Gene Duplication, exon Shuffling, (and Concerted Evolution)

Lecture 6
Segmental & Contextual Changes
Point substitution
Reciprocal Recombination
Gene Conversion
Tandem duplication
Deletion
Conservative transposition
Duplicative transposition
Horizontal gene transfer
“A redundant duplicate of a gene may acquire divergent mutations and eventually emerge as a new gene.”

J. B. S. Haldane (1932)
The first duplication mutation described in the literature:

<table>
<thead>
<tr>
<th>Chromosome Structure</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>X–16A</td>
<td>Normal eye female (800 facets)</td>
</tr>
<tr>
<td>X–16A</td>
<td></td>
</tr>
<tr>
<td>X–16A–16A</td>
<td>Bar eye female (350 facets)</td>
</tr>
<tr>
<td>X–16A</td>
<td></td>
</tr>
<tr>
<td>X–16A–16A</td>
<td>Bar eye female (68–70 facets)</td>
</tr>
<tr>
<td>X–16A–16A</td>
<td></td>
</tr>
<tr>
<td>X–16A–16A–16A</td>
<td>Double bar eye female (45 facets)</td>
</tr>
<tr>
<td>X–16A</td>
<td></td>
</tr>
<tr>
<td>X–16A–16A–16A</td>
<td>Double bar eye female (25 facets)</td>
</tr>
</tbody>
</table>
Extent of DNA duplication

1. internal gene duplication
2. complete gene duplication
3. partial chromosomal duplication (partial polysomy)
4. complete chromosomal duplication (polysomy)
5. genome duplication (polyploidy)
<table>
<thead>
<tr>
<th>Extent of DNA Duplication</th>
<th>Mutational Occurrence</th>
<th>Effects on Fitness</th>
</tr>
</thead>
<tbody>
<tr>
<td>partial gene</td>
<td>very frequent</td>
<td>Deleterious if it affects reading frame.</td>
</tr>
<tr>
<td>whole gene</td>
<td>frequent</td>
<td>Deleterious only in organisms in which the genome is replicated as unit.</td>
</tr>
<tr>
<td>partial polysomy</td>
<td>rare</td>
<td>Almost invariably deleterious.</td>
</tr>
<tr>
<td>polysomy</td>
<td>common</td>
<td>Almost invariably deleterious.</td>
</tr>
<tr>
<td>polyploidy</td>
<td>common</td>
<td>Deleteriousness determined by the mode of reproduction and sex-determination.</td>
</tr>
</tbody>
</table>
Complete chromosomal duplication $\downarrow$ deleterious
e.g., trisomy 13
Partial chromosomal duplication
↓
deleterious

Partial Trisomy 16q+
A partial duplication of 16q from band q13 to q24
Equal & Unequal Crossing Over

Equal

Unequal
The main mechanism responsible for gene duplication is unequal crossing-over.
## Possibilities for unequal crossing over with four tandem genes

<table>
<thead>
<tr>
<th>Type of overlap</th>
<th>Probability of occurring</th>
<th>Products</th>
</tr>
</thead>
<tbody>
<tr>
<td>X X X X</td>
<td>1/16</td>
<td>Singlet and Septuplet</td>
</tr>
<tr>
<td>X X X X</td>
<td>2/16</td>
<td>Doublet and Sextuplet</td>
</tr>
<tr>
<td>X X X X</td>
<td>3/16</td>
<td>Triplet and Quintuplet</td>
</tr>
<tr>
<td>X X X X</td>
<td>4/16</td>
<td>2 Quadruplets</td>
</tr>
<tr>
<td>X X X X</td>
<td>3/16</td>
<td>Quintuplet and Triplet</td>
</tr>
<tr>
<td>X X X X</td>
<td>2/16</td>
<td>Sextuplet and Doublet</td>
</tr>
<tr>
<td>X X X X</td>
<td>2/16</td>
<td>Septuplet and Singlet</td>
</tr>
<tr>
<td>X X X X</td>
<td>1/16</td>
<td>Singlet and Septuplet</td>
</tr>
<tr>
<td>X X X X</td>
<td>1/16</td>
<td>Septuplet and Singlet</td>
</tr>
</tbody>
</table>
DOMAINS

A well defined region within a protein that performs a specific function: FUNCTIONAL DOMAIN.

