Basic Genomics

1. How do you sequence DNA?
Two methods developed in the mid-1970's:
Maxam-Gilbert (chemical) – rarely used
Sanger method (dideoxy, enzymatic) – Developed by Frederick Sanger and is still used today with little change to the basic method, although great improvements have been made in efficiency and automation.

Requirements:

1. Template DNA (to be sequenced), typically purified plasmid + insert DNA
2. Specific primer DNA (≈20 nucleotides long)
3. DNA polymerase
4. Deoxynucleotides (dATP, dGTP, dCTP, dTTP = dNTPs), high conc.
5. Dideoxynucleotides (ddATP, ddGTP, ddCTP, ddTTP = ddNTPs), low conc = "terminators"
6. Labeled deoxynucleotides (radioactive dNTPs or fluorescent ddNTPs), low conc.

Steps:

1. DNA and primers heated to denature, then cooled to anneal
2. Polymerase adds nucleotides specific to template on end of primer
3. Occasionally a ddNTP is incorporated and the enzyme stops
4. Fragments of different lengths are separated and the sequence is "read"

Original Method:

1. 4 reactions run, 1 with each ddNTP, radioactive labeling
2. Fragments separated on polyacrylamide "slab" gel, 1 lane per ddNTP
3. Entire gel dried and exposed to X-ray film
4. Sequence interpreted from band order by human inspection

Modern Method:

1. All 4 fluorescently-labeled ddNTPs used in 1 reaction, each a different "color"
2. Fragments separated in matrix-filled capillary tubes, 1 capillary per sample
3. Laser detects fluorescence automatically as each fragment exits capillary
4. Computer software "calls bases" and processes sequence files
   (if sequences were processed by human at 15 min. per sample, it would take 7 people a full-time week to process 1 day's output from an automated sequencer)

The biggest problem in genomics still remains: the longest continuous stretch of sequence that can be read by a single sequencing reaction is ≈1,000 bp, and the high-quality portion is typically only 500-700 bp.

Sequencing larger pieces of DNA:

1. Sequence Walking - new primer designed to match end of previous sequence
   a) pro: minimum amount of sequencing, no assembly required
   b) con: very slow and expensive. must wait for results of 1 reaction before performing the next. Must design custom primers each time.

2. Shotgun Sequencing - sequence all pieces at once, then assemble
    a) pro: much faster, cost-efficient, high-throughput parallel processing, universal primers
    b) con: requires much more (redundant) sequencing, must be assembled

"Next Generation" sequencing:
Over the past few years, new sequencing methods have been developed. These include commercial methods such as 454 (Roche FLX), Solexa (Illumina), and SOLiD (ABI). In general, these use massively-parallel methods to simultaneously sequence millions of short pieces of DNA (read lengths are usually around18-200 bases). At present, the read lengths are shorter than those of the traditional Sanger method and the error rates are higher, but these methods have already been applied to genome-scale projects.

2. 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. As you might expect, the list of organisms with completely sequenced genomes is biased towards those with lower C-values.

Some examples from sequenced genomes:

Species	**Genome size**
Mycoplasma genitalium (bacteria)	580 kb
Haemophilis influenzae (bacteria)	1.8 Mb
Escherichia coli (bacteria)	4.7 Mb
Saccharomyces cerevisea (yeast)	12.5 Mb
Caenorhabditis elegans (worm)	97 Mb
Arabidopsis thaliana (mustard weed)	125 Mb
Drosophila melanogaster (fruit fly)	180 Mb
Fugu rubripes (puffer fish)	400 Mb
Oryza sativa (rice)	400 Mb
Homo sapiens (human)	3.2 Gb

3. How do you sequence a genome?
"Clone by clone" – Genome is cloned into large-insert vectors (BAC, YAC, P1; 100-200 kb) and these are mapped to form a minimal overlapping set. Then each clone broken up and sequenced individually by shotgun method. This is also known as "hierarchical shotgun sequencing". Used by publicly-funded human genome project.
pro: less redundant sequencing, easily subdivided, easier assembly, clones available for "finishing" and further research
con: clone mapping is difficult and requires a lot of time

