Applied Genomics

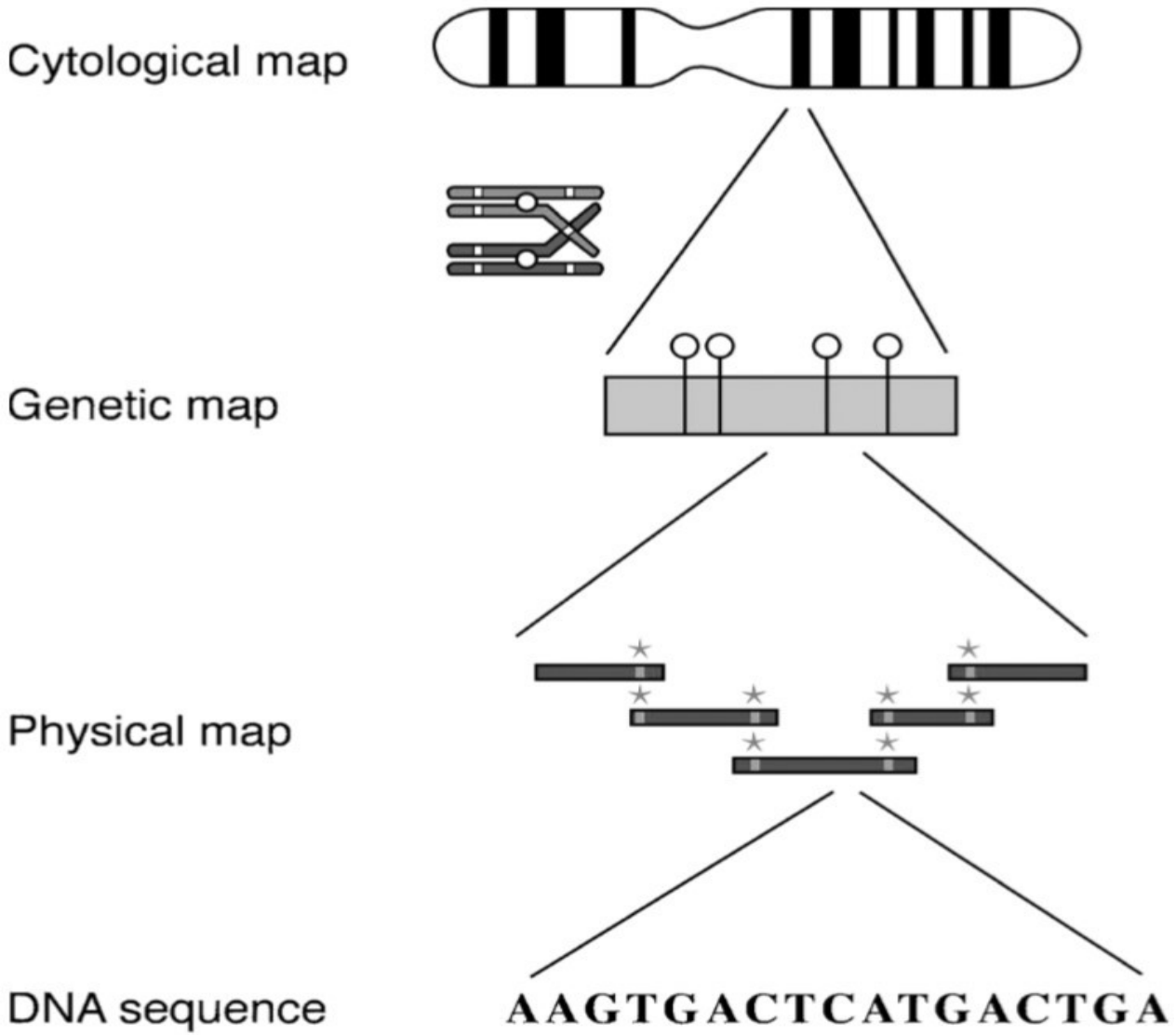
By

Aparna Patil Kose

Dept of Bioinformatics

Genome Mapping

- Genome mapping is a process of identifying relative locations of genes, mutations or traits on a chromosome.
- A low-resolution approach is to describe the order and relative distances of genetic markers on a chromosome.
  - **Genetic markers** are identifiable portions of a chromosome whose inheritance patterns can be followed
- **Genetic linkage** maps, **physical maps** and **cytologic maps** describe genomes at different levels of resolution.


