Origins of New Genes: Exon Shuffling

By Carl Hillstrom

- The talk is about how the shuffling of exons can give rise to new genes.

Merriam-Webster Online Dictionary

Main Entry: **ex·on**
Pronunciation: 'ek-"sän
Function: *noun*

: a polynucleotide sequence in a nucleic acid that codes information for protein synthesis and that is copied and spliced together with other such sequences to form messenger RNA -- compare INTRON

http://www.m-w.com/dictionary/exon

Exon shuffling

Recombination, exclusion, or duplications of exons can drive the evolution of new genes.

The general idea of exon shuffling is typically attributed to Walter Gilbert (e.g. Long et al. 2003)

The definition of exon shuffling used in this presentation encompass:

--exon assumes a new function after it has been moved

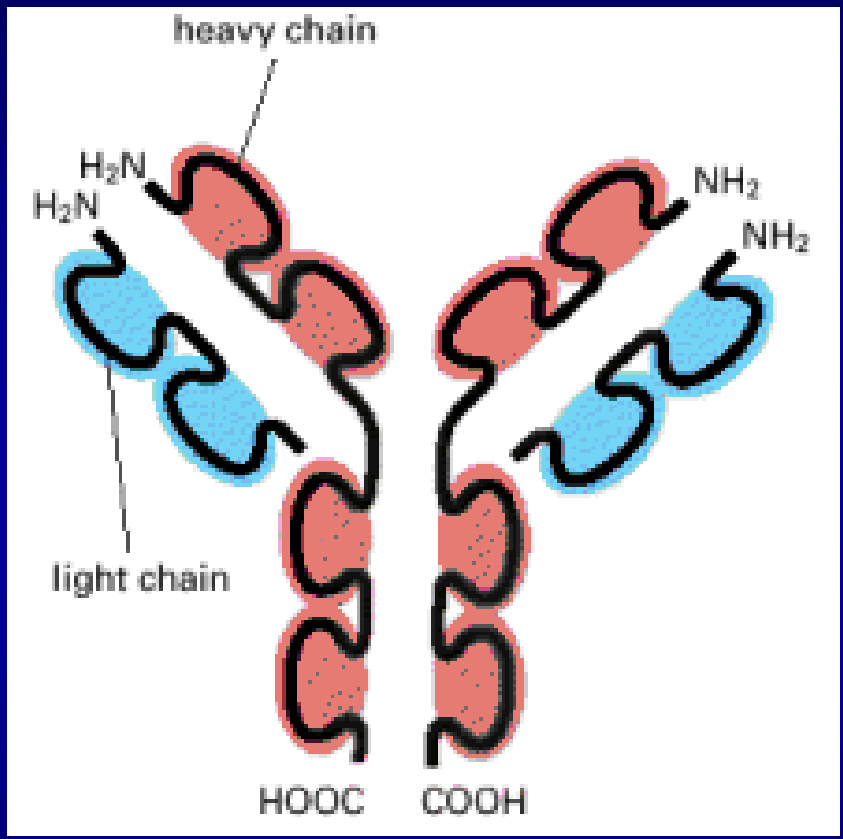
--exon retains its original function after it has been moved

- So what is exon shuffling?
- It is basically the idea that recombination or exclusion of exons can drive the evolution of new genes.

  “Recombination, exclusion, or duplications of exons can drive the evolution of new genes.” –this is a very general definition that I have adopted for the purpose of this presentation.

- The definition of exon shuffling used in this presentation encompass:
- --exon assumes a new function after it has been moved
- --exon retains its original function after it has been moved
- there is disagreement whether exon shuffling applies to both of these definition—for simplicity, I will use the concept of exon shuffling as if it applies to both of these definitions

Outline of a typical antibody



Alberts et al. (2002)

A concrete example of how exon shuffling is physiologically crucial. The immunoglobulin genes of undifferentiated carries broad coding capacity. But through deletions and rearrangements of the gene as B lymphocytes differentiate, considerable functional diversity can be conferred.

http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=mboc4.figgrp.1421

A concrete example of how exon shuffling is physiologically crucial. The immunoglobulin genes of undifferentiated carries broad coding capacity. But through deletions and rearrangements of the gene as B lymphocytes differentiate, considerable functional diversity can be conferred. This is a very simple example of exon shuffling that I think we all can relate to. I just wanted to use this antibody example to show that exon shuffling has very real implications. It is by no means an exclusively theoretical concept.

