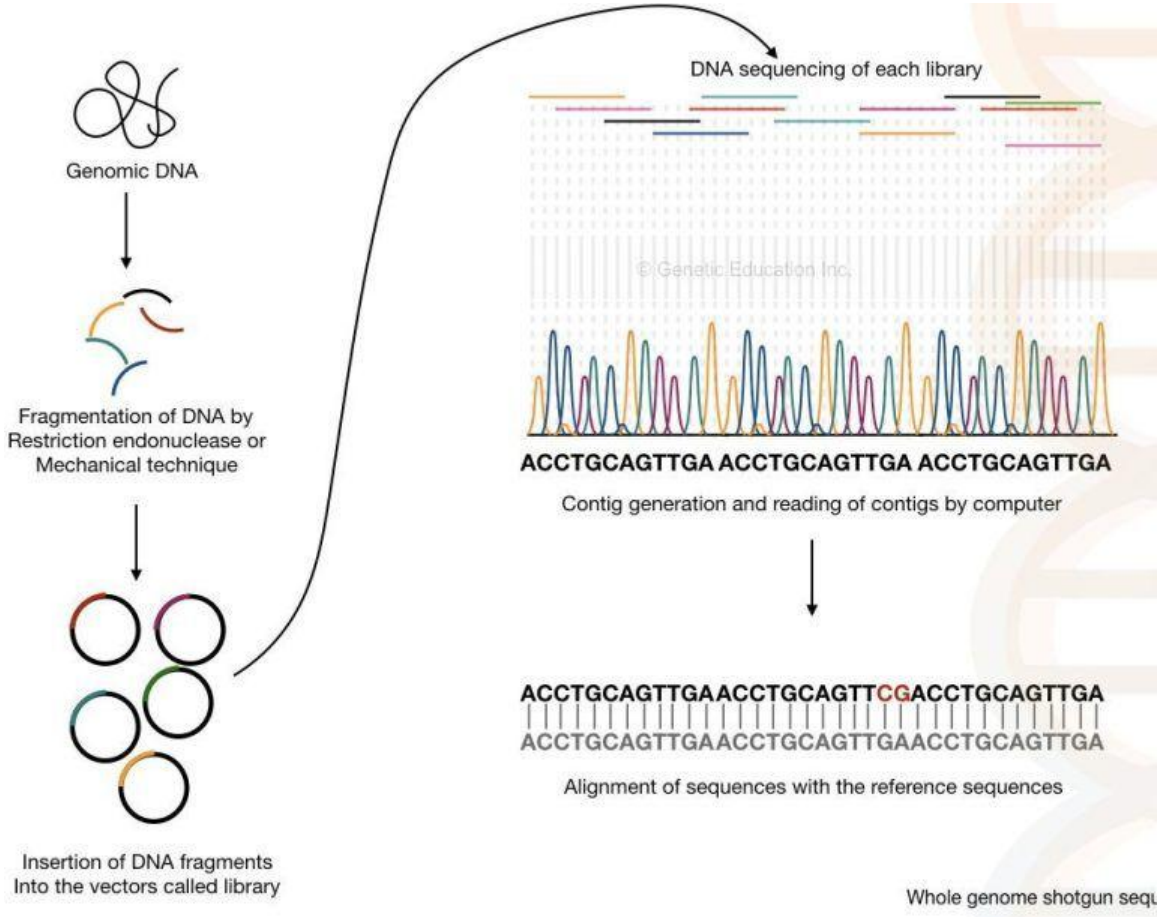
Fragment assembly of DNA

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

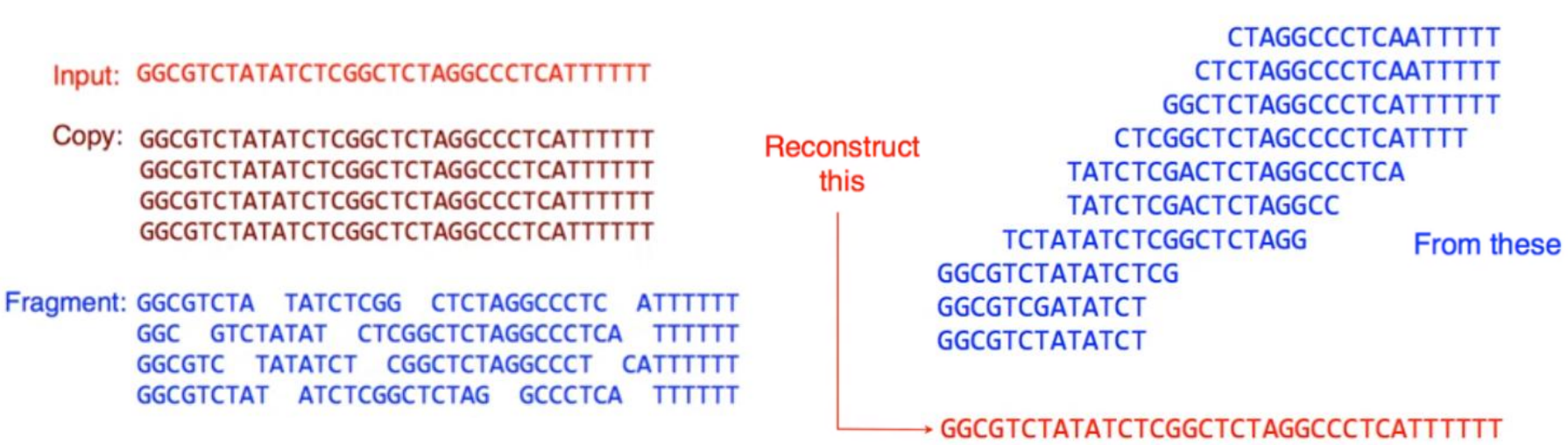
Dr. Jaspreet Kaur Dhanjal
Assistant Professor, Center for Computational Biology
