**PHAR2811: Genome Organisation**

**Synopsis:** C-value paradox, different classes of DNA, repetitive DNA and disease. If protein-coding portions of the human genome make up only 1.5% what is the rest doing?

---

**The C-value paradox**

- The C-value is the total number of DNA nucleotide residues in the genome (per haploid set of chromosomes).
- When you compare this to the complexity of the organism you find a massive disparity.
- Clearly the amount of DNA is not proportional to that required to produce all the proteins made by the organism.

---

**What do these life forms have in common?**

---

**Is there too much DNA?**

- Only about 1 - 2% of the DNA in the genome actually codes for proteins. What is the rest of it doing?
- Some clues come from re-annealing experiments.
- The time it takes for DNA to re-anneal depends on the complexity of the sequence

---

**C_0t plots**

- You isolate the DNA from a species and cut it up to manageable lengths and then you melt it.
- Once it is fully single stranded you monitor, by absorbance at 260 nm, the re-annealing kinetics of the sample.
- The C_0t_1/2 is the time taken (*C_0*) for half the DNA to be reannealed.

---

**C_0t plots**

- You need to account for the starting concentration (*C_0*).
- Obviously if you started with more of a sequence it would anneal quicker….more matches to find each other.
- By using the C_0t value you can plot (on the same graph) different annealing experiments with different starting concentrations (10 – 2 000 ug/mL) from the same source and they will all lie of the graph.
Re-annealing kinetics

$C_0 t$ plots

- The rate of re-association, hence the time taken to renature is dependent on the complexity of the DNA sequence.
- The complexity is defined as the number of bases in each unique sequence e.g. poly (U)+poly (A) has a complexity of 1, the repeating sequence AGTGC<sub>n</sub> has a complexity of 5.

Different DNA samples

$C_0 t$ plots

- The $C_0 t_{1/2}$ for a given DNA depends on the complexity.
- Eukaryotic genomic DNA can be divided up into 4 classes: highly repetitive (hundreds to millions of copies), moderately repetitive (10s to hundreds of copies), slightly repetitive (1 – 10 copies) and single copy sequences.
- The last 2 are often combined

Cot plot for a mixed population: similar to the human genome

Highly repetitive DNA

- Short sequences arranged in tandem repeats, sometimes thousands of times.
- Short Tandem Repeats (STRs) or satellite DNA
- Microsatellites 1 – 13 nucleotides
- Minisatellites 14 – 500 nucleotides
- Often found clustered around the centromere or the telomere.
**Moderately repetitive DNA**
- Segments of 100 to several thousand base pairs repeated
- Repeated groups of genes whose products are needed by cells in large quantities e.g. histones, ribosomal and transfer RNA (although these are sometimes classified in the highly repetitive group)
- Retrotransposons, DNA which has been transcribed in reverse back from RNA

**Retrotransposons**
- Around 40% of the human genome
- LINES (long interspersed nuclear elements) 6 – 8 kb segments that encode the proteins that enable the transposition
- SINES (short interspersed nuclear elements) 100 – 400 bp sections containing remnants of tRNA transcription machinery.
- LTR retrotransposons or long terminal repeats

**Gene Families**
- Most genes in the genome are only represented once.
- Some have a few copies on the genome.
- One example is the globin family. This set of genes contains a number of closely related sequences which vary by only a few changes in the code.
- Sometimes found clustered together on the one chromosome (but not always!)

**Single copy genes**
- Most of the genes of the organism are single copy genes
- But they only make up a small proportion of the total genome.
- They are the most complex group and hence take the longest to re-anneal.

**What else helps explain the C-value paradox?**
- So our genome contains highly repetitive DNA which doesn’t code for proteins.
- It also contains some multi-gene families and multiple copies of some genes.

**What else helps explain the C-value paradox?**
- This makes up 40% of the genome by Cot plot analysis.
- What of the other 60% unique sequences?? Remember only 1 – 2% of the genome is coding sequence.
- **How do we account for this discrepancy??**
What else helps explain the C-value paradox?

- Eukaryotic genes have introns which are removed after transcription by processing.
- These introns can comprise some 90% of the gene.

What else helps explain the C-value paradox?

- Pseudogenes; thought to be relics from ancient gene duplications or reverse transcription events.
- They do not produce functional proteins.
- They are sometimes considered as part of the retrotransposon group.

Repetitive DNA and disease.