Disclaimer: This example does not meet many definitions of exon shuffling. The exon shuffling concept is mainly applied to the recombination of exons from distinct genes (Long et al. 2003).

The macroevolution connection

Comparisons of the yeast and *C. elegans* genomes have revealed that domains associated with intracellular proteins in yeast have found a place in extracellular domains in *C. elegans.*

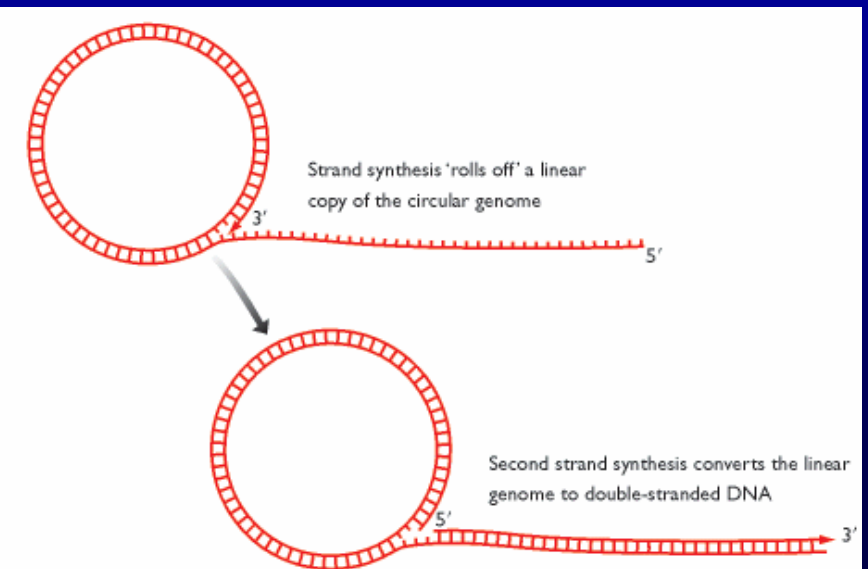
Did exon shuffling faciliate the evolution of extracellular proteins necessary for multicellularity?

(Patthy 2003)

Did exon shuffling faciliate the evolution of extracellular proteins necessary for multicellularity?—no clear example of this among plants (Patthy 2003).

Mechanisms of exon shuffling

- The basic mechanisms are believed to origins in an RNA world (Long et al. 2003b)
- Transposon mediated
  - -long-terminal repeat (LTR) retrotransposons (Wang 2006)
  - -long interspersed element (LINE)-1 (Ejima and Yang 2003)
  - -helitron like (Morgante et al. 2005)



http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=genomes.figgrp.8120

- **<u>Mechanisms of exon shuffling</u>**
- The basic function of exon shuffling, i.e. the origin new genetic material through rearrangement of already existing genetic material, is presumably very old. The ribozyme activities of certain RNA molecules are likely to have had a role in re arranging RNA genetic material in a pre DNA world

- In DNA, it is clear that transposons play vital roles in mediating sequence rearrangements

- **LTRs , LINEs** , and helitron – three types of transposons that can facilitate evolution of new genes

Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize

Michele Morgante[1], Stephan Brunner[2], Giorgio Pea[2,3], Kevin Fengler[2], Andrea Zuccolo[1,3] & Antoni Rafalski[2]

We report a whole-genome comparison of gene content in allelic BAC contigs from two maize inbred lines. Genic content polymorphisms involve as many as 10,000 sequences and are mainly generated by DNA insertions. The termini of eight of the nine genic insertions that we analyzed shared the structural hallmarks of helitron rolling-circle transposons[1–3]. DNA segments defined by helitron termini contained multiple gene-derived fragments and had a structure typical of nonautonomous helitron-like transposons. Closely related insertions were found in multiple genomic locations. Some of these produced transcripts containing segments of different genes, supporting the idea that these transposition events have a role in exon shuffling and the evolution of new proteins. We identified putative autonomous helitron elements and found evidence for their transcription. Helitrons in maize seem to continually produce new nonautonomous elements responsible for the duplicative insertion of gene segments into new locations and for the unprecedented genic diversity. The maize genome is in constant flux, as transposable elements continue to change both the genic and nongenic fractions of the genome, profoundly affecting genetic diversity.