Email ID: jaspreet@iiitd.ac.in

September 08, 2025

Whole genome shortgun sequencing



Whole genome shortgun sequencing



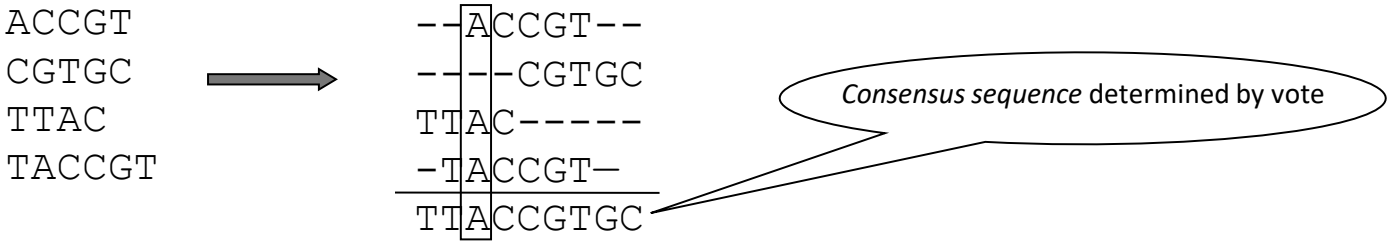
Important terms

- 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.