A well defined region within a protein that constitutes a stable, compact structural unit that can be distinguished from all other such units: STRUCTURAL DOMAIN or MODULE.
DOMAINS

A well defined region within a protein that performs a specific function: FUNCTIONAL DOMAIN.

Defining the boundaries of a functional domain is often difficult because functionality is in many cases conferred by amino-acid residues that are scattered throughout the polypeptide.
Base pair scale

DNA

Precursor RNA

RNA processing (nucleus)

Mature messenger RNA

Translation into protein (cytoplasm)

β-globin amino acids

Numerical position on polypeptide

DNA source
Structural modules are collinear with the amino-acid sequence of a protein, i.e., a module consists of a continuous stretch of amino acids.

A well defined region within a protein that constitutes a stable, compact structural unit that can be distinguished from all other such units: STRUCTURAL DOMAIN or MODULE.
If functionality is conferred by a **module**, a duplication will increase the number of functional segments.

If functionality is conferred by **amino-acid residues scattered among different modules**, a duplication may not be functionally desirable.
Module definition and identification:

Mitiko Go
A + sign is entered in the matrix if the distance between two corresponding residues is greater than a preset value. For globular proteins, the value is usually $R = \text{the radius of the sphere containing the protein.}$
Modules are encoded by exons

Intron sliding

Intron loss

Intron acquisition

Modules are not encoded by exons
Where did the missing intron go?
A central intron was found in...

Legumes  Nematodes  Chironomids
A panoply of exon/intron permutations
internal gene duplication

\[\downarrow\]

gene elongation
PATHWAYS FOR GENE ELONGATION

1. Obliteration of stop codon.
2. Substitution in splicing site.
3. Insertion.
4. Random internal duplication.
5. Internal duplication of asymmetrical exon.
6. Internal duplication of symmetrical exon.
internal gene duplication
↓
whole-exon duplication

internal gene duplication
↓
↓
new functions
Evolution of a dinucleotide-binding domain in glyceraldehyde-3-phosphate dehydrogenase
Internal duplication?
internal gene duplication

↓

repetitive structure
PROPEPTIDES  |  COLLAGEN MOLECULE (1050 amino acids)  |  PROPEPTIDES
N-TERMINAL    |                                 |  C-TERMINAL

TRIPLE-HELICAL DOMAIN (basic repeat = gly-X-Y)
54-bp exon

Intragenic Duplications

If it involves active site
\[\downarrow\]
  enhancement
  or
  refinement of activity

If it does not involves active site
\[\downarrow\]
  confer added spatial stability
  or
  prolong longevity of product
Duplication affects composition:

PEST proteins, i.e., proteins rich in proline (P), glutamic acid (E), serine (S), and threonine (T), are degraded rapidly inside eukaryotic cells. Duplicated PEST domains are found in proteins, such as nuclear factors that have a transient function, thus ensuring a very rapid degradation.
Haptoglobin alleles

Haptoglobin is a \textit{tetrameric protein} made out of \textbf{two $\alpha$ and two $\beta$ chains}.

Both chains are \textbf{produced by the same gene} as a single polypeptide, which is then \textit{cleaved at an arginine residue} to generate the $\alpha$ and $\beta$ subunits.

Haptoglobin is found in the \textbf{blood serum}, and it functions as a \textit{transport glycoprotein} which \textbf{removes free heme} from the circulation of vertebrates.
Haptoglobin alleles

In humans, haptoglobin α is polymorphic due to the existence of four alleles: slow α1 (α1S), fast α1 (α1F), α2 and α3.

α1S and α1F differ by having at position 54 glutamic acid and lysine, respectively.
Haptoglobin alleles

The $\alpha^2$ allele was created by an unequal crossing-over within different introns of $\alpha^1S$ and $\alpha^1F$ within a heterozygous individual.

The internal duplication of about 1.7 kb, of which 177 bp are exonic, increased the length of the polypeptide from 84 to 143 amino acids.
Haptoglobin alleles

α2 has an increased stability in complex with heme, and is more efficient at rendering the heme group susceptible to degradation.

The α2 allele is of recent origin, more recent than the human-chimpanzee split (~5 MYA), but has a high frequency (30-70%) in Europe and in parts of Asia. This rapid increase in frequency implies that individuals carrying the α2 allele enjoy a selective advantage over carriers of the α1 allele.

