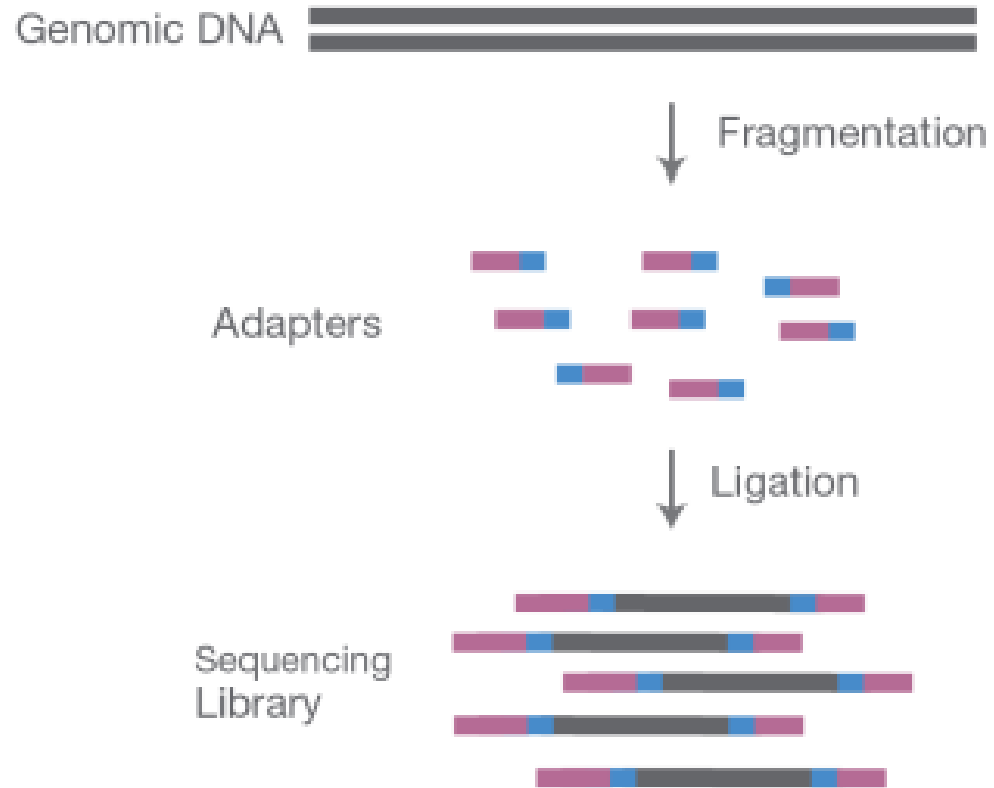
● Region containing novel patches

Next-generation sequencing- ILLUMINA

Illumina NGS workflow includes four basic steps:



A. Library Preparation

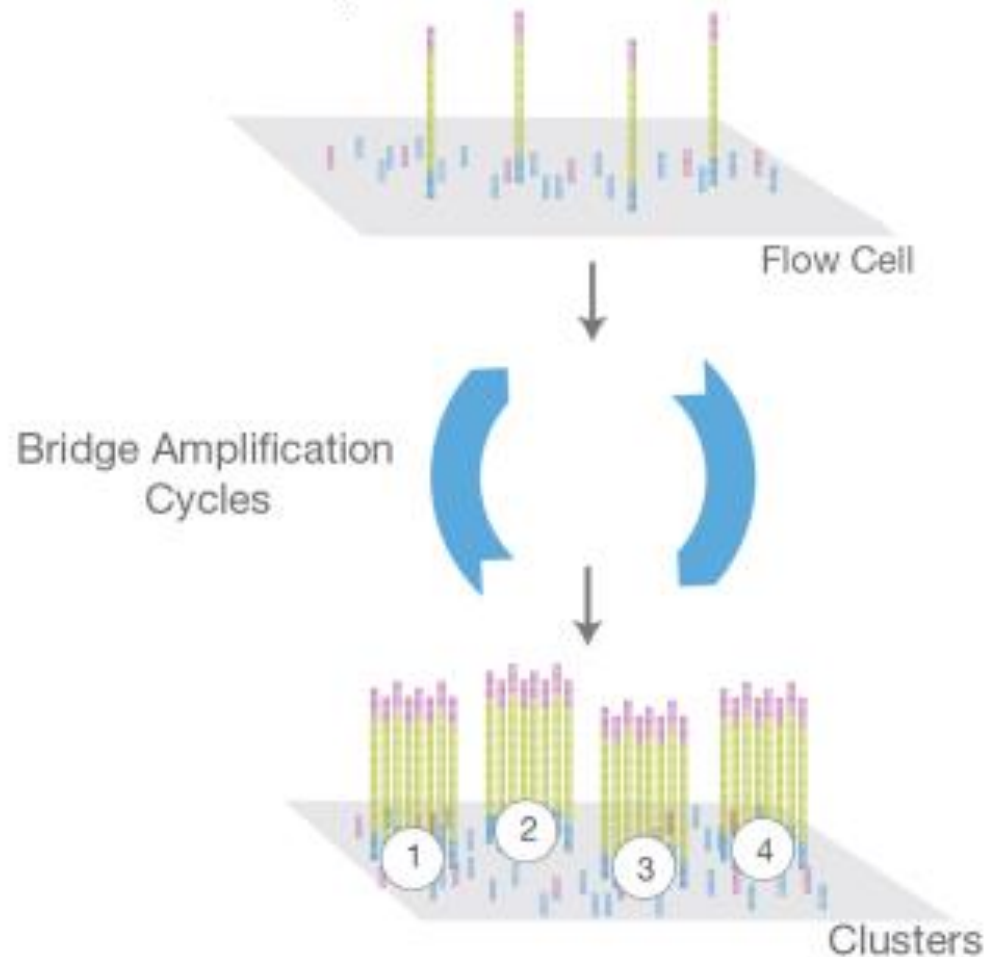


NGS library is prepared by fragmenting a gDNA sample and ligating specialized adapters to both fragment ends.