Cytological map
Genetic map
Physical map
DNA sequence
AAGTGACTCATGACTGA

Genetic linkage maps

- Genetic linkage maps, also called genetic maps, identify the **relative positions of genetic markers** on a chromosome and are based on how frequent the markers are inherited together.
- The rationale behind genetic mapping is that the closer the two genetic markers are, the more likely it is that they are inherited together and are not separated in a genetic crossing event
  - When genes are found on different chromosomes or far apart on the same chromosome, they assort independently and are said to be **unlinked**.
  - When genes are close together on the same chromosome, they are said to be **linked**.
- Genetic crossing experiment can be performed to calculate the **recombination frequency**.
- The distance between the two genetic markers is measured in centiMorgans (cM), which is the frequency of recombination of genetic markers.

Physical and Cytological maps

- Physical maps are **maps of locations of identifiable landmarks** on a genomic DNA regardless of inheritance patterns.
- The distance between genetic markers is measured directly as kilobases (Kb) or megabases (Mb).
- Physical maps are constructed by using a chromosome walking technique, which uses a number of radiolabeled probes to hybridize to a library of DNA clone fragments.
- By identifying overlapping clones probed by common probes, a relative order of the cloned fragments can be established.

- **Cytologic maps** refer to banding patterns seen on stained chromosomes, which can be directly observed under a microscope. The observable light and dark bands are the visually distinct markers on a chromosome. A genetic marker can be associated with a specific chromosomal band or region

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 from both a physical and a functional standpoint.
- The project formally launched in 1990, and was declared complete on April 14, 2003.
- The HGP has revealed that there are probably about 20,500 human genes.
- This ultimate product of the HGP has given the world a resource of detailed information about the structure, organization and function of the complete set of human genes.

Whole genome sequencing

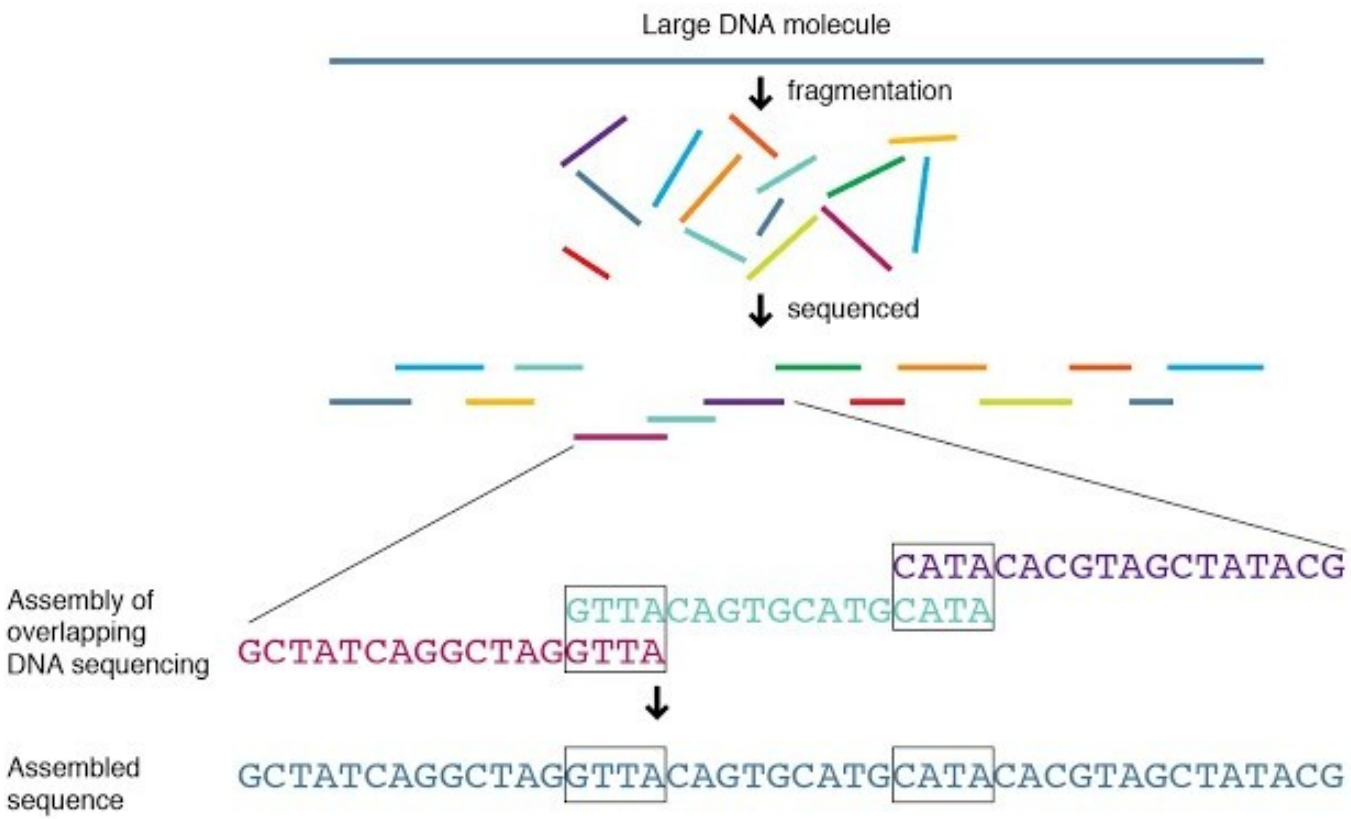
- There are two major strategies for whole genome sequencing:

1. The shotgun approach: Bottom-up approach

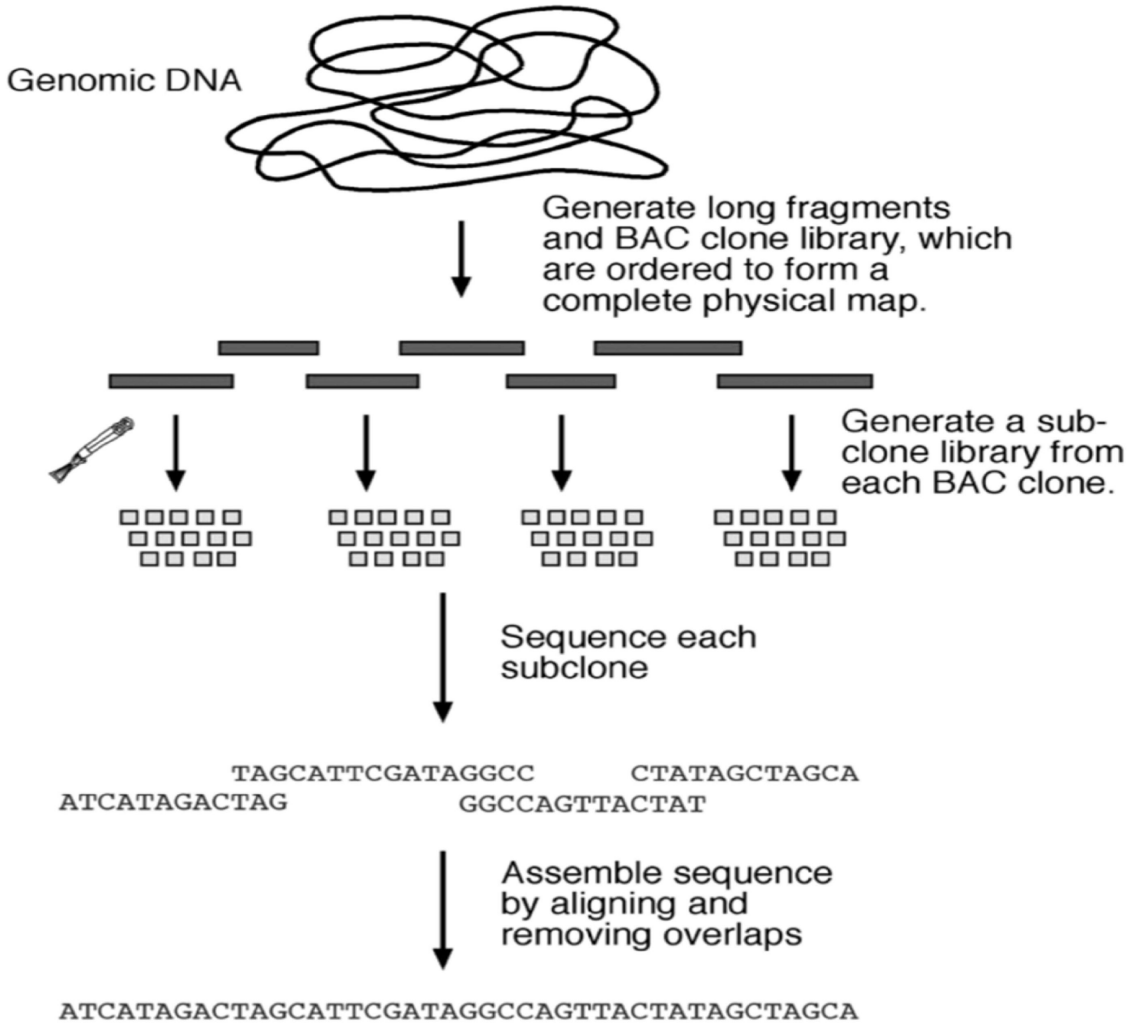
   Generates a large number of sequenced DNA fragments. The number of random fragments has to be very large, so large that the DNA fragments overlap sufficiently to cover the entire genome.

