FUNCTIONAL DIVERGENCE TOPIC 4:  *Evolution of new genes and new functions*

In topic 3 we considered cases where adaptive alterations of protein function were *modulating its existing function* in such a way as to confer fitness benefits on the individual. This does not explain how “genic diversity” has evolved; i.e., *how do new genes, with completely novel functions at the molecular level, evolve?* As before we will only concern ourselves with protein molecules at this time.

Consider the complexity of developmental processes (e.g., organ and tissue development, etc.), the interactions in metabolic pathways (e.g., the Pantothenate biosynthesis pathway of *E. coli* has 10 steps, with each reaction catalysed by one or more enzymes) or the regulation of physiological processes (organismal and cellular). Although we can not rule out *de novo* “invention” of the molecular genetic basis of these features of life, we can understand that evolution of novelty in such systems is much “easier” if pre-existing genetic systems can be used as a framework for evolving new functions. Whether a “new protein function” or a “new metabolic pathway”, it must be easier to modify an existing protein or pathway, as compared with building a completely new system from scratch. We refer to this evolutionary mechanism for novelty as GENETIC CO-OPTION; more formally, it is defined as the tailoring of pre-existing genetic systems to new uses by natural selection.

Genetic co-option can occur by altering the function of a protein encoded by a gene, altering the regulatory pattern of a protein, or both. A major role for genetic co-option has long been assumed for the evolution of developmental novelty; this notion has been supported by recent work on certain cases.

**Mixing and matching, the power of combinatorial evolution:**

*Exon shuffling:*  
As you know, eukaryotic protein-coding genes are comprised of exons and introns. Recognition that exons structure can correspond to the functional and structural domains of a protein led to the hypothesis that *non-homologous recombination of exons from different genes could lead to the evolution of a new gene with new functions*. There is some controversy as to the frequency of this mechanism for evolutionary novelty. Work on mammalian genes suggests that hundreds of families of mammalian proteins may have arisen this way. Although the relative importance of exon shuffling in all groups is subject to debate, it seems clear that it has been important in certain cases. Proteins that result from exon shuffling are sometime called MOSAIC PROTEINS or HYBRID PROTEINS. The LDL receptor gene is a classic example of evolution by exon shuffling, as it originated...
by the mixing and matching of exons from two other genes, C9 compliment and EGF precursor (See figure).

Protein modularity:
Protein modularity describes the organization of the exon structure of eukaryotic genes into modules, or units, which correspond to structural or functional domains of the 3D protein product of the gene. A well known example of such modularity is the central exon of both myoglobin and haemoglobin, which encode the heme-binding region of the mature protein. The figure below illustrates the relationship between exons and modularity in β-haemoglobin.

Note that exon shuffling is not the same as DOMAIN SHUFFLING. Exons and domains might not coincide. Domain shuffling is thought to be more likely to succeed than exon shuffling. Also note that an exon may be successfully shuffled if it does not correspond to a structural or functional unit; in such cases the resulting mosaic protein is even less likely to be evolutionary significant.
The mechanisms for non-homologous recombination that could drive exon shuffling are only recently being investigated and understood. Two major forms are proposed: (i) ILLEGITIMATE RECOMBINATION, where recombination occurs between sequences with little or no homology, and (ii) RETROPOSITION, the integration of a sequence derived from RNA into a DNA genome.

Assume that an exon will be inserted into an intron of another gene; its fitness effects will be impacted by its effect on the reading frame. Hence, exon shuffling is limited by the phase of the introns. If an intron is PHASE 0, it lies between two intact codons. If a foreign exon is inserted within a phase 0 intron, the reading frames of the pre-existing exons are not impacted. However, some introns lie between the 1st and 2nd positions of a codon (PHASE 1) or between the 2nd and 3rd positions of a codon (PHASE 2). If a foreign exon is inserted in either a phase 1 or a phase 2 intron, it will disrupt the reading frame. If one exon is exchanged with another, and one or both of those exons do not have intact reading frames, i.e., they are adjacent to at least one phase 1 or phase 2 intron, the reading frame will be disrupted.

Exon shuffling also is constrained by folding. The protein must be able to fold into an appropriate 3D conformation. If one or more exons are exchanged that disrupt or prevent the protein from folding, the resulting mosaic protein will not function in any way. 3D domains that are compatible with proper protein folding are called SCHEMA. Thus the compatibility of exons involved in exon shuffling is limited by the schematic composition of the proteins 3D structure.

Domain and Exon duplication:  
Note that DOMAIN DUPLICATION and EXON DUPLICATION refer to internal duplication events within a gene, and do not involve illegitimate recombination among non-homologous genes. They are included here because they represent another mechanism for mixing and matching the content of exons or domains within a protein. Combination of duplication and loss of exons can allow proteins to grow and shrink in size over time. The mechanism of change in this case is UNEQUAL CROSSING OVER.

