Genome Evolution

Outline

1. **What**: Patterns of Genome Evolution

2. **Why**: Evolution of Genome Complexity and the interaction between Natural Selection and Genetic Drift

Genome

- **Definition**: The entirety of an organism’s hereditary information, encoded in DNA or in RNA, for many types of viruses. The genome includes both the genes and non-coding sequences of the DNA/RNA

“New Evolutionary Synthesis”

- Comparative genomics has the potential to measure the strength of constraints on different classes of sites in genomes and to elucidate the biological nature of these constraints.
- Genome comparisons also help to address higher-level questions, including the degree to which constraints act on gene repertoires, genome architecture and the evolution rate itself.
- The avalanche of systems biology data allows researchers to ask new, qualitative questions, such as how do constraints affect metabolic fluxes and the ‘molecular phenome’

Genome Architecture

- The totality of non-random arrangements of functional elements (genes, regulatory regions etc.) in the genome
- Genome architecture is highly variable across taxa

Constraints on Genome Evolution

- Most protein coding genes evolve under purifying selection of widely varying strength, to preserve protein function, and few evolve under positive selection
- Strength of constraints also differs between coding + non-coding but regulatory sequences versus non-coding (and non-regulatory) sequences
- **Note that non-coding sequence includes both nonfunctional and functional (e.g. regulatory) regions**
Purifying Selection

- = Negative Selection

Acts to select out deleterious mutations in coding or regulatory sequences

Tends to maintain constancy of DNA sequences... More constant than would be expected if genetic drift were acting alone ($K_s/K_a < 1$)

General Principles

- Most conserved feature of Prokaryotes is the operon

Gene Order: Prokaryotic gene order is not conserved (aside from order within the operon), whereas in Eukaryotes gene order tends to be conserved across taxa

Intron-exon genomic organization: The distinctive feature of eukaryotic genomes that sharply separates them from prokaryotic genomes is the presence of spliceosomal introns that interrupt protein-coding genes

Evolutionary constraints acting on different regions of the genome

1. Least constrained: “Junk DNA” – noncoding and nonfunctional sequence
2. Less constrained: Introns - noncoding, but affect splice sites in genes and exon shuffling
3. Weaker constraint: Noncoding, but regulatory sequence
4. Most constrained: Sequences encoding structural RNAs and nonsynonymous sites in protein-coding sequences

Comparison of Genome Size:

Bacteria
E. coli (4.6MB)
Minibacillus
(0.2MB)
H. influenzae
(1.8MB)
Archaebacteria
Methanococcus
(1.0MB)
Fungi
Yeast
(6.6MB)
Mammals
H. sapiens
(3.0GB)
Comparative Genomics

Genome Architecture

Fig. 1. The distinct genome architectures of prokaryotes, unicellular eukaryotes, and multicellular eukaryotes. Genes are shown by rectangles with the arrowhead indicating the direction of transcription. Exons are shown by rectangles without arrowheads. Intergenic regions are shown by straight lines, and introns are shown by sacligly lines. The schematic is not to scale. (For interpretation of the references to color in this artwork, the reader is referred to the web version of the article.)
Small vs. Large Genomes

1. **Compact**, relatively small genomes of viruses, archaea, bacteria (typically, <10 Mb), and many unicellular eukaryotes (typically, <20 Mb). In these genomes, protein-coding and RNA-coding sequences occupy most of the genomic sequence.

2. **Expansive**, large genomes of multicellular and some unicellular eukaryotes (typically, >100 Mb). In these genomes, the majority of the nucleotide sequence is non-coding, and contain introns, transposons, etc.

Genome Compactness and Constraint

- Evolutionary constraints on compact genomes, particularly those of prokaryotes, are much stronger than the constraints on the genomes of multicellular eukaryotes (median $K_a/K_s$ values for prokaryotes and multicellular organisms are typically 0.01–0.1 and 0.1–0.5, respectively).
- In viruses and prokaryotes, nearly all genomic sites are evolutionarily constrained, as most of the genome is functional.