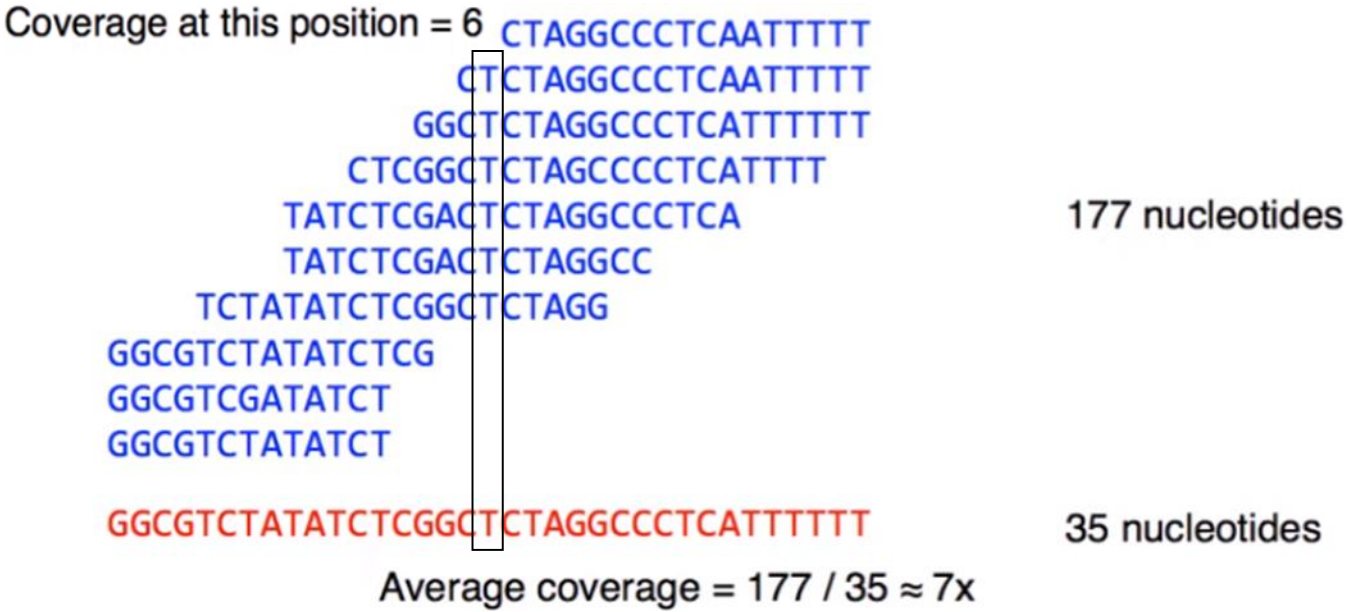
  The spaces at the extremities are ignored.

  Consensus sequence is built based on majority vote.



Coverage

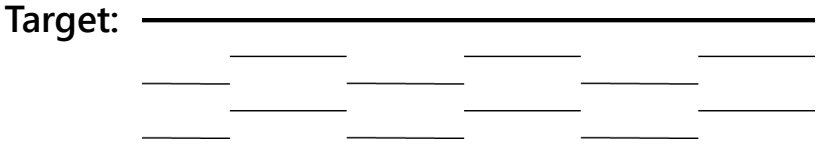
Coverage/ average coverage: Number of reads covering a position in the genome



Linkage

Linkage: The degree of overlaps between the fragments

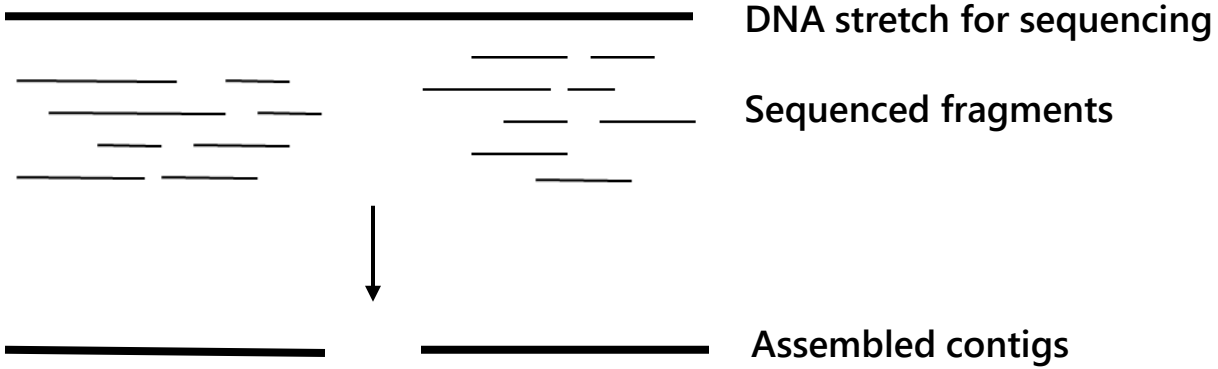
```
                    CTAGGCCCTCAATTTTT
                  CTCTAGGCCCTCAATTTTT
                GGCTCTAGGCCCTCATTTTTT
             CTCGGCTCTAGCCCCTCATTTT
          TATCTCGACTCTAGGCCCTCA
          TATCTCGACTCTAGGCC
      TCTATATCTCGGCTCTAGG
GGCGTCTATATCTCG
GGCGTCGATATCT
GGCGTCTATATCT

GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
```



