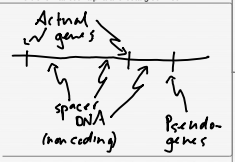
Chromosomes and Chromosome Biology

Genome composition

- The human genome contains large amounts of repetitive DNA (see lecture 1)
  - These are dominated by types of **transposable elements** and their derivatives
    - LINEs (long interspersed nuclear elements) - 20% of human genome
      - Retro-transposon type sequences between 6kb and 8kb long
      - Contain 2 open reading frames, encoding a **reverse transcriptase** and an **integrase**
      - LINEs copy themselves and insert into new genomic locations as follows
        - The LINE element is transcribed by cullular RNA polymerase
        - The LINE transcript is then copied into DNA by the element-encoded reverse transcriptase
        - The integrase then inserts this DNA back into the genome at a new site
      - Most lines in the human genome are partially deleted and therefore no longer transpose
      - Some elements of the LINE1 family appear active
    - SINEs (short interspersed nuclear elements) - 13% of human genome
      - Hitchhike on the enzymatic activity of the LINEs
      - Transcripted by RNA polymerase III and use the LINE encoded eenzymes to produce DNA and integrate themselves into the genome
      - The most abundant SINEs are Alu repeats (so called caus' they have a recognition site for AluI)
    - LTR-like sequences
      - These resemble retroviruses
      - They encode **reverse transcriptase** (*pol*), **group specific antigen** (*gag*) and **envelope** (*env*) proteins
      - They are flanked by **long terminal repeats** (LTR)
    - So far, we've been seing **retrotransposons** - those which originate from reverse transcriptase copied RNA
    - A final class move by a DNA-based cut and paste mechanism
      - (Or by a DNA based replicative mechanism)
      - They use transposon-encoded **transposase** protein
      - No RNA intermediate
  - Simple sequence repeats (SSRs)
    - Perfect or imperfect repeats of short sequences occupy ~3% of the human genome, occuring on average every 33 kb
    - Many SSRs are **polymorphic** in the population, meaning that there are different alleles on each repeat locus
      - One of the most common is the CA repeat, with n between 10 and 100
      - i.e.: different number of repeats
      - This means that
        - An individual is likely to have different alleles on each of a pair of chromosomes, allowing us to follow particular regions of the genome by the transmission of SSRs
        - Individuals are likely to have different spectrums of SSRs if a number of loci are compared
    - This is immensely useful in DNA fingerprinting
  - Repeated genes
    - The repeats we have encountered so far are all apparently junk
    - Sometimes, however, the genome needs a lot of a particular gene product (e.g.: ribosomal RNA or histones), and in many cases, the genes coding for these products have been amplified
      - Tandem arrays (e.g.: rDNA)
        - Large tandem arrays, each repeat consist of the 3 rRNA found in the ribosome
        - Each repeat is separated bya "non-transcribed spacer"
        - rDNA genes are usually localized to the nucleolus
        - In some species, the transcription can be visualised as a Christmas tree-like structure
        - Recombination between arrays can lead to deletions and duplications
      - More complex arrangements (e.g.: histones)
        - The organisation of repeats of DNA coding for histones is far more complex than that for DNA
        - It can also differen between species
        - Recombination between arrays can lead to deletions and duplications, and also to new arrangements of histone genes
    - Another kind of gene replication is interesting from an evolutionary perspective
      - Classic examples
        - The globin gene family
          - This has been one of the driving forces behind the generation of diversity
          - Genes are duplicated and then, freed from the constraints of selection, can evolve new functions
          - Both alpha and beta globin genes have been duplicated during evolution and have evolved slightly different functions
          - Include specialised genes expressed in the embryo and foetus
          - Remains of this process can be seen in the degenerate non-functional *pseudo-genes* "left-over" at both alpha and beta globin loci
          - 

          - These form a large portion of the 11% "misc." part of the human genome
        - Hox gene clusters
          - Found in all metazoans (animals)
          - Control development
          - Regulate how sets of genes are turned on in different parts of the body
          - All are clearly related, and it is believed that they have been maintained as a group to facilitate precise control of expression
  - Actual genes
    - Very little space left for these!
    - The actual number in a given organism is a throny issue, depending on the quality of the sequence and its annotation
    - In unicellular organisms, gene numbers are fairly low and the genes are very densely packed in the genome
      - These divide failry quickly (E. Coli > every 20 min under good conditions)
      - Their DNA therefoer also replicated very quickly, and they have "eliminated waste" from their genome
    - In metazoans, genes are more numerous and "space out"
      - However, even here there can be up to 5-fold differences in gene densities between related species (cf rice and maize)
      - In part, this is due to the fact the genes themselves are rather more complex
    - If we look at the number of exons a gene contains...
      - Almost all budding yeast genes have a single exon
      - 83% of drosophila genes have at least one intron
      - 5% of mammalian genes have over 30 introns
      - Some are very complex!
    - This makes the annotation of metazoan (and in particular mammalian) genes difficult
      - We are limited to a set of experimental and comparative tools
      - Compare with processed mRNAs (obtained from cDNA libraries)
      - Compare with the primary sequence of proteins
      - Compare with information derived from simpler organisms

