Gene and Genome Duplication

1. Genome Duplication

Polyploidization – addition of 1 or more complete sets of chromosomes to the genome. This is more common in plants, but also occurs in animal species. The most common situation is a diploid species (2N) becoming tetraploid (4N), although it is possible to become hexaploid (6N), octoploid (8N) or more.

Types of polyploidy:

autopolyploidy – multiplication of 1 set of chromosomes (same species). Results from a lack of disjunction between all chromosomes during meiosis.

allopolyploidy – combination of 2 different sets of chromosomes (different species).
example: common wheat is an allohexaploid. It has 3 distinct sets of chromosomes derived from 3 different species of grass.

Genome sizes in grass show a multimodal distribution with peaks in multiples of genome size, suggesting that genome duplication is a major mechanism in the evolution of their genome size.

Polyploidization will initially duplicate the whole genome. However, over time the different genome copies will undergo mutations, translocations, chromosome rearrangements, deletions, *etc.* Eventually it will be very difficult to distinguish a polyploid from a diploid.

cryptopolyploidy – an ancient polyploid that is no longer recognizable. It looks like a diploid.

Examples:
Is yeast an ancient tetraploid? (Wolfe and Shields, 1997. Nature 387: 708-713)
The approach: Look for duplicated regions found spread throughout the yeast genome.
Criteria for identification:

1) Strong sequence similarity (BLAST E < 10^{-18})
2) At least 3 genes per region
3) Gene order is conserved

The result: 54 duplicated regions

Two possibilities:

a) Each region duplicated separately at different times.

b) There was one complete genome duplication, followed by gene loss and chromosome rearrangement. The authors favor this possibility. 50/54 regions maintained the same orientation with respect to the centromere. Statistically, independent duplications would be expected to produce 7 triplicated regions, but none were observed (Poisson). They propose a yeast genome duplication occurred ≈100 million years ago, and ≈92% of the duplicates have been lost by mutation/deletion.

Has the vertebrate genome undergone polyploidization?
The 2R hypothesis (2 Rounds of genome duplication)

The early observation: Hox genes = a group of genes arranged in order along the chromosome that is involved in the establishment of the body plan during early development. There is 1 cluster of Hox genes in Drosophila, 4 clusters in human. This could be explained by two separate genome duplication events. 1H -> 2H -> 4H.

Predictions of 2R hypothesis:

1) Many gene clusters should be found in 4 copies
2) Phylogenetic relationship should be "2 x 2"

Some genes follow this pattern, while others do not. So there has been some controversy about the 2R hypothesis. The latest results from genome sequencing support the 2R hypothesis. For example, the amphioxis genome (Nature vol. 453, pages 1064-1071, 2008).

2. Chromosome Duplication

Aneuploidy – increase in chromosome number, but not by an integral of the typical haploid set. ex: not N=6, 12, 18, 24 …, but N=6, N=7, or N=8 …

Polysomy – duplication of 1 complete chromosome.

Polysomy is almost always deleterious. Example: Down's syndrome in human (trisomy 21) = 3 copies of chromosome 21.

Polysomy is not thought to play a major role in genome size evolution.

Why is duplicating one chromosome more deleterious that duplicating all of them?
This is probably due to an imbalance in gene expression.
In polysomy, expression of genes on one chromosome is increased, thus these genes are overexpressed relative to other genes in the genome. In polyploidy, all genes are increased in number, so their expression levels remain relatively constant.

3. Gene Duplication

Where do new genes come from? In 1970, Ohno suggested that most evolution occurs by gene duplication. The duplicate copy provides a template for the evolution of new gene function.

Abundance of duplicate genes (Lynch and Conery, 2000. Science 290:1151-1155).
The approach: search complete genomes for pairs of duplicates genes.
Duplicates = pairs of genes (protein-encoding) in same genome with BLAST score < 10^{-10}.
They must have intact ORF (to eliminate suspected pseudogenes). Transposable elements and multigene families (genes with >5 matches) were also eliminated. This resulted in a conservative set of duplicates.