us to identify allelic pairs of contigs covering ~67% of the total map size. In each pair of contigs, we identified shared and nonshared genes on the basis of the probe hybridization using a conservative criterion to minimize the effect of false positive and false negative hybridizations. Of the 20,656 qualifying hybridization instances, 79% were shared between the two lines, 11% were found only in line B73 and 9% were found only in line Mo17 (**Table 1**). We concluded that a large fraction (20%) of genome segments hybridizing with gene-derived probes was not shared between inbred lines B73 and Mo17, in agreement with observations from fully sequenced genomic regions[4–6]. Assuming that maize has at least 40,000 functional shared genes (on the basis of previous estimates for rice[9] and maize[10,11]), ~10,000 genes or gene fragments are not shared between the two lines. The fraction of nonshared sequences varies widely among different contigs, even when considering only large contigs (**Fig. 1**). We observed no substantial skewing toward regions with or without a high proportion of nonshared genes, but 8% of the contigs shared all genes.

Analysis of the nonshared genic segments in the five sequenced regions showed that they correspond to fragments of genes (pseudogenes) usually present in clusters. Gene fragments in clusters tend to have the same orientation with respect to the direction of transcrip-

1. this is a paper from last year, characterizing helitron type transposons in corn. I've underlined a key point in the abstract

Mechanisms of exon shuffling (cont'd)

- Crossover during sexual recombination of parental genomes

  -exons favored (Kolkman and Stemmer 2001)

- Gene fusion/fission, lateral gene transfer, non-homologous recombination-- (van Rijk and Bloemendal 2003)

- Crossover during sexual recombination of parental genomes
  - -exons favored
  - In humans, exons occupy 1% of the genome and introns occupy 24%--yet, far more crossovers occur between exons

(Kolkman and Stemmer 2001)

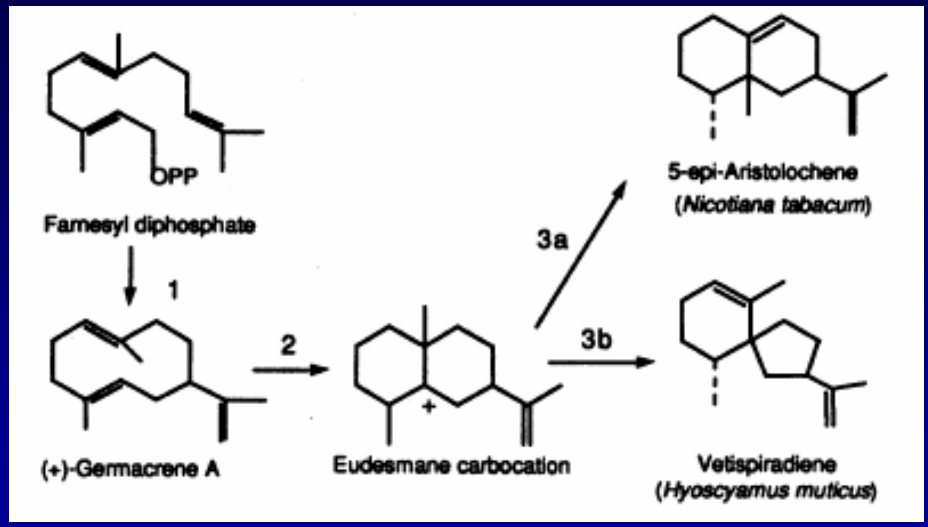
The study of exon shuffling as an evolutionary driving force

- Highly bioinformatics driven
  - one can look for duplications, retrotranspositions, transposable elements, etc.

- Genetic engineering approaches to trace evolutionary developments

Hyoscyamus muticus L.

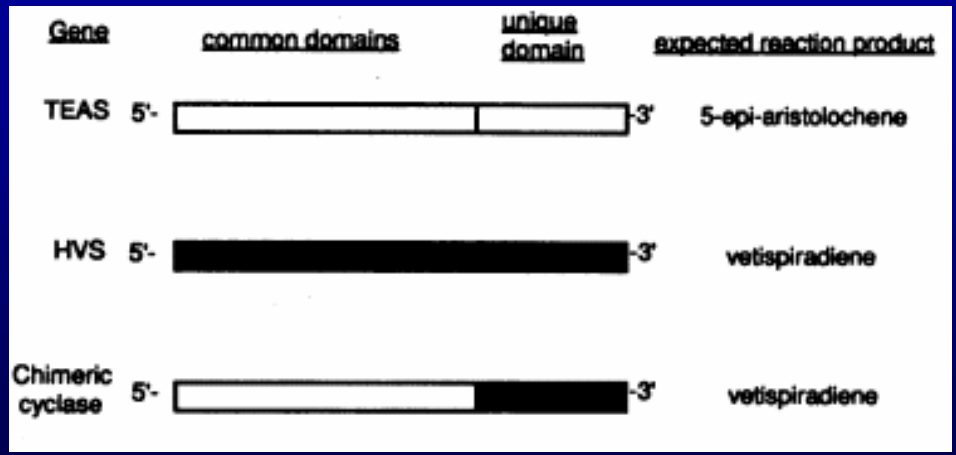
http://www.giftpflanzen.com/hyoscyamus_muticus.html