The power of combinatorial evolution:  
The notion of modularity and evolution was first introduced in relation to regulatory signals (FD TOPIC 2A, A PRIMER ON THE STRUCTURE AND FUNCTION OF GENES). There we saw that COMBINATORIAL EVOLUTION was a very powerful way of achieving fast evolutionary change. By taking advantage of the vast combinatorial space of possibilities, through mixing and matching modules, evolution was not limited by the relatively slow process of mutation. Combinatorial evolution of regulatory elements and protein modules are examples of evolution by genetic co-option.

NOTES ON RETROPOSITION:  Retroposition is a type of TRANSPOSITION based on RNA: a common form of retroposition is the reverse transcription of an mRNA into a DNA genome. Retroposition is considered an important, genome scale, mechanism for copying genetic material. One estimate suggests that it has accounted for over 10,000 duplications in the human genome.
Starting with a template, the power of gene duplication:

The vast majority of genes are members of families or superfamilies of genes. As gene families, and hence genic diversity, arises by the process of gene duplication, this mechanism is considered a major driver of gene diversity evolution. The attractiveness of this model is that a gene duplication event results in a minimum of two copies of a fully functional gene; thereby providing one copy that can perform the original function while the alternate copy is free to evolve a new function without negatively impacting the original function of the gene product. Moreover, evolution is provided with the template of a fully functional gene with which to work; i.e. it folds properly, and perhaps has other desirable functional properties such as it associates with other proteins, it binds a ligand, or it possesses enzymatic properties. It is important to understand that not all duplication events lead to functional divergence.

The evolutionary fates of gene duplicates are often placed in one of four following categories:

1. Nonfunctionalization: mutational meltdown of one duplicate following a duplication event. This is the fate of the vast majority of new duplicates. As there are two copies in the genome, there is no selection pressure against fixing a deleterious mutation (unless a beneficial one has also occurred). As there are two copies, fixing such a deleterious mutation is not negative, as the alternate copy is capable of proving full functionality. Because the vast majority of most new mutations will be deleterious to function, nonfunctionalization is the most common evolutionary outcome of a gene duplication event. Such non-functional genes are called PSEUDOGENES.

A copy of proto $\beta$-globin initially diverges from a copy of the proto $\varepsilon$-globin gene. This is believed to be a case of NEOFUNCTIONALIZATION as the proto $\varepsilon$-globin gene evolved novel oxygen affinity.

$\eta$-globin becomes a pseudogene in stem primates (NONFUNCTIONALIZATION). This event occurred independently of an earlier conversion of $\eta$-globin to a pseudogene in rodents and lagomorphs (NONFUNCTIONALIZATION).
2. **Neofuntionalization:** This represents the evolution of a new function by fixation of one or more beneficial mutations (i.e., that confer a new or modified function). Globin genes represent a classic example of such a model of evolution. Divergence of oxygen storage and oxygen transport capabilities originated via an ancient duplication of the ancestor to modern day myoglobin and haemoglobin molecules. Subsequent functional specialization of haemoglobins into two families was mediated by the process of gene duplication (see figure above).

3. **Neospecialization:** This refers to the case where a gene has evolved divergent roles prior to a gene duplication event (Gene sharing). Perhaps it has one role in one tissue and another role in another tissue. The important point is that such a gene would be evolving under conflicting selection pressure. In such a case there would be very strong selection pressure following a duplication event for partitioning this function into two separate genes, thereby relieving the conflicting selection pressures.

An extreme example of divergent roles for a single gene is GENE SHARING. An excellent example is the ε-crystallin protein in birds and crocodiles and the LDH-B protein in the same organisms. The ε-crystallin protein functions to maintain transparency and diffraction in the eye lens (a structural role) whereas the LDH-B protein functions as a lactate dehydrogenase enzyme (a catalytic role). These two highly divergent proteins are encoded by the same gene sequence. Thus these proteins are subject to two different, conflicting, sets of evolutionary selection pressures!

4. **Subfunctionalization:** This model assumes that a protein contains functional modules that are essential to a gene's function in different tissues, or at different times of expression. Subfunctionalization occurs after a duplication event when complimentary regulatory elements degrade such that the ancestral pattern of gene expression is partitioned between the duplicates (see APPENDIX I). If this happens, some modules in one copy of the gene are no longer needed because that copy is no longer expressed in the relevant tissue, or at the relevant time. Such superfluous modules then degrade by the accumulation of mutations because these modules have been release from purifying selection pressure.