2. The hierarchical approach: Top-down approach

   The chromosomes are initially mapped using the physical mapping strategy. Longer fragments of genomic DNA (100 to 300 kB) are obtained and cloned into a high-capacity bacterial vector called bacterial artificial chromosome(BAC). Physical mapping determines the locations and orders of the BAC clones on a chromosome. By successively sequencing adjacent BAC clone fragments, the entire genome can be covered

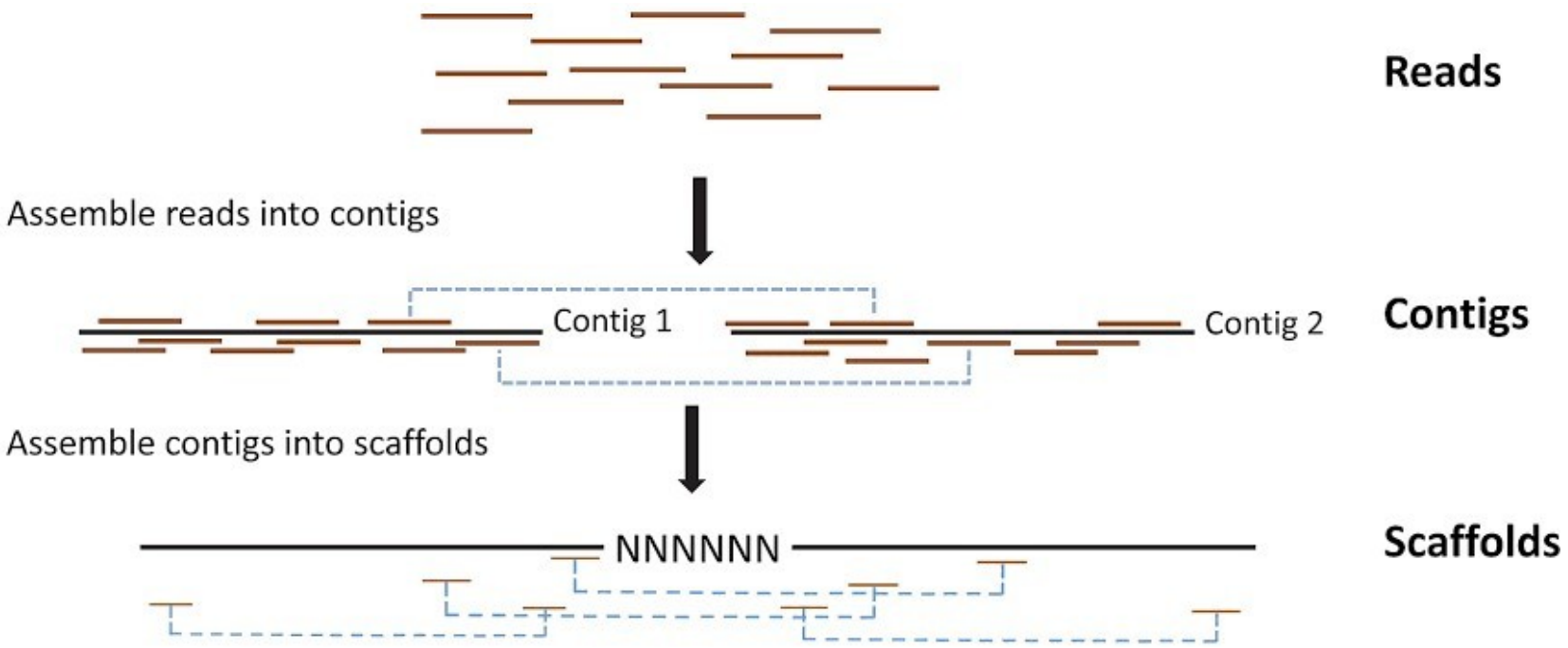

Large DNA molecule
fragmentation
sequenced
Assembly of overlapping DNA sequencing
GCTATCAGGCTAGGTTA
GTTACAGTGCATGCATA
CATACACGTAGCTATACG
Assembled sequence
GCTATCAGGCTAGGTTACAGTGCATGCATACACGTAGCTATACG

Hierarchical Sequencing Approach



Whole Genome sequencing

- To assemble a whole genome sequence, short fragments are joined to form larger fragments after removing overlaps.
- These longer, merged sequences are termed contigs, which are usually 5,000 to 10,000 bases long.
- A **contig** (from contiguous) is a set of overlapping DNA segments that together represent a consensus region of DNA. (top-down sequencing projects, contig refers to the overlapping clones)
- A series of contigs that are in the right order but not necessarily connected in one continuous stretch of sequence is called **scaffold**
- Overlapping scaffolds are then connected to create the final highest resolution map of the genome.


Reads
Assemble reads into contigs
Contig 1
Contig 2
Contigs
Assemble contigs into scaffolds
NNNNNN
Scaffolds

DNA polymorphisms

- DNA polymorphisms are the different DNA sequences among individuals, groups, or populations.

- Polymorphism at the DNA level includes a wide range of variations from **single base pair change, many base pairs, and repeated sequences.**