Number of duplicate pairs found:
Human = 336
Mouse = 225
Drosophlia = 462
Arabidopsis = 2,671
C. elegans = 1,933
Yeast = 326

Age estimation
The age of each duplication can be estimated from the number of differences between the two gene copies at "silent" sites = sites that do not change amino acid sequence (i.e. synonymous sites). Initially the two copies are identical, but over time they accumulate mutations. More sequence differences = more time since duplication. The assumption is that there is little or no selection on silent sites, so they change at a relatively constant rate over time.

The duplicate genes appear to be under selective constraint (functional), because in general they have fewer amino acid replacement changes than silent changes. That is, most duplicates have Ka/Ks < 1, suggesting they are maintained by selection and are not pseudogenes.

However, a more conservative criterion should be used when comparing duplicates within a genome. If we assume that at least one copy is essential and is maintained by selection, then in the extreme case Ka/Ks = 0 on one (functional) lineage after duplication and Ka/Ks = 1 on the other (nonfunctional) lineage. Thus, Ka/Ks between the two copies would be 0.5. To infer that both copies are maintained by selection, one would need to observe Ka/Ks < 0.5.

4. Fate of Duplicate Genes

a) Degeneration into pseudogene (pseudogenization) – one copy is “knocked out” by mutation, eventually degenerates and is lost. This is the most common fate.

b) Maintenance of two redundant genes – this may be a short-term adaptation to increase gene activity (expression), but is unlikely to be maintained over long evolutionary timescales.

Examples:
Bacteria (Riehle, Bennett, and Long, 2001. PNAS USA 98:525-530).
E. coli cells were adapted to grow at 37 C for 2,000 generations, then independent lines were adapted to 41.5 C for 2000 generations.

The genomes of the heat-adapted lines were compared to the original 37 C lines by array hybridization of DNA sequences. This detects changes in gene copy number (deletions and duplications). Do particular genes increase in copy number during heat adaptation?

Three independent duplications occurred in the same region. This region contains candidate genes involved in stress and starvation resistance.

Drosophila Metallothionine (MT) genes?
MTs are small, metal-binding proteins (Cu, Zn, Ca) that regulate metals in the cell and also are involved in metal detoxification. Some *D. melanogaster* individuals have a recent duplication of the MT gene, *Mtn*. These flies produce more *Mtn* mRNA and are more tolerant to heavy metals in laboratory experiments. It has been suggested that flies with the duplication are selectively favored in areas where there is heavy metal pollution, although this is difficult to demonstrate in nature.

c) Adoption of new function (neofunctionalization) – one copy accumulates random mutations until it gains a new function by chance. If the new function is advantageous, the new gene will be favored by selection and will be maintained. It can then further adapt to its new function. This could be a new protein function, new expression pattern, or both.
Examples:

Sdic = Sperm-specific Dynein Intermediate Chain
Arose as a duplication of the Cdic gene (Cytoplasmic Dynein Intermediate Chain)
Sdic is only present in one Drosophila species, *D. melanogaster*, so it is a very recent duplication (< 2.5 million years old). It arose through a complex series of duplications and fusion of two genes.

After duplication, Sdic gained a new, sperm-specific promoter and appears to be maintained by selection for a new function in sperm.

Janus-Ocnus
Three Drosophila male reproductive genes, *janusA*, *janusB*, and *ocnus* arose through two separate gene duplication events. Following duplication, each gene evolved at a different rate (as measured by Ka/Ks across multiple Drosophila species), suggesting that each gene is under different selective constraints and that the three genes have diverged in function.

d) Subfunctionalization – if the original gene has two different functions, then after duplication each copy may lose one of the functions by mutation. Afterwards, both copies must be maintained. No new function is gained, however the two copies may become more specialized. The subfunctions could be at the protein or gene expression level.

5. Does duplication make a gene less essential?

It is usually observed that the knock-out of a single copy gene is more likely to have a major effect on phenotype (or be lethal) than the knock-out of a gene that has a close paralog in the genome. This is thought to be explained by an overlap of function between the paralogs, such that one copy can mask the effect of knocking out the other copy. However, a study in the worm, *C. elegans*, found that phenotypic masking between paralogs is rare. In most cases, both paralogs can be knocked-out together and there is still no phenotypic effect. Thus, it appears that non-essential genes are more likely to undergo duplication.

Reference:

Woods et al. (2013) Duplication and Retention Biases of Essential and Non-Essential Genes Revealed by Systematic Knockdown Analyses. PLoS Genet 9(5): e1003330.