Perfect coverage, poor average linkage
poor minimum linkage

Contigs

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



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

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

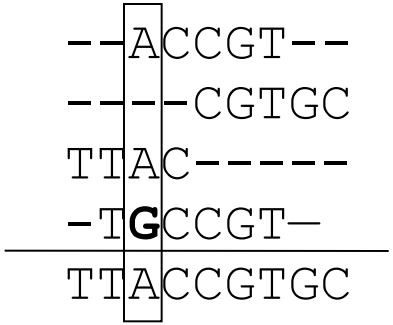
Complications in the assembly of fragments

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

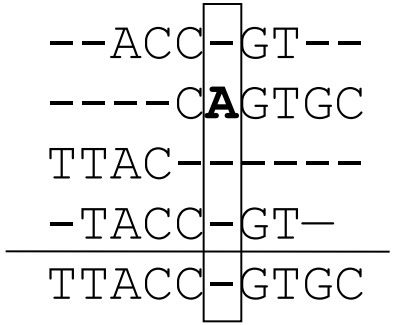
- Error
- Unknown orientation
- Repeated regions
- Lack of coverage

Errors

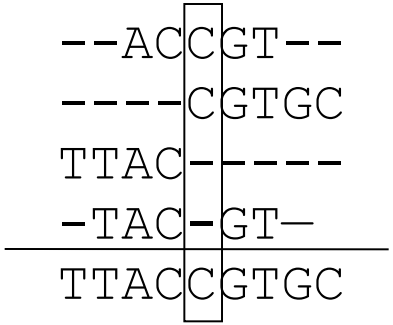
- The simplest errors are called *base call errors* and comprise base substitutions, insertions and deletions in the fragments.
- Base call errors occurs in practice at rates varying from 1 to 5 errors every 100 characters.



Base Call Error



Insertion Error

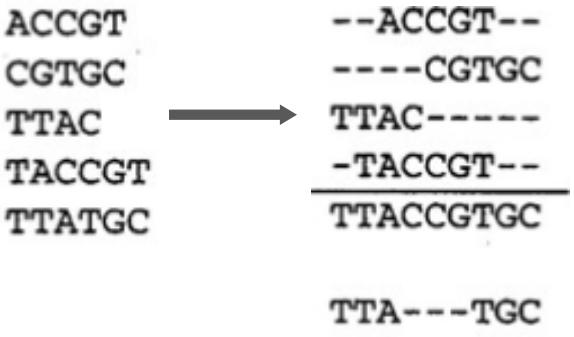


Deletion Error

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.

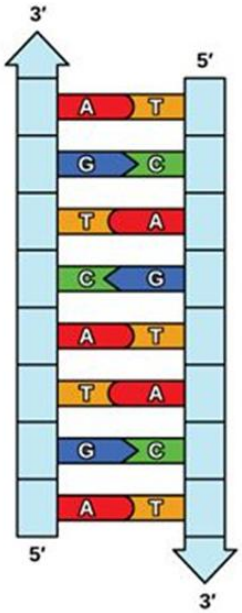


- Contamination is from host or vector DNA

  Solution: Most vectors are well know, 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
```