MIC3011 Notes

Bacterial Genomics & Genome Sequencing

What is genomics?

- Basic genomics involves determining and analysing whole genome sequences, and getting as much information as possible (initially this was simply: what are the genes? What proteins are expressed? What do these proteins do?)
- Now because we can analyse genome sequences more rapidly, we can do more:

1. Comparative genomics – sequencing the genome of a number of species or strains for an individual species, and comparing the gene content and making evolutionary assessments (to give phylogenetic trees)
    - Comparison of genomes from different species, strains, or individuals with different phenotypes
    - Analysis of differences can give information on genetics underlying phenotypes (i.e. what genes are unique that a whole species, and therefore contribute to the specifics of that species in being able to live in their host or niche)
    - Identify virulence genes by comparing pathogens and non-pathogens
    - Identify genes involved in disease states
2. Metagenomics – can determine the genome sequence of a whole complex of organisms in an environmental or biological sample without having to culture the organisms
    - Identify: all genomes in very complex samples, organisms that can't be cultured, complexity of environmental samples
    - Identify human microbiome → comparison with health and disease states
3. Functional genomics – analysing the level of transcription and protein production of all of the genes on a particular genome
    - Uses genomics information to impart functional knowledge as DNA itself does not provide functional knowledge
    - Determining transcripts produced by an organism, translation and protein expression
    - Analysis of protein-protein interactions
4. Epigenomics – can now do whole genome scale analysis of epigenetics
    - Epigenetics is the study of reversible non-sequence changes to DNA that affect gene expression
    - Understanding whole genome epigenetic modifications and how these affect gene expression
        - DNA methylation affects gene expression (but does not change the sequence of the DNA)
        - Histone acetylation affects how tightly compacted the chromosome is, altering the level of gene expression

What is the genome?

- Genome: the entire genetic material of an organism including the main chromosome(s), plasmids, and mobile DNA elements (bacteriophage, transposons)
- Genome determines the total genetic potential of an organism, but both the genome and environment determine the phenotype
    - Note that not all genes are expressed at any given time

Why sequence a genome?

- All the genetic information is stored in DNA sequence—by containing all the genes, it therefore carries the entire genetic potential
- Determination and analysis of the genome sequence will allow prediction of all of the proteins and organism can produce, as well as give an indication of the function of those proteins
- Further analyses will allow understanding of disease states, development of rationally designed drugs (e.g. if a person lacks a particular enzyme, you could design a drug that can overcome this), and understanding of the basis of pathogen virulence (and create a drug that targets this particular virulence factor)

STRATEGIES USED FOR GENOME SEQUENCING: Sequencing by Synthesis

- Classical method is Sanger dideoxy chain termination sequencing and was used for almost all previous genome projects up until ~2005 (still used today, but not for genome sequencing due to high cost and time)
- New 'high-throughput methods' have recently been developed including: pyrosequencing (Roche/454) and short read sequencing (illumina, SoLiD)

Sanger Dideoxy Method

- Dideoxy chain termination sequencing is a DNA synthesis reaction and requires a **DNA template** (the piece of DNA you want to sequence), **DNA primer** (oligonucleotide) to start synthesis, **DNA polymerase** (to do the synthesis reaction itself) and **dNTPs** (Required to make the new strand)
- High accuracy sequencing method (up to 1000 bp per reaction) and requires sequencing thousands of DNA fragments to assemble whole genomes
- Uses both normal deoxynucleotides (dNTPs) and modified dideoxynucleotides (ddNTPs)
- ddNTPs always terminate DNA synthesis
- During DNA replication, the 3' OH group is necessary for the new base to come in and make a chemical bond—involved in phosphodiester bond formation with the 5' phosphate on the incoming dNTP
- dNTPs have the normal 3' hydroxyl (OH) group which is essential for DNA synthesis
- ddNTPs, however, are chemically synthesised dNTP derivatives that lack the 3' OH group (only has a Hydrogen)
- Therefore, once a ddNTP is added to the growing chain, additional nucleotides cannot be added and DNA synthesis stops
- Sequencing involves adding DNA template, DNA polymerase, DNA primer and dNTPs/ddNTPs mix (ratio must be right)
    - All the ddNTPs are fluorescently labelled
    - Extension reactions terminate with the insertion of a ddNTP at all possible bases (based on random chance of when the ddNTP gets incorporated into one of the growing strands)
    - Then analysis of chromatogram output, looking at colour and intensity of fluorescence at every point

Problems with Sanger Sequencing

- Each sequencing **reaction gives only ~1000 bp** of sequence data as longer fragments are difficult to resolve on a gel (Sanger sequencing method requires having to separate DNA fragments that differ by 1 bp on the gel)
- Need to start from a known piece of sequence (since the primer/oligonucleotide must bind in order to start the sequencing reaction)

Two main methods in overcoming the first problem:

1. Primer walking – still requires a known bit of the sequence, but after this you can determine the next part of the sequence, make a second oligonucleotide associated with the new sequence you have found, and extend the sequence
    - Does not require DNA to be cloned prior to sequencing but is very time-consuming as it requires having to make a new oligonucleotide after every ~1000 bps
2. Random cloning approach – break genomic DNA into small pieces, clone each into a bacterial plasmid, and start the sequencing reaction from the plasmid (since the plasmid has known sequences)
    - This requires you to clone the entire genome in fragments into different plasmid molecules

"Random" Approach

1. **Genomic DNA is mechanically sheared into small fragments** (< 2kb) (note that some fragments will overlap)—mechanical shearing is preferred because restriction enzymes often result in non-random cleavage patterns
2. Each DNA fragment is ligated into a plasmid vector (that we know the sequence of) in the same position of the set of plasmids = "genomic library"
3. Amplify the ligated vectors in *E. coli*, and then sequence each different plasmid
4. Each of the insert sequences are different, but since we know the sequence of the plasmid around the cloned DNA fragment, we can design an oligonucleotide that starts at the end of the plasmid
5. Assemble the final sequence by finding fragments with overlapping sequences (**repeats can be a significant problem**, particularly assembling sequences that fall entirely in a repeat region)
    - In bacterial genomes, transposons mostly cause repeat regions
    - Eukaryotic genomes also have large numbers of transposon-related and retroviral-related elements

NEXT GENERATION GENOME SEQUENCING TECHNOLOGIES

- The aim of next generation sequencing (NGS) methods is to be able to sequence quicker and more cheaply
- Methods were developed to remove the gel separation portion of sequencing with reduced costs and allowing many reactions to be carried out simultaneously
- The necessity to know a piece of sequence was overcome by ligating a known piece of sequence on the end of the unknown genome sequence
- There is a range of new low cost, high throughput sequencing technologies including:
    - Roche 454 pyrosequencing (read length: ~500-700 bp)
    - Ion Torrent (read length: ~400 bp)
    - Illumina/Solexa (read length: ~150-300 bp)
    - SMRT sequencing (read length: ~1,000-40,000 bp)—this will overcome the assembly problem

High-Throughput Pyrosequencing (454 Sequencing)

- DNA synthesis reaction is linked to **measurement of amount of light released**
- Uses a credit card sized reaction plates with **>200,000 reactions** carried out simultaneously (sanger sequencing could only have 96 at once on one gel)
- Each sequencing reaction yields 500-700 bp of sequence data
- 98% coverage of a bacterial genome in a single run (~5 hours)

- No longer running—out of date
- Short read length makes data assembly more difficult as DNA repeats longer than 500-700 bp cannot be resolved

1. Genomic DNA is fragmented
2. Linkers are added (known pieces of DNA) to the ends of each of the fragments
3. Each DNA fragment is attached to individual beads (one piece of DNA per bead)
4. DNA on each bead is PCR amplified (this method measures amount of light released with the addition of each nucleotide using coupled reactions, and one molecule will not release enough light to be detected)
    - PCR reaction is carried out in an oil base, meaning that the oil forms around the bead and the PCR reaction is confined to the bubble of oil—this way, there is no cross-contamination between the beads
5. DNA coated beads placed in separate wells of the plate (and so each well contains millions of the one DNA fragment)
6. When a new base is added on, pyrophosphate is released (and a hydrogen)—this is linked with enzymes, sulfurylase and luciferase; the former takes the pyrophosphate and makes ATP, and the latter takes the ATP and, together with luciferin, makes light
7. Each dNTP is flowed sequentially over all wells at once—if light is released from a well it means that nucleotide was added in that well
    - Light release is linear with the number of bases added, i.e. if two Gs are added, then there will be double the amount of light released (e.g. GTA**GG**)
    - Not perfectly linear, and thus **long tracks of the same base is a problem**

Ion Torrent

- Uses a similar methodology to pyrosequencing (sequencing by synthesis, and flows one base at a time across a chip) but **allows 5,000,000 parallel reactions**
- Does not measure light, but instead H+ that is released with pyrophosphate with the addition of a nucleotide
- Ion torrent **measures the pH change** which results from H+ release (voltage change)
- Voltage and pH change is more linearly accurate than light change

Illumina Short Read Sequencing

- Very high-throughput sequencing method
- Method relies on a single base addition per cycle and **fluorescent imaging**

1. Ligate common oligonucleotides/adaptor sequences to both ends of each fragment
2. Bind DNA fragments to a chip
    - On the slide/chip, pieces of DNA sit there with a lawn of primers that have complementarity to the oligonucleotides that were added to the fragments
    - DNA fragments are heated such that they become single-stranded and then are added to the chip—this must be done at a relatively low density so that individual DNA fragments are bound at a distant from each other
    - Amplify each individual DNA fragment in the area where it was initially bound to form clusters via bridge PCR
        - The linker not bound to the chip can bend over and form a hydrogen bond with a primer nearby on the chip, thus forming a bridge
        - The complementary strand can then be synthesised since there is now a double-stranded section
        - Heat up, hydrogen bonds will break, and the double-strand will become two full lengths of the original fragment
        - Reverse strands are cleaned and washed off, leaving the forward strands (or vice versa)