DATE: 30-03-22

A field guide to whole-genome sequencing, assembly and annotation

Introduction:

Genome sequencing projects were long confined to biomedical model organisms and required the concerted effort of large consortia. Rapid progress in high-throughput sequencing technology and the simultaneous development of bioinformatic tools have democratized the field. It is now within reach for individual research groups in the eco-evolutionary and conservation community to generate de novo draft genome sequences for any organism of choice. Because of the cost and considerable effort involved in such an endeavour, the important first step is to thoroughly consider whether a genome sequence is necessary for addressing the biological question at hand. Once this decision is taken, a genome project requires careful planning with respect to the organism involved and the intended quality of the genome draft.

The Steps involved in this are Genome Sequencing, Genome Assembly and Genome annotation. We will discuss these in detail ahead.

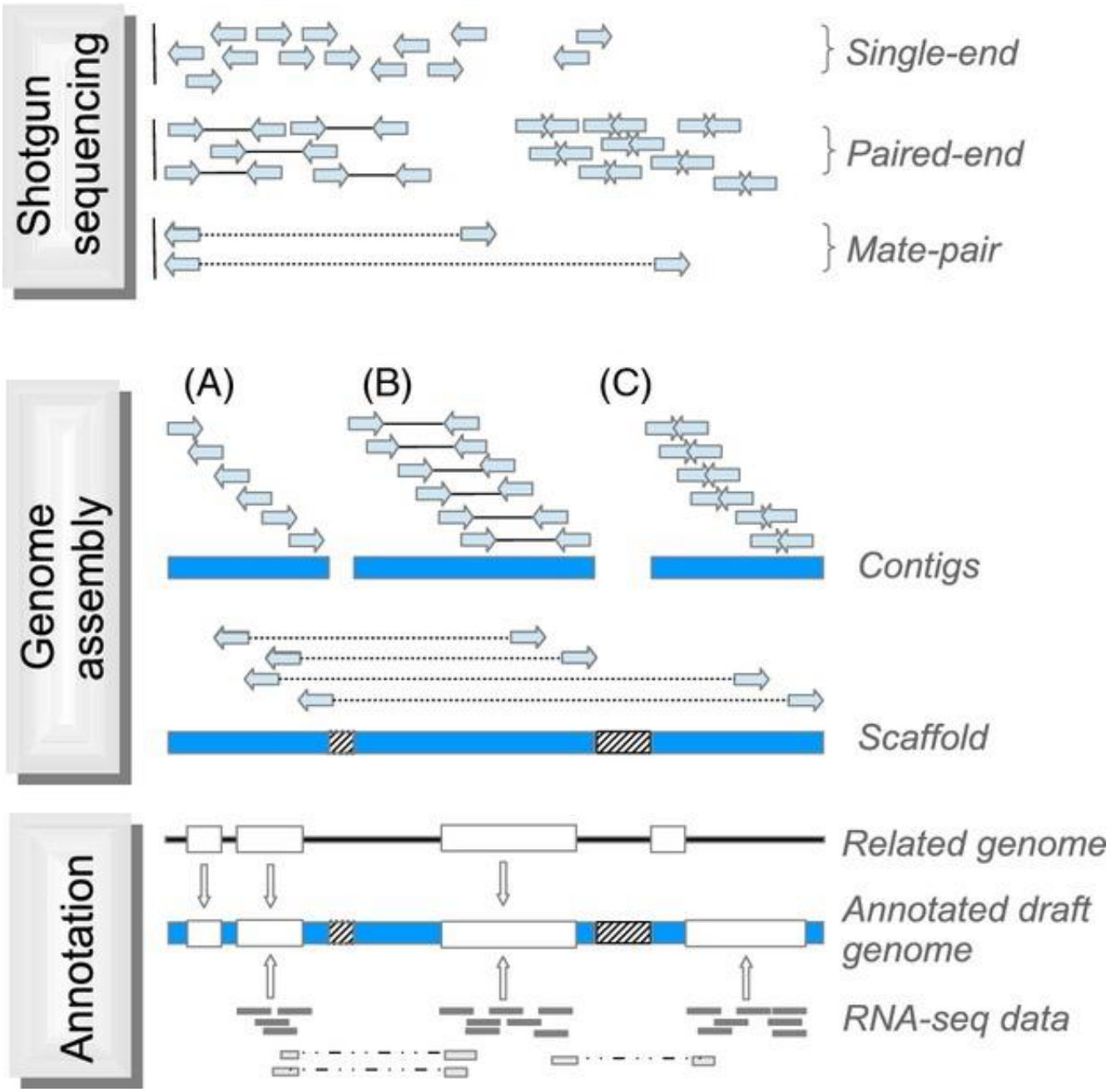

Fig1. Simplified illustration of the assembly process

GENOME SEQUENCING:

- The highest resolution genome map is the genomic DNA sequence that can be considered as a type of physical map describing a genome at the single base-pair level.
- DNA sequencing is now routinely carried out using the Sanger method. This involves the use of DNA polymerases to synthesize DNA chains of varying lengths.
- There are two major strategies for whole genome sequencing:
    - Shotgun approach
        - The shotgun approach randomly sequences clones from both ends of cloned DNA. This 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.
        - This approach does not require knowledge of physical mapping of the clone fragments, but rather a robust computer assembly program to join the pieces of random fragments into a single, whole-genome sequence.
        - Generally, the genome has to be redundantly sequenced in such a way that the overall length of the fragments covers the entire genome multiple times.
        - This is designed to minimize sequencing errors and ensure correct assembly of a contiguous sequence.
        - Despite the multiple coverage, sometimes certain genomic regions remain un-sequenced, mainly owing to cloning difficulties.
        - In such cases, the remainder gap sequences can be obtained through extending sequences from regions of known genomic sequences using a more traditional PCR technique, which requires the use of custom primers and performs genome walking in a stepwise fashion.
    - Hierarchical approach
        - The hierarchical genome sequencing approach is similar to the shotgun approach, but on a smaller scale.
        - 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).
        - Based on the results of physical mapping, the locations and orders of the BAC clones on a chromosome can be determined.
        - By successively sequencing adjacent BAC clone fragments, the entire genome can be covered. The complete sequence of each individual BAC clone can be obtained using the shotgun approach.
        - Overlapping BAC clones are subsequently assembled into an entire genome sequence.

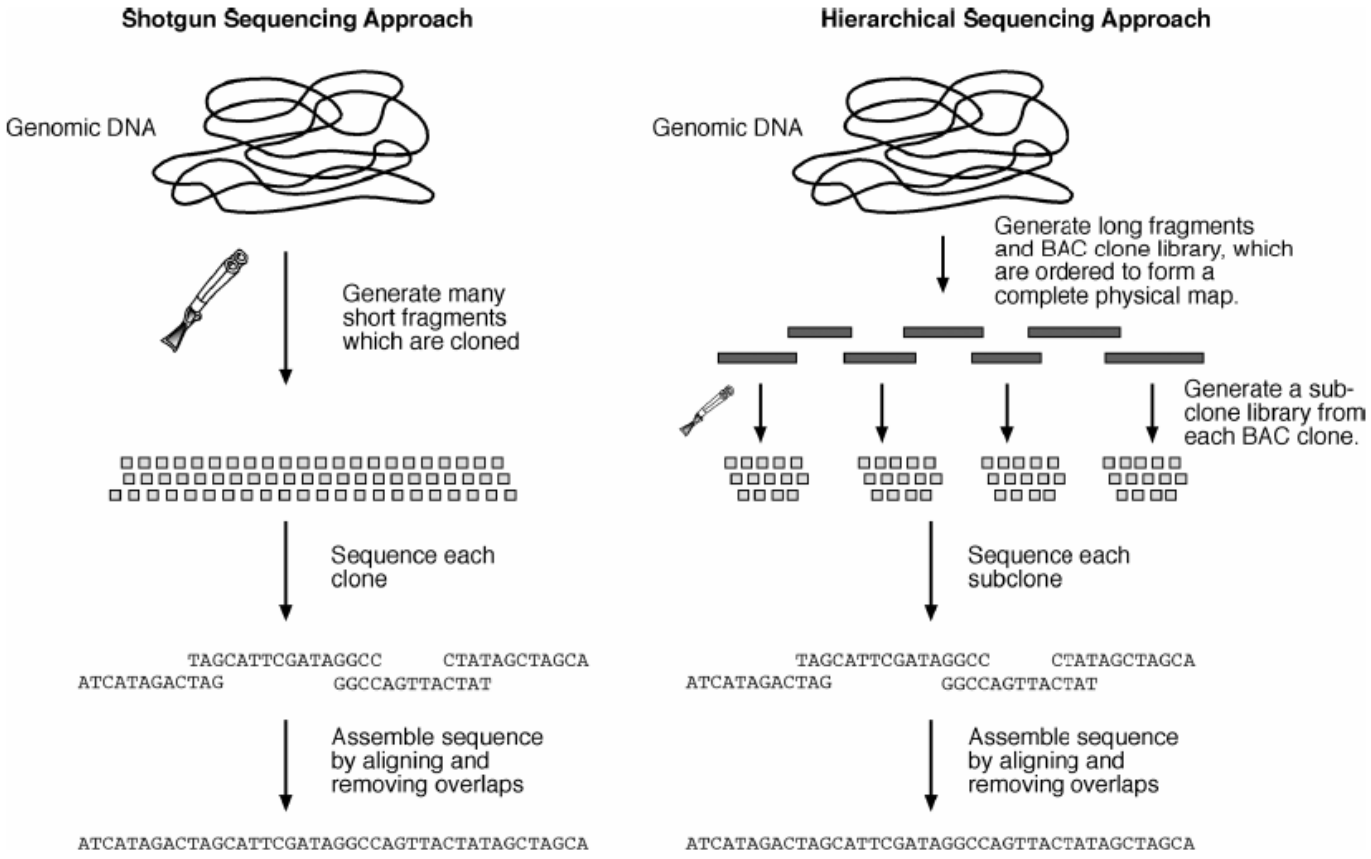

Fig2. Differences between Shotgun sequencing and hierarchical sequencing

GENOME ASSEMBLY:

- As described, initial DNA sequencing reactions generate short sequence reads from DNA clones. The average length of the reads is about 500 bases.
- To assemble a whole genome sequence, these 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 number of overlapping contigs can be further merged to form scaffolds (30,000–50,000 bases, also called super contigs), which are unidirectionally oriented along a physical map of a chromosome.
- Overlapping scaffolds are then connected to create the final highest resolution map of the genome.
- Correct identification of overlaps and assembly of the sequence reads into contigs are like joining jigsaw puzzles, which can be very computationally intensive when dealing with data at the whole-genome level.
- The major challenges in genome assembly are sequence errors, contamination by bacterial vectors, and repetitive sequence regions. Sequence errors can often be corrected by drawing a consensus from an alignment of multiple overlapped sequences.
- Bacterial vector sequences can be removed using filtering programs prior to assembly. To overcome the problem of sequence repeats, programs such as RepeatMasker can be used to detect and mask repeats. Additional constraints on the sequence reads can be applied to avoid misassembly caused by repeat sequences.
- A commonly used constraint to avoid errors caused by sequence repeats is the so- called forward–reverse constraint.
- When a sequence is generated from both ends of a single clone, the distance between the two opposing fragments of a clone is fixed to a certain range, meaning that they are always separated by a distance defined by a clone length (normally 1,000 to 9,000 bases).

- When the constraint is applied, even when one of the fragments has a perfect match with a repetitive element outside the range, it is not able to be moved to that location to cause miss assembly.
- The first step toward genome assembly is to derive base calls and assign associated quality scores. The next step is to assemble the sequence reads into contiguous sequences.
- This step includes identifying overlaps between sequence fragments, assigning the order of the fragments and deriving a consensus of an overall sequence. Assembling all shotgun fragments into a full genome is a computationally very challenging step.
- There are a variety of programs available for processing the raw sequence data.
- The following is a selection of base calling and assembly programs commonly used in genome sequencing projects:
    - **Phred** (https://www.phrap.org/) is a UNIX program for base calling. It uses a Fourier analysis to resolve fluorescence traces and predict actual peak locations of bases
    - **Phrap** (https://www.phrap.org/) is a UNIX program for sequence assembly. It takes Phred base-call files with quality scores as input and aligns individual fragments in a pairwise fashion using the Smith–Waterman algorithm.
    - **VecScreen** (https://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html) is a web-based program that helps detect contaminating bacterial vector sequences.
    - **TIGR Assembler** (https://www.tigr.org/) is a UNIX program from TIGR for assembly of large shotgun sequence fragments. It treats the sequence input as clean reads without consideration of the sequence quality.
    - **ARACHNE** (https://www.genome.wi.mit.edu/wga/) is a free UNIX program for the assembly of whole-genome shotgun reads. Its unique features include using a heuristic approach similar to FASTA to align overlapping fragments, evaluating alignments using statistical scores, correcting sequencing errors based on multiple sequence alignment, and using forward–reverse constraints.
    - **EULER** (http://nbcr.sdsc.edu/euler/) is an assembly algorithm that uses a Eulerian Superpath approach, which is a polynomial algorithm for solving puzzles such as the famous "traveling salesman problem": finding the shortest path of visiting a given number of cities exactly once and returning to the starting point.

GENE ANNOTATION

- Before the assembled sequence is deposited into a database, it has to be analyzed for useful biological features. The genome annotation process provides comments for the features.
- This involves two steps:
    - Gene prediction.
    - Functional assignment.
- As a real-world example, gene annotation of the human genome employs a com- bination of theoretical prediction and experimental verification.
- Gene structures are first predicted by ab initio exon prediction programs such as GenScan or FgenesH.
- The predictions are verified by BLAST searches against a sequence database. The predicted genes are further compared with experimentally determined cDNA and EST sequences using the pairwise alignment programs such as GeneWise, Spidey, SIM4, and EST2Genome.
- All predictions are manually checked by human curators. Once open reading frames are determined, functional assignment of the encoded proteins is carried out by homology searching using BLAST searches against a protein database.
- Further functional descriptions are added by searching protein motif and domain databases such as Pfam and InterPro as well as by relying on published literature.