

Fragment assembly of DNA

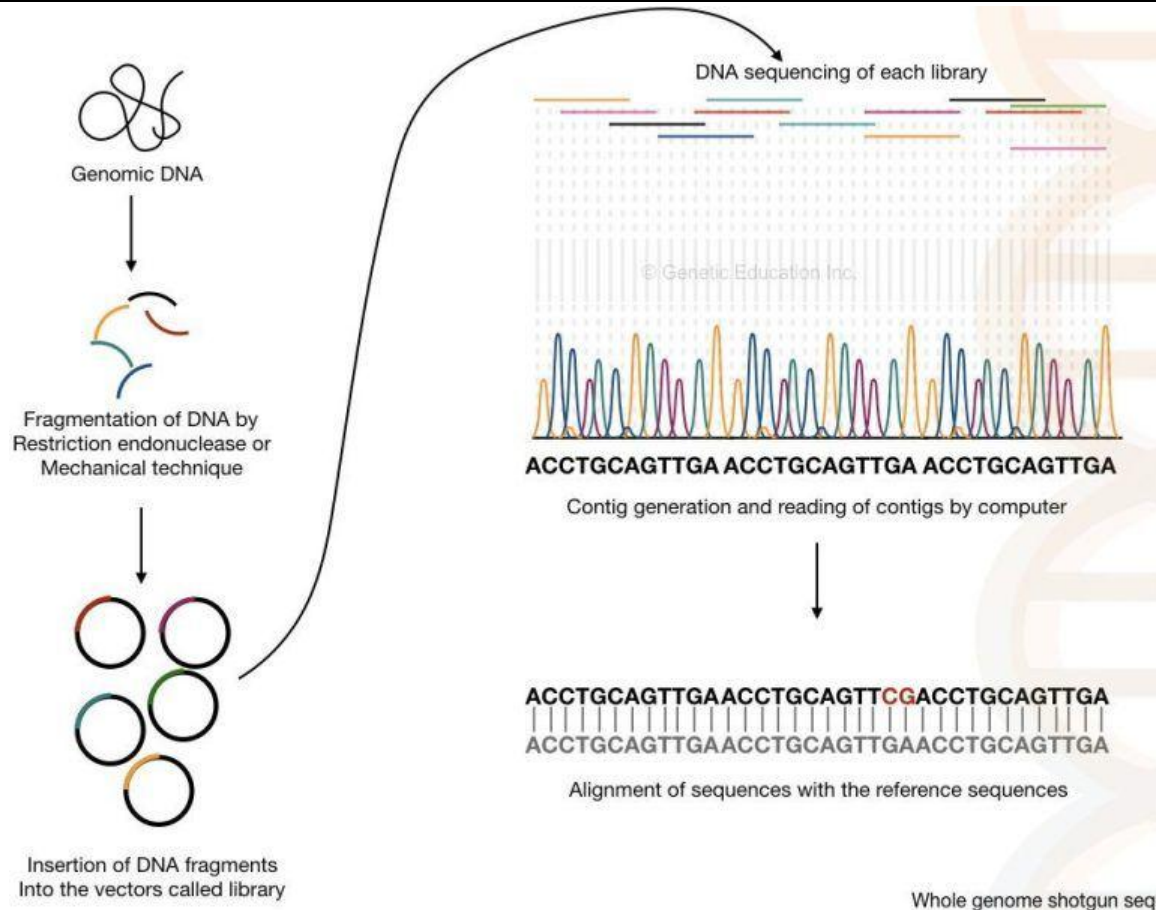


INDRAPRASTHA INSTITUTE of
INFORMATION TECHNOLOGY **DELHI**

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

September 08, 2025

Whole genome shotgun sequencing



Whole genome shotgun sequencing

Input: GCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT

Copy: GCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT

Fragment: GCGTCTA TATCTCGG CTCTAGGCCCTC ATTTTTT
GGC GTCTATAT CTCGGCTCTAGGCCCTCA TTTTTT
GGCGTC TATATCT CGGCTCTAGGCCCT CATTTTTT
GGCGTCTAT ATCTCGGCTCTAG GCCCTCA TTTTTT

Reconstruct
this

CTAGGCCCTCAATTTTT
CTCTAGGCCCTCAATTTTT
GGCTCTAGGCCCTCATTTTT
CTCGGCTCTAGGCCCTCATTTT
TATCTCGACTCTAGGCCCTCA
TATCTCGACTCTAGGCC
TCTATATCTCGGCTCTAGG
GGCGTCTATATCTCG
GGCGTCGATATCT
GGCGTCTATATCT