"Whole Genome Shotgun (WGS)"- Entire genome cloned into small-insert vectors and all are sequenced. Raw sequence is then assembled by computer to reconstruct genome. Used by Craig Venter and Celera Genomics for human genome.
pro: can start almost immediately, save time of clone mapping, faster, less expensive?
con: much harder to assemble, may be many gaps, gaps difficult to close

Assembly statistics:
Coverage = average number of times each base is sequenced
Since sequenced clones are chosen at random from a large pool, this can be approximated by a sampling-with-replacement scheme. The probability that a base is not sequenced is estimated from the Poisson distribution as $P_0 = e^{-m}$, where e is the base of the natural logarithm and m is the sequence coverage.

Assumptions:

1. Sampling with replacement (starting genomic DNA >> sequenced DNA)
2. All pieces of DNA clone with equal frequency. This is not 100% true. Repetitive DNA, inverted repeats, high %GC or low %GC fragments often do not clone well in bacteria. In general, the highly-repetitive parts of a genome (usually around the centromeres and telomeres) that contain few genes are known as heterochromatin and are not cloned or sequenced in genome projects.

Examples:

Coverage	e^{-m}	**% of genome not sequenced**
1-fold	0.37	37%
2-fold	0.14	14%
3-fold	0.05	5%
4-fold	0.02	2%
5-fold	0.0067	0.67%

Total gap length is Le^{-m},
average gap size is L/n,
where L is genome length and n is the number of random fragments sequenced

Examples: assuming 500 bp per read

Bacteria (genome size = 2 Mb)

Coverage	**Reads**	**Unsequenced**	**Avg. Gap Size**	**Gap Number**
1-fold	4,000	740,000 bp	500 bp	1,480
5-fold	20,000	13,400 bp	100 bp	134

Vertebrate (genome size = 2 Gb)

Coverage	**Reads**	**Unsequenced**	**Avg. Gap Size**	**Gap Number**
1-fold	4,000,000	740,000,000 bp	500 bp	1,480,000
5-fold	20,000,000	13,400,000 bp	100 bp	134,000

4. Where are the genes?

de novo (or *ab initio*) prediction – use computer programs (such as Genscan or Genie) to identify genes from raw DNA sequence data. Look for long open reading frames (ORFs) that begin with a start codon (ATG) and end with a stop codon (TAA, TAG, TGA), but contain no internal stop codons. Can also incorporate intron splice signals, or codon bias information.
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 – mRNA is isolated from the organism and converted to cDNA, then sequenced. These sequences are often referred to as ESTs (Expressed Sequence Tags).
Pro – experimental evidence that genes are expressed. Intron/Exon boundaries can be determined.
Con – requires much experimental work. Genes expressed at low levels or regulated temporally or spatially may be overlooked.

5. What else is there?

Protein-coding sequence represents less than 2% of the human genome.

Repetitive DNA – long stretches of the same DNA sequence repeated many times in tandem. This makes up most of the heterochromatin. It is enriched at centromeres and telomeres.

Transposable elements (TE's) – pieces of DNA that can replicate and move within the genome. Also known as "Interspersed repetitive DNA", "jumping genes" or "selfish DNA". Make up almost 1/2 of the human genome. Many copies are "dead" or partial TE sequences that can no longer "jump" and are just relics of previously-active TE's.

Pseudogenes – genes that are no longer functional (often duplicates of functional genes). Typically have a stop codon or frame-shift within their ORF. May have lost their promoter and not be expressed.

Other sequences, such as introns and intergenic sequences, do not encode proteins, but may contain important gene regulatory information or may be transcribed into functional, non-coding RNA. Some of the non-coding regions are highly conserved across species, which suggests that they have an important function. However, the function of most non-coding DNA is unknown and it is possible that much of it is functionless "junk DNA".