Nicotiana tabacum L.

http://www.luciolongo.it/semi%20e%20piante/immagini/nicotiana%20tabacum/nicotiana%20tabacum%20virginia.jpg

TEAS is from *Nicotiana tabacum*



HVS is from *Hyoseyamus muticus*

- Examples of how genetic engineering approaches could help in tracing evolutionary developments

- Back and Chappell (1996) demonstrated that distinct exons were responsible for the products synthesized by the otherwise highly similar *Nicotiana tabacum* 5-epi-aristolochene synthase and the *Hyoscyamus muticus* vetispiradiene synthase genes, suggesting that exon shuffling of a common ancestor gene may have given rise to functionally distinct lineages.

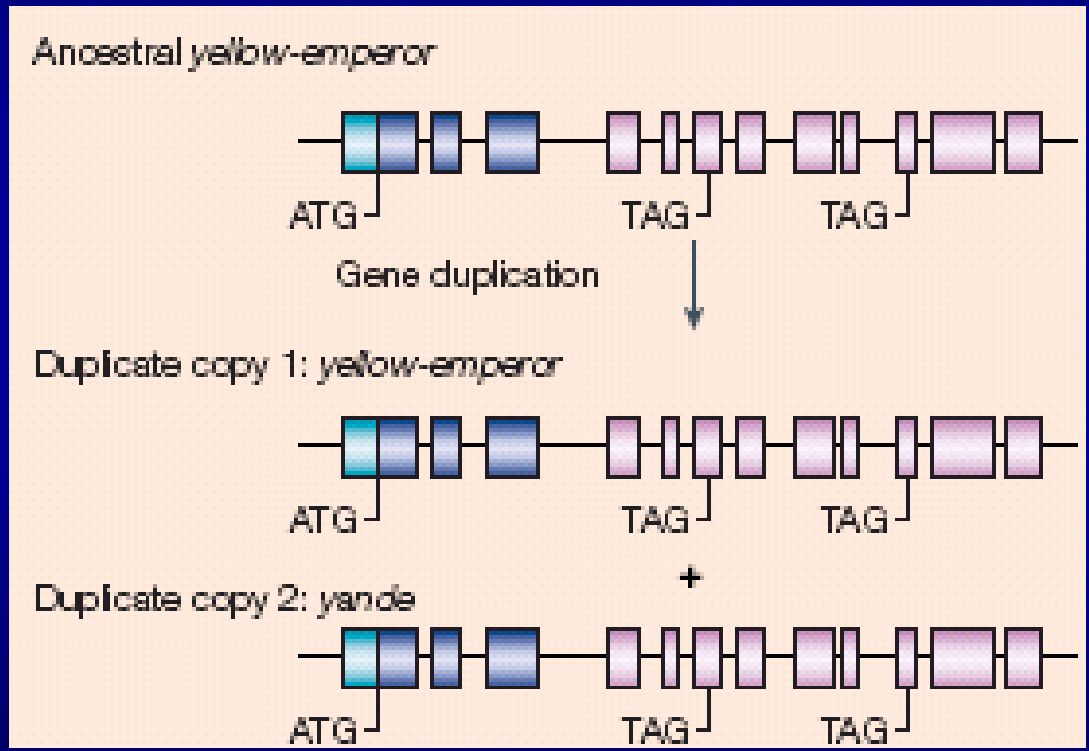
- Disclaimer regarding the appropriateness of the exon shuffling concept may apply (see slide 8).

The *jingwei* example

This chimeric gene arose 2.5 million years ago in the common ancestor of two African *Drosophila* species, *Drosophila yakuba* and *Drosophila teissieri*

An ancestral species have single copies of *yellow-emperor* (*ymp*) and the alcohol dehydrogenase encoding gene *Adh*.