Genome Compactness and Constraint

- Noncoding regions constitute only 10–15% of the genomes of most free-living prokaryotes, and a considerable fraction of these sequences encompasses regulatory elements that are substantially constrained in their evolution.
- The genomes of most viruses are even more compact, with almost all of the genome sequence taken up by protein coding genes.

Multicellular Eukaryotes

- The estimated fractions (%) of constrained nucleotides in a genome differ substantially even between animals.
- In *Drosophila melanogaster*, ~70% of sites in the genome, including ~65% of the non-coding sites, seem to be subject to selection, whereas in mammals this fraction is estimated at only 5–6% or even ~3%.
- The absolute numbers of sites subject to selection in these animal genomes of widely different size are quite close.

Gene Order Conservation

- More synteny in Eukaryotes than in Prokaryotes.
- Comparison of gene orders between eukaryotic genomes reveals considerable conservation of synteny over long evolutionary spans (hundreds of millions of years), e.g., among vertebrates or insects.

**Gene size vs # Protein-coding Genes**

*Fig. 2.* Genome size versus the number of protein-coding genes for selected organisms from different kingdoms of life. The plot is on a log-logarithmic scale. Black squares, large DNA viruses; green diamonds, bacteria; red diamonds, archaea; blue circles, unicellular eukaryotes (including fungi); orange circles, multicellular eukaryotes (plants and animals). The data were from the NCBI Genome Project database (NCBI genomes, 2008). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)
Introns

- The average number of introns per gene in most multicellular species is 4-7, whereas the average number for most unicellular eukaryotes is less than two.

- Below a threshold genome size of 10 Mb, introns are very rare and above 10 Mb, they approach an asymptote of about seven per gene.

Genome Size Variation

- Vertebrate genomes are "veritable junkyards of selfish genetic elements where only a small fraction of the genetic material is dedicated to encoding biologically relevant information." (Koonin 2009, your reading)

- Microbial genomes are more compact, with most of the genetic material assigned to distinct biological functions.

Why???

- So what might be the cause of genome size variation?

- Why do larger organisms, on average, tend to have larger and more complex genomes?
Genome Size Variation

Genome size generally increases as body size increases.

In general, larger organisms tend to have larger genomes and greater genome complexity (though there are exceptions).

As we go to higher trophic levels (up the food web), organisms become larger and fewer (smaller effective population size).

Population size declines at higher trophic levels.

Body size scales inversely with population size

- Large animals are less abundant (Damuth 1981)

In general, total biomass declines ~10-fold with increasing trophic level and because average body size increases at higher levels in the food chain, total population size must decline even more sharply.

So what are the evolutionary consequences of this inverse relationship between body size and population size?
Recall our Discussion of Natural Selection in the presence of Genetic Drift…

Genetic Drift and Natural Selection

- Because of the randomness introduced by Genetic Drift, Natural Selection is less efficient when there is genetic drift
- Thus, Natural Selection is more efficient in larger populations, and less effective in smaller populations

The Origins of Genome Complexity


Lynch & Connery’s argument on causes of evolution of genome size and complexity

- Transitions from prokaryotes to unicellular eukaryotes to multicellular eukaryotes are associated with orders-of-magnitude reductions in population size
- Reduced population size increases the power of Genetic Drift, weakening the effect of natural selection to remove various genomic features that would tend to proliferate (such as transposons, introns, gene families) – purifying selection

Lynch & Connery’s argument on causes of evolution of genome size and complexity

- Thus, purifying selection would tend to be more effective in organisms with large population size, which are also organisms that tend to be small in body size
- This action of purifying selection would result in smaller and more efficient genomes in organisms that have large population size
- Negative selection or purifying selection is the selective removal of alleles that are deleterious. New alleles that arise that alter the phenotype would be purged
Prokaryotes: \( N_e \) is generally \( >10^8 \)

Unicellular eukaryotes: \( \sim 10^7 - 10^8 \)

Invertebrates: \( 10^4 - 10^6 \)

Vertebrates: \( 10^4 - 10^5 \)

Larger organisms have greater genome complexity

- Larger organisms with small population sizes have much more “complex” genomes, with introns, vast amounts of non-coding DNA, transposons, etc.