From these

→ GCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT

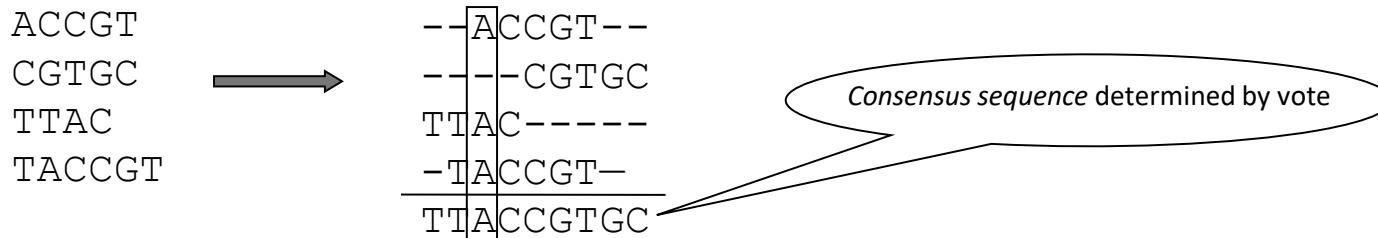
Important terms

- 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.

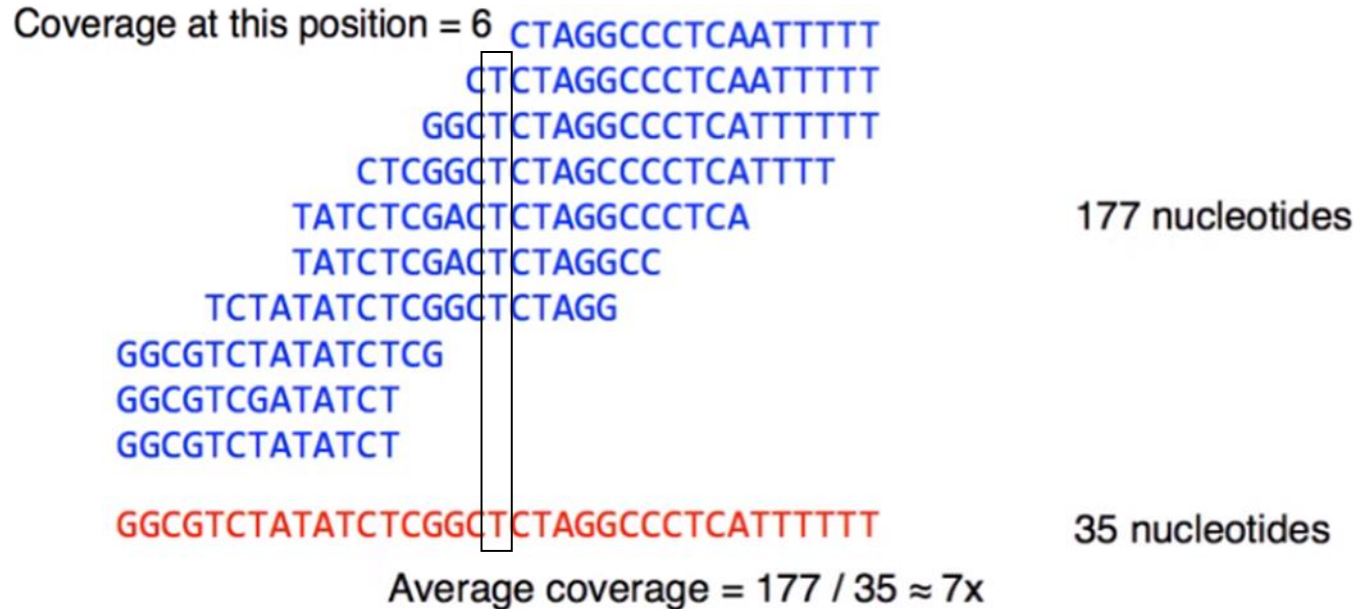
The spaces at the extremities are ignored.

Consensus sequence is built based on majority vote.