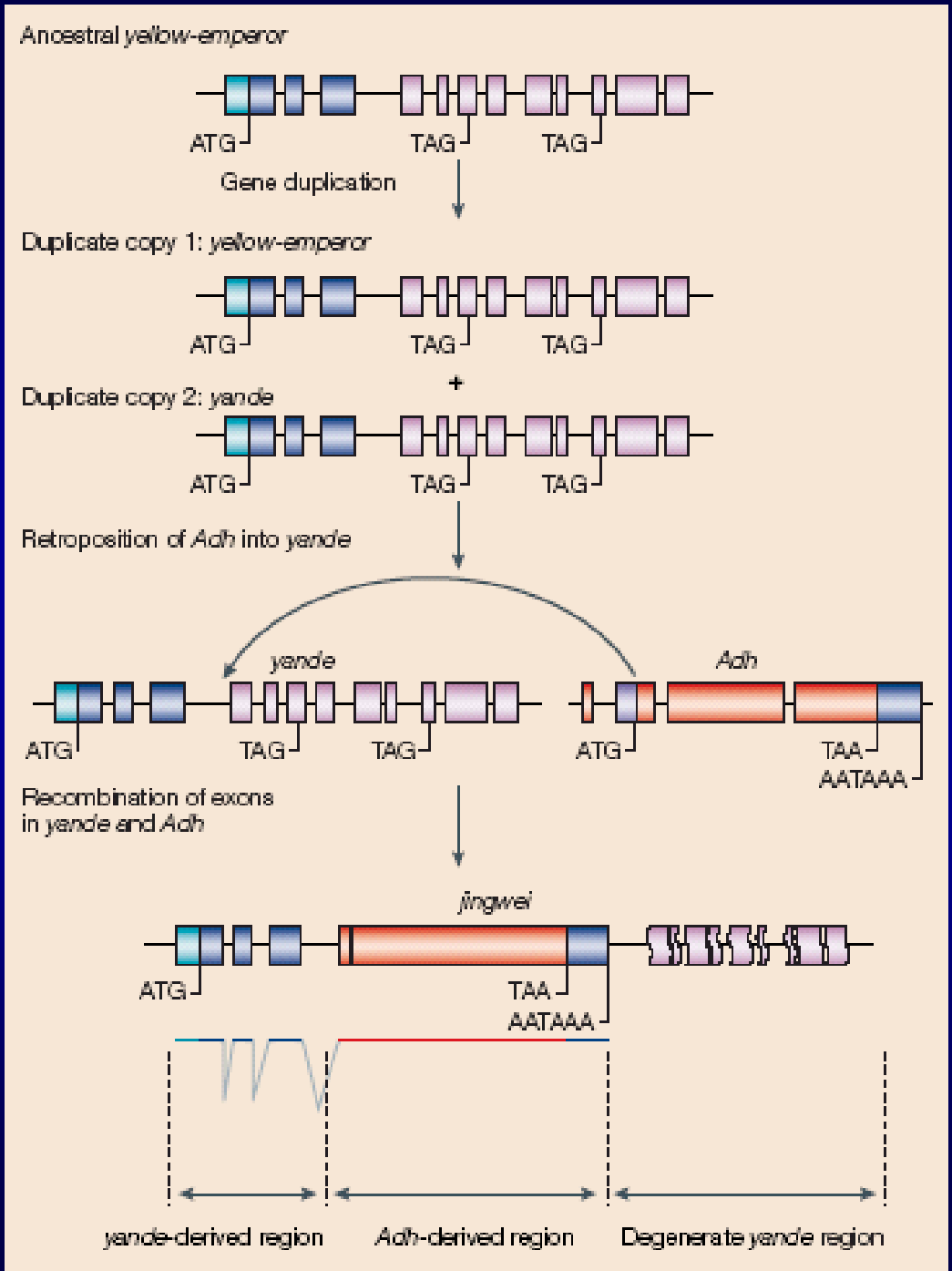
Ymp was duplicated; one copy retained the original name (*Ymp*) while the other was called *yande* (*ynd*). The copy with the original name (*Ymp*)



- Figure in slides 19 and 21comes from Long et al. (2003). I have taken the liberty to cut it up.

Adh mRNA retroposed into the third intron of *yande* as a fused exon and recombined with the first three *yande* exons.

Adh terminate the readthrough transcription and downstream *yande* exons degenerate.



"The origin of *jingwei* has highlighted the creative roles of several molecular processes acting in combination: exon shuffling, retroposition and gene duplication."