In the future, it is likely that the α2 allele will become fixed in the human population at the expense of α1.
Haptoglobin alleles

α3 (or haptoglobin Johnson) contains a three-fold tandem repeat of the same 1.7 kb segment implicated in the α2 allele duplication.

Is α3 better than α2?
Origin of an antifreeze glycoprotein gene:
Origin of an antifreeze glycoprotein gene:

The body fluids of most teleosts (ray-finned fish) freeze at $-1.0^\circ\text{C}$ to $-0.7^\circ\text{C}$. Therefore, most fish cannot survive the temperatures of the Antarctic Ocean ($-1.9^\circ\text{C}$).

Freezing resistance in Antarctic fish is due to the existence in the blood of a protein that lowers the freezing temperature by inhibiting the growth of small ice crystals, which otherwise may break the cell membranes.
Origin of an antifreeze glycoprotein gene:
The giant Antarctic toothfish (*Dissostichus mawsoni*)
Proteins with internal domain duplications taking up 50% or more of the total length of the protein

<table>
<thead>
<tr>
<th>Sequence (organism)</th>
<th>Length protein&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Length repeat&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Number repeats&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Percent repetition&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>α1β-glycoprotein (human)</td>
<td>474</td>
<td>91</td>
<td>5</td>
<td>96</td>
</tr>
<tr>
<td>Angiotensin I-converting enzyme (human)</td>
<td>1306</td>
<td>357</td>
<td>2</td>
<td>55</td>
</tr>
<tr>
<td>Calbindin (human, bovine)</td>
<td>260</td>
<td>43</td>
<td>6</td>
<td>99</td>
</tr>
<tr>
<td>Calcium-dependent regulator protein (human)</td>
<td>148</td>
<td>74</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td>Ferredoxin (&lt;i&gt;Azobacter vineandii&lt;/i&gt;)</td>
<td>70</td>
<td>30</td>
<td>2</td>
<td>86</td>
</tr>
<tr>
<td>Ferredoxin (&lt;i&gt;Azobacter pasteurianum&lt;/i&gt;)</td>
<td>55</td>
<td>28</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td>Homopexin (human)</td>
<td>439</td>
<td>207</td>
<td>2</td>
<td>94</td>
</tr>
<tr>
<td>Hexokinase (human)</td>
<td>917</td>
<td>447</td>
<td>2</td>
<td>97</td>
</tr>
<tr>
<td>Immunoglobulin γ chain C region (human)</td>
<td>329</td>
<td>108</td>
<td>3</td>
<td>98</td>
</tr>
<tr>
<td>Immunoglobulin ε chain C region (human)</td>
<td>423</td>
<td>108</td>
<td>4</td>
<td>100</td>
</tr>
<tr>
<td>Interleukin-2 receptor (human)</td>
<td>251</td>
<td>68</td>
<td>2</td>
<td>54</td>
</tr>
<tr>
<td>Interstitial retinol-binding protein (bovine)</td>
<td>1263</td>
<td>302</td>
<td>4</td>
<td>96</td>
</tr>
<tr>
<td>Lactase-phlorizin hydrolase (human)</td>
<td>1927</td>
<td>480</td>
<td>3</td>
<td>79</td>
</tr>
<tr>
<td>Lymphocyte-activation gene-3 protein (human)</td>
<td>470</td>
<td>138</td>
<td>2</td>
<td>59</td>
</tr>
<tr>
<td>Multidrug resistance-1 P-glycoprotein (human)</td>
<td>1280</td>
<td>609</td>
<td>2</td>
<td>95</td>
</tr>
<tr>
<td>Ovo inhibitor (chicken)</td>
<td>472</td>
<td>64</td>
<td>7</td>
<td>95</td>
</tr>
<tr>
<td>Parvalbumin (human)</td>
<td>108</td>
<td>39</td>
<td>2</td>
<td>72</td>
</tr>
<tr>
<td>Plasminogen (human)</td>
<td>790</td>
<td>79</td>
<td>5</td>
<td>50</td>
</tr>
<tr>
<td>Preproglucagon (rat)</td>
<td>180</td>
<td>36</td>
<td>3</td>
<td>60</td>
</tr>
<tr>
<td>Pre-pro-von Willebrand factor (human)&lt;sup&gt;d&lt;/sup&gt;</td>
<td>3817</td>
<td>586</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Protease inhibitor, Bowman-Birk type (soybean)</td>
<td>71</td>
<td>28</td>
<td>2</td>
<td>79</td>
</tr>
<tr>
<td>Protease inhibitor, submandibular-gland type (rat)</td>
<td>115</td>
<td>54</td>
<td>2</td>
<td>94</td>
</tr>
<tr>
<td>Ribonuclease/angiogenin inhibitor (human)</td>
<td>461</td>
<td>57</td>
<td>8</td>
<td>99</td>
</tr>
<tr>
<td>Serum albumin (human)</td>
<td>584</td>
<td>195</td>
<td>3</td>
<td>100</td>
</tr>
<tr>
<td>Tropomyosin α chain (human)</td>
<td>284</td>
<td>42</td>
<td>7</td>
<td>100</td>
</tr>
<tr>
<td>Twitchin (&lt;i&gt;Caenorhabditis elegans&lt;/i&gt;)&lt;sup&gt;d&lt;/sup&gt;</td>
<td>6049</td>
<td>100</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>Villin (human)</td>
<td>826</td>
<td>360</td>
<td>2</td>
<td>87</td>
</tr>
<tr>
<td>Vitamin D-dependent calcium-binding protein (bovine)</td>
<td>260</td>
<td>54</td>
<td>3</td>
<td>63</td>
</tr>
</tbody>
</table>

