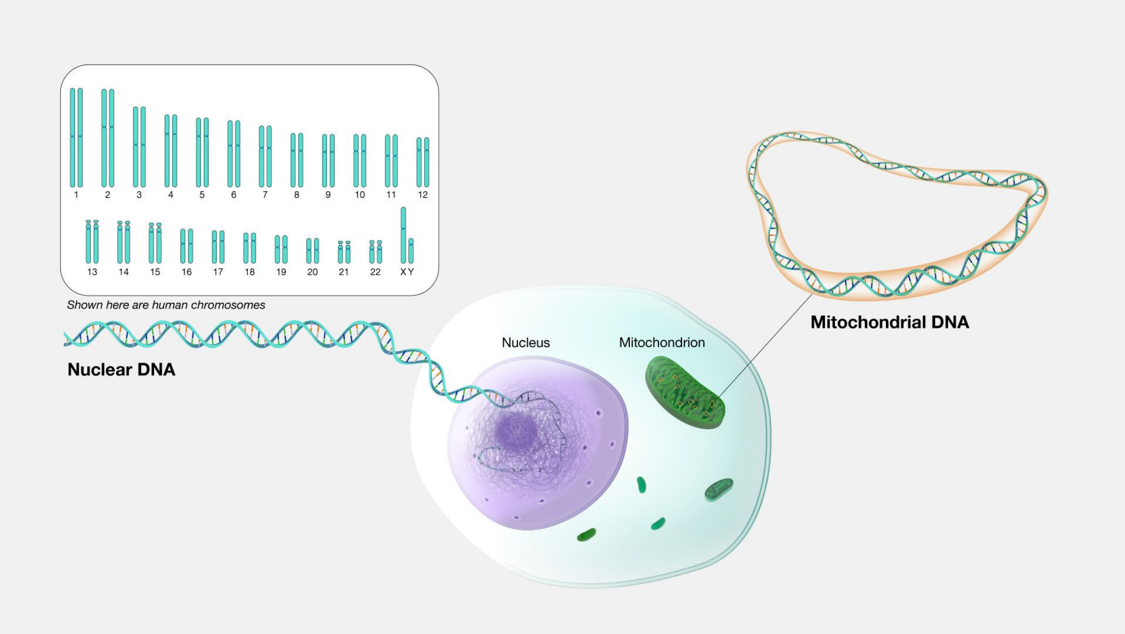
Genome and Sequencing Basics

Genome Biology (BIOL7263)
5Sept24

What is a genome?

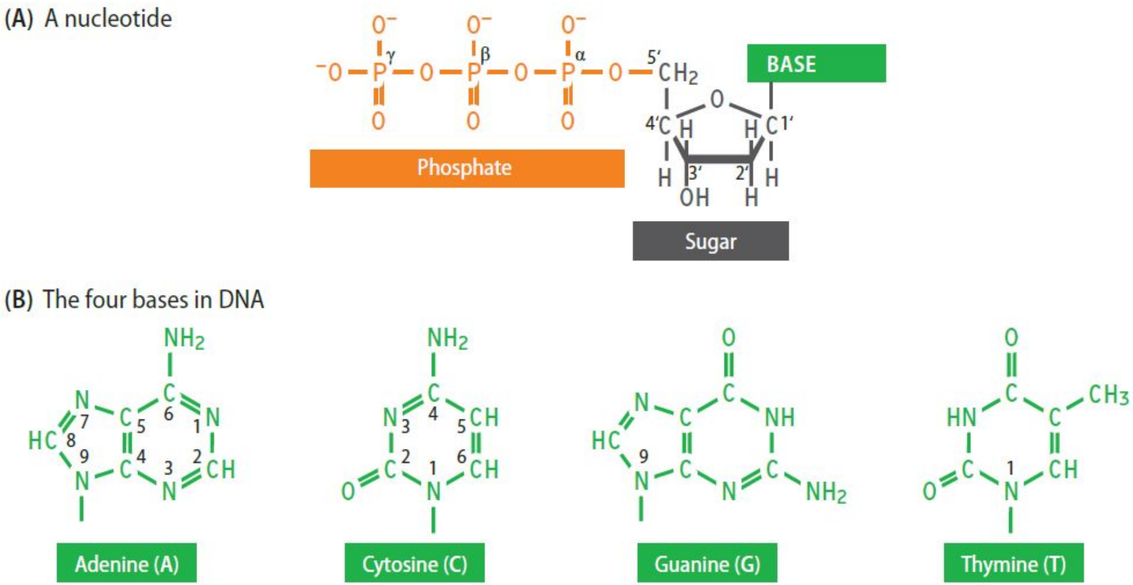
A complete set of genetic information.



What are genomes made of?

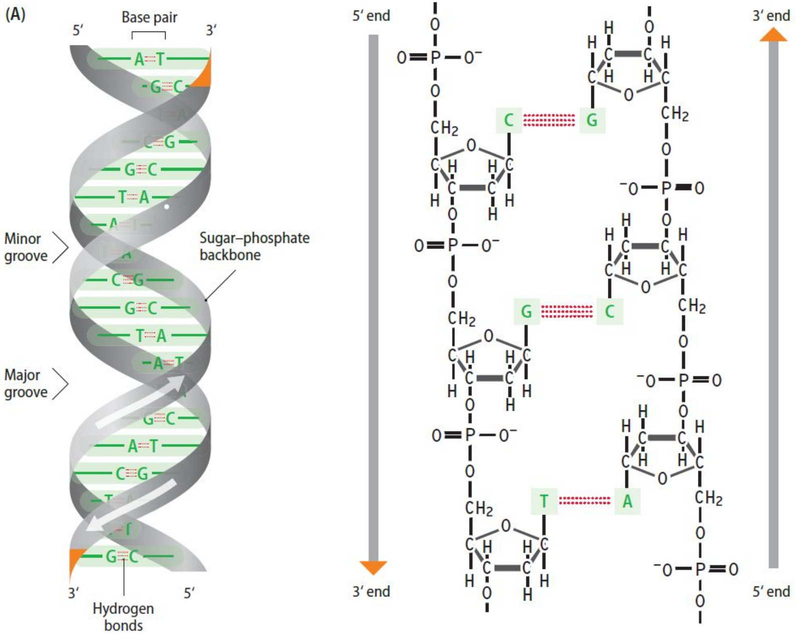
Nucleic Acids

- deoxyribonucleic acid (DNA) - most genomes
- ribonucleic acid (RNA) - some viruses



What are genomes made of?

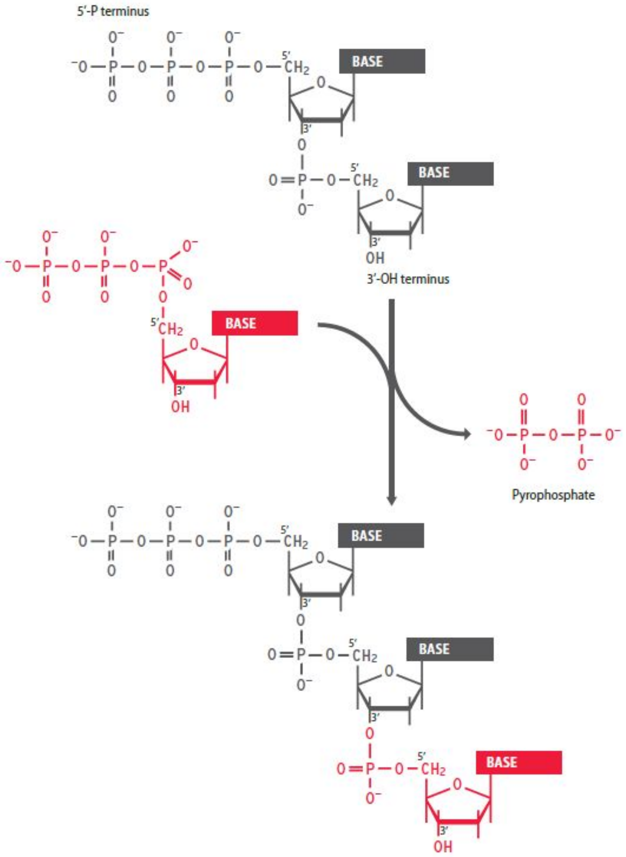
- Nucleotides are assembled into chains (polymers) that are 10s to 10^6 units long.



What are genomes made of?

- Nucleotides are assembled into chains (polymers) that are 10s to 10^6 units long.

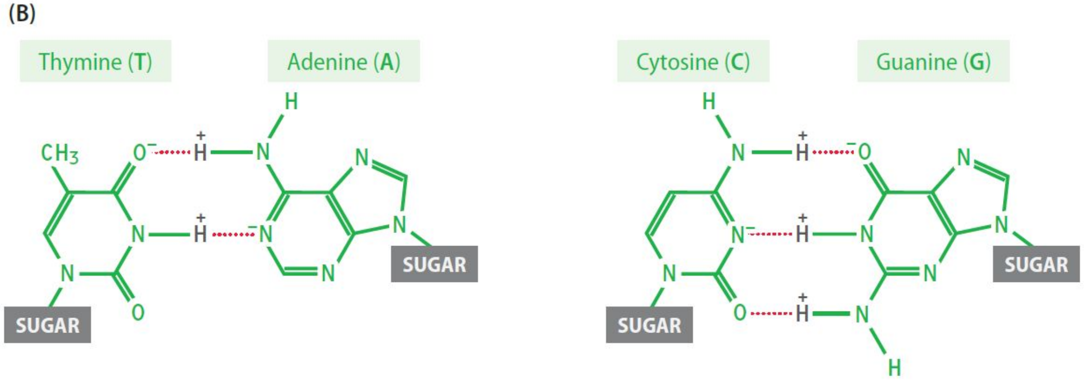
- Nucleotides are linked 5’ to 3’ through the formation of phosphodiester bonds.



DNA is usually a double helix with specific base-pairing

- Strands held together by hydrogen-bonds.

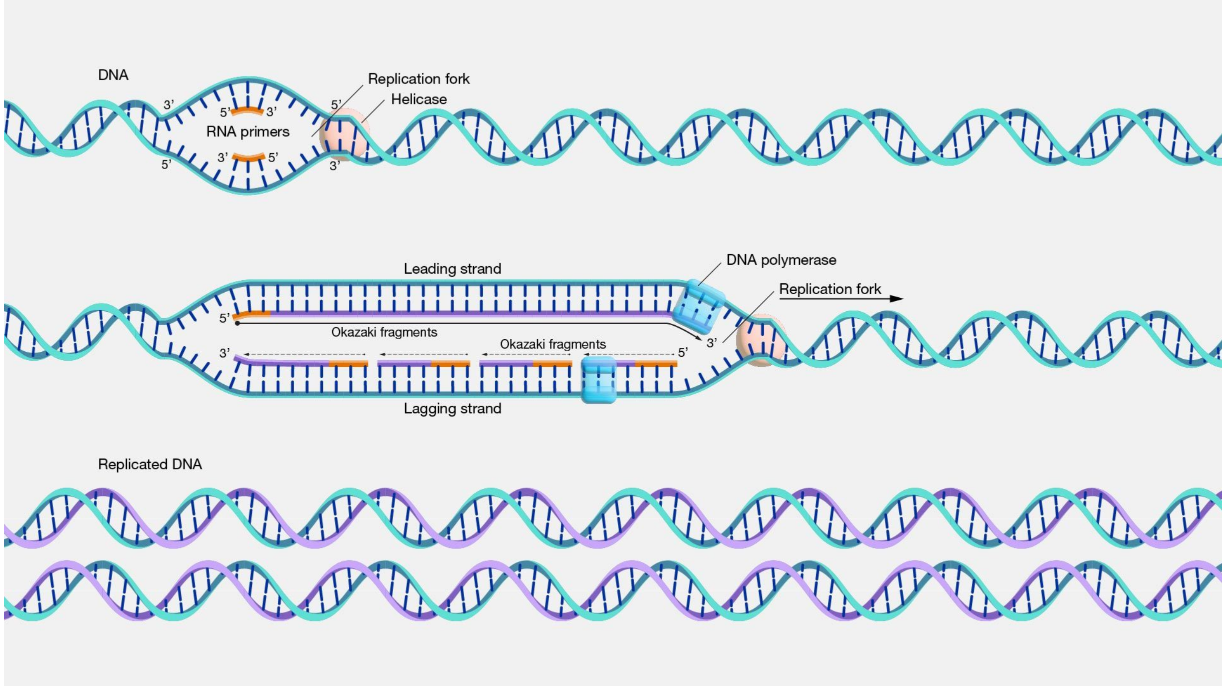
- G-C have a stronger bond
  - High GC content can cause problems for variety of molecular biology techniques.