- Trinucleotide repeats (TNR) are a specialised type of repeat sequence found in the genome.
- They arise from mutations during replication, repair or recombination.
- Both germline cells (sperm and ova, meiotic) and somatic (mitotic)

Static Mutations

- A mutation event in the germline which is stably transmitted to all somatic cells in the next generation.
- Each somatic cell has a copy of the mutation in their genome.
- Examples are sickle cell anemia, Cystic fibrosis, PKU

Dynamic Mutations

- Unlike static mutations dynamic mutations change; they continue to mutate between different tissues and across generations.
- The longer the tract length (i.e. number of repeats) the more likely the repeat is going to continue to mutate.

Dynamic Mutations

- This leads to increased severity with successive generations or in some diseases, the age of onset decreases.
- In other words with each generation the disorder becomes worse and/or you start to get the symptoms earlier. This leads to genetic anticipation.
Repetitive DNA and disease.

- There are an increasing number of genetic disorders: many of them are neurological disorders.
- They result from expansion of trinucleotide repeats (this means multiple copies of a 3 nucleotide sequence).
- The repeat is characteristically CNG

The result of this repeat mutation

- If the expansion occurs in a coding region the result is usually gain of function for the protein
- The protein has an altered function as it now has a whole bunch of amino acids added to the code
- Example is Huntington’s disease

Fragile X syndrome (FRAXA)

- One of the most common inheritable forms of mental retardation
- The fragile X syndrome (FRAXA) results from multiple copies of the sequence CGG (the expansion) in the 5’ UTR of the fragile X syndrome gene, FMR1
- Results in gene silencing of the FMR1 gene product, FMRP.
Fragile X syndrome

- The number of repeats is very important to the final severity of the disease.
- 5 – 50 copies has no effect,
- 50 – 200 results in an intermediate and distinct syndrome, fragile X tremor/ataxia (FXTAS)
- >200 copies gives rise to the full blown mutation.

FMRP: the protein product of the FMR1 gene

- It is located largely in the cytoplasm but does make excursions to the nucleus.
- It has 3 RNA binding domains; it associates with ribosomes and seems to be involved in translational regulation of a group of RNA targets.

FMRP: the protein product of the FMR1 gene

- The protein and its mRNA localise in dendritic spines.
- The protein binds to certain “target” RNA species and represses translation.
- These are thought to be mRNAs coding for proteins involved in neuronal development, synaptic transmission and cytoskeleton

Huntington’s disease

“George Huntington (1850-1916) described the condition while working as a newly qualified doctor in the rural general practice of his father and grandfather on Long Island, New York State. Together their observations covered 78 years. George Huntington did not continue working on Hereditary chorea but went into general practice in Ohio.” He never published another paper in his life - yet his name is remembered from the single one he did write.
Huntington’s disease

- A second example is the expansion of the repeats in the coding region of a gene.
- This alters the function of the protein product, referred to as “gain of function”.
- There are 9 separate disorders are associated with the expansion of a CAG repeat in the coding region of various proteins.

Huntington’s disease

- CAG codes for glutamine and the expansion results in multiple copies of glutamine in the affected protein (polyglutamine disorders).
- This gain of function outcome accounts for the neurodegenerative symptoms of Huntington’s disorder.

How does the mutation account for the symptoms or the pathogenesis?

- The result of the polyglutamine in the protein is a misfolded protein.
- The protein aggregates and is sequestered into inclusion bodies complete with the chaperones.
- This is thought to eventually overload the chaperone and ubiquitin systems.

Huntington’s Disease

- The HD gene was discovered in 1993
- There is now a direct genetic test to make or confirm a diagnosis of HD in an individual who is exhibiting HD-like symptoms.
- Using a blood sample, the genetic test analyzes DNA for the HD mutation by counting the number of repeats in the HD gene region.
Huntington’s Disease

- Individuals who do not have HD usually have 28 or fewer CAG repeats.
- Individuals with HD usually have 40 or more repeats.
- A small percentage of individuals, however, have a number of repeats that fall within a borderline region.

<table>
<thead>
<tr>
<th>No. of CAG repeats</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 28</td>
<td>Normal range; individual will not develop HD</td>
</tr>
<tr>
<td>29-34</td>
<td>Individual will not develop HD but the next generation is at risk</td>
</tr>
<tr>
<td>35-39</td>
<td>Some, but not all, individuals in this range will develop HD; next generation is also at risk</td>
</tr>
<tr>
<td>≥ 40</td>
<td>Individual will develop HD</td>
</tr>
</tbody>
</table>