<sup>a</sup>Number of amino acid residues. <sup>b</sup>Some of the repeats may be truncated in comparison with the repeated consensus unit. <sup>c</sup>Percent of the total length of the protein occupied by repeated sequences. <sup>d</sup>In these protein several unrelated types of repeats are present.
Post-duplication evolutionary processes
Emergence of a new function:  

In theory, three pathways can lead to the creation of a new function:

1. *de novo* appearance from nonfunctional sequence due to accumulation of mutations.
2. replacement due to change of one function into another.
3. creation of a novel function from a redundant copy of an old function following duplication.
In theory, three pathways can lead to the creation of a new function:

1. *de novo* appearance from nonfunctional sequence due to accumulation of mutations.

2. Replacement due to change of one function into another.

3. Creation of a novel function from a redundant copy of an old function following duplication.
complete gene duplication

↓

gene families
Types of Molecular Homology:

ORTHOLOGY: due to speciation
PARALOGY: due to gene duplication
XENOLOGY: due to horizontal gene
Following gene duplication three things may happen to the copies:

1. All copies retain the same function.
2. Some copies die.
3. Some copies evolve into new functions.
Repeated genes

Variant repeats

Dose-independent repeats

Invariant repeats

Dosage repeats
Prevalence of gene duplication

Gene duplications arise spontaneously at high rates in bacteria, bacteriophages, insects and mammals, and are generally viable.

Mutation is NOT the rate-limiting step in the evolutionary process of gene duplication.

Only a small fraction of all duplicated genes are retained, and an even smaller fraction evolves new functions.

The probability of nonfunctionalization is much higher than that of evolving a new function.
An increase in gene number can occur quite rapidly under selection pressure for increased amounts of a gene product.
Example:

The genome of the wild-type strain of the peach-potato aphid (*Myzus persicae*) contains two genes encoding esterases E4 and FE4. The two genes are very similar in sequence (98%), indicating that they have been duplicated recently.
Following exposure to organophosphorous insecticides, which can be hydrolized and sequestered by esterase, resistant strains of *Myzus persicae* were found to contain multiple copies of E4 and FE4 (up to 80 copies). The increase in the frequency of the carriers of these duplications is likely to have occurred within the last 50 years, with the introduction of the selective agent.

\[
\text{malathion}
\]
RIBOSOMAL TRANSCRIPTION UNITS

Human ≈ 13.0 kb

X. laevis ≈ 7.9 kb

D. melanogaster ≈ 7.7 kb

S. cerevisiae ≈ 6.6 kb

~300

~760

~240

~140

TANDEM ARRAY

Transcription unit
The diagram shows the number of unique species for Bacteria and Archaea across different 16S rRNA gene copy number categories. Bacteria have a significantly higher number of unique species compared to Archaea.
\[ R^2 = 0.80 \]
\[ P < 0.001 \]
Numbers of rRNA and tRNA genes per haploid genome in various organisms

