

# Fragment assembly of DNA

---

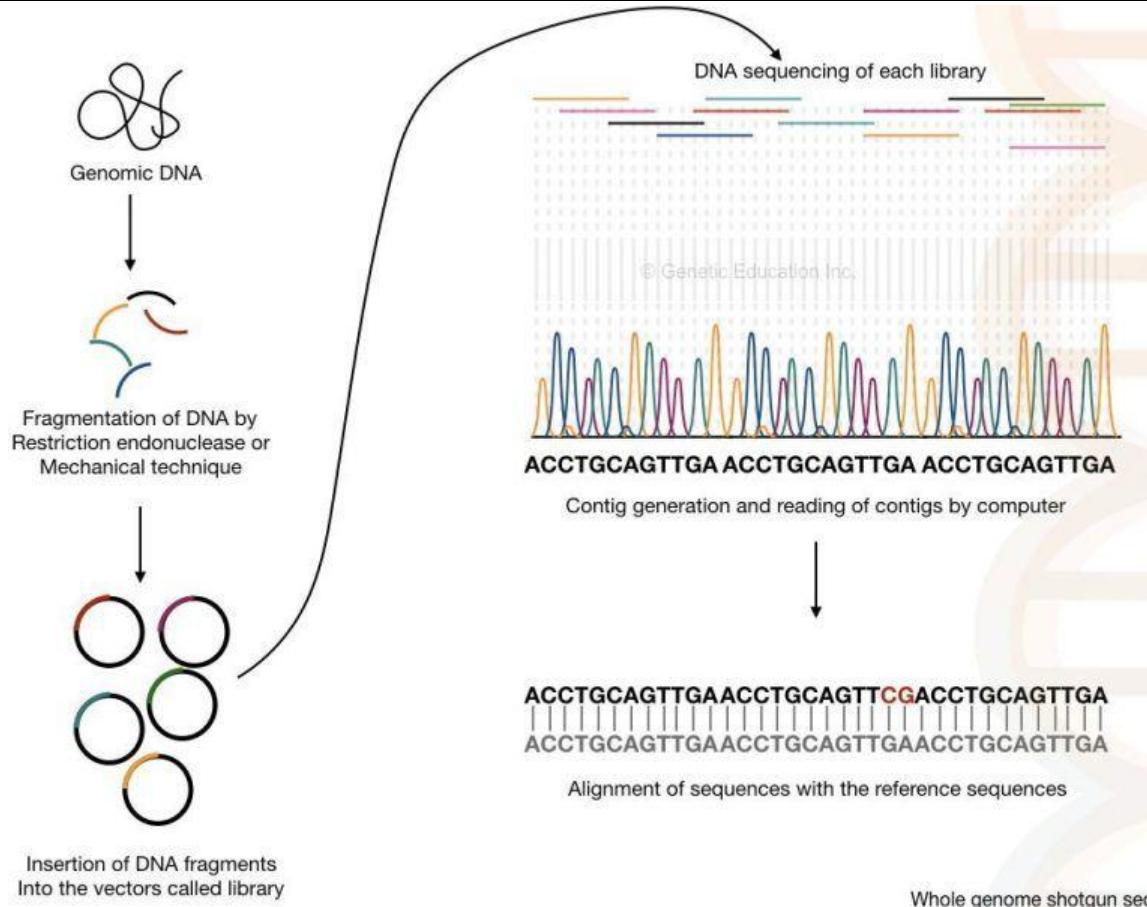


INDRAPRASTHA INSTITUTE *of*  
INFORMATION TECHNOLOGY **DELHI**

Dr. Jaspreet Kaur Dhanjal  
Assistant Professor, Center for Computational Biology  
Email ID: [jaspreet@iiitd.ac.in](mailto:jaspreet@iiitd.ac.in)

*September 08, 2025*

# Whole genome shotgun sequencing



# Whole genome shotgun sequencing

Input: GGC GTCT ATAT CTC GGCT TAGG CCTC ATTTTTT

Copy: GGC GTCT ATAT CTC GGCT TAGG CCTC ATTTTTT  
GGC GTCT ATAT CTC GGCT TAGG CCTC ATTTTTT  
GGC GTCT ATAT CTC GGCT TAGG CCTC ATTTTTT  
GGC GTCT ATAT CTC GGCT TAGG CCTC ATTTTTT