DNA is usually a double helix with specific base-pairing

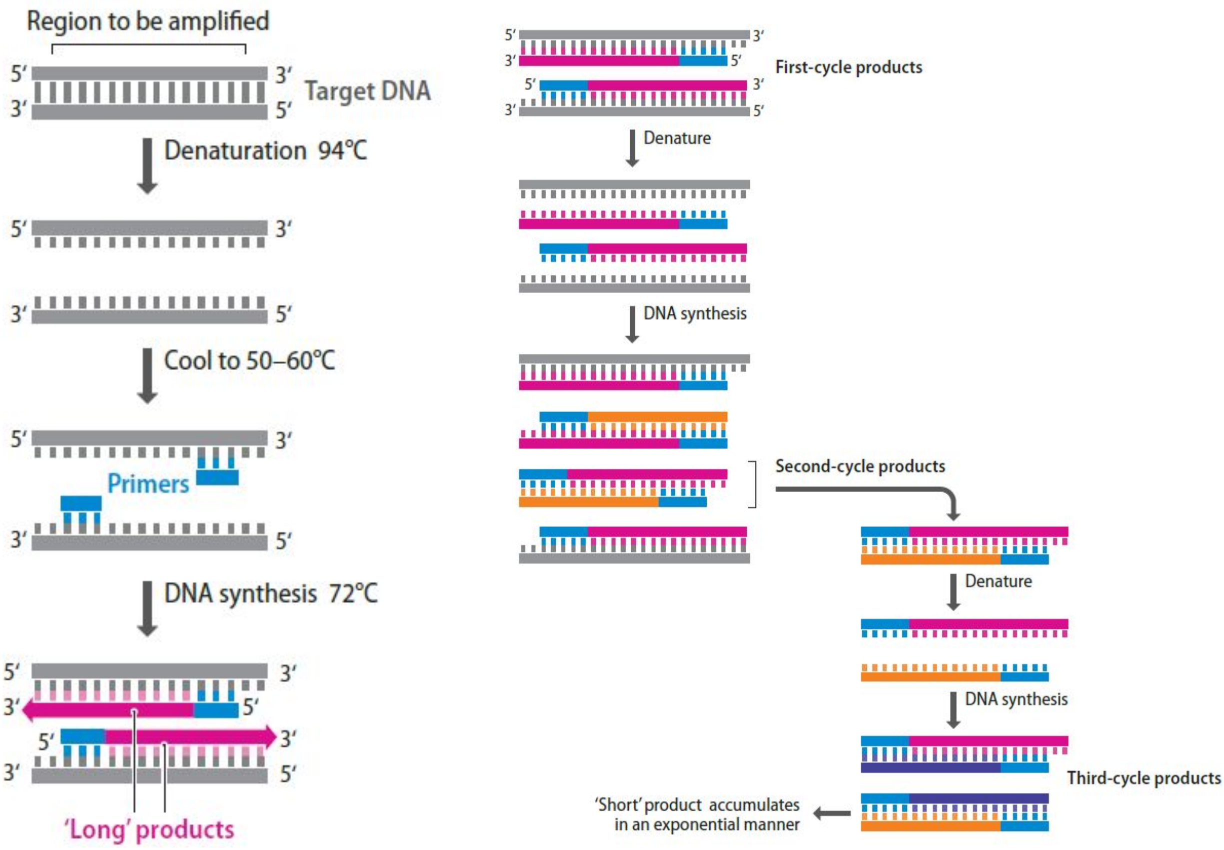
Each strand is entirely dependent of the sequence of the other strand.

Each strand can serve as template for replication.



The structure of DNA allows for in vitro amplification

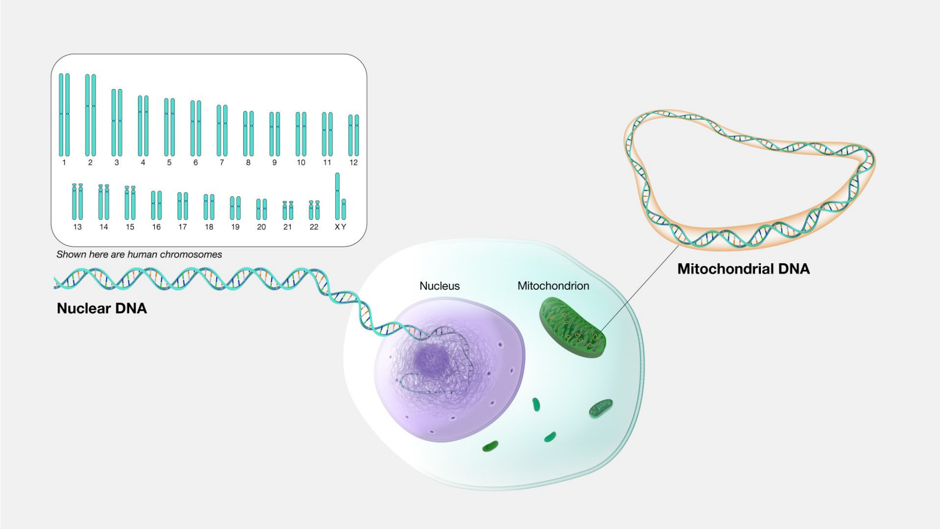
Polymerase chain reaction (PCR) is an essential component of most sequencing technologies



How many genomes?

All eukaryotes carry at least two genomes:

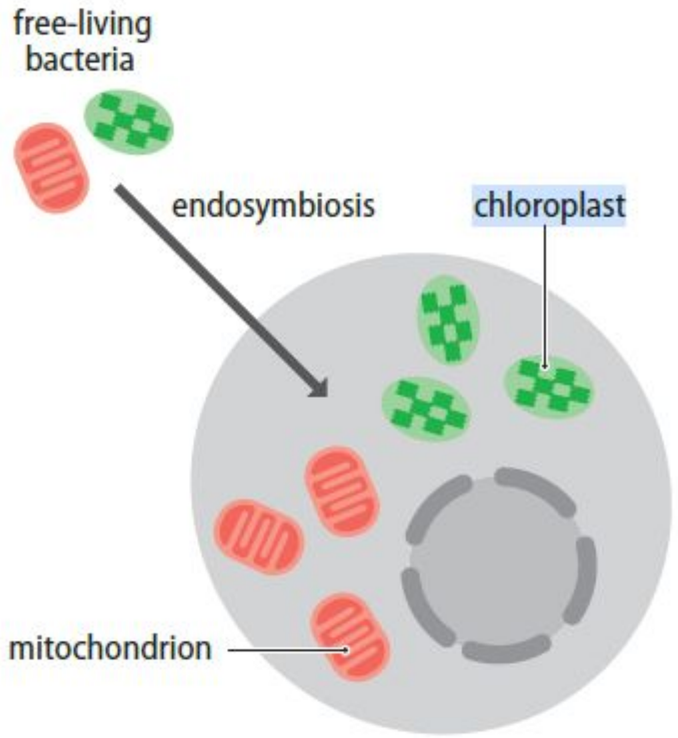
- Nuclear - bi-parental inheritance
- Mitochondria - maternally inherited - prokaryotic-like in structure



How many genomes?

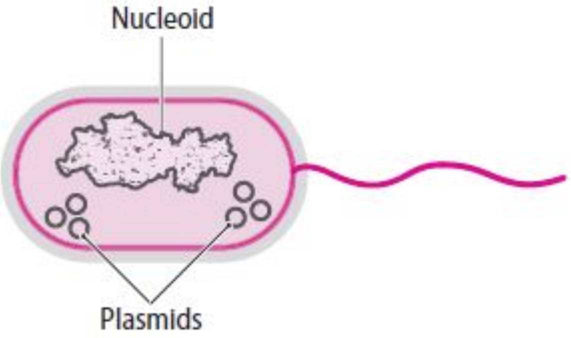
Plants have three genomes:

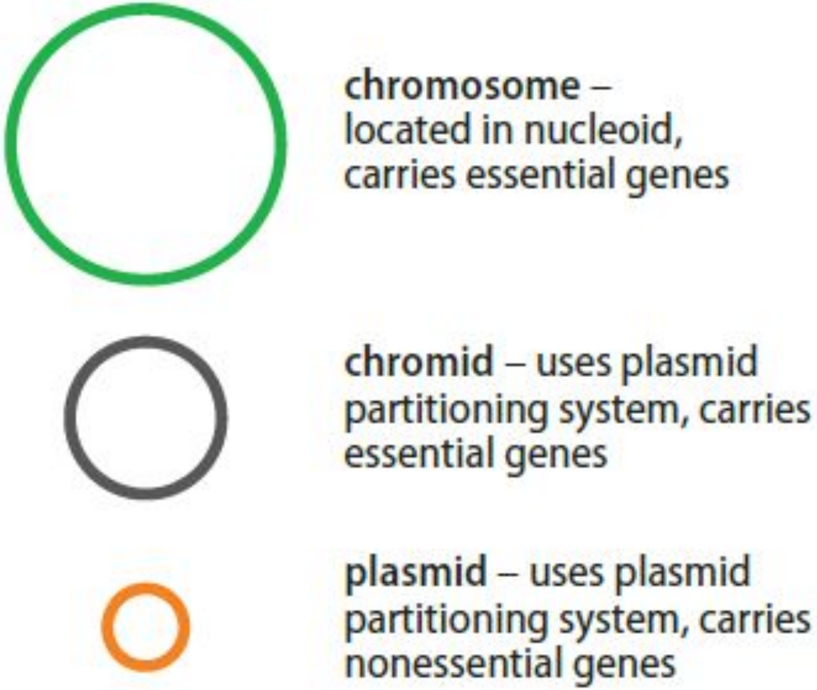
- Nuclear - biparental inheritance
- Mitochondria - maternally inherited - prokaryotic-like in structure
- chloroplast - maternally inherited - prokaryotic-like in structure



How many genomes?

- Prokaryotes





chromosome – located in nucleoid, carries essential genes

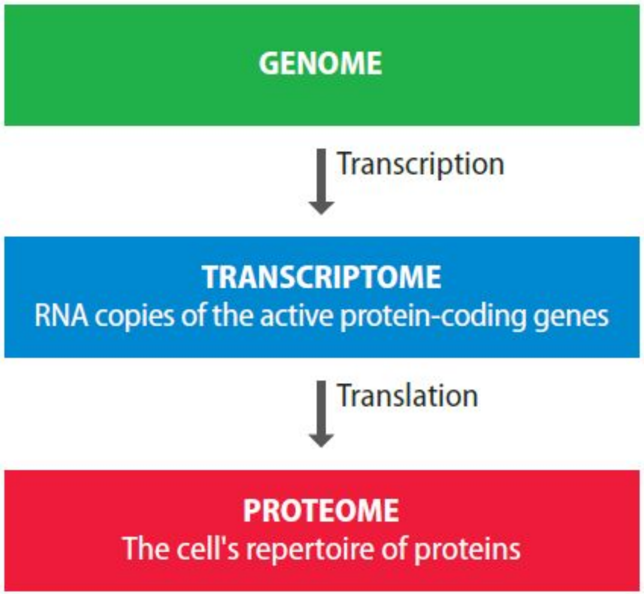
chromid – uses plasmid partitioning system, carries essential genes

plasmid – uses plasmid partitioning system, carries nonessential genes

How do genomes encode information?