- Genomic variability can be present in many forms, including single nucleotide polymorphisms (SNPs), variable number of tandem repeats (VNTRs, e.g., mini- and microsatellites), transposable elements (e.g., Alu repeats), structural alterations, and copy number variations.

- Different forms of DNA polymorphisms can be tracked using a variety of techniques; some of these techniques include restriction fragment length polymorphisms (RFLPs) with Southern blots, polymerase chain reactions (PCRs), hybridization techniques using DNA microarray chips, and genome sequencing.

- During the last years, the recent advance of molecular technologies revealed new discoveries of DNA polymorphisms.

- Mapping the human genome requires a set of genetic markers and they serves as a genetic marker for its own location in the chromosome; thus, they are convenient for analysis and are often used as in molecular genetic studies.

DNA sequences and analysis

Nucleotide sequence databases such as :
GenBank, EMBL, DDBJ

Resource for restriction enzyme: REBASE/ Restriction Analyzer

REBASE (http://rebase.neb.com) :The Restriction Enzyme Database is a collection of information about restriction enzymes, methylases, the microorganisms from which they have been isolated, recognition sequences, cleavage sites, methylation specificity, the commercial availability of the enzymes, and references - both published and unpublished observations.

Ref: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2808884/

Restriction Analyzer: (https://molbiotools.com/restrictionanalyzer.php)
Restriction Analyzer is a free software tool for comprehensive restriction analysis of a DNA sequence. It detects all present and absent restriction sites and presents the results both as tabular listings and graphical output (annotated sequence). Furthermore, it provides a DNA digest electropherogram simulation with unlimited number of restriction enzymes.