Genetics Notes Chapter 14: Genomics

Intro.

- **Genomics**-The study of genomes in their entirety.
- Having the entire genetic sequence allows scientists to study both forward and reverse genetics.
- **Bioinformatics**-The analysis of genome's information content including gene number, gene type, gene products, binding sites.
- **Comparative Genomics**- Considers the genomes of related species.
- **Functional Genomics**- Use of reverse genetics to understand gene functions and protein networks.

14.1-The Genomics Revolution

- Not much important

14.2- Obtaining the Sequence of a Genome

- There are 3 billion bases in the human genome.
- To sequence the genome DNA is cloned over and over again and then cut up into tiny segments of 300-600 base pair.
- These smaller sequences are easier to read and these individual reads are put together in the process of **sequence assembly**.
- **Consensus sequences** and overlaps are found between the various segments, and each segment is fit together like a puzzle to read out a complete sequence of the DNA of an entire chromosome.
- The sheer amount of data resulting from this project necessitated **automation** in order to be able to reduce human error.
- The goal of the human genome project was to produce a highly accurate human DNA sequence that could be used as a reference for all of science.
- Genome sequences can range in level of accuracy and quality from **draft level** (basic sequence there but some inaccuracies) to **finished** (very high accuracies but may be some errors) level, to **complete level** (complete accuracy).
- The general strategy for sequencing the genome is **whole-genome shotgun sequencing**, where the long chromosomes have been broken up into many small DNA segments and then read and rebuilt from there.
- There are two types of whole-genome shotgun sequencing (WGS)
    - **Traditional WGS** which relied on the cloning to DNA in microbial cells and the Sanger dideoxy sequencing technique.
    - **New-Age WGS** are cell free methods that are designed for very high levels of material.
- In traditional WGS, genomic libraries are constructed, made from short cut up segments of DNA sequence.
- These short sequences have been input into accessory chromosomes (such as plasmids, artificial chromosomes, or modified bacteriophages) and are propagated in microbes like bacteria or yeast.
- The insert carrying accessory chromosomes are called **vectors.**

- To generate the genomic library researchers use restriction enzymes that cleave the DNA in multiple places.
- These fragments are cut up into single stranded segments which bind to complementary bases in the accessory chromosome.
- In order to ensure the entire genome is represented, multiple copies of the whole genomes are fragmented.
- The resulting recombinant DNA molecules are propagated through normal DNA replication during the host's growth.
- The host produces identical copies of the sequence, amplifying it and creating **clones.**
- These clones create a **shotgun library.**
- These short sequences are partially read starting from a primer sequence in the vector DNA.
- These short sequences are scanned for overlap and put together into a **consensus sequence.**
- These overlapping reads are assembled into units called **sequence contigs.**
- In next-generation WGS DNA segments are prepared for sequencing in cell free environments without cloning or microbial hosts.
- DNA fragments are isolated and sequenced in parallel with each other.
- Advanced Fluid handling techniques, cameras, and software can detect sequencing products in very low volumes.
- One main Next Generation WGS method utilizes a DNA library of single stranded DNA segments.
- The molecules are amplified using polymerase chain reaction into small "beads" which are inserted into small wells.
- Each well has some deoxyribonucleotides added to it in a certain sequence, if they are complementary to the next base pair in the sequence they release an inorganic phosphate which reacts with special enzymes to produce light which is then detected by a camera. Pyrosequencing
- This approach allows for a lot of sequences to be run at once.
- The main issue with genome sequencing is not the shotgun sequencing and readings but the assembly of the contigs.
- Contig assembly is easier in bacteria because they have fewer base pairs and because they don't have repeating sequences so one sequence comes from one locus.
- DNA sequences of eukaryotes are highly repetitive leading to difficulty in determining the length of the repeated segment in the draft.
- Paired end reads?

14.3-Bioinformatics

- **Bioinformatics**-The study of the information encoded into genomes.
- The genome's information can be thought of the sum of all **protein coding regions, all functional RNA coding regions, and all binding sites** that govern these actions.
- The process of determining the function of each element is called **annotation.**
- The total collection of an organism's proteins is called its **proteome.**
- However determining the kinds of proteins created from a DNA sequence is difficult especially in eukaryotic organisms that have introns, exons, and alternative splicing.

- One approach is **Open Reading Frame** detection which looks at DNA segments that have all the necessary requirements to become coding mRNAs like the proper size and sense codons after intron removal, proper 5' and 3' sequences, and start and stop codons.
- Also the use of libraries of DNA molecules called **cDNA** (complementary DNA molecules) already have their introns spliced out and so when lined up with the complementary genomic DNA researchers can see exactly what sections of DNA are introns and which are exons.
- Short cDNA sequences called **Expressed Sequence Tags** have only the 5' and 3' ends sequenced. This allows us to determine the boundaries of a gene in complementary genomic DNA.
- Consensus motifs are used for finding various DNA segments that code for binding sites but this process is very imperfect.
- Gene function can be determined by trying to match the DNA sequence to DNA sequenced with known function in related animals.
- **Synonymous codons** are two different codons that code for the same amino acid. However, organisms have a unique **codon bias** as to which codons they use the most to get a particular amino acid. If an mRNA's codon frequencies match the biases typical of the animal then it is likely a genuine ORF.

14.4-The Structure of the Human Genome

- It is difficult to pin down which sections of DNA are genes and which are just exons of larger gene regions.
- There are thousands of **psuedogenes** which appear as open reading frames but are not functional due to some sort of mutation.
- Many of these psuedogenes are processed psuedogenes which are reverse transcribed from RNA and inserted into a gene.
- These discoveries and more have caused a drop in the estimates of protein coding genes in the human genome.
- Proteins belong to larger families of related proteins that seem to carry out similar biological function. These families of proteins are much larger in humans than in invertebrates.
- Proteins are composed of **modular domains** which can be mixed and matched to form new proteins. The number of modular domains per protein is higher in humans than invertebrates.
- Chromosomal mapping has led to the discovery of chromosomal rearrangement mutations. Where sections of chromosomes break off at a "rearrangement break point" and rejoin either another area of the chromosome or a different chromosome.

14.5-Comparitive Genomics

- Comparative genomics between related species allows researchers to identify **conserved regions** of DNA sequence (which is likely to support a vital function) and also plays an important role is showing the phylogeny of how the species diverged.
- **Phylogeny** is the evolutionary history of a related group.
- Closely related genes are called **homologs**. Two classes:
    - **Orthologs**- Homolog genes at the same genetic locus of a different species. Are inherited from a common ancestor.