- Because in smaller populations, natural selection would be less efficient, and less likely to take out new mutations that arise, even if they might be mildly deleterious

The implication is that organisms that rise in trophic level in a food web would tend to acquire a more complex genome (due to lower efficiency of selection acting)

Duplicated genes last longer in smaller populations

- There is a clear tendency for the half-life of duplicate genes to increase with genome size, again with a continuous transition between prokaryotes and eukaryotes (Fig. 2).

- Thus, by correlation, the ability of a newly arisen gene to survive the accumulation of mutations increases with decreasing effective population size.

Duplicated genes last longer in smaller populations

- Much of the increase in gene number in multicellular species may not have been driven by adaptive processes

- But, rather as a passive response to reduced population size (and reduced purifying selection) more conducive to duplicate-gene preservation by subfunctionalization (the subfunctionalized copies would all have to be retained)

Introns

- The average number of introns per gene in most multicellular species is 4-7, whereas the average number for most unicellular eukaryotes is less than two.

- Below a threshold genome size of 10 Mb, introns are very rare and above 10Mb, they approach an asymptote of about seven per gene
Transposon abundance increases with genome size

**Example**: Genome Composition in Humans

- An example of genome architecture of a larger organism, which had a small population size for most of its evolutionary history (~100,000) – (our current huge population size is an anomaly, atypical of organisms our size)

Components of the Human Genome

- Less than 1.5% of the human genome consists of the suspected ~30,000 protein-coding sequences. By contrast, a large majority is made up of non-coding sequences such as introns (almost 26%) and (mostly defunct) transposable elements (nearly 45%).

Synergy between sequence and size in Large-scale genomics


Features that arose by accident could then be subjected to selection→ opportunities for evolutionary novelty

- Although the mechanisms responsible for the initial restructuring of eukaryotic genomes may have been nonadaptive in nature, this would not preclude the new features from undergoing selection and then contributing to phenotypic evolution
- Introns sustained a reliable mechanism for alternative splicing, and in at least some lineages, they provide an orientation mechanism for the surveillance of defective mRNAs
- Subfunctionalization of duplicated genes provides a mechanism for eliminating pleiotropic constraints on ancestral genes, thereby opening up previously inaccessible evolutionary pathways

Some Critiques of Lynch’s Argument

- Did not take phylogenetic history into account; the negative correlation between effective population size and genome size might simply be a result of closely related taxa having more similar genome size. (although, archaea and bacteria having similar architecture supports his argument)
- There are probably other factors also operating… For example, parasitic species often have small effective population sizes but also have small genomes
Questions:
• Despite other potential factors that could contribute to the evolution of genome size
• Lynch’s argument remains a very useful null model against which to test for evidence of other factors affecting genome size evolution

Questions:
• What are the relationships between body size, population size and genome size and architecture?
• What are the potential causes of the evolution of genome size and architecture?
• What are some key differences in architecture between viral, prokaryotic, and eukaryotic genomes?
• Why should bacterial and archaeal genomes share genomic features when they are not evolutionarily more closely related to one another than to eukaryotes?

Questions:
• What are some distinctive features of prokaryotic genomes?
• Why is gene order not conserved in prokaryotes?
• Why do eukaryotes have a lot of introns, transposons, etc.?
• What is purifying selection, and what does it have to do with genome size evolution?

1. When comparing DNA sequences that encode a protein between two species, the ratio of substitutions at nonsynonymous relative to synonymous sites was found to be higher than the ratio of nonsynonymous relative to synonymous polymorphic sites. This result provided evidence for:
   (a) Non-adaptive evolution
   (b) Adaptive evolution
   (c) Negative selection
   (d) Evolutionary constraint
   (e) Preferential fixation of synonymous sites

2. Genome size differences from prokaryotes to multicellular eukaryotes are mostly attributable to:
   (a) The amount of "junk" (non-coding and non-functional) DNA in the genome
   (b) The amount of coding sequences
   (c) The number of rRNA coding genes
   (d) "Junk" DNA being selectively advantageous in prokaryotes
   (e) "Junk" DNA being selectively advantageous in eukaryotes