Information is primarily encoded in the linear sequence of nucleotides:

- Genes:
    - Portions of the genome the encode instructions to assemble proteins.
    - Gene expression is a multi-step process involving transcription and translation (“Central Dogma”)



GENOME

Transcription

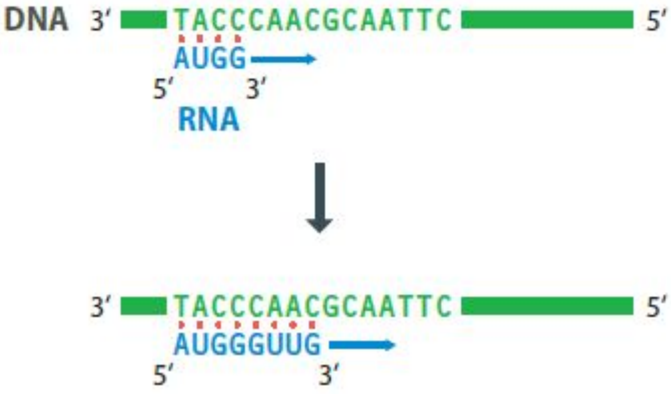
TRANSCRIPTOME
RNA copies of the active protein-coding genes

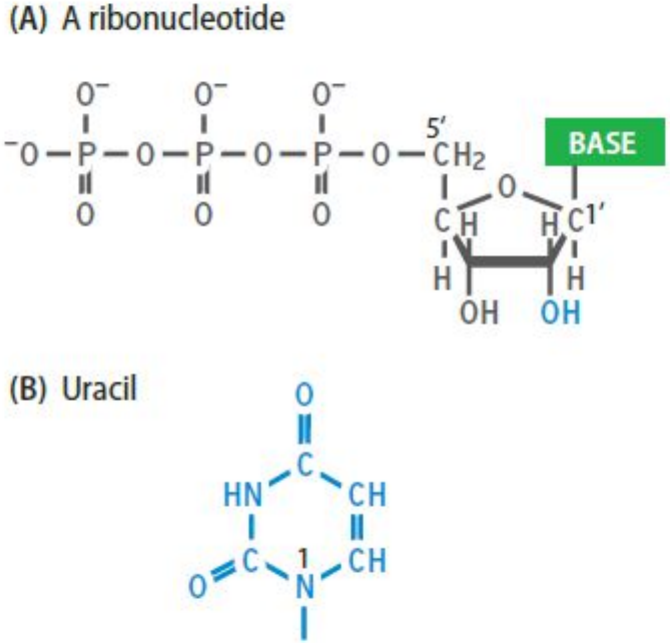
Translation

PROTEOME
The cell's repertoire of proteins

Figure 1.2 Genome expression. The genome specifies the transcriptome, and the transcriptome specifies the proteome.

Translation - DNA to RNA

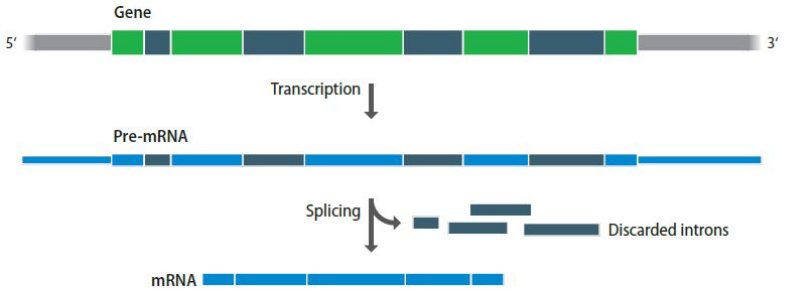


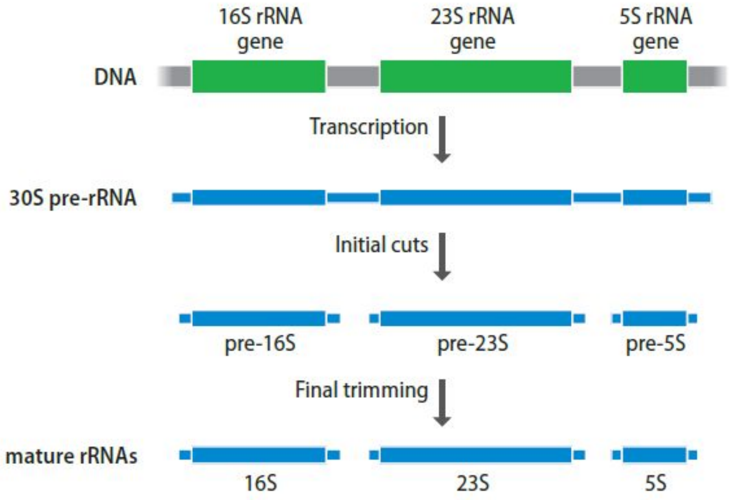

(A) A ribonucleotide

(B) Uracil

Translation - DNA to RNA

RNA is processed in different ways in different organisms and organelles



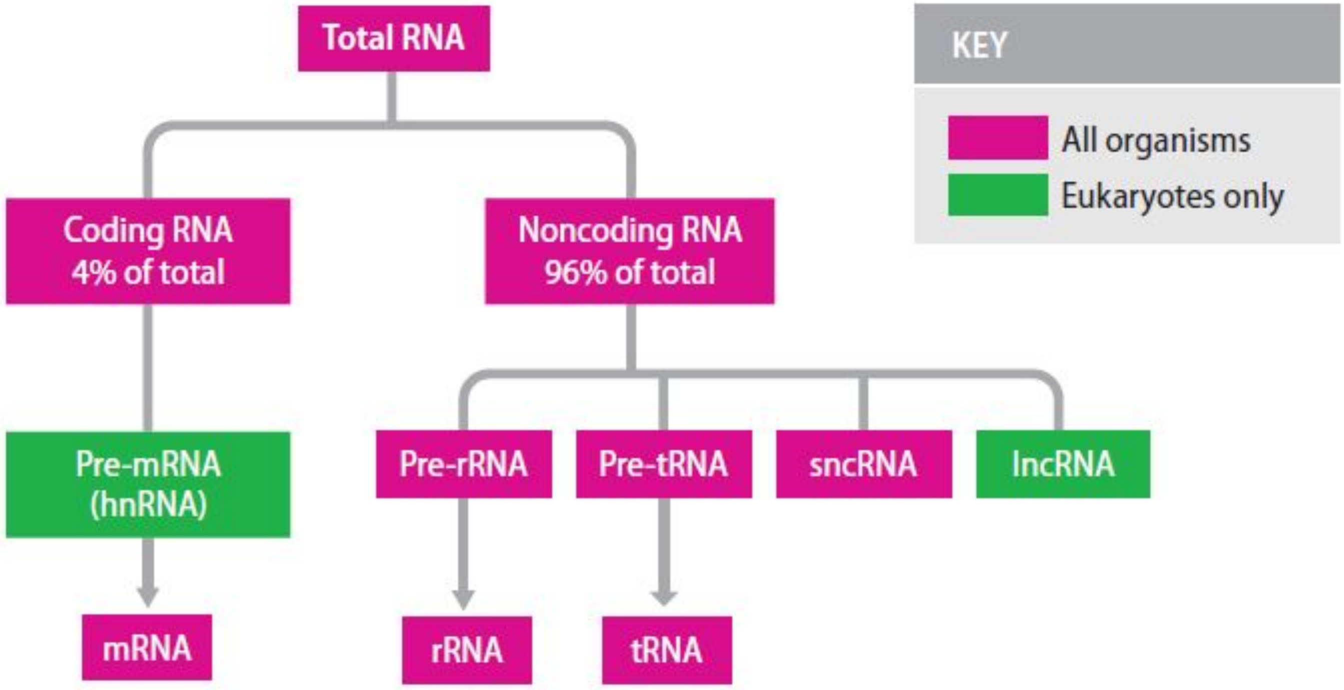


Eukaryotic mRNA

Prokaryotic rRNA

There are many types of RNAs in the cell

- Protein coding mRNAs are relatively rare!!



Total RNA

Coding RNA
4% of total

Noncoding RNA
96% of total

Pre-mRNA
(hnRNA)

Pre-rRNA

Pre-tRNA

sncRNA

lncRNA

mRNA

rRNA

tRNA

KEY

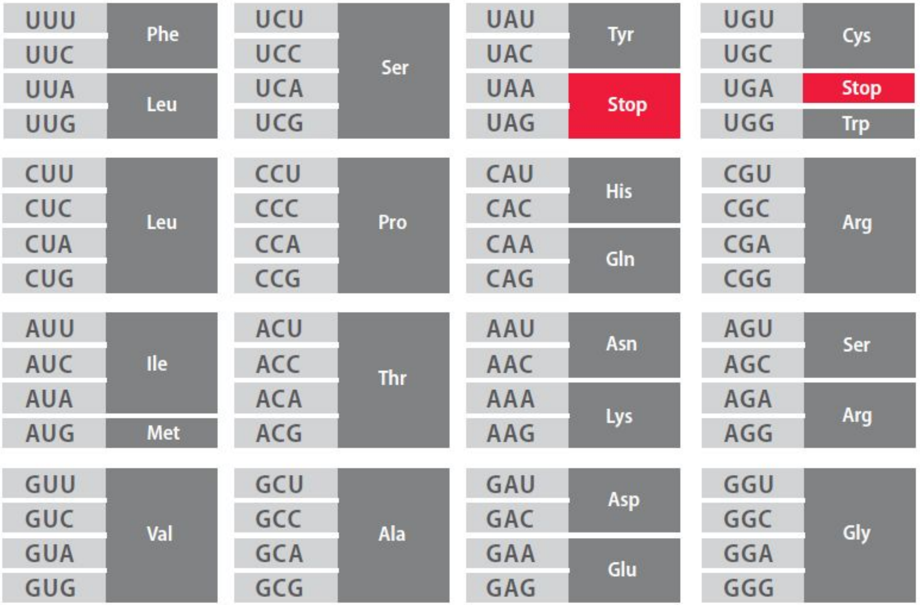
🟪 All organisms

🟩 Eukaryotes only

How do genomes encode information?

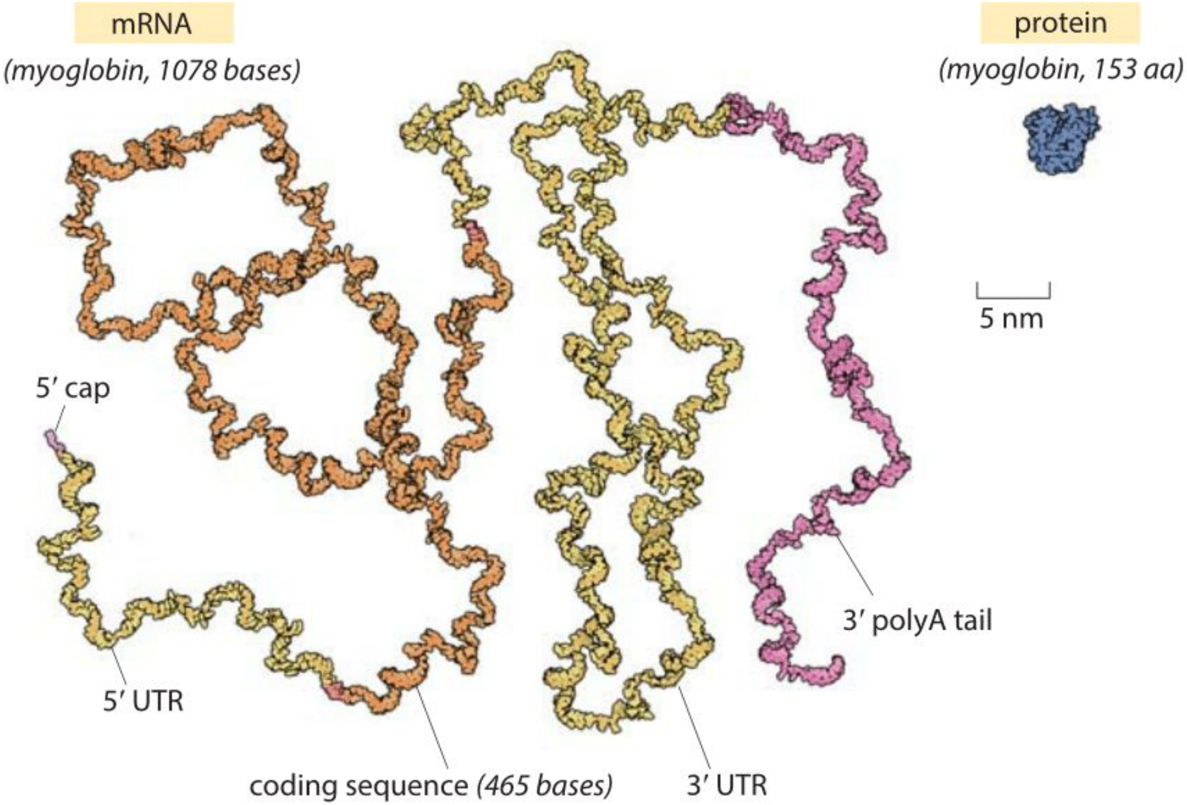
Information is primarily encoded in the linear sequence of nucleotides:

- Genetic code is triplicate - 3 nucleotides = 1 amino acid



UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC		UCC		UAC		UGC	
UUA	Leu	UCA		UAA	Stop	UGA	Stop
UUG		UCG		UAG		UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC		CCC		CAC		CGC	
CUA		CCA		CAA	Gln	CGA	
CUG		CCG		CAG		CGG	
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC		ACC		AAC		AGC	
AUA		ACA		AAA	Lys	AGA	Arg
AUG	Met	ACG		AAG		AGG	
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC		GCC		GAC		GGC	
GUA		GCA		GAA	Glu	GGA	
GUG		GCG		GAG		GGG	

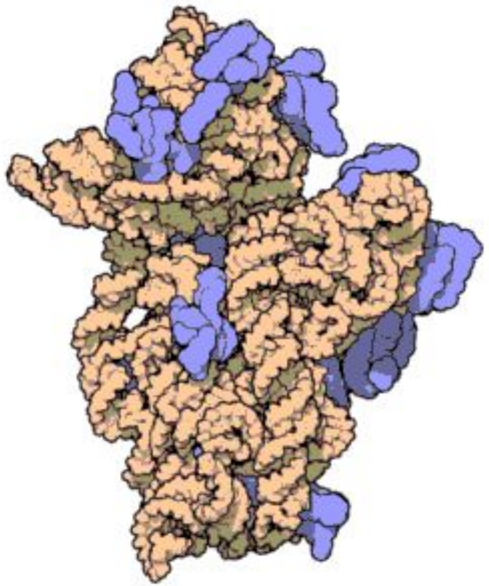
Transcripts are huge! Proteins are small



How do genomes encode information?

Information is primarily encoded in the linear sequence of nucleotides:

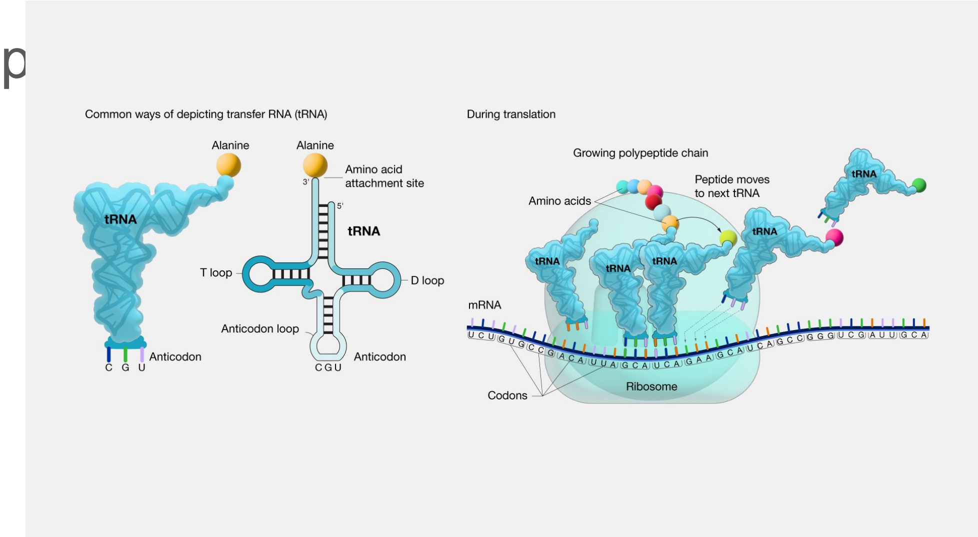
- rRNA - not translated - structural and functional elements of the ribosome



How do genomes encode information?

Information is primarily encoded in the linear sequence of nucleotides:

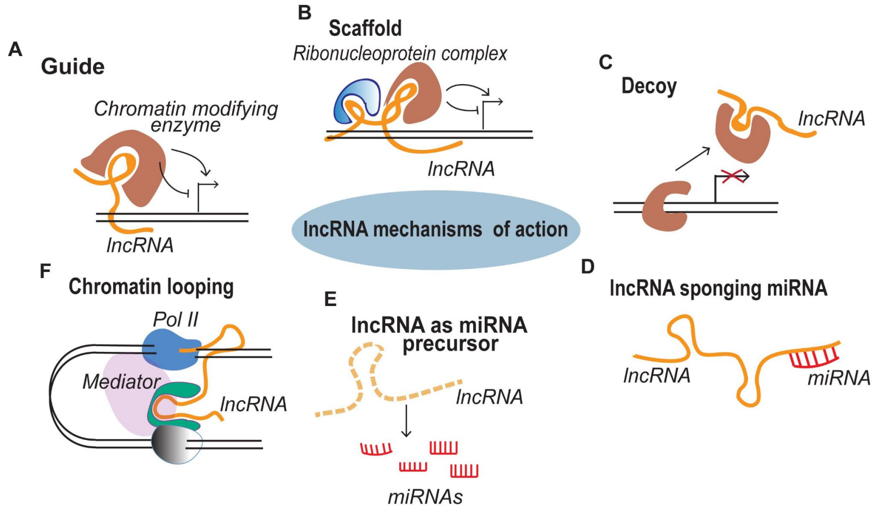
- tRNA - not translated - adap for the process translation


Common ways of depicting transfer RNA (tRNA)
Alanine
tRNA
Anticodon
C G U
Alanine
3'
Amino acid attachment site
5'
tRNA
T loop
D loop
Anticodon loop
Anticodon
C G U
During translation
Growing polypeptide chain
Peptide moves to next tRNA
Amino acids
tRNA
tRNA
tRNA
tRNA
tRNA
tRNA
mRNA
U C U G U G C C G A C A U U A G C A U C A G A A G C A U C A G C C G G G U C G A U U G C A
Codons
Ribosome

How do genomes encode information?

Information is primarily encoded in the linear sequence of nucleotides:

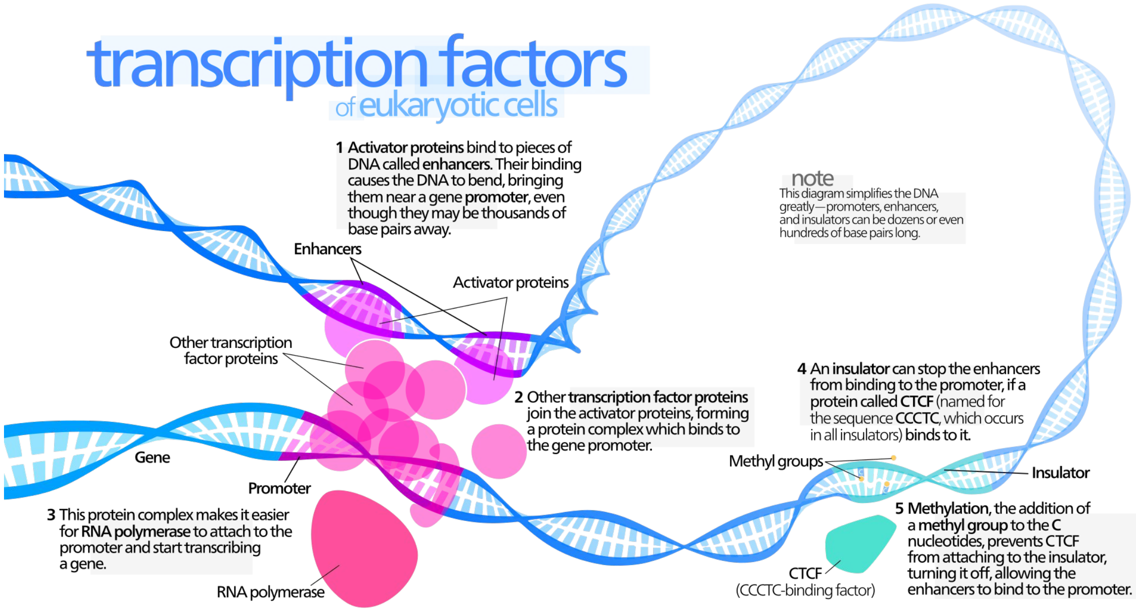
- Noncoding RNAs - incompletely understood roles regulating transcription, translation, and cellular physiology



How do genomes encode information?

Cis-regulatory elements

- Promoters
- Terminators
- Enhancers
- Repressors
- Insulators



transcription factors of eukaryotic cells

1 **Activator proteins** bind to pieces of DNA called **enhancers.** Their binding causes the DNA to bend, bringing them near a gene **promoter,** even though they may be thousands of base pairs away.

2 Other **transcription factor proteins** join the activator proteins, forming a protein complex which binds to the gene promoter.

3 This protein complex makes it easier for **RNA polymerase** to attach to the promoter and start transcribing a gene.

4 An **insulator** can stop the enhancers from binding to the promoter, if a **protein called CTCF** (named for the sequence **CCCTC,** which occurs in all insulators) **binds to it.**

5 **Methylation,** the addition of a **methyl group** to the **C** nucleotides, prevents CTCF from attaching to the insulator, turning it off, allowing the enhancers to bind to the promoter.

note
This diagram simplifies the DNA greatly—promoters, enhancers, and insulators can be dozens or even hundreds of base pairs long.

By Kelvin13 - Own work, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=23272278

Genome Sequencing

Three major approaches

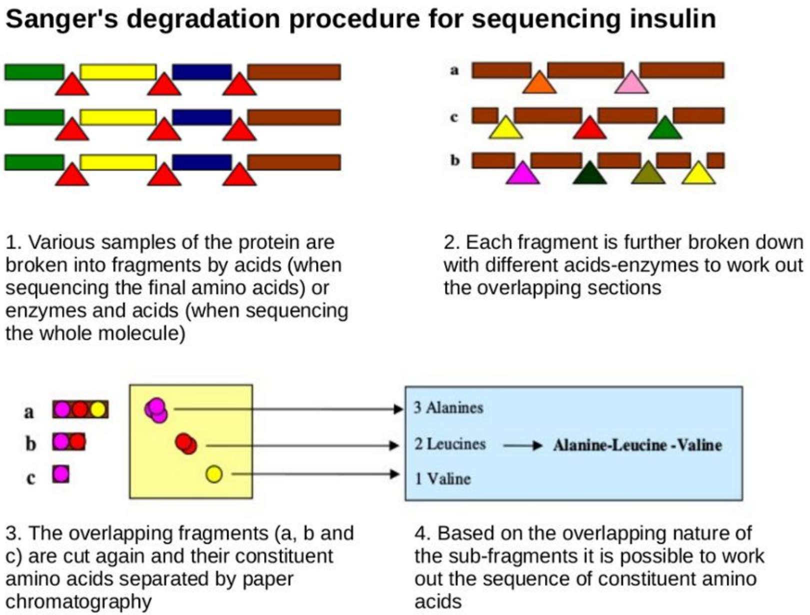
1. Chain-termination (Sanger)
2. Short-read (Illumina)
3. Long-read (PacBio and Nanopore)

Genome Sequencing - terminology

- **Read** - A single sequence from one fragment in the sequencing library (one cluster, bead, etc.)
- **Library** - A collection of DNA fragments that have been prepared to be sequenced
- **Coverage** - number of reads spanning a region of the genome

Fred Sanger (1918-2013)

- Two-time nobel prize winner
- 1958 - Protein sequencing
- 1980 - DNA sequencing



