20.106J – Systems Microbiology
Lecture 12
Prof. DeLong

- Chapter 15 – Brock
- Demonstration: Prof. Delong shows us a 48 capillary array DNA sequencing device with an optical detector.
    - You can do some 500 runs on the same polymer
    - Then Prof. Delong shows us a 1.5 million well plate – you can do 1.5 million sequencing reactions all at once
        - Each well is around 60 microns across, and you get one bead in each well

- Genomics II
    - Comparative genomics basics
    - Gene calling
    - Bioinformatics web resources – there are a huge number of tools
    - Genomics of insect endosymbionts

- We each have around 10^{14} microbes associated with us, that we carry around in our bodies
    - We all have our own genes that predispose us to certain diseases
    - It's not that well understood how the human-associated microbe interacts with human genetics
    - There are some really neat studies going on right now involving how mice genetics determine which microbes live in the mouse's gut

- Genomics is the starting point

- What do we do with a sequence, once assembled?
    - The first thing to do is to translate it into the reading frame – look at which codons, and start to discover what the genes are
    - The range of the number of base pairs and genes stretch from around 400 protein encoding genes up to around 10,000.
        - The organisms with the fewest genes are obligate endosymbionts
        - The ones with the most genes are more metabolically versatile
    - Remember that bacterial and archeal genomes are circular.

- If you look at a bacterial sequence…
    - Table: Gene functional groups identified in the DNA sequence of E. coli K-12
    - To this day, only around 60% of E. coli genes have been identified and described – and E. coli is the most well understood microorganism in the world
        - This is getting better, but it's still kind of a sobering thought.

- Table: Gene function in bacterial genomes: percentage of genes on chromosome in each functional category
  - A larger proportion of the genes in smaller genomes are related to transcription and translation, because those are essential.
  - In larger genomes, a greater percentage is devoted to other functions, which are expendable in the smaller genomes.
- Graph: Relative percent of ORFs vs. Total ORFs in genome
- Figure: Different strains of E. coli can differ by ~1Mbp!
  - Non-pathogenic, uropathogenic, and enterohaemorrhagic
  - Only around 40% of their genes are found in all three strains
  - Around 47% of the genes are found in only one strain.
  - E. coli can have many different phenotypes depending on these genes

- Diagram: Human genes are shared with…
  - Less than one percent of human genes are unique to humans

- Kyoto Encyclopedia of genes and genomes (KEGG)
  - Metabolic maps that you can pull off the Web
  - KEGG map – Prochlorococcus marinus
  - Figure 15-7 from Brock: a diagram of the many identified genomic activities going on inside a cell – a model of what the cell is doing with its DNA

- How do you make sense of a DNA sequence? How do you find the genes from just the list of A, T, C, and G?
  - You can recognize the genetic code from the start and stop sites – this lets you know where there might be gene sites
  - You have to figure out what the right reading frame are – there are always three possible frames.
    - You can use a program to find open reading frames
  - Gene finding – current methods.
    - Homology method/Extrinsic method – use other genomes that have already been sequenced, and compare
    - Gene prediction method/Intrinsic method – look at the codon content
  - Content sensors and gene prediction tools – there are many different programs
  - Demonstration: Prof. Delong pastes a DNA sequence (around 70 kilobase pairs) from a marine microorganism into a sample commercial program on the web
    - It gives back a list of possible genes, with around one gene for every thousand base pairs
  - These possible genes are then pasted into another program: BLAST.
    - It compares the sequences with known genes

        - As it turns out, the sequence in question was already in the database, because the genes all match up perfectly with genes that have already been sequenced
        - These genes are probably involved in cell wall synthesis
    - BLAST lets you compare nucleotides to nucleotides and protein sequences to protein sequences
    - The possibilities are out there for comparing almost any sequence that you might have.
    - Bit score
    - The statistics depend on:
        - The size of the database you're comparing to (this can be corrected for)
        - The length of your sequence (this is what it really depends on)
    - There are a number of different ways that you can compare genomes
        - Operons are different in different organisms – this is useful
    - COGs – Clusters of Orthologous Groups
    - You can categorize genes in different ways

- What else can you do with all these protein-encoding genes?
    - DNA microarrays
    - Proteomics – isolate proteins from cells, cleave proteins with proteases or CNBr, fractionate proteins, detect and identify proteins
        - We used to do this with gels. Now we can put the whole tissue in a mass spectrometer.