Centromeres

- After DNA replication, chromosomes must be separated during mitosis
- To achieve this, chromosomes are coordinated with the **mitotic spindle** via their **centromeres**
- These are specialised structures on each chromosome that interact with a large number of proteins and the **microtubules** of the **spindle** to form a structure known as the **kinetochore**
- The best characterised centromeres are those of the budding yeast **Saccharomyces cerevisiae**
  - If is a 125bp long sequences (roughly 0.1% of chromosome) with an AT rich core
  - The core coordinates a set of protein complexes that secure the spindles to the chromosome
- Those of the fission yeast **schizosaccharomyces pombe** are less understood
  - They're much longer (25kb to 110kb - up to 4% of the chromosome)
  - They're composed of a complex arrangement of repeats flanking the inner core bordered by direct repeats
  - The reason for this dramatic difference is unknown
- Most poorly characterised are **mammalian centromeres**
  - Since they are extremely repetitive, they haven't yet been fully sequenced
  - They appear to be composed of long tandem arrays of 171bp alpha satelite repeats
  - As with yeast, a complex array of proteins are assembled at the centromere (a very large number of which have been identified!)
- A 420kb drosophila centromere has been described at the sequence level
  - A functional drosophila centromere was identified and isolated by virtue of its ability to confer mitotic stability of a marker gene
  - Sequence analysis reveals a complex organisation of simple sequence repeats interspersed with several different types of transposable elements

Introduction

- Chromosomes
  - In eukaryotes, the genome is partitioned into chromosomes
  - Composed of a complex arrangement of DNA and proteins, including a set of functional elements necessary to ensure that DNA is efficiently replicated and partitioned into daughter cells
  - Replication is initiated at **DNA replication origins**
  - At the end of chromosomes, **telomeres** maintain the integrity of the chromosome
  - At cell division, the **centromere** and **kinetochore** facilitate chromosome segregation
  - "Junk" DNA is pacakged into highly compacted chromatin (**heterochromatin**) and genes are found in less compact **euchromatin**
- Banding patterns
  - Cytogeneticists have utilised a variety of staining techniques to generates maps based on invariant **banding patterns**
  - G-Bands are revealed by staining with **giemsa** after treating with a protease to digest some of the proteins
    - Such darkly stained band have low G+C content
    - They also generally replicated late in the cycle
  - R-Bands, however, are obtained by Geimsa staining after incorporation of the nucleotide analogue BromodeoxyUridine
    - Dark bands hare are G+C rich
    - And they usually replicate early
  - Even today, this cytological analysis can remain useful
    - For example, some cancers involve translocations or delections, which modify banding patterns

Replication origins

- The structures of some bacterial replication origins are well characterised and understood
- The best characterised eukaryotic replication origins are **yeast autonomously replicated sequences** (ARS) which binds a complex set of ORC proteins (origin recognition complex) that act to control DNA replication
- Very few metazoan replication origins have been defined
  - The situation is more complex than in yeast, though
    - For example, in early embryos, each chromosome originally replicates from multiple origins
    - Eventually, as the cell cycle gets longer, the number of origins used per chromosomes deplines
    - Similarly, different parts of the chromosome replicate at different times (*cf* banding analysis)

Telomeres

- In eukaryotes, DNA exists as a linear molecule, and information is therefore lost from the end of each chromosome at each DNA replication
- The ends of chromosomes are protected by special terminal structures known as **telomeres**
- These consist of arrays of short G-rich repeats
- They stabilise the ends of chromosomes by forming a complex T-loop structure that binds several telomere specific proteins
- When telomeres become too short, cell death or senescence is triggered

DNA fingerprinting

- Introduction
  - DNA samples from two unrelated people differs on average by only one base per thousand
  - Since the diploid genome contains 6 billion bp, this amounts to 6 million differences
    - Crudely, the chance that two unrelated people have an identical total DNA sequence is 1 in 6 million
    - This becomes more and more probably for closely related people
    - And is definite for identical twins
  - This allows us to use DNA to identify people
  - Twould be a bit tough to sequence a person's entire DNA, though!
  - We can instead take advantage of polymorphisms associated with SSRs
- The method
  - We take specific primers that bind to unique sequences **flaking** a polymorphic SSR
  - We use those in a PCR reaction with the DNA to be identified
  - Since each chromosome in the pair will have a different allele (length) of SSR, each individual should produce 2 fragments in a PCR reaction
  - These can easily be resolved on a polyacrylamide gel (capable of base-pair resolution)
  - Two people compared are unlikely to have the same bands
- Probabilities
  - However, the polymorphisms displayed at each polymorphic region are by thesmelves very common
  - typically each polymorphism will be shared by around 5 - 20 % of the population
  - However, if we look at multiple loci, the the **unique combination of several polymorphisms** in an individual makes the method discriminating enough to be a robust identification tool
  - The national DNA database in the UK is the SGM+ DNA profiling system - it includes 10 different regions and a sex-specific test
  - The probability of an unrelated match is 1 in 1 billion