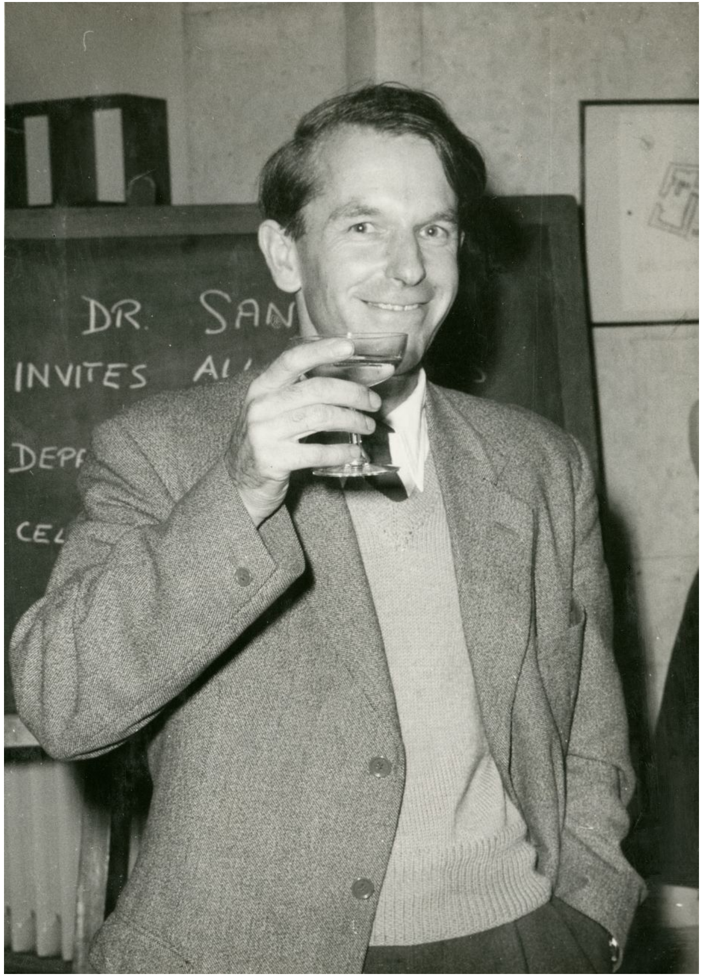
Sanger's degradation procedure for sequencing insulin

1. Various samples of the protein are broken into fragments by acids (when sequencing the final amino acids) or enzymes and acids (when sequencing the whole molecule)

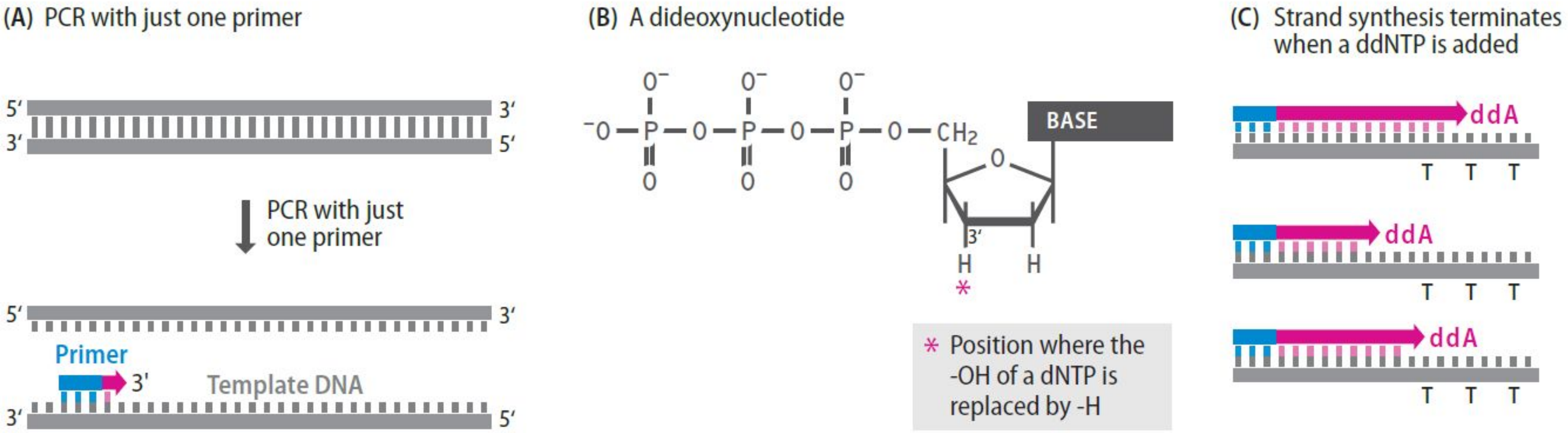
2. Each fragment is further broken down with different acids-enzymes to work out the overlapping sections

3. The overlapping fragments (a, b and c) are cut again and their constituent amino acids separated by paper chromatography

4. Based on the overlapping nature of the sub-fragments it is possible to work out the sequence of constituent amino acids

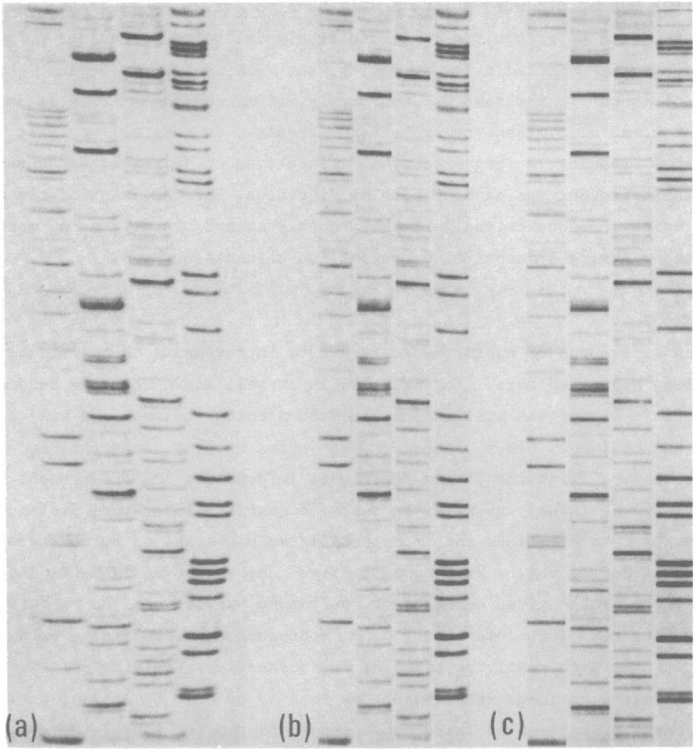


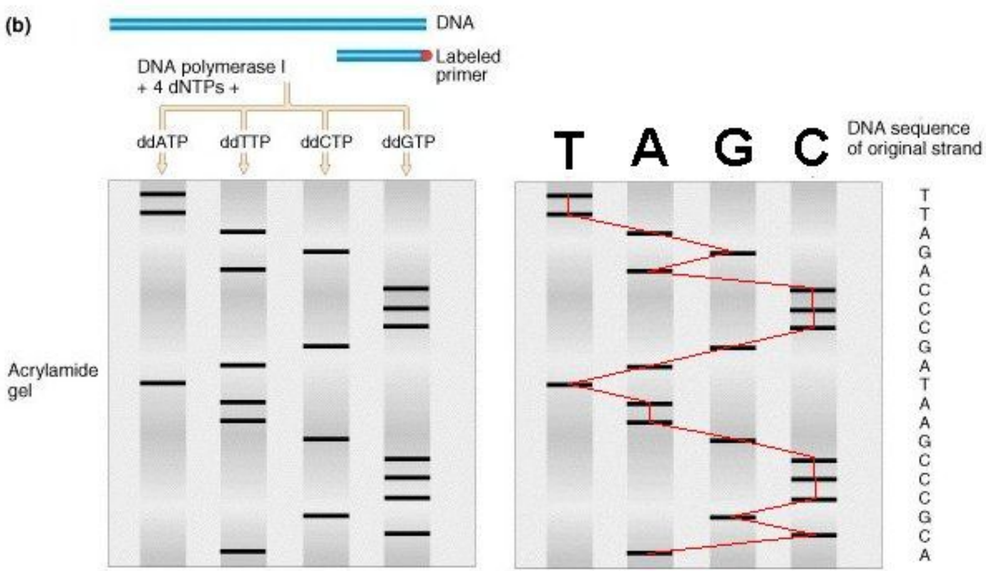
Chain-termination (Sanger)



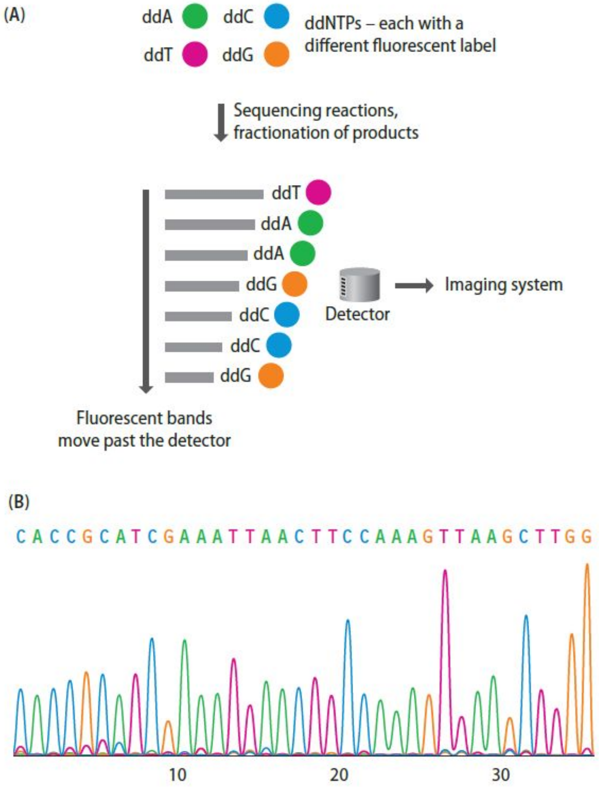
terminator nucleotide

DNA autoradiograph



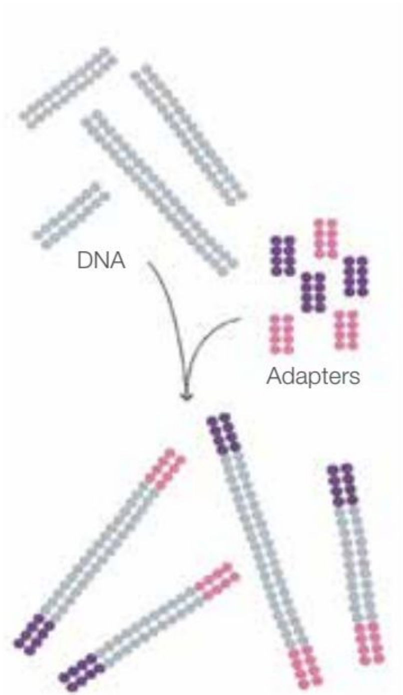


Chain-termination (Sanger) - Modern


(A)
ddA
ddC
ddT
ddG
ddNTPs – each with a different fluorescent label
Sequencing reactions, fractionation of products
ddT
ddA
ddA
ddG
ddC
ddC
ddG
Detector
Imaging system
Fluorescent bands move past the detector
(B)
CACCGCATCGAAATTAACTTCCAAAGTTAAGCTTGG
10
20
30

- Read length: 700-1000 bp
- Throughput: 96 reads per run (~1 run per hour)
- Error rate: < 1/100,000 bases
- Cost: $1 per read

Short-read sequencing (Illumina)



Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

Short-read sequencing (Illumina)

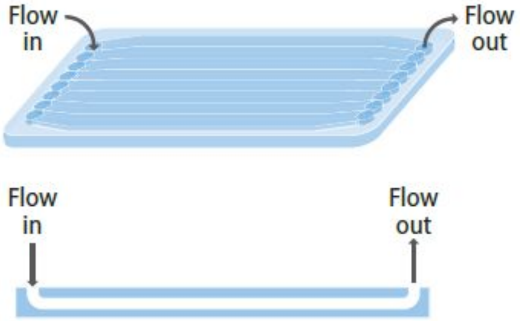
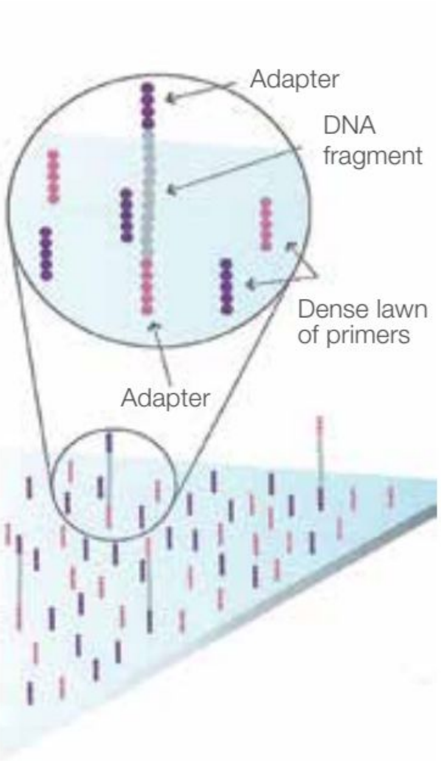
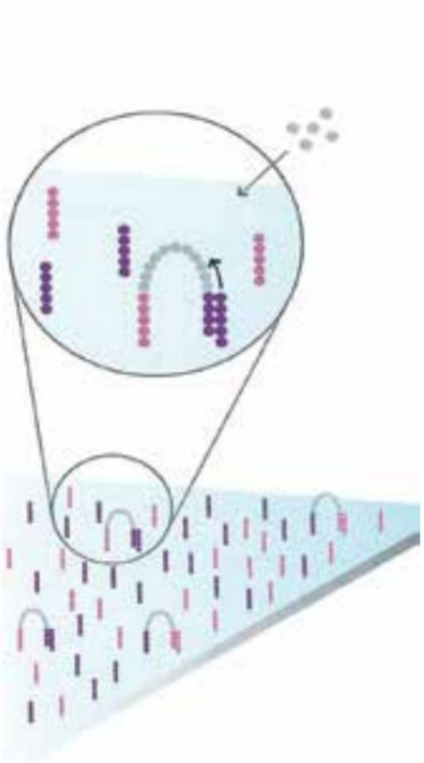


Figure 4.7 A typical flow cell used in DNA sequencing. The sequencing library is immobilized within the channels of the flow cell. To carry out the sequencing reactions, the necessary reagents flow through



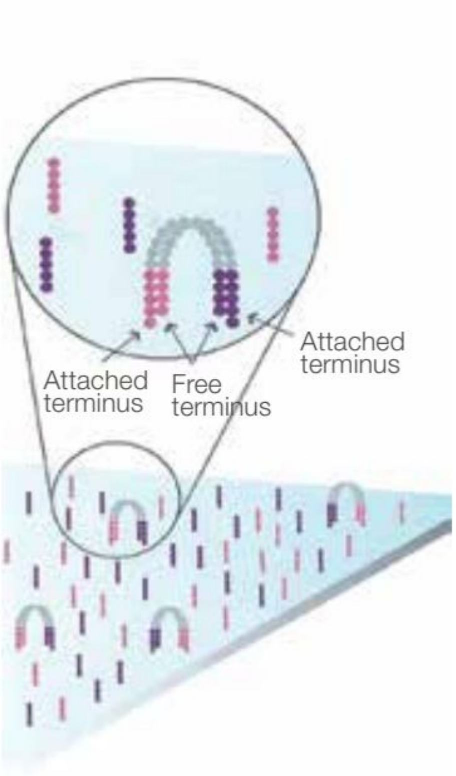
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

Short-read sequencing (Illumina)



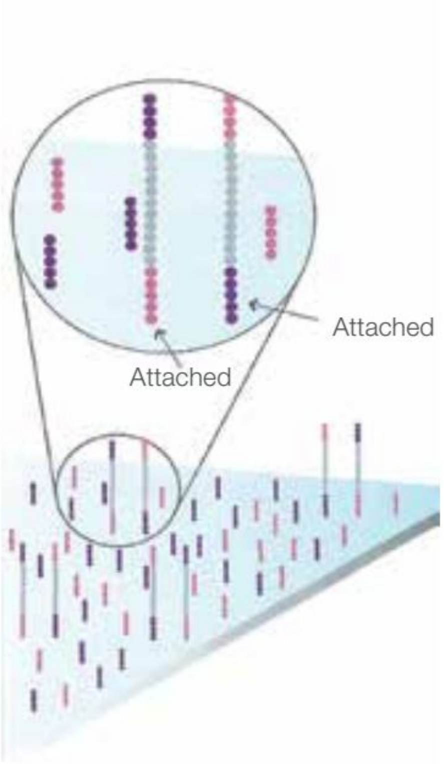
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

Short-read sequencing (Illumina)



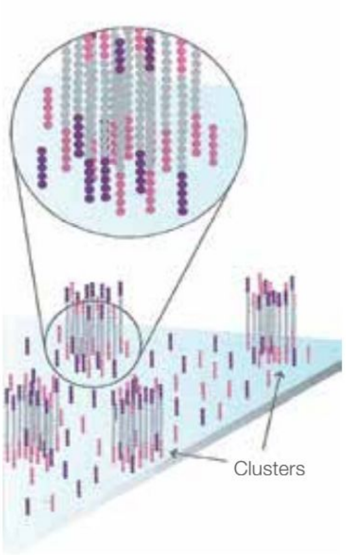
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

Short-read sequencing (Illumina)

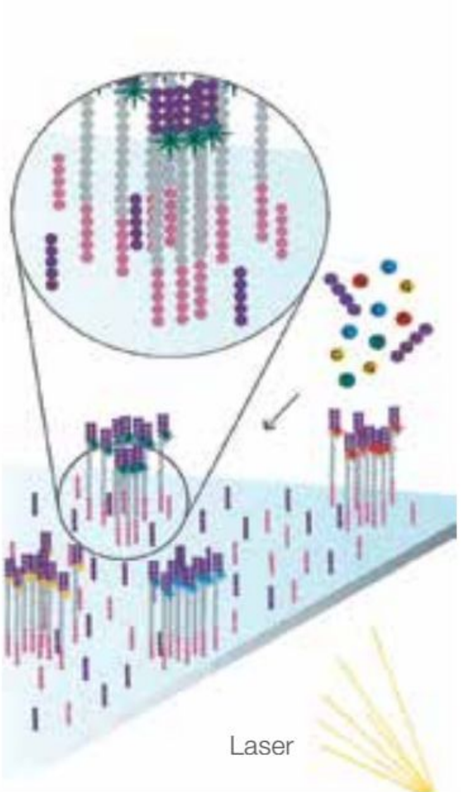


Denaturation leaves single-stranded templates anchored to the substrate.

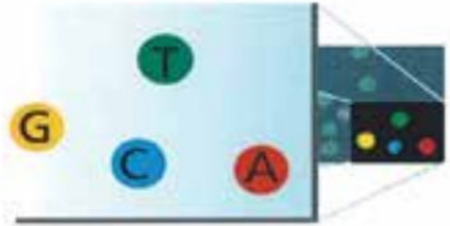
- Bridge PCR repeated 35x to create clusters of identical fragments



Short-read sequencing (Illumina)

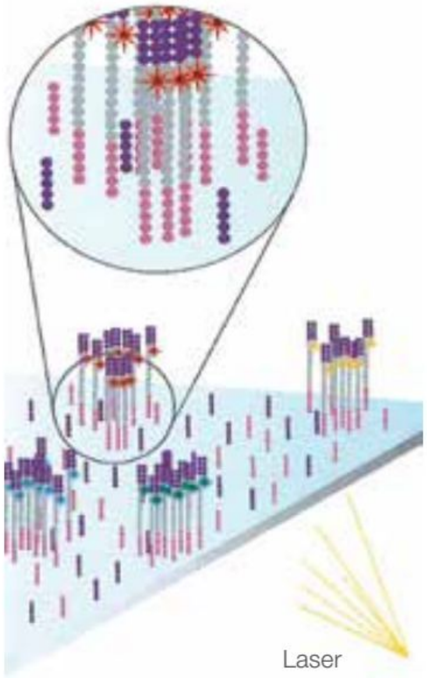


ſhe first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.

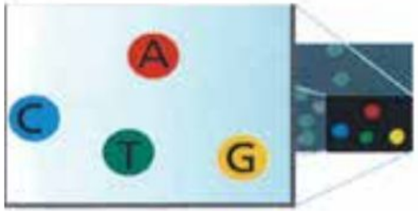


After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.

Short-read sequencing (Illumina)

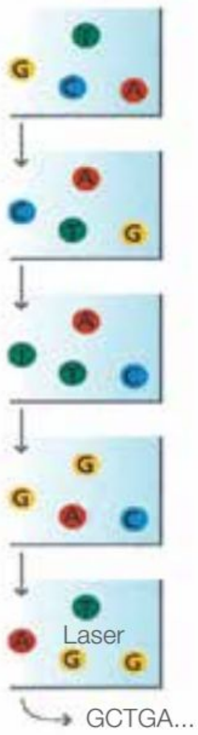


The next cycle repeats the incorporation of four labeled reversible terminators, primers, and DNA polymerase.



After laser excitation, the image is captured as before, and the identity of the second base is recorded.

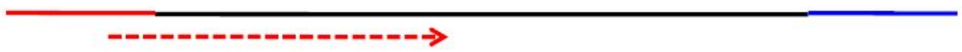
Short-read sequencing (Illumina)



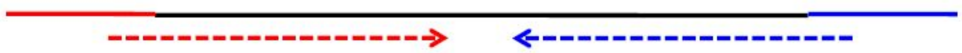
The sequencing cycles are repeated to determine the sequence of bases in a fragment, one base at a time.

Fragments can be sequenced from either end or both

- Fragment reads (come from fragment libraries)
  - Single read in one direction from a fragment



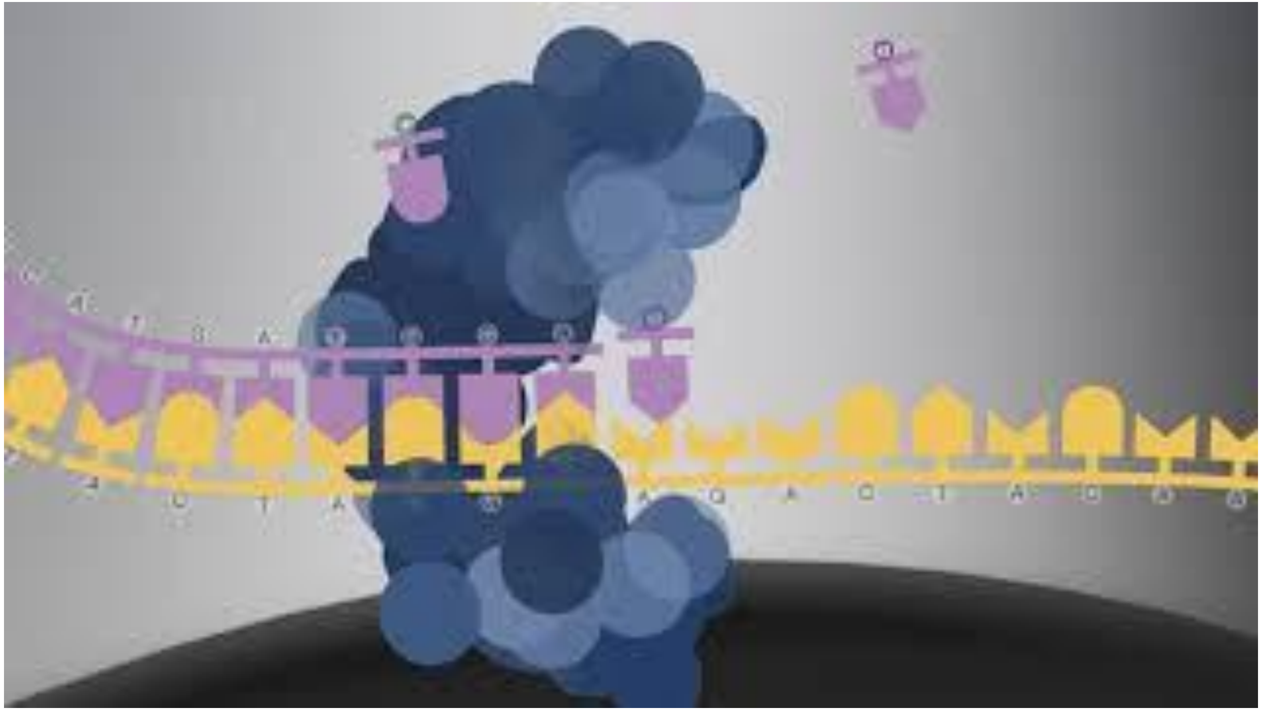
- Paired end reads (come from fragment libraries)
  - Two reads from opposite ends of the same fragment
  - Reads point towards each other



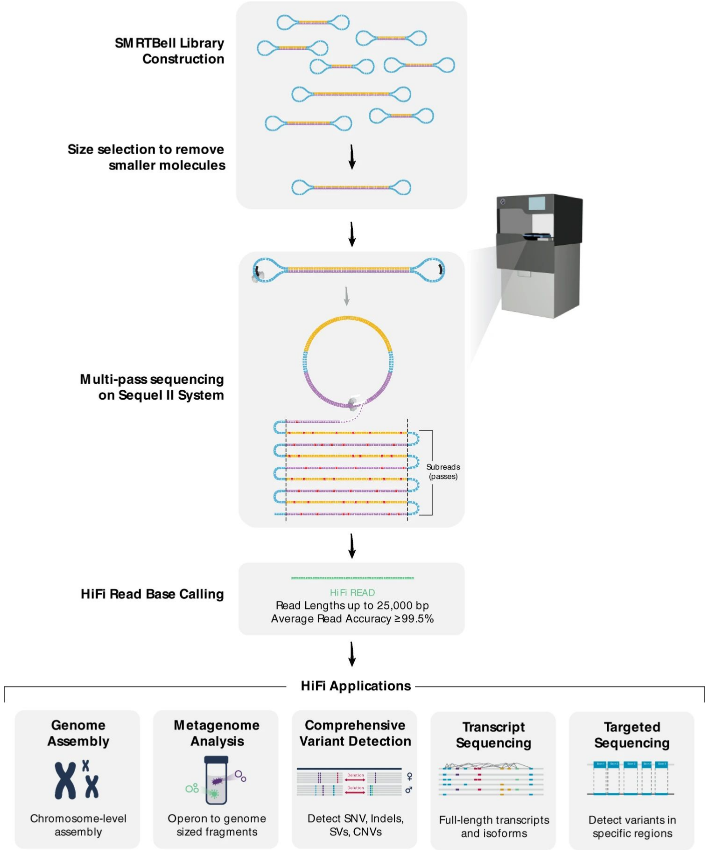
Short-read sequencing (Illumina)

- Read length: 100-300 bp
- Throughput: 25 billion read pairs per run
- Error rate: ~ 1/100 - 1/1000 bases
- Cost: $0.00000005 per read

Long-read sequencing - Pacific Bioscience



Long-read sequencing - Pacific Bioscience



SMRTBell Library Construction

Size selection to remove smaller molecules

Multi-pass sequencing on Sequel II System

HiFi Read Base Calling

HiFi READ

Read Lengths up to 25,000 bp

Average Read Accuracy ≥99.5%

HiFi Applications

Genome Assembly

Chromosome-level assembly

Metagenome Analysis

Operon to genome sized fragments

Comprehensive Variant Detection

Detect SNV, Indels, SVs, CNVs

Transcript Sequencing

Full-length transcripts and isoforms

Targeted Sequencing

Detect variants in specific regions

- Read length: 10 - 20 kb
- Throughput: up to 4 million reads per run
- Error rate: ~ 1/1000 bases
- Cost: $0.005 per read

Long-read sequencing - Oxford Nanopore

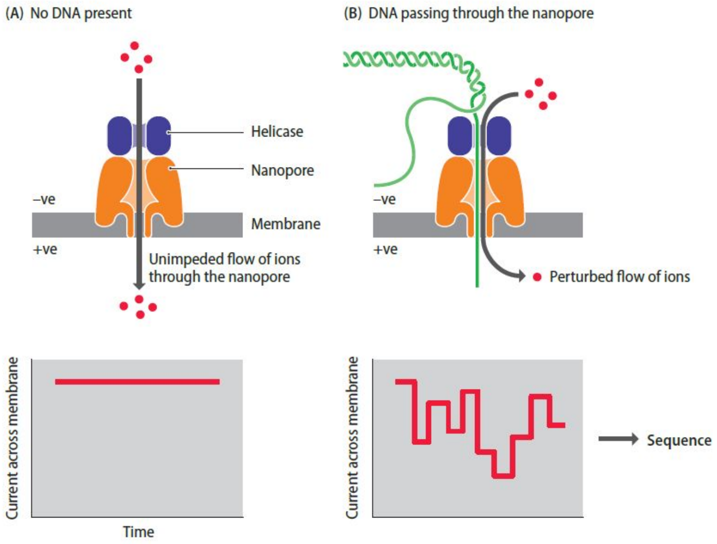


Figure 4.15 Nanopore sequencing. (A) In the absence of DNA, the flow of ions through the nanopore is unimpeded and the electrical current across the membrane is constant. (B) Passage of a polynucleotide through the nanopore perturbs the ion flow. Each nucleotide, or combination of adjacent nucleotides, perturbs the ion flow in a different way, resulting in fluctuations in the current from which the DNA sequence can be deduced.

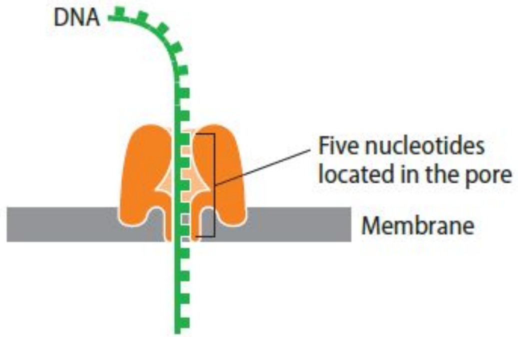
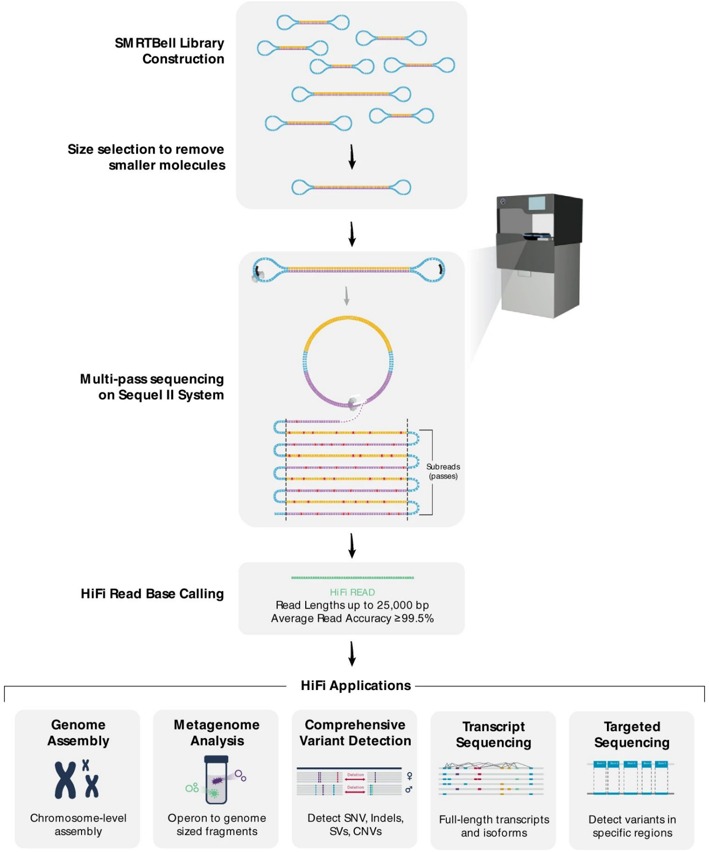


Figure 4.16 More than one nucleotide is present in the nanopore at a single time.

Long-read sequencing - Oxford Nanopore



SMRTBell Library Construction

Size selection to remove smaller molecules

Multi-pass sequencing on Sequel II System

Subreads (passes)

HiFi Read Base Calling

HiFi READ
Read Lengths up to 25,000 bp
Average Read Accuracy ≥99.5%

HiFi Applications

Genome Assembly — Chromosome-level assembly

Metagenome Analysis — Operon to genome sized fragments

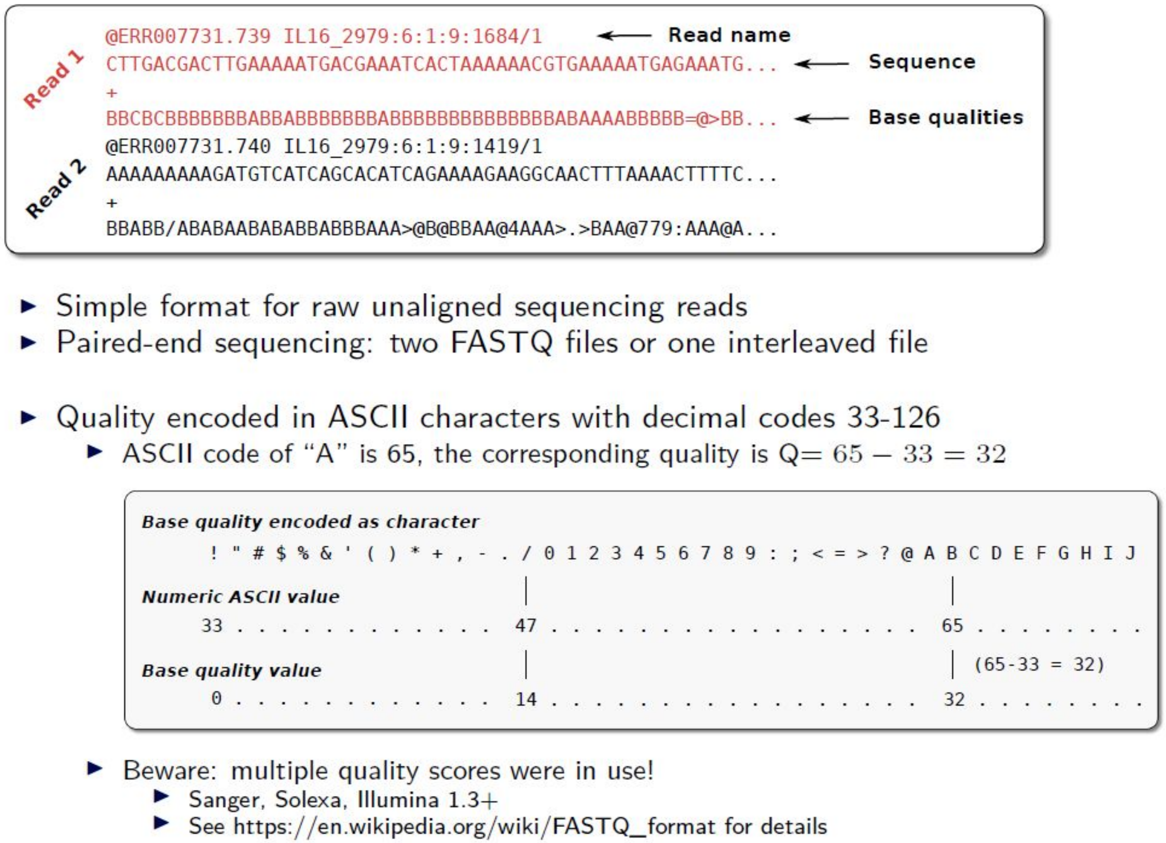
Comprehensive Variant Detection — Detect SNV, Indels, SVs, CNVs

Transcript Sequencing — Full-length transcripts and isoforms

Targeted Sequencing — Detect variants in specific regions

- Read length: >1 Mbs
- Throughput: 7-12 million reads
- Error rate: ~ 1/10 bases
- Cost: $0.005 per read

Sequencing data - FastQ file



```
@ERR007731.739 IL16_2979:6:1:9:1684/1
CTTGACGACTTGAAAAATGACGAAATCACTAAAAAACGTGAAAAATGAGAAATG...
+
BBCBCBBBBBBBABBABBBBBBBABBBBBBBBBBBBBBABAAAABBBBB=@>BB...
@ERR007731.740 IL16_2979:6:1:9:1419/1
AAAAAAAAAGATGTCATCAGCACATCAGAAAAGAAGGCAACTTTAAAACTTTTC...
+
BBABB/ABABAABABABBABBBAAA>@B@BBAA@4AAA>.>BAA@779:AAA@A...
```