- **The *jingwei* example**
- This chimeric gene arose 2.5 million years ago in the common ancestor of two African *Drosophila* species, *Drosophila yakuba* and *Drosophila teissieri*

- Quote from Long et al. (2003):
- Long et al. (2003):
- “The origin of jingwei has highlighted the creative roles of several molecular processes acting in combination: exon shuffling, retroposition and gene duplication.”

Retrotransposition implicated in yielding new genes in rice (*Oryza sativa*)

- Several hundred retrogenes identified in the rice genome (898 are defined as intact retrogenes—not pseudogenes, and more than half of them have been found to be expressed with support of either full-length cDNAs, ESTs, microarray analysis, or RT-PCR
- Many have chimerical structures (like *jingwei*)

Wang et al. (2006)

Distribution of intron phases --indication of exon shuffling?

Table 3 | **Intron-phase correlation in eukaryotic genomes**

	Symmetrical			Asymmetrical					
	(0,0)	(1,1)	(2,2)	(0,1)	(0,2)	(1,2)	(1,0)	(2,0)	(2,1)
Observed number*	3,051	1,303	620	1,321	1,184	749	1,408	1,219	704
Expected number‡	2,709	1,013	558	1,657	1,230	752	1,657	1,229	752

*The frequencies in an exon database extracted form GenBank. ‡Calculated as a product of $E(i,j) = Pi \times Pj \times N$, assuming that the association of two introns in the same gene is random: Pi is the proportion of intron phase i actually observed ($P_0 = 0.48$; $P_1 = 0.30$; $P_2 = 0.22$); P_0, P_1 and P_2 are the frequencies of phase zero introns (between two codons), phase one introns (after the first nucleotide within a codon) and phase two introns (after the second nucleotide in a codon); N is the total observed number of intron associations (i,j) (N = 11,559). When i = j the association is called *symmetrical exon*; when i ≠ j the association is called asymmetrical exon. The observed intron-phase frequencies are significantly different from the expected distribution. Modified with permission from REF. 122 © (2000) Nature Publishing Group.

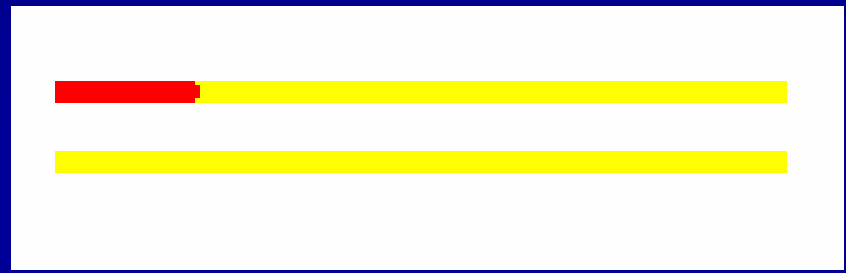
- Biased intron insertions?
  - unlikely
- Signatures of exon shuffling?
  - more likely
  - length of an inserted exon should be a multiple of three
  - insertion of a symmetric exon into an intron of the same phase does not disrupt the reading frame

INTRON PHASE
The relative position of an intron within or between codons. Phase zero, one and two are defined by the position of an intron between two codons or after the first or second nucleotide of a codon, respectively.

Long et al. (2003)

Exon shuffling in prokaryotic genes?

- Long et al. (1995) investigated 296 intron phase correlations using introns lying in the region of match between eukaryotes and prokaryotes.
- Eukaryotic intron sequences are obviously not intron sequences in prokaryotes
- 55% phase zero introns, 24% phase one introns, 21% phase two introns



No footprints of primordial introns in a eukaryotic genome

The presence of introns in the protein-coding genes of eukaryotes as opposed to their absence in those of archaea and bacteria has been explained in two ways. The 'introns-early' hypothesis holds that most, if not all pro-

into the eukaryotic nuclear genome from bacterial genomes, primarily those of endosymbionts (mitochondria and chloroplasts), and newly invented genes[8]. Because mitochondria and chloroplasts have evolved from distinct

Footprints of primordial introns on the eukaryotic genome

A recent paper in this journal[1] presented an analysis of PHASE DISTRIBUTION OF INTRONS (see Glossary) in *Caenorhabditis elegans* with respect to the EXON THEORY OF

recent ones. Here, we show this reanalysis and, furthermore, provide new evidence for the existence of ancient phase-zero introns: those intron positions identified as putatively ancient by virtue of a wide phylogenetic distribution lie preferentially in phase zero.

