Evolution of Genome Size

1. The C-Value Paradox

The size of the genome (C-value) depends on the organism. It is essentially constant within species, but varies widely among species. There is not a strong correlation between organism complexity and genome size. This is known as the C-value paradox.

In general, bacterial genomes are smaller than eukaryotic genomes. Within bacteria, genomes range from 580 kb to 13 Mb, thus there is 20-30 fold size variation within prokaryotes

A few eukaryotic genomes fall in the size range of bacteria (e.g. yeast), but most are much larger. The size range of eukaryotic genomes is 8.8 Mb to ≈700 Gb. This is 80,000 fold size variation

Is there a correlation between genome size and gene number?

In bacteria, yes.

In eukaryotes, no. Of course, there is some correlation in the genomes that are sequenced (e.g., yeast < Drosophila < human), but the range of variation in eukaryotic gene number is estimated to be <50 fold. It may actually be much less. An example from well-annotated genomes:

Yeast, 15 Mb total, 6,000 genes
Human, 3,000 Mb (3 Gb) total, 25,000 genes

For DNA, the C ratio (human to yeast) = 3,000/15 = 200
For genes, the G ratio = 25,000/6,000 = 4.2

Variation in C is much greater than variation in G.
Conclusion: most of the C-value variation is due to the amount of non-coding DNA.

Heterochromatin = large regions of the genome with no (or very few) genes. It is difficult to clone and usually not sequenced in genome projects.

In general, there is a steep decline in the fraction of genic DNA (coding DNA) as genomes become larger.

Example: only around 2% of the human genome is protein-encoding sequence
Example: the Norway spruce (tree) has a genome of 20 Gb, but has about the same number of genes as *Arabidopsis thaliana* and human (around 20,000-30,000).

2. Repetitive DNA

a) Satellite DNA – first identified as distinct bands of DNA that are heavier or lighter than the majority of genomic DNA by density centrifugation. These are repeated sequences that have either high GC (heavy) or high AT (light) content. They are fairly short sequences (2-2000 bp) repeated 1000's of times in a row. They are found in heterochromatic regions and around centromeres.

b) Minisatellites – sequences of 9-100 bp repeated 10-100 times. Found in subtelomeric regions and (rarely) dispersed throughout chromosomes.

c) Microsatellites (SRS "short repetitive sequences", STR "short tandem repeats", SSR "simple sequence repeats") – very short sequences of 1-5 bp repeated 10-100 times. Found dispersed throughout chromosomes, often in and around genes.
For example, the dinucleotide repeat CA is very common in the human genome (≈50,000 copies)

Microsatellites have very high mutation rates (where a "mutation" means a change in repeat number). Thus they are often variable within a population and useful for population genetics. This property also makes them useful for "DNA fingerprinting".

The expansion of tri-nucleotide repeats (increase in repeat number) in or near genes is often associated with inherited diseases. Some examples include:
Fragile-X syndrome (CCG)
Huntington's disease (CAG)
Schizophrenia? (CAG)
Myotonic Dystrophy (CTG)
… plus many other neuro-muscular disorders

3. Transposable elements (TEs)
Also known as interspersed repetitive elements or "jumping genes"
TEs are pieces of DNA that can move within the genome and increase in number. About 50% of the human genome is made up of TEs and remnants of TEs.

There are two major types of TEs: transposons and retrotransposons, which are classified by their mechanism of transposition.

a) Conservative transposition – A TE moves from one place in the genome to another. This does not necessarily lead to an increase in copy number. Copy number can be increased through recombination between elements at different chromosomal locations. However this should lead to an equal number of gains and losses. ("cut-and-paste").
b) Replicative transposition – copy number is increased because the original element remains at donor site, while a new copy inserts into a new site. ("copy-and-paste").
c) Retrotransposition – the TE is transcribed into RNA, then reverse transcribed into cDNA, then inserts into new chromosomal location. The copy number increases. These elements are typically the most abundant in a genome. ("copy-and-paste").

Mechanisms A and B are used by transposons and involve only DNA, mechanism C is used by retrotransposons and requires an RNA intermediate.

Transposons – ≈2,500-7,000 bp long, DNA -> DNA
autonomous – have inverted repeats at ends, encode a single gene (transposase), can move by themselves.

non-autonomous – have inverted repeats at ends, but no transposase gene. Cannot move by themselves, but can move if there is another element in the genome producing transposase.