Fragment: GGC GTCTA TATCTCGG CTCTAGGCCCTC ATTTTTT  
GGC GTCT ATAT CTC GGCT CTAGGCCCTCA TTTTTT  
GGC GTCT ATATCT CGGCTCTAGGCCCT CATTTTTT  
GGC GTCTAT ATCTCGGCTCTAG GCCCTCA TTTTTT

Reconstruct  
this

From these

CTAGGCCCTCAATTTTT  
CTCTAGGCCCTCAATTTTT  
GGCTCTAGGCCCTCATTTTTT  
CTCGGCTCTAGGCCCTCATTTT  
TATCTGACTCTAGGCCCTCA  
TATCTGACTCTAGGCC  
TCTATATCTCGGCTCTAGG  
GGCGTCTATATCTCG  
GGCGTCTGATATCT  
GGCGTCTATATCT  
→ GGC GTCT ATAT CTC GGCT TAGG CCTC ATTTTTT

# Important terms

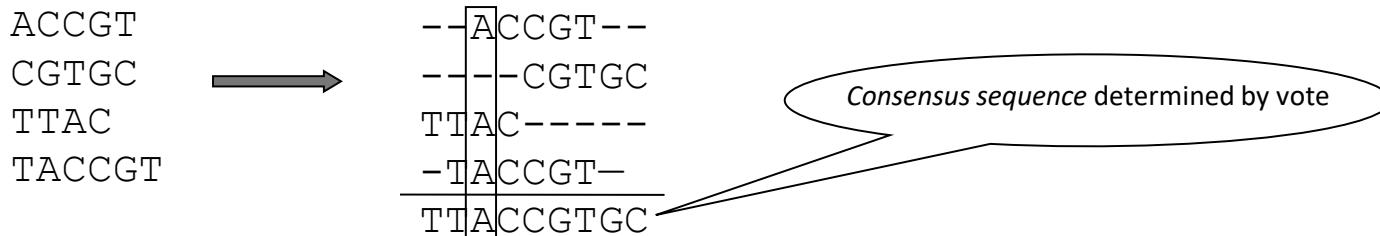
---

- Target: The long sequence to reconstruct.
- Fragment: A small stretch or substring/subsequence of the target.
- Fragment assembly: A collection of fragments to put together.
- Overlaps: The end part of one fragment similar to the beginning of another fragment.
- In ideal case:

The input set is aligned.

The spaces at the extremities are ignored.

Consensus sequence is built based on majority vote.



# Coverage

Coverage/ average coverage: Number of reads covering a position in the genome

Coverage at this position = 6

The diagram shows a reference sequence in blue: C T A G G C C T C A A T T T T T. Below it, several reads are shown in blue, some overlapping the reference. A vertical bar highlights a segment where multiple reads overlap, indicating high coverage at that position.

CTAGGCCCTCAATT  
CTCTAGGCCCTCAATT  
GGCTCTAGGCCCTCATTTT  
CTCGGCTCTAGCCCTCATTT  
TATCTCGACTCTAGGCCCTCA  
TATCTCGACTCTAGGCC  
TCTATATCTCGGCTCTAGG  
GGCGTCTATATCTCG  
GGCGTCGATATCT  
GGCGTCTATATCT  
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTT

177 nucleotides

Average coverage = 177 / 35 ≈ 7x

35 nucleotides

# Linkage

Linkage: The degree of overlaps between the fragments

CTAGGCCCTCAATTTT  
CTCTAGGCCCTCAATTTT  
GGCTCTAGGCCCTCATTTTT  
CTCGGCTCTAGCCCCTCATTTT  
TATCTCGACTCTAGGCCCTCA  
TATCTCGACTCTAGGCC  
TCTATATCTGGCTCTAGG  
GGCGTCTATATCTCG  
GGCGTCGATATCT  
GGCGTCTATATCT  
  
GGCGTCTATATCTGGCTCTAGGCCCTCATTTTT

Target:

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

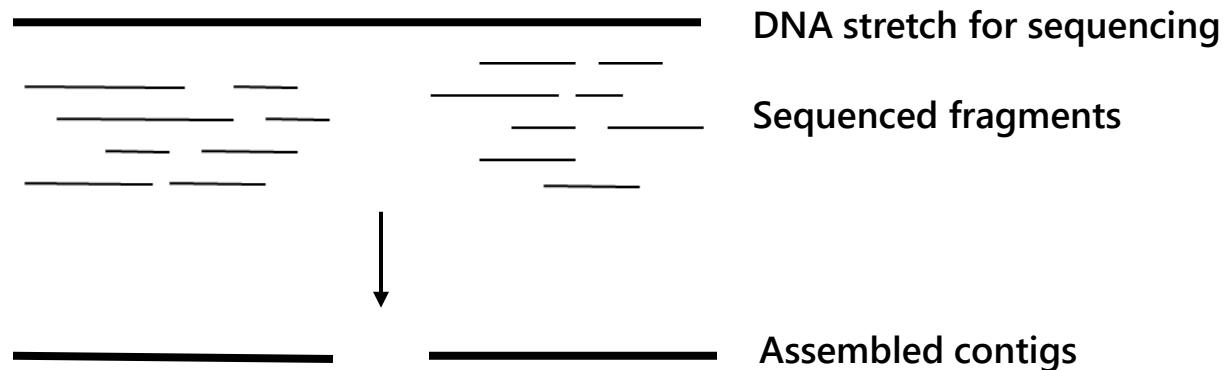
\_\_\_\_\_

\_\_\_\_\_

*Perfect coverage, poor average linkage  
poor minimum linkage*

# Contigs

Sometimes you just can't put all of the fragments together into one contiguous sequence.



No way to tell how much sequence is missing between them.

No way to tell the order of these two *contigs*.

# Complications in the assembly of fragments

---

The main factors that add to the complexity of the problem of fragment assembly are:

- Error
- Unknown orientation
- Repeated regions
- Lack of coverage

# Errors

- The simplest errors are called *base call errors* and comprise base substitutions, insertions and deletions in the fragments.
- Base call errors occurs in practice at rates varying from 1 to 5 errors every 100 characters.

--**A**CCGT--  
-----CGTGC  
TT**A**C-----  
-T**G**CCGT-  
-----  
TTAC**CC**GTGC

Base Call Error

--ACC**G**T--  
-----**A**GTGC  
TTAC-----  
-TACC**G**T-  
-----  
TTACC**G**TGC

Insertion Error

--AC**C**GT--  
-----CGTGC  
TTAC-----  
-TAC**G**T-  
-----  
TTAC**C**GTGC

Deletion Error

# Errors

- Two other types of errors: *Chimera* and *Contamination*
- Chimeras, arise when two regular fragments from distinct parts of the target molecule join end-to-end to form a fragment that is not a contiguous part of the target

*Solution:* Must be recognized as such and removed from the fragment set in a preprocessing stage.

|        |                 |
|--------|-----------------|
| ACCGT  | --ACCGT--       |
| CGTGC  | ----CGTGC       |
| TTAC   | TTAC-----       |
| TACCGT | -TACCGT--       |
| TTATGC | <hr/> TTACCGTGC |
|        | TTA---TGC       |

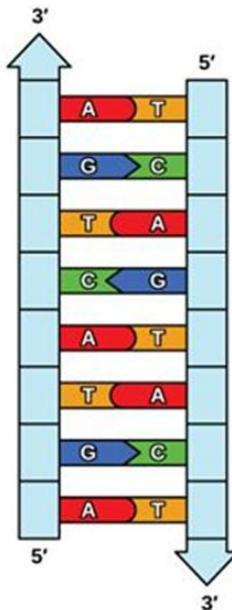


- Contamination is from host or vector DNA

*Solution:* Most vectors are well known, so we can screen the data before starting assembly.

# Unknown Orientation

- We generally do not know to which strand a particular fragment belongs to.
- The input fragments as being all approximate substrings of the consensus sought either as given or in reverse complement.



|        |   |            |
|--------|---|------------|
| CACGT  | → | CACGT      |
| ACGT   | → | -ACGT      |
| ACTACG | ← | --CGTAGT   |
| GTACT  | ← | -----AGTAC |
| ACTGA  | → | -----ACTGA |
| CTGA   | → | -----CTGA  |