Several years ago, Long *et al.*[3] argued that the skewed phase distribution (more introns in phase zero than in phase one or two) in ancient conserved genes could be

Footprints of primordial introns on the eukaryotic genome: still no clear traces

Response from Yuri I. Wolf, Fyodor A. Kondrashov and Eugene V. Koonin

The slightly lower fraction of PHASE-ZERO INTRONS (see Glossary) and the slightly lower excess of SYMMETRICAL EXONS over non-symmetrical exons in *Caenorhabditis elegans* genes that are thought to be recent transfers from prokaryotes compared with ANCIENT GENES has been proposed to be evidence for persistence of PRIMORDIAL INTRONS in ancient genes. We show here that there is no significant difference, by both of

Article

Phylogenetically Older Introns Strongly Correlate With Module Boundaries in Ancient Proteins

Alexei Fedorov,[1,2] Scott Roy,[1] Xiaohong Cao,[1,3] and Walter Gilbert[1,4]

[1] *Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA*

The hypothesis that some (but not all) introns were used to construct ancient genes by exon shuffling of modules at the earliest stages of evolution is supported by the finding of an excess of phase-zero intron positions in the boundary regions of such modules in 276 ancient proteins (defined as common to eukaryotes and prokaryotes). Here we show further that as phase-zero intron positions are shared by distant taxa, and thus are truly phylogenetically ancient, their excess in the boundaries becomes greater, rising to an 80% excess if shared by four out of the five taxa: vertebrates, invertebrates, fungi, plants, and protists.

We recently studied the distribution of introns in homologs of 276 ancient unrelated proteins of known three-dimensional structure and found a significant, but small, excess of phase-zero intron positions in the boundary regions of modules 15–35 Å in diameter (Fedorov et al. 2001). This correlation holds only for phase-zero introns, which lie between codons, and not for phase-one or phase-two introns, lying

can be no exon shuffling in their history, because the eukaryotic forms are colinear to the prokaryotic sequence. However, in an introns-early model (Gilbert 1987), some or all of these introns might be left over from the exon-shuffling events that created the gene before the separation of prokaryotes and eukaryotes. This picture of the modular substructure of domains being created by exon shuffling using phase-zero introns is

Examples of competing views

Evolution of new genes by other mechanisms

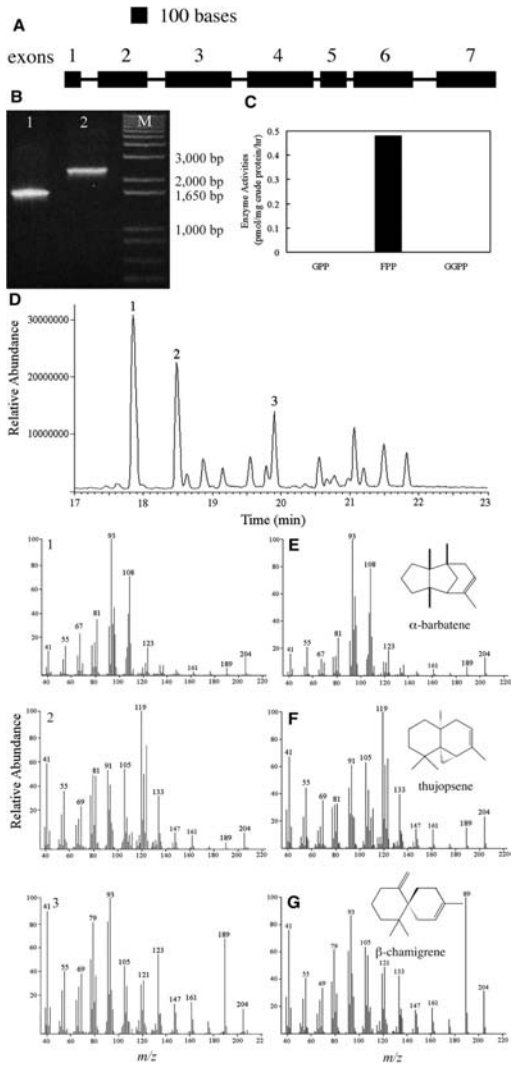
- de novo recruitment of exons from intronic regions
- de novo recruitment of exons for 5’ untranslated regions

Zhang and Chasin (2006)

If any of you attended Dr. Chasin's presentation (BIO dept. seminar, fall 2006), you know that he generally sounded somewhat skeptical about the idea of exon shuffling. He argues for some other ideas regarding how genes may have evolved.

