Genomics

1. The world of "-omics"

The field of "Genomics" has seen explosive growth over the past 10-15 years, giving rise to new terminology and new disciplines. For example:

Individual item	Collection in an organism	Field of study
Gene	Genome	Genomics
Protein	Proteome	Proteomics
Transcript (mRNA)	Transcriptome	Transcriptomics
Metabolite	Metabolome	Metabolomics
Phenotype	Phenome	Phenomics
Interacting proteins/genes	Interactome	Interactomics

These last four fields of study also can be put in the category of "functional genomics"

There are also sub-fields within genomics:

Genomics – Now a broad category covering several sub-disciplines. Initially referred to the process of determining the complete DNA sequence of an organism's genome.

Functional Genomics – Experimental identification of functional regions of genome.

a) Patterns of gene expression – where/when is the gene expressed? (transcription profile)
b) What is the function of the gene product?
c) What other genes does it interact with?

Comparative Genomics – Comparison of genomes between species. Identification of evolutionarily conserved (functional) sequences (genes, regulatory sequences) and functional differences between species (*e.g.* pathogenic *vs.* nonpathogenic bacteria).

Evolutionary Genomics – How do genomes change over time? What forces drive these changes?

2. Genomics

a) How big is a genome?

The scale of genomes:
1,000 base pairs (bp) = 1 kilobase (kb); scale of individual genes
1,000,000 bp = 1000 kb =1 megabase (Mb); scale of bacterial genomes
1,000,000,000 bp = 1,000,000 kb = 1000 Mb =1 gigabase (Gb); scale of vertebrate genomes

The size of the genome (C-value = the amount of DNA in a single haploid nucleus) depends on the organism. However, there is not a strong correlation between organism complexity and genome size. This is known as the C-value paradox.

b) How do you sequence a genome?

Problem: Standard (Sanger) DNA sequencing methods can only read 500-1000 bases per reaction.

"Clone by clone" – Genome is cloned into large-insert vectors, usually BAC (bacterial artificial chromosome) vectors of ≈ 100 kb. These are mapped to form a minimal overlapping set. Then each clone is broken up and sequenced individually. This method was used by publicly-funded human genome project.
Pro – easy assembly, less redundant sequencing, clones available to "finish" sequence
Con – requires initial mapping of clones

"Shotgun"- Entire genome cloned into small-insert vectors, usually plasmids of ≈ 2kb. A large number of these are sequenced at random. The raw sequence is then assembled by computer to reconstruct the genome. This method was used by Celera Genomics to sequence the human genome.
Pro – fast, mapping of clones not necessary
Con – assembly and "finishing" are difficult

As an example of the scale of the assembly problem for shotgun sequencing, consider the following:

Celera Genomics sequenced the *Drosophila melanogaster* genome to ≈10-fold coverage. The *Drosophila* euchromatic genome is 120 Mb. This means 1.2 Gb were sequenced. With an average read length of ≈ 500 bp, there were ≈ 2.4 million sequences.

Step 1: Pairwise comparison of each sequence to every other sequence to search for overlaps. The total number of pairwise comparisons is given by the formula:

$n\,(n\text{-}1)\,/\,2$, where n is the number of sequences.

Filling in the numbers gives ≈ 3×10^{12} comparisons.

There are ≈ 3×10^{7} seconds in a year. So if you did one pairwise comparison per second it would take ≈ 100,000 years. The Celera supercomputer was capable of 32 million comparisons per second. Still this would require ≈ 24 hours for initial comparison. The human genome (27 million reads) would take ≈ 100 days. This can be shortened with parallel processing.

c) Where are the genes?
de novo (or *ab initio*) prediction – use computer programs (*e.g*. Genscan, Genie) to identify genes from raw DNA sequence data. Look for long open reading frames (ORFs) that start with ATG end with stop codon. Can also incorporate intron splice signals, or codon bias information, *etc*.
Pro – fast and easy to implement, no additional experimental work required.
Con – computer algorithms are not good when there are many, long introns. Are predicted genes "real"?

Comparative prediction – Look for sequences sharing significant homology with other, known genes. Can compare different species. If ORFs are conserved between species, they are likely functional.
Pro – fast and easy to implement. Homology often gives hint to gene function.
Con – overlooks unique or fast evolving genes. Requires sequences from related organisms.

Experimental identification – mRNAs are isolated from the organism and converted to cDNAs, then sequenced on a large scale.
Pro – experimental evidence that genes are expressed. Inron/Exon boundaries can be determined.
Con – requires much experimental work. Genes expressed at low levels or regulated temporally or spatially may be overlooked.

d) What else is there?
Protein-coding sequence represents only ≈ 2% of the human genome.

repetitive DNA – long stretches of the same DNA sequence repeated many times in tandem. It is enriched at centromeres and telomeres and referred to as "heterochromatin". It does not clone easily in bacterial vectors and is mostly excluded from genome projects.

transposable elements (TE's) – Interspersed repetitive DNA. TE's are pieces of DNA that can replicate and move within the genome. They are sometimes called "jumping genes" or "selfish DNA". TE's make up about 1/2 of the human genome.

pseudogenes – genes that are no longer functional. These are often duplicated genes that are not expressed and/or have a disrupted ORF (internal stop codon or frameshift).