Genomics II

1. Functional Genomics

a) Microarrays – used for transcriptomics. When, where, and to what extent is a gene expressed? In most cases, two or more different samples are compared (male *vs*. female, healthy *vs*. cancerous cells, different stages of development, *etc*.) and one looks for genes that are expressed at different levels in the two samples.

Microarrays are made by attaching many DNA sequences to a small surface, typically a glass microscope slide. Each unique DNA sequence is placed in a "spot" of known location and there may be thousands (or even millions) of spots on a single array. The DNA placed on the array may come from cDNA clones, synthesized oligonucleotides (25-80 bases), or PCR-amplified genomic DNA.

In most applications, mRNA is purified from the two samples of interest and reverse-transcribed into cDNA. The cDNA from one sample (A) is labeled with a "red" fluorescent dye; cDNA from the other sample (B) is labeled with a "green" fluorescent dye. The cDNA from the two samples is mixed in equal amounts and hybridized to the microarray. The cDNAs will anneal to the DNA spot complementary to their sequence. Using a laser scanner that can quantify the fluorescence of the two dyes, one can measure the relative amount of expression of each gene by the red/green signal at each spot.

Genes with higher expression in sample A: stronger red signal
Genes with higher expression in sample B: stronger green signal
Genes with equal expression samples A and B: equal red and green signal = yellow

b) Gene "knockouts"- this is also known as reverse genetics. One starts with a particular target gene sequence and looks for the resulting phenotype when this gene is mutated or its expression is knocked-out. This is the opposite of forward (or classical) genetics, where one knows the mutant phenotype and tries to find the responsible gene.

Homologous recombination – works well in yeast, mouse, other mammals.
A "knockout" DNA sequence (usually a plasmid) that shares homologous end sequences with the target gene is constructed *in vitro*, then introduced into the nucleus of a cell. In some cases, recombination will occur between the plasmid and the chromosomal DNA. This will result in the corresponding segment of chromosomal DNA being replaced by the knockout sequence.

RNA interference (RNAi) – works well in *C. elegans*, *Drosophila*, other insects, plants, mammalian cells, *etc*.
Double-stranded RNA (dsRNA) complementary to the target gene is introduced into the cell (through injection, feeding, soaking, or transgenic methods). This dsRNA activates an innate defense pathway (probably evolved as a response to TE's or retroviruses) that leads to the degradation of the corresponding mRNA. This is post-transcriptional gene silencing (PTGS). That is, the gene is still intact in the nucleus, but its expression is prevented because the mRNA is degraded before it can be translated into protein.

Many knockouts by either of the above methods have no obvious effect!

For example, less than 20% of all yeast genes are required for growth in rich medium and only 40% show measurable growth defects in rich or minimal medium when knocked-out.

Only about one-third of all mouse genes lead to inviable or infertile mice when knocked-out.

This has led some researchers to classify genes as "essential" and "dispensable".

2. Comparative Genomics

What is the same? What is different?

a) Genes conserved among organisms. Conservation across diverse species suggests a shared, important function.

Example: human disease genes present in model organisms
Of human genes known to cause diseases, 80% have homologs in the fly (*D. melanogaster*), 73% in the worm (*C. elegans*), and 42% in yeast (*S. cerevisiea*). This suggests that many human disease-causing genes perform essential cellular functions that are conserved across eukaryotes. Furthermore, it suggests that the study of distantly-related model organisms is relevant to human health and medicine.

b) Genes NOT conserved - may be responsible for phenotypic differences between species.

Example 1: *Mycobacterium leprae vs. M. tuberculosis*
These are two closely-related bacteria with very different phenotypes. *M. leprae* causes leprosy, *M. tuberculosis* causes tuberculosis. *M. leprae* has the longest doubling time of any known bacteria and cannot be cultured in the lab.

M. leprae appears to have lost function of half of its genes. Most of these are still present in the genome as pseudogenes.

Example 2: Human *vs*. chimpanzee. What makes us human?

Some findings from the chimp genome project:

- overall, human and chimp are ≈ 98.75% identical in DNA sequence
- on average, human and chimp proteins differ by 2 amino acids
- the greatest divergence (in both protein sequence and gene expression) is in the testes
- the least divergence (in both protein sequence and gene expression) is in the brain

3. Evolutionary Genomics

How do genomes change over time? What forces are responsible for these changes?

a) Rates of DNA loss
Laupala (Hawaiian crickets) have a genome 11 times larger than that of *Drosophila*. Rates of DNA deletion in "neutral" dead-on-arrival (DOA) transposable elements indicate that there is a much faster rate of DNA loss in *Drosophila* through spontaneous deletion mutations. Is this why the *Drosophila* genome is so small and has no pseudogenes? Can this explain the C-value paradox?

b) Gene duplication
Possibilities following duplication:

- one copy loses function through mutation and becomes a pseudogene
- selection favors keeping multiple copies of the same gene -> more product
- one copy gains new function that is favored by selection (neofunctionalization)
- the original gene had two functions, one of which is lost in each of the two copies. Then both copies are retained, but no new function has been gained (subfunctionalization).

Example: *janusA-janusB-ocnus* in *Drosophila*

After duplication, the three genes evolved at different rates (Ka/Ks), suggesting they are under different selective constraints and have diverged in function.

c) Patterns of DNA sequence variation
Patterns of variation along a chromosome can be used to identify selective sweeps and determine the location of the selected site(s).

Example: *janus-ocnus* region in *Drosophila simulans*.

Here there is a haplotype in high frequency, suggesting positive selection.