http://www.anbg.gov.au/fungi/images-captions/marchantia-gemma-0259.html

AT5G44630 locus (Arabidopsis) --an α-barbatene synthase gene



Barbatene synthase genes

-not previously found in higher plants

-barbatene previously mainly considered a characteristic of *Bazzania*

http://www.science.siu.edu/landplants/Hepatophyta/images/Bazzania.dor.JPEG

Wu, S., et al. Plant Physiol. 2005;138:1322-1333

Liverwort and tracheophyte terpene synthase comparisons

- Terpene synthase evolution
  - origins of terpene synthases uncertain
  - liverworts evolutionary conserved
- Liverworts as progenitors of tracheophytes?
  - what would liverwort and tracheophyte terpene synthase homology suggest regarding the evolution of terpene synthases among tracheophytes?
- Synthesis specificity
  - single or multiproduct?
  - what determines this?
  - what determines enantiomeric orientation?
- Distinct exons determining products?

- As noted in my previous reference to a study my advisor and many of his collaborators have been involved in trying to determine what certain exon means for the function of enzymes. My project is also centered around this general theme, and it will require that I consider issues pertaining to the evolution of how exons are distributed.

Literature cited

- Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. 2002. Molecular Biology of the Cell - Fourth Edition. *On-line version*. http://www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowTOC&rid=mboc4.TOC&depth=10

- Back, K. and J. Chappell. 1996. Identifying functional domains within terpene cyclases using a domain-swapping strategy. *Biochemistry* 93: 6841-6845.

- Fedorov, A., S. Roy, X. Cao, and W. Gilbert. 2003. Phylogenetically Older Introns Strongly Correlate With Module Boundaries in Ancient Proteins. *Genome Research* 13: 1155-1157.

- Ejima, Y. and L. Yang. 2003. Trans mobilization of genomic DNA as a mechanism for retrotransposon-mediated exon shuffling. *Human Molecular Genetics* 12: 1321-1328.

- Kolkman, J. A. and Stemmer, W. P.C. Directed evolution of proteins by exon shuffling. *Nature Biotechnology* 19: 423-428.
- Long, M., E. Betrán, K. Thornton, and W. Wang. 2003. The origin of new genes: glimpses from the young and old. *Nature Reviews Genetics* 4: 865-875.

- Long, M., M. Deutsch, W. Wang, E. Betrán, F. G. Brunet, and J. Zhang. 2003b. Origin of new genes: evidence from experimental and computational analyses. *Genetica* 118: 171–182, 2003.
  —it is admittedly a little unclear whether it is appropriate to reference this paper here. I am not sure if they are citing Gilbert (1987), or if they are putting forth an idea based on their interpretation of another idea put forth by Gilbert (1987). I have not been able to locate Gilbert (1987).

- M. Long, Rosenberg, C., and Gilbert, W. Intron phase correlations and the evolution of the intron/exon structure of genes. *Proceedings of the National Academy of Sciences* 92: 12495–12499.

- Morgante, M., S. Brunner, G. Pea, K. Fengler, A. Zuccolo, and A. Rafalski. 2005. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nature Genetics* 37: 997-1002.

- Patthy, L. Modular assembly of genes and the evolution of new functions. *Genetica* 118: 217–231, 2003.

- Roy, S. W., B. P. Lewis. A. Fedorov, W. Gilbert. 2001. Footprints of primordial introns on the eukaryotic genome. *Trends in Genetics* 17: 496-499.

- van Rijk, A. and H. Bloemendal. 2003. Molecular mechanisms of exon shuffling: illegitimate recombination. *Genetica* 118: 245–249.

- Wang, W., H. Zheng, C. Fan, J. Li, J. Shi, Z. Cai, G. Zhang, D. Liu, J. Zhang, S. Vang, Z. Lu, G. Ka-Shu Wong, M. Long, and J. Wang. High Rate of Chimeric Gene Origination by Retroposition in Plant Genomes. *The Plant Cell* 18: 1791–1802.

- Wolf, Y. I., F. A. Kondrashov, E. V. Koonin. 2000. No footprints of primordial introns in a eukaryotic genome. *Trends in Genetics* 16, 333-334.

- Wolf, Y. I., F. A. Kondrashov, E. V. Koonin. 2001. Footprints of primordial introns on the eukaryotic genome: still no clear traces. *Trends in Genetics* 17: 499-501

- Zhang, X. H.-F. and L. A. Chasin. 2006. Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. *Proceedings of the National Academy of Sciences* 103: 13427-13432.

All URLs are given in the presentation links to the pictures used or Merriam-Webster's on-line dictionary.