<table>
<thead>
<tr>
<th>Genome Source</th>
<th>Number of rRNA sets</th>
<th>Number of tRNA genes</th>
<th>Approximate genome size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human mitochondrion</td>
<td>1</td>
<td>22</td>
<td>$2 \times 10^4$</td>
</tr>
<tr>
<td><em>Nicotiana tabacum</em> chloroplast</td>
<td>2</td>
<td>37</td>
<td>$2 \times 10^5$</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>7</td>
<td>~ 100</td>
<td>$4 \times 10^6$</td>
</tr>
<tr>
<td><em>Neurospora crassa</em></td>
<td>~ 100</td>
<td>~ 2,600</td>
<td>$2 \times 10^7$</td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td>~ 140</td>
<td>~ 360</td>
<td>$5 \times 10^7$</td>
</tr>
<tr>
<td><em>Caenorhabditis elegans</em></td>
<td>~ 55</td>
<td>~ 300</td>
<td>$8 \times 10^7$</td>
</tr>
<tr>
<td><em>Tetrahymena thermophila</em></td>
<td>1</td>
<td>~ 800</td>
<td>$2 \times 10^8$</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>120-240</td>
<td>590-900</td>
<td>$2 \times 10^8$</td>
</tr>
<tr>
<td><em>Physarum polycephalum</em></td>
<td>80-280</td>
<td>~ 1,050</td>
<td>$5 \times 10^8$</td>
</tr>
<tr>
<td><em>Euglena gracilis</em></td>
<td>800-1,000</td>
<td>~ 740 $2 \times 10^9$</td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td>~ 300</td>
<td>~ 1,300</td>
<td>$3 \times 10^9$</td>
</tr>
<tr>
<td><em>Rattus norvegicus</em></td>
<td>150-170</td>
<td>~ 6,500</td>
<td>$3 \times 10^9$</td>
</tr>
<tr>
<td><em>Xenopus laevis</em></td>
<td>500-760</td>
<td>6,500-7,800</td>
<td>$8 \times 10^9$</td>
</tr>
</tbody>
</table>
Following gene duplication three things may happen to the copies:

1. All copies retain the same function.

2. Some copies die.

3. Some copies evolve into new functions.
gene death (nonfunctionalization) ↓ pseudogene
<table>
<thead>
<tr>
<th>Pseudogene</th>
<th>TATA box</th>
<th>Init. codon</th>
<th>Frame shift</th>
<th>Premature stop</th>
<th>Essential amino acid</th>
<th>Splice GT/AG</th>
<th>Stop codon</th>
<th>AATAAA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human ψα1</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Human ψξ1</td>
<td></td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouse ψα3</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouse ψα4</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouse βh3</td>
<td>?</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Goat ψβ^X</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Goat ψβ^Z</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Rabbit ψβ2</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
<td>+</td>
</tr>
</tbody>
</table>
GENE LOSS

There are \(~7,000\) genetic diseases that have been documented in the literature attesting to the fact that mutations can easily destroy the function of a protein-coding gene.

The vast majority of such mutations are deleterious, and are either eliminated quickly from the population or maintained at very low frequencies due to overdominant selection or genetic drift.

A few such mutations may be neutral or even advantageous and they may be fixed in the population.
“As long as there are other copies of a gene that function normally, a duplicate gene may accumulate deleterious mutations and become nonfunctional without adversely affecting the fitness of the organism.”

J. B. S. Haldane (1933)
“Because deleterious mutations occur far more often than advantageous ones, a redundant duplicate gene is more likely to become nonfunctional than to evolve into a new gene.”

Susumu Ohno (1972)
*Dead genes (as opposed to dead “anything else”) can reproduce and multiply.
Unprocessed Pseudogenes:

unequal crossing-over

gene death
Unprocessed Pseudogenes:

Chromosome 16
- ζ
- ψζ
- ψα2
- ψα1
- α2
- α1
- θ1

Chromosome 17
- ε
- Gγ
- Λγ
- ψη
- δ
- β
“By extrapolation [from chromosomes 21 and 22], we predict that there could be up to 20,000 pseudogenes in the whole human genome...”