3. In very large populations, genomic elements that comprise "Junk DNA" tend to be:
   (a) Selectively advantageous
   (b) Selectively neutral
   (c) Selectively removed (via purifying selection)
   (d) Pleiotropic
   (e) Under epigenetic control
5. Which of the following is most TRUE regarding the evolution of genome architecture?
(a) Over evolutionary time, natural selection would tend to favor the evolution of larger and more complex genomes.
(b) Bacterial operons are analogous to the intron-exon organization in eukaryotes.
(c) Eukaryotic genomes tend to have introns, transposons, and other non-coding genomic elements due to the larger body sizes of eukaryotes.
(d) Much of viral and bacterial genomes are under greater evolutionary constraint (than eukaryotes) because a greater proportion of their genomes consist of coding sequences, which experience purifying selection.
(e) Gene order tends to be highly conserved in bacteria (compared to eukaryotes) due to operons.

6. Which of the following is LEAST likely to lead to the evolution of increased genome size?
(a) Gene duplications
(b) Small population size
(c) Transposons
(d) Genetic Drift
(e) Purifying selection

answers
- 1B
- 2A
- 3C
- 5D
- 6E

1. Genome size differences from prokaryotes to eukaryotes are mostly attributable to:
(a) The amount of "junk" (non-coding and non-functional) DNA in the genome
(b) The amount of coding sequences
(c) The number of tRNA coding genes
(d) "Junk" DNA being selectively advantageous in prokaryotes
(e) "Junk" DNA being selectively advantageous in eukaryotes

2. Which of the following is most TRUE regarding the evolution of genome architecture?
(a) Over evolutionary time, natural selection would tend to favor the evolution of larger and more complex genomes.
(b) Bacterial operons are analogous to the intron-exon organization in eukaryotes.
(c) Eukaryotic genomes tend to have introns, transposons, and other non-coding genomic elements due to the larger body sizes of eukaryotes.
(d) Much of viral and bacterial genomes are under greater evolutionary constraint (than eukaryotes) because a greater proportion of their genomes consist of coding sequences, which experience purifying selection.
(e) Gene order tends to be highly conserved in bacteria (compared to eukaryotes) due to operons and horizontal gene transfer.

- 1A
- 2D
Components of the Human Genome

- **Protein-coding genes**: Although most prokaryotic chromosomes consist almost entirely of protein-coding genes, such elements make up a small fraction of most eukaryotic genomes (see figure). As a prime example, the human genome might contain as few as 20,000 genes, comprising less than 1.5% of the total genome sequence.

- **Introns**: Shortly after their discovery, the non-coding intervening sequences within coding genes (introns) were suggested to account for the pronounced discrepancy between gene number and genome size. It has also recently been suggested that most non-coding DNA in animals (but not plants) is intronic, which would imply that most of the genome is transcribed even though protein-coding regions represent a tiny minority. At the very least, introns were found to account for more than a quarter of the draft human sequence. Over a broad taxonomic scale, intron size and genome size are positively correlated, although within genera a correlation might (for example, Drosophila) or might not (for example, Gossypium) be observed.

- **Pseudogenes**: Non-functional copies of coding genes, the original meaning of the term 'junk DNA', were once thought to explain variation in genome size. However, it is now apparent that even in combination, 'classical pseudogenes' (direct DNA to DNA duplicates), 'processed pseudogenes' (copies that are reverse transcribed back into the genome from RNA and therefore lack introns) and 'Numts' (nuclear pseudogenes of mitochondrial origin) comprise a relatively small portion of mammalian genomes. The human genome is estimated to contain about 19,000 pseudogenes.

- **Transposable elements**: In eukaryotes, transposable elements are divided into two general classes according to their mode of transposition. Class I elements transpose through an RNA intermediate. This class comprises long interspersed nuclear elements (LINEs), endogenous retroviruses, short interspersed nuclear elements (SINEs) and long terminal repeat (LTR) retrotransposons. Class II elements transpose directly from DNA to DNA, and include DNA transposons and miniature inverted repeat transposable elements (MITEs).