"Helper element" – Does not have inverted repeats, but does have transposase gene. Cannot move, but can cause non-autonomous elements to move. These are very useful for experiments in organisms like Drosophila.

For example, the transposase gene of a TE can be replaced with any gene, then a helper element can be used to make transposase and insert this gene into the Drosophila genome. Then the helper element is removed so the new gene becomes a stable part of the genome.

Retrotransposons
Active – have intact promoter, are transcribed, and can retrotranspose.

"Dead" or "Dead On Arrival (DOA)" – retroelements are often truncated at the 5' end when inserting into DNA. When this happens they lose their promoter and no longer can be transcribed or retrotransposed. They are thought to be "junk DNA" that is under no selective constraint and accumulate mutations at random.

4. Pseudogenes
Previously functional genes that have lost their function due to mutation (usually by a mutation that introduces a stop codon into the ORF or an insertion/deletion that disrupts the reading frame). In rare cases, genes may lose function due to parasitic or symbiotic relationship with their host. In these cases, the genes are not needed and can be lost through mutations (ex. pathogenic bacteria, *M. tuberculosis*). Most cases, however, involve some type of gene duplication.

unprocessed pseudogenes – often arise through tandem duplication, where an entire section of DNA is duplicated during replication, producing two copies of a gene. They are usually adjacent in the genome. If only one copy is required, the other copy may accumulate mutations and become a non-functional pseudogene

processed pseudogenes (or retrotransposed genes) – the mRNA of a nuclear gene is reverse transcribed into cDNA, then re-inserts into the genome. Most likely this uses the reverse transcriptase and integrase enzymes encoded by a retroelement.
Key features:
- Does not have introns present in the "parental" gene
- If recent, may have a poly(A) sequence at 3' end
- Usually lacks promoter sequences (thus "Dead on arrival" = not expressed)

Some genes appear to retrotranspose more than others.

Why?
a) Expression level – highly-expressed genes have more mRNA and thus have a greater chance of being reverse transcribed.

b) Gene size – short mRNAs may retrotranspose better than long mRNAs.

c) Sequence specific – the primary sequence of some genes may be better for retrotransposition.

5. The C-Value Paradox Revisited

Mutational and selective forces affect genome size. Why is there such great variation in genome size? There are two major classes of explanation:

a) adaptive – the non-coding DNA is functionally important to the organism.

b) junk DNA – most of the non-coding DNA serves no purpose. It may even be parasitic or "selfish DNA".

Can C-value variation be explained by differences in mutation rates? Specifically, by the rate of spontaneous DNA deletion?

The approach: *Laupala* (Hawaiian crickets) have a genome 11X that of Drosophila. The Drosophila genome is small and has almost no pseudogenes. Rates of DNA deletion can be estimated in both species by comparing sequences of DOA transposable elements, which are similar to pseudogenes.

The result: Spontaneous DNA loss is faster in Drosophila than in Laupala. This may explain why the Drosophila genome is small and has no pseudogenes – because pseudogenes are lost very rapidly by deletion mutations rendering them undetectable.
Does this result extend to other taxa with larger genome sizes?

The approach: Grasshoppers (genus *Podisma*) have even larger genomes (≈20 Gb) – over 10X greater than Laupaula and 100X greater than Drosophila. In Grasshoppers and many other species, there are many pseudogenes in the nuclear DNA that are derived from mitochondrial genes (NUMTs = Nuclear copies of mitochondrial genes = "new mites").

Why are NUMTs non-functional?
a) the genetic code is different between mitochondria and nucleus
b) they often lack a promoter
c) they do not have a signal sequence to target them to mitochondria

These are "Dead-on-arrival" and can be used to estimate mutation (and deletion) rates.

The result: Grasshoppers have a very low rate of DNA loss, lower than Laupaula and Drosophila. Thus the inverse correlation between genome size and rate of spontaneous DNA deletion holds for three insect groups with greatly different genome sizes.

Deletion rate: Dros > Lau > Pod

Deletion size: Dros > Lau > Pod

Genome size: Pod > Lau > Dros

References:

Petrov et al. (2000). Evidence for DNA loss as a determinant of genome size. Science 287: 1060-1062.

Bensasson et al. (2001). Genomic gigantism: DNA loss is slow in mountain grasshoppers. Mol Biol Evol. 18: 246-253.