Currently in the annotated pseudogene database: 11,072 processed, 2,282 duplicated, 6 unitary, 9291 ambiguous. Because of rapid decay, these numbers underestimate the true extent of pseudogenes in the genome.
Nonfunctionalization time

Species 1

Species 2
Let $l_i$, $m_i$, and $n_i$ be the numbers of nucleotide substitutions per site at codon position $i$ between points O and $\psi A$, O and A, and O and B, respectively.
\[ d_{(\psi AA)i} = l_i + m_i \]

\[ l_i = \frac{d_{(\psi AA)i} + d_{(\psi AB)i} - d_{(AB)i}}{2} \]

\[ m_i = \frac{d_{(\psi AA)i} - d_{(\psi AB)i} + d_{(AB)i}}{2} \]
\[
\begin{align*}
I_i &= \frac{d_{(\psi AA)i} + d_{(\psi AB)i} - d_{(AB)i}}{2} \\
m_i &= \frac{d_{(\psi AA)i} - d_{(\psi AB)i} + d_{(AB)i}}{2} \\
n_i &= \frac{-d_{(\psi AA)i} + d_{(\psi AB)i} + d_{(AB)i}}{2}
\end{align*}
\]
Assumptions: The rate of nucleotide substitution differs among codon positions in the functional genes. The rate of nucleotide substitution is the same for all the three codon positions in pseudogenes.
Solving $T_N$ is based on the assumption that the rate of nucleotide substitution differs among codon positions in the functional genes & that the rate of nucleotide substitution is the same for all the three codon positions in pseudogenes.
Species 1 = human
Species 2 = mouse
\[ T = 80 \text{ million years ago} \]
\[ \downarrow \]
\[ T_D = 49 \text{ million years ago} \]
\[ T_N = 45 \text{ million years ago} \]
gene death (nonfunctionalization) ↓ pseudogene

organ death ↓ vestigial organ
Following gene duplication three things may happen to the copies:

1. All copies retain the same function.

2. Some copies die.

3. Some copies evolve into new functions.
<table>
<thead>
<tr>
<th>X chromosome</th>
<th>Autosome</th>
<th>Vision</th>
</tr>
</thead>
<tbody>
<tr>
<td>♀ Human</td>
<td></td>
<td>Trichromatic</td>
</tr>
<tr>
<td>♂ Human</td>
<td></td>
<td>Trichromatic</td>
</tr>
<tr>
<td>♂ Human (color blind)</td>
<td></td>
<td>Dichromatic protanopia</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dichromatic deuteranopia</td>
</tr>
<tr>
<td>♀ NWM homozygote</td>
<td></td>
<td>Dichromatic</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>♂ NWM heterozygote</td>
<td></td>
<td>Trichromatic</td>
</tr>
<tr>
<td>♂ NWM</td>
<td></td>
<td>Dichromatic</td>
</tr>
</tbody>
</table>
New-World monkeys possess only two opsin loci, one autosomal and one X-linked. However, the X-linked opsin locus is highly polymorphic. Two of these alleles have maximal-sensitivity peaks similar to those of human red and green opsin, while the third allele has an intermediate peak. A heterozygous female will be is trichromatic, while males and homozygous females are dichromatic.
The **orthodox model** for the evolution of functionally novel proteins

(Ohno 1970 & Kimura 1983)

After gene duplication, one copy is rendered redundant and can accumulate substitutions at random.

By chance, some of these substitutions may result in a new function.
Difficulties with the orthodox model:

1. Unless the new function can be acquired through one or a few nucleotide substitutions, it is more than likely that the copy will become a pseudogene rather than a new functional gene.
Difficulties with the orthodox model:

2. Evidence from tetraploid organisms, such as *Xenopus laevis* and *Cyprinus carpio*, indicate that after gene duplication, both copies continue to be subjected to purifying selection (as inferred from the ratio of synonymous to nonsynonymous substitutions) as intense as before the duplication.
Difficulties with the orthodox model:

3. For a number of divergent multigene families, there is evidence that functionally distinct proteins have arisen not as a result of chance fixation of neutral variants, but as a result of positive Darwinian selection (as inferred from the ratio of nonsynonymous to synonymous substitutions > 1).