Coverage

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



Linkage

Linkage: The degree of overlaps between the fragments

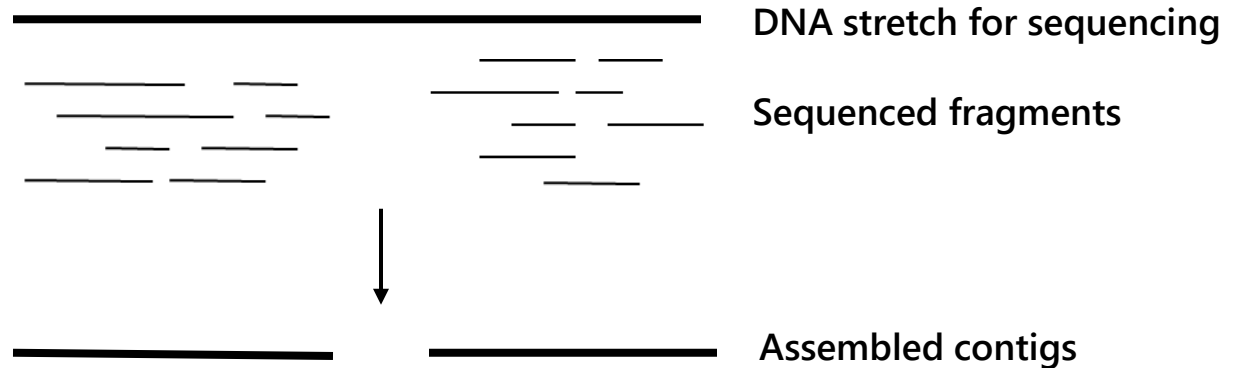
CTAGGCCCTCAATTTT
CTCTAGGCCCTCAATTTT
GGCTCTAGGCCCTCATTTTT
CTCGGCTCTAGCCCCTCATTTT
TATCTCGACTCTAGGCCCTCA
TATCTCGACTCTAGGCC
TCTATATCTCGGCTCTAGG
GGCGTCTATATCTCG
GGCGTCGATATCT
GGCGTCTATATCT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTT

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.

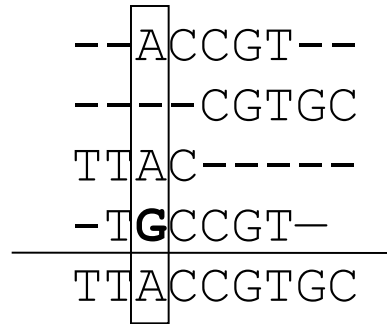


Diagram illustrating a Base Call Error. A vertical box highlights the second column of bases in two DNA fragments. The top fragment is --ACCGT-- and the bottom fragment is TTACCGTGC. In the bottom fragment, the 'G' at the second position is highlighted in bold, indicating a substitution error from 'A' to 'G'.

Base Call Error

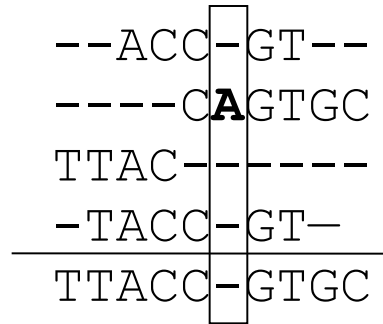


Diagram illustrating an Insertion Error. A vertical box highlights the second column of bases in two DNA fragments. The top fragment is --ACC--GT-- and the bottom fragment is TTACC--GTGC. In the bottom fragment, an 'A' is inserted at the second position, highlighted in bold, which is not present in the top fragment.

Insertion Error

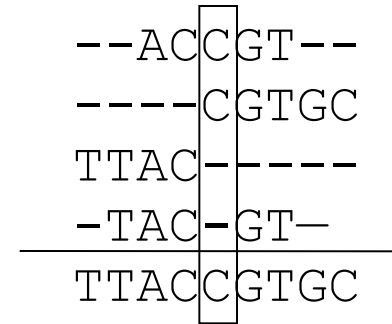


Diagram illustrating a Deletion Error. A vertical box highlights the second column of bases in two DNA fragments. The top fragment is --ACCGT-- and the bottom fragment is TTACC--GTGC. In the bottom fragment, the 'G' at the second position is missing, highlighted in bold, indicating a deletion error.

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

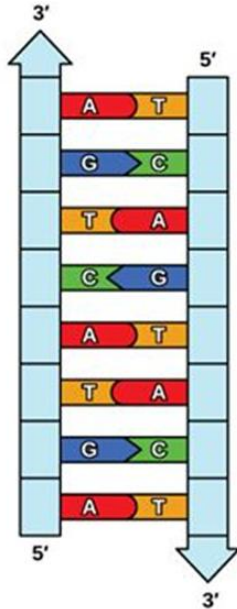


- Contamination is from host or vector DNA

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