This model is attractive because it predicts that duplicates of genes with complex regulatory controls will be less likely to be lost to the process of nonfunctionalization. The larger the complexity of the involved regulatory genes, the higher the probability that subfunctionalization will occur. Because subfunctionalization is hypothesized to be a force for stabilizing the outcome of duplication events, it provides a mechanism for the slower process of beneficial mutation to occur while avoiding nonfunctionalization. Thus, once partitioned, such duplicate genes may become functionally divergent later in evolutionary time. This model does not include pre-existing conflicts in selection pressure (as in neospecialization) and hence does NOT predict rapid functional divergence following gene duplication.

The globin example above illustrates the expansion of a gene family by tandem duplications (and contraction by nonfunctionalization). Note that members of such gene families are typically found on the same chromosome and in a more or less tandem array (In humans, the β-globin family is located in a tandem array on chromosome 11; see above figure). Duplications also can occur by mechanisms that place the duplicate
gene in a new, and very different, genomic complex. Mechanisms for such duplication are RETROPOSTION and TRANSLOCATION. Such a new genomic environment undoubtedly means that the gene will be placed in a novel regulatory environment. Such genes could then be recruited for completely new functions under completely new patterns of expression. Although most such events will be evolutionary dead ends, leading to nonfuntionalization, the process of retroposition has been referred to as “sowing the seeds for novel gene function”.

**Rapid evolution of gene family content, Birth-Death evolution:**
Birth-death evolution refers to the rapid rate of gene duplication and deletion relative to the organism diversification process (See PHYLOGENETIC TOPIC 2,: PHYLOGENETIC HOMOLOGY, for a review). Such families are rapidly change their gene content, and in some cases simply changing this can have adaptive significance.
Examples are the odorant receptor genes of rodents. These genes are members of large and diverse gene families that are evolving under a birth-death process; hence they are rapidly changing content over time. Rodents use the proteins encoded by these genes to detect and distinguish between complex mixtures of many chemical signatures, including pheromones. These complex mixtures of pheromones provide and individual with information about other individuals such as; (i) if he is the same species, (ii) if he is ready to mate, (iii) if he is aggressive, (iv) if he is sick and (v) how physically fit he might be. Distinguishing among such signals is clearly beneficial, and a large family of odorant receptor genes make this possible. Changes among species are thought to be particularly adaptive, as it prevents wasting reproductive effort on the wrong species.

**Gene conversion:**
Gene conversion is a force for homogenization of genetic material (See MUTATION AND RECOMBINATION notes for a review). However, it can also be a force to repair a region of a paralogous gene that contains a deleterious mutation. Consider a duplication event where one of the duplicates has acquired both a beneficial mutation and a deleterious mutation, say a deletion that causes a frameshift thereby rendering the duplicate non-functional. A partial gene conversion event that restores the reading frame in the duplicate gene would essentially “resurrect” that gene.

**What if existing coding sequence simply can’t do the job? Is there any evolutionary hope?**
Until now all cases we have considered utilized co-option of existing genetic systems, and in particular protein-coding sequences, as a starting point for modification for a new function. What if the current genome simply does not contain a suitable template for co-option? No doubt there are many cases where such a lack of pre-existing genetic machinery constrains the rate, and perhaps the opportunities, for evolution of novelty. For example Human beings will undoubtedly NEVER evolve into angels; they lack both the developmental machinery for wings and a suitable moral character (in the spirit of a concept originally raised by J.B.S Haldane 1932).

At least two options do exist, however, for obtaining genetic material from other sources [i.e., other than protein coding genes contained by the present genome]: (i) LGT (lateral gene transfer) and (ii) frame-shift mutations.
LGT opens up the genome of other organism as possible sources of protein coding sequences that could be co-opted to new uses. Prokaryotes exhibit a high frequency of LGT events associated with adaptive evolution (remember the PATHOGENICITY ISLANDS, METABOLISM ISLANDS etc.). As these islands often comprise complete OPERONS, co-option events involve simultaneous acquisition of templates for sets of genes and their involved regulatory sequences.

Recent work has revealed that frame-shift mutations, when coupled with a second frame-shift that restores the reading frame, have resulted in expanded coding capacity of genes that can be modified, via a process of mutation and selection, to provide a gene with a new function. In particular, many genes encoding TRANSCRIPTION FACTORS and TRANSMEMBRANE PROTEINS appear to have evolved divergent functions by this mechanism.

A final note:
Although we have invoked several mechanisms for the origin of diversity (illegitimate recombination, gene duplication, transposition, LGT, etc), we need not assume anything new regarding how the genetic variation is ultimately fixed in a population (selection or drift). In all these cases the standard microevolutionary forces we covered previously in this course are sufficient to explain the evolution of novelty (by genetic co-option).