The sequencing library is prepared by random fragmentation of the DNA or cDNA sample, followed by 5' and 3' adapter ligation.

Adapter-ligated fragments are then PCR amplified and gel purified.

B. Cluster Amplification



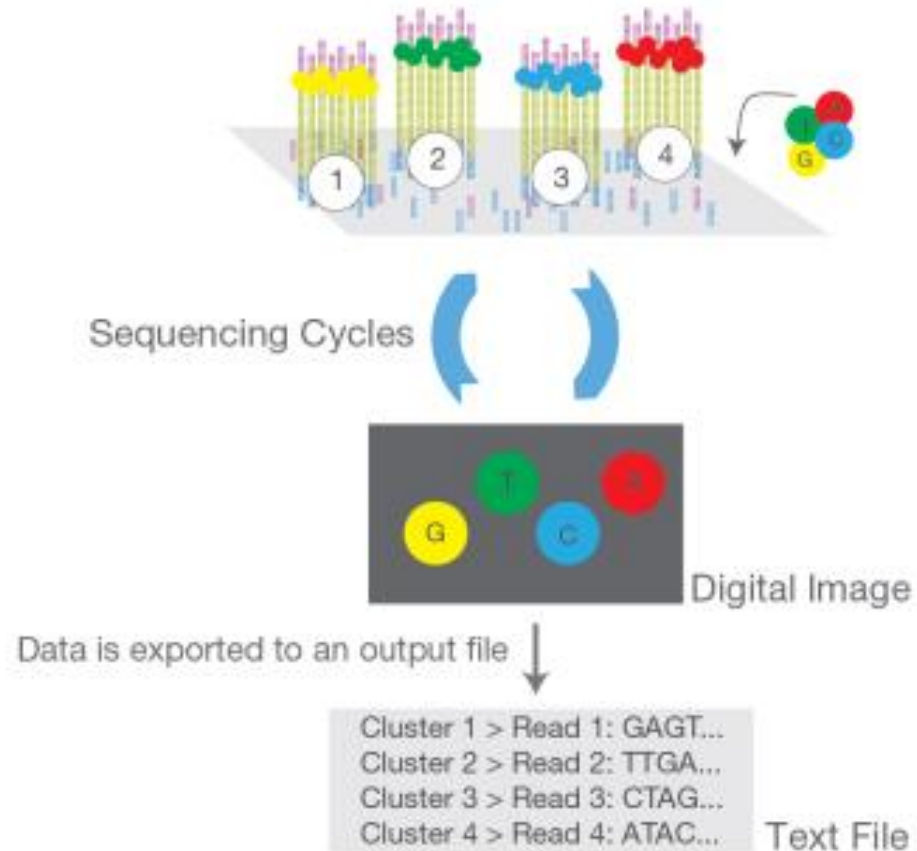
Library is loaded into a flow cell and the fragments are hybridized to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification.

The library is loaded into a flow cell where fragments are captured on a lawn of surface-bound oligos complementary to the library adapters.

Each fragment is then amplified into distinct, clonal clusters through bridge amplification.

When cluster generation is complete, the templates are ready for sequencing.

C. Sequencing



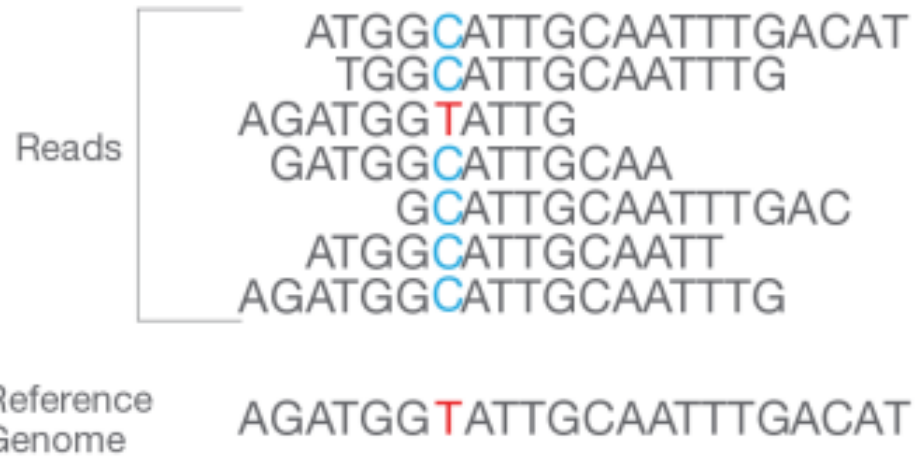
Sequencing reagents, including fluorescently labeled nucleotides, are added and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated “n” times to create a read length of “n” bases.

Illumina SBS technology uses a proprietary reversible terminator–based method that detects single bases as they are incorporated into DNA template strands.

As all four reversible terminator–bound dNTPs are present during each sequencing cycle, natural competition minimizes incorporation bias and greatly reduces raw error rates compared to other technologies.

The result is highly accurate base-by-base sequencing that virtually eliminates sequence context–specific errors, even within repetitive sequence regions and homopolymers.

D. Alignment and Data Analysis



The newly identified sequence reads are aligned to a reference genome.

Reads are aligned to a reference sequence with bioinformatics software. After alignment, differences between the reference genome and the newly sequenced reads can be identified.