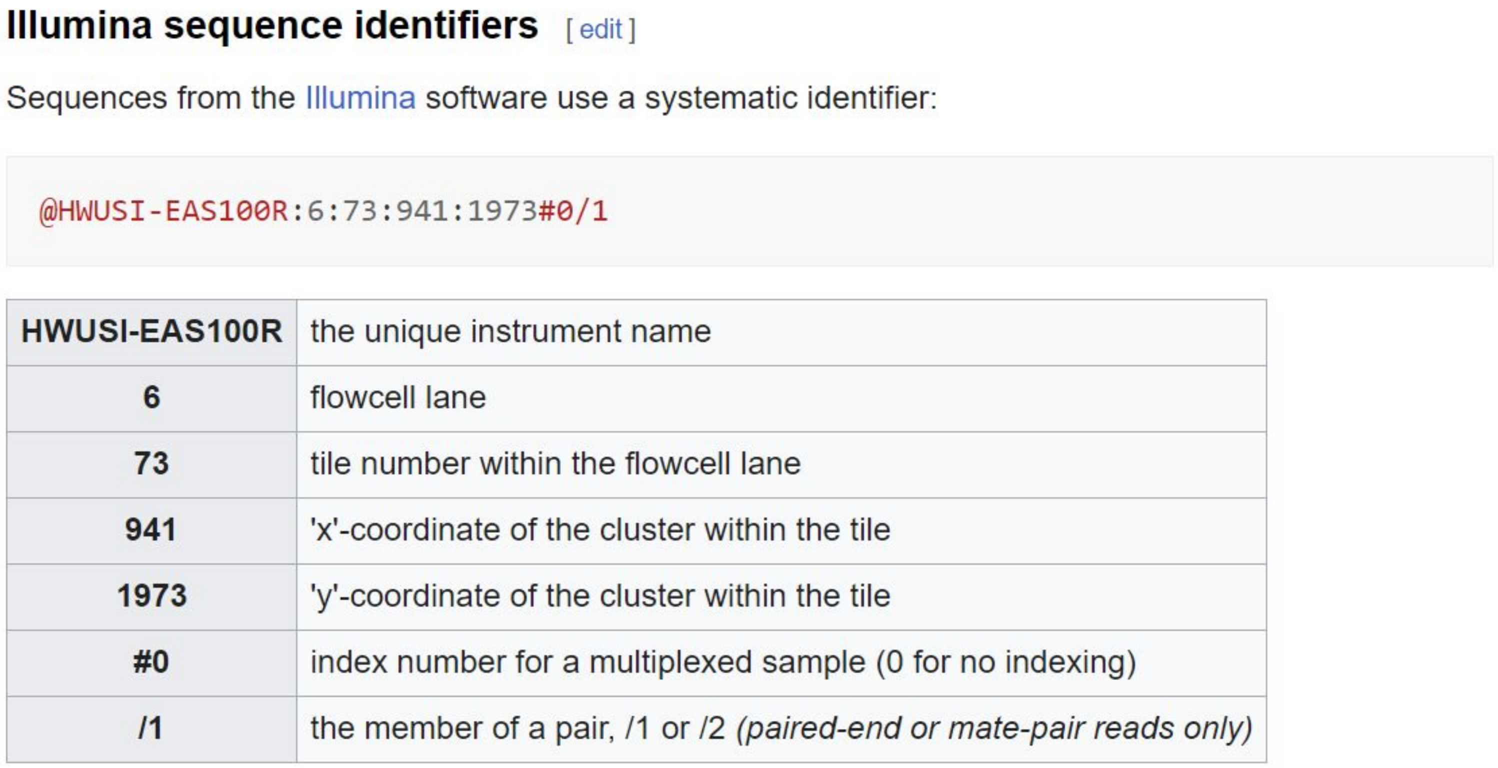
- Simple format for raw unaligned sequencing reads
- Paired-end sequencing: two FASTQ files or one interleaved file

- Quality encoded in ASCII characters with decimal codes 33-126
  - ASCII code of “A” is 65, the corresponding quality is Q= $65 - 33 = 32$

```
Base quality encoded as character
     ! " # $ % & ' ( ) * + , - . / 0 1 2 3 4 5 6 7 8 9 : ; < = > ? @ A B C D E F G H I J
                                 |                                     |
Numeric ASCII value
    33 . . . . . . . . . . . . . 47 . . . . . . . . . . . . . . . . . 65 . . . . . . . .
                                 |                                     |  (65-33 = 32)
Base quality value
     0 . . . . . . . . . . . . . 14 . . . . . . . . . . . . . . . . . 32 . . . . . . . .
```

  - Beware: multiple quality scores were in use!
    - Sanger, Solexa, Illumina 1.3+
    - See https://en.wikipedia.org/wiki/FASTQ_format for details

FastQ - Read name info



Illumina sequence identifiers [edit]

Sequences from the Illumina software use a systematic identifier:

```
@HWUSI-EAS100R:6:73:941:1973#0/1
```

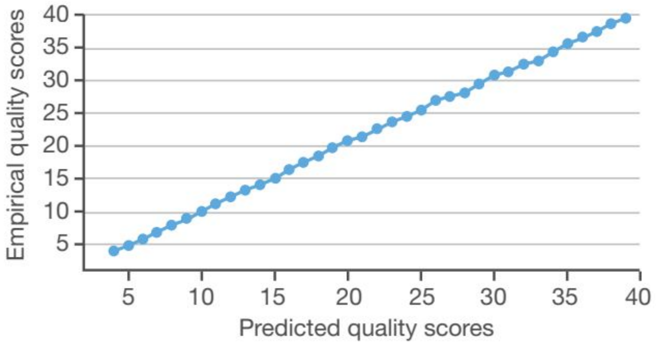
HWUSI-EAS100R	the unique instrument name
6	flowcell lane
73	tile number within the flowcell lane
941	'x'-coordinate of the cluster within the tile
1973	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)*

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

FastQ - Quality Info - Phred scores

- Metrics produced by assessing the signal from the sequencing instrument

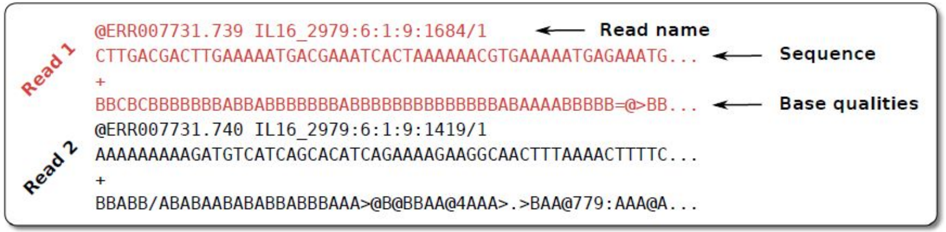
Figure 1: High Correlation of Empirical and Predicted Q Scores



Illumina sequencing Q scores are highly accurate. This example shows that predicted Q scores for a HiSeq 2000 run correlate well to empirically derived Q scores.

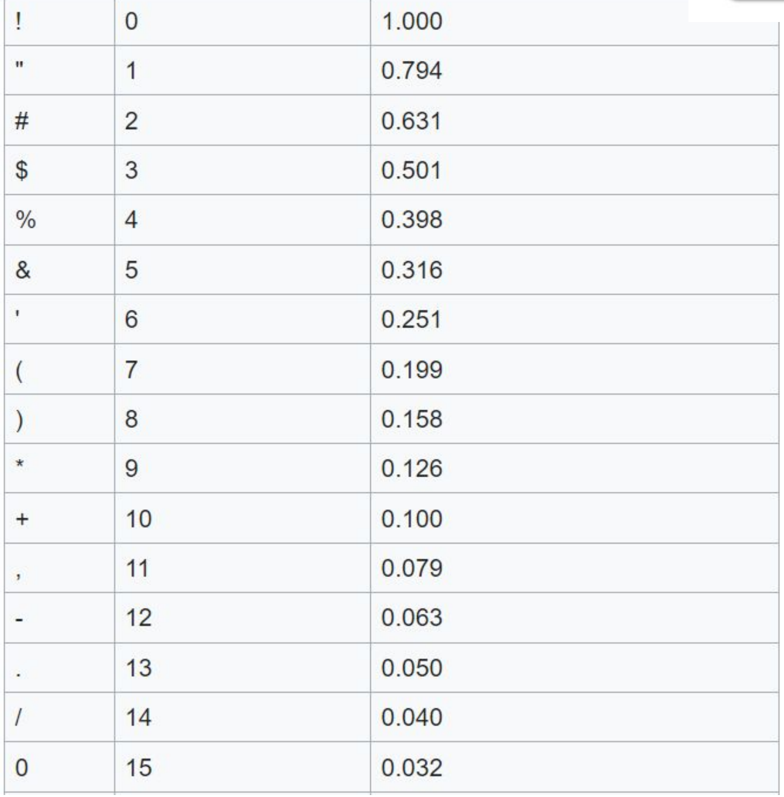
https://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf

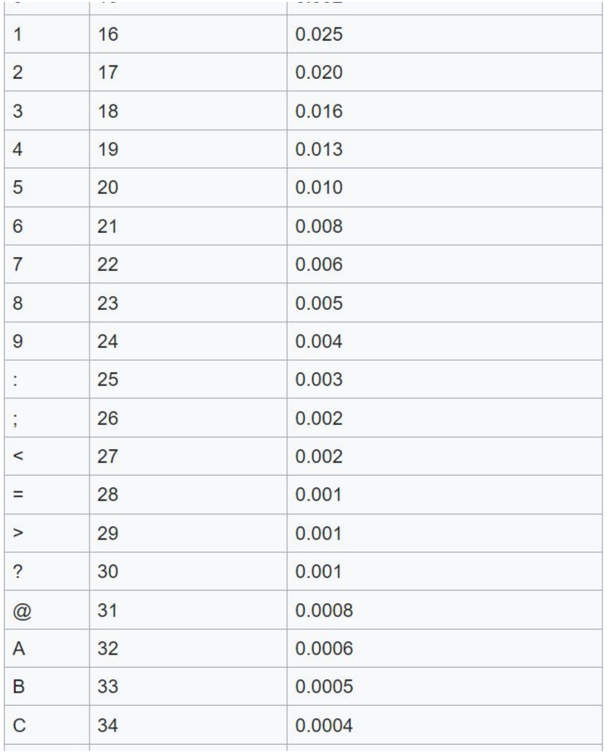
FastQ - Quality Info



Read 1
Read name
Sequence
Base qualities
Read 2

```
@ERR007731.739 IL16_2979:6:1:9:1684/1
CTTGACGACTTGAAAAATGACGAAATCACTAAAAAACGTGAAAAATGAGAAATG...
+
BBCBCBBBBBBBABBABBBBBBBABBBBBBBBBBBBBBABAAAABBBBB=@>BB...
@ERR007731.740 IL16_2979:6:1:9:1419/1
AAAAAAAAAGATGTCATCAGCACATCAGAAAAGAAGGCAACTTTAAAACTTTTC...
+
BBABB/ABABAABABABBABBBAAA>@B@BBAA@4AAA>.>BAA@779:AAA@A...
```





Symbol	Phred Quality Score	Probability of Incorrect Ba
!	0	1.000
"	1	0.794
#	2	0.631
$	3	0.501
%	4	0.398
&	5	0.316
'	6	0.251
(	7	0.199
)	8	0.158
*	9	0.126
+	10	0.100
,	11	0.079
-	12	0.063
.	13	0.050
/	14	0.040
0	15	0.032
1	16	0.025
2	17	0.020
3	18	0.016
4	19	0.013
5	20	0.010
6	21	0.008
7	22	0.006
8	23	0.005
9	24	0.004
:	25	0.003
;	26	0.002
<	27	0.002
=	28	0.001
>	29	0.001
?	30	0.001
@	31	0.0008
A	32	0.0006
B	33	0.0005
C	34	0.0004

https://en.wikipedia.org/wiki/Phred_quality_score

Let’s start the genomics adventure and take a look a some typical short reads!