1. A protein-coding gene evolves into a gene that encodes a multifunctional (or *moonlighting*) protein.

2. Gene duplication occurs.

3. Each copy evolves toward increased specialization for one of the functions of the ancestral gene.
There is no novelty at the molecular level. New genes originate from old genes.

New functions are produced through modifications of preexisting functions.

Evolution != Revolution
Creation = Recombination
Dating gene-duplication events through phylogenetic distribution of duplicated genes in conjunction with paleontological data pertinent to the divergence date of the species in question.

Example: All vertebrates with the exception of hagfishes and lampreys encode α- and β-globin chains.
Following gene duplication three things may happen to the copies:

1. All copies retain the same function.
2. Some copies die.
3. Some copies evolve into new functions.

The three processes are independent of one another.
Chromosome 16

Chromosome 11

Chromosome 22

Chromosome 14

Chromosome 17

Myoglobin

Neuroglobin

Cytoglobin
jawless fish (lamprey)
Xenopus laevis has $\alpha$ and $\beta$ on the same chromosome
Homo sapiens has $\alpha$ and $\beta$ on different chromosome

\[\downarrow\]

Chromosomal separation occurred after gene duplication
transposition
(next chapter)
Ancestral \( \beta \)-globin gene

\[ \bullet = \text{Gene duplication} \]

\[ \triangledown = \text{Speciation} \]

Aves Monotremes Marsupials Eutherians
We’ll be discussed in the context of Genome Evolution
Exon Shuffling
Mosaic (or chimeric) protein = a protein encoded by a gene that contains regions also found in other genes. The existence of such proteins is evidence for exon shuffling.
exon shuffling
↓
mosaic proteins
3 types of exon shuffling

exon duplication = the duplication of one or more exons within a gene (internal duplication)

exon insertion = exchange of domains between genes or insertions into a gene

exon deletion = the removal of a segment from a gene.
Phases of introns and classes of exons

[Diagram showing various phases of introns and exons with labels and symbols.]

\[\text{\textcolor{black}{\square}} = \text{symmetrical exon}\]
Consequences of exon duplication & deletion

(a) Asymmetrical exon

1

2 / 3

1

Duplication

Deletion

(b) Symmetrical exon

3

1 / 3

1

Duplication

No frameshift

Deletion
Consequences of exon insertion into introns

(a) 

(b) 

(c) 

(d)
Exonization & Pseudoexonization

**Exonization** is the process through which an intronic sequence becomes an exon.

**Pseudoexonization** is the process through which an exonic sequence becomes an intron.
Pseudoexon
For exonization and pseudoexonization to occur without disruption of the reading frame, the rules pertaining to exon insertion and exon deletion must be respected.
“introns-early” versus “introns-late”
The introns-early hypothesis (the exon theory of genes)

**Ancient genes** possessed *self-splicing* introns.

Most self-splicing introns were lost in **Bacteria** and **Archaea**.

In **Eucarya**, self-splicing introns evolved into spliceosomal introns.
The introns-late hypothesis

Ancient genes had no introns.

The addition of introns occurred after the emergence of the eukaryotic cell and the endosymbiotic process that gave rise to the mitochondria.
The **introns-late hypothesis**

Nuclear **spliceosomal introns** were derived from **group II self-splicing introns**, which in turn were derived from **transposable elements** (e.g., retrons).
Molecular mechanisms for both intron gain and intron loss are known.
The introns-early hypothesis is supported by the correspondence between exons and modules (concordant positions).

However, a huge number of introns are found in positions not corresponding with the borders of modules (discordant positions).
However, not all introns in discordant positions can be taken as evidence against the introns-early hypothesis.
Intron gain and loss rates. Node sizes are proportional to their (known or inferred) intron density, and the branches are color-coded: (green) predominant intron gain; (red) predominant intron loss; (blue) balanced gain and loss; (brown) extensive (significantly greater than the mean over the tree) gains and losses.

Alternative pathways for producing new functions
1. Overlapping genes
2. Alternative *cis* splicing
3. Alternative *trans* splicing
4. Intron encoded proteins (*cis*)
5. Nested genes (*trans*)
6. Functional convergence
7